

# Importance Sampling for Rare Events

Soren Asmussen  
Department of Mathematical Sciences  
Aarhus University  
Ny Munkegade  
DK-8000 Aarhus C, Denmark  
asmus@imf.au.dk

Paul Dupuis  
Division of Applied Mathematics  
Brown University  
Providence, R.I. 02912, U.S.A.  
dupuis@dam.brown.edu

Reuven Rubinstein  
Industrial Engineering and Management  
Technion – Israel Institute of Technology  
Technion City, Haifa 32000, Israel  
ierrr01@ie.technion.ac.il

Hui Wang  
Division of Applied Mathematics  
Brown University  
Providence, R.I. 02912, U.S.A.  
huiwang@dam.brown.edu

## 1 Introduction and Background

The estimation of rare event probabilities is probably one of the most challenging topics in Monte Carlo simulation. Interest in rare events arises from many branches of science. Examples include performance analysis in communication theory and computer science where extremely small buffer overflow probabilities are of concern, chemical physics where the transition probabilities from one metastable state to another plays a key role, and risk

management where measuring rare but catastrophic losses is a prerequisite. Under these circumstances, one is often interested in both qualitative and quantitative information directly related to the rare event, such as how likely is the rare event and, given that it does occur, how does it happen.

To illustrate the inefficiency of standard Monte Carlo in simulating rare events, consider a simple example. Let  $X$  be a random variable defined on some probability space  $(\Omega, \mathcal{F}, P)$ . Suppose that one is interested in estimating the probability that  $X$  is in some given set  $A$ :

$$p = P(X \in A).$$

Standard Monte Carlo would generate  $k$  independent identically distributed samples  $\{X_i : i = 1, \dots, k\}$  from the distribution of  $X$  and form an unbiased estimate

$$\hat{p}_k = \frac{1}{k} \sum_{i=1}^k 1_A(X_i),$$

where  $1_A$  is the indicator function of the set  $A$ :

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

As the sample size  $k$  tends to infinity, the estimate  $\hat{p}_k$  converges to  $p$  with probability one by the strong law of large numbers. The rate of convergence is determined by the variance of  $1_A(X)$ . More precisely, by the central limit theorem, the distribution of  $\hat{p}_k$  is approximately normal with mean  $p$  and variance

$$\text{Var}[\hat{p}_k] = \frac{1}{k} \text{Var}[1_{\{X \in A\}}] = \frac{1}{k} p(1-p).$$

Even though this variance is very small when  $p$  is very small, the *relative error* associated with the estimate  $\hat{p}_k$

$$\text{relative error} = \frac{\text{standard deviation of } \hat{p}_k}{\text{mean of } \hat{p}_k} = \frac{\sqrt{p-p^2}}{\sqrt{k}p}$$

can be very large. Indeed, the relative error is unbounded as the event  $A$  becomes rarer. Therefore, a large number of samples are required in order to achieve a fixed relative error bound.

Two major classes of techniques to improve the efficiency of estimating small probabilities are importance sampling and particle splitting. When properly designed, both algorithms can dramatically reduce the number of samples needed to achieve the desired precision. Due to space constraints,

this paper will focus on importance sampling, even though it should be said that particle splitting is related to importance sampling in an unexpected way via subsolutions. See Glasserman et al. (1999), Dean and Dupuis (2009), Rubinstein (2010), and the references therein. For the same reason, we will not be able to discuss other interesting variance reduction techniques such as conditional Monte Carlo (Asmussen and Glynn, 2007).

The basic idea of importance sampling is to simulate the system based on an alternative probability distribution [i.e., change of measure] and an unbiased estimate is formed by multiplying the original estimate by an appropriate likelihood ratio. This technique was first applied to nuclear-physics calculation around 1940's and has been an area of active research for the last two decades. See, e.g., Heidelberger (1995) and Asmussen and Rubinstein (1995), for surveys. In this paper we review some of the recent developments of this methodology. Our purpose is to emphasize basic concepts and innovative ideas, without much attention paid to mathematical rigor. The precise statements of the theorems and their rigorous proofs can be found in the relevant references.

The paper is organized as follows. In Section 2, we describe two efficiency criteria for Monte Carlo simulation algorithms. In Section 3 we set up two examples that will be frequently used throughout the paper. In Section 4, we discuss several different techniques for the design of the change of measure in importance sampling, including the cross-entropy method, the game/subsolution approach in dynamic importance sampling, and the Lyapunov function method for heavy tailed distributions.

## 2 Efficiency Criteria

There are two commonly used criteria for the performance of a Monte Carlo algorithm in rare event simulation. Consider a family of rare event probabilities  $\{p_n\}$  such that  $p_n \rightarrow 0$  as  $n \rightarrow \infty$ . One can think of  $n$  as an index for rarity. For example,  $p_n$  may denote the probability that a one dimensional simple random walk with negative drift ever crosses a large threshold  $n$ , starting at the origin.

Consider a Monte Carlo algorithm for estimating  $p_n$ , where the estimate is the sample mean of independent copies of some random variable  $Y_n$  that satisfies  $E[Y_n] = p_n$ . Then the estimate is unbiased. We say that the estimate has *bounded relative error* if

$$\limsup_{n \rightarrow \infty} \frac{\text{Var}[Y_n]}{p_n^2} < \infty.$$

It is not difficult to show that the number of samples required to achieved a fixed relative error remains bounded as  $n$  increases.

Since  $Y_n$  is unbiased, minimizing its variance is equivalent to minimizing its second moment. By Jensen's inequality,  $E[Y_n^2] \geq (EY_n)^2 = p_n^2$ . This motivates a weaker notion of efficiency, namely, the *logarithmic asymptotic efficiency* or *asymptotic efficiency*, which holds if

$$\lim_{n \rightarrow \infty} \frac{\log E[Y_n^2]}{\log p_n} = 2. \quad (2.1)$$

This criterion is particularly convenient when the rare event probabilities  $\{p_n\}$  satisfy the large deviation asymptotics

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log p_n = -\gamma,$$

where  $\gamma > 0$  is some constant. In this situation, logarithmic asymptotic efficiency amounts to

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log E[Y_n^2] = -2\gamma,$$

and implies that the number of samples required to achieve a fixed relative error grow sub-exponentially as  $n$  increases. In the literature, logarithmic asymptotic efficiency is sometimes referred to as *asymptotic optimality*.

### 3 Two illustrative examples

Even though the methodologies we are going to discuss can be applied to general settings, it is perhaps best to convey the main ideas through some concrete examples. The purpose of this section is to describe two examples that will be used repeatedly later in the paper to illustrate various Monte Carlo schemes.

**Simple Random Walk [SRW].** Let  $\{Z_i\}$  be a sequence of  $\mathbb{R}^d$ -valued, independent identically distributed random variables with distribution  $\mu$ , and assume that the log-moment generating function

$$H(\alpha) = \log E[e^{\langle \alpha, Z_1 \rangle}] = \log \int_{\mathbb{R}^d} e^{\langle \alpha, x \rangle} \mu(dx)$$

is finite for every  $\alpha \in \mathbb{R}^d$ . Define for  $n \geq 1$ ,  $S_n = Z_1 + \dots + Z_n$ . For some Borel set  $A \subset \mathbb{R}^d$ , we are interested in estimating the probability

$$p_n = P\left(\frac{S_n}{n} \in A\right).$$

Under some mild conditions, the large deviations asymptotics hold, namely,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log p_n = - \inf_{\beta \in A} L(\beta), \quad (3.2)$$

where  $L$  is the Legendre transform of  $H$ :

$$L(\beta) = \sup_{\alpha \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - H(\alpha)]. \quad (3.3)$$

**Tandem Queueing Network [TQN].** Consider a two-node tandem Jackson queueing network, where the arrival process is Poisson with rate  $\lambda$  and the service times are exponentially distributed with rate  $\mu_1$  and  $\mu_2$ , respectively. The system is assumed to be stable, that is,  $\lambda < \min\{\mu_1, \mu_2\}$ . See Figure 1.

