Adv. Appl. Prob. **49,** 981–1010 (2017)
*doi:*10.1017/apr.2017.31

MODERATE DEVIATIONS-BASED IMPORTANCE SAMPLING FOR STOCHASTIC RECURSIVE EQUATIONS

PAUL DUPUIS,* *Brown University*
DANE JOHNSON,** *University of North Carolina at Chapel Hill*

Abstract

Subsolutions to the Hamilton–Jacobi–Bellman equation associated with a moderate deviations approximation are used to design importance sampling changes of measure for stochastic recursive equations. Analogous to what has been done for large deviations subsolution-based importance sampling, these schemes are shown to be asymptotically optimal under the moderate deviations scaling. We present various implementations and numerical results to contrast their performance, and also discuss the circumstances under which a moderate deviation scaling might be appropriate.

Keywords: Moderate deviations; importance sampling; rare event; accelerated Monte Carlo

2010 Mathematics Subject Classification: Primary 65C05
Secondary 60F10

1. Introduction

Accurately estimating a probability using Monte Carlo can be computationally demanding if the event is sufficiently rare. What is meant by 'sufficiently' depends on the context and is related to the computing resources needed for the generation of a single sample. Importance sampling can reduce the number of samples needed by changing the distribution of the dynamics, and then adjusting the estimate using the likelihood ratio. However, one must be careful when selecting the alternative measure, and poor choices produce inaccurate and misleading results.

In this paper we investigate the effectiveness (including asymptotic optimality) of importance sampling schemes based on moderate deviations asymptotics for $\mathbb{R}^d$-valued discrete-time processes of the form

$$X_{i+1}^n = X_i^n + \frac{1}{n} b(X_i^n) + \frac{1}{n} u_i(X_i^n), \qquad X_0^n = x_0, \tag{1.1}$$

where $\{u_i(\cdot)\}_{i\in\mathbb{N}_0}$ are zero-mean random independent and identically distributed (i.i.d.) vector fields. We consider the continuous-time piecewise linear interpolation $\{X^n(t)\}_{0\le t\le T}$ with $X^n(i/n) = X_i^n$, which takes values in $C([0,T]\colon \mathbb{R}^d)$ (see (2.5) for the precise definition). Importance sampling can also be used to evaluate expected values, which we assume are expressed in the canonical form $\mathbb{E}e^{-F(X^n(T))}$, with F a lower-semicontinuous function from $\mathbb{R}^d$ to $\mathbb{R}$ that is bounded from below. (While in this paper we consider functionals that depend only

Received 11 February 2016; revision received 13 February 2017.
* Postal address: Division of Applied Mathematics, Brown University, Box F, 182 George St., Providence, RI 02912, USA.
** Postal address: Department of Statistics and Operations Research, University of Carolina at Chapel Hill, 318 Hanes Hall, Chapel Hill, NC 27599, USA. Email address: danedane@email.unc.edu.

on $X^n(T)$, various functionals of the trajectory can be handled by the same adaptations as those used for the large deviation problem as in, e.g. [10].)

The large deviation theory of the continuous-time linear interpolations $X^n \in C([0,T]: \mathbb{R}^d)$ has been extensively studied; see, e.g. [2], [5], [6], [15], and [20]–[23]. If $\mathbb{E}e^{-F(X^n(T))}$ is largely determined by rare events and if the true model can be embedded in a sequence satisfying a large deviations principle, then it is natural to expect that information contained in the large deviations rate function can be used to suggest effective changes of measure. (While it is customary to assume that F includes the large deviation scaling, in which case the problem is to estimate $\mathbb{E}e^{-nF(X^n(T))}$; for reasons made clear below, we assume that the problem of interest takes the form $\mathbb{E}e^{-F(X^n(T))}$ for a specific value $n = n^*$.) The first use of large deviation ideas in the context of rare events appeared in [19]. This inspired a number of related works, and for a discussion on some nonrigorous (and in some sense misleading) early approaches to the problem of algorithm design; see [16]. It turns out that a rigorous and systematic approach to design can be based on subsolutions to a Hamilton–Jacobi–Bellman (HJB) equation associated with the large deviation rate function [10] (see also the related notion of the Lyapunov inequality [3]). An important result of these analyses is the observation that, in general, *state feedback* is needed for changes of measure to perform well. This approach has been further developed in various ways and for many different process models; see, e.g. [9]–[13].

Since for complicated process models (e.g. those with state dependence) nonasymptotic bounds are either not available or too conservative, one typically uses some asymptotic performance measure to evaluate and compare difference schemes. In this case, one expects that the 'true' process corresponds to some value of the asymptotic parameter (e.g. n^*), and hopes that the asymptotic approximation is good enough that the scheme based on it is effective for the true model. With the introduction of a new form of asymptotic optimality (i.e. one based on a moderate deviation approximation) these issues should be revisited.

Let X^0 solve the ordinary differential equation (ODE) $\dot{X}^0 = b(X^0),\ X^0(0) = x_0$. An alternative asymptotic approximation is as follows. Let $a(n)$ be a sequence satisfying $a(n) \to 0$ and $a(n)\sqrt{n} \to \infty$, and consider the scaled or amplified difference between X^n and the noiseless version X^0:

$$Y^n \doteq a(n)\sqrt{n}(X^n - X^0).$$

It was shown in [8], under weaker conditions on the noise $u_i(\cdot)$ than are necessary when proving large deviation asymptotics for X^n, that Y^n satisfies a large deviation principle on $C([0,T]: \mathbb{R}^d)$ with a 'Gaussian'-type rate function. As is customary for this type of scaling, we refer to this as moderate deviations. In this paper we use the large deviation approximation for Y^n (the moderate deviation approximation for X^n) rather than the large deviation approximation for X^n to design importance sampling schemes. This will naturally limit the range of functionals F to those for which a moderate deviations approximation for X^n captures the distributional properties that are important in determining $\mathbb{E}e^{-F(X^n)}$. Intuitively, these should be functionals and events determined by trajectories falling somewhere between a functional central limit approximation and the full large deviation approximation (i.e. rare but not too rare). In exchange for this restriction on the class of functionals, one has to work with an approximation (i.e. a rate function) that is typically more tractable than the large deviation rate function.

One may ask if in this situation of 'rare but not too rare' events it is really worthwhile to use importance sampling, rather than standard Monte Carlo. That will depend on the difficulty and cost of generating samples. One can imagine at least three scenarios where it is worth the effort. One is when the generation of even a single trajectory of X^n is time consuming, e.g. if, say, X^n corresponds to a discrete-time approximation of a stochastic partial differential equation, or a

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

system with multiple (time or space) scales. The second is when a quantity such as $\mathbb{E}e^{-F(X^n)}$ must be computed many times, such as when the output is used as part of an optimization scheme. A third is when the event of interest is only marginally within the moderate deviations regime, but schemes based on the large deviation rate function are simply intractable, and the improvement may be sufficient to make the numerical problem feasible.

It is important to note that we do *not* propose the use of a diffusion approximation followed by importance sampling. While the diffusion approximation may lead to a simpler process model, it will also lead to bias in the estimates that cannot be removed. The use of moderate deviation for purposes of importance sampling makes use of a Gaussian-type approximation to determine the change of measure, but works with the true dynamics under a change of measure and thus gives an unbiased estimate. Of course, the moderate deviation approximation is not a panacea, and there are many situations (e.g. in the analysis of metastability issues) for which it is simply inadequate. However, for those situations described previously it may have merit.

For both the large deviation and moderate deviation scalings, the approach we consider for construction of importance sampling schemes is based on subsolutions to a nonlinear partial differential equation (PDE) associated with the rate function (see [10] for the case of large deviations). In general, one expects the moderate deviations rate function to involve a linear approximation of the dynamics and a quadratic approximation of the costs found in the large deviations rate function, centered around the law of large numbers limit. This corresponds to a 'local Gaussian' approximation, and the PDEs are those associated with the famous 'linear-quadratic regulator,' whose solution can be constructed in terms of the solution to an appropriate Riccati equation [1].

In analogy with the proof of the moderate deviations principle found in [8], proving the asymptotic properties of importance sampling schemes in the moderate deviations setting involves challenges not found in the large deviations setting. The main difficulties are due to issues of tightness. In analyzing the large deviation properties of a process such as $\{X^n\}$, one usually assumes that the noise has a finite-moment generating function everywhere, which implies a tightness result that is used in the proof. With the scaling of moderate deviations this assumption is not sufficient for the analogous tightness, and alternative methods are needed.

The paper is organized as follows. In Section 2 we provide details on the types of problems considered, summarize importance sampling based on moderate deviations subsolutions, and state the asymptotic performance bound for these schemes that will be proved later on. In Section 3 we describe several different ways importance sampling schemes can be based on a particular subsolution, all of which result in the same asymptotic performance bound. In Section 4 we outline a flexible approach to constructing subsolutions with optimal decay rates in terms of solutions to the linear-quadratic regulator with affine terminal costs. Numerical examples are provided in Section 5, and Section 6 contains the proof of the result stated in Section 2 on asymptotic performance.

2. Preliminaries

We consider the processes model

$$X_{i+1}^n = X_i^n + \frac{1}{n} b(X_i^n) + \frac{1}{n} u_i(X_i^n), \qquad X_0^n = x_0, \tag{2.1}$$

where the $\{u_i(\cdot)\}_{i\in\mathbb{N}_0}$ are zero-mean i.i.d. random vector fields with distribution given by the stochastic kernel $\mu(\cdot \mid x)$: $\mathbb{P}\{u_i(x) \in A\} = \mu(A \mid x)$ for $A \in \mathcal{B}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$. We use the

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

following assumptions on $\mu(\cdot \mid x)$ and $b(x)$. Define

$$H_c(x,\alpha) \doteq \log\left(\int_{\mathbb{R}^d} \mathrm{e}^{\langle u,\alpha\rangle}\mu(\mathrm{d}u \mid x)\right) \quad \text{for } \alpha \in \mathbb{R}^d. \tag{2.2}$$

The subscript c reflects the fact that this log moment-generating function uses the centered distribution $\mu(\cdot \mid x)$, rather than the usual $H(x,\alpha) = H_c(x,\alpha) + \langle \alpha, b(x)\rangle$.

