Chapter 14
Rare Event Monte Carlo and Importance Sampling

Suppose that in the analysis of some system, the value of a probability or expected value that is largely determined by one or a few events is important. Examples include the data loss in a communication network; depletion of capital reserves in a model for insurance; motion between metastable states in a chemical reaction network; and exceedance of a regulatory threshold in a model for pollution in a waterway. In previous chapters we have described how large deviation theory gives approximations for such quantities. The approximations take the form of logarithmic asymptotics, i.e., exponential decay rates.[1] For some purposes, especially when one is seeking *qualitative* information on how a rare event occurs, these approximations may be sufficient. For other purposes they may be inadequate, and a more accurate estimate is needed.

In this situation it is natural to turn to Monte Carlo approximation. However, as we will explain in some detail, the Monte Carlo approximation of small probabilities and related expected values also has difficulties owing to the role of rare events, and the design of reliable schemes requires great care. It turns out that many of the tools and constructions used for the large deviation analysis of a given problem can be used for the problem of designing Monte Carlo schemes that are efficient and reliable.

14.1 Example of a Quantity to be Estimated

To set the context, we consider a particular problem that arises frequently in various systems, especially communication theory. Let X^n be a Markov process with small noise of the form analyzed in Chap. 4. Thus we are given iid random vector fields $\{v_i(x), i \in \mathbb{N}_0, x \in \mathbb{R}^d\}$ on some probability space, and for each $x \in \mathbb{R}^d$ $v_i(x)$ has

[1] For certain special structures one can obtain more accurate approximations, e.g., approximations which identify both the exponential rate of decay as well as "pre-exponential" terms.

A. Budhiraja and P. Dupuis, *Analysis and Approximation of Rare Events*, Probability Theory and Stochastic Modelling 94,
https://doi.org/10.1007/978-1-4939-9579-0_14

distribution $\theta(\cdot|x)$, where $\theta(dy|x)$ is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d$. Then the discrete time Markov process $\{X_i^n\}_{i\in\mathbb{N}_0}$ is constructed through the recursion

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \quad X_0^n = x, \tag{14.1}$$

and the continuous time interpolation is defined by

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in [i/n, i/n + 1/n], i \in \mathbb{N}_0. \tag{14.2}$$

Let P_x denote probability conditioned on $X_0^n = x$ and integration with respect to P_x by E_x. The problem of interest is then to evaluate

$$p^n(x) \doteq P_x\left\{X^n \text{ enters } B \text{ before entering } A\right\}, \tag{14.3}$$

where A is open, B is closed, and $A \cap B = \emptyset$. For reasons that will become clear later, we explicitly record the initial condition in the notation, even though for many problems there may be only one initial condition of interest.

Under Conditions 4.3 and 4.7 or Conditions 4.3 and 4.8, Theorem 4.9 shows that for every $T \in (0, \infty)$, $\{X^n\}_{n\in\mathbb{N}}$ regarded as a collection of stochastic processes over the time horizon $[0, T]$ satisfies an LDP. Let I_T denote the corresponding rate function, and recall that $I_T(\phi) = 0$ characterizes the LLN limit trajectories of $\{X^n\}_{n\in\mathbb{N}}$. We will assume that A is an attractor of the LLN limit with nonempty interior and that B is in some sense rare. See Fig. 14.1. The trajectories in the figure are assumed to satisfy $I_T(\phi) = 0$ for all $T \in (0, \infty)$, and for all initial conditions $\phi(t)$ enters A as $t \to \infty$.

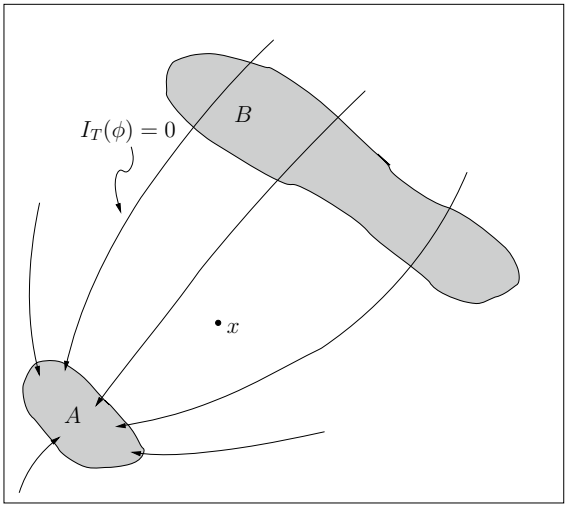

Fig. 14.1 Stability of the zero cost trajectories

Recall the notation

$$H(y,\alpha) = \log E \exp\{\langle \alpha, v_i(y)\rangle\}, \quad L(y,\beta) = \sup_{\alpha\in\mathbb{R}^d} [\langle \alpha,\beta\rangle - H(y,\alpha)]$$

from Chap. 4. Under appropriate regularity conditions the large deviation principle for $\{X^n\}$ implies

$$-\frac{1}{n}\log P_x\left\{X^n \text{ enters } B \text{ before entering } A\right\} \to V(x), \tag{14.4}$$

where

$$V(x) \doteq \inf\left[\int_0^T L(\phi(t),\dot{\phi}(t))dt : \phi \in C_{x,T}, T < \infty\right],$$

and with

$$C_{x,T} \doteq \{\phi(0) = x, \phi(t) \in B \text{ for some } t \in [0,T] \text{ and } \phi(s) \notin A \text{ for } s \in [0,t]\}.$$

The proof of (14.4) is typically carried out by reducing the analysis to that over a finite time interval and then invoking the large deviation principle for $\{X^n\}$ over finite time intervals (see Condition 15.18 and Proposition 15.19).

14.1.1 Relative Error

Recall that the problem of interest is to estimate the probability

$$p^n(x) \doteq P_x\left\{X^n \text{ enters } B \text{ before entering } A\right\}.$$

Let $C_x \doteq \cup_{T\in(0,\infty)} C_{x,T}$ be the trajectories that enter B before entering A after starting at x. To apply straightforward Monte Carlo, one would simulate K independent copies $\left\{X^{n,k}\right\}_{k=1,\ldots,K}$ of X^n, and then form the estimate

$$\hat{p}^n_K(x) \doteq \frac{1}{K}\sum_{k=1}^{K} 1_{\{X^{n,k}\in C_x\}}.$$

Note that k here is the index of the sample and *not* the time step, and that depending on the problem, the computational expense of simulating a single trajectory can vary greatly.

The variance of a single sample is

$$\begin{aligned}\mathrm{Var}\left(1_{\{X^{n,k}\in C_x\}}\right) &= E_x\left[1_{\{X^{n,k}\in C_x\}} - E_x 1_{\{X^{n,k}\in C_x\}}\right]^2 \\ &= E_x 1_{\{X^{n,k}\in C_x\}} - \left(E_x 1_{\{X^{n,k}\in C_x\}}\right)^2 \\ &= p^n(x) - [p^n(x)]^2,\end{aligned}$$

and if $p^n(x)$ is small $[p^n(x)]^2$ can be neglected. The *relative error*, which is defined by the ratio of the standard deviation of $\hat{p}^n_K(x)$ and $p^n(x)$, is then

$$\frac{\sqrt{\mathrm{Var}(\hat{p}^n_K(x))}}{p^n(x)} \approx \sqrt{\frac{p^n(x)}{K}} \cdot \frac{1}{p^n(x)} = \sqrt{\frac{1}{Kp^n(x)}}.$$

When considering rare events it is essential to use relative error as the figure of merit, since the variance can be small (or conversely big in some situations involving expected values) in absolute terms, and yet provide an estimate that is orders of magnitude off, and therefore quite inaccurate in a relative sense.

For the example problem, to obtain a relative error of roughly size 1 requires $K \approx (p^n(x))^{-1}$ samples. This is computationally infeasible when $p^n(x)$ is very small (e.g., 10^{-5}), or even when $p^n(x)$ is not so small if the computational effort needed to generate samples of X^n is great. For example, consider the problem of estimating the probability of an unusually large concentration of pollutant in a model for ground water contamination. The generation of each sample would typically involve solving a time dependent stochastic partial differential equation, and hence each sample is computationally expensive.

An alternative to standard Monte Carlo is to construct iid random variables $\gamma^n_1, \ldots, \gamma^n_K$ with $E_x\gamma^n_1 = p_n(x)$, and use the estimator

$$\hat{q}^n_K(x) \doteq \frac{\gamma^n_1 + \cdots + \gamma^n_K}{K}.$$

The performance as with ordinary Monte Carlo is determined by variance of γ^n_1, and since the estimator is unbiased [i.e., $E_x\gamma^n_1 = p_n(x)$], minimizing the variance is equivalent to minimizing $E_x\left(\gamma^n_1\right)^2$.

It is straightforward to obtain bounds on the best possible performance. For example, by Jensen's inequality and (14.4)

$$-\frac{1}{n}\log E_x\left(\gamma^n_1\right)^2 \le -\frac{2}{n}\log E_x\gamma^n_1 = -\frac{2}{n}\log p_n(x) \to 2V(x). \tag{14.5}$$

Hence the decay rate for the second moment cannot possible exceed $2V(x)$. An estimator is called **asymptotically efficient** if

$$\liminf_{n\to\infty} -\frac{1}{n}\log E_x\left(\gamma_1^n\right)^2 = 2V(x),$$

i.e., the optimal decay rate is achieved.

One could consider more stringent measures of performance, such as **bounded relative error**: there is $K < \infty$ such that

$$\limsup_{n\to\infty} \frac{\sqrt{\mathrm{Var}(\gamma_1^n)}}{p^n(x)} \le K,$$

or vanishing relative error [183]. While bounded relative error is certainly a desirable feature, it can be achieved only for the most elementary process models and events, such as the probability that a homogeneous random walk escapes from a set with simple structure. State dependence in the dynamics or even more complex situations (e.g., Markov modulated noise, multiscale dynamics) usually make it very difficult to construct schemes with (provable) bounded relative error. However, while asymptotic efficiency may be a more practical figure of merit, the logarithmic scaling can wipe out important terms in the variance that depend on other system parameters, such as another exponential scaling in terms of a time variable. Thus while more flexible and realistic than bounded relative error, it must be used with caution, and in all cases no single performance measure can replace a careful analysis of the variance and its dependence on all important system parameters. If different methods vary significantly with regard to the computational cost of implementation, then that aspect should also be factored into the performance measure.

