Rare-Event Simulation

Background: Read Chapter 6 of text.

1 Why is Rare-Event Simulation Challenging?

Consider the problem of computing $\alpha = \mathrm{P}(A)$ when $\mathrm{P}(A)$ is small (i.e. "rare"). The crude Monte Carlo estimator estimates α via the proportion α_n of times on which the event A is observed to occur over n independent trials, so that

$$\alpha_n = \frac{1}{n}\sum_{j=1}^{n} I_j,$$

where $I_1, I_2, \ldots, I_n$ are iid copies of $I = I(A)$. The CLT asserts that

$$n^{1/2}(\alpha_n - \alpha) \Rightarrow \sqrt{\alpha(1-\alpha)}\mathcal{N}(0,1)$$

as $n \to \infty$, so that

$$\alpha_n \overset{\mathcal{D}}{\approx} \alpha + \sqrt{\frac{\alpha(1-\alpha)}{n}}\mathcal{N}(0,1) \tag{1.1}$$

for n large. The approximation (1.1) asserts that the "easy problem instances" arise when α is either close to 0 or 1, and that the "hardest problem instances" arise when $\alpha \approx 1/2$. This (perhaps surprising) conclusion arises because (1.1) is describing the absolute error associated with the computation.

But in rare-event simulation, we are concerned with relative error, not absolute error (i.e. we need to know whether the probability is of order 10^{-3} or 10^{-7}, not that it is just close to 0). This relative error is described by the approximation

$$\frac{\alpha_n}{\alpha} - 1 \overset{\mathcal{D}}{\approx} \sqrt{\frac{1-\alpha}{\alpha n}}\mathcal{N}(0,1) \tag{1.2}$$

for n large (which follows immediately from (1.1)). The approximation (1.2) makes clear that crude Monte Carlo requires that

$$n \gg 1/\alpha,$$

in order that the relative error be small. In many settings, this makes rare-event simulation prohibitively expensive.

2 Importance Sampling

The most widely used alternative to crude Monte Carlo computation of rare-event probabilities is "importance sampling". It is most easily illustrated in the setting of a one-dimensional integration

problem, namely in computing

$$\alpha = \int_0^1 f(x)\,\mathrm{d}x.$$

The most obvious way of representing α as an expectation is to write

$$\alpha = \mathrm{E}f(U),$$

where U is uniform on [0,1]. But there are infinitely many alternative representations. Specifically, if $q(\cdot)$ is a density that is positive on [0,1], then we can write

$$\alpha = \int_0^1 \frac{f(x)}{q(x)} q(x)\,\mathrm{d}x = \mathrm{E}\left[\frac{f(Z)}{q(Z)} I(0 \le Z \le 1)\right],$$

where Z has density $q(\cdot)$. Note that

$$\mathrm{var}\left[\frac{f(Z)}{q(Z)} I(0 \le Z \le 1)\right] = \int_0^1 \frac{f(x)^2}{q(x)}\,\mathrm{d}x - \alpha^2.$$

The variance-minimizing choice of q is

$$q^*(x) = \begin{cases} |f(x)| / \int_0^1 |f(y)|\,\mathrm{d}y, & 0 \le x \le 1 \\ 0, & \text{else;} \end{cases}$$

see Exercise 2.1. When f is non-negative, observe that

$$\frac{f(Z)}{q^*(Z)} = \alpha,$$

so that $f(Z)/q^*(Z)$ is deterministic and is a zero-variance estimator of α. Of course, this is not practically implementable. However, it points to a practically useful insight.

Attempt to choose q so that $q(x)$ is proportional to $|f(x)|$.

Importance sampling can easily be generalized to the setting of computing general expectations. Suppose that we wish to compute

$$\alpha = \mathrm{E}X.$$

Note that the expectation operator depends on the probability P, namely

$$\alpha = \int_\Omega X(\omega)\,P(\mathrm{d}\omega). \tag{2.1}$$

To compute α, we can instead choose to sample outcomes ω from an alternative probability, call it Q. The sampling probability (or "measure") Q must be such that there exists a random variable L for which

$$X(\omega)L(\omega)Q(\mathrm{d}\omega) = X(\omega)P(\mathrm{d}\omega) \tag{2.2}$$

for $\omega \in \Omega$. In the presence of (2.2), we may then re-write (2.1) as

$$\alpha = \int_\Omega X(\omega)L(\omega)Q(\mathrm{d}\omega) = \widetilde{\mathrm{E}}XL,$$

where $\widetilde{\mathrm{E}}(\cdot)$ is the expectation operator corresponding to Q. Note that

$$\widetilde{\mathrm{var}}(XL) = \widetilde{\mathrm{E}}(X^2L^2) - \alpha^2.$$

Exercise 2.1: Suppose that $\mathrm{E}X^2 < \infty$. Prove that $\widetilde{\mathrm{var}}(XL)$ is minimized over probabilities Q satisfying (2.2) by the choice of

$$Q^*(\mathrm{d}\omega) = \frac{|X(\omega)|P(\mathrm{d}\omega)}{\mathrm{E}|X|},$$

and that when X is non-negative, XL is a zero-variance estimator for α. (Hint: Use the Cauchy-Schwarz inequality.)

When $X = I(A)$, Exercise 2.1 guarantees that

$$Q^*(\mathrm{d}\omega) = \frac{I_A(\omega)P(\mathrm{d}\omega)}{P(A)} = P(\mathrm{d}\omega|A),$$

so that Q^* is the conditional distribution given the event A. This suggests that in applying importance sampling to rare-event simulation problems, we should select the sampling distribution Q so that Q is close to the conditional distribution, given the rare-event A.