Condition 2.1. *We have the following conditions:*

- *there exist $\lambda > 0$ and $K_{\mathrm{mgf}} < \infty$ such that*

$$\sup_{x\in\mathbb{R}^d}\sup_{\|\alpha\|\le\lambda} H_c(x,\alpha) \le K_{\mathrm{mgf}}; \tag{2.3}$$

- *$x \to \mu(\cdot \mid x)$ is continuous with respect to the topology of weak convergence;*
- *$b(x)$ is continuously differentiable, and the norms of both $b(x)$ and its derivative are uniformly bounded by some constant $K_b < \infty$.*

Throughout this paper we let $\|\alpha\|_A^2 \doteq \langle \alpha, A\alpha\rangle$ for any $\alpha \in \mathbb{R}^d$ and symmetric, nonnegative definite matrix A. Define

$$A_{ij}(x) \doteq \int_{\mathbb{R}^d} u_i u_j \mu(\mathrm{d}u \mid x), \tag{2.4}$$

and note that the weak continuity of $\mu(\cdot \mid x)$ with respect to x and (2.3) ensure that $A(x)$ is continuous in x and its norm is uniformly bounded by some constant K_A. Note also that

$$\frac{\partial H_c(x,0)}{\partial \alpha_i} = \int_{\mathbb{R}^d} u_i \mu(\mathrm{d}u \mid x) = 0 \quad \text{and} \quad \frac{\partial^2 H_c(x,0)}{\partial\alpha_i\partial\alpha_j} = \int_{\mathbb{R}^d} u_i u_j \mu(\mathrm{d}u \mid x) = A_{ij}(x)$$

for all $i, j \in \{1,\dots,d\}$ and $x \in \mathbb{R}^d$, and that $A(x)$ is nonnegative definite and symmetric.

Definition 2.1. For a symmetric, nonnegative definite matrix A we can write $A = Q\Lambda Q^\top$, where Q is an orthogonal matrix whose columns are the eigenvectors of A and Λ is the diagonal matrix consisting of the nonnegative eigenvalues of A. Throughout the paper we use the convention that $A^{1/2} = Q\Lambda^{1/2}Q^\top$, where $\Lambda^{1/2}$ is the diagonal matrix consisting of the positive square roots of the nonnegative eigenvalues of A. In addition, let $\|\alpha\|_A^2 = \langle Q\alpha, \Lambda Q\alpha\rangle$, and let $\|\alpha\|_{A^{-1}}^2 = \langle Q\alpha, \Lambda^{-1}Q\alpha\rangle$, where Λ^{-1} is a diagonal matrix consisting of the inverse of the eigenvalues for the positive eigenvalues and ∞ for the zero eigenvalues. Consequently, $\langle Q\alpha, \Lambda^{-1}Q\alpha\rangle$ has a well-defined and finite value for α in the span of the eigenvalues corresponding to positive eigenvalues, and is equal to ∞ for α outside of that linear span.

Define the continuous-time linear interpolation of X_i^n by $X^n(i/n) = X_i^n$ for $i \in \mathbb{N}_0$ and

$$X^n(t) = (i+1-nt)X_i^n + (nt-i)X_{i+1}^n \quad \text{for } t \in (i/n, i/n+1/n). \tag{2.5}$$

In addition, define

$$X_{i+1}^{n,0} = X_i^{n,0} + \frac{1}{n}b(X_i^{n,0}), \qquad X_0^{n,0} = x_0$$

and let $X^{n,0}(t)$ be the analogously defined continuous-time linear interpolation. Then as is well known, $X^{n,0} \to X^0$ in $C([0,T]\colon \mathbb{R}^d)$, where

$$X^0(t) = \int_0^t b(X^0(s))\,\mathrm{d}s + x_0.$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

Since $\mathbb{E}u_i(x) = 0$ for all $x \in \mathbb{R}^d$, we know that $X^n \to X^0$ in $C([0, T]: \mathbb{R}^d)$ in probability [17, Theorem 3.3.1]. Under stronger assumptions, including the assumption that

$$\sup_{x\in\mathbb{R}^d} \sup_{\alpha\in\mathbb{R}^d} H_c(x, \alpha) < \infty,$$

it has been shown that X^n satisfies the large deviation principle on $C([0, T]: \mathbb{R}^d)$ with sequence $r(n) = 1/n$ and rate function

$$I_L(\phi) \doteq \inf\left\{\int_0^T L_c(\phi(s), u(s))\,\mathrm{d}s : \phi(t) = x_0 + \int_0^t b(\phi(s))\,\mathrm{d}s + \int_0^t u(s)\,\mathrm{d}s,\ t \in [0, T]\right\},$$

where

$$L_c(x, \beta) \doteq \sup_{\alpha\in\mathbb{R}^d} \{\langle\alpha, \beta\rangle - H_c(x, \alpha)\}$$

is the Legendre transform of $H_c(x, \alpha)$ [2], [5], [6], [15], [20]–[23]. When stated in terms of the equivalent Laplace principle [6, Section 1.2], this means that for any bounded and continuous function M on $C([0, T]: \mathbb{R}^d)$,

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{E}\mathrm{e}^{-nM(X^n)} = \inf_{\phi\in C([0,T]:\,\mathbb{R}^d)} \{M(\phi) + I_L(\phi)\}.$$

Assume that $a(n)$ satisfies

$$a(n) \to 0 \quad \text{and} \quad a(n)\sqrt{n} \to \infty \quad \text{as } n \to \infty.$$

We define the rescaled difference

$$Y^n(t) \doteq a(n)\sqrt{n}(X^n(t) - X^{n,0}(t)). \tag{2.6}$$

Note that

$$Y^n_{i+1} = Y^n_i + \frac{a(n)}{\sqrt{n}}(b(X^n_i) - b(X^{n,0}_i)) + \frac{a(n)}{\sqrt{n}}u_i(X^n_i), \qquad Y^n_0 = 0.$$

Let $D_x b(x)$ be the matrix whose kth row is $(\partial b_k(x)/\partial x_1, \ldots, \partial b_k(x)/\partial x_d)$. It was shown in [8] that Y^n satisfies the large deviations principle on $C([0, T]: \mathbb{R}^d)$ with sequence $r(n) = a(n)^2$ and rate function

$$I_M(\phi) \doteq \inf\left\{\frac{1}{2}\int_0^T \|u(t)\|^2\,\mathrm{d}t : \right.$$
$$\left. \phi(t) = \int_0^t D_x b(X^0(s))\phi(s)\,\mathrm{d}s + \int_0^t A^{1/2}(X^0(s))u(s)\,\mathrm{d}s,\ t \in [0, T]\right\},$$

where $A^{1/2}(x)$ is specified in Definition 2.1. Due to the scaling, this is typically referred to as a moderate deviations principle. For some lower-semicontinuous function G that is bounded below, let

$$V(t, y) \doteq \inf\left\{\frac{1}{2}\int_t^T \|u(s)\|^2\,\mathrm{d}s + G(\phi(T)) : \right.$$
$$\phi(r) = y + \int_t^r D_x b(X^0(s))\phi(s)\,\mathrm{d}s$$
$$\left. + \int_t^r A^{1/2}(X^0(s))u(s)\,\mathrm{d}s,\ r \in [t, T]\right\} \tag{2.7}$$

for all $(t, y) \in [0, T] \times \mathbb{R}^d$.

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

The goal of the numerical procedures to be developed is the approximation of a functional of the form $\mathbb{E}e^{-F(X^n(T))}$ for a specific value $n = n^*$. To do this, we must rewrite the functional in terms of Y^n and with the moderate deviations scaling. Since the original problem was of interest for $n = n^*$, $\mathbb{E}e^{-F(X^n(T))}$ should be replaced by $\mathbb{E}e^{-G(Y^n(T))/a(n)^2}$, where G is defined by

$$F(\cdot) = \frac{G(a(n^*)\sqrt{n^*}[\cdot - X^0(T)])}{a(n^*)^2}. \tag{2.8}$$

Thus, G implicitly depends on n^*. However, it is important to note that an analogous issue arises when applying large deviations approximations to the design of importance sampling, since $\mathbb{E}e^{-F(X^n(T))}$ would have to be replaced by $\mathbb{E}e^{-n\bar{F}(X^n(T))}$ with $\bar{F}$ defined by $F(\cdot) = n^*\bar{F}(\cdot)$. Fortunately, it can be shown that the choice of embedding, meaning the choice of $\{a(n)\}$ and G such that (2.8) is satisfied, does not affect in any way the importance sampling schemes that result from a moderate deviations approximation. The details are omitted for brevity because the notation is cumbersome and the proof is straightforward [18]. Consequently, the choice of embedding can be based on convenience, and with this in mind we use the convention $a(n^*)\sqrt{n^*} = 1$ for the numerical examples of Section 6 to make the comparison to large deviations-based schemes more straightforward.

We next give the construction of an importance sampling scheme and introduce notation that will be used throughout the paper. In the construction, various processes are defined that depend on a control. The explicit dependence on the control is suppressed in the notation, save for the appearance of an overbar or related notation. Thus, whenever objects using an overbar are present, the reader is warned that a control such as η introduced below has been used in their construction. The particular control used will be spelled out. Let $\lceil a \rceil$ denote the smallest integer greater than a.