We will discuss two well known methods used to design random variables $\{\gamma_k^n\}$ that are unbiased, which can be simulated with reasonable effort, and for which one may hope to get good performance: **importance sampling** and **splitting schemes**. For the remainder of this chapter and in Chap. 15 we focus on importance sampling (IS), and then in Chap. 16 turn to splitting. While many of the constructions needed for the successful design and analysis are essentially the same for both approaches, there are also interesting differences, some of which will be discussed at the end of Chap. 17.

We stress that for any approach to problems of rare event estimation a rigorous and *independent* analysis of performance is very important, since typical methods one would use to assess accuracy of the estimates (e.g., the empirical variance) are prone to the same difficulties and errors which can affect the estimates themselves. This point will be illustrated via a numerical example in the next section.

14.2 Importance Sampling

The basic formulae of importance sampling are as follows. Suppose that X has distribution θ, where X takes values in a Polish space S. Suppose that $G : S \to \mathbb{R}$ is Borel measurable and integrable with respect to θ, and the goal is to estimate

$m = EG(X)$. Consider an alternative sampling distribution π. It is required that θ be absolutely continuous with respect to π, so that the Radon-Nikodym derivative (also called the likelihood ratio in this context) $f(x) \doteq (d\theta/d\pi)(x)$ exists. Iid samples $Y_0, Y_1, \ldots$ with distribution π are generated, and the estimate

$$\bar{m}_K \doteq \frac{1}{K} \sum_{k=0}^{K-1} G(Y_k) f(Y_k)$$

is formed. Since

$$EG(Y_k) f(Y_k) = \int_S G(x) f(x) \pi(dx) = \int_S G(x) \theta(dx) = m,$$

$\bar{m}_K$ is an unbiased estimate of m, with a rate of convergence determined by

$$\text{var}\left[G(Y_0) f(Y_0)\right] = \int_S G(x)^2 f(x) \theta(dx) - \left[\int_S G(x) \theta(dx)\right]^2.$$

Standard Monte Carlo corresponds to $f = 1$, and the goal of importance sampling is to choose f in such a way that: (i) the variance is lowered significantly, and (ii) sampling from π is not too difficult. Note that minimizing the variance with respect to f is equivalent to minimizing the second moment, and so if posed as an optimization problem, one can use the simpler second moment in lieu of variance. Note also that without further restriction on the class of sampling measures the problem is in some sense ill-posed. For example, suppose θ is supported on $[0, \infty)$, and $\theta(dx) = g(x)\,dx$. Let $G(x) \doteq x$ so that $m = EX \neq 0$ and let $\pi(dx) \doteq m^{-1} x g(x)\,dx$. Then θ is absolutely continuous with respect to π, with $f(x) = m/x$. Furthermore,

$$\text{var}\left[Y_0 f(Y_0)\right] = \int_{[0,\infty)} x^2 f(x) \theta(dx) - m^2 = 0.$$

However, such a distribution π is of little use in practice since it requires knowledge of m, the very thing we want to estimate! Instead of this unconstrained optimization, one typically seeks to minimize over parameterized families of alternative sampling distributions.

14.2.1 Importance Sampling for Rare Events

We now return to the discrete time model of Sect. 14.1. Recall the notation

$$H(y, \alpha) = \log E \exp\left\{\langle \alpha, v_i(y) \rangle\right\}, \quad L(y, \beta) = \sup_{\alpha \in \mathbb{R}^d} \left[\langle \alpha, \beta \rangle - H(y, \alpha)\right], \quad (14.6)$$

and consider the problem of estimating $p^n(x)$ as defined in (14.3). When $p^n(x)$ is small (e.g., on the order of say 10^{-6}) ordinary Monte Carlo attempts to estimate this number as a convex combination of 0's and 1's. The goal of importance sampling (and indeed any accelerated Monte Carlo scheme) is to produce estimators whose distribution is more closely clustered around the target value of 10^{-6}.

As we have just noted, the problem of optimizing over all changes of measure is in some sense ill-posed, and thus the first question is, "what are natural changes of measure?" A hint is provided by the analysis of Chap. 4. The control measures $\bar{\mu}_i^n$ of the weak convergence approach correspond to a change of measure for the noise sequence. An a posteriori conclusion of the large deviation analysis is that *exponential* changes of measure are asymptotically optimal in the representation. (See, for example, the measures γ defined in part (g) of Lemma 4.16, and their use in the proof of the Laplace lower bound proof in Sect. 4.7.) Exponential changes of measure have a finite dimensional parameterization, and thus are convenient to work with. Recalling that $\{v_i(x), i \in \mathbb{N}\}$ are iid with distribution $\theta(dv|x)$ and associated log moment generating functions $H(x, \alpha)$, this suggests that measures of the form

$$\eta_\alpha(dv|x) = e^{\langle \alpha, v\rangle - H(x,\alpha)}\theta(dv|x)$$

be used to generate the noise sequence under the new distribution. We will show later on that changes of measure within this class are sufficient for asymptotic optimality. The parameter α can be thought of as a control, which is selected to produce good performance of the resulting Monte Carlo scheme. In this context η_α is sometimes referred to as an **exponential tilt** of θ, with α the **tilt parameter**.

While more complicated dependencies could be considered, it will turn out (for the models of Chap. 4) that allowing α to depend on time and the current state of the simulated trajectory will be sufficient for asymptotic optimality. Thus a control scheme (i.e., a change of measure) will be characterized as a collection of measurable mappings $\alpha_i^n : \mathbb{R}^d \to \mathbb{R}^d$, defined for $i \in \mathbb{N}_0$. The generation of a single sample as well as the likelihood ratio needed to estimate $p^n(x)$ then proceeds as follows.

We initialize with $Y_0^n = x$. A sequence of noises w_i^n and states Y_{i+1}^n are then generated recursively by

$$P_x\left\{w_i^n \in dv \middle| \mathscr{F}_i^n\right\} = \eta_{\alpha_i^n(Y_i^n)}(dv|Y_i^n), \text{ with } \mathscr{F}_i^n = \sigma\left(w_j^n, j = 0, \ldots, i-1\right)$$

and

$$Y_{i+1}^n = Y_i^n + \frac{1}{n}w_i^n.$$

The simulation proceeds up until

$$N^n \doteq \inf\left\{i : Y_i^n \in A \cup B\right\},$$

and we define $Y^n(t)$ to be the piecewise linear interpolation, so that $1_{\{Y^n \in C_x\}}$ means B was entered before A. The likelihood ratio is then

$$\prod_{i=0}^{N^n-1} \frac{d\theta(\cdot|Y_i^n)}{d\eta_{\alpha_i^n(Y_i^n)}(\cdot|Y_i^n)}(w_i^n) = \prod_{i=0}^{N^n-1} e^{-\left\langle \alpha_i^n(Y_i^n), w_i^n \right\rangle + H(Y_i^n, \alpha_i^n(Y_i^n))},$$

and the estimate based on a single sample is thus

$$1_{\{Y^n \in C_x\}} \prod_{i=0}^{N^n-1} e^{-\left\langle \alpha_i^n(Y_i^n), w_i^n \right\rangle + H(Y_i^n, \alpha_i^n(Y_i^n))}. \tag{14.7}$$

As discussed previously, one then simulates K independent copies of (14.7) and takes the sample average, where K depends on the variance of a single sample and the desired accuracy.

We recall that performance is determined by the variance of a single sample, and minimizing this is the same as minimizing the second moment. The second moment of (14.7) is

$$E_x \left[1_{\{Y^n \in C_x\}} \prod_{i=0}^{N^n-1} e^{-2\left\langle \alpha_i^n(Y_i^n), w_i^n \right\rangle + 2H(Y_i^n, \alpha_i^n(Y_i^n))} \right],$$

which when rewritten in terms of the distribution of the *original process* $\left\{X_i^n\right\}$ takes the form

$$E_x \left[1_{\{X^n \in C_x\}} \prod_{i=0}^{N^n-1} e^{-\left\langle \alpha_i^n(X_i^n), v_i(X_i^n) \right\rangle + H(X_i^n, \alpha_i^n(X_i^n))} \right].$$

14.2.2 Controls Without Feedback, and Dangers in the Rare Event Setting

Since one of the classical approaches to the large deviation lower bound involves a change of measure argument, it is natural to ask if there is a connection between the change of measure (equivalently control measure) used there to prove bounds for a particular event or expected value, and a change of measure that would produce a good IS scheme for that same event. Note that there are actually many changes of measure that *could* be used to prove the lower bound. Here we mean the one that is typically used in the proof, and which uses a deterministic sequence $\alpha_i^n(x)$ that depends on i but not x, which we refer to as an "open loop" control. It turns out that in some special circumstances one can achieve asymptotic optimality within the class of open loop controls (e.g., [232]), and for some time it was generally thought that using this lower bound change of measure would work well in general. This turned out to be false, and indeed the class of schemes that had been considered up to that time turned out to be, in general, inadequate. In this section we illustrate the issue through an example due to [150]. The techniques we develop to understand the particular example are broadly useful for understanding rare event importance sampling. Of

special importance is the game characterization of performance described in the next section.

The example is as follows. Suppose that $v_i(X_i^n)$ are in fact independent of X_i^n, i.e., that they are just an iid sequence with distribution θ. We further assume $d = 1$ and that $X_0^n = 0$ (for the rest of this section we write P and E rather than P_0 and E_0). Then X_i^n is a random walk, and $X_n^n = \frac{1}{n}\sum_{i=0}^{n-1} v_i$ is just the sample mean, i.e., we are in the setting of Cramér's theorem with rate function $L(\beta)$ (see Sect. 3.1.6). Let $B \subset \mathbb{R}$, and suppose we want to estimate $P\left\{X_n^n \in B\right\}$ by importance sampling.