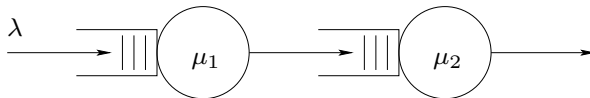


Figure 1: Two-node tandem queue

Assume that the two queues share a single buffer with total capacity  $n$ . We are interested in the buffer overflow probability

$$p_n = P \{ \text{network total population reaches } n \text{ before returning to } 0, \\ \text{starting from } 0 \}.$$

Glasserman and Kou (1995) established the large deviation limit

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log p_n = - \log \frac{\min\{\mu_1, \mu_2\}}{\lambda}.$$

## 4 Importance Sampling

The basic setup of importance sampling is as follows. Suppose that we are interested in estimating

$$p = P(X \in A),$$

where  $X$  is a random variable with distribution  $\mu$ . Importance sampling generates samples from a different probability distribution  $\nu$  and uses the sample mean of independent copies of

$$Y = 1_{\{X \in A\}} \frac{d\mu}{d\nu}(X)$$

as the estimate. One usually requires that  $\mu$  be absolutely continuous with respect to  $\nu$  so that the likelihood ratio  $d\mu/d\nu$  is well defined. This requirement can be relaxed as long as the absolute continuity holds for the restrictions of  $\mu$  and  $\nu$  to the set  $A$ . Note that the estimate  $Y$  is unbiased since

$$E_\nu[Y] = \int_A \frac{d\mu}{d\nu}(x) d\nu(x) = \int_A d\mu(x) = P(X \in A).$$

Here we have used  $E_\nu[\cdot]$  to denote the expectation taken under the probability distribution  $\nu$ .

The key question in importance sampling is the choice the sampling distribution  $\nu$ . Ideally, one would like to find the one that minimizes the variance of  $Y$ . To this end, define a measure  $\nu^*$  such that

$$\frac{d\nu^*}{d\mu}(x) = \frac{1}{p} \cdot 1_{\{x \in A\}}.$$

It is not difficult to verify that  $\nu^*$  is a probability distribution and the corresponding importance sampling estimator  $Y$  has variance zero. However, such a probability measure is of little practical use since it requires the knowledge of  $p$ , the quantity we wish to estimate. Therefore, instead of this unconstrained optimization, it is typical to search within a parameterized family of alternative probability measures. When the problem can be cast into the framework of Section 2, it is desirable for the estimator to achieve logarithmic asymptotic efficiency or bounded relative error.

**Remark 4.1.** For future analysis, observe that the second moment of the importance sampling estimate  $Y$  admits a very simple form

$$E_\nu[Y^2] = \int_A \left( \frac{d\nu}{d\mu} \right)^2(x) d\nu(x) = \int_A \frac{d\nu}{d\mu}(x) d\mu(x) = E_\mu[Y].$$

## 4.1 Classical results in importance sampling

Siegmund (1976) was the first to argue that, using an exponential change of measure, asymptotically efficient importance sampling schemes can be built for estimating gambler's ruin probabilities. The analysis was related to the theory of large deviations, which has since become an indispensable tool for the design of efficient Monte Carlo algorithms.

To illustrate the idea, consider the example of SRW where  $A$  is assumed to be a closed *convex* set. Suppose that instead of generating the increments

$\{Z_i\}$  according to  $\mu$ , we sample  $\{Z_i\}$  from an exponential change of measure  $\nu_\alpha$  where

$$\nu_\alpha(dx) = e^{\langle \alpha, x \rangle - H(\alpha)} \mu(dx)$$

for some  $\alpha \in \mathbb{R}^d$ . The importance sampling estimate is

$$Y_n = 1_{\{S_n/n \in A\}} \prod_{i=1}^n e^{-\langle \alpha, Z_i \rangle + H(\alpha)} = 1_{\{S_n/n \in A\}} e^{-\langle \alpha, S_n \rangle + nH(\alpha)}.$$

Taking into account Remark 4.1 and application of Varadhan's Lemma, the second moment of  $Y_n$  satisfies

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log E_{\nu_\alpha}[Y_n^2] = \lim_{n \rightarrow \infty} \frac{1}{n} \log E_\mu[Y_n] = - \inf_{\beta \in A} [\langle \alpha, \beta \rangle - H(\alpha) + L(\beta)].$$

The  $\alpha^*$  that minimizes the right-hand-side of the above display yields the asymptotically *most efficient* exponential change of measure [say  $\nu^* = \nu_{\alpha^*}$ ], and is the solution to the min/max problem

$$\sup_{\alpha \in \mathbb{R}^d} \inf_{\beta \in A} [\langle \alpha, \beta \rangle - H(\alpha) + L(\beta)]. \quad (4.4)$$

Since  $A$  is closed and convex, it is valid to exchange of the order of the sup and inf in the above expression. Then it follows from (3.3) that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log E_{\nu^*}[Y_n^2] = - \inf_{\beta \in A} \sup_{\alpha \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - H(\alpha) + L(\beta)] = -2 \inf_{\beta \in A} L(\beta).$$

In other words,  $\nu^* = \nu_{\alpha^*}$  is logarithmic asymptotically efficient. Furthermore, if  $\beta^*$  minimizes  $L(\beta)$  over  $\beta \in A$ , then  $\alpha^*$  can be identified as the conjugate point of  $\beta^*$  or the point that maximizes  $\langle \alpha, \beta^* \rangle - L(\beta^*)$  over  $\alpha \in \mathbb{R}^d$ .

It turns out that  $\nu^*$  coincides with the change of measure used in the classical proof of the large deviations lower bound for the rare event probabilities  $P(S_n/n \in A)$ . This formal connection between importance sampling and the theory of large deviations has been subsequently explored by many and made rigorous under certain circumstances. See, e.g., Asmussen (1985), Heidelberger (1993), Asmussen and Glynn (2007) and the references therein. These investigations gave rise to an entirely new community using exponential change of measure as the driving force for importance sampling.

Glasserman and Kou (1995) was the first to challenge the standard heuristic that the change of measure used in the proof of the large deviation lower bound should perform well. The paper considered a change of

measure proposed by Parekh and Walrand (1989) for the example of TQN, which amounts to interchanging the arrival rate and the smallest service rate, and showed that it failed to be asymptotically efficient in general. In Glasserman and Wang (1997), counterexamples were constructed, such as SRW with a non-convex target set  $A$ , to show that the importance sampling estimator based on the standard heuristic can be less efficient than the standard Monte Carlo. In retrospect, the failure of the standard heuristic even in very simplistic settings is not surprising. In the previous analysis of the SRW model, a key assumption is that  $A$  is convex so that the sup and inf in (4.4) can be interchanged. This is clearly not true when  $A$  is a general non-convex set. The work of Glasserman and Kou (1995) and Glasserman and Wang (1997) made it clear that the standard heuristic had to be applied with great caution and motivated the development of general methodologies such as dynamic importance sampling. We will review some of these development later in the paper.

## 4.2 Cross-entropy method

The cross-entropy method is a relatively new Monte Carlo technique that originated from a sequence of papers Rubinstein (1997, 1999). It is a very powerful and versatile technique that can be used not only for estimating rare event probabilities, but also for solving difficult combinatorial optimization problems. See de Boer et al. (2005) for a tutorial and Rubinstein and Kroese (2004) for a comprehensive treatment.