Construction 2.1. Given a probability measure $\eta \in \mathcal{P}((\mathbb{R}^d)^{\lceil nT \rceil})$, let $(\bar{u}_0^n, \ldots, \bar{u}_{\lceil nT \rceil - 1}^n)$ be random variables with distribution η. Define

$$\bar{X}_{i+1}^n = \bar{X}_i^n + \frac{1}{n}b(\bar{X}_i^n) + \frac{1}{n}\bar{u}_i^n, \qquad \bar{X}_0^n = x_0.$$

Define also the continuous-time linear interpolations $\bar{X}^n(t)$ as in (2.5) and the scaled difference

$$\bar{Y}^n(t) \doteq a(n)\sqrt{n}(\bar{X}^n(t) - X^{n,0}(t)).$$

Let $\eta_i(\mathrm{d}u \mid \bar{u}_0^n, \ldots, \bar{u}_{i-1}^n)$ denote the conditional distribution of $\bar{u}_i^n$ given $(\bar{u}_0^n, \ldots, \bar{u}_{i-1}^n)$ with the understanding that we generally suppress the dependence on $(\bar{u}_0^n, \ldots, \bar{u}_{i-1}^n)$ in the notation. Define the conditional means

$$w^n(t) \doteq \int_{\mathbb{R}^d} u\eta_i(\mathrm{d}u) \quad \text{for } t \in [i/n, i/n + 1/n),$$

the amplified conditional means $\bar{w}^n(t) \doteq a(n)\sqrt{n}w^n(t)$, and the associated random measures on $\mathbb{R}^d \times [0, 1]$:

$$\bar{\varsigma}^n(\mathrm{d}w \times \mathrm{d}t) \doteq \delta_{\bar{w}^n(t)}(\mathrm{d}w)\,\mathrm{d}t = \delta_{a(n)\sqrt{n}w^n(t)}(\mathrm{d}w)\,\mathrm{d}t.$$

Construction 2.1 involves the use of new driving noises. With respect to the original system, η corresponds to the measure on $(\mathbb{R}^d)^{\lceil nT \rceil}$ given by

$$\mu^n(u_0, \ldots, u_{n-1}) = \prod_{i=0}^{n-1} \mu(\mathrm{d}u_i \mid x_i^n),$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

where x_i^n is the position given recursively by

$$x_{i+1}^n = x_i^n + \frac{1}{n} b(x_i^n) + \frac{1}{n} u_i, \qquad x_0^n = x_0.$$

Let $\mathrm{d}\mu^n/\mathrm{d}\eta^n$ be the Radon–Nikodym derivative of μ^n with respect to η^n and let $\bar{Y}^n$ be as in Construction 2.1 for η^n. We consider the unbiased estimate of $\mathbb{E}\exp\{-G(Y^n(T))/a(n)^2\}$ given by averaging a number (say K) of independent copies of $\exp\{-G(\bar{Y}^n(T))/a(n)^2\}(\mathrm{d}\mu^n/\mathrm{d}\eta^n)$. Thus, the variance of the average is proportional to the variance of a single sample, and since unbiasedness implies that minimizing the variance is equivalent to minimizing the second moment, we can characterize the performance by the magnitude of the second moment. To study asymptotic performance, we consider the decay rate of the second moment under a moderate deviations scaling, i.e. the limit as $n \to \infty$ of

$$-a(n)^2 \log \mathbb{E}\left[\left(\exp\left\{-\frac{1}{a(n)^2} G(\bar{Y}^n(T))\right\}\frac{\mathrm{d}\mu^n}{\mathrm{d}\eta^n}\right)^2\right].$$

As usual, Jensen's inequality gives an *a priori* bound on the best possible rate:

$$\begin{aligned}
&\limsup_{n\to\infty} -a(n)^2 \log \mathbb{E}\left[\left(\exp\left\{-\frac{1}{a(n)^2} G(\bar{Y}^n(T))\right\}\frac{\mathrm{d}\mu^n}{\mathrm{d}\eta^n}\right)^2\right] \\
&\quad \le \limsup_{n\to\infty} -2a(n)^2 \log \mathbb{E}\left[\exp\left\{-\frac{1}{a(n)^2} G(\bar{Y}^n(T))\right\}\frac{\mathrm{d}\mu^n}{\mathrm{d}\eta^n}\right] \\
&\quad \le 2\limsup_{n\to\infty} -a(n)^2 \log \mathbb{E}\left[\exp\left\{-\frac{1}{a(n)^2} G(Y^n(T))\right\}\right] \\
&\quad \le 2V(0,0),
\end{aligned}$$

where the last inequality comes from the large deviations principle for the moderate deviations scaling [8, Theorem 2.3] and the definition of $V(t, y)$ in (2.7). With Df indicating the gradient of f, we note that V is a weak sense solution to the HJB PDE

$$V_t(t, y) = \tfrac{1}{2}\|D_y V(t, y)\|^2_{A(X^0(t))} - \langle D_y V(t, y), D_x b(X^0(t))y\rangle \tag{2.9}$$

and

$$V(T, y) = G(y). \tag{2.10}$$

By a smooth subsolution to (2.9) and (2.10), we mean a function $W\colon \mathbb{R}^d \times [0, 1] \to \mathbb{R}$ that is C^1 and satisfies

$$W_t(t, y) \ge \tfrac{1}{2}\|D_y W(t, y)\|^2_{A(X^0(t))} - \langle D_y W(t, y), D_x b(X^0(t))y\rangle \tag{2.11}$$

and

$$W(T, y) \le G(y). \tag{2.12}$$

In the next section we generalize the notion of subsolutions and describe several different implementations of importance sampling schemes based on a subsolution, but the general idea is always the following. As in the previous papers on the use of subsolutions (e.g. [10]), changes of measure are suggested by each subsolution through the associated HJB equation, and, in particular, through feedback controls defined through the duality formula

$$\tfrac{1}{2}\|p\|^2_{A(X^0(t))} = \inf_u \left\{\langle p, A^{1/2}(X^0(t))u\rangle + \tfrac{1}{2}\|u\|^2\right\}$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

that relates (2.7) and (2.9) via dynamic programming. As we will discuss later on, a verification argument shows that if $\{\eta^n\}$ is a sequence of importance sampling changes of measure based on a subsolution W satisfying (2.11) and (2.12), then the asymptotic decay rate of the second moment of the corresponding importance sampling estimate satisfies

$$\liminf_{n\to\infty} -a(n)^2 \log \mathbb{E}\left[\left(\exp\left\{-\frac{1}{a(n)^2}G(\bar{Y}^n)\right\}\frac{\mathrm{d}\mu^n}{\mathrm{d}\eta^n}\right)^2\right] \geq V(0,0) + W(0,0),$$

where V is given by (2.7). In general, $W(0,0) \leq V(0,0)$, and an optimal rate of decay is obtained if $W(0,0) = V(0,0)$. In the next section we present a generalized definition of subsolutions, as well as two importance sampling schemes associated with each such subsolution.

3. Subsolutions and importance sampling

We first define the generalized class of subsolutions to be used in constructing importance sampling schemes. The approach we use involves building subsolutions out of relatively simple component functions which satisfy the HJB subsolution dynamics condition (2.11), but not necessarily the terminal condition (2.12). The change of measure based on the subsolution is a weighted combination of the changes of measure suggested by these component functions (i.e. a mixture), and the generalized class of subsolutions we define specifies the components and weights along with the subsolution itself. Throughout the following, if A is a closed subset of some Euclidean space then $f \in C^1(A\colon \mathbb{R})$ means that there is an open neighborhood of A on which f is continuously differentiable.

The following collection of functions are the components used in building the subsolution. Let

$$\mathcal{S} \doteq \left\{W \in C^1([0,T]\times\mathbb{R}^d\colon \mathbb{R})\colon W \text{ satisfies (2.11) and } \sup_{(t,y)\in[0,T]\times\mathbb{R}^d} \|DW\| < \infty\right\}$$

and let

$$\mathcal{W}([0,T]\times\mathbb{R}^d) \tag{3.1}$$
$$\doteq \left\{\rho \in C([0,T]\times\mathbb{R}^d\colon [0,1]^K), \sum_{k=1}^K \rho_k(t,y) = 1 \text{ for all } (t,y)\in[0,T]\times\mathbb{R}^d\right\}$$

be the class of continuous weightings. Finally, we define the subsolutions along with their component functions and their weightings:

$$\mathcal{MS} \doteq \left\{(U,\rho,\{W^k\}_{k=1}^K)\colon U\in\mathcal{S},\ \rho\in\mathcal{W}([0,T]\times\mathbb{R}^d),\ \{W^k\}_{k=1}^K\in\mathcal{S},\right. \tag{3.2}$$
$$\left. D_yU(t,y) = \sum_{k=1}^K \rho_k(t,y)D_yW^k(t,y) \text{ and } U_t(t,y) = \sum_{k=1}^K \rho_k(t,y)W_t^k(t,y)\right\}.$$

It follows from Jensen's inequality that if $(U,\rho,\{W^k\}_{k=1}^K)\in\mathcal{MS}$ then (2.11) holds with W replaced by U, and, thus, $(U,1,U)\in\mathcal{MS}$. In the following subsections we describe two methods for implementing importance sampling schemes based on subsolutions. Although the U component of any element of $\mathcal{MS}$ always satisfies (2.11), to be used for importance sampling, U must also satisfy the terminal condition inequality (2.12).

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

3.1. Standard importance sampling

We first describe the importance sampling schemes analogous to those of [10] that can be based on a given subsolution from $\mathcal{MS}$.

Construction 3.1. Given $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{S}$ with U satisfying (2.12), define the corresponding 'standard' importance sampling as follows. The distribution of $\bar{u}_i^n$, given $\bar{u}_j^n$, $j = 0, \dots, i-1$, is defined by

$$\gamma_i^n(\mathrm{d}u \mid \bar{X}_i^n) \doteq \sum_{k=1}^K \rho_k\left(\bar{Y}_i^n, \frac{i}{n}\right) \times \exp\left\{\left\langle u, -\frac{1}{a(n)\sqrt{n}} D_y W^k\left(\frac{i}{n}, \bar{Y}_i^n\right)\right\rangle - H_c\left(\bar{X}_i^n, -\frac{1}{a(n)\sqrt{n}} D_y W^k\left(\frac{i}{n}, \bar{Y}_i^n\right)\right)\right\} \mu(\mathrm{d}u \mid \bar{X}_i^n), \tag{3.3}$$

where the controlled processes $(\bar{X}^n, \bar{Y}^n)$ are defined in terms of the $\bar{u}_i^n$ as in Construction 2.1. The unbiased estimator for this change of measure is then defined by

$$r^n \doteq \exp\left\{-\frac{1}{a(n)^2} G(\bar{Y}^n(T))\right\} \prod_{i=0}^{\lfloor Tn \rfloor} \left(\frac{\mathrm{d}\gamma_i^n(\cdot \mid \bar{X}_i^n)}{\mathrm{d}\mu(\cdot \mid \bar{X}_i^n)}(\bar{u}_i^n)\right)^{-1}.$$

We refer to a sampling distribution of the form $\exp\{\langle u, \alpha\rangle - H_c(x, \alpha)\}\mu(\mathrm{d}u \mid x)$ as an *exponential tilt*, and to α as the *tilt parameter*. The importance sampling change of measure in Construction 3.1 is a convex combination of exponential tilt changes of measure using the weights given by ρ. This can be thought of as a mixture distribution. We refer to this as the 'randomized' implementation of the subsolution U since we randomly choose which component function determines the change of measure based on the probabilities given by ρ. Recall that $(U, 1, U) \in \mathcal{MS}$. The implementation of $(U, 1, U) \in \mathcal{MS}$ under Construction 3.1 is a single exponential tilt which is the average of the tilts given by the $\{W^k\}$ according to the weights given by ρ. We call this the 'deterministic' implementation of the subsolution U. Thus, two distinct implementations correspond to each subsolution U when $K > 1$. In the numerical section of the paper we use only the deterministic implementation. However, the statement and proof of performance of the schemes is given for the (more general) randomized implementation. The randomized schemes can be particularly useful for some classes of problem, e.g. problems with multiscale aspects [10, Remark 7.1].