The heuristic just described to construct an alternative sampling distribution is straightforward to implement. Let β^* solve $\inf[L(\beta) : \beta \in B]$ (and assume the infimum over the interior and closure of B are the same). If α^* is dual to β^*, i.e., if α^* is the point that maximizes in the relation $L(\beta^*) = \sup_{\alpha\in\mathbb{R}}[\alpha\beta^* - H(\alpha)]$, then as discussed in Chap. 4 [see part (g) of Lemma 4.16], the mean of $\eta_{\alpha^*}(dv) \doteq e^{\alpha^* v - H(\alpha^*)}\theta(dv)$ is exactly β^*, and $\mu_i^n = \eta_{\alpha^*}$ is the control one could use to prove the large deviation lower bound. Since this problem is over a fixed time horizon, the single sample estimate is just

$$1_{\{Y_n^n \in B\}} \prod_{i=0}^{n-1} e^{-\alpha^* w_i^n + H(\alpha^*)} = 1_{\{Y_n^n \in B\}} e^{-n\left[\alpha^* Y_n^n - H(\alpha^*)\right]}.$$

One can now describe the shortcomings of the open loop heuristic. Assume that θ is Gaussian $N(0, 1)$ and consider the *nonconvex* set $B = (-\infty, -0.25] \cup [0.2, \infty)$ (see Fig. 14.2). For this process $L(\beta) = \beta^2/2$, $H(\alpha) = \alpha^2/2$, and $\alpha^* = \beta^* = 0.2$, and the change of measure will shift the mean to this value. If all goes according to plan and the simulated trajectory ends up near β^*, then the likelihood ratio will be near $\exp\{-n\left[\alpha^*\beta^* - H(\alpha^*)\right]\} = \exp\{-nL(\beta^*)\}$. Thus the estimator is either zero or close to the large deviation approximation to the probability, which is just the sort of qualitative behavior that is needed. However, it is also possible that an event that is rare under the $\eta_{\alpha^*}(dv)$ distribution may occur, and one can end up with Y_n^n that is in the interval $(-\infty, -0.25]$. Such an occurrence is labeled the "rogue" simulation in Fig. 14.2 . When this happens, the likelihood ratio will be approximately

$$\exp\left\{-n\left[\alpha^*\bar{\beta} - H(\alpha^*)\right]\right\} = \exp\left\{n\left[0.2 \times 0.25 + \frac{1}{2}(0.2)^2\right]\right\}.$$

This quantity grows exponentially in n and, while the event itself might be rare, it happens often enough that the variance of the estimate is very large, and even larger than standard Monte Carlo!

In this example the true probability for $n = 60$ is $p^n = 8.71 \times 10^{-2}$, which can be calculated using the known distribution of X_n^n. The data in Table 14.1 reflect four trials of $K = 5000$ replications. The "standard error" is the estimated standard deviation for the entire trial, and $\hat{\mathfrak{S}}^n$ is the estimate of the second moment based on the data.

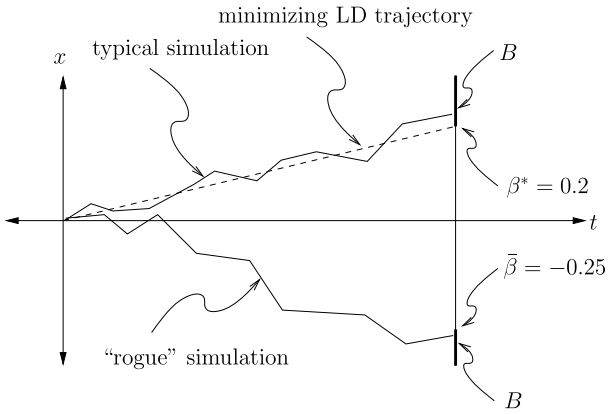

Fig. 14.2 An expected trajectory and a rogue trajectory

Table 14.1 Importance sampling implementation based on an open loop control

	No. 1	No. 2	No. 3	No. 4
Estimate $\hat{p}^n$ ($\times 10^{-2}$)	18.36	7.51	5.95	8.02
Standard error ($\times 10^{-2}$)	7.41	1.37	0.118	1.80
95% confidence interval ($\times 10^{-2}$)	[3.83, 32.89]	[4.82, 10.20]	[5.72, 6.18]	[4.49, 11.55]
Number of "rogue" trajectories	3	1	0	1
$(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$	−1.96	0.0210	1.62	−0.193

The first, second and fourth trials have 3, 1 and 1 "rogue" trajectories, respectively. In contrast the third has none. While the third estimate has a small standard error and associated confidence interval, the interval does not contain the true value. The estimate is smaller than the true value, reflecting the fact that the estimate has never sampled from the interval $(-\infty, -0.25]$, and is therefore in some sense providing an estimate of only the probability to end in $[0.2, \infty)$. Both the estimate and the estimate of the standard deviation are misleading, and it is in fact the same difficulties that affect the estimation of p_n that make the confidence interval essentially useless, though one does not a priori know this is the case. Because of this, an independent theoretical (and not only data driven) analysis of errors is important for rare event Monte Carlo estimation. All of the other trials include at least one rogue trajectory, which is needed to avoid the bias of trial 3. The estimates may be far from the true value, but in this case at least the confidence intervals are correctly indicating this fact. If the estimates were accurate, $(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$ should be close to 2 for asymptotic optimality. This appears to be to some degree valid for trial 3, but for reasons mentioned previously this is misleading.

One could argue that the difficulties encountered in this example can be avoided by splitting the problem into that of estimating two half-infinite intervals. While such an approach would work here, it will fall apart as soon as one considers problems in

higher dimensions or even slightly more complicated dynamics. What is needed is a *global* approach that properly controls the likelihood ratio for any possible simulated trajectory.

14.2.3 A Dynamic Game Interpretation of Importance Sampling

Further insight into the difficulties of IS in the rare event setting can be obtained by modeling the performance in terms of prelimit *small noise stochastic game* and limiting *deterministic differential game*. Although in this section we develop this connection only for the simple random walk model just discussed, it is easily adapted to other situations. Suppose that for the iid random walk model and problem of the last section we consider, instead of the constant control α^* suggested by the standard heuristic, a collection of sampling controls of the general form $\alpha_i^n(x)$, and in particular assume

$$\alpha_i^n(Y_i^n) = u(Y_i^n, i/n)$$

for some smooth function $u : \mathbb{R} \times [0, 1] \to \mathbb{R}$ (we assume $d = 1$ for simplicity). In this case, the second moment of a single sample, and hence the performance of the scheme, is given by the exponential integral

$$E\left[1_{\{Y_n^n \in B\}} \prod_{i=0}^{n-1} e^{-2u(Y_i^n, i/n)w_i^n + 2H(u(Y_i^n, i/n))}\right],$$

which we can rewrite in terms of the original process as

$$E\left[1_{\{X_n^n \in B\}} \prod_{i=0}^{n-1} e^{-u(X_i^n, i/n)v_i + H(u(X_i^n, i/n))}\right].$$

Note that this is the sort of Laplace functional for which relative entropy representations are derived in Chaps. 3 and 4 (see for example Proposition 3.1 and Theorem 4.5), although previously we have assumed (e.g., in Proposition 2.3) that the quantity appearing in the exponent was at least bounded either from above or below. This boundedness will not hold if the support of $\{v_i\}$ is unbounded above and below. Setting aside the issue of boundedness, the quantity in the last display is still expected to scale exponentially in n, and thus it is natural to consider the log transform. Using the same notation for the controls (measures) and controlled processes as in Sect. 4.2, we formally have

$$-\frac{1}{n}\log E\left[1_{\{X_n^n\in B\}}\prod_{i=0}^{n-1}e^{-u(X_i^n,i/n)v_i+H(u(X_i^n,i/n))}\right]=\inf_{\{\bar{\mu}_i^n\}}E\left[\frac{1}{n}\sum_{i=0}^{n-1}R\left(\bar{\mu}_i^n\|\theta\right)\right.$$
$$\left.+\frac{1}{n}\sum_{i=0}^{n-1}\left[u(\bar{X}_i^n,i/n)\bar{v}_i-H(u(\bar{X}_i^n,i/n))\right]+\infty 1_{B^c}\left(\bar{X}_n^n\right)\right].$$

Keeping in mind that to minimize the variance we will supremize the right hand side over $u(\cdot,\cdot)$, the optimal variance is then characterized in terms of a discrete time small noise stochastic game. One player (corresponding to u) is given in feedback form as a function of the state and seeks to maximize. The other player (with controls $\{\bar{\mu}_i^n\}$) arises from the representation, and seeks to minimize. This player's controls can depend on the state (in fact the whole history) as well as u, and since it seeks to minimize the cost, must drive the process into B at time n with probability one. Note that the class of open loop controls as they would be used in IS schemes correspond to eliminating state feedback in u, i.e., restricting to the form $u(x,s)=u(s)$.

One can calculate the limit in the last display using the same weak convergence methods as those used in Chap. 4 to study the LDP for $\{X^n\}$, and for a fixed bounded and continuous control u the limit is characterized by the optimization problem

$$J[u]=\inf_{\phi}\left[\int_0^1\left[u(\phi(t),t)\dot{\phi}(t)-H(u(\phi(t),t))+L(\dot{\phi}(t))\right]dt+\infty 1_{B^c}\left(\phi(1)\right)\right],$$

where the infimum is over absolutely continuous ϕ with $\phi(0)=0$.

The quantity $J[u]$ gives the rate of decay of the second moment of the IS scheme that uses the sampling control $\alpha_i^n(Y_i^n)=u(Y_i^n,i/n)$ to dynamically choose the change of measure. For the purposes of IS scheme selection, one can consider this simpler limit problem which characterizes the rate of decay. Thus we consider $U=\sup_{u(\cdot,\cdot)}J[u]$. This is a type of deterministic differential (or dynamic) game, where $\dot{\phi}$ (replacing $\{\bar{\mu}_i^n\}$) attempts to minimize (in open loop form) and u attempts to maximize (in feedback form, but u must be selected before ϕ is chosen).

Suppose we extend the definition to allow for an arbitrary initial condition (x,t) (i.e., we consider the cost over $[t,1]$ and with $\phi(t)=x$), and denote the corresponding optimal rate of decay by $U(x,t)$. Let U_t be the partial with respect to t and $DU(x,t)$ the gradient in x. Then $U(x,t)$ will be a viscosity solution to

$$U_t(x,t)+\sup_{\alpha\in\mathbb{R}}\inf_{\beta\in\mathbb{R}}\left[DU(x,t)\beta+\alpha\beta-H(\alpha)+L(\beta)\right]=0 \tag{14.8}$$

and the terminal condition

$$U(x,1)=\infty \text{ for } x\in B^c \text{ and } U(x,1)=0 \text{ for } x\in B. \tag{14.9}$$

For properties of viscosity solutions that will be used here (though these arguments are only intended to be formal), we refer to [14].

