Importance Sampling and Rare Event Simulation

Pierre L'Ecuyer, Michel Mandjes, and Bruno Tuffin

June 7, 2008

Chapter 1

Importance Sampling and Rare Event Simulation

1.1 Introduction

As described in the introductory chapter, crude (also called standard, or naive) Monte Carlo simulation is inefficient for simulating rare events. Recall that crude Monte Carlo consists in considering a sample of n independent copies of the random variable or process at hand, and estimating the probability of a rare event by the proportion of times the rare event occurred over that sample. The resulting estimator can be considered useless when the occurrence probability γ is very small, unless n is much larger than $1/\gamma$. Indeed, if for instance $\gamma = 10^{-9}$, a frequent target in rare event applications, this would require on average a sample of size $n = 10^9$ to observe just a single occurrence of the event, and much more if we expect a reliable estimation of the mean and variance to obtain a confidence interval.

Importance Sampling (IS) has come up in the literature as a powerful tool to reduce the variance of an estimator, which, in the case of rare-event estimation, also means increasing the occurrence of the rare event. The generic idea of IS is to change the probability laws of the system under study to sample more frequently the events that are more "important" for the simulation. Of course, using a new distribution results in a biased estimator if no correction is applied. Therefore the simulation output needs to be translated in terms of the original measure; this is done by multiplication with a so-called likelihood ratio. IS has received substantial theoretical attention,

see [13, 12, 27], among many others, and in the rare event context, [15] or the more up-to-date tutorial [16].

IS is one of the most widely used variance reduction technique in general, and for rare event estimation in particular. Typical and specific applications will be more extensively described in the second part of the book. The goal of this chapter is to give an overview of the technical framework and the main underlying ideas. It is organized as follows. Section 1.2 reviews the very basic notions of importance sampling. It describes what the ideal (zero-variance) estimator looks like, and why it is, except in situations where simulation is not needed, infeasible to implement it exactly. That section also provides illustrative examples and outlines some properties leading to a good IS estimator, the main message being that the zero-variance estimator has to be approximated as closely as possible. In Section 1.3 the focus is on application of IS in the context of a Markov chain model. Since every discrete-event simulation model can be seen as a Markov chain (albeit over a high-dimensional state space), this setting is very general. We show how to define a zero-variance change of probabilities in that context. It is noted that, in general, the zero-variance change of probabilities must depend on the state of the chain. We compare this type of change of probabilities with a more restricted class of IS called *state-independent*, in which the probabilities are changed independently of the current state of the chain. This type of state-independent IS originates mainly from asymptotic approximations based on large deviations theory [3, 15, 26], and has been developed in applications areas such as queueing and finance [10, 15, 16, 20]. However, in many situations, any good IS scheme *must* be state-dependent [3] (as state-independent IS leads to estimators with large, or even infinite, variance). Note that in computational physics (the application area from which it originates) and in reliability, IS has traditionally been state-dependent [5, 14, 16]. Finally, Section 1.4 describes various methods used to approximate the zero-variance (that is, optimal) change of measure. Some just use intuitive approximations, whereas others are based on the asymptotic behavior of the system when the events of interest become more and more rare (this includes methods based on large deviations theory, and other techniques as well). Another option is to use adaptive techniques that learn (and use) approximations of the zero-variance change of measure, or optimal parameter values within a class of parameterized IS strategies: the results of completed runs can be used as inputs of strategies for the next runs, but those IS strategies can also be updated at each step of a given run [16, 23].

The accuracy assessment of the resulting confidence interval, and the robustness properties of the estimator with respect to rarity, is the focus of the next chapter. To avoid overlap, we limit our discussion of these aspects to a minimum here.

1.2 Static problems

We want to compute the expected value of a random variable $X = h(Y)$, $\mathbb{E}[h(Y)]$, where Y is assumed to be a random variable with density f (with respect to the Lebesgue measure) in the d-dimensional real space $\mathbb{R}^d$. (In our examples, we will have $d = 1$.) Then, the crude Monte Carlo method estimates

$$\mathbb{E}[h(Y)] = \int h(y)f(y)dy \quad \text{by} \quad \frac{1}{n}\sum_{i=1}^{n} h(Y_i),$$

where $Y_1, \ldots, Y_n$ are i.i.d. copies of Y, and the integral is over $\mathbb{R}^d$.

IS, on the other hand, samples Y from another density $\tilde{f}$ rather than f. Of course, the same estimator $\frac{1}{n}\sum_{i=1}^{n} h(Y_i)$ then becomes biased in general, but we can recover an unbiased estimator by weighing the simulation output, as follows. Assuming that $\tilde{f}(y) > 0$ whenever $h(y)f(y) \neq 0$,

$$\begin{aligned}\mathbb{E}[h(X)] &= \int h(y)f(y)dy = \int h(y)\frac{f(y)}{\tilde{f}(y)}\tilde{f}(y)dy \\ &= \int h(y)L(y)\tilde{f}(y)dy = \tilde{\mathbb{E}}[h(Y)L(Y)],\end{aligned}$$

where $L(y) = f(y)/\tilde{f}(y)$ is the *likelihood* ratio of the density $f(\cdot)$ with respect to the density $\tilde{f}(\cdot)$, and $\tilde{\mathbb{E}}[\,\cdot\,]$ is the expectation under density $\tilde{f}$. An unbiased estimator of $\mathbb{E}[h(Y)]$ is then

$$\frac{1}{n}\sum_{i=1}^{n} h(Y_i)L(Y_i), \tag{1.1}$$

where $Y_1, \ldots, Y_n$ are i.i.d. random variables sampled from $\tilde{f}$.

The case where Y has a discrete distribution can be handled analogously; it suffices to replace the densities by probability functions and the integrals by sums. That is, if $\mathbb{P}[Y = y_k] = p_k$ for $k \in \mathbb{N}$, then IS would sample n copies of Y, say $Y_1, \ldots, Y_n$, using probabilities $\tilde{p}_k$ instead of p_k, for $k \in \mathbb{N}$, where

$\tilde{p}_k > 0$ whenever $p_k h(y_k) \neq 0$. An unbiased IS estimator of $\mathbb{E}[h(Y)]$ is again (1.1), but with $L(y_k) = p_k/\tilde{p}_k$. Indeed,

$$\tilde{\mathbb{E}}[h(Y)L(Y)] = \sum_{k\in\mathbb{N}} h(y_k)\frac{p_k}{\tilde{p}_k}\tilde{p}_k = \sum_{k\in\mathbb{N}} h(y_k)p_k = \mathbb{E}[h(Y)].$$

In full generality, if Y obeys some probability law (or measure) $\mathbb{P}$, and IS replaces $\mathbb{P}$ by another probability measure $\tilde{\mathbb{P}}$, we must multiply the original estimator by the likelihood ratio (or Radon-Nikodým derivative) $L = d\mathbb{P}/d\tilde{\mathbb{P}}$.

Clearly, the above procedure leaves us with a huge amount of freedom: any alternative $\tilde{\mathbb{P}}$ yields an unbiased estimator (as long as the above-mentioned regularity conditions are fulfilled). Therefore, the next question is: based on what principle should we choose the IS measure $\tilde{\mathbb{P}}$? The goal is to find a change of measure for which the IS estimator has small variance, preferably much smaller than for the original estimator, and is also easy (and not much more costly) to compute (in that the new probability law should be easy to generate variates from). We denote these two variances by

$$\tilde{\sigma}^2(h(Y)L(Y)) = \tilde{\mathbb{E}}[(h(Y)L(Y))^2] - (\mathbb{E}[h(Y)])^2$$

and

$$\sigma^2(h(Y)) = \mathbb{E}[(h(Y))^2] - (\mathbb{E}[h(Y)])^2,$$

respectively. Under the assumptions that the IS estimator has a normal distribution (which is often a good approximation—but not always), a confidence interval at level $1-\alpha$ for $\mathbb{E}[h(Y)]$ is given by

$$\left[\frac{1}{n}\sum_{i=1}^{n} h(Y_i)L(Y_i) - z_{\alpha/2}\frac{\tilde{\sigma}(h(Y)L(Y))}{\sqrt{n}},\ \frac{1}{n}\sum_{i=1}^{n} h(Y_i)L(Y_i) + z_{\alpha/2}\frac{\tilde{\sigma}(h(Y)L(Y))}{\sqrt{n}}\right]$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and Φ is the standard normal distribution function. For fixed α and n, the width of the confidence interval is proportional to the standard deviation (the square root of the variance). So reducing the variance by a factor K improves the accuracy by reducing the width of the confidence interval by a factor $\sqrt{K}$. The same effect is achieved if we multiply n by a factor K, but this requires (roughly) K times more work.