Returning to (2.2), note that it somehow suggests that

$$L(\omega) = \frac{P(\mathrm{d}\omega)}{Q(\mathrm{d}\omega)}$$

on the support of the rv X. In other words, the rv $L(\omega)$ should represent the ratio of the likelihood of ω under the nominal distribution P to that under the sampling distribution Q. For this reason, L is called the *likelihood ratio* of P relative to Q (when P and Q have the "same support"). The notion of the support of a probability is therefore relevant to this discussion. We now clarify this notion.

Definition 2.1: A probability P_1 is *absolutely continuous* with respect to P_2 if whenever $P_2(B) = 0$, then $P_1(B)$.

Thus, absolute continuity of P_1 with respect to P_2 (written $P_1 \ll P_2$) means that the support of P_1 is contained within that of P_2. Two probabilities P_1 and P_2 have the "same support" if $P_1 \ll P_2$ and $P_2 \ll P_1$.

Suppose that

$$P_2(B) \triangleq \int_B |X(\omega)|Q(\mathrm{d}\omega) = 0$$

implies that

$$P_1(B) \triangleq \int_B |X(\omega)|P(\mathrm{d}\omega) = 0,$$

so that $P_1 \ll P_2$. The Radon-Nikodym Theorem then guarantees the existence of a non-negative rv L such that (2.2) holds. In the great majority of applications, the likelihood ratio L can be directly computed (without the need to appeal to the Radon-Nikodym Theorem to assert its existence). Because of the connection of L to the Radon-Nikodym Theorem, L is sometimes called the Radon-Nikodym derivative of P with respect to Q, and is denoted by

$$L(\omega) = \left[\frac{\mathrm{d}P}{\mathrm{d}Q}\right](\omega).$$

3 The Likelihood Ratio for Common Stochastic Models

As asserted above, the relevant likelihood ratio can typically be directly computed in the setting of most stochastic models.

Example 3.1: Let $S = (S_n : n \geq 0)$ be a *random walk*, so that $S_n = S_0 + Z_1 + \cdots + Z_n$, with the Z_i's iid, $\mathbb{R}^d$-valued, and having common density g. To compute

$$\alpha = \mathrm{E}_x f(S_0, S_1, \ldots, S_n),$$

we could alternatively sample $(Z_1, Z_2, \ldots, Z_n)$ from a positive (joint) density h on $\mathbb{R}^{nd}$, and write

$$\alpha = \widetilde{\mathrm{E}}_x \left[f(S_0, S_1, \ldots, S_n) \frac{\prod_{i=1}^n g(S_i - S_{i-1})}{h(S_1 - S_0, \ldots, S_n - S_{n-1})} \right], \tag{3.1}$$

where $\widetilde{E}_x(\cdot)$ is the expectation operator under which $S_0 = x$ and $(Z_1, \ldots, Z_n)$ is sampled from density h. In this case,

$$L = \frac{\prod_{i=1}^n g(S_i - S_{i-1})}{h(S_1 - S_0, \ldots, S_n - S_{n-1})}.$$

Note that if $f(S_0, \ldots, S_n)$ is the indicator rv of a complicated event involving $(S_0, \ldots, S_n)$, the zero-variance importance distribution will generally correspond to a complex non-Markovian distribution. One additional difficulty with the zero-variance distribution is that its non-Markovian structure may make generating random variates from h highly non-trivial from a computational viewpoint.

As a result, one would prefer to use sample distributions that are no more complex than the distribution that governs the nominal model. In the random walk setting, this suggests that we consider a density h under which the increments $Z_1, \ldots, Z_n$ continue to be iid. Specifically, suppose that

$$h(z_1, \ldots, z_n) = \prod_{i=1}^n h(z_i),$$

where h is a positive density on $\mathbb{R}^d$. In this case,

$$\alpha = \widetilde{\mathrm{E}}_x f(S_0, \ldots, S_n) L_n, \tag{3.2}$$

where

$$L_n = \prod_{i=1}^n \frac{g(S_i - S_{i-1})}{h(S_i - S_{i-1})} = \prod_{i=1}^n \frac{g(Z_i)}{h(Z_i)}.$$

When the Z_i's are "light-tailed", there is a particularly natural set of alternative sampling densities that can be used in rare-event simulation settings.

Definition 3.1: An $\mathbb{R}^d$-valued rv Z is said to be *light-tailed* if

$$\mathrm{E} \exp(\theta^T Z) < \infty \tag{3.3}$$

for $\theta \in \mathbb{R}^d$ in some neighborhood of the origin. Otherwise, Z is said to be *heavy-tailed*.

For each θ for which (3.3) holds, note that if $\psi(\theta) = \log \mathrm{E} \exp(\theta^T Z)$, then

$$P_\theta(Z \in \mathrm{d}z) = \exp(\theta^T z - \psi(\theta)) P(Z \in \mathrm{d}z), \tag{3.4}$$

defines an alternative sampling distribution for Z. With this choice of P_θ, the rv L_n further simplifies to

$$L_n = \exp(-\theta^T (S_n - S_0) + n\psi(\theta)).$$

Exercise 3.2: Argue (non-rigorously) that if L_n is to be a function only of $S_n - S_0$ (i.e. of the displacement of the random walk between times 0 and n), then Z_i must have a distribution of the form (3.4) under $\widetilde{\mathrm{P}}_x$.

The sampling distribution (3.4) (also known as "change-of-measure") is said to be obtained from P via "exponential tilting" (or "exponential twisting") and P_θ is called an "exponential change-of-measure".