3.2. Corrected importance sampling

The HJB equation (2.9) and inequality (2.11) directly relate to a small-noise Gaussian Markov processes with time-dependent coefficients. They are naturally suggested for the system (2.6) due to the moderate deviations Gaussian-type rate function. However, the moderate deviation approximation is an asymptotic result, and it is natural to ask if the impact of prelimit 'errors' in the approximation can be reduced in some obvious way.

Note that exactly the same issue could arise for importance sampling based on large deviation asymptotics. For example, if b is Lipschitz continuous and $\tilde{b}^\varepsilon$ is uniformly bounded, then the systems

$$\mathrm{d}Z^\varepsilon = b(Z^\varepsilon)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}W, \qquad Z^\varepsilon(0) = z$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

and

$$\mathrm{d}\tilde{Z}^\varepsilon = b(\tilde{Z}^\varepsilon)\,\mathrm{d}t + \varepsilon\tilde{b}^\varepsilon(\tilde{Z}^\varepsilon)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}W, \qquad \tilde{Z}^\varepsilon(0) = z, \tag{3.4}$$

satisfy a large deviation principle with the same rate function, and this rate function does not depend on $\tilde{b}^\varepsilon$. In designing an importance sampling scheme for $\tilde{Z}^\varepsilon$ based on the large deviation approximation, one could accommodate the $\tilde{b}^\varepsilon$ term in ways analogous to those we will use to treat such 'errors' for the moderate deviation problem. However, systems with such perturbations do not commonly appear in prior work on the design and analysis of importance sampling using large deviation ideas. In contrast, the analogous issue is ubiquitous in the moderate deviations setting.

For reasons discussed at length in [9] and [10], the feedback provided through the subsolution (see (3.3)) is needed to control the variance when using importance sampling in all but the simplest of situations. The feedback occurs because a dynamic programming analysis applied to a control problem associated with the moderate deviation approximation shows that the tilt parameter should be obtained from the optimization problem

$$\inf_\alpha \big\{ \langle D_y U(t,y), A(X^0(t))\alpha\rangle + \tfrac{1}{2}\|\alpha\|^2_{A(X^0(t))} \big\},$$

i.e. $\alpha(t,y) = -D_y U(t,y)$.

Note that this tilt parameter can be equivalently characterized as the unique tilt parameter leading to the mean velocity $-A(X^0(t))D_yU(t,y)$, in that if $\varsigma(\mathrm{d}u \mid x)$ is Gaussian with mean $D_x b(x)y$ and variance $A(x)$, then

$$\int_{\mathbb{R}^d} u \exp\big\{\langle u, \alpha(t,y)\rangle - \tfrac{1}{2}\langle \alpha(t,y), A(x)\alpha(t,y)\rangle - \langle D_x b(x)y, \alpha(t,y)\rangle\big\}\varsigma(\mathrm{d}u \mid x)$$
$$= -A(x)D_yU(t,y).$$

When this tilt parameter is applied in the prelimit, a scaling $1/a(n)\sqrt{n}$ is used because the driving noise appears in the equation for $\bar{Y}$ in the form $a(n)\sqrt{n}u$. Since the true model need not be Gaussian, the correct change of measure for the prelimit is thus given by

$$\exp\left\{\left\langle u, -\frac{1}{a(n)\sqrt{n}} D_y U\left(\frac{i}{n}, y\right)\right\rangle - H_c\left(\bar{X}^n_i, -\frac{1}{a(n)\sqrt{n}} D_y U\left(\frac{i}{n}, y\right)\right)\right\}\mu(\mathrm{d}u \mid \bar{X}^n_i), \tag{3.5}$$

where the term involving $D_x b$ no longer appears because we use H_c rather than H. Differentiating with respect to α in the definition (2.2) of $H_c(x,\alpha)$ yields

$$\int_{\mathbb{R}^d} u \exp\{\langle u, \alpha\rangle - H_c(x,\alpha)\}\mu(\mathrm{d}u \mid x) = D_\alpha H_c(x,\alpha),$$

and so the mean of the driving noise $a(n)\sqrt{n}u$ under (3.5) is, in fact,

$$a(n)\sqrt{n} D_\alpha H_c\left(\bar{X}^n_i, -\frac{1}{a(n)\sqrt{n}} D_y U\left(\frac{i}{n}, y\right)\right)$$

rather than $-A(\bar{X}^n_i)D_yU(i/n, y)$. This 'error' is analogous to the $\tilde{b}^\varepsilon$ in (3.4). There is also a difference between the discrete-time 'drift' obtained under the standard construction, which is

$$a(n)\sqrt{n}\left[b\left(X^{n,0}_i + \frac{1}{a(n)\sqrt{n}}\bar{Y}^n_i\right) - b(X^{n,0}_i)\right],$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

and the corresponding drift one would obtain with the limit 'linear Gaussian' approximation, which would be $\langle D_x b(X^0(i/n)), \bar{Y}_i^n \rangle$. Due to these errors, the controlled process at the prelimit follows a path whose mean behavior deviates from what would have been obtained if the controls had been applied to a true Gaussian and linear system model. There is a mismatch between what is suggested by the subsolution and what is actually happening with the true dynamics.

The errors just described disappear in the limit, and indeed as expected one obtains asymptotic optimality. However, there is an obvious opportunity for improvement at the prelimit level, which amounts to correcting for these errors. While a detailed analysis, as in [7], might suggest more elaborate ways to improve performance, the 'corrected' scheme introduced below provides better performance than the standard scheme with essentially no additional analysis or assumptions. The difference between the two is not always substantial, as in some cases where G is a smooth functional. However, when G is not continuous and, in particular, in the problem of estimating escape probabilities where the discontinuity takes the relatively severe form $G(y) = \infty \mathbf{1}_{B^c}(y)$, computational accuracy can be significantly improved by this correction.

Let

$$\mathcal{A}_{y,i}^n = \left\{ \xi \in \mathbb{R}^d : H_c\left(X_i^{n,0} + \frac{y}{a(n)\sqrt{n}}, \xi \right) < \infty \right\}$$

be the set of feasible tilt parameters at a particular position and time.

Definition 3.1. Given $W \in \mathcal{S}$ and some positive constant $K_T \in (0, \infty)$, define the state- and time-dependent tilt parameter

$$\xi_i^{n,W,T}(y) \doteq \arg\min \left\{ \left\| D_\alpha H\left(X_i^{n,0} + \frac{y}{a(n)\sqrt{n}}, \alpha \right) + \frac{1}{a(n)\sqrt{n}} A\left(X^0\left(\frac{i}{n}\right)\right) D_y W\left(\frac{i}{n}, y\right) + \theta_i^n(y) \right\| \right\},$$

where the minimization is over $\alpha \in \mathcal{A}_{y,i}^n$ such that $\|\alpha + D_y W(i/n, y)/a(n)\sqrt{n}\| \le K_T/a(n)^2 n$, and

$$\theta_i^n(y) \doteq b\left(X_i^{n,0} + \frac{y}{a(n)\sqrt{n}} \right) - b(X_i^{n,0}) - \frac{1}{a(n)\sqrt{n}} D_x b\left(X^0\left(\frac{i}{n}\right)\right) y.$$

The corresponding exponential tilts will be used in what we refer to as the 'corrected' subsolution-based importance sampling scheme. The purpose of restricting the tilt parameters to the set $\mathcal{A}_{y,i}^n$ is simply to ensure they can be implemented. Forcing them to be close to $D_y W^k(i/n, y)/[a(n)\sqrt{n}]$ as $n \to \infty$ guarantees the same asymptotic performance as the standard scheme. The preasymptotic role of the constant K_T is to prevent the corrected scheme from choosing extremely large exponential tilts, and, in general, if very large exponential tilts are necessary to obtain the desired conditional mean then the moderate deviation asymptotic approximation is probably not appropriate for the problem of interest. The corrected approach presented here is a simple (and effective) way to account for preasymptotic inaccuracies in the moderate deviations approximation by directly adjusting the dynamics.

Construction 3.2. Given $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{S}$ with U satisfying (2.12) and $K_T \in (0, \infty)$, define the corresponding corrected importance sampling as follows. The distribution of $\bar{u}_i^n$, given $\bar{u}_j^n$, $j = 0, \dots, i-1$, is defined by

$$\gamma_i^n(\mathrm{d}u \mid \bar{X}_i^n) = \sum_{k=1}^K \rho_k\left(\frac{i}{n}, \bar{Y}_i^n\right) \exp\{\langle u, \xi_i^{n,W^k,T}(\bar{Y}_i^n)\rangle - H_c(\bar{X}_i^n, \xi_i^{n,W^k,T}(\bar{Y}_i^n))\} \mu(\mathrm{d}u \mid \bar{X}_i^n),$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

where $\xi_i^{n,W,T}(y)$ is given by Definition 3.1 and the controlled processes $(\bar{X}^n, \bar{Y}^n)$ are defined in terms of the $\bar{u}_i^n$ as in Construction 2.1. The unbiased estimator for this change of measure is then defined by

$$r^n \doteq \exp\left\{-\frac{1}{a(n)^2}G(\bar{Y}^n(T))\right\}\prod_{i=0}^{\lfloor Tn\rfloor}\left(\frac{\mathrm{d}\gamma_i^n(\cdot \mid \bar{X}_i^n)}{\mathrm{d}\mu(\cdot \mid \bar{X}_i^n)}(\bar{u}_i^n)\right)^{-1}.$$

In Construction 3.2 we present the randomized implementation, which as noted previously includes as a special case the deterministic implementation. The proof of the following result appears in Section 6.