In the rare-event context, one usually simulates until the relative accuracy of the estimator, defined as the ratio of the confidence-interval half-width and the quantity γ to be estimated, is below a certain threshold. For this, we

need $\tilde{\sigma}^2(h(Y)L(Y))/n$ approximately proportional to γ^2. Thus, the number of samples needed is proportional to the variance of the estimator. In the case where γ is a small probability and $h(Y)$ is an indicator function, without IS, $\tilde{\sigma}^2(h(Y)L(Y)) = \sigma^2(h(Y)) = \gamma(1-\gamma) \approx \gamma$, so the required n is roughly inversely proportional to γ and often becomes excessively large when γ is very small.

The optimal change of measure is to select the new probability law $\tilde{\mathbb{P}}$ so that

$$L(Y) = \frac{d\mathbb{P}}{d\tilde{\mathbb{P}}} = \frac{\mathbb{E}[|h(Y)|]}{|h(Y)|},$$

which means $\tilde{f}(y) = f(y)|h(y)|/\mathbb{E}[|h(Y)|]$ in the continuous case, and $\tilde{p}_k = p_k|h(y_k)|/\mathbb{E}[|h(Y)|]$ in the discrete case. Indeed, for any alternative IS measure $\mathbb{P}'$ leading to the likelihood ratio L' and expectation $\mathbb{E}'$, we have

$$\tilde{\mathbb{E}}[(h(Y)L(Y))^2] = (\mathbb{E}[|h(Y)|])^2 = (\mathbb{E}'[|h(Y)|L'(Y)|])^2 \leq \mathbb{E}'[(h(Y)L'(Y))^2].$$

In the special case where $h \geq 0$, the optimal change of measure gives $\tilde{\mathbb{E}}[(h(Y)L(Y))^2] = (\mathbb{E}[h(Y)])^2$, i.e., $\tilde{\sigma}^2(h(Y)L(Y)) = 0$. That is, IS provides a *zero-variance estimator*. We call the corresponding change from $\mathbb{P}$ to $\tilde{\mathbb{P}}$ the *zero-variance change of measure*. In many typical rare-event settings, one indeed has $h \geq 0$; for example, this is obviously the case when the focus is on estimating the probability of a rare event (h is then an indicator function).

All of this is nice in theory, but in practice there is a crucial drawback: implementing the optimal change of measure requires knowledge of $\mathbb{E}[|h(Y)|]$, the quantity that we wanted to compute; if we knew it, no simulation would be needed! But the expression for the zero-variance measure provides a hint on the general form of a "good" IS measure, that is, a change of measure that leads to substantial variance reduction. As a rough general guideline, it says that $L(y)$ should be small when $|h(y)|$ is large.

In particular, if there is a constant $\kappa \leq 1$ such that $L(y) \leq \kappa$ for all y such that $h(y) \neq 0$, then

$$\tilde{\mathbb{E}}[(h(Y)L(Y))^2] \leq \kappa\tilde{\mathbb{E}}[(h(Y))^2L(Y)] = \kappa\mathbb{E}[(h(Y)^2], \tag{1.2}$$

so the second moment is guaranteed to be reduced at least by the factor κ. If h is also an indicator function, say $h(y) = 1_A(y)$ for some set A, and $\mathbb{E}[h(Y)] = \mathbb{P}[A] = \gamma$, then we have

$$\tilde{\sigma}^2(1_A(Y)L(Y)) \;=\; \tilde{\mathbb{E}}[(1_A(Y)L(Y))^2] - \gamma^2 \leq \kappa^2 - \gamma^2.$$

This implies that we always have $\kappa \geq \gamma$, but, evidently, we want to have κ as close as possible to γ.

In theoretical analysis of rare-event simulation, it is customary to parameterize the model by a rarity parameter $\epsilon > 0$ so that the important events occur (in the original model) with a probability that converges to 0 when $\epsilon \to 0$. In that context, an IS estimator based on a change of measure that may depend on ϵ is said to have *bounded relative variance* (or *bounded relative error*) if $\tilde{\sigma}^2(h(Y)L(Y))/\mathbb{E}^2[h(Y)]$ is bounded uniformly in ϵ. This important property means that estimating $\mathbb{E}[h(Y)]$ with a given relative accuracy can be achieved with a bounded number of replications even if $\epsilon \to 0$.

In the special case where $h(y) = 1_A(y)$ and $\gamma = \mathbb{P}[A]$, if we can find a constant κ' such that $L(y) \leq \kappa'\gamma$ when $y \in A$, then

$$\tilde{\sigma}^2(1_A(Y)L(Y)) \quad \leq \quad (\kappa'\gamma)^2 - \gamma^2 = \gamma^2((\kappa')^2 - 1),$$

which means that we have bounded relative variance: the relative variance remains bounded by $(\kappa')^2 - 1$ no matter how rare then event A is. This type of property will be studied in more detail in a separate (dedicated) chapter.

Example 1 To illustrate the ideas and the difficulty of finding a good IS distribution, we first consider a very simple example for which a closed-form expression is known. Suppose that the failure time of a system follows an exponential distribution with rate λ and that we want to compute the probability γ that the system fails before T. We can write $h(y) = 1_A(y)$ where $A = [0, T]$, and we know that $\gamma = \mathbb{E}[1_A(Y)] = 1 - e^{-\lambda T}$. This quantity is small (i.e., A is a rare event) when λT is close to 0. The zero-variance IS here consists in sampling Y from the same exponential density, but truncated to the interval $[0, T]$: $\tilde{f}(y) = \lambda e^{-\lambda y}/(1 - e^{-\lambda T})$ for $0 \leq y \leq T$.

But suppose that we insist on sampling from an exponential density with a different rate $\tilde{\lambda}$ instead of truncating the distribution. The second moment of that IS estimator will be

$$\tilde{\mathbb{E}}[(1_A(Y)L(Y))^2] = \int_0^T \left(\frac{\lambda e^{-\lambda y}}{\tilde{\lambda} e^{-\tilde{\lambda} y}}\right)^2 \tilde{\lambda} e^{-\tilde{\lambda} y} dy = \frac{\lambda^2}{\tilde{\lambda}(2\lambda - \tilde{\lambda})}(1 - e^{-(2\lambda - \tilde{\lambda})T}).$$

Figure 1.1 displays the variance ratio $\tilde{\sigma}^2(1_A(Y)L(Y))/\sigma^2(1_A(Y))$ as a function of $\tilde{\lambda}$, for $T = 1$ and $\lambda = 0.1$. The variance is minimized with $\tilde{\lambda} \approx 1.63$, i.e., with a 16-fold increase of the failure rate, and its minimal value is about 5.3% of the value with $\tilde{\lambda} = \lambda$. If we increase $\tilde{\lambda}$ too much, then the variance

increases again. With $\tilde{\lambda} > 6.01$ (approximately), it becomes larger than with $\tilde{\lambda} = \lambda$. This is due to the fact that for very large values of $\tilde{\lambda}$, the likelihood ratio takes huge values when Y is smaller than T but close to T.

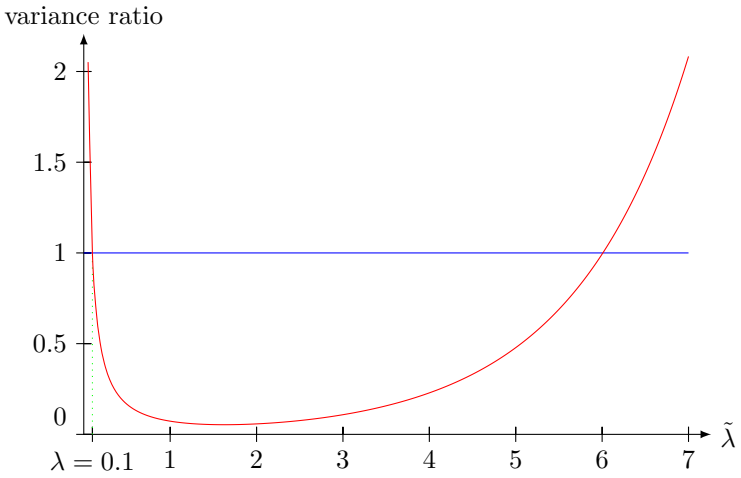

Figure 1.1: Variance ratio (IS vs non-IS) as a function of $\tilde{\lambda}$ for Example 1 with $\lambda = 0.1$ and $A = [0, 1]$.