We will not delve deeply into the nuances of differential games, since this game has a special structure which allows a reduction to a much simpler problem. Using the Minimax theorem [233] and that L is the Legendre-Fenchel transform of H, we observe that

$$\begin{aligned}\sup_{\alpha\in\mathbb{R}}\inf_{\beta\in\mathbb{R}}\left[p\beta+\alpha\beta-H(\alpha)+L(\beta)\right] &= \inf_{\beta\in\mathbb{R}}\sup_{\alpha\in\mathbb{R}}\left[p\beta+\alpha\beta-H(\alpha)+L(\beta)\right]\\ &= \inf_{\beta\in\mathbb{R}}\left[p\beta+2L(\beta)\right].\end{aligned}$$

Not surprisingly then, the PDE (14.8) is closely related to ones that are connected with the large deviation rate function for the original process. Define $\mathbb{H}(p)\doteq\inf_{\beta\in\mathbb{R}}[p\beta+L(\beta)]$. This form of the Legendre transform, which is natural when discussing PDEs, is related to the form usually used in large deviation theory (e.g., a log moment generating function) by $\mathbb{H}(p)=-H(-p)$. Then the Isaacs equation (14.8) can be rewritten as

$$U_t(x,t)+2\mathbb{H}(DU(x,t)/2)=0.$$

Suppose we consider the probability $P\{X_n^n\in B\}$, but generalize to allow an arbitrary initial point x and starting time i/n. Let

$$V^n(x,i/n)\doteq-\frac{1}{n}\log P\left\{x+\frac{1}{n}\sum_{j=i}^{n-1}v_i\in B\right\}.$$

If $i/n\to t$ as $n\to\infty$, then by Cramér's theorem $V^n(x,i/n)\to V(x,t)\doteq\inf[(1-t)L(\beta):x+(1-t)\beta\in B]$, and it is straightforward to verify that $V(x,t)$ is a viscosity solution to the problem with the same terminal condition (14.9) as U and the PDE

$$V_t(x,t)+\mathbb{H}(DV(x,t))=0. \tag{14.10}$$

Using the fact that a comparison principle holds for viscosity solutions to these PDE, is follows that $U(x,t)=2V(x,t)$, which is consistent with the claim made previously (see Sect. 14.1.1) that the best possible rate of decay for the second moment of any IS is precisely twice the large deviation rate. It in fact suggests more, which is that within the class of IS schemes based on feedback and exponential changes of measure, one can in fact achieve this best decay rate.

The situation just described, which we have presented here in the context of Cramér's theorem and for a particular event, is in fact generic under the small noise large deviation scaling [116]. Note the remarkable fact that the equation for U, which models a *game*, turns out to be equivalent to the equation for V, which models a *calculus of variations* or *control* problem [14]. The Isaacs equation for U identifies (at least for smooth solutions) optimal controls for both players. Evaluating the infimum in β in (14.8) gives

$$\begin{aligned}
&\sup_{\alpha\in\mathbb{R}} \inf_{\beta\in\mathbb{R}} [DU(x,t)\beta + \alpha\beta - H(\alpha) + L(\beta)] \\
&\quad = \sup_{\alpha\in\mathbb{R}} \left[-\sup_{\beta\in\mathbb{R}} [-(DU(x,t)+\alpha)\beta - L(\beta)] - H(\alpha) \right] \\
&\quad = -\inf_{\alpha\in\mathbb{R}} [H(-DU(x,t)-\alpha) + H(\alpha)] .
\end{aligned}$$

Suppose that the distribution of v_i does not concentrate on a single point, so that H is strictly convex. Then the optimal α is given by the unique solution of $H'(\alpha) = H'(-DU(x,t)-\alpha)$, which is $\alpha = -DU(x,t)/2$. In terms of the value function associated with the large deviation control problem this is simply $\alpha = -DV(x,t)$. Although it is not needed or used, the optimal control for the large deviation player [but for the second moment, and not for the original event!] can be found by solving $L'(\beta) = -DU(x,t)/2$.

It turns out that one does not need to solve the game or control problem, and in fact the construction of suitable *subsolutions* to the associated PDE (14.10) will be sufficient for a certain level of performance, in a sense that will be made precise in Chap. 15. This is a significant simplification, because for many interesting classes of problems such subsolutions can be constructed explicitly. The reason subsolutions suffice is because the goal in algorithm design is lower bounds on the rate of decay of the second moment. The verification of these one-sided bounds require only certain inequalities, which coincide with the subsolution definition.

In the next section the definitions of classical and piecewise classical subsolution are given. It will turn out to be much easier for many problems to find appropriate piecewise classical subsolutions, so this generalization is important. We also spell out how the various subsolutions generate sampling schemes.

14.3 Subsolutions

We will describe the subsolutions needed for both finite time problems (as in Sect. 14.2.2) and exit probability problems (as in Sect. 14.1.1). We begin with the finite time problem, which generalizes the example used in Sect. 14.2.2. Processes will be of interest on a continuous time interval of the form $[0, T]$, $T < \infty$, and to simplify the notation we assume Tn is an integer. As in Section 14.1 let $\{v_i(x), i \in \mathbb{N}_0, x \in \mathbb{R}^d\}$ be iid random vector fields given on some probability space with the property that for each $x \in \mathbb{R}^d$ $v_i(x)$ has distribution $\theta(\cdot|x)$, where $\theta(dy|x)$ is a stochastic kernel on $\mathbb{R}^d$ given $\mathbb{R}^d$. Recall the discrete time Markov process $\{X_i^n\}_{i=0,\ldots,Tn}$ defined by the recursion

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \quad X_0^n = x_0,$$

and the continuous time interpolation defined by

$$X^n(t) = X_i^n + \left[X_{i+1}^n - X_i^n\right](nt - i), \quad t \in [i/n, i/n + 1/n], i = 0, 1, \ldots, Tn.$$

Also, we assume $H(x, \alpha) = \log E \exp\{\langle \alpha, v_i(x)\rangle\} < \infty$ for all $x \in \mathbb{R}^d$ and $\alpha \in \mathbb{R}^d$.

The importance sampling problem of interest is to estimate

$$P_{x_0}\left\{X^n(T) \in B\right\},$$

where $B \subset \mathbb{R}^d$. As for the one dimensional setting considered in Sect. 14.2.3, the PDE that characterizes the large deviation rate and half the optimal rate of decay for an asymptotically optimal importance sampling scheme is

$$V_t(x, t) + \mathbb{H}(x, DV(x, t)) = 0 \tag{14.11}$$

for $(x, t) \in \mathbb{R}^d \times [0, T)$, where $\mathbb{H}(x, p) = -H(x, -p)$. The terminal condition is

$$V(x, T) = \infty \text{ for } x \in B^c \text{ and } V(x, T) = 0 \text{ for } x \in B. \tag{14.12}$$

Definition 14.1 A function $\bar{V} : \mathbb{R}^d \times [0, T] \to \mathbb{R}$ is a **classical sense subsolution** (or simply a classical subsolution) if it is continuously differentiable in both variables and if

$$\bar{V}_t(x, t) + \mathbb{H}(x, D\bar{V}(x, t)) \geq 0$$

for all $(x, t) \in \mathbb{R}^d \times [0, T)$ and

$$\bar{V}(x, T) \leq \infty \text{ for } x \in B^c \text{ and } \bar{V}(x, T) \leq 0 \text{ for } x \in B.$$

Note that the condition $\bar{V}(x, T) \leq \infty$ for $x \in B^c$ is vacuous. Let $\wedge_{j=1}^J a_j$ denote the minimum of real numbers a_j, $j = 1, \ldots, J$.

Definition 14.2 A function $\bar{V} : \mathbb{R}^d \times [0, T] \to \mathbb{R}$ is a **piecewise classical sense subsolution** (or simply a piecewise classical subsolution) if the following hold. There are $J \in \mathbb{N}$ and functions $\bar{V}^{(j)} : \mathbb{R}^d \times [0, T] \to \mathbb{R}$, $j = 1, \ldots, J$, that are continuously differentiable in both variables and satisfy

$$\bar{V}_t^{(j)}(x, t) + \mathbb{H}(x, D\bar{V}^{(j)}(x, t)) \geq 0$$

for all $(x, t) \in \mathbb{R}^d \times [0, T)$. Moreover $\bar{V}(x, t) \doteq \wedge_{j=1}^J \bar{V}^{(j)}(x, t)$ satisfies

$$\bar{V}(x, T) \leq \infty \text{ for } x \in B^c \text{ and } \bar{V}(x, T) \leq 0 \text{ for } x \in B.$$

Example 14.3 Consider again the iid random walk example of Sect. 14.2.2, where $H(\alpha) = \log E e^{\alpha v_i}$ and $\{v_i\}_{i \in \mathbb{N}}$ are iid real random variables with mean zero. Without loss of generality we take the time horizon $T = 1$. The set B in the example was of the form $(-\infty, \bar{\beta}] \cup [\beta^*, \infty)$, with $\bar{\beta} < 0 < \beta^*$. The solution to (14.11) and (14.12) is

$$V(x,t) = \inf\left[(T-t)L(\beta) : x + (T-t)\beta \in (-\infty, \bar{\beta}] \cup [\beta^*, \infty)\right].$$

For this example it is natural to look for a piecewise classical subsolution as the minimum of two functions. One can easily construct solutions to the PDE by assuming the simple form $-ax + bt + c$ and requiring that $b + \mathbb{H}(-a) = b - H(a) = 0$ hold. If $\hat{\alpha}$ and $\hat{\beta}$ are convex dual points, i.e.,

$$L(\hat{\beta}) = \sup_{\alpha \in \mathbb{R}} \left[\alpha\hat{\beta} - H(\alpha)\right] = \hat{\alpha}\hat{\beta} - H(\hat{\alpha}),$$

we obtain the solution $-\hat{\alpha}(x - \hat{\beta}) + (L(\hat{\beta}) - \hat{\alpha}\hat{\beta})[1-t]$, which corresponds to the terminal condition $-\hat{\alpha}(x - \hat{\beta})$. Note that since $Ev_i = 0$ Jensen's inequality implies $H(\hat{\alpha}) \geq 0$, and so $\hat{\alpha}\hat{\beta} \geq 0$. Thus $\hat{\beta} > 0$ if and only if $\hat{\alpha} > 0$.

We conclude that the two solutions

$$\bar{V}^{(1)}(x,t) = -\alpha^*(x - \beta^*) + (L(\beta^*) - \alpha^*\beta^*)[1-t],$$
$$\bar{V}^{(2)}(x,t) = -\bar{\alpha}(x - \bar{\beta}) + (L(\bar{\beta}) - \bar{\alpha}\bar{\beta})[1-t],$$

which correspond to the terminal conditions indicated in Fig. 14.3, generate the piecewise classical subsolution $\bar{V} \doteq \bar{V}^{(1)} \wedge \bar{V}^{(2)}$. Note that since the (α, β) pairs are convex dual points, α^* and $\bar{\alpha}$ generate changes of measure with the means β^* and $\bar{\beta}$, respectively. See Fig. 14.4. The dotted line in the figure represents points (x,t) for which $\bar{V}^{(1)}(x,t) = \bar{V}^{(2)}(x,t)$. Note that the *subsolution* $\bar{V}(x,t)$ has a much simpler structure than the *solution* $V(x,t)$, but it also has the same (maximal) value at $(0,0)$, namely $\left[L(\beta^*) \wedge L(\bar{\beta})\right]$.