Consider the generic importance sampling problem for estimating  $p = P(X \in A)$ , where  $X$  is a random variable with distribution  $\mu$ . When looking for an alternative sampling distribution, we will restrict ourselves to a prescribed, parameterized family of distributions, say  $\{\mu_\theta : \theta \in \Theta\}$ , that contains the original distribution  $\mu$ . The reference parameter  $\theta$  is sometimes termed the *tilting parameter*. As discussed previously, the zero-variance change of measure  $\nu^*$  is defined by

$$\frac{d\nu^*}{d\mu} = \frac{1}{p} \cdot 1_{\{x \in A\}}. \quad (4.5)$$

Under the natural assumption that a sampling distribution “close” to  $\nu^*$  should be a good choice for importance sampling, the cross-entropy method aims to solve for the distribution  $\mu_\theta$  that is closest to  $\nu^*$  under the Kullback-Leibler distance. This leads to the minimization problem

$$\min_{\theta \in \Theta} R(\nu^* \parallel \mu_\theta), \quad (4.6)$$

where  $R(\cdot\|\cdot)$  is the *Kullback-Leibler cross-entropy*, or *relative entropy*, defined by

$$R(\nu\|\mu) = \int \log \frac{d\nu}{d\mu}(x) d\nu(x)$$

if  $\nu$  is absolutely continuous with respect to  $\mu$  and  $\infty$  otherwise. Note that  $R(\nu\|\mu)$  is always non-negative and equals zero if and only if  $\nu = \mu$ .

The cross-entropy method provides a simple iterative procedure to obtain a solution to the optimization problem (4.6). Every iteration involves two phases: (1) samples are generated from the distribution  $\mu_{\theta_t}$  where  $\theta_t$  is the current candidate of the tilting parameter; (2) based on these samples, the tilting parameter  $\theta_t$  is updated to  $\theta_{t+1}$  in order to produce better samples in the next iteration. The iteration is terminated when the convergence of  $\{\theta_t\}$  is reached. Suppose that  $\theta^*$  is the final tilting parameter. Then  $\mu_{\theta^*}$  is used as the importance sampling change of measure to estimate  $p$ , the probability of interest.

A big advantage of the cross-entropy method is that  $\theta_{t+1}$  can often be solved *analytically*. In particular, this happens when the distributions  $\{\mu_{\theta}\}$  belong to the family of exponential changes of measure. See Subsection 4.2.1 for more details.

The initialization  $\theta_0$  of the cross-entropy algorithms is quite flexible in general. For example, in many situations one can simply choose  $\theta_0$  that corresponds to the original distribution  $\mu$ . However, in the context of rare event simulation, the choice of  $\theta_0$  becomes less straightforward. We will discuss these issues in Subsection 4.2.2.

### 4.2.1 The adaptive updating of $\theta$

Consider the minimization problem (4.6). Denote by  $W_{\theta}$  the likelihood ratio function

$$W(x; \theta) = \frac{d\mu}{d\mu_{\theta}}(x).$$

Plugging in the formula (4.5) we have

$$R(\nu^*\|\mu_{\theta}) = \int \log \frac{d\nu^*}{d\mu_{\theta}}(x) d\nu^*(x) = -\log p + \frac{1}{p} \int 1_{\{x \in A\}} \log W(x; \theta) d\mu(x).$$

It follows that the minimization problem (4.6) amounts to minimizing over  $\theta \in \Theta$  the integral

$$\int 1_{\{x \in A\}} \log W(x; \theta) d\mu(x).$$

Now let  $\gamma \in \Theta$  be an *arbitrary* reference parameter. Then the above integral equals

$$\int 1_{\{x \in A\}} W(x; \gamma) \log W(x; \theta) d\mu_\gamma(x) = E_\gamma [1_{\{X \in A\}} W(X; \gamma) \log W(X; \theta)],$$

where  $E_\gamma[\cdot]$  means that the expectation is taken with  $X$  distributed according to  $\mu_\gamma$ . Therefore the minimization problem (4.5) is equivalent to the minimization problem

$$\min_{\theta \in \Theta} E_\gamma [1_{\{X \in A\}} W(X; \gamma) \log W(X; \theta)], \quad (4.7)$$

for any arbitrarily fixed  $\gamma \in \Theta$ . In the cross-entropy method, the minimizing  $\theta$  is estimated by solving the corresponding stochastic program

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \gamma) \log W(X_i; \theta), \quad (4.8)$$

where  $\{X_1, \dots, X_N\}$  are independent samples from the distribution  $\mu_\gamma$ . The function in (4.8) is convex and differentiable with respect to  $\theta$  in typical applications. Thus the minimizing  $\theta$  is the solution to the equation

$$\frac{1}{N} \sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \gamma) \nabla \log W(X_i; \theta) = 0,$$

where the gradient  $\nabla$  is with respect to  $\theta$ .

Now we can state the basic adaptive updating rule for the tilting parameter  $\theta$  in the cross-entropy method. We will use the stochastic program (4.8) in lieu of the deterministic program (4.7).

**The basic updating rule of  $\theta$ .** Suppose  $\hat{\theta}_t$  is the value of the tilting parameter at the end of the last iteration. Generate independent samples  $\{X_1, \dots, X_N\}$  from the distribution  $\mu_{\hat{\theta}_t}$ . Define

$$\hat{\theta}_{t+1} = \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{\theta}_t) \log W(X_i; \theta). \quad (4.9)$$

The iteration continues until a prescribed convergence criterion of  $\hat{\theta}_t$  is satisfied.

As we have mentioned previously, the minimization problem (4.9) can often be solved analytically (the connection between this minimization problem and the maximum likelihood estimate can be found in Asmussen and Glynn 2007). For illustration, we will consider the case where  $\{\mu_\theta\}$  is the family of exponential changes of measures of the original distribution  $\mu$ . That is,

$$\frac{d\mu_\theta}{d\mu}(x) = e^{\langle \theta, x \rangle - H(\theta)}.$$

Then  $\log W(X_i; \theta) = H(\theta) - \langle \theta, X_i \rangle$  and  $\nabla W(X_i; \theta) = \nabla H(\theta) - X_i$ . It follows easily that the minimizing  $\hat{\theta}_{t+1}$  satisfies

$$\nabla H(\hat{\theta}_{t+1}) = \frac{\sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{\theta}_t) X_i}{\sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{\theta}_t)}.$$

Note that  $\nabla H(\theta)$  equals the expected value of a random variable with distribution  $\mu_\theta$ . Therefore if we reparametrize the distributions  $\{\mu_\theta\}$  by the mean  $v$ , then we obtain the classical cross-entropy updating formula

$$\hat{v}_{t+1} = \frac{\sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{v}_t) X_i}{\sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{v}_t)}. \quad (4.10)$$

This formula actually holds when the distributions  $\{\mu_\theta\}$  belongs to a more general *natural exponential family* that is reparametrized by the mean; see Appendix A.3 of Rubinstein and Kroese (2008).

EXAMPLE 1. Consider the SRW model. Let  $X = (Z_1, \dots, Z_n)$  where  $Z_j$ 's are independent with common distribution  $\mu$ . Denote by  $\{\mu_\theta : \theta \in \Theta\}$  the family of exponential change of measure of  $\mu$ , that is,

$$\frac{d\mu_\theta}{d\mu}(z) = e^{\langle \theta, z \rangle - H(\theta)}$$

Suppose that the family of candidate sampling distributions of  $X$  is  $\{\nu_\theta : \theta \in \Theta\}$  such that under  $\nu_\theta$ ,  $Z_j$ 's are independent with common distribution  $\mu_\theta$ . Then the likelihood ratio  $W(x; \theta)$  is given by

$$W(x; \theta) = \prod_{j=1}^n e^{\langle \theta, z_j \rangle - H(\theta)} = e^{n\langle \theta, \bar{S}(x) \rangle - nH(\theta)},$$

where  $x = (z_1, \dots, z_n)$  and  $\bar{S}(x) = (z_1 + \dots + z_n)/n$ . It is not difficult to solve the updating formula (4.9) to obtain

$$\nabla H(\hat{\theta}_{t+1}) = \frac{\sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{\theta}_t) \bar{S}(X_i)}{\sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{\theta}_t)},$$

where  $X_i = (Z_1^{(i)}, \dots, Z_n^{(i)})$ ,  $i = 1, \dots, N$ , and  $Z_j^{(i)}$  are independent samples from the common distribution  $\mu_{\hat{\theta}_t}$ . As before, if we reparametrize the distribution  $\nu_\theta$  by the mean  $v$ , the formula becomes

$$\hat{v}_{t+1} = \frac{\sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{v}_t) \bar{S}(X_i)}{\sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{v}_t)}.$$

In other words, the mean of the updated sampling distribution is the weighted average of the sample path means.