Theorem 3.1. *Assume that G is bounded from below and is lower semicontinuous. For any $n \in \mathbb{N}$, let r^n be the unbiased estimate of $\mathbb{E}\exp\{-G(Y^n(T))/a(n)^2\}$ based on the subsolution $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{WS}$, with U satisfying the terminal inequality (2.12) as in either Construction 3.1 or Construction 3.2 with the deterministic or randomized implementation. Then*

$$\liminf_{n\to\infty} -a(n)^2 \log \mathbb{E}[(r^n)^2] \geq U(0,0) + V(0,0),$$

where V is given by (2.7).

4. Constructing subsolutions

According to Theorem 3.1, the lower bound on the decay rate of the second moment of the importance sampling estimator based on a subsolution U depends on $U(0,0)$ and, therefore, subsolutions with larger values at the origin are (at least asymptotically) preferable. Our approach to the construction of subsolutions is to identify a relatively simple class of elements satisfying (2.9) (e.g. solutions with an affine terminal condition), and then construct $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ with W^k selected from this class and U satisfying (2.12) by using the minimum (and also possibly the maximum) mollification described below in Definition 4.1. Although this is similar to prior work in the large deviations setting, such as [10], it is typically easier to find subsolutions to the moderate deviations PDE given by (2.9) and (2.10).

Let $\Omega(s,t)$ be the matrix-valued solution to

$$\frac{\mathrm{d}}{\mathrm{d}t}\exp\{\Omega(s,t)\} = D_x b(X^0(t))\exp\{\Omega(s,t)\} \quad \text{for } s < t \text{ and } \Omega(s,s) = 0.$$

Given a matrix or vector M, let $M^\top$ denote its transpose. It is easily verified that

$$\frac{\mathrm{d}}{\mathrm{d}t}\exp\{\Omega(t,T)^\top\} = -D_x b(X^0(t))^\top \exp\{\Omega(t,T)^\top\}. \tag{4.1}$$

There is a large literature concerned with the numerical approximation of $\Omega(s,t)$, such as [4].

Given arbitrary $c \in \mathbb{R}$ and $\xi \in \mathbb{R}^d$, let

$$W^{c,\xi}(t,y) \doteq \langle \mathrm{e}^{\Omega(t,T)^\top}\xi, y\rangle - \frac{1}{2}\left\langle \xi, \left(\int_t^T \mathrm{e}^{\Omega(s,T)}A(X^0(s))\mathrm{e}^{\Omega(s,T)^\top}\,\mathrm{d}s\right)\xi\right\rangle + c. \tag{4.2}$$

The following theorem asserts that the functions given in (4.2) are classical sense solutions with an affine terminal condition.

Theorem 4.1. *It holds that $W^{c,\xi}(t,y)$ defined in (4.2) is a solution to (2.9) with the terminal condition $W^{c,\xi}(T,y) = \langle \xi, y\rangle + c$.*

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

Proof. Using $\Omega(T, T) = 0$,

$$W^{c,\xi}(T, y) = \langle \exp\{\Omega(T, T)^\top\}\xi, y\rangle + c = \langle \xi, y\rangle + c.$$

It remains to show that

$$W_t^{c,\xi}(t, y) = -\langle D_y W^{c,\xi}(t, y), D_x b(X^0(t))y\rangle + \tfrac{1}{2}\|D_y W^{c,\xi}(t, y)\|^2_{A(X^0(t))}.$$

By (4.2), $D_y W^{c,\xi}(t, y) = \mathrm{e}^{\Omega(t,T)^\top}\xi$ and

$$W_t^{c,\xi}(t, y) = \left\langle y, \frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{\Omega(t,T)^\top}\xi\right\rangle + \frac{1}{2}\langle \xi, \mathrm{e}^{\Omega(t,T)}A(X^0(t))\mathrm{e}^{\Omega(t,T)^\top}\xi\rangle.$$

When combined with (4.1), these expressions yield

$$\begin{aligned} W_t^{c,\xi}(t, y) &= -\langle \exp\{\Omega(t, T)^\top\}\xi, D_x b(X^0(t))y\rangle + \tfrac{1}{2}\|\exp\{\Omega(t, T)^\top\}\xi\|^2_{A(X^0(t))} \\ &= -\langle D_y W^{c,\xi}(t, y), D_x b(X^0(t))y\rangle + \tfrac{1}{2}\|D_y W^{c,\xi}(t, y)\|^2_{A(X^0(t))}, \end{aligned}$$

which proves the result. □

Note that

$$\sup_{(t,y)\in[0,T]\times\mathbb{R}^d} \|D_y W^{c,\xi}(t, y)\| \le \|\xi\| \sup_{t\in[0,T]} \|\exp\{\Omega(t, T)\}\| < \infty, \tag{4.3}$$

so $W^{c,\xi} \in \mathcal{S}$. To obtain an approximation to $W^{c,\xi}$, we need to approximate $\exp\{\Omega(t, T)^\top\}$ and

$$\int_t^T \exp\{\Omega(s, T)\}A(X^0(s))\exp\{\Omega(s, T)^\top\}\,\mathrm{d}s.$$

Note that these do not depend on $c \in \mathbb{R}$ or $\xi \in \mathbb{R}^d$. Consequently, we can use a large number of solutions W^k of the form W^{c_k,ξ_k} to construct $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ without substantially increasing the numerical work.

Next we define the mollifications that are used, which are smooth approximations of the pointwise minima of a finite collection of elements of $\mathcal{S}$. It is also possible to use pointwise maxima [18], which generalizes the framework of [10]. Owing to a closure property of $\mathcal{S}$ stated below in Lemma 4.1, this allows considerable flexibility in approximating the terminal condition. However, since the examples we consider do not benefit greatly from this generalization, discussion is limited to the case of pointwise minima to simplify the presentation.

Definition 4.1. Given K functions $f^k\colon [0, T]\times\mathbb{R}^d \to \mathbb{R}$, define the exponential mollification of $\min(\{f^k\}_{k=1}^K)$ with parameter $\delta > 0$ by

$$U^\delta(\{f^k\}_{k=1}^K)(t, y) \doteq -\delta \log\left\{\sum_{k=1}^K \exp\left\{-\frac{1}{\delta}f^k(t, y)\right\}\right\},$$

and the associated weightings

$$\rho_k^\delta(\{f^k\}_{k=1}^K)(t, y) \doteq \frac{\exp\{-(1/\delta)f^k(t, y)\}}{\sum_{k=1}^K \exp\{-(1/\delta)f^k(t, y)\}}.$$

It is not difficult to show that

$$-\delta\log K + \min(\{f^k\}_{k=1}^K) \le U^\delta(\{f^k\}_{k=1}^K) \le \min(\{f^k\}_{k=1}^K).$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

In addition, if $U(t,y) \doteq U^\delta(\{W^k\}_{k=1}^K)(t,y)$ and $\rho_k(t,y) \doteq \rho_k^\delta(\{W^k\}_{k=1}^K)(t,y)$, then

$$D_y U(t,y) = \sum_{k=1}^K \rho_k(t,y) D_y W^k(t,y) \quad \text{and} \quad U_t(t,y) = \sum_{k=1}^K \rho_k(t,y) W_t^k(t,y) \tag{4.4}$$

for all $(t,y) \in [0,T] \times \mathbb{R}^d$.

Hence, first derivatives of these mollifications are just weighted averages of the first derivatives of the component functions. The weights are given by ρ^δ, which belongs to the class of smooth weightings $\mathcal{W}$ given by (3.1), whenever $W^k \in C^1([0,T] \times \mathbb{R}^d : \mathbb{R})$ for $k = 1, \ldots, K$.

Lemma 4.1. *Suppose that the* $\{W^k\}_{k=1}^K \in \mathcal{S}$ *and that* $U^\delta(\{W^k\}_{k=1}^K)$ *is defined as in Definition 4.1. Then* $(U^\delta, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$.

Proof. Recall that U and $\{W^k\}_{k=1}^K$ satisfy (4.4). Clearly, $\rho_k \in C([0,T] \times \mathbb{R}^d : [0,1])$ for $k = 1, \ldots, K$, and since by assumption $W^k \in C^1([0,T] \times \mathbb{R}^d : \mathbb{R})$ for $k = 1, \ldots, K$, it follows that $U \in C^1([0,T] \times \mathbb{R}^d : \mathbb{R})$. In addition, $\|D_y U\|_\infty < \infty$ follows from the fact that $\|D_y W^k\|_\infty < \infty$ for $k = 1, \ldots, K$ and $\rho \in \mathcal{W}([0,T] \times \mathbb{R}^d)$. Also, U automatically also satisfies (2.11), since the convexity of $\|\alpha\|_A^2$ implies that

$$\begin{aligned}
U_t(t,y) &= \sum_{k=1}^K \rho_k(t,y) W_t^k(t,y) \\
&\geq \sum_{k=1}^K \rho_k(t,y) \left(\frac{1}{2} \|D_y W^k(t,y)\|^2_{A(X^0(t))} - \langle D_y W^k(t,y), D_x b(X^0(t)) y \rangle \right) \\
&\geq \frac{1}{2} \left\| \sum_{k=1}^K \rho_k(t,y) D_y W^k(t,y) \right\|^2_{A(X^0(t))} - \left\langle \sum_{k=1}^K \rho_k(t,y) D_y W^k(t,y), D_x b(X^0(t)) y \right\rangle \\
&= \tfrac{1}{2} \|D_y U(t,y)\|^2_{A(X^0(t))} - \langle D_y U(t,y), D_x b(X^0(t)) y \rangle.
\end{aligned}$$

This completes the proof. □

Thus, if one can approximate a given terminal condition well from below in terms of the minima of affine functions then the mollification will satisfy both the inequality in (2.11) and the terminal inequality (2.12) and provide good asymptotic performance when used for importance sampling schemes due to Theorem 3.1. As noted previously, one could also consider maxima of functions and combinations of maxima and minima if that were useful or necessary [18].

Given any $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ with U satisfying (2.12), it follows that $U(t,y) \leq V(t,y)$ for all $(t,y) \in [0,T] \times \mathbb{R}^d$, where V is given by (2.7). It is natural to ask if one could use information about V to assist in constructing a subsolution $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ with U satisfying (2.12) and with $U(0,0)$ close to $V(0,0)$. The following theorem gives an explicit quadratic form for the minimal point-to-point cost in the control problem (2.7), and can give information of this sort.