Exercise 3.3:

a.) Prove that $\psi(\cdot)$ is strictly convex and infinitely differentiable on the interior of $\{\theta : \psi(\theta) < \infty\}$ whenever Z has a positive density on $\mathbb{R}^d$ and is light-tailed.

b.) (continuation of a.)) Prove that for θ_0 in the interior of $\{\theta : \psi(\theta) < \infty\}$,

$$\nabla\psi(\theta)^T|_{\theta=\theta_0} = \mathrm{E}_{\theta_0} Z.$$

c.) (continuation of a.)) Prove that for θ_0 in the interior of $\{\theta : \psi(\theta) < \infty\}$,

$$\left(\frac{\partial^2 \psi}{\partial\theta_i \partial\theta_j}(\theta)\bigg|_{\theta=\theta_0} : 1 \le i,j \le d \right) = (\mathrm{cov}_{\theta_0}(Z_i, Z_j) : 1 \le i,j \le d)$$

The expectations of some random variables involving random walk involve "infinite-dimensional" computations. For example, consider computing

$$\alpha = \mathrm{E}_0 \max_{n\ge 0} S_n,$$

where $(S_n : n \ge 0)$ is a real-valued random walk for which $\mathrm{E}Z_1 < 0$. This is a special case of the more general problem of computing

$$\alpha = \mathrm{E}_x f(S_0, S_1, \ldots)$$

for a non-negative function f that involves the "infinite history" of $(S_n : n \ge 0)$. It is tempting to hope that one can then generalize (3.2) to

$$\alpha = \widetilde{\mathrm{E}}_x f(S_0, S_1, \ldots) L_\infty \tag{3.5}$$

where

$$L_\infty = \prod_{i=1}^{\infty} \frac{g(Z_i)}{h(Z_i)} \tag{3.6}$$

and $(Z_n : n \ge 1)$ is iid with common density h under $\widetilde{\mathrm{P}}_x$.

Exercise 3.4:

a.) Prove that $(L_n : n \ge 0)$ is a non-negative martingale adapted to $(S_n : n \ge 0)$.

b.) Prove that there exists a finite-valued rv L_∞ such that $L_n \to L_\infty$ a.s. as $n \to \infty$. (Hence, L_∞, as defined through (3.6), is a well-defined rv).

We will shortly see that the rv L_∞ must equal 0 a.s. under $\widetilde{\mathrm{P}}_x$. Of course, this implies that the right-hand side of (3.5) is zero, and hence it must be that the generalization of (3.2) to (3.5) fails. Note that

$$\log L_n = \sum_{i=1}^{n} \log\left(\frac{g(Z_i)}{h(Z_i)}\right).$$

It follows that whenever $h \neq g$, Jensen's inequality implies that

$$\frac{1}{n}\log L_n \to \widetilde{\mathrm{E}}_x\left[\log\left(\frac{g(Z_1)}{h(Z_1)}\right)\right] = \int_{\mathbb{R}^d} \log\left(\frac{g(x)}{h(x)}\right) h(x)\,\mathrm{d}x < \log\left(\int_{\mathbb{R}^d} \frac{g(x)}{h(x)} h(x)\,\mathrm{d}x\right) = 0 \quad \text{a.s.}$$

and hence

$$L_n \to 0 \quad \text{a.s.}$$

(at an exponentially fast rate), so that $L_\infty = 0$ $\widetilde{\mathrm{P}}_x$ a.s..

The fundamental reason that (3.5) fails is that P_x is not absolutely continuous with respect to $\widetilde{\mathrm{P}}_x$ when we permit B's that are infinite-dimensional. In other words, we can find an infinite-dimensional event B for which $\widetilde{\mathrm{P}}_x(B) = 0$ but $\mathrm{P}_x(B) > 0$.

Exercise 3.5: Find an infinite-dimensional event B for which $\widetilde{\mathrm{P}}_x(B) = 0$ but $\mathrm{P}_x(B) > 0$.

The above discussion makes clear that importance sampling typically fails on rare-events that depend on the infinite history of the process being simulated. (Of course, this is only a theoretical question, since any such event typically takes infinitely long to simulate, so it is practically impossible to simulate such events.)

On the other hand, while applying importance sampling to computing

$$\alpha = \mathrm{E}_0 \max_{n\geq 0} S_n$$

is impossible, we can apply importance sampling to computing

$$\alpha = \mathrm{P}_0(\max_{n\geq 0} S_n > b). \tag{3.7}$$

In particular, (3.7) equals

$$\alpha = \mathrm{P}_0(T_b < \infty),$$

where

$$T_b = \inf\{n \geq 0 : S_n > b\}$$

is the "level-crossing time" of level b for the random walk $(S_n : n \geq 0)$.

Definition 3.2: A random time $T \in \mathbb{Z}_+$ is said to be a *stopping time* adapted to $(S_n : n \geq 0)$ if for each $n \geq 0$, there exists a (deterministic) function k_n such that

$$I(T = n) = k_n(S_0, \ldots, S_n).$$

Note that T_b is a stopping time that is adapted to $(S_n : n \geq 0)$.

Exercise 3.6: Prove that if f is non-negative, then

$$\mathrm{E}_x f(S_0,\ldots,S_T)I(T<\infty) = \widetilde{\mathrm{E}}_x f(S_0,\ldots,S_T)I(T<\infty)L_T,$$

where

$$L_T = \prod_{i=1}^{T} \frac{g(Z_i)}{h(Z_i)}.$$

It follows that

$$\mathrm{P}_0(T_b<\infty) = \widetilde{\mathrm{E}}_0 I(T_b<\infty)L_{T_b}.$$

Hence, $\mathrm{P}_0(T_b<\infty)$ can be computed by modifying the random walk so that the iid increments are generated under h rather than g, followed by calculating $I(T_b<\infty)L_{T_b}$.