**EXAMPLE 2.** Consider the TQN model. Suppose that the family of candidate sampling distributions is  $\{P_v : v = (v_1, v_2, v_3), v_i > 0\}$  such that under  $P_v$  the system is a Jackson network with exponential interarrival times of mean  $v_1$ , and exponential service times of mean  $v_2$  and  $v_3$ , respectively. The original distribution corresponds to  $v_0 = (1/\lambda, 1/\mu_1, 1/\mu_2)$ .

Let  $X_1, \dots, X_N$  be independent sample paths generated from the distribution  $P_{\hat{v}_t}$ , each of which starts from the origin and stops at the first time either the total population hits the level  $n$  or the system becomes empty again. For a sample path  $X$ , denote by  $\tau_1(X)$ ,  $\tau_2(X)$ , and  $\tau_3(X)$  the total number of interarrivals, service completion at node 1, and service completion at node 2, respectively. Let  $\{Y_{1j}(X) : j = 1, \dots, \tau_1(X)\}$  be the interarrival times. Similarly, let  $\{Y_{2j}(X) : j = 1, \dots, \tau_2(X)\}$  and  $\{Y_{3j}(X) : j = 1, \dots, \tau_3(X)\}$  be the service times at node 1 and node 2, respectively. Then the density of a sample path  $X$  under the distribution  $P_v$  equals

$$f(X; v) = \prod_{k=1}^3 \prod_{j=1}^{\tau_k(X)} \frac{1}{v_k} e^{-Y_{kj}(X)/v_k}$$

and the likelihood ratio  $W$  is given by

$$W(X; v) = \frac{f(X; v_0)}{f(X; v)}.$$

Denote by  $A$  the buffer overflow event. It is not difficult to solve the stochastic program (4.9) to obtain the analytic formula

$$\hat{v}_{t+1,k} = \frac{\sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{v}_t) \sum_{j=1}^{\tau_k(X_i)} Y_{kj}(X_i)}{\sum_{i=1}^N 1_{\{X_i \in A\}} W(X_i; \hat{v}_t) \tau_k(X_i)}$$

for  $k = 1, 2, 3$ . This updating formula is actually valid for much more complicated queueing networks. See de Boer, Kroese, and Rubinstein (2004) for more details.

### 4.2.2 The initialization in rare event simulation

The initialization of the cross-entropy algorithm, or the choice of  $\hat{\theta}_0$ , can be quite flexible in general. It usually suffices to set  $\hat{\theta}_0 = \theta_0$  where  $\theta_0$  corresponds to the original distribution. However, this recipe is problematic in the context of rare event simulation, since most likely the indicator  $1_{\{X_i \in A\}}$  will be zero for all  $i$  if  $A$  is a rare event, rendering the minimization problem (4.9) meaningless.

A possible approach is as follows. Choose a set  $B \supseteq A$  so that it is much less rare than  $A$  but shares the same qualitative flavor, e.g., in the TQN example one may choose  $B$  to be the event of total population overflow with buffer size  $m \ll n$ . Setting  $\hat{\theta}_0 = \theta_0$ , a pilot cross-entropy algorithm delivers the nearly optimal tilting parameter (say)  $\theta^*$  for estimating  $P(X \in B)$ . Next with  $\hat{\theta}_0 = \theta^*$ , the main cross-entropy algorithm is performed to yield the optimal tilting parameter for estimating the actual probability of interest  $P(X \in A)$ . De Boer, Kroese, and Rubinstein (2004) used this approach to estimate buffer overflow probabilities in queueing networks, where  $B$  is chosen as the buffer overflow event with a *small* buffer level.

Generalizing this idea, a *two-stage* iterative scheme where *both the set  $B$  and the tilting parameter  $\theta$  are updated* seems to be more convenient for most problems. To describe the idea, assume that the probability we wish to estimate is

$$p = P(S(X) \geq \gamma) = P(X \in A_\gamma)$$

where  $\gamma$  is a fixed level,  $S$  is some performance measure, and  $A_\gamma = \{x : S(x) \geq \gamma\}$ . It is assumed that  $\gamma$  is large and  $A_\gamma$  is a rare event. As before, the distribution of  $X$  is denoted by  $\mu$  and  $\{\mu_\theta : \theta \in \Theta\}$  is a parametrized family of candidate sampling distribution. In this two-stage approach for estimating  $p$ , one generates a sequence of tilting parameters  $\{\hat{\theta}_t\}$ , as well as a sequence of levels  $\{\hat{\gamma}_t\}$  that are determined by the samples and generally increase to the actually fixed large level  $\gamma$ . In essence, these *artificial* intermediate levels divide the original difficult rare event  $A$  into a sequence of easier, less rare events  $A_{\hat{\gamma}_t}$ .

The algorithm is as follows. Fix a priori a fraction  $\rho$  that is not too small, usually between 1% and 10%. Setting  $\hat{\theta}_0 = \theta_0$ , we generate  $N$  samples  $X_1, \dots, X_N$  from the distribution  $\mu_{\hat{\theta}_0}$ . Estimate the  $(1 - \rho)$ -quantile of  $S(X)$  by the sample quantile. That is, order the performances  $S(X_i)$  from the smallest to the largest:  $S_{(1)} \leq \dots \leq S_{(N)}$  and define

$$\hat{\gamma}_1 = S_{(N_e)}, \quad N_e = \lceil (1 - \rho)N \rceil,$$

where  $\lceil x \rceil$  is the ceiling of  $x$  or the smallest integer that is greater than or equal to  $x$ . Then we update the tilting parameter as in (4.9) with the set  $A$  replaced by  $A_{\hat{\gamma}_1}$ . In other words,  $\hat{\theta}_1$  is estimated on the basis of those samples  $X_i$  that satisfies  $S(X_i) \geq \hat{\gamma}_1$  and there are about  $\rho N$  of them (*elite samples*). Iterating these steps until  $\hat{\gamma}_t \geq \gamma$ , we have the following algorithm:

MAIN CROSS-ENTROPY ALGORITHM FOR RARE EVENT SIMULATION.

1. Let  $\hat{\theta}_0 = \theta_0$  and  $t = 0$  (iteration counter).
2. Generate samples  $X_1, \dots, X_N$  from the distributions  $\mu_{\hat{\theta}_t}$ . Calculate the performances  $S(X_i)$  and order them from the smallest to the largest:  $S_{(1)} \leq \dots \leq S_{(N)}$  and define

$$\hat{\gamma}_{t+1} = \min\{S_{(N_e)}, \gamma\}.$$

3. Use these samples  $X_1, \dots, X_N$  to solve the stochastic program (4.9) with the set  $A$  replaced by  $A_{\hat{\gamma}_{t+1}}$ .
4. If  $\hat{\gamma}_{t+1} < \gamma$ , set  $t = t+1$  and reiterate from Step 2. Otherwise, proceed with Step 5.
5. Let  $T$  be the final iteration counter. Estimate the rare event probability  $p$  by importance sampling, with the final tilting parameter  $\hat{\theta}_T$ .

Sometimes between Step 4 and Step 5, one can refine the final tilting parameter by running a few extra iterations of the standard cross-entropy updating program (4.9) with  $\hat{\theta}_T$  as the initial tilting parameter and the set  $A$  fixed as  $A_\gamma$ . The analysis of the convergence properties of this algorithm can be found in R.Y. Rubinstein and D.P. Kroese (2004) and Costa, Jones, and Kroese (2007).