Suppose now that $A = [T, \infty)$ instead, i.e., $\gamma = \mathbb{P}[Y \geq T]$. The zero-variance density is the exponential with rate λ, truncated to $[T, \infty)$. If we use an exponential with rate $\tilde{\lambda}$ instead, the second moment of the IS estimator is finite if and only if $0 < \tilde{\lambda} < 2\lambda$, and is

$$\tilde{\mathbb{E}}[(1_A(Y)L(Y))^2] = \int_T^\infty \left(\frac{\lambda e^{-\lambda y}}{\tilde{\lambda} e^{-\tilde{\lambda} y}}\right)^2 \tilde{\lambda} e^{-\tilde{\lambda} y} dy = \frac{\lambda^2}{\tilde{\lambda}(2\lambda - \tilde{\lambda})} e^{-(2\lambda - \tilde{\lambda})T}.$$

In this case, the variance is minimized for $\tilde{\lambda} = \lambda + 1/T - (\lambda^2 + 1/T^2)^{1/2} < \lambda$. When $\tilde{\lambda} > 2\lambda$, the variance is infinite because the squared likelihood ratio grows exponentially with y at a faster rate than the exponential rate of decrease of the density. Figure 1.2 shows the variance ratio (IS vs non-IS) as a function of $\tilde{\lambda}$, for $T = 3$ and $\lambda = 1$. We see that the minimal variance is attained with $\tilde{\lambda} \approx \lambda/4$.

Another interesting situation is if $A = [0, T_1] \cup [T_2, \infty)$ where $0 < T_1 < T_2 < \infty$. The zero-variance density is again the exponential truncated to A, which is now split in two pieces. If we just change λ to $\tilde{\lambda}$, then the variance

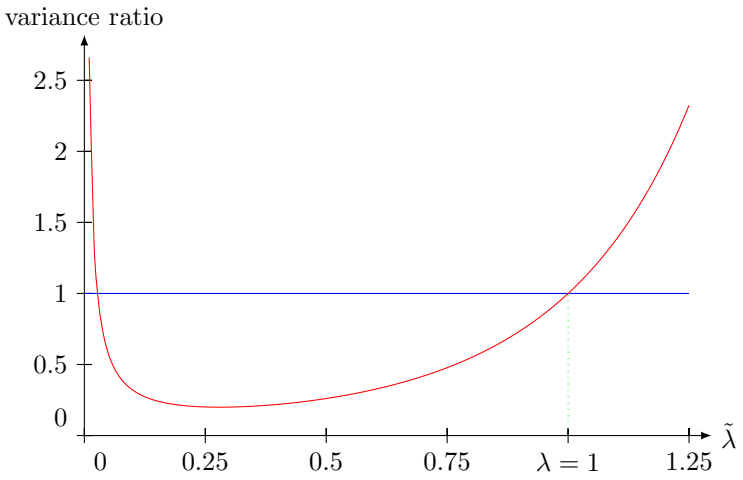

Figure 1.2: Variance ratio as a function of $\tilde{\lambda}$ for Example 1 with $\lambda = 1$ and $A = [3, \infty)$.

associated with the first [second] piece increases if $\tilde{\lambda} < \lambda$ [$\tilde{\lambda} > \lambda$]. So one of the two variance components increases, regardless of how we choose $\tilde{\lambda}$. One way of handling this difficulty is to use a mixture of exponentials: take $\tilde{\lambda}_1 < \lambda$ with probability p_1 and $\tilde{\lambda}_2 > \lambda$ with probability $p_2 = 1 - p_1$. We now have three parameters to optimize: $\tilde{\lambda}_1$, $\tilde{\lambda}_2$, and p_1.

Example 2 For a second illustrative example, let X be binomially distributed with parameters (n, p), and suppose we wish to estimate $\gamma = \mathbb{P}[X \geq na]$ for some constant $a > 0$, where na is assumed to be an integer. Again, the zero-variance IS samples from this binomial distribution truncated to $[na, \infty)$. But if we restrict ourselves to the class of non-truncated binomial changes of measure, say with parameters $(n, \tilde{p})$, following the same line of reasoning as in the previous example, we want to find the $\tilde{p}$ that minimizes the second moment

$$\sum_{i=na}^{n} \binom{n}{i} \left(\frac{p^2}{\tilde{p}}\right)^i \left(\frac{(1-p)^2}{1-\tilde{p}}\right)^{n-i}.$$

Figure 1.3 shows the variance ratio as a function of $\tilde{p}$, for $a = 3/4$, $p = 1/4$, and $n = 20$. It shows that the best choice of $\tilde{p}$ lies around a. If $a > p$ is fixed and $n \to \infty$, then large deviations theory tells us that γ decreases

exponentially with n and that the optimal $\tilde{p}$ (asymptotically) is $\tilde{p} = a$. The intuitive interpretation is that when $a > p$, conditional on the event $\{X/n \geq a\}$, most of the density of X/n is concentrated very close to a when n is large. By selecting $\tilde{p} = a$, IS mimics this conditional density. We also see from the plot that the variance ratio is a very flat function of $\tilde{p}$ in a large neighborhood of a; the ratio is approximately 1.2×10^{-5} from 0.72 to 0.78, and is still 5.0×10^{-5} at 0.58 and 0.88. On the other hand, the IS variance blows up quickly when $\tilde{p}$ approaches 1 or goes below p.

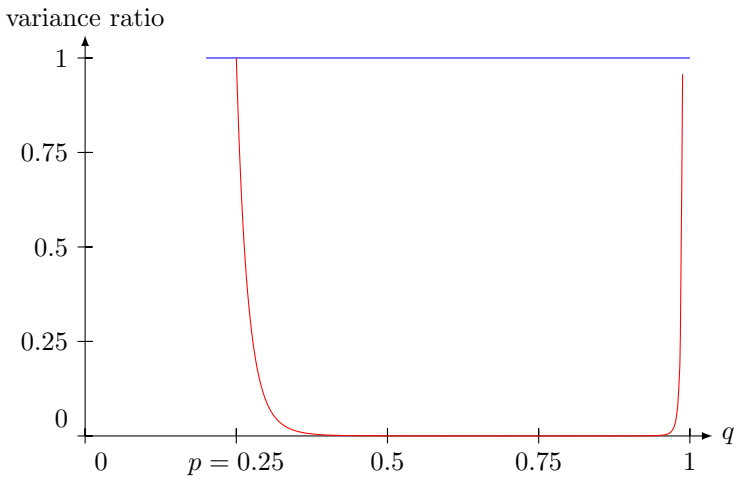

Figure 1.3: Variance ratio of IS vs non-IS estimators, as a function of $\tilde{p}$, for Example 2 with $n = 20$, $p = 1/4$, and $A = [15, \infty)$.

1.3 Markov chains

Having dealt in the previous section with the case of a general random variable, we focus here on the specific case where this random variable is a function of the sample path of a Markov chain. We introduce IS in this context, both for discrete-time and continuous-time chains, as well as the form of the corresponding zero-variance change of measure. Approximation algorithms for this change of measure are discussed in the next section.

1.3.1 Discrete-time Markov chains

Consider now a discrete-time Markov chain (DTMC), say $\{Y_j,\, j \geq 0\}$, with discrete state space $\mathcal{Y}$ (possibly infinite and high-dimensional). The chain evolves up to a stopping time τ defined as the first time the chain hits a given set of states, $\Delta \subset \mathcal{Y}$; that is, $\tau = \inf\{j \geq 0 : Y_j \in \Delta\}$. We assume that $\mathbb{E}[\tau] < \infty$. The chain has a transition probability matrix whose elements are $P(y,z) = \mathbb{P}[Y_j = z \mid Y_{j-1} = y]$ for all $y, z \in \mathcal{Y}$, and the initial probabilities are $\pi_0(y) = \mathbb{P}[Y_0 = y]$ for all $y \in \mathcal{Y}$. We consider the random variable $X = h(Y_0, \ldots, Y_\tau)$, where h is a given function of the trajectory of the chain, with values in $[0, \infty)$. Let $\gamma(y) = \mathbb{E}_y[X]$ denote the expected value of X when $Y_0 = y$, and define $\gamma = \mathbb{E}[X] = \sum_{y \in \mathcal{Y}} \pi_0(y)\gamma(y)$, the expected value of X for the initial distribution π_0.