Consider next the problem of entering a rare set B before a typical set A (Fig. 14.5). Thus the importance sampling problem is to estimate

$$P_{x_0}\left\{X^n \text{ enters } B \text{ before entering } A\right\}.$$

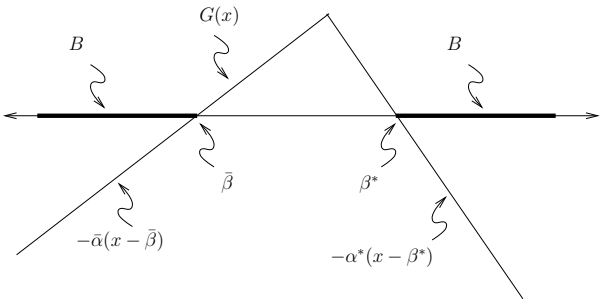

Fig. 14.3 Terminal condition corresponding to a subsolution

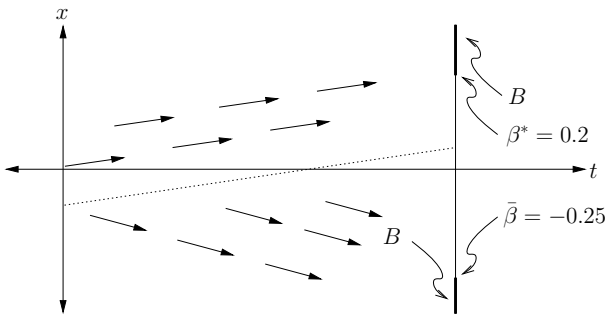

Fig. 14.4 Partition of the domain by a piecewise classical subsolution

The definitions for classical and piecewise classical subsolutions are similar to the finite time case. The relevant PDE is

$$\mathbb{H}(x, DV(x)) = 0, \tag{14.13}$$

with the boundary condition

$$V(x) = 0 \text{ for } x \in \partial B. \tag{14.14}$$

Definition 14.4 A function $\bar{V} : \mathbb{R}^d \to \mathbb{R}$ is a **classical sense subsolution** (or simply a classical subsolution) of (14.13)–(14.14) if it is continuously differentiable and if

$$\mathbb{H}(x, D\bar{V}(x)) \geq 0$$

for all $x \in (A \cup B)^c$, and if

$$\bar{V}(x) \leq 0 \text{ for } x \in B.$$

Definition 14.5 A function $\bar{V} : \mathbb{R}^d \to \mathbb{R}$ is a **piecewise classical sense subsolution** (or simply a piecewise classical subsolution) if the following hold. For some $J \in \mathbb{N}$ there are functions $\bar{V}^{(j)} : \mathbb{R}^d \to \mathbb{R}$, $j = 1, \ldots, J$, that are continuously differentiable and satisfy

$$\mathbb{H}(x, D\bar{V}^{(j)}(x)) \geq 0$$

for all $x \in (A \cup B)^c$. Moreover $\bar{V}(x) \doteq \wedge_{j=1}^{J} \bar{V}^{(j)}(x)$ satisfies

$$\bar{V}(x) \leq 0 \text{ for } x \in B.$$

Remark 14.6 (boundary condition for ∂A) In general one should specify a boundary condition for A as well. Since we have taken A to be open, the appropriate

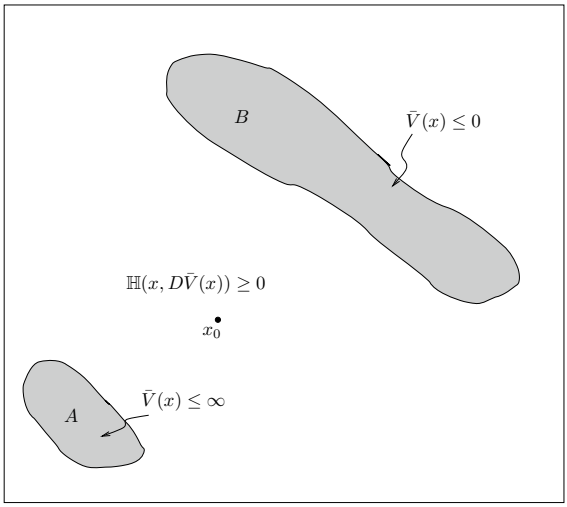

Fig. 14.5 Subsolution for the exit problem

boundary condition is the one which corresponds to a "state space constraint" [234]. For example, if ∂A were smooth with $x \in \partial A$ and n an outward normal to A at x, then the classical formulation of the state space constraint is

$$\inf_{\beta:\langle\beta,n\rangle\ge 0} [\langle DV(x), \beta\rangle + L(x,\beta)] = 0,$$

which reflects the fact that any candidate trajectory in the definition of $V(x)$ cannot enter A. Since our approach to rare event simulation is based on the construction of suitable classical and piecewise classical subsolutions, this boundary condition is vacuous. Indeed, we will assume that $\bar{V}$ is a subsolution in the sense of either Definition 14.4 or 14.5, and as a consequence the boundary condition for a subsolution will hold automatically. For example, in the context of Definition 14.4

$$\begin{aligned}\inf_{\beta:\langle\beta,n\rangle\ge 0} \left[\left\langle D\bar{V}(x), \beta\right\rangle + L(x,\beta)\right] &\ge \inf_{\beta\in\mathbb{R}^d} \left[\left\langle D\bar{V}(x), \beta\right\rangle + L(x,\beta)\right] \\ &= -H(x, -D\bar{V}(x)) \\ &= \mathbb{H}(x, D\bar{V}(x)) \\ &\ge 0.\end{aligned}$$

For a piecewise classical subsolution the concavity of $\mathbb{H}(x, p)$ gives the analogous bound. If instead we had assumed that A is closed with the attractor in the interior of A, the appropriate boundary condition ($V(x) = \infty$ for $x \in \partial A$) would again be vacuous when used to characterize a subsolution ($\bar{V}(x) \le \infty$ for $x \in \partial A$). The fact that these boundary conditions are vacuous can also be seen from the proofs of asymptotic optimality, where they play no role.

Of course there are many other types of events and (risk-sensitive) expected values that one could consider, and the interested reader can find the appropriate definitions of subsolutions for many of these in the references and Sect. 14.5. However, the two examples of this section will suffice to illustrate the main points.

14.4 The IS Scheme Associated to a Subsolution

We next discuss importance sampling schemes associated with a particular subsolution. Consider first the finite time problem. As discussed at the end of Sect. 14.2.3, if a smooth solution $V(x,t)$ to the HJB equation were available, then the correct change of measure if the current state of the simulated trajectory is at Y_i^n would be to replace the original distribution on the noise $v_i(Y_i^n)$, i.e., $\theta(dv|Y_i^n)$, by

$$\eta_\alpha(dv|Y_i^n) = e^{\langle\alpha,v\rangle - H(Y_i^n,\alpha)}\theta(dv|Y_i^n) \text{ with } \alpha = -DV(Y_i^n, i/n).$$

If one is using a classical subsolution $\bar{V}$ to design a scheme we follow exactly the same recipe, and the resulting second moment, rewritten in terms of the original random variables and process model, will equal

$$\mathfrak{S}^n(\bar{V}) \doteq E_{x_0}\left[1_{\{X_{Tn}^n \in B\}} \prod_{i=0}^{Tn-1} e^{\left\langle D\bar{V}(X_i^n, i/n), v_i(X_i^n)\right\rangle + H(X_i^n, -D\bar{V}(X_i^n, i/n))}\right]. \tag{14.15}$$

Rigorous asymptotic bounds on $\mathfrak{S}^n(\bar{V})$ will be derived in Sect. 15.2. It is shown in Theorem 15.1 that the decay rate of the second moment is bounded below by $V(x_0, 0)$ (the large deviation decay rate for the starting point x_0) plus $\bar{V}(x_0, 0)$. If $\bar{V}(x_0, 0) = V(x_0, 0)$ (the maximum possible value) then we have asymptotic optimality.

If dealing with a piecewise classical sense subsolution, the situation is different. In such a case the gradient $D\bar{V}$ is not smooth, and the analysis used to prove asymptotic performance bounds on $\mathfrak{S}^n(\bar{V})$ for the smooth case does not apply. In this case we mollify $\bar{V}$ and consider two associated importance sampling schemes. To be precise, for a small parameter $\delta > 0$ the standard mollification

$$\bar{V}^\delta(x,t) \doteq -\delta \log\left(e^{-\frac{1}{\delta}\bar{V}^{(1)}(x,t)} + \cdots + e^{-\frac{1}{\delta}\bar{V}^{(J)}(x,t)}\right) \tag{14.16}$$

is used. The properties of this mollification are summarized in the following lemma. The straightforward proof is omitted.

Lemma 14.7 *Let $\bar{V}^\delta$ be as in (14.16) where each function $\bar{V}^{(j)}$, $j = 1, \ldots, J$ is continuously differentiable on $\mathbb{R}^d \times [0, T]$. Define the weights*

$$\rho_j^\delta(x,t) \doteq \frac{e^{-\frac{1}{\delta}\bar{V}^{(j)}(x,t)}}{e^{-\frac{1}{\delta}\bar{V}^{(1)}(x,t)} + \cdots + e^{-\frac{1}{\delta}\bar{V}^{(J)}(x,t)}}.$$

Then

$$D\bar{V}^{\delta}(x,t) = \sum_{j=1}^{J} \rho_j^{\delta}(x,t) D\bar{V}^{(j)}(x,t) \text{ and } \bar{V}_t^{\delta}(x,t) = \sum_{j=1}^{J} \rho_j^{\delta}(x,t) \bar{V}_t^{(j)}(x,t). \tag{14.17}$$

Moreover

$$e^{-\frac{1}{\delta}\bar{V}(x,t)} \leq e^{-\frac{1}{\delta}\bar{V}^{\delta}(x,t)} \leq J e^{-\frac{1}{\delta}\bar{V}(x,t)},$$

and therefore

$$\bar{V}(x,t) \geq \bar{V}^{\delta}(x,t) \geq \bar{V}(x,t) - \delta \log J. \tag{14.18}$$

Recall that given an initial condition x_0, we seek a subsolution for which the value at $(x,t) = (x_0, 0)$ is as large as possible. From the convexity of H and the properties (14.17) it is easily checked that $\bar{V}^{\delta}$ is a classical subsolution in the sense of Definition 14.1 whenever $\bar{V}$ is a piecewise subsolution in the sense of Definition 14.2. The inequality (14.18) together with Theorem 15.1 in Chap. 15 then says that the mollification may lead to a loss of performance (a lowering of the decay rate of the second moment) that is at most $\delta \log J$ (see Theorem 15.1). Thus the role of the mollification is to define a mixture whose performance is very close to that of a classical subsolution, without giving up the flexibility and convenience of piecewise subsolutions. There are (at least) two schemes generated by a subsolution of the form (14.16), which we call the *ordinary implementation* and the *randomized implementation*.