### 4.3 Dynamic importance sampling

The notion of *dynamic*, or *state-dependent* importance sampling was introduced in Dupuis and Wang (2004). The development of this methodology was partly motivated by the counterexamples in Glasserman and Kou (1995) and Glasserman and Wang (1997) that had challenged the validity of the standard heuristic. It was shown in Dupuis and Wang (2004) that the second moment of an importance sampling estimator can be interpreted as the value of a small noise stochastic game. In this context it was obvious that the heuristic approach, which amounted to allowing only those state

independent changes of measure [or open loop controls in the language of stochastic games], could not possibly be asymptotically efficient in general. See also Bassamboo, Juneja, and Zeevi (2006). This connection also linked importance sampling to the Isaacs equation of a limiting differential game, which turned out to be *equivalent* to the Hamilton-Jacobi-Bellman (HJB) equation associated with the corresponding large deviation rate function. As a consequence, the solution to this HJB equation can be used to construct asymptotically efficient importance sampling schemes.

Dupuis and Wang (2007) explored this connection in further depth and showed that the design and analysis of dynamic importance sampling algorithms could be based on the *classical subsolutions* to the HJB equation. One can often construct subsolutions that are structurally much simpler than the actual solution, but which correspond to asymptotically efficient importance sampling schemes that reflect this simplicity. Subsolutions provide a unifying and flexible tool and can be used to study a broad range of process models. See, e.g., Dupuis, Sezer, and Wang (2007) and Dupuis and Wang (2007, 2009).

### 4.3.1 Limit differential game and its Isaacs equation

We will use the SRW model to formally illustrate the connection between importance sampling and small noise stochastic games. Recall that  $\{Z_1, \dots, Z_n\}$  is a sequence of independent random variables with common distribution  $\mu$ . Define the scaled random walk process

$$X_j = \frac{1}{n} \sum_{i=1}^j Z_i, \quad j = 1, \dots, n, \quad (4.11)$$

with  $X_0 = 0$ . The probability of interest is  $p_n = P(X_n \in A)$ . As before, one can define an exponential change of measure  $\nu_\alpha$  by

$$\nu_\alpha(dx) = e^{\langle \alpha, x \rangle - H(\alpha)} \mu(dx)$$

for every  $\alpha \in \mathbb{R}^d$ .

Consider a *state-dependent* change of measure in the following sense. For each  $j = 0, 1, \dots, n-1$ , conditional on the simulation history  $\{Z_i : i = 1, \dots, j\}$ ,  $Z_{j+1}$  is sampled from a distribution  $\mu_{\alpha_j}$ , where  $\alpha_j$  is a function of both the *scaled time*  $j/n$  and the *scaled state*  $X_j$  as defined in (4.11). The corresponding importance sampling estimator is given by

$$Y_n = 1_{\{X_n \in A\}} \prod_{j=0}^{n-1} e^{-\langle \alpha_j, Z_{j+1} \rangle + H(\alpha_j)}.$$

The estimate  $Y_n$  is unbiased. Our goal is to minimize the variance, or equivalently, the second moment of  $Y_n$ .

We will recast this minimization problem as a stochastic control problem with  $\{\alpha_j\}$  being the control, and make a natural connection to a partial differential equation. To this end we must extend the problem slightly to allow a general initial time and state. For  $i \geq 0$  and  $x \in \mathbb{R}^d$ , define  $X_j$  for  $j = i, \dots, n$  as above except that  $X_i = x$ , and then define

$$V_n(x, i) = \inf_{\{\alpha_j\}} \bar{E} \left[ 1_{\{X_n \in A\}} \prod_{j=i}^{n-1} e^{-\langle \alpha_j, Z_{j+1} \rangle + H(\alpha_j)} \right]^2,$$

where  $\bar{E}$  denotes the expectation taken under the change of measure determined by the control  $\{\alpha_j\}$ . In other words,  $V_n(x, i)$  is the minimal second moment of the importance sampling estimators given that the state process  $\{X_j\}$  starts at time  $i$  with initial state  $x$ . It will be more convenient to express this in terms of the original distributions as in Remark 4.1:

$$V_n(x, i) = \inf_{\{\alpha_j\}} E \left[ 1_{\{X_n \in A\}} \prod_{j=i}^{n-1} e^{-\langle \alpha_j, Z_{j+1} \rangle + H(\alpha_j)} \right],$$

where the expected value is taken such that  $\{Z_j, \dots, Z_n\}$  are independent with common distribution  $\mu$ .

As the value function of a discrete time stochastic control problem,  $V_n$  satisfies the dynamic programming equation

$$V_n(x, i) = \inf_{\alpha \in \mathbb{R}^d} \int e^{H(\alpha) - \langle \alpha, y \rangle} V_n \left( x + \frac{y}{n}, i + 1 \right) \mu(dy). \quad (4.12)$$

Owing to the exponential scaling in  $n$ , it is natural to consider the logarithmic transform of  $V_n$  and assume that

$$-\frac{1}{n} \log V_n(x, i) \approx W(x, i/n) \quad (4.13)$$

for some smooth function  $W : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}$ . This leads to the approximation

$$V_n \left( x + \frac{y}{n}, i + 1 \right) \cdot V_n^{-1}(x, i) \approx \exp \left\{ -\langle \nabla W(x, t), y \rangle - \frac{\partial W}{\partial t}(x, t) \right\}$$

where  $\nabla$  is the gradient with respect to  $x$ , and  $t = i/n$ . Plugging the above approximation into equation (4.12), taking log on both sides, and recalling

the definition of  $H$ , we arrive at

$$0 = -\frac{\partial W}{\partial t} + \inf_{\alpha \in \mathbb{R}^d} [H(\alpha) + H(-\nabla W - \alpha)]. \quad (4.14)$$

Since  $V_n(x, n) = 1_{\{x \in A\}}$ ,  $W$  satisfies the boundary condition  $W(x, 1) = 0$  if  $x \in A$  and  $\infty$  otherwise.

We wish to make a few observations regarding equation (4.14). Even though a very special model has been used here, these claims actually hold in much greater generality.

1. Equation (4.14) is the *Isaacs equation* associated with a two-person zero-sum game. Indeed, since  $H$  and  $L$  are convex duals, for every  $\alpha \in \mathbb{R}^d$

$$H(\alpha) = \sup_{\beta \in \mathbb{R}^d} [\langle \alpha, \beta \rangle - L(\beta)].$$

Thus equation (4.14) can be written as

$$0 = \frac{\partial W}{\partial t} + \sup_{\alpha} \inf_{\beta} [\langle \nabla W, \beta \rangle + L(\beta) + \langle \alpha, \beta \rangle - H(\alpha)].$$

This is the Isaacs equation corresponds to the following zero-sum differential game. The dynamics  $\dot{\phi}(t) = \beta(t)$  only involves the  $\beta$ -player. The running cost  $L(\beta) + \langle \alpha, \beta \rangle - H(\alpha)$  is affected by both players, and the terminal cost is  $\infty \cdot 1_{A^c}$ . Because of the intervening minus sign, the maximizing  $\alpha$ -player indeed tries to minimize the variance.

2. Thanks to the convexity of  $H$ , the maximizing  $\alpha$  in equation (4.14) is

$$\alpha^*(x, t) = -\frac{1}{2} \nabla W(x, t). \quad (4.15)$$

This is the basic formula for computing the state-dependent change of measure.