We turn next to computing the likelihood ratio that arises in the setting of Markov chains. We start with the case in which $X=(X_n:n\geq 0)$ is a Markov chain taking values in a discrete state space S and having transition matrix $P=(P(x,y):x,y\in S)$. Suppose that we wish to compute

$$\alpha = \mathrm{E}_x f(X_0,X_1,\ldots,X_n)$$

for some non-negative function f. In the spirit of wishing to sample X using an importance distribution that is as easy to sample from as is the nominal distribution, we consider only Markovian changes-of-measure. In particular, let $(Q_n:n\geq 1)$ be a sequence of transition matrices, and suppose that under the importance distribution $\widetilde{\mathrm{P}}_x$,

$$\widetilde{\mathrm{P}}_x(X_0=x_0,\ldots,X_n=x_n) = \begin{cases} \prod_{i=1}^n Q_i(x_{i-1},x_i), & x_0=x \\ 0, & \text{else.} \end{cases}$$

If $Q_i(x,y)>0$ whenever $P(x,y)>0$, then $P_x(B)=0$ whenever $\widetilde{\mathrm{P}}_x(B)=0$ for any event B determined by $(X_0,X_1,\ldots,X_n)$ (so that $P_x \ll \widetilde{\mathrm{P}}_x$ for such events). Furthermore,

$$\mathrm{E}_x f(X_0,X_1,\ldots,X_n) = \widetilde{\mathrm{E}}_x f(X_0,X_1,\ldots,X_n)L_n, \tag{3.8}$$

where

$$L_n = \prod_{i=1}^{n} \frac{P(X_{i-1},X_i)}{Q_i(X_{i-1},X_i)}.$$

In fact, (3.8) can be generalized to the equality

$$\mathrm{E}_x f(X_0,\ldots,X_T)I(T<\infty) = \widetilde{\mathrm{E}}_x f(X_0,\ldots,X_T)I(T<\infty)L_T$$

where

$$L_T = \prod_{i=1}^{T} \frac{P(X_{i-1},X_i)}{Q_i(X_{i-1},X_i)},$$

provided that T is a stopping time adapted to $(X_n:n\geq 0)$.

Definition 3.3: Let $(S_n:n\geq 0)$ be a real-valued sequence for which $S_n = S_0+Z_1+\cdots+Z_n$, where $Z_i=r(X_i)$ and $X=(X_n:n\geq 0)$ is a Markov chain. Then $(S_n:n\geq 0)$ is a *Markov additive process.*

Exercise 3.7: Suppose that $X = (X_n : n \geq 0)$ is an irreducible finite state Markov chain.

a.) Use the Perron-Frobenius Theorem to conclude that for each $\theta \in \mathbb{R}$, there exists a quantity $\psi(\theta)$ and a function $q_\theta = (q_\theta(x) : x \in \mathbb{R})$ for which

$$q_\theta(x) = \sum_y \exp(\theta r(x) - \psi(\theta)) P(x,y) q_\theta(y)$$

b.) Put

$$Q_\theta(x,y) = \exp(\theta r(x) - \psi(\theta)) P(x,y) \frac{q_\theta(y)}{q_\theta(x)}$$

Prove that $(Q_\theta(x,y) : x, y \in S)$ is a stochastic matrix.

c.) Suppose that X evolves under Q_θ when simulated under $\widetilde{\mathrm{P}}_x$. Prove that

$$L_T = \exp(-\theta(S_T - S_0) + T\psi(\theta)) \frac{q_\theta(X_0)}{q_\theta(X_T)}.$$

This is the analog to exponential twisting for a Markov addition process.

We conclude this section with a further generalization to Markov chains taking values in a continuous state space. Suppose that $X = (X_n : n \geq 0)$ is an $\mathbb{R}^d$-valued Markov chain with transition density $(p(x,y) : x, y \in \mathbb{R}^d)$, so that

$$\mathrm{P}(X_{n+1} \in |X_n) = \int_B p(X_n, y)\, \mathrm{d}y.$$

Suppose that $(q_i(x,y) : x, y \in \mathbb{R}^d)$ is a sequence of transition densities selected so that $q_i(x,y) > 0$ whenever $p(x,y) > 0$, and assume that $\widetilde{\mathrm{P}}_x$ is such that

$$\widetilde{\mathrm{P}}_x(X_{n+1} \in B | X_n) = \int_B q_{n+1}(X_n, y)\, \mathrm{d}y.$$

Then, for any non-negative f,

$$\mathrm{E}_x f(X_0, \ldots, X_T) I(T < \infty) = \widetilde{\mathrm{E}}_x f(X_0, \ldots, X_T) I(T < \infty) L_T$$

where

$$L_T = \prod_{i=1}^{T} \frac{p(X_{i-1}, X_i)}{q_i(X_{i-1}, X_i)},$$

and T is a stopping time adapted to X. Hence, importance sampling can be easily applied to continuous state space Markov chains, with a corresponding likelihood ratio L_T that can be readily (recursively) computed.

4 Zero-Variance Conditional Distribution

In Section 2, we noted that the zero-variance conditional distribution $Q^*(\cdot) = \mathrm{P}(\cdot|A)$ is typically non-Markovian, even when the underlying dynamics are those associated with a random walk. However, in this section, we will see that for certain types of events, the zero-variance conditional distribution is Markovian so that restricting the importance distribution to Markovian changes-of-measure can be done without loss of generality.