Theorem 4.2. *Define*

$$\begin{aligned}
C(z,t,y,r) \doteq \inf \Bigg\{ &\frac{1}{2} \int_t^r \|u(q)\|^2 \, \mathrm{d}q : \phi(r) = y, \ \phi(s) = z \\
&+ \int_t^s [D_x b(X^0(q)) \phi(q) + A^{1/2}(X^0(q)) u(q)] \, \mathrm{d}q, \ s \in [t,r] \Bigg\}
\end{aligned}$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

for $0 \le t < r \le T$ *and* $y, z \in \mathbb{R}^d$. *Then*

$$C(z,t,y,r) = \frac{1}{2}\left\langle (y - \mathrm{e}^{\Omega(t,r)}z), \left(\int_t^r \mathrm{e}^{\Omega(q,r)}A(X^0(q))\mathrm{e}^{\Omega(q,r)^\top}\,\mathrm{d}q\right)^{-1}(y - \mathrm{e}^{\Omega(t,r)}z)\right\rangle, \quad (4.5)$$

where we use Definition 2.1 for the inverse of a symmetric, nonnegative definite matrix.

Proof. The proof is straightforward when $A(X^0(t))$ is nondegenerate for all $t \in [0,T]$. The degenerate case is addressed by approximating with $A(X^0(t)) + \epsilon I$ and sending $\epsilon \to 0$. The details are omitted. □

We slightly abuse notation and write $C(T,y)$ for $C(0,0,T,y)$, i.e. when the initial time and position are both 0. Recall that the method of building subsolutions approximates G from below by a function $\bar{G}$ that is given as the minimum of affine functions, and then uses the mollification as given in Definition 4.1 to produce a classical sense subsolution U^δ for PDE (2.9) with the mollified version of $\bar{G}$ instead of G. Note that

$$V(0,0) = \inf_{y\in\mathbb{R}^d}\{C(T,y) + G(y)\}$$

due to Theorem 4.2. If $\bar{V}$ denotes the solution to the control problem (2.7) with terminal condition $\bar{G}$ then likewise

$$\bar{V}(0,0) = \inf_{y\in\mathbb{R}^d}\{C(T,y) + \bar{G}(y)\}.$$

Recall that according to Theorem 3.1, the rate of decay of the second moment of the importance sampling estimate based on subsolution U^δ is given by $U^\delta(0,0) + V(0,0)$, and that with the construction of U^δ just described, $U^\delta(0,0) \uparrow \bar{V}(0,0)$ as $\delta \to 0$. The bound in Theorem 3.1 assumes a fixed subsolution, and, hence, $\delta > 0$ is fixed before sending $n \to \infty$. However, one can often justify letting δ tend to 0 while $n \to \infty$; see [12], [14]. In any case, the focus here is on choosing $\bar{G}$ (given as the minimum of affine functions) to make $\bar{V}(0,0)$ as large as possible, and recall that $\bar{V}(0,0) \le V(0,0)$ because $\bar{G} \le G$. The following simple example indicates how the explicit quadratic form of $C(T,y)$ given by (4.5) can help in this process.

Example 4.1. Consider the case with $A(x) = \sqrt{2}$, $b(x) = 0$, terminal time $T = 1$, and

$$G(y) = \begin{cases} (y-1)^2 & \text{for } y > -1, \\ 0 & \text{for } y \le -1. \end{cases}$$

Thus, there is a discontinuity at $y = -1$. For this problem, $C(1,y) = y^2$ and

$$\inf_{y\in\mathbb{R}}\{C(1,y) + G(y)\} = C\big(1, \tfrac{1}{2}\big) + G\big(1, \tfrac{1}{2}\big) = \tfrac{1}{2}.$$

We would like to construct $\bar{G}$ so that

$$\inf_{y\in\mathbb{R}}\{C(1,y) + \bar{G}(y)\} = \tfrac{1}{2}. \quad (4.6)$$

Consider the affine function $g_1(y) = \frac{3}{4} - y$. Since g_1 agrees with G at $y = \frac{1}{2}$ and the same is true of their derivatives, this function satisfies

$$\inf_{y\in\mathbb{R}}\{C(1,y) + g_1(y)\} = \tfrac{1}{2}$$

with $y = \frac{1}{2}$ as the unique minimizer. However, $g_1(y) > G(y)$ for $y \le -1$, and so it does not approximate G from below. As a result, we look for a second affine function $g_2(y)$ that will

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

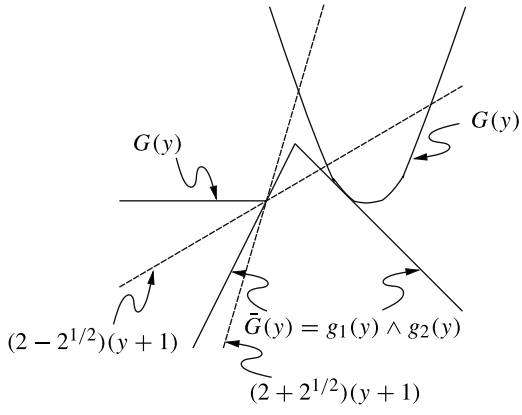

FIGURE 1: Range of slopes possible for $g_2(y)$.

intersect $G(y)$ at $y = -1$ (i.e. $g_2(-1) = 0$), and such that $\bar{G}(y) = \min\{g_1(y), g_2(y)\}$ satisfies $\bar{G}(y) \le G(y)$ for all $y \in \mathbb{R}$. The requirement $g_2(-1) = 0$ implies that $g_2(y) = c(y + 1)$, and if $0 \le c \le \infty$ then $\bar{G}(y) \le G(y)$ for all $y \in \mathbb{R}$. However, this does not guarantee (4.6), since a big gap between $\bar{G}(y)$ and $G(y)$ for $y \le -1$ might lead to a value that is strictly smaller than $\frac{1}{2}$. Using the explicit form $C(1, y) = y^2$, if either $c < 2 - 2^{1/2}$ or $c > 2 + 2^{1/2}$ then $\inf_{y\in\mathbb{R}}\{C(1, y) + \bar{G}(y)\} < \frac{1}{2}$ with minimizers in $(-1 + 1/2^{1/2}, 0]$ and $(-\infty, -1 - 1/2^{1/2})$, respectively. If $2 - 2^{1/2} < c < 2 + 2^{1/2}$ then (4.6) applies with a unique minimizer at $y = \frac{1}{2}$. The range of allowed g_2 is depicted in Figure 1.

Remark 4.1. In the moderate deviations control problem, it is always the case that the drift is linear in the position and the control cost is quadratic (and independent of the position). For terminal conditions that are quadratic, the classical sense solution to (2.9) and (2.10) is given in terms of the famous linear-quadratic regulator problem [1]. Hence, it would seem natural for $\mathcal{S}$ to include functions of this form. However, we typically restrict $\mathcal{S}$ even further, and consider only affine terminal conditions. The reason for this, which was made precise earlier in this section, is that the computations needed to produce a solution for just one terminal condition immediately give all such solutions since the associated Riccati equation needs to be solved only once. This is not true (in general) for different quadratic terminal conditions.

5. Numerical examples

In this section we test the performance of importance sampling schemes based on moderate deviations subsolutions. All the simulations presented here use the deterministic implementation as opposed to the randomized implementation. An aim of the section is to compare the performance of the standard importance sampling scheme of Construction 3.1 to that of the corrected importance sampling scheme of Construction 3.2. For convenience, in all of the examples we use an embedding satisfying $a(n^*)\sqrt{n^*} = 1$ (see the remarks above Construction 2.1).

In the table of numerical results we use the following notation, where we assume that K is the number of samples and $\{e_k\}_{k=1}^K$ are the individual samples. Thus,

(i) estimate: $\text{est} = (1/K)\sum_{k=1}^K e_k$;

(ii) standard deviation estimate: $\text{SDe} = ((1/(K-1))\sum_{k=1}^K (e_k - \text{est})^2)^{1/2} K^{-1/2}$;

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

(iii) confidence interval: $\mathrm{CI} = [\mathrm{est} - (1.96) \times \mathrm{SDe}, \mathrm{est} + (1.96) \times \mathrm{SDe}]$;

(iv) relative error: $\mathrm{RE} = K^{1/2}(\mathrm{SDe}/\mathrm{est})$;

(v) ratio: $\mathrm{ratio} = \{-\log((1/K)\sum_{k=1}^{K} e_k^2)\}/\{-\log((1/K)\sum_{k=1}^{K} e_k)\}$.

Recall that when using an asymptotically optimal scheme, the quantity that 'ratio' is estimating will converge to 2 as $n \to \infty$. The 'RE' is normalized to give the ratio of the standard deviation of a single sample and the estimate.

5.1. One dimension, exponential noise, and linear drift

Recall that results presented under the moderate deviations scaling require only that the moment-generating function be finite in a neighborhood of the origin. Here we consider an example where the moment-generating function is not finite everywhere. Although a standard process-level large deviations theory is not currently available for this model (in particular the rate function would not have compact level sets in $C([0, T]: \mathbb{R}^d)$), we can still implement a scheme based on a formal use of the corresponding equations, and owing to the particular form of the event of interest, a theoretical justification could be provided. Let $X_i^n \in \mathbb{R}$ and

$$X_{i+1}^n = X_i^n - \frac{1}{n} X_i^n + \frac{1}{n} u_i, \qquad X_0^n = 5,$$

where $u_i = \theta_i - 1$, and the $\{\theta_i\}_{i=1}^{\infty}$ are i.i.d. exponential random variables with mean 1. The law of large numbers limit is $X^0(t) = 5\mathrm{e}^{-t}$ for $t \in [0, 1]$, and we are interested in estimating the probability

$$p_n^{\alpha} = \mathbb{P}(X^n(1) - X^0(1) \notin (-\alpha, \alpha))$$

for various values of $\alpha > 0$ and $n^* = 200$. Because we use the embedding $a(n^*)\sqrt{n^*} = 1$, and since $G(y)$ should satisfy (2.8), we take $F(x) = \infty \mathbf{1}_{(5\mathrm{e}^{-1}-\alpha,\, 5\mathrm{e}^{-1}+\alpha)}(x)$ and $G(y) = \infty \mathbf{1}_{(-\alpha,\alpha)}(y)$.