3. Plugging the formula of  $\alpha^*$  into equation (4.14), we arrive at

$$0 = \frac{\partial W}{\partial t} + 2H(-\nabla W/2). \quad (4.16)$$

This is equivalent to the HJB equation associated with the corresponding large deviation rate function. To see this, we abuse the notation and extend the definition of  $p_n$  to

$$p_n(x, t) = P(X_n \in A | X_{[nt]} = x)$$

for  $x \in \mathbb{R}$  and  $t \in [0, 1]$ . Clearly the probability of interest  $P(X_n \in A)$  equals  $p_n(0, 0)$ . Then under suitable conditions

$$-\lim_n \frac{1}{n} \log p_n(x, t) = \inf_{\phi} \int_t^1 L(\dot{\phi}(s)) ds,$$

where the infimum is taken over all absolutely continuous functions  $\phi$  such that  $\phi(t) = x$  and  $\phi(1) \in A$ . Denote by  $U(x, t)$  the value function of this minimization problem. Then  $U$  satisfies the HJB equation

$$0 = \inf_{\beta} \left[ \frac{\partial U}{\partial t} + \langle \nabla U, \beta \rangle + L(\beta) \right] = \frac{\partial U}{\partial t} + H(-\nabla U)$$

with terminal condition  $U(x, 1) = 0$  if  $x \in A$  and  $\infty$  otherwise. Clearly it is equivalent to (4.16) by a change of variable  $W = 2U$ . This equivalence also indicates that the state-dependent change of measure based on the solution to the Isaacs equation (4.16) is asymptotically efficient since by equation (4.13)

$$-\lim_n \frac{1}{n} \log V_n(0, 0) = W(0, 0) = 2U(0, 0) = -2 \lim_n \frac{1}{n} \log p_n(0, 0),$$

and  $V_n(0, 0)$  is the second moment of the corresponding importance sampling estimator. A rigorous proof can be found in Dupuis and Wang (2004).

### 4.3.2 The idea of subsolutions

From the previous discussion, it follows that the solution to a related Isaacs equation can be used to build asymptotically efficient importance sampling schemes. A difficulty with this approach is that a solution to a nonlinear partial differential equation such as Isaacs equation is hard to compute. To circumvent this, Dupuis and Wang (2007) proposed importance sampling schemes based on the subsolutions to the Isaacs equation. Subsolutions are functions that satisfy the partial differential equation with inequality instead of equality, and allow much greater flexibility in the design of importance sampling schemes.

In order to understand the sufficiency of subsolution, let us examine the criterion of logarithmic asymptotic efficiency more closely. Recall the definition of logarithmic asymptotic efficiency (2.1). Jensen's inequality implies that

$$\log E[Y_n^2] \geq 2 \log E[Y_n] = 2 \log p_n.$$

Therefore, if for some  $\gamma > 0$  the large deviation asymptotics

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log p_n = -\gamma$$

hold, then (2.1) is equivalent to the inequality

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log E[Y_n^2] \leq -2\gamma. \quad (4.17)$$

In other words, in order to show that  $Y_n$  is asymptotically efficient, it suffices to establish the upperbound (4.17) only. The inequalities in the definition of a subsolution [see below] are consistent with this upper-bound, when the subsolution is combined with a verification argument to bound the second moment of  $Y_n$ .

To give the definition of a subsolution, we consider a family of Isaacs equations of a given form. The definition easily extends to other types of Isaacs equations. For a broad collection of problems, the probability of interest is of form  $p_n = P(S_n/n \in A)$ , where  $S_n$  is the partial sum of independent identically distributed random variables or functionals of Markov chains. Then under suitable conditions, the large deviations asymptotics

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log p_n = - \inf_{\beta \in A} L(\beta) := -\gamma$$

hold for some convex rate function  $L$ . Denoting by  $H$  the Legendre transform of  $L$ , then the Isaacs equation takes the familiar form

$$\frac{\partial W}{\partial t} + \sup_{\alpha} \inf_{\beta} [\langle \nabla W, \beta \rangle + L(\beta) + \langle \alpha, \beta \rangle - H(\alpha)] = 0, \quad (4.18)$$

with boundary condition  $W(x, 1) = 0$  if  $x \in A$  and  $\infty$  otherwise.

DEFINITION. A classical subsolution (4.18) to the Isaacs equation is a smooth function  $W$  that satisfies

$$\frac{\partial W}{\partial t} + \sup_{\alpha} \inf_{\beta} [\langle \nabla W, \beta \rangle + L(\beta) + \langle \alpha, \beta \rangle - H(\alpha)] \geq 0$$

with boundary inequality  $W(x, 1) \leq 0$  for  $x \in A$ .

Given a classical subsolution  $W$ , the corresponding change of measure is determined by the maximizing  $\alpha^*$  for the min/max term, which has exactly the same form as (4.15). The following theorem is the key result in

the performance analysis of those importance sampling schemes based on subsolutions. See Dupuis and Wang (2007) for more details.

**THEOREM 1.** Let  $W$  be a classical subsolution to the Isaacs equation and  $Y_n$  the corresponding importance sampling estimate of  $p_n$ . Then under suitable conditions

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log E[Y_n^2] \leq -W(0, 0). \quad (4.19)$$

In particular, if  $W(0, 0) = 2\gamma$ , then  $Y_n$  is asymptotically efficient.

### 4.3.3 Construction of subsolutions

Theorem 1 reduces the problem of building an asymptotically efficient or nearly asymptotically efficient importance sampling scheme to that of a classical subsolution  $W$  with  $W(0, 0)$  equal or close to  $2\gamma$ , respectively. For systems with piecewise homogeneous dynamics, a particularly useful technique is to build a piecewise affine subsolution at first and then obtain a classical subsolution by mollification. The construction of a piecewise affine subsolution, which is usually identified as the minimum of a collection of affine functions, is the key step. Once such a piecewise affine subsolution is given, say,

$$\bar{W} = W_1 \wedge \cdots \wedge W_m,$$

we use a mollification technique called *exponential weighting*:

$$W_\varepsilon = -\varepsilon \log \sum_{i=1}^m e^{-W_i/\varepsilon},$$

where  $\varepsilon$  is a small positive number. It is not difficult to show that  $W_\varepsilon$  approximates  $\bar{W}$  as  $\varepsilon$  approaches zero. Furthermore, analytic formulas for quantities such as  $\nabla W_\varepsilon$  are readily available. There are two remarks we wish to make: (1) In general,  $W_\varepsilon$  is not exactly a subsolution, but an approximate one in the sense that the inequality (4.19) is satisfied with the right hand side replaced by a vanishing negative number. This is usually sufficient for asymptotic efficiency; (2) It is sometimes more convenient to use a change of measure slightly different from the one determined by  $\alpha^* = -\nabla W_\varepsilon/2$ . It is essentially a state dependent mixture of the changes of measure determined by  $\{W_i\}$ . Theorem 1 still holds in this case. See Dupuis and Wang (2007) for details.

To finish the discussion, we will illustrate the construction of piecewise affine subsolutions through two examples. In general, it is accomplished by

carefully analyzing the properties of the system dynamics and the relevant large deviation properties.

EXAMPLE 3. Consider the SRW model. Without loss of generality we assume that  $E[X_1] = 0$ . We first assume that  $A = [\beta, \infty)$  for some  $\beta > 0$ . Denote by  $\alpha$  the conjugate point of  $\beta$ . Then the affine function

$$W(x) = -2\langle \alpha, x - \beta \rangle - 2(1-t)H(\alpha)$$

is a subsolution to the Isaacs equation. Since  $-\nabla W/2 = \alpha$ , the corresponding change of measure is exactly the classical one as in Section 4.1.

A more interesting case is when  $A = (-\infty, \bar{\beta}] \cup [\beta, \infty)$  when  $\bar{\beta} < 0 < \beta$ . Let  $\bar{\alpha}$  be the conjugate point of  $\bar{\beta}$ . Define

$$\bar{W}(x) = -2\langle \bar{\alpha}, x - \bar{\beta} \rangle - 2(1-t)H(\bar{\alpha}).$$

Then it is not difficult to check that  $W^* = W \wedge \bar{W}$  is a two-piece affine subsolution. Note that

$$-\nabla W^*/2 = \begin{cases} \alpha & \text{if } W < \bar{W} \\ \bar{\alpha} & \text{if } W > \bar{W} \end{cases}$$

is piecewise constant. See Figure 2 for an illustration of how this would partition the space-time domain.