Specifically, let

$$T = \inf\{n \geq 0 : X_n \in B\}$$

for a Markov chain $X = (X_n : n \geq 0)$, and set

$$\alpha = \mathrm{P}_x(X_T \in A, T < \infty)$$

for $A \subset B$; such a probability is called an "exit probability". If we set

$$h(x) = \mathrm{P}_x(X_T \in A, T < \infty),$$

note that

$$h(x) = \int_S \mathrm{P}_x(X_1 \in \mathrm{d}y) h(y) \tag{4.1}$$

subject to the boundary conditions

$$h(y) = \begin{cases} 1, & \text{if } y \in A \\ 0, & \text{if } y \in B - A. \end{cases}$$

Equation (4.1) implies that

$$Q^*(x, \mathrm{d}y) = \mathrm{P}_x(X_1 \in \mathrm{d}y) \frac{h(y)}{h(x)}$$

is a Markov transition kernel. Hence, if $\widetilde{\mathrm{P}}_x$ is such that X evolves as a Markov chain having transition kernel Q^*, it follows that for f non-negative,

$$\mathrm{E}_z f(X_0, \ldots, X_T) I(T < \infty) = \widetilde{\mathrm{E}}_z f(X_0, \ldots, X_T) I(T < \infty) L_T \tag{4.2}$$

for $z \in (B - A)^c$, where

$$L_T = \prod_{i=1}^{T} \frac{h(X_{i-1})}{h(X_i)} = \frac{h(X_0)}{h(X_T)}.$$

Since $h = 0$ on $B - A$ and $h = 1$ on A, $h(X_T) = 1$ under $\widetilde{\mathrm{P}}_z$, so that

$$L_T = h(X_0) = h(z)$$

under $\widetilde{\mathrm{P}}_z$. It follows from (4.2) that

$$h(z) = \widetilde{\mathrm{P}}_z(T < \infty) \cdot h(z)$$

and hence $\widetilde{\mathrm{P}}_z(T < \infty) = 1$.

Exercise 4.1: Prove that X, conditional on $\{X_T \in A, T < \infty\}$, evolves according to the transition kernel Q^*.

Exercise 4.2: Prove that X, conditional on $\{X_n \in A\}$, has conditional dynamics

$$\mathrm{P}_x(X_1 \in \mathrm{d}x_1, \ldots, X_n \in \mathrm{d}x_n | A) = \prod_{i=1}^{n} Q_i(x_{i-1}, \mathrm{d}x_i),$$

where $x_0 = x$ and

$$Q_i(x, \mathrm{d}y) = \mathrm{P}_x(X_1 \in \mathrm{d}y) \frac{u(n-i, y)}{u(n-i+1, x)},$$
$$u(j, x) = \mathrm{P}_x(X_j \in A).$$

Exercise 4.1 and 4.2 make clear that under the conditionings postulated in these exercises, the zero-variance conditional dynamics are Markovian.

5 The Level Crossing Problem for Random Walk

We now discuss the application of importance sampling to compute

$$\alpha(b) = \mathrm{P}(T_b < \infty),$$

where $T_b = \inf\{n \geq 0 : S_n > b\}$ and $(S_n : n \geq 0)$ is a light-tailed random walk with $\mathrm{E}Z_1 < 0$. This problem arises in multiple applications settings:

Queueing: If W_n = waiting time (not including service) for the n'th customer to arrive to a single-server infinite capacity waiting room queue in which customers are served according to a FIFO queue discipline,

$$W_n = [W_{n-1} + V_{n-1} - \chi_n]^+$$

where V_{n-1} is the service time for customer $n-1$ and χ_n is the inter-arrival time separating customer $n-1$ and n. If $(V_n : n \geq 0)$ and $(\chi_n : n \geq 1)$ are independent iid sequences with $\mathrm{E}V_0 < \mathrm{E}\chi_1$, then

$$W_n \Rightarrow W_\infty$$

as $n \to \infty$, where

$$\mathrm{P}(W_\infty > b) = \mathrm{P}(T_b < \infty)$$

and T_b is the level crossing for b corresponding to the random walk in which $Z_i = V_{i-1} - \chi_i$.

Insurance: Suppose that b is the capital reserve for an insurance company at time $t = 0$. If the company collects total premiums p in each period but pays out claims Y_i in period i, then the reserve after n periods is given by

$$R_n = b + (p - Y_1) + \cdots + (p - Y_n).$$

Let $\tau = \inf\{n \geq 0 : R_n \leq 0\}$ be the "ruin" time for the company. Note that

$$\mathrm{P}(\tau < \infty) = \mathrm{P}(T_b < \infty),$$

where $(S_n : n \geq 0)$ is a random walk with $S_0 = 0$ and $S_n = Z_1 + \cdots + Z_n$ with $Z_i = Y_i - p$ for $i \geq 1$. Hence, if the Y_i's are iid, computing this ruin probability is a special case of the level crossing problem.

The level crossing problem also arises in sequential analysis, "cusum tests", and in the analysis of certain financial trading strategies.

Note that the spatial homogeneity of random walk guarantees that

$$\alpha(b) = \mathrm{P}_{-b}(T < \infty),$$

where $T = \inf\{n \geq 0 : S_n > 0\}$. It follows that the zero-variance conditional distribution involves Markovian dynamics in which the transition kernel of the random walk is given by

$$Q^*(x, \mathrm{d}y) = \mathrm{P}_x(S_1 \in \mathrm{d}y)\frac{h(y)}{h(x)}, \tag{5.1}$$

where $h(z) = \mathrm{P}_z(T < \infty)$. Suppose that the Z_i's have a density and that there exists $\theta^* > 0$ for which

$$\mathrm{E}\exp(\theta^* Z_1) = 1.$$

Then, the Cramér-Lundberg approximation asserts that

$$\alpha(b) \sim c \exp(-\theta^* b) \tag{5.2}$$

as $b \to \infty$ for some positive constant c; the quantity θ^* is called the Cramér-Lundberg root. It follows that

$$h(y) \sim c \exp(\theta^* y)$$

as $y \to -\infty$.