Our discussion could be generalized to broader classes of state spaces. For a continuous state space $\mathcal{Y}$, the transition probabilities would have to be replaced by a probability transition kernel and the sums by integrals, and we would need some technical measurability assumptions. Any discrete-event simulation model for which we want to estimate the expectation of some random variable $X = h(Y_0, \ldots, Y_\tau)$ as above can fit into this framework. For simplicity, we stick to a discrete state space.

The basic idea of IS here is to replace the probabilities of sample paths $(y_0, \ldots, y_n)$,

$$\mathbb{P}[(Y_0, \ldots, Y_\tau) = (y_0, \ldots, y_n)] = \pi_0(y_0) \prod_{j=1}^{n} P(y_{j-1}, y_j),$$

where $n = \min\{j \geq 0 : y_j \in \Delta\}$, by new probabilities $\tilde{\mathbb{P}}[(Y_0, \ldots, Y_\tau) = (y_0, \ldots, y_n)]$ such that $\tilde{\mathbb{E}}[\tau] < \infty$ and $\tilde{\mathbb{P}}[\cdot] > 0$ whenever $\mathbb{P}[\cdot]h(\cdot) > 0$. This is extremely general.

To be more practical, we might want to restrict ourselves to changes of measure under which $\{Y_j,\, j \geq 0\}$ remains a DTMC with the same state space $\mathcal{Y}$. That is, we replace the transition probabilities $P(y,z)$ by new transition probabilities $\tilde{P}(y,z)$ and the initial probabilities $\pi_0(y)$ by $\tilde{\pi}_0(y)$. The new probabilities must be chosen so that any sample path having a positive contribution to γ must still have a positive probability, and $\tilde{\mathbb{E}}[\tau] < \infty$. The likelihood ratio becomes

$$L(Y_0, \ldots, Y_\tau) = \frac{\pi_0(Y_0)}{\tilde{\pi}_0(Y_0)} \prod_{j=1}^{\tau} \frac{P(Y_{j-1}, Y_j)}{\tilde{P}(Y_{j-1}, Y_j)}$$

and we have

$$\gamma = \tilde{\mathbb{E}}[XL(Y_0, \ldots, Y_\tau)].$$

A question that comes to mind: Is there a zero-variance change of measure in this setting? What is it?

To answer this question, following [5, 18, 19, 21], we restrict ourselves to the case where the cost function X is additive:

$$X = \sum_{j=1}^{\tau} c(Y_{j-1}, Y_j) \tag{1.3}$$

for some function $c : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$. Note that in this case, we can multiply the term $c(Y_{j-1}, Y_j)$ by the likelihood ratio only up to step j. This gives the estimator

$$\tilde{X} \;=\; \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{P(Y_{j-1}, Y_j)}{\tilde{P}(Y_{j-1}, Y_j)}.$$

We now show that in this setting, if we take $\tilde{P}(y, z)$ proportional to

$$P(y, z)[c(y, z) + \gamma(z)]$$

for each $y \in \mathcal{Y}$, then we have zero variance. (Without the additivity assumption (1.3), to get zero variance, the probabilities for the next state must depend in general of the entire history of the chain.) Suppose that

$$\tilde{P}(y, z) = \frac{P(y, z)(c(y, z) + \gamma(z))}{\sum_{w \in \mathcal{Y}} P(y, w)(c(y, w) + \gamma(w))} = \frac{P(y, z)(c(y, z) + \gamma(z))}{\gamma(y)}, \tag{1.4}$$

where the denominator acts as a normalization constant (the probabilities add up to 1 from the first equality; the second equility results from conditionning with respect to a one-step transition). Then,

$$\begin{aligned}
\tilde{X} &= \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{P(Y_{j-1}, Y_j)}{\tilde{P}(Y_{j-1}, Y_j)} \\
&= \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{P(Y_{j-1}, Y_j)\gamma(Y_{j-1})}{P(Y_{j-1}, Y_j)(c(Y_{j-1}, Y_j) + \gamma(Y_j))} \\
&= \sum_{i=1}^{\tau} c(Y_{i-1}, Y_i) \prod_{j=1}^{i} \frac{\gamma(Y_{j-1})}{c(Y_{j-1}, Y_j) + \gamma(Y_j)} \\
&= \gamma(Y_0)
\end{aligned}$$

by induction on the value taken by τ, using the fact that $\gamma(Y_\tau) = 0$. In other words, the estimator is a constant, so it has zero variance.

Another way to show this property is by looking at the variance and using the classical decomposition

$$\tilde{\sigma}^2[X|Y_0] = \tilde{\sigma}^2[\tilde{\mathbb{E}}[X|Y_0, Y_1]|Y_0] + \tilde{\mathbb{E}}[\tilde{\sigma}^2[X|Y_0, Y_1]|Y_0]. \tag{1.5}$$

Define $v(y) = \tilde{\sigma}^2[\tilde{X}|Y_0 = y]$. Then

$$\begin{aligned} v(Y_0) &= \tilde{\sigma}^2[\tilde{\mathbb{E}}[\tilde{X} \mid Y_1]|Y_0] + \tilde{\mathbb{E}}[\tilde{\sigma}^2[\tilde{X} \mid Y_1]|Y_0] \\ &= \tilde{\sigma}^2[(c(Y_0, Y_1) + \gamma(Y_1))L(Y_0, Y_1)|Y_0] + \tilde{\mathbb{E}}[L^2(Y_0, Y_1)v(Y_1)|Y_0] \\ &= \tilde{\mathbb{E}}[(c(Y_0, Y_1) + \gamma(Y_1))^2 L^2(Y_0, Y_1)|Y_0] - \gamma^2(Y_0) + \tilde{\mathbb{E}}[L^2(Y_0, Y_1)v(Y_1)|Y_0] \\ &= \tilde{\mathbb{E}}[((c(Y_0, Y_1) + \gamma(Y_1))^2 + v(Y_1))L^2(Y_0, Y_1)|Y_0] - \gamma^2(Y_0). \end{aligned}$$

From the change of measure, we have

$$\tilde{\mathbb{E}}[(c(Y_0, Y_1) + \gamma(Y_1))^2 L^2(Y_0, Y_1)|Y_0] = \gamma^2(Y_0),$$

leading to

$$\begin{aligned} v(Y_0) &= \tilde{\mathbb{E}}[((c(Y_0, Y_1) + \gamma(Y_1))^2 + v(Y_1))L^2(Y_0, Y_1)|Y_0] - \gamma^2(Y_0) \\ &= \tilde{\mathbb{E}}[v(Y_1)L^2(Y_0, Y_1)|Y_0]. \end{aligned}$$

Applying induction, we again obtain

$$v(Y_0) = \tilde{\mathbb{E}}\left[v(Y_\tau)\prod_{j=1}^{\tau} L(Y_{i-1}, Y_i)\right] = 0$$

because $v(Y_\tau) = 0$.

The change of measure (1.4) is actually the *unique* Markov chain implementation of the zero-variance change of measure. To see that, suppose we are in state $Y_j = y \notin \Delta$. Since

$$v(y) \geq \tilde{\sigma}^2[\tilde{\mathbb{E}}[\tilde{X} \mid Y_1]|Y_0 = y] = \tilde{\sigma}^2[(c(y, Y_1)+\gamma(Y_1))P(y, Y_1)/\tilde{P}(y, Y_1) \mid Y_0 = y],$$

zero-variance implies that $(c(y, Y_1) + \gamma(Y_1))P(y, Y_1)/\tilde{P}(y, Y_1) = K_y$ for some constant K_y that does not depend on Y_1. But since the probabilities $\tilde{P}(y, Y_1)$ must sum to 1 for any fixed y, the constant K_y must take the value $\gamma(y)$ as in (1.4). The same argument can be repeated at each step of the Markov chain.

It is important to emphasize that in (1.4), the probabilities are changed in a way that depends in general on the current state of the chain.