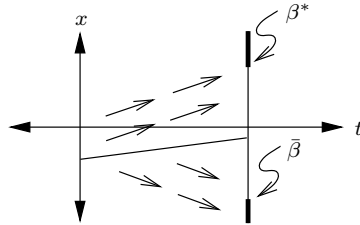


Figure 2: Domain decomposition and corresponding drifts

EXAMPLE 4. Consider the TQN model. Without loss of generality assume  $\lambda + \mu_1 + \mu_2 = 1$ . Define by  $\{Z_k = (Z_{k,1}, Z_{k,2}) : k = 0, 1, \dots\}$  the embedded discrete time Markov chain, where  $Z_{k,i}$  represents the length of the  $i$ -th queue at the  $k$ -th transition epoch of the network,  $i = 1, 2$ . The space of the possible jumps is

$$\mathbb{V} = \{v_0 = e_1, v_1 = -e_1 + e_2, v_2 = -e_2\}.$$

The system dynamics can be described as  $Z_{k+1} = Z_k + \pi[Z_k, Y_{k+1}]$ , where  $\{Y_k\}$  are random variables taking values in  $\mathbb{V}$  and  $\pi$  is the mapping due to the non-negativity constraint on the queue lengths: for  $x = (x_1, x_2) \in \mathbb{R}_+^2$  and  $y \in \mathbb{V}$

$$\pi[x, y] = \begin{cases} 0 & \text{if } x_i = 0 \text{ and } y = v_i \text{ for some } i = 1, 2 \\ y & \text{otherwise.} \end{cases}$$

See Figure 3. Let  $\mathcal{P}$  be the space of strictly positive probability measures on  $\mathbb{V}$ , i.e.,

$$\mathcal{P} = \{\theta = (\theta_0, \theta_1, \theta_2) : \theta_0 + \theta_1 + \theta_2 = 1, \theta_i > 0\}.$$

Under the original distribution,  $\{Y_k\}$  are independent identically distributed with distribution  $\Theta = (\lambda, \mu_1, \mu_2)$ . Recall that  $R(\cdot\|\cdot)$  denotes the relative entropy. The relevant Isaacs equation is such that for  $x \in \{(x_1, x_2) \in \mathbb{R}_+^2 : x_1 + x_2 < 1\}$

$$\sup_{\bar{\Theta} \in \mathcal{P}} \inf_{\theta \in \mathcal{P}} \left[ \langle \nabla W(x), \sum_{i=0}^2 \theta_i \cdot \pi[x, v_i] \rangle + \sum_{i=0}^2 \theta_i \log \frac{\bar{\Theta}_i}{\Theta_i} + R(\theta\|\Theta) \right] = 0,$$

with the boundary condition  $W(x) = 0$  when  $x_1 + x_2 = 1$ . Here  $\bar{\Theta}$  corresponds to the change of measure. Given  $W$ , the optimal (maximizing)  $\bar{\Theta}$  admits an analytic formula.

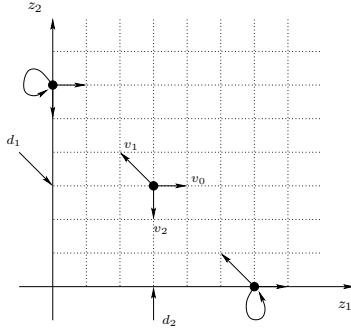


Figure 3: The system dynamics

The definition of a subsolution is just to replace the “=” by “ $\geq$ ” in the Isaacs equation and replace the boundary condition “ $W(x) = 0$ ” by “ $W(x) \leq 0$ ”. Simple piecewise affine subsolutions can be constructed. For example, when  $\mu_2 \leq \mu_1$ , define vectors

$$r_1 = 2\gamma(-1, -1), \quad r_2 = 2\gamma(-1, 0), \quad r_3 = (0, 0).$$

Let  $\delta$  be a small positive number. Then  $\bar{W} = W_1 \wedge W_2 \wedge W_3$  defines a subsolution, where

$$W_k(x) = \langle r_k, x \rangle + 2\gamma - k\delta, \quad k = 1, 2, 3.$$

This subsolution divides the region into three pieces:  $R_1$ ,  $R_2$ , and  $R_3$ , such that  $\bar{W}(x) = W_k(x)$  for  $x \in R_k$ . See Figure 4. The regions  $R_2$  and  $R_3$  are sometimes called “boundary layers”. They are closely related to the discontinuity of the dynamics on the boundary  $\{x_2 = 0\}$  and the origin, and the large deviations properties of the rare event. Details of the algorithms can be found in Dupuis, Sezer, and Wang (2007).

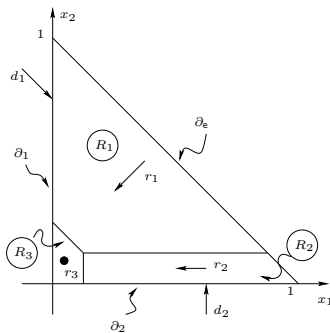


Figure 4: Piecewise affine subsolution

#### 4.4 Lyapunov function method for heavy tailed distribution

Much of the previous discussion has assumed that the distributions involved are light-tailed in the sense that their moment generating functions are finite in a small neighborhood of the origin, and thus the exponential changes of measure are meaningful. On the contrary, for a large class of distributions emerging from practice, the tail probabilities decay much more slowly. These *heavy-tailed distributions* have very different large deviation properties and the exponential scaling is in general not valid. As a consequence, fast rare event simulation algorithms can look very different from those for light-tailed distributions. See Asmussen, Binswanger, and Hojgaard (2000) and references therein.

Some of the recent works on rare event simulation involving heavy-tailed distributions have been concerned with state-dependent algorithms. See, e.g., Dupuis, Leder, and Wang (2007). In this section we review a general technique proposed by Blanchet and Glynn (2008) that is based on Lyapunov

functions. These Lyapunov functions are closely related to the subsolutions in Section 4.3 [they are in some sense the exponential of subsolutions]. Even though the method is applicable to light-tailed distributions as well, we will discuss it in the context of heavy-tailed distributions, via the example of estimating level crossing probabilities for heavy-tailed random walks.

Let  $\{X_i\}$  be a sequence of independent identically distributed heavy-tailed random variables with common distribution  $\mu$  and strictly negative mean. Define the simple random walk

$$S_n = y + \sum_{i=1}^n X_i,$$

with initial condition  $S_0 = y$ . Here we assume that  $y$  is a very negative number and the quantity of interest is the level crossing probability

$$p^*(y) = P(S_n \geq 0 \text{ for some } n).$$

Under suitable conditions, as  $y \rightarrow -\infty$ ,  $p^*(y)$  has the asymptotics

$$p^*(y) \sim \frac{1}{-E[X_1]} \int_{|y|}^{\infty} P(X_1 > s) ds. \quad (4.20)$$

A useful observation is that if  $Q^*$  is a probability measure (Doob's h-transform) that satisfies

$$Q^*(X_{n+1} \in dz | S_n = x) = \frac{p^*(z+x)}{p^*(x)} \mu(dz) \quad (4.21)$$

for  $x < 0$ , then  $Q^*$  is the zero variance importance sampling change of measure. As before,  $Q^*$  is impractical since  $p^*(\cdot)$  is unknown.

However, (4.21) does motivate the use of a change of measure  $Q$  such that for  $x < 0$

$$Q(X_{n+1} \in dz | S_n = x) = \frac{v(z+x)}{w(x)} \mu(dz), \quad (4.22)$$

where  $v$  is a function that is *close* to  $p^*$  and  $w(x)$  is the normalization constant such that

$$w(x) = \int_{\mathbb{R}} v(z+x) \mu(dz). \quad (4.23)$$

The corresponding importance sampling estimator is just

$$Y = 1_{\{T < \infty\}} \prod_{i=0}^{T-1} \frac{w(S_i)}{v(S_{i+1})},$$

where  $T = \inf\{n \geq 1 : S_n \geq 0\}$ .

To aid the design of  $Q$ , one also need some means to analyze its performance. This is where the Lyapunov function comes into play. Even though the definition here is slightly different from that of Blanchet and Glynn (2008) in the form, they are indeed equivalent.