Using Theorem 4.2 to evaluate $C(y, 1)$,

$$V(0, 0) = \inf_{y \in \mathbb{R}} \{\infty \mathbf{1}_{(-\alpha,\alpha)}(y) + C(y, 1)\} = \inf_{y \in \mathbb{R}} \{\infty \mathbf{1}_{(-\alpha,\alpha)}(y) + y^2(1 - \mathrm{e}^{-2})^{-1}\},$$

and clearly the minimizers are $y = \pm\alpha$. We need to approximate $\infty \mathbf{1}_{(-\alpha,\alpha)}(y)$ from below by $\bar{G}$, where $\bar{G}$ is the minimum of affine functions, in such a way that

$$\inf_{y \in \mathbb{R}} \{\infty \mathbf{1}_{(-\alpha,\alpha)}(y) + C(y, 1)\} = \inf_{y \in \mathbb{R}} \{\bar{G}(y) + C(y, 1)\}.$$

This can be achieved by choosing an affine function (a line since $d = 1$) for each minimizer $\pm\alpha$ equal to 0 at the minimizer and with slope equal to $-D_y C(y, 1)$ at the minimizer. Define classical solutions to (2.9), $W^{c_1,\beta_1}(t, y)$ and $W^{c_2,\beta_2}(t, y)$, as in (4.2) with

$$\beta_1 = 2\alpha(1 - \mathrm{e}^{-2})^{-1}, \qquad c_1 = \alpha\beta_1, \qquad \beta_2 = -2\alpha(1 - \mathrm{e}^{-2})^{-1}, \qquad c_2 = -\alpha\beta_2.$$

These functions have terminal values $W^{c_i,\beta_i}(y, 1) = c_i + \beta_i y, i = 1, 2$. Note that

$$\bar{G}(y) = \min\{W^{c_1,\beta_1}(y, 1), W^{c_2,\beta_2}(y, 1)\} \le \infty \mathbf{1}_{(-\alpha,\alpha)}(y) \quad \text{for all } y \in \mathbb{R},$$

and

$$\inf_{y \in \mathbb{R}} \{\bar{G}(y) + y^2(1 - \mathrm{e}^{-2})^{-1}\} = V(0, 0),$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

with minimizers $y = \pm\alpha$. For comparison purposes we also implement a scheme based on a large deviations subsolution. This large deviations subsolution was created analogously to the moderate deviations subsolution. It is a mollification of the minimum of two exact solutions with affine (linear) terminal conditions which have the same value at the origin as the true solution.

The numerical results in the top half of Table 1 were computed with $n^* = 200$, $\delta = 0.01$, and using 100,000 samples.

The corrected moderate deviations importance sampling substantially outperforms the standard version for larger values of α. In fact, for $\alpha = 0.4$ the standard moderate deviations importance sampling scheme is not at all accurate. This is because the moderate deviations control problems treat the noise as if it is Gaussian, which for $\alpha = 0.4$ suggests an exponential tilt close to 1 (the moment-generating function is finite only for values less than 1), which results in an enormous conditional mean. The corrected scheme takes into account the true noise and adjusts for this. The corrected moderate deviations importance sampling scheme does not perform quite as well as the large deviations counterpart, but does well considering the limited information regarding the process that is used for its design, and for the fairly broad range of probabilities.

5.2. Finite-state mean-field interacting particle system

We next consider a process which lies outside the scope of the results proved here, in that the process model evolves in continuous rather than discrete time. We can still formally apply the importance sampling approach proposed in this paper, and extending the results on asymptotic efficiency to this setting would not be difficult. The computational effort needed to generate samples for the model of this section is substantial for large n. Hence, the moderate deviations importance sampling might be of interest even if the estimated probabilities are not very small, so long as there is a significant improvement over standard Monte Carlo. For the three-state model considered here one could, with some effort, identify a suitable large deviations-based sampling scheme. However, we have not done so and only compare with standard Monte Carlo. Depending on the model, it could be difficult to identify a suitable large deviations scheme when there are more than three states.

Consider n particles $\{Z_i^n(\cdot)\}_{i=1}^n$ each taking values in the finite space $\{1, 2, 3\}$. Let $Z^n(\cdot)$ evolve as a càdlàg $\{1, 2, 3\}^n$-valued jump Markov process. The associated empirical measure is $X^n(\cdot) = (1/n)\sum_{i=1}^n \delta_{Z_i^n(\cdot)}$ and $X^n(\cdot)$ is a stochastic process taking values in

$$\mathcal{S} = \left\{x \in \mathbb{R}^3 \colon x_i \geq 0,\ \sum_{i=1}^3 x_i = 1\right\}.$$

Let the wait times until the next jump of the particles be independent, conditional on the empirical measure, and let the jump rates of each particle be given by $r_{ij}(x) = 2 - x_i$ when $X^n(\cdot) = x$ is the measure the empirical distribution puts on state i. This gives transition rates for the empirical distribution $X^n(\cdot)$ in the form

$$R\left(x, x + \frac{1}{n}(e_j - e_i)\right) = nx_i(2 - x_j),$$

where $R(x, y)$ is the transition rate from state x to state y with $x, y \in \mathcal{S}$, and e_i is the unit vector in the ith direction in $\mathbb{R}^3$. Given an initial distribution $x_0 \in \mathcal{S}$, it can be shown that the law of large numbers limit $X^0(\cdot)$ satisfies the ODE

$$X_t^0(t) = 6\left(\left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right) - X^0(t)\right), \qquad X^0(0) = x_0.$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

TABLE 1.

Corrected moderate deviation importance sampling.			
	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$
est	3.42×10^{-5}	6.69×10^{-9}	2.11×10^{-13}
SDe	4.40×10^{-7}	1.41×10^{-10}	5.98×10^{-15}
CI	$[3.33, 3.51] \times 10^{-5}$	$[6.41, 6.97] \times 10^{-9}$	$[1.99, 2.23] \times 10^{-13}$
RE	4.06	6.67	8.96
ratio	1.72	1.80	1.85
Standard moderate deviation importance sampling.			
est	3.40×10^{-5}	7.84×10^{-9}	7.24×10^{-42}
SDe	6.55×10^{-7}	2.79×10^{-9}	5.15×10^{-42}
CI	$[3.27, 3.53] \times 10^{-5}$	$[2.37, 13.3] \times 10^{-9}$	$[-2.86, 17.3] \times 10^{-42}$
RE	6.09	1.13×10^{2}	2.25×10^{2}
ratio	1.65	1.49	1.89
Large deviation importance sampling.			
est	3.32×10^{-5}	6.75×10^{-9}	2.14×10^{-13}
SDe	2.39×10^{-7}	5.98×10^{-11}	2.20×10^{-15}
CI	$[3.27, 3.37] \times 10^{-5}$	$[6.63, 6.87] \times 10^{-9}$	$[2.10, 2.18] \times 10^{-13}$
RE	2.27	2.80	3.25
ratio	1.82	1.88	1.92
Corrected moderate deviation importance sampling.			
	$\alpha = 0.12$	$\alpha = 0.16$	$\alpha = 0.2$
est	9.04×10^{-5}	2.14×10^{-7}	1.23×10^{-10}
SDe	7.78×10^{-7}	2.40×10^{-9}	1.59×10^{-12}
CI	$[8.89, 9.19] \times 10^{-5}$	$[2.09, 2.19] \times 10^{-7}$	$[1.20, 1.26] \times 10^{-10}$
RE	2.72	3.54	4.09
ratio	1.77	1.83	1.87
Standard moderate deviation importance sampling.			
est	9.13×10^{-5}	2.16×10^{-7}	1.22×10^{-10}
SDe	7.91×10^{-7}	2.44×10^{-9}	1.62×10^{-12}
CI	$[8.98, 9.29] \times 10^{-5}$	$[2.11, 2.21] \times 10^{-7}$	$[1.19, 1.25] \times 10^{-10}$
RE	2.74	3.58	4.20
ratio	1.77	1.83	1.87
Standard Monte Carlo.			
est	5.00×10^{-5}	–	–
SDe	2.24×10^{-5}	–	–
CI	$[.617, 9.38] \times 10^{-5}$	–	–
RE	1.41×10^{2}	–	–
ratio	1.00	–	–

Under the moderate deviations scaling, we consider the quantity $Y^n(\cdot) = a(n)\sqrt{n}(X^n(\cdot) - X^0(\cdot)) \in \mathbb{R}^3$. We use initial condition $x_0 = (1, 0, 0)$ and are interested in the probability

$$p_n^\alpha = \mathbb{P}(X_1^n(1) - X_1^0(1) \notin (-\alpha, \alpha))$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

for various values of $\alpha > 0$ and $n^* = 200$. Subsolutions to the associated HJB equation must satisfy (2.11) with

$$A_{i,j}(X^0(t)) = -2(X_i^0(t) + X_j^0(t) - X_i^0(t)X_j^0(t)) \quad \text{for } i \neq j,$$
$$A_{i,i}(X^0(t)) = 2(1 + (X_i^0(t))^2),$$

and $Db(X^0(t)) = 6I$, where I is the identity matrix.

In the moderate deviations subsolution we base our importance sampling scheme on is the mollified minimum of two exact solutions to (2.9), $W^{c_1,\beta_1}(t, y)$ and $W^{c_2,\beta_2}(t, y)$, as in (4.2). We numerically approximated $C(y, 1)$ based on Theorem 4.2 and chose parameters

$$\beta_1 = (-5.39, 0, 0)\alpha, \quad c_1 = -5.39\alpha^2 \quad \text{and} \quad \beta_2 = (5.39, 0, 0)\alpha, \quad c_2 = 5.39\alpha^2$$

so that

$$\bar{G}(y) = \min\{W^{c_1,\beta_1}(y, 1), W^{c_2,\beta_2}(y, 1)\} \leq \infty \mathbf{1}_{(-\alpha,\alpha)}(y)$$

and

$$\inf_{y\in\mathbb{R}} \{\bar{G}(y) + C(y, 1)\} = V(0, 0).$$

The numerical results in the lower half of Table 1 were computed with $n^* = 200$, $\delta = 0.01$, and using 100,000 samples.