Given the Cramér-Lundberg approximation and the asymptotic (5.2), it is natural to try to approximate $Q^*(x, \mathrm{d}y)$ via

$$Q(x, \mathrm{d}y) = \frac{\mathrm{P}_x(S_1 \in \mathrm{d}y) c \exp(\theta^* y)}{\int_{\mathbb{R}} \mathrm{P}_x(S_1 \in \mathrm{d}z) c \exp(\theta^* z)}.$$

Note that

$$\int_{\mathbb{R}} \mathrm{P}_x(S_1 \in \mathrm{d}z) c \exp(\theta^* z) = c \mathrm{E}_x \exp(\theta^* S_1) = c \exp(\theta^* x) \mathrm{E} \exp(\theta^* Z_1) = c \exp(\theta^* x).$$

Hence,

$$Q(x, \mathrm{d}y) = \mathrm{P}_x(S_1 \in \mathrm{d}y) \exp(\theta^*(y - x)) = \mathrm{P}(Z_1 \in \mathrm{d}y - x) \exp(\theta^*(y - x))$$

so that the dynamics of $(S_n : n \geq 0)$ under Q are that of a random walk with iid increments in which the increment distribution is modified to

$$\mathrm{P}(Z_1 \in \mathrm{d}z) \exp(\theta^* z). \tag{5.3}$$

If $\widetilde{\mathrm{P}}_x$ is the probability under which the Z_i's are iid with common distribution (5.3), then

$$\mathrm{P}_x(T < \infty) = \widetilde{\mathrm{E}}_x[I(T < \infty) \exp(-\theta^*(S_T - S_0))] \tag{5.4}$$

for $x \leq 0$. According to Exercise 3.3,

$$\widetilde{\mathrm{E}}_x Z_1 = \frac{\partial \psi}{\partial \theta}(\theta)\Big|_{\theta = \theta^*}. \tag{5.5}$$

Since $\psi(\cdot)$ is strictly convex with $\mathrm{E}Z_1 = \frac{\partial \psi}{\partial \theta}(\theta)|_{\theta=0} < 0$, evidently the derivative of $\psi(\cdot)$ must be positive at the θ^* satisfying $\psi(\theta^*) = 0$. Hence, $(S_n : n \geq 0)$ is a positive drift random walk under $\widetilde{\mathrm{P}}_x$, so that $\widetilde{\mathrm{P}}_x(T < \infty) = 1$. Consequently, (5.4) reduces to

$$\mathrm{P}_x(T < \infty) = \widetilde{\mathrm{E}}_x \exp(-\theta^*(S_T - S_0)) = \exp(\theta^* x) \widetilde{\mathrm{E}}_x \exp(-\theta^* S_T).$$

This leads to the following algorithm (due to Siegmund (1976)) for computing $\alpha(b)$:

1. Simulate $Z_1, Z_2, \ldots$ as an iid sequence with distribution (5.3) until $Z_1 + \cdots + Z_n$ exceeds b at time T_b.

2. Compute

$$\beta = \exp(-\theta^* S_{T_b}).$$

3. Replicate $\beta_1, \beta_2, \ldots, \beta_n$ and estimate $\alpha(b)$ via $\overline{\beta}_n$.

Exercise 5.1: Suppose that $(S_n : n \geq 0)$ is a Markov additive process associated with a finite state space irreducible Markov chain. Assume that

$$\sum_x \pi(x) r(x) < 0,$$

where $\pi = (\pi(x) : x \in S)$ is the stationary distribution of X. Develop an analog to Siegmund's algorithm for computing $\alpha(b)$ when b is large.

Exercise 5.2: Use (5.4) to develop an algorithm for sampling $(S_j : 0 \leq j \leq T)$, conditional on $\{T < \infty\}$.

6 Efficiency for Rare-Event Simulations

In the level-crossing problem, the formulation leads naturally to the consideration of a parameterized family of problem instances $(\alpha(b) : b > 0)$. In general, given such a family of problem instances, we say that a corresponding family of estimators $(\beta(b) : b > 0)$ exhibits *bounded relative variance* if

$$\sup_b \frac{\mathrm{var}\beta(b)}{\alpha(b)^2} < \infty. \tag{6.1}$$

Note that in the presence of bounded relative variance, Chebyshev's inequality implies that

$$\mathrm{P}\left(\left|\frac{\overline{\beta}_n(b)}{\alpha(b)} - 1\right| > \epsilon\right) \leq \frac{\mathrm{var}\beta(b)}{n\alpha(b)^2\epsilon^2}$$

Hence, bounded relative variance implies that the number of samples n of $\beta(b)$ required to compute $\alpha(b)$ to relative precision ϵ is uniformly bounded in b. An algorithm having this boundedness property is said to be *efficient*.