Again, knowing the zero-variance IS measure requires the knowledge of $\gamma(y)$ for all y; that is, of the values we are trying to estimate. In practice, we can try to approximate the zero-variance IS by replacing γ by an accurate proxy, and using this approximation in (1.4) [1, 4, 5, 18]. Some methods restrict themselves to a parametric class of IS distributions and try to optimize the parameters, instead of trying to approximate the zero-variance IS. We will return to this in Section 1.4.

Example 3 Consider a Markov chain with state-space $\{0, 1, \ldots, B\}$, for which $P(y, y+1) = p_y$ and $P(y, y-1) = 1 - p_y$, for $y = 1, \ldots, B-1$, and $P(0, 1) = P(B, B-1) = 1$. Note that a birth-and-death process with bounded state space has an embedded DTMC of this form. We take $\Delta = \{0, B\}$, and we define $\gamma(y) = \mathbb{P}[Y_\tau = B \mid Y_0 = y]$. This function γ satisfies the recurrence equations

$$\gamma(y) = p_y \gamma(y+1) + (1 - p_y)\gamma(y-1)$$

for $y = 1, \ldots, B-1$, with the boundary conditions $\gamma(0) = 0$ and $\gamma(B) = 1$. This gives rise to a linear system of equations that is easy to solve. In the case where $p_y = p < 1$ for $y = 1, \ldots, B-1$, this is known as the gambler's ruin problem, and $\gamma(y)$ is given by the explicit formula $\gamma(y) = (1-\rho^{-y})/(1-\rho^{-B})$ if $\rho = p/(1-p) \neq 1/2$, and $\gamma(y) = y/B$ if $\rho = 1/2$.

But for the sake of illustration, suppose we want to estimate $\gamma(1)$ by simulation with IS. The zero-variance change of measure in this case replaces each p_y, for $1 \leq y < B$, by

$$\tilde{p}_y = \frac{p_y \gamma(y+1)}{\gamma(y)} = \frac{p_y \gamma(y+1)}{p_y \gamma(y+1) + (1-p_y)\gamma(y-1)}.$$

Since $\gamma(0) = 0$, this gives $\tilde{p}_1 = 1$, which means that this change of measure cuts the link that returns to 0, so it brings us to B with probability 1. For the special case where $p_y = p$ for $y = 1, \ldots, B-1$, by plugging the formula for $\gamma(y)$ into the expression for $\tilde{p}_y$, we find that the zero-variance probabilities are

$$\tilde{p}_y = \frac{1 - \rho^{-y-1}}{1 - \rho^{-y}}\, p.$$

Note that all the terms $1 - \rho^{-B}$ have canceled out, so the new probabilities $\tilde{p}_y$ do not depend on B. On the other hand, they depend on y even though the original probabilities p did not depend on y.

One application that fits this framework is an $M/M/1$ queue with arrival rate λ and service rate $\mu > \lambda$. Let $\rho = \lambda/\mu$ and $p = \lambda/(\lambda + \mu)$. Then $\gamma(y)$ represents the probability that the number of customers in the system reaches level B before the system empties, given that there are y customers in the system.

1.3.2 Continuous-time Markov chains

We now examine how the previous framework applies to continuous-time Markov chains (CTMC). Following [13], let $Y = \{Y(t), t \geq 0\}$ be a CTMC evolving in $\mathcal{Y}$ up to some stopping time $T = \inf\{t \geq 0 : Y(t) \in \Delta\}$, where $\Delta \subset \mathcal{Y}$. The initial distribution is π_0 and the jump rate from y to z, for $z \neq y$, is $a_{y,z}$. Let $a_y = \sum_{z \neq y} a_{y,z}$ be the departure rate from y. The goal is to estimate $\mathbb{E}[X]$, where $X = h(Y)$ is a function of the entire sample path of the CTMC up to its stopping time T. A sample path for this chain is determined uniquely by the sequence $(Y_0, V_0, Y_1, V_1, \ldots, Y_\tau, V_\tau)$ where Y_j is the jth visited state of the chain, V_j the time spent in that state, and τ is the index of the jump that corresponds to the stopping time (the first jump that hits Δ). Therefore $h(Y)$ can be re-expressed as $h^*(Y_0, V_0, Y_1, V_1, \ldots, Y_n, V_n)$, and a sample path $(y_0, v_0, y_1, v_1, \ldots, y_n, v_n)$ has density (or likelihood)

$$\begin{aligned} p(y_0, v_0, \ldots, y_n, v_n) &= \prod_{j=0}^{n-1} \frac{a_{y_j, y_{j+1}}}{a_{y_j}} \prod_{j=0}^{n} a_{y_j} \exp[-a_{y_j} v_j] \\ &= \prod_{j=0}^{n-1} a_{y_j, y_{j+1}} \exp[-\sum_{j=0}^{n} a_{y_j} v_j], \end{aligned}$$

each term $a_{y_j, y_{j+1}}/a_{y_j}$ being the probability of moving from y_j to y_{j+1} and $a_{y_j} \exp[-a_{y_j} v_j]$ the density for leaving y_j after a sojourn time v_j. Then we have

$$\mathbb{E}[X] = \sum_{y_0, \ldots, y_n} \int_0^\infty \cdots \int_0^\infty h^*(y_0, v_0, \ldots, y_n, v_n) p(y_0, v_0, \ldots, y_n, v_n) dv_0 \cdots dv_n.$$

Suppose that the cost function has the form

$$X = h(Y) = \sum_{j=1}^{\tau} c'(Y_{j-1}, V_{j-1}, Y_j)$$

where $c' : \mathcal{Y} \times [0, \infty) \times \mathcal{Y} \to [0, \infty)$. In this case, a standard technique that always reduces the variance, and often reduces the computations as well, is to replace the estimator X by

$$X_{\text{cmc}} = \mathbb{E}[X \mid Y_0, \ldots, Y_\tau] = \sum_{j=1}^{\tau} c(Y_{j-1}, Y_j),$$

where $c(Y_{j-1}, Y_j) = \mathbb{E}[c'(Y_{j-1}, V_{j-1}, Y_j) \mid Y_{j-1}, Y_j]$ [9]. In other words, we would never generate the sojourn times. We are now back in our previous DTMC setting and the zero-variance transition probabilities are given again by (1.4).

Consider now the case of a *fixed* time horizon T, which therefore no longer has the form $T = \inf\{t \geq 0 : Y(t) \in \Delta\}$. We then have two options, either we again reformulate the process as a DTMC, or keep a CTMC formulation. In the first case, we can redefine the state as (Y_j, R_j) at step j, where R_j is the remaining clock time (until we reach time T), as in [6]. Then the zero-variance scheme is the same as for the DTMC setting if we replace the state Y_j there by (Y_j, R_j), and if we redefine Δ. We then have a non-denumerable state space, so the sums must be replaced by combinations of sums and integrals. In this context of a finite time horizon, effective IS schemes will typically use *non-exponential* (often far from exponential) sojourn time distributions. This means that we will no longer have a CTMC under IS. Assume now that we want to stick with a CTMC formulation, and that we restrict ourselves to choosing a Markovian IS measure with new initial distribution $\tilde{\pi}_0$ and new generator $\tilde{A}$ such that $\tilde{\pi}_0(y) > 0$ (respectively $\tilde{a}_{y,z} > 0$) whenever $\pi_0(y) > 0$ (respectively $a_{y,z} > 0$). Let τ be the index j of the first jump to a state $Y_j = Y(t_j)$ at time t_j such that $t_j \geq T$. Then, similarly to the discrete-time case, it can be shown that, provided τ is a stopping time with finite expectation under $(\tilde{\pi}_0, \tilde{A})$,

$$\mathbb{E}[X] = \tilde{\mathbb{E}}[X L_\tau],$$

with L_τ the likelihood ratio given by

$$L_\tau = \frac{\pi_0(Y_0)}{\tilde{\pi}_0(Y_0)} \prod_{j=0}^{\tau-1} \frac{a_{Y_j, Y_{j+1}}}{\tilde{a}_{Y_j, Y_{j+1}}} \exp\left[\sum_{j=0}^{\tau} (\tilde{a}_{Y_j} - a_{Y_j}) V_j\right].$$

In the above formula, we can also replace the likelihood $a_{Y_{\tau-1},Y_\tau} \exp[-a_{Y_{\tau-1}} V_{\tau-1}]$ of the last occurrence time V_τ by the probability that this event occurs after the remaining time $T - \sum_{j=0}^{\tau-1} V_j$, which is $\exp[-a_{Y_{\tau-1}}(T - \sum_{j=0}^{\tau-1} V_j)]$, the same being done for the IS measure: instead of considering the exact occurrence time after time T, we consider its expected value given that it happens after T. This reduces the variance of the estimator, because it replaces it by its conditional expectation.