Ordinary Implementation. Using the fact that $\bar{V}^{\delta}$ is a classical subsolution whenever $\bar{V}$ is a piecewise subsolution, we follow the standard procedure for classical subsolutions. Given that the state of the current simulated trajectory is Y_i^n, we use the sampling distribution $\eta_{\alpha}(dv|Y_i^n) = e^{\langle \alpha, v\rangle - H(Y_i^n, \alpha)}\theta(dv|Y_i^n)$ with tilt parameter $\alpha = -D\bar{V}^{\delta}(Y_i^n, i/n)$ to generate a random variable w_i^n with the given (conditional) distribution. The state of the system is then updated according to $Y_{i+1}^n = Y_i^n + w_i^n/n$, and we repeat. The likelihood ratio is

$$R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,Tn-1}) = \prod_{i=0}^{Tn-1} e^{\langle D\bar{V}^{\delta}(Y_i^n, \frac{i}{n}), w_i^n\rangle + H(Y_i^n, -D\bar{V}^{\delta}(Y_i^n, \frac{i}{n}))}, \tag{14.19}$$

and the resulting estimator is $1_{\{Y^n(T)\in B\}} R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,Tn-1})$, where $Y^n(t)$ is the continuous time interpolation.

Randomized Implementation. In this case, the estimator is constructed as follows. Given that the state of the current simulated trajectory is Y_i^n, we generate an independent random variable $\kappa_i^n \in \{1, \ldots, J\}$ with probabilities $\rho_j^{\delta}(Y_i^n, i/n)$, and if $\kappa_i^n = j$ then use the sampling distribution with the tilt parameter $\alpha = -D\bar{V}^{(j)}(Y_i^n, i/n)$ to generate w_i^n. In this case the likelihood ratio is

$$R^n(\{Y_i^n, w_i^n\}_{i=0,\dots,Tn-1}) \tag{14.20}$$
$$= \prod_{i=0}^{Tn-1} \left(\sum_{j=1}^{J} \rho_j^\delta \left(Y_i^n, \frac{i}{n} \right) e^{-\langle D\bar{V}^{(j)}(Y_i^n, \frac{i}{n}), w_i^n \rangle - H(Y_i^n, -D\bar{V}^{(j)}(Y_i^n, \frac{i}{n}))} \right)^{-1},$$

and the estimator takes the same form as in the ordinary case.

For both implementations the resulting second moment, rewritten in terms of the original process and noises, is

$$\mathfrak{S}^n(\bar{V}^\delta) \doteq E_{x_0}\left[1_{\{X^n(T)\in B\}} R^n(\{X_i^n, v_i(X_i^n)\}_{i=0,\dots,Tn-1}) \right]. \tag{14.21}$$

where R^n is given by (14.19) or (14.20) depending on which implementation is used. Since $\bar{V}^\delta$ is a classical subsolution, the randomized case includes the ordinary case with $J = 1$ and taking $\bar{V}^{(1)} = \bar{V}^\delta$. It is shown in Theorem 15.1 that the decay rate of the second moment for both implementations is bounded below by $V(x_0, 0)$ (the large deviation decay rate for the starting point x_0) plus $\bar{V}^\delta(x_0, 0)$.

Example 14.8 (Example 14.3 continued) In Example 14.3 a piecewise subsolution was constructed for the problem of Sect. 14.2.2 with a nonconvex set B. We apply this subsolution for the same data ($\beta^* = 0.2$ and $\bar{\beta} = -0.25$) as in Sect. 14.2.2. As before, each trial is based on $K = 5,000$ simulated trajectories. We give the number of "rogue" trajectories (those ending in $(-\infty, -0.25]$) even though that name is no longer appropriate. Recall that the true value for $n = 60$ is $p^n = 8.70 \times 10^{-2}$. Table 14.2 presents data using the ordinary implementation. The estimates are much more stable across the different trials, with confidence intervals that are both small and which contain the true value. Table 14.3 gives the analogous data for the randomized implementation, which is qualitatively very similar to that of the ordinary case. Table 14.4 considers the same model and escape set for the randomized implementation, but for various values of n. The analogous results for the ordinary implementation are omitted since they are similar. Each trial used $K = 20,000$ simulated trajectories. As with $n = 60$, the results are stable and accurate. Note that the ratio $(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$ is increasing in n (though since $\delta > 0$ is fixed it will never reach 2), and that the number of "rogue" trajectories is decreasing in n, reflecting

Table 14.2 Ordinary implementation of mollified subsolution with $\delta = 0.02$

	No. 1	No. 2	No. 3	No. 4
Estimate $\hat{p}^n$ ($\times 10^{-2}$)	8.55	8.73	8.72	8.61
Standard error ($\times 10^{-2}$)	0.183	0.184	0.182	0.182
95% confidence interval ($\times 10^{-2}$)	[8.19, 8.91]	[8.37, 9.10]	[8.36, 9.08]	[8.25, 8.97]
Number of "rogue" trajectories	751	727	833	807
$(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$	1.51	1.52	1.53	1.52

Table 14.3 Randomized implementation of mollified subsolution with $\delta = 0.02$

	No. 1	No. 2	No. 3	No. 4
Estimate $\hat{p}^n$ ($\times 10^{-2}$)	9.02	8.76	8.62	8.91
Standard error ($\times 10^{-2}$)	0.183	0.182	0.181	0.183
95% confidence interval ($\times 10^{-2}$)	[8.66, 9.38]	[8.40, 9.11]	[8.26, 8.97]	[8.55, 9.26]
Number of "rogue" trajectories	802	782	823	883
$(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$	1.54	1.53	1.52	1.53

Table 14.4 Randomized implementation of mollified subsolution with $\delta = 0.02$

	$n = 100$	$n = 200$	$n = 500$
Exact value p^n	2.90×10^{-2}	2.54×10^{-3}	3.88×10^{-6}
Estimate $\hat{p}^n$	2.93×10^{-2}	2.59×10^{-3}	3.81×10^{-6}
Standard error	3.76×10^{-4}	4.82×10^{-5}	1.35×10^{-7}
95% confidence interval	$[2.86, 3.00] \times 10^{-2}$	$[2.49, 2.68] \times 10^{-3}$	$[3.55, 4.08] \times 10^{-6}$
Number of "rogue" trajectories	2176	935	107
$(-\log \hat{\mathfrak{S}}^n)/(-\log \hat{p}^n)$	1.59	1.65	1.74

the fact that the probability associated with $(-\infty, -0.25]$ conditioned on ending in B is decreasing in n.

Remark 14.9 (role of smoothness) The theoretical bounds on performance derived in Chap. 15 make use of the fact that $\bar{V}^\delta$ smooth, and in particular that it is a classical sense subsolution (and not just a viscosity sense subsolution [14, 134]). A natural question is whether this smoothness is necessary. From the perspective of implementation it is certainly convenient, since the change of measure for the increments is based on the gradient of the subsolution. However, one could ask if there is some generalized implementation (e.g., based on sub or superdifferentials) that might allow for less regular subsolutions. Such a construction would require that in the analysis of the second moment we consider the large deviation theory for processes with "discontinuous statistics." The theory for such processes is not well understood in great generality, and in particular there is no rigorous analysis of importance sampling for nonsmooth subsolutions. Given the subtlety in applying importance sampling to rare event estimation, it seems prudent to use the mollification presented previously, which is very easy to implement and for which a rigorous analysis is available. This difference in the properties of subsolutions is one of the key qualitative distinctions between importance sampling and the analogous splitting algorithms to be considered in Chap. 16, for which a weak sense subsolution is known to be sufficient.

Remark 14.10 (achieving asymptotic optimality) Since the mollification can reduce the value of the subsolution at the starting point [i.e., $\bar{V}^\delta(x_0, 0) < V(x_0, 0)$ is possible even when $\bar{V}(x_0, 0) = V(x_0, 0)$], this would seem to be a significant drawback for

importance sampling. However, while there may be other issues to consider when comparing importance sampling and splitting, it is easy to remedy this objection, and in general one can allow $\delta \to 0$ as $n \to \infty$ so as to achieve asymptotic optimality. This issue is discussed in Remark 15.7 and Theorem 15.14.

Remark 14.11 (randomized versus ordinary) When dealing with noise models such as those of (14.1) one may prefer the ordinary implementation over the randomized implementation, since the appropriate change of measure is simply defined by an exponential tilt, and there is no need to generate random variables according to the weights $\rho_j^\delta(\cdot, i/n)$. Note that for these models the distribution of the noise, conditioned on the state X_i^n, is independent in the time variable. For more complex models (e.g., the Markov modulated models discussed in Sect. 7.3) there may be an advantage to using the randomized implementation, since the change of measure is more complex, and requires, for each distinct value of the gradient, the solution of an eigenvalue problem. In particular, if the component functions $\bar{V}^{(j)}$ all have a constant gradient then one must solve at most J eigenvalue problems for the randomized implementation, while the ordinary implementation will typically require that such a problem be solved for each time $i = 0, \ldots, Tn - 1$ of the simulation. An example of this sort appears in Sect. 14.5.5.

Remark 14.12 The implementation of the importance sampling scheme and resulting form of the second moment are entirely analogous for the problem of hitting a rare set before a typical set, save that the scheme has no explicit dependence on time, and Tn is replaced by the first exit time N^n.

14.5 Generalizations

In this section we briefly comment on generalizations with respect to various aspects of the model, including expected values besides probabilities, continuous time models, and more complex noise models. Some generalizations that are very straightforward (e.g., when the local rate function also depends on time) are not discussed.

14.5.1 Functionals Besides Probabilities

Straightforward and natural generalizations in the context of both the finite time problem and the problem of hitting a rare set prior to a typical set involve the computation of risk-sensitive functionals. For example, in the setting of the finite time problem, we may want to compute a quantity such as

$$V^n(x, 0) = -\frac{1}{n} \log E_x \exp\left\{-nF(X_{Tn}^n)\right\},$$

where F is a suitably regular (e.g., continuous) function, and where for convenience in the notation we assume Tn is an integer. Under appropriate conditions $V^n(x,0) \to V(x,0)$, where

$$V(x,t) \doteq \inf\left[\int_t^T L(\phi(s),\dot{\phi}(s))ds + F(\phi(T)) : \phi(t) = x\right], \tag{14.22}$$

and the only difference in the definition of the various forms of subsolution occur in the terminal condition. Thus in Definition 14.1, the condition $\bar{V}(x,T) \le 0$ for $x \in B$ is replaced by $\bar{V}(x,T) \le F(x)$ for $x \in \mathbb{R}^d$.