DEFINITION. A function  $H : \mathbb{R} \rightarrow [0, \infty)$  is said to be a *Lyapunov function* associate with the probability measure  $Q$  if for every  $x < 0$ ,

$$H(x) \geq \int_{\mathbb{R}} \frac{w(x)}{v(z+x)} H(z+x) \mu(dz).$$

and  $H(x) \geq 1$  for  $x \geq 0$ .

THEOREM. Let  $Y$  be the importance sampling estimator and  $H$  a Lyapunov function associated with  $Q$ , then

$$E_Q[Y^2 | S_0 = y] \leq H(y).$$

*Proof.* Define the process

$$R_n \doteq H(S_n) \cdot \prod_{i=0}^{n-1} \frac{w^2(S_i)}{v^2(S_{i+1})}.$$

Then it is straightforward to check that the definition of the Lyapunov function  $H$  is equivalent to the claim that  $R_{T \wedge n}$  is a supermartingale under  $Q$ . Therefore by the Optional Sampling Theorem, for  $y < 0$

$$E_Q[R_T | S_0 = y] \leq E_Q[R_0 | S_0 = y] = H(y).$$

We conclude the proof by observing that  $R_T \geq Y^2$  since  $H(S_T) \geq 1$ . Q.E.D.

The idea of the Lyapunov function method is to find a pair  $(v, H)$  such that (i)  $v$  is close to  $p^*$  in the sense that they are asymptotically equivalent; (ii)  $H$  is a Lyapunov function of the form  $H(x) = h(x)v^2(x)$ . Then the preceding theorem asserts that the performance of the importance sampling algorithm associated with the change of measure  $Q$  is characterized by  $h$ . For example, if  $h$  is bounded then it is of bounded relative error.

Now let us discuss the objectives (i) and (ii). An immediate problem is that how one can tell if  $v$  is close to  $p^*$  when  $p^*$  is unknown in the first place. The idea is that, comparing (4.21) and (4.22), if  $v = p^*$  then  $w$  defined in (4.23) should equal  $p^*$  as well and thus  $w - v = 0$ . Therefore,  $w - v$  can be

used as a criterion to measure how close  $v$  and  $p^*$  are. With this in mind, it is natural to start with the function on the right-hand-side of (4.20). Define a non-negative random variable  $Z$  that is independent of  $\{X_i\}$  and such that for  $t > 0$ ,

$$P(Z > t) = 1 \wedge \int_t^\infty \frac{1}{-E[X_1]} P(X_1 > s) ds.$$

Define  $\bar{v}(x) = P(Z > -x)$  for all  $x \in \mathbb{R}$ . Note that  $\bar{v}(x) = 1 = p^*(x)$  if  $x > 0$ . Furthermore, with  $\bar{w}$  defined as in (4.23) with  $v$  replaced by  $\bar{v}$ , it can be shown that  $\bar{w}(x)$  and  $\bar{v}(x)$  are asymptotically very close as  $x \rightarrow -\infty$ . Therefore, there exists an  $a^* < 0$  such that  $v(x) = \bar{v}(x + a^*)$  and  $w(x) = \bar{w}(x + a^*)$  are very close for all  $x < 0$ . Given the function  $v$ , one can find a bounded piecewise constant function  $h$  such that  $H(x) = h(x)v^2(x)$  defines a Lyapunov function. Thus we obtain an importance sampling scheme with bounded relative error.

We want to mention in the end that the sampling distribution is determined by (4.22), and it is not difficult to check that

$$Q(X_{n+1} \in dz | S_n = x) = P(X_1 \in dz | X_1 + Z > -x - a^*).$$

Samples from this conditional distribution are typically generated by suitable acceptance/rejection schemes, where the acceptance probability remains uniformly bounded away from 0. The design of such schemes is based on the tail behavior of the distribution  $\mu$ . See Blanchet and Glynn (2007) for the case when  $\mu$  is regularly varying.

## References

- [1] Asmussen, S. (1985). Conjugate processes and the simulation of ruin probability. *Stochastic Process. Appl.*, 20:213–229.
- [2] Asmussen, S., Binswanger, K., and Hojgaard, B. (2000). Rare event simulation for heavy-tailed distributions. *Bernoulli*, 6:303–322.
- [3] Asmussen, S. and Glynn, P. (2007). *Stochastic Simulation: Algorithms and Analysis*. Springer, New York.
- [4] Asmussen, S. and Rubinstein, R.Y. (1995). Steady state rare event simulation in queueing model and its complexity properties. *Advances in Queueing: Theory, Methods and Open Problems*, J. Dshalalow (editor), Volume 1, CRC Press, 429-462.

- [5] Bassamboo, A., Juneja, S. and Zeevi, A. (2006). On the efficiency loss of state-independent importance sampling in the presence of heavy-tails. *Oper. Res. Lett.*, 34:521-531.
- [6] Blanchet, J.H. and Glynn, P. (2008). Efficient rare-event simulation for the maximum of heavy-tailed random walks. *Ann. Appl. Probab.*, 18:1351–1378.
- [7] Costa, A., Jones, O.D., and Kroese, D.P. (2007). Convergence properties of the cross-entropy method for discrete optimization. *Oper. Res. Letters*, 35:573–580.
- [8] De Boer, P.T., Kroese D.P., and Rubinstein, R.Y. (2004). A fast cross-entropy method for estimating buffer overflows in queueing networks. *Manage. Sci.*, 50:883–895.
- [9] De Boer, P.T., Kroese D.P., Mannor, S., and Rubinstein, R.Y. (2005). A tutorial on the cross-entropy method. *Ann. Oper. Res.*, 134:19–967.
- [10] Dean, T. and Dupuis, P. (2009). Splitting for rare event simulation: A large deviation approach to design and analysis. *Stoch. Proc. Appl.*, 119:562–587.
- [11] Dupuis, P., Leder, K., and Wang, H. (2007). Importance sampling for sums of random variables with regularly varying tails. *TOMACS*, Vol. 17, Article 14.
- [12] Dupuis, P., Sezer, A., and Wang, H. (2007). Dynamic importance sampling for queueing networks. *Ann. Appl. Probab.*, 17:1306–1346.
- [13] Dupuis, P. and Wang, H. (2004). Importance sampling, large deviations, and differential games. *Stoch. and Stoch. Reports.*, 76:481–508.
- [14] Dupuis, P. and Wang, H. (2007). Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.*, 32:1–35.
- [15] Glasserman, P., Heidelberger, P., Shahabuddin, P., and Zajic, T. (1999). Multilevel splitting for estimating rare event probabilities. *Oper. Res.*, 47:585–600.
- [16] Glasserman, P. and Kou, S. (1995). Analysis of an importance sampling estimator for tandem queues. *ACM Trans. Modeling Comp. Simulation*, 4:22–42.

- [17] Glasserman, P. and Wang, Y. (1997). Counter examples in importance sampling for large deviations probabilities. *Ann. Appl. Prob.*, 7:731–746.
- [18] Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM Trans. Modeling Comp. Simulation*, 4:43–85.
- [19] Parekh, S. and Walrand, J. (1989). Quick simulation of rare events in networks. *IEEE. Trans. Autom. Control TAC*, 34:54–66.
- [20] Rubinstein, R.Y. (1997). Optimization of computer simulation models with rare events. *Euro. J. Oper. Res*, 99:89–112.
- [21] Rubinstein, R.Y. (1999). The simulated entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 2:127–190.
- [22] Rubinstein, R.Y. (2010). Randomized algorithm with splitting: why the classic randomized algorithms do not work and how to make them work. *Methodology and Computing in Applied Probability*, 12:1–41.
- [23] Rubinstein, R.Y. and Kroese, D.P. (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte Carlo Simulation and Machine Learning*. Springer-Verlag, New York.
- [24] Rubinstein, R.Y. and Kroese, D.P. (2007). *Simulation and the Monte Carlo Method*. John Wiley & Sons, Hoboken, New Jersey.
- [25] Siegmund, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.*, 4:673–684.