A weaker (and generally easier to be prove) notion of efficiency involves requiring that for each $\epsilon > 0$,

$$\sup_b \frac{\mathrm{var}\beta(b)}{\alpha(b)^{2-\epsilon}} < \infty. \tag{6.2}$$

Of course, (6.1) and (6.2) are equivalent when $\inf\{\alpha(b) : b > 0\} > 0$. But this corresponds to the case in which there is a lower bound on the rarity of the event being computed. The more interesting and important setting is that in which $\alpha(b) \to 0$ as $b \to \infty$. In this context, (6.2) is equivalent to requiring that

$$\frac{\log \mathrm{var}\beta(b)}{\log \alpha(b)} \to 2 \tag{6.3}$$

as $b \to \infty$. Consequently, an algorithm satisfying (6.3) is said to be *logarithmically efficient*.

Exercise 6.1: Prove that Siegmund's algorithm of Section 5 is efficient.

Exercise 6.2: Prove that the generalization of Siegmund's algorithm developed in Exercise 5.1 is efficient.

Because $\mathrm{var}\beta(b) = \mathrm{E}\beta(b)^2 - \alpha(b)^2$, (6.3) is seen to be equivalent to

$$\frac{\log \mathrm{E}\beta(b)^2}{\log \alpha(b)} \to 2$$

as $b \to \infty$, reducing (6.3) to a criterion involving only the second moment of the estimator.

7 Large Deviations for Random Walk

Because so many stochastic models are built up from random walk, it is instructive to first develop a comprehensive understanding of the rare-event behavior of random walk.

In Section 4, we computed the zero-variance conditional distribution of $(S_j : 0 \leq j \leq n)$ given $S_n \in A$. We now analyze this conditional distribution in greater detail in the setting where $A = [na, \infty)$ with $a > \mathrm{E}Z_1$. Observe that the law of large numbers guarantees that

$$\mathrm{P}_0(S_n > na) \to 0$$

as $n \to \infty$, so that $\{S_n > na\}$ is an unusual "large deviation" for the random walk.

Suppose that $(S_n : n \geq 0)$ is a light-tailed random walk. Observe that we can bound $\mathrm{P}_0(S_n > na)$ as follows:

$$\mathrm{P}_0(S_n > na) \leq \exp(-\theta na)\mathrm{E}_0 \exp(\theta S_0) = \exp(-\theta na + n\psi(\theta)),$$

for $\theta > 0$. Hence,

$$\mathrm{P}_0(S_n > na) \leq \inf_{\theta>0} \exp(-\theta na + n\psi(\theta)) = \exp(-n \sup_{\theta>0}(\theta a - \psi(\theta))) = \exp(-n(\theta_a a - \psi(\theta_a))), \tag{7.1}$$

where θ_a is such that

$$\psi'(\theta_a) = a. \tag{7.2}$$

But recall that $\psi'(\theta_a) = \mathrm{E}_{\theta_a} Z$, so (7.2) asserts that θ_a is the exponential twisting parameter for which

$$\mathrm{E}_{\theta_a} Z = a,$$

so that the random walk has mean drift exactly equal to a when twisted under θ_a. This calculation suggests that perhaps the twisted random walk with twisting parameter θ_a is connected somehow to the conditional distribution of $(S_j : 0 \leq j \leq n)$ conditional on $S_n > na$. To verify this, suppose that $\widetilde{\mathrm{P}}_0(\cdot)$ is the probability under which $S_0 = 0$ and the iid increments $Z_1, Z_2, \ldots, Z_n$ are generated from the distribution

$$\exp(\theta_a z - \psi(\theta_a))\mathrm{P}(Z_i \in \mathrm{d}z).$$

Then,

$$\mathrm{P}_0(S_n > na) = \widetilde{\mathrm{E}}_0 I(S_n > na) \exp(-\theta_a S_n + n\psi(\theta_a))$$

For each $\epsilon > 0$, the central limit theorem ensures that

$$\widetilde{\mathrm{P}}_0(na < S_n < n(a+\epsilon)) \to \frac{1}{2}$$

as $n \to \infty$, so that

$$\widetilde{\mathrm{E}}_0 I(S_n > na) \exp(-\theta_a S_n + n\psi(\theta_a)) \geq \widetilde{\mathrm{E}}_0 I(na < S_n < n(a+\epsilon)) \exp(-\theta_a n(a+\epsilon) + n\psi(\theta_a)).$$

It follows that for each $\epsilon > 0$,

$$\varliminf_{n\to\infty} \frac{1}{n} \log \mathrm{P}_0(S_n > na) \geq -\theta_a(a+\epsilon) + \psi(\theta_a). \tag{7.3}$$

Combining (7.3) with (7.2) yields the conclusion that

$$\frac{1}{n} \log \mathrm{P}_0(S_n > na) \to -\theta_a a + \psi(\theta_a) \tag{7.4}$$

as $n \to \infty$. The "logarithmic asymptotic" (7.4) is the most basic result in that branch of probability theory known as "large deviations theory" (basically, the probability of rare events for light-tailed stochastic models).

A more delicate analysis proves that

$$\mathrm{P}_0(Sn > na) \sim \frac{1}{\sqrt{2\pi\psi''(\theta_a)n}} \exp(-n(\theta_a a - \psi(\theta_a))) \tag{7.5}$$

as $n \to \infty$.

Exercise 7.1:

a.) Use (7.5) to prove that

$$\mathrm{P}_0(Z_1 \in \mathrm{d}z_1, \ldots, Z_l \in \mathrm{d}z_l | S_n > na) \Rightarrow \prod_{i=1}^{l} \exp(\theta_a z_i - \psi(\theta_a))\mathrm{P}(Z_i \in \mathrm{d}z_i)$$

as $n \to \infty$.

b.) Relate part a.) to the functions $u(j, x)$ and kernels $Q_1, Q_2, \ldots$ introduced in Exercise 4.2.