A single sample of the estimator, with $R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,Tn-1})$ defined by either (14.19) or (14.20) depending on which implementation is used, is

$$F(Y_{Tn}^n)R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,Tn-1}).$$

14.5.2 Continuous Time

When considering continuous time process models a basic issue is numerical implementation. For example, trajectories of the solution to an SDE are usually approximated, e.g., by the Euler-Maruyama method. Since this returns the problem to the discrete time setting, it can be dealt with using the same notions of importance sampling and subsolutions as those already given. (Note that there is still the problem of quantifying the impact of the time discretization, but that is a topic we do not consider here.) In contrast, for continuous time models that are of pure jump form there is no need to discretize time, and one can formulate both the importance sampling and related analysis directly in continuous time.

To keep the presentation brief we will consider just one class of models, but the ideas can easily be generalized. Thus suppose that X^n is a continuous time Markov process of the following form. There is $J \in \mathbb{N}$ and bounded and Lipschitz continuous functions

$$v_j : \mathbb{R}^d \to \mathbb{R}^d, \quad r_j : \mathbb{R}^d \to (0,\infty), \quad j = 1,\ldots,J,$$

where $nr_j(x)$ is the jump intensity of a jump to the point $x + v_j(x)/n$, given that $X^n(t) = x$. Thus X^n has the infinitesimal generator

$$(\mathscr{L}^n f)(x) \doteq \sum_{j=1}^J nr_j(x)\left[f(x + v_j(x)/n) - f(x)\right]$$

for bounded functions $f : \mathbb{R}^d \to \mathbb{R}$. Hence the process waits at the location x for an exponentially distributed time τ with inverse mean $\sum_{j=1}^J nr_j(x)$. After τ units of time, it jumps to the location $x + v_j(x)/n$ with probability proportional to $r_j(x)$

for $j = 1, \ldots, J$. The weighed serve-the-longest queue model of Chap. 13 is of this sort, except that $r_j(x)$ can be equal to zero for some x values.

If we consider the problem of hitting a rare set before a typical one, the continuous time aspect is unimportant, and this problem can be reduced to the discrete time problems considered previously by working with the imbedded discrete time process. This is the approach taken in [105, 110, 117]. In the notation of this chapter, the discrete time model corresponds to

$$\theta(A|x) = \sum_{j=1}^{J} r_j(x)\delta_{v_j(x)}(A) \Big/ \sum_{j=1}^{J} r_j(x).$$

This simplification is not possible with the finite time problem, since a rare outcome depends on the holding time and not just on which jump type is selected at the time of a transition. In this case, we need to stay in the continuous time framework.

The processes X^n take values in $\mathscr{D}([0, T] : \mathbb{R}^d)$, and the local rate function for the sequence $\{X^n\}_{n\in\mathbb{N}}$ is given by

$$L(x, \beta) \doteq \inf \left[\sum_{j=1}^{J} r_j(x)\ell(\bar{r}_j/r_j(x)) : \sum_{j=1}^{J} \bar{r}_j v_j(x) = \beta, \bar{r}_j \in [0, \infty), j = 1, \ldots, J \right],$$

where $x \in \mathbb{R}^d$, $\beta \in \mathbb{R}^d$ and as usual $\ell(z) \doteq z \log z - z + 1$ for $z \in [0, \infty)$. The analogue of the log moment generating function is given by

$$\begin{aligned} H(x, \alpha) &= \sup_{\beta\in\mathbb{R}^d} [\langle \alpha, \beta\rangle - L(x, \beta)] \\ &= \sup_{\bar{r}_j\in[0,\infty), j=1,\ldots,J} \left[\sum_{j=1}^{J} \bar{r}_j \left\langle v_j(x), \alpha \right\rangle - \sum_{j=1}^{J} r_j(x)\ell(\bar{r}_j/r_j(x)) \right] \\ &= \sum_{j=1}^{J} r_j(x) \left[e^{\langle v_j(x),\alpha\rangle} - 1 \right], \end{aligned}$$

with the supremum achieved at $\bar{r}_j = r_j(x)e^{\langle v_j(x),\alpha\rangle}$. As before the PDE that is relevant takes the form (14.11), where $\mathbb{H}(x, p) = -H(x, -p)$, and the terminal condition is as in the discrete time setting. Given a classical subsolution $\bar{V}(x, t)$, the simulated process under the ordinary implementation uses the rates $\bar{r}_j(x, t) = r_j(x)e^{\langle v_j(x),-D\bar{V}(x,t)\rangle}$. The estimate is then $1_{\{Y^n(T)\in B\}}R^n(Y^n)$, where

$$\log R^n(Y^n) = \int_0^T \sum_{j=1}^{J} r_j(Y^n(t)) \left[e^{\langle v_j(Y^n(t)),-D\bar{V}(Y^n(t),t)\rangle} - 1 \right] dt$$

$$- \sum_{i:t_i^n \le T} \left[\langle v_{j_i^n}(Y^n(t_i^n-)), D\bar{V}(Y^n(t_i^n-), t_i^n-) \rangle \right],$$

with t_i^n the jump times of Y^n and with j_i^n identifying the type of jump.

Remark 14.13 For the problem of hitting a rare set before a typical set one could also use the PDE (14.13) and boundary condition (14.14), with

$$\mathbb{H}(x, p) = -\sum_{j=1}^{J} r_j(x)[e^{-\langle v_j(x), p \rangle} - 1].$$

This form of $\mathbb{H}$ differs from the discrete time analogue, but characterizes the same set of subsolutions if used for an exit type problem.

14.5.3 Level Crossing

Problems such as level crossings, which appear in ruin problems from insurance, are of the same general sort as that of hitting a rare set before hitting a typical set. The main distinction is that the "typical set" is not part to the state space, and instead corresponds to the process drifting infinitely far in some direction. For example, consider once again the discrete time setting, suppose that $\theta(dy|x) = \theta(dy)$, $H(\alpha) < \infty$ for all $\alpha \in \mathbb{R}^d$ and also that

$$\int_{\mathbb{R}^d} y_k \theta(dy) < 0 \text{ for } k = 1, \ldots, d. \tag{14.23}$$

Then each component of $X^n(t)$ tends to $-\infty$ as $t \to \infty$. Let $M_k \in (0, \infty)$ for $i = 1, \ldots, d$ and consider the problem of estimating the level crossing probability

$$P\left\{ \sup_{m \in \mathbb{N}} \max_{k=1,\ldots,d} \frac{1}{M_k} \sum_{i=0}^{m-1} (v_i)_k \ge n \right\},$$

where v_i are iid with distribution θ and $(v_i)_k$ denotes the kth component of v_i. This quantity is the same as

$$P_0\left\{ \sup_{t \in [0,\infty)} \max_{k=1,\ldots,d} \frac{[X^n(t)]_k}{M_k} \ge 1 \right\},$$

and can be thought of has hitting the rare set G^c, where $G \doteq \times_{k=1}^d (-\infty, M_k)$, before wandering off to $-\infty$ in each component (the "typical" set). With this analogy in place, the definitions of subsolution and their use are exactly as before. In particular, if $H(\alpha)$ is the log moment generating function of θ and $\mathbb{H}(p) = -H(-p)$, then a smooth function $\bar{V} : \mathbb{R}^d \to \mathbb{R}$ is a classical subsolution if

$$\mathbb{H}(D\bar{V}(x)) \geq 0 \text{ for } x \in G \text{ and } \bar{V}(x) \leq 0 \text{ for } x \in G^c. \tag{14.24}$$

The (now state dependent) alternative sampling distribution for the next increment w_i^n given $Y_i^n = x$ is $e^{-\langle D\bar{V}(x),v\rangle - H(-D\bar{V}(x))}\theta(dv)$, and the estimate is

$$1_{\left\{Y^n_{\bar{N}^n} \in G^c\right\}} \prod_{i=0}^{\bar{N}^n - 1} e^{\langle D\bar{V}(Y_i^n), w_i^n\rangle + H\left(-D\bar{V}(Y_i^n)\right)},$$

where $\bar{N}^n \doteq \inf\left\{i : Y_i^n \in G^c\right\}$. For such problems it is natural to consider piecewise classical subsolutions with one component $\bar{V}^{(k)}$ for each index $k = 1, \ldots, d$. $\bar{V}^{(k)}(x)$ should be of the form $= -\langle \alpha^{(k)}, x\rangle + c^{(k)}$, where $\alpha^{(k)}$ is of the form $a^{(k)}e_k$, $H(\alpha^{(k)}) = 0$, and $c^{(k)} = a^{(k)}M_k$. One can check that under (14.23), for each $k = 1, \ldots, d$ there is exactly one positive number $a^{(k)}$ such that $H(a^{(k)}e_k) = 0$, that with these choices $\bar{V}(x) \doteq \min_{k=1,\ldots,d} \bar{V}^{(k)}(x)$ is a piecewise classical subsolution with $V(0) = \bar{V}(0) = \min_{k=1,\ldots,d} a^{(k)}M_k$. For any problem where the simulation time is potentially unbounded it is important to know that a proposed scheme is practical. In the present setting, for the process that is simulated the increments have conditional distribution $e^{-\langle D\bar{V}^\delta(x),v\rangle - H(-D\bar{V}^\delta(x))}\theta(dv)$. The mean of this distribution points towards the target set, and it follows that $\bar{N}^n < \infty$ and $Y^n_{\bar{N}^n} \in G^c$ with probability one. One can in fact show more, for example that $E\bar{N}^n < \infty$.

14.5.4 Path Dependent Events

In some situations one may be interested in probabilities and related quantities in which the occurrence or not of the rare event is determined by the path of X^n over an interval $[0, T]$. To simplify notation we will consider a homogeneous random walk as in the last section [i.e., $\theta(dy|x) = \theta(dy)$], $T = 1$, and the case of one dimension. Then an example of this type of problem is to compute

$$E_0\left[e^{-nF(X^n(1))} 1_{\{\max_{t\in[0,1]} X^n(t) \geq h\}}\right], \tag{14.25}$$

where $h \in (0, \infty)$ and F is bounded and continuous. Let $l < h$ and define $\tau_h^n \doteq \inf\{t \geq 0 : X^n(t) \geq h\}$ and $\tau_l^n \doteq \inf\{t \geq \tau_h^n : X^n(t) \leq l\}$. A second example is computing $P_0\left\{\tau_l^n \leq 1\right\}$. Of course this is only particularly difficult if the indicated events are rare, and to make this so we assume $\int_{\mathbb{R}} y\theta(dy) < h$.