1.3.3 State-independent versus state-dependent changes of measure

In the context of simulating a Markov chain, we often distinguish two types of IS strategies:

- State-independent IS, where the change of measure does not depend on the current state of the Markov chain;
- State-dependent IS where at each step of the Markov chain, a new IS change of measure is used that takes into account the current state of the Markov chain. In case where the state of the chain must contain the current simulation time (e.g., if the simulation stops at a fixed clock time in the model), then the change of measure will generally depend on the current time.

Example 4 In Example 3, even though the birth-and-death process had original transition probabilities p and $1-p$ that did not depend on the current state y, the zero-variance probabilities $\tilde{p}_y$ did depend on y (although not on B). These probabilities satisfy the equations

$$\tilde{p}_y(1 - \tilde{p}_{y-1}) = p(1-p)$$

for $y \geq 2$, with boundary condition $\tilde{p}_1 = 1$. For $p < 1/2$, we have $1 - p < \tilde{p}_y < \tilde{p}_{y-1} < 1$ for all $y > 2$, and $\tilde{p}_y \to 1 - p$ when $y \to \infty$. That is, the optimal change of measure is very close to simply permuting p and $1-p$, i.e., taking $\tilde{p} = 1 - p > 1/2$. For the $M/M/1$ queue, this means exchanging the arrival rate and the service rate, which gives an unstable queue (i.e., the event under consideration is not rare anymore). This simple permutation is an example of a *state-independent* change of measure: It does not depend on the current state y.

With $\tilde{p} = 1 - p$, the likelihood ratio associated with any sample path that reaches level B before returning to 0 is ρ^{B-1}, so, when estimating $\gamma(1)$, the second moment is reduced at least by that factor, as shown by Inequality (1.2). This reduction can be quite substantial. Moreover, the probability $\tilde{\gamma}(1)$ of reaching B under the new measure must satisfy $\tilde{\gamma}(1)\rho^{B-1} = \gamma(1)$, which implies that

$$\tilde{\gamma}(1) = \gamma(1)\rho^{1-B} = \frac{1-\rho}{1-\rho^B}.$$

Then the relative variance is

$$\frac{\tilde{\gamma}(1)\rho^{2B-2}}{\gamma^2(1)} - 1 = \frac{(1-\rho)(\rho^{2B} - 2\rho^B + 1)}{(\rho-1)^2} - 1 = \frac{1-\rho^B}{1-\rho} - 1 \approx \frac{\rho}{1-\rho}$$

when B is large. We have the remarkable result that the number of runs needed to achieve a predefined precision remains bounded in B, i.e., we have bounded relative error as $B \to \infty$, even with a state-independent change of measure.

Example 5 Suppose now that our birth-and-death process evolves over the set of non-negative integers and let $\gamma(y)$ be the probability that the process ever reaches 0 if it starts at $y > 0$. This $\gamma(y)$ can be seen as the ruin probability if we start with y Euros in hand and win [lose] one Euro with probability p $[1-p]$ at each step. For $p \leq 1/2$, $\gamma(y) = 1$, so we assume that $p > 1/2$. In this case, we have that $\gamma(1) = (1-p) + p\gamma(2) = (1-p) + \gamma^2(1)$. For $j \geq 2$, $\gamma(j+1) = \gamma(1)\gamma(j)$ because the probability of reaching 0 from $j+1$ is the probability of eventually reaching j from $j+1$, which equals $\gamma(1)$, multiplied by the probability of reaching 0 from j. From this, we find that $\gamma(1) = (1-p)/p$. Still from $\gamma(j+1) = \gamma(1)\gamma(j)$, we find easily that the zero-variance probabilities are $\tilde{p}_j = 1 - p$ for all $j \geq 1$. In this case, the zero-variance change of measure is state-independent.

Example 6 We come back to Example 2, where we wish to estimate $\gamma = \mathbb{P}[X \geq na]$, for X binomially distributed with parameters (n, p), and for some constant $a > p$. If we view X as a sum of n independent Bernoulli random variables and define Y_j and the partial sum of the first j variables, then $X = Y_n$ and we have a Markov chain $\{Y_j,\, j \geq 0\}$. We observed in Example 2 that when we restricted ourselves to a state-independent change of measure that replaced p by $\tilde{p}$ for this Markov chain, the variance was approximately minimized by taking $\tilde{p} = a$. In fact, this choice turns out to

be optimal asymptotically when $n \to \infty$ [25]. But even this optimal choice fails to provide a bounded relative error. That is, a state-independent change of p cannot provide a bounded relative error in this case. The only way of getting a bounded relative error is via state-dependent IS. However, when p is replaced by $\tilde{p} = a$, the relative error increases only very slowly when $n \to \infty$: the second moment decreases exponentially at the same exponential rate as the square of the first moment. When this property holds, the estimator is said to have *logarithmic efficiency*. In this example, it holds for no other value of $\tilde{p}$. All these results have been proved in a more general setting by Sadowsky [25].

1.4 Algorithms

A general conclusion from the previous section is that to accurately approximate the zero variance IS estimator, a key ingredient is a good approximation of the function $\gamma(\cdot)$. In fact, there are several ways of finding a good IS strategy. Most of the good methods can be classified in two large families: those that try to directly approximate the zero-variance change of measure via an approximation of the function $\gamma(\cdot)$, and those that restrict a priori the change of measure to a parametric class, and then try to optimize the parameters. In both cases, the choice can be made either via simple heuristics, or via a known asymptotic approximation for $\gamma(y)$, or by adaptive methods that learn (statistically) either the function $\gamma(\cdot)$ or the vector or parameters that minimizes the variance. In the remainder of this section, we briefly discuss these various approaches.

In the scientific literature, IS has often been applied in a very heuristic way, without making any explicit attempt to approximate the zero-variance change of measure. One heuristic idea is simply to change the probabilities so that the system is pushed in the direction of the rare event, by looking at what could increase its occurrence. However, Example 1 shows very well how pushing too much can have the opposite effect; in fact, it can easily lead to an infinite variance. Changes of measure that may appear promising a priori can eventually lead to a variance increase. In situations where the rare event can be reached in more than one direction, pushing in one of those directions may easily inflate the variance by reducing the probability or density of paths that lead to the rare event via other directions. The last part of Example 1 illustrates a simplified case of this. Other illustrations can

be found in [17, 2, 3, 11], for example. Generally speaking, good heuristics should be based on reasonable understanding of the shape of $\gamma(\cdot)$ and/or the way the likelihood ratio will behave under IS. We give examples of these types of heuristics in the next subsection.

1.4.1 Heuristic approaches

Here, the idea is to use an heuristic approximation of $\gamma(\cdot)$ in the change of measure (1.4).

Example 7 We return to Example 3, with $p_y = p$. Our aim is to estimate $\gamma(1)$. Instead of looking at the case where B is large, we rather focus on the case where p is small, $p \to 0$ for fixed B. This could be seen as a (simplified) dependability model where each transition from y to $y+1$ represents a component failure, each transition from y to $y-1$ corresponds to a repair, and B is the minimal number of failed components for the whole system to be in a state of failure. If $p \ll 1$, each failure transition (except the first one) is rare and we have $\gamma(1) \ll 1$ as well. Instead of just blindly increasing the failure probabilities, we can try to mimic the zero-variance probabilities (1.4) by replacing $\gamma(\cdot)$ in this expression by an approximation, with $c(y,z) = 0$, $\gamma(0) = 0$ and $\gamma(B) = 1$. Which approximation $\hat{\gamma}(y)$ could we use instead of $\gamma(y)$? Based on the asymptotic estimate $\gamma(y) = p^{B-y} + o(p^{B-y})$, taking $\hat{\gamma}(y) = p^{B-y}$ for all $y \in \{1, \ldots, B-1\}$, with $\hat{\gamma}(0) = 0$ and $\hat{\gamma}(B) = 1$, looks like a good option. This gives

$$\tilde{P}(y, y+1) = \frac{p^{B-y}}{p^{B-y} + (1-p)p^{B-y+1}} = \frac{1}{1 + (1-p)p}$$

for $y = 2, \ldots, B-2$. Repairs then become rare while failures are no longer rare.