This suggests that the most efficient means of computing $\mathrm{P}_0(S_n > na)$ for large n involves the following algorithm:

1. Simulate $Z_1, Z_2, \ldots, Z_n$ as an iid sequence with common distribution

$$\exp(\theta_a z - \psi(\theta_a))\mathrm{P}(Z \in \mathrm{d}z).$$

2. Compute

$$\beta(n) = I(S_n > na) \exp(-\theta_a S_n + n\psi(\theta_a))$$

3. Replicate $\beta(n)$ m iid times, thereby generating $\beta_1(n), \ldots, \beta_m(n)$ and estimate $\mathrm{P}_0(S_n > na)$ via $\overline{\beta}_m(n)$.

Exercise 7.2: Prove that the above algorithm is logarithmically efficient.

Remark 7.1: The above algorithm is not efficient. However, in practice, it behaves well and leads to (very) large variance reductions.

Exercise 7.3:

a.) Explicitly describe the above algorithm when the Z_i's are Gaussian rv's (including variate generation).

b.) Explicitly describe the above algorithm when the Z_i's are gamma distributed rv's (including variate generation).

c.) Explicitly describe the above algorithm when the Z_i's are uniform rv's (including variate generation)

Exercise 7.4: Generalize the above algorithm to a Markov additive process for which $(X_n : n \geq 0)$ is an irreducible finite state Markov chain.

Exercise 7.5: Prove that under the conditions above, there exists $\gamma > 0$ such that

$$\mathrm{P}_0(S_n > na + x | S_n > na) \to \exp(-\gamma x)$$

as $n \to \infty$, so that the "overshoot" $S_n - na$ conditional on $S_n > na$ is bounded in n as $n \to \infty$. (Hence, conditioning on $S_n \in [na, na + x]$ should lead to results and algorithms quite similar to those involving conditioning on $S_n > na$.)

8 More Complex Computations Involving Random Walk

Let us see how we can use the large deviations theory of Section 7 to see why the exponential twist θ^* naturally appears in the level crossing problem.

We know that when b is large, the approximation (7.4) suggests that

$$\mathrm{P}_0(S_{\lfloor t \rfloor} > b) \approx \exp\left(-\theta(t)\left(\frac{b}{t}\right)t + t\psi(\theta(t))\right),$$

where $\theta(t)$ is the root of

$$\psi'(\theta(t)) = \frac{b}{t}.$$

We now search for the "most likely" value of t by seeking to maximize

$$\exp(-\theta(t)b + t\psi(\theta(t)))$$

over t. Note that

$$\frac{\mathrm{d}}{\mathrm{d}t}(\theta(t)b - t\psi(\theta(t))) = \theta'(t)b - \psi(\theta(t)) - t\psi'(\theta(t))\theta'(t) = \theta'(t)b - \psi(\theta(t)) - t\left(\frac{b}{t}\right)\theta'(t) = -\psi(\theta(t))$$

so that the maximizing t^* satisfies

$$\psi(\theta(t^*)) = 0.$$

In other words, we should select the twisting parameter θ^* for which $\psi(\theta^*) = 0$. Hence, large deviations permits us to compute the appropriate exponential twist to use for our importance distribution.

Exercise 8.1: Let $(S_n : n \geq 0)$ be an $\mathbb{R}^2$-valued random walk with light-tailed increments with $\mathrm{E}Z_1 < 0$. Suppose that $B \subseteq \mathbb{R}^2_+$ is a convex set with a smooth boundary.

a.) Develop an importance sampling algorithm for computing

$$\alpha = \mathrm{P}_0(S_n \in nB)$$

when n is large.

b.) Provide conditions under which your algorithm is logarithmically efficient.

Exercise 8.2: Let $(S_n : n \geq 0)$ be a real-valued random walk with light-tailed increments for which $\mathrm{E}Z_1 = 0$. Suppose $B = (-\infty, b] \bigcup (a, \infty)$ with $b < 0 < a$. (Note that B is non-convex.)

a.) Compute the limit of

$$\frac{1}{n} \log \mathrm{P}_0(S_n \in nB)$$

as $n \to \infty$.

b.) Compute the limit distribution of

$$\mathrm{P}_0((Z_1, \ldots, Z_k) \in \cdot | S_n \in nB)$$

as $n \to \infty$.

c.) Suppose that you use the importance sampling algorithm suggested by your computation in b.). Prove that your algorithm is not logarithmically efficient.

d.) How does the algorithm in c.) compare in efficiency to crude Monte Carlo?

Exercise 8.3: Let

$$p_1(j, x) = \exp(-\theta_1(j, x)(b - x) + (n - j)\psi(\theta_1(j, x)))$$
$$p_2(j, x) = \exp(-\theta_2(j, x)(b - x) + (n - j)\psi(\theta_2(j, x)))$$

be the approximations to the exit probabilities at the boundaries bn and an, respectively. We then generate Z_{j+1} from

$$\exp(\theta_1(j, x)z - \psi(\theta_1(j, x)))\mathrm{P}(Z \in \mathrm{d}z)$$

with probability $p_1(j, x)/(p_1(j, x) + p_2(j, x))$ and from

$$\exp(\theta_2(j, x)z - \psi(\theta_2(j, x)))\mathrm{P}(Z \in \mathrm{d}z)$$

with probability $p_2(j, x)/(p_1(j, x)+p_2(j, x))$. (In other words, this algorithm uses a *state-dependent twist* in which the algorithm attempts to push the random walk towards the exit boundaries with probability proportional to the large deviations exit probability approximations.) Compare this algorithm to that of Exercise 8.2 empirically.