It is easy to write down the variational problem for the large deviation approximations to these quantities. For example, for the expected value in (14.25) the corresponding variational problem is

$$\inf\left[\int_0^1 L(\dot{\phi}(t))dt + F(\phi(1)) : \phi(0) = 0, \phi(s) \geq h \text{ for some } s \in [0, 1]\right],$$

where the infimum is over absolutely continuous ϕ. To identify the PDE that is related to this problem we introduce a state variable that will indicate whether or not h has been crossed. Denote the simulated process by $Y^n(t)$ and consider the associated indicator process $Z^n(t) \doteq 1_{[h,\infty)}(\max_{s\in[0,t]} Y^n(s))$. Suppose we are given that $(Z^n(t), Y^n(t)) = (1, x)$. Then the event $\{\max_{t\in[0,1]} Y^n(t) \geq h\}$ is certain, and importance sampling schemes for time instants after t can be generated by subsolutions of the PDE

$$\bar{V}_t(1, x, t) + \mathbb{H}(D\bar{V}(1, x, t)) \geq 0, \quad x \in \mathbb{R}, t \in (0, 1), \tag{14.26}$$

with terminal condition

$$\bar{V}(1, x, 1) \leq F(x), \quad x \in \mathbb{R} \tag{14.27}$$

(here we use variables (z, x, t), and $\bar{V}(1, x, t)$ indicates that $z = 1$). If on the other hand we are given $(Z^n(t), Y^n(t)) = (0, x)$, then for the cost to be finite the event $Y^n(s) \geq h$ must occur for some $s \in [t, 1]$, and by the usual logic of dynamic programming the asymptotic optimal future costs after that time will be bounded below by any subsolution $\bar{V}(1, \cdot, \cdot)$ to (14.26). The characterization of a subsolution for times prior to this event is given by

$$\bar{V}_t(0, x, t) + \mathbb{H}(D\bar{V}(0, x, t)) \geq 0, \quad x \in (-\infty, h), t \in (0, 1), \tag{14.28}$$

and

$$\bar{V}(0, x, t) \leq \bar{V}(1, x, t), \quad x \in [h, \infty), t \in (0, 1). \tag{14.29}$$

Note that one must construct the subsolutions in the order first $\bar{V}(1, x, t)$, then $\bar{V}(0, x, t)$. Given classical subsolutions $\bar{V}(0, x, t)$ and $\bar{V}(1, x, t)$, the simulated trajectory $\{Y^n(t)\}$ is defined as follows. Given that the state of the current simulated trajectory is Y_i^n, we use the sampling distribution $\eta_\alpha(dv|Y_i^n) = e^{\langle\alpha,v\rangle - H(\alpha)}\theta(dv)$ with tilt parameter $\alpha = -D\bar{V}(0, Y_i^n, i/n)$ to generate a random variable w_i^n with the given (conditional) distribution if $i < N^n$, where $N^n \doteq \inf\{i : Y_i^n \geq h\} \wedge n$. If $i \geq N^n$ we instead use the tilt parameter $\alpha = -D\bar{V}(1, Y_i^n, i/n)$ to generate a random variable w_i^n. The state of the system is then updated according to $Y_{i+1}^n = Y_i^n + w_i^n/n$, and we repeat. The likelihood ratio is

$$\begin{aligned} R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,n-1}) = {} & \textstyle\prod_{i=0}^{N^n-1} e^{\langle D\bar{V}(0,Y_i^n,\frac{i}{n}),w_i^n\rangle + H(-D\bar{V}(0,Y_i^n,\frac{i}{n}))} \\ & \times \textstyle\prod_{i=N^n}^{n-1} e^{\langle D\bar{V}(1,Y_i^n,\frac{i}{n}),w_i^n\rangle + H(-D\bar{V}(1,Y_i^n,\frac{i}{n}))}, \end{aligned}$$

and the resulting estimator is

$$e^{-nF(Y^n(1))} 1_{\{\max_{t\in[0,1]} Y^n(t)\geq h\}} R^n(\{Y_i^n, w_i^n\}_{i=0,\ldots,n-1}),$$

where $Y^n(t)$ is the continuous time interpolation.

The corresponding set of PDEs for $P_0\left\{\tau_l^n \le 1\right\}$ is similar, save (14.26) and (14.27) are replaced by

$$\bar{V}_t(1,x,t) + \mathbb{H}(D\bar{V}(x,1,t)) \ge 0, \quad x \in (l,\infty), t \in (0,1),$$

and

$$\bar{V}(1,x,t) \le 0, \quad x \in (-\infty, l], t \in [0,1].$$

This construction can be generalized in many directions. For example, with a level crossing problem as in the last section one could consider the event that a particular level is crossed (i.e., one component exceeds its threshold) prior to a level crossing in some other coordinate direction.

14.5.5 Markov Modulated Models

As a final example we consider problems where there are two times scales, as was the case with the models of Sect. 7.3. To keep the discussion simple we consider the case

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(\Xi_{i+1}), \quad X_0^n = x_0, \Xi_1 = \xi,$$

with X_i^n taking values in $\mathbb{R}^d$ and the probability of interest a level crossing as in Sect. 14.5.3. However, the constructions generalize greatly, and other examples can be found in [116]. Recall from Sect. 7.3 that $\{\Xi_i\}_{i\in\mathbb{N}}$ is an S-valued Markov chain with transition probability kernel p and that $\{v_i(\xi)\}_{i\in\mathbb{N}_0}$ is a sequence of iid random vector fields with distribution given by $\theta(\cdot|\xi)$. We assume that the moment generating functions $Ee^{\langle\alpha, v_i(\xi)\rangle}$ are bounded from above uniformly in $\xi \in S$, and the other conditions of Sect. 7.3. The local rate function for this model is

$$L(\beta) \doteq$$
$$\inf\left[\int_S R(\nu(\cdot|\xi)\|\theta(\cdot|\xi))\mu(d\xi) + R\left(\gamma \| \mu \otimes p\right) : \int_{S\times\mathbb{R}^d} y\nu(dy|\xi)\mu(d\xi) = \beta\right],$$

where the infimum is over $\gamma \in \mathscr{P}(S\times S)$ such that $[\gamma]_1 = [\gamma]_2 = \mu$ and stochastic kernels $\nu(dw|\xi)$ on $\mathbb{R}^d$ given S.

Let $\mathbb{H}(p) = \inf_{\beta\in\mathbb{R}^d}[\langle p,\beta\rangle + L(\beta)]$. Then the correct notion of subsolution for this problem is again (14.24). There is an alternative characterization of $\mathbb{H}(p) = -H(-p)$ in terms of an eigenvector/eigenvalue problem. For $\alpha \in \mathbb{R}^d$ let $H(\alpha)$ and $r(\cdot;\alpha)$ solve

$$\int_S \int_{\mathbb{R}^d} e^{\langle\alpha,w\rangle}\theta(dw|\eta)r(\eta;\alpha)p(\xi,d\eta) = e^{H(\alpha)}r(\xi;\alpha), \quad \xi \in S,$$

where $r(\cdot;\alpha) : S \to [0,\infty)$ is the corresponding eigenfunction [116]. One can interpret $H(\alpha)$ in terms of a large time risk-sensitive (i.e., multiplicative) cost, in that

$$\frac{1}{k}\log E\left[e^{n\langle\alpha,X_k^n\rangle}\,\middle|\,X_0^n=0,\,\Xi_1=\xi\right]\to H(\alpha)\text{ as }k\to\infty,$$

and $r(\xi;\alpha)$ plays the role of the cost potential. One can in fact prove this limit using the weak convergence arguments of Chap. 6.

Given a subsolution $\bar{V}$, we generate processes $\{(Y_i^n,\Theta_{i+1}^n)\}$ by setting $Y_0^n=x_0$ and $\Theta_1^n=\xi$, using

$$e^{-\langle D\bar{V}(Y_i^n),w\rangle-H(-D\bar{V}(Y_i^n))}\theta(dw|\eta)\frac{r(\eta;-D\bar{V}(Y_i^n))}{r(\Theta_i^n;-D\bar{V}(Y_i^n))}p(\Theta_i^n,d\eta)$$

to identify the conditional distribution of w_i^n and Θ_{i+1}^n given Y_i^n and Θ_i^n, and then setting $Y_{i+1}^n=Y_i^n+w_i^n/n$. The estimator for the level crossing problem is then

$$1_{\left\{Y_{\bar{N}^n}^n\in(\times_{k=1}^d(-\infty,M_k))^c\right\}}\prod_{i=0}^{\bar{N}^n-1}e^{\langle D\bar{V}(Y_i^n),w_i^n\rangle+H(-D\bar{V}(Y_i^n))}\frac{r(\Theta_i^n;-D\bar{V}(Y_i^n))}{r(\Theta_{i+1}^n;-D\bar{V}(Y_i^n))},$$

where $\bar{N}^n\doteq\inf\left\{i:Y_i^n\in(\times_{k=1}^d(-\infty,M_k))^c\right\}$. As in the iid case the resulting algorithm is practical, in that $E\bar{N}^n<\infty$.

14.6 Notes

The references [6, 190, 224] present Monte Carlo methods in a general setting, and also discuss various aspects of rare event estimation. A nice overview of the use of Monte Carlo in the rare event setting specifically can be found in [223], which discusses other methods that are widely used, such as interacting particle methods (see also [51, 78]) and the cross entropy method for the design of importance sampling (see also [225]).

As noted previously the first paper to apply importance sampling in the rare event context is Siegmund [232]. The material of this chapter is mostly taken from [114, 116, 150], though the last section includes examples from other papers as well. The notion of Lyapunov inequality as used in [27] is closely related to that of subsolution in the context of importance sampling, and more information on this connection can be found in [28].

We consider only the light tailed cases (i.e., distributions for which moment generating functions are finite at least in a neighborhood of the origin). Problems with heavy tailed distributions are also important. A survey of developments up to 2012 on importance sampling for rare event estimation that includes the heavy tailed case is [26], and more recent developments for the heavy tailed case (including new classes of problem formulation not discussed previously) appear in [54].

For background on the Hamilton-Jacobi-Bellman equations used in this chapter we refer to [14, 134].