We can extend the previous example to a multidimensional state space, which may correspond to the situation where there are different types of components, and a certain subset of the combinations on the numbers of failed components of each type corresponds to the failure state of the system. Several IS heuristics have been proposed for this type of setting [16] and some of them are examined in Chapter ?????? of this book. One heuristic suggested in [21] approximates $\gamma(y)$ by considering the probability of the most likely path to failure. In numerical examples, it provides a drastic variance reduction with respect to previously known IS heuristics.

1.4.2 Learning the Function $\gamma(\cdot)$

Various techniques that try to approximate the function $\gamma(\cdot)$, often by adaptive learning, and plug the approximation (1.4), have been developed in the literature [16]. Old proposals of this type can be found in the computational physics literature, for example; see the references in [5]. We outline examples of such techniques taken from recent publications.

One simple type of approach, called *adaptive Monte Carlo* in [18, 8], proceeds iteratively as follows. At step i, it replaces the exact (unknown) value $\gamma(x)$ in (1.4) by a guess $\gamma^{(i)}(x)$, and it uses the probabilities

$$\tilde{P}^{(i)}(y,z) = \frac{P(y,z)(c(y,z)+\gamma^{(i)}(z))}{\sum_{w\in\mathcal{Y}} P(y,w)(c(y,w)+\gamma^{(i)}(w))} \tag{1.6}$$

in n_i independent simulation replications, to obtain a new estimation $\gamma^{(i+1)}(y)$ of $\gamma(y)$, from which a new transition matrix $\tilde{P}^{(i+1)}$ is defined. These iterations could go on until we feel that the probabilities have converged to reasonably good estimates.

A second type of approach is to try to approximate the function $\gamma(\cdot)$ by stochastic approximation. The *adaptive stochastic approximation* (ASA) method proposed in [1] for the simulation of discrete-time finite-state Markov chains falls in that category. One starts with a given distribution for the initial state y_0 of the chain, an initial transition matrix $\tilde{P}^{(0)}$ (it can be the original transition matrix of the chain), and an initial guess $\gamma^{(0)}(\cdot)$ of the value function $\gamma(\cdot)$. The method simulates a single sample path as follows. At each step n, given the current state y_n of the chain, if $y_n \notin \Delta$, we use the current transition matrix $\tilde{P}^{(n)}$ to generate the next state y_{n+1}, we update the estimate of $\gamma(y_n)$ by

$$\begin{aligned}\gamma^{(n+1)}(y_n) &= (1-a_n(y_n))\gamma^{(n)}(y_n) \\ &\quad + a_n(y_n)\left[c(y_n,y_{n+1}) + \gamma^{(n)}(y_{n+1})\frac{P(y_n,y_{n+1})}{\tilde{P}^{(n)}(y_n,y_{n+1})}\right],\end{aligned}$$

where $\{a_n(y),\ n\ge 0\}$, is a sequence of *step sizes* such that $\sum_{n=1}^{\infty} a_n(y) = \infty$ and $\sum_{n=1}^{\infty} a_n^2(y) < \infty$ for each state y, and we update the probability of the current transition by

$$\tilde{P}^{(n+1)}(y_n,y_{n+1}) = \max\left(P(y_n,y_{n+1})\frac{\left[c(y_n,y_{n+1})+\gamma^{(n+1)}(y_{n+1})\right]}{\gamma^{(n+1)}(y_n)},\ \delta\right)$$

where $\delta > 0$ is a constant whose role is to ensure that the likelihood ratio remains bounded (to rule out the possibility that it takes huge values). For the other states, we take $\gamma^{(n+1)}(y) = \gamma^{(n)}(y)$ and $\tilde{P}^{(n+1)}(y, z) = P^{(n)}(y, z)$. We then normalize via

$$P^{(n+1)}(y_n, y) = \frac{\tilde{P}^{(n+1)}(y_n, y)}{\sum_{z \in \mathcal{Y}} \tilde{P}^{(n+1)}(y_n, z)}$$

for all $y \in \mathcal{Y}$. When $y_n \in \Delta$, i.e., if the stopping time is reached at step n, y_{n+1} is generated again from the initial distribution, the transition matrix and the estimate of $\gamma(\cdot)$ are kept unchanged, and the simulation is resumed. In [1], batching techniques are used to obtain a confidence interval.

Experiments reported in [1] show that these methods can be quite effective when the state space has small cardinality. However, since they require storing the approximation $\gamma^{(n)}(y)$ for each state y, their direct implementation quickly becomes impractical as the number of states increases (e.g., for continuous state spaces, or for multidimensional state spaces such as those of Example 7).

In the case of large state spaces, one must rely on interpolation or approximation instead of trying to estimate $\gamma(y)$ directly at each state y. One way of doing this is by selecting a set of k predefined basis functions $\gamma_1(y), \ldots, \gamma_k(y)$, and search for a good approximation of $\gamma(\cdot)$ within the class of linear combinations of the form $\hat{\gamma}(y) = \sum_{j=1}^{k} \alpha_j \gamma_j(y)$, where the weights $(\alpha_1, \ldots, \alpha_k)$ can be learned or estimated in various ways, for instance by stochastic approximation. It therefore consists in a parametric approach, where the parameter is the vector of weights.

1.4.3 Optimizing within a parametric class

Most practical and effective IS strategies in the case of large state spaces restrict themselves to a parametric class of IS measures, either explicitly or implicitly, and try to estimate the parameter vector that minimizes the variance. More specifically, we consider a family of measures $\{\tilde{\mathbb{P}}_\theta, \theta \in \Theta\}$, which may represent a family of densities $\tilde{f}_\theta$, or a family of probability vectors $\tilde{p}_\theta$ for a discrete distribution, or the probability measure associated with the transition matrix $\tilde{P}_\theta$ or the transition kernel of a Markov chain. Then, we look for a θ that minimizes the variance of the IS estimator under $\tilde{\mathbb{P}}_\theta$, or some other measure of distance to the zero-variance measure, over the set Θ.

Of course, a key issue is a clever selection of this parametric class, so that it includes good IS strategies within the class. The value of θ can be selected either via a separate prior analysis, for example based on asymptotically valid approximations, or can be learned adaptively. We briefly discuss these two possibilities in what follows.

Non-adaptive parameter selection

Examples 2 and 6 illustrate the popular idea of fixing θ based on an asymptotic analysis. The parametric family there is the class of binomial distributions with parameters $(n, \tilde{p})$. We have $\theta = \tilde{p}$. Large deviations theory shows that twisting the binomial parameter p to $\tilde{p} = a$ is asymptotically optimal [25]. This choice works quite well in practice for this type of example. On the other hand, we also saw that it cannot provide a bounded relative variance. Several additional examples illustrating the use of large deviations theory to select a good change of measure can be found in [3, 15, 16], for example.

Adaptive learning of the best parameters

The value of θ that minimize the variance can be learned adaptively in various ways. For example, the ASA method described earlier can be adapted to optimize θ by stochastic approximation. Another type of approach is based on *sample average approximation*: write the variance or the second moment as a mathematical expectation that depends on θ, replace the expectation by a sample average function of θ obtained by simulation, and optimize this sample function with respect to θ. These simulations are performed under an IS measure $\tilde{P}$ that may differ from P and does not have to belong to the selected family. The optimizer $\hat{\theta}^*$ is used in a second stage to estimate the quantity of interest using IS.

A more general way of formulating this optimization problem is to replace the variance by some other measure of distance between $\tilde{\mathbb{P}}_\theta$ and the optimal (zero-variance) change of measure $\tilde{\mathbb{P}}^*$, which is known to satisfy $d\tilde{\mathbb{P}}^* = (|X|/\mathbb{E}[|X|])d\mathbb{P}$ when we want to estimate $\gamma = \mathbb{E}[X]$. Again, there are many ways of measuring this distance and some are more convenient than others.

Rubinstein [24] proposed and motivated the use of the Kullback-Leibler

(or cross-entropy) "distance", defined by

$$\mathcal{D}(\tilde{\mathbb{P}}^*, \tilde{\mathbb{P}}_\theta) \;=\; \tilde{\mathbb{E}}^* \left[\log \frac{d\tilde{\mathbb{P}}^*}{d\tilde{\mathbb{P}}_\theta} \right]$$

(this is not a true distance, because it is not symmetric and does not satisfy the triangle inequality, but this causes no problem), and called the resulting technique the *cross-entropy* (CE) method [7, 23, 24].

Easy manipulations lead to

$$\mathcal{D}(\tilde{\mathbb{P}}^*, \tilde{\mathbb{P}}_\theta) = \mathbb{E}\left[\frac{|X|}{\mathbb{E}[|X|]} \log\left(\frac{|X|}{\mathbb{E}[|X|]} d\mathbb{P} \right) \right] - \frac{1}{\mathbb{E}[|X|]} \mathbb{E}\left[|X| \log d\tilde{\mathbb{P}}_\theta \right].$$

Since only the last expectation depends on θ, minimizing the above expression is equivalent to solving

$$\max_{\theta\in\Theta} \mathbb{E}\left[|X| \log d\tilde{\mathbb{P}}_\theta \right] = \max_{\theta\in\Theta} \tilde{\mathbb{E}}\left[\frac{d\mathbb{P}}{d\tilde{\mathbb{P}}} |X| \log d\tilde{\mathbb{P}}_\theta \right]. \tag{1.7}$$

The CE method basically solves the optimization problem on the right side of (1.7) by sample-average approximation, replacing the expectation $\tilde{\mathbb{E}}$ in (1.7) by a sample average over simulations performed under $\tilde{\mathbb{P}}$.

How should we select $\tilde{\mathbb{P}}$? In the case of rare events, it is often difficult to find a priori a distribution $\tilde{\mathbb{P}}$ under which the optimizer of the sample average approximation does not have too much variance and is sufficiently reliable. For this reason the CE method is usually applied in an iterative manner, starting with a model under which the rare events are not so rare, and increasing the rarity at each step. We start with some $\theta_0 \in \Theta$ and a random variable X_0 whose expectation is easier to estimate than X, and having the same shape. At step $i \geq 0$, n_i independent simulations are performed using IS with parameter θ_i, to approximate the solution of (1.7) with $\tilde{\mathbb{P}}$ replaced by $\tilde{\mathbb{P}}_{\theta_i}$ and X replaced by X_i, where X_i becomes closer to X as i increases, and eventually becomes identical when $i = i_0$, for some finite i_0. In Example 3, for instance, we could have $X_i = 1_{Y_\tau = B_i}$ with $B_i = a + ib$ for some fixed positive integers a and b such that $B = a + i_0 b$ for some i_0. The solution of the corresponding sample average problem is

$$\theta_{i+1} = \arg\max_{\theta\in\Theta} \frac{1}{n_i} \sum_{j=1}^{n_i} |X_i(\omega_{i,j})| \log(d\tilde{\mathbb{P}}_\theta(\omega_{i,j})) \frac{d\mathbb{P}}{d\tilde{\mathbb{P}}_{\theta_i}}(\omega_{i,j}), \tag{1.8}$$

where $\omega_{i,j}$ represents the jth sample at step i. This θ_{i+1} is used for IS at the next step.

By a quick glance at (1.8), we find that the specific choice of the Kullback-Leibler distance is convenient for the case where $\tilde{\mathbb{P}}_\theta$ is from an exponential family, because the log and the exponential cancel, simplifying the solution to (1.8) considerably.

In some specific contexts, the parametric family can be a very rich set of IS measures. For example, in the case of a DTMC over a finite state space, one can define the parametric family as the set of all transition probability matrices over that state space [22]. In this case, CE serves as a technique to approximate the zero-variance change of measure, but at the higher cost of storing an entire transition matrix instead of just the vector $\gamma(\cdot)$.

Bibliography

[1] I. Ahamed, V. S. Borkar, and S. Juneja. Adaptive importance sampling for Markov chains using stochastic approximation. *Operations Research*, 54(3):489–504, 2006.

[2] S. Andradóttir, D. P. Heyman, and T. J. Ott. On the choice of alternative measures in importance sampling with Markov chains. *Operations Research*, 43(3):509–519, 1995.

[3] S. Asmussen. Large deviations in rare events simulation: Examples, counterexamples, and alternatives. In K.-T. Fang, F. J. Hickernell, and H. Niederreiter, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2000*, pages 1–9, Berlin, 2002. Springer-Verlag.

[4] N. Bolia, S. Juneja, and P. Glasserman. Function-approximation-based importance sampling for pricing American options. In *Proceedings of the 2004 Winter Simulation Conference*, pages 604–611. IEEE Press, 2004.

[5] T. E. Booth. Generalized zero-variance solutions and intelligent random numbers. In *Proceedings of the 1987 Winter Simulation Conference*, pages 445–451. IEEE Press, 1987.

[6] P. T. De Boer, P. L'Ecuyer, G. Rubino, and B. Tuffin. Estimating the probability of a rare event over a finite horizon. In *Proceedings of the 2007 Winter Simulation Conference*. IEEE Press, 2007.

[7] P. T. De Boer, V. F. Nicola, and R. Y. Rubinstein. Adaptive importance sampling simulation of queueing networks. In *Proceedings of the 2000 Winter Simulation Conference*, pages 646–655. IEEE Press, 2000.

[8] P. Y. Desai and P. W. Glynn. A Markov chain perspective on adaptive Monte Carlo algorithms. In *Proceedings of the 2001 Winter Simulation Conference*, pages 379–384. IEEE Press, 2001.

[9] B. L. Fox and P. W. Glynn. Discrete-time conversion for simulating semi-Markov processes. *Operations Research Letters*, 5:191–196, 1986.

[10] P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York, 2004.

[11] P. Glasserman and Y. Wang. Counterexamples in importance sampling for large deviations probabilities. *The Annals of Applied Probability*, 7(3):731–746, 1997.

[12] P. W. Glynn. Efficiency improvement techniques. *Annals of Operations Research*, 53:175–197, 1994.

[13] P. W. Glynn and D. L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35:1367–1392, 1989.

[14] G. Goertzel and M. H. Kalos. Monte carlo methods in transport problems. *Progress in Nuclear Energy, Series I*, 2:315–369, 1958.

[15] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation*, 5(1):43–85, 1995.

[16] S. Juneja and P. Shahabuddin. Rare event simulation techniques: An introduction and recent advances. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, Handbooks in Operations Research and Management Science, pages 291–350. Elsevier, Amsterdam, The Netherlands, 2006. Chapter 11.

[17] J. P. C. Kleijnen. Communication: Reply to Fox and Schruben. *Management Science*, 24:1772–1774, 1978.

[18] C. Kollman, K. Baggerly, D. Cox, and R. Picard. Adaptive importance sampling on discrete Markov chains. *Annals of Applied Probability*, 9(2):391–412, 1999.

[19] I. Kuruganti and S. Strickland. Importance sampling for Markov Chains: computing variance and determining optimal measures. In *Proceedings of the 1996 Winter Simulation Conference*, pages 273–280. IEEE Press, 1996.

[20] P. L'Ecuyer and Y. Champoux. Estimating small cell-loss ratios in atm switches via importance sampling. *ACM Transactions on Modeling and Computer Simulation*, 11(1):76–105, 2001.

[21] P. L'Ecuyer and B. Tuffin. Effective approximation of zero-variance simulation in a reliability setting. In *Proceedings of the 2007 European Simulation and Modeling Conference*, pages 48–54, Ghent, Belgium, 2007. EUROSIS.

[22] A. Ridder. Importance sampling simulations of Markovian reliability systems using cross-entropy. *Annals of Operations Research*, 134:119–136, 2005.

[23] R. Rubinstein and D. P. Kroese. *A Unified Approach to Combinatorial Optimization, Monte Carlo Simulation, and Machine Learning*. Springer Verlag, Berlin, 2004.

[24] R. Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operations Research*, 99:89–112, 1997.

[25] J. S. Sadowsky. On the optimality and stability of exponential twisting in Monte Carlo estimation. *IEEE Transactions on Information Theory*, IT-39:119–128, 1993.

[26] D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics*, 4:673–684, 1976.

[27] R. Srinivasan. *Importance sampling – Applications in communications and detection*. Sringer Verlag, Berlin, 2002.