In this example, the performance is comparable between the corrected and the standard moderate deviations importance sampling schemes, in part because the moment-generating function is not infinite at any point. Both schemes have small relative errors for all three values of α, and standard Monte Carlo is not useful using this sample size. Note that '–' in the results indicates that no escapes occurred. Hence, for this problem, a scheme based on a moderate deviation makes the computations quite feasible, even though this is not true for standard Monte Carlo. At the same time, the effort needed to construct the scheme is less (and in some cases will be much less) than that which would be needed for an analogous large deviation-based scheme.

6. Proof of Theorem 3.1

The proof of this theorem differs from its large deviations counterpart found in [10] because of difficulties with tightness. As in the proof of the moderate deviations principle [8], we can only obtain tightness of the occupation measure of the conditional means of the controlled noises, rather than tightness of the occupation measure of the controlled noises themselves. To obtain this tightness, we use Theorem 6.1 (Theorem 2.5 of [8]), which assumes a bound on the relative entropy of the control measure with respect to the original measure. This bound is also required for the tightness proved in [10], however it is attained more easily there since it is assumed that the moment-generating function of the noise $u_i(x)$ is finite everywhere. Without this assumption, the approach used in [10] becomes significantly more complicated, and we found it more convenient to use a second change of measure, a technique first used in [13].

The proof given below applies to the importance sampling schemes of both Construction 3.1 and Construction 3.2. We recall that the setup involves a subsolution $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ (see (3.2)). For the nth process in Construction 3.1, the tilt parameter based on $W^k \in \mathcal{S}$ is given by

$$-\frac{1}{a(n)\sqrt{n}} D_y W^k\left(\frac{i}{n}, y\right),$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

but in Construction 3.2 it is given by Definition 3.1. To avoid specifying a construction we will refer in both cases to the tilt parameter based on W^k for the nth process at step i and position y by $\xi_i^{n,k}(y)$. Note that because of Definition 3.1, regardless of which construction is chosen, $\xi_i^{n,k}(y)$ satisfies

$$\left\| -D_y W^k\left(\frac{i}{n}, y\right) - a(n)\sqrt{n}\xi_i^{n,k}(y)\right\| \le \frac{K_T}{a(n)\sqrt{n}}. \tag{6.1}$$

We label the importance sampling estimators $\{r^n\}$, the importance sampling changes of measure $\{\gamma^n\}$, and corresponding process-level random variables $\{(\bar{u}^n, \bar{X}^n, \bar{Y}^n)\}$.

Recall that we want to prove that

$$\liminf_{n\to\infty} -a(n)^2 \log \mathbb{E}[(r^n)^2] \ge U(0,0) + V(0,0),$$

where V is given by (2.7). Assuming without loss of generality that $\liminf_{n\to\infty} -a(n)^2 \log \mathbb{E}[(r^n)^2] = L < \infty$, we consider any subsequence $n(m)$ such that

$$\lim_{m\to\infty} -a(n(m))^2 \log \mathbb{E}[(r^{n(m)})^2] = L.$$

We will show that for this subsequence the corresponding process-level random variables are tight, and that along any convergent subsubsequence (for convenience again labeled by $n(m)$)

$$\lim_{m\to\infty} -a(n(m))^2 \log \mathbb{E}[(r^{n(m)})^2] \ge U(0,0) + V(0,0).$$

The proof is divided into first proving tightness and then weak convergence analysis. We will use the following to prove tightness. For probability measures η and μ on the same measurable space let $R(\eta\|\mu)$ denote the relative entropy of η with respect to μ [6, Section 1.4].

Theorem 6.1. *Let $\{\eta^n\}$ be a sequence of measures with each $\eta^n \in \mathcal{P}((\mathbb{R}^d)^{\lceil nT\rceil})$, and define the corresponding random variables $\{(\hat{\varsigma}^n, \hat{X}^n, \hat{Y}^n)\}$ based on $\{\eta^n\}$ as in Construction 2.1. If*

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R(\eta_i^n \| \mu(\cdot \mid \hat{X}_i^n))\right] < \infty$$

then $\{(\hat{\varsigma}^n, \hat{Y}^n)\}$ is tight in $\mathcal{P}((\mathbb{R}^d \times [0,T])) \times C([0,T]: \mathbb{R}^d)$. In addition, along any convergent subsequence (maintaining n as the index for convenience) with limit $(\hat{\varsigma}, \hat{Y})$, the $\{\hat{\varsigma}^n\}$ are uniformly integrable in the sense that

$$\lim_{C\to\infty} \limsup_{n\to\infty} \mathbb{E}\left[\int_{\{w:\, \|w\|\ge C\}} \|w\| \hat{\varsigma}^n(\mathrm{d}w \times \mathrm{d}t)\right] = 0. \tag{6.2}$$

The limit $\hat{\varsigma}$ has with probability 1 the decomposition $\hat{\varsigma}^n(\mathrm{d}w \times \mathrm{d}t) = \hat{\varsigma}_{1\,|\,2}(\mathrm{d}w \mid t)\,\mathrm{d}t$, and

$$\hat{Y}(t) = \int_0^t Db(X^0(s))\hat{Y}(s)\,\mathrm{d}s + \int_0^t \hat{w}(s)\,\mathrm{d}s,$$

where $\hat{w}(t) = \int_{\mathbb{R}^d} y\hat{\varsigma}_{1\,|\,2}(\mathrm{d}y \mid t)$.

Proof. See [8, Theorem 2.5]. □

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

Throughout this section we will also use the following inequalities. The bound (2.3) implies that there exists $K_{DA} < \infty$ and $\lambda_{DA} \in (0, \lambda]$ (independent of x) such that

$$\sup_{x\in\mathbb{R}^d} \sup_{\|\alpha\|\le\lambda_{DA}} \max_{i,j,k} \left| \frac{\partial^3 H_c(x,\alpha)}{\partial\alpha_i \partial\alpha_j \partial\alpha_k} \right| \le \frac{K_{DA}}{d^3}.$$

Consequently, for all $\|\alpha\| \le \lambda_{DA}$ and all $x \in \mathbb{R}^d$,

$$\tfrac{1}{2}\|\alpha\|^2_{A(x)} - \|\alpha\|^3 K_{DA} \le H_c(x,\alpha) \le \tfrac{1}{2}\|\alpha\|^2_{A(x)} + \|\alpha\|^3 K_{DA}. \tag{6.3}$$

6.1. Tightness

First note that if $\liminf_{n\to\infty} -a(n)^2 \log \mathbb{E}[(r^n)^2] = \infty$ then the result is true trivially, so we assume that

$$\liminf_{n\to\infty} -a(n)^2 \log \mathbb{E}[(r^n)^2] = L < \infty. \tag{6.4}$$

We recall the form of r^n and the likelihood ratios from either Construction 3.1 or 3.2, and that in both cases $\xi_i^{n,k}$ denotes the state- and time-dependent tilt parameters. Since removing one likelihood ratio amounts to replacing the process under γ^n by the original process, it follows that

$$\begin{aligned}
\mathbb{E}[(r^n)^2] &= \mathbb{E}\left[\exp\left\{-\frac{1}{a(n)^2} 2G(\bar{Y}^n(T))\right\}\left(\prod_{i=0}^{\lfloor Tn\rfloor} \frac{\mathrm{d}\mu(\cdot \mid \bar{X}_i^n)}{\mathrm{d}\gamma_i^n(\cdot \mid \bar{X}_i^n)}(\bar{u}_i^n)\right)^2\right] \\
&= \mathbb{E}\left[\exp\left\{-\frac{1}{a(n)^2} 2G(Y^n(T))\right\}\right. \\
&\qquad \left. \times \prod_{i=0}^{\lfloor Tn\rfloor}\left(\sum_{k=1}^K \rho_k\left(\frac{i}{n}, Y_i^n\right)\exp\{\langle u_i(X_i^n), \xi_i^{n,k}(Y_i^n)\rangle - H_c(X_i^n, \xi_i^{n,k}(Y_i^n))\}\right)^{-1}\right].
\end{aligned}$$

This is the first change of measure, which equates the second moment with an expectation under the measure of the original process. In contrast to the approach taken in [10] (which involves a large deviations scaling as opposed to a moderate deviations one), we find it convenient to make a second change of measure to remove the exponential tilt term as in [13]. Note that Jensen's inequality yields

$$\begin{aligned}
&\sum_{k=1}^K \rho_k\left(\frac{i}{n}, Y_i^n\right)\exp\{\langle u_i(X_i^n), \xi_i^{n,k}(Y_i^n)\rangle - H_c(X_i^n, \xi_i^{n,k}(Y_i^n))\} \\
&\qquad \ge \exp\left\{\sum_{k=1}^K \rho_k\left(\frac{i}{n}, Y_i^n\right)(\langle u_i(X_i^n), \xi_i^{n,k}(Y_i^n)\rangle - H_c(X_i^n, \xi_i^{n,k}(Y_i^n)))\right\}.
\end{aligned}$$

We now introduce the second change of measure. Define $(\tilde{u}^n, \tilde{X}^n, \tilde{Y}^n)$ in terms of θ^n as in Construction 2.1, where the conditional distribution of $\tilde{u}_i^n$, given $\tilde{u}_j^n$, $j < i$, is given by

$$\theta_i^n(\mathrm{d}u \mid \tilde{X}_i^n) = \exp\{\langle u, -\bar{\xi}_i^n(\tilde{Y}_i^n)\rangle - H_c(\tilde{X}_i^n, -\bar{\xi}_i^n(\tilde{Y}_i^n))\}\mu(\mathrm{d}u \mid \tilde{X}_i^n) \tag{6.5}$$

and

$$\bar{\xi}_i^n(y) = \sum_{k=1}^K \rho_k\left(\frac{i}{n}, y\right)\xi_i^{n,k}(y). \tag{6.6}$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

The use of this second change of measure is a technical device to make a term in the exponent in the expression for r^n uniformly bounded. In particular, using Jensen's inequality for the inequality and (6.5) for the equality,

$$\begin{aligned}
&\mathbb{E}\Bigg[\exp\Bigg\{-\frac{1}{a(n)^2}2G(Y^n(T))\Bigg\}\prod_{i=0}^{\lfloor Tn\rfloor}\Bigg(\sum_{k=1}^{K}\rho_k\Bigg(\frac{i}{n},Y_i^n\Bigg)\exp\{\langle u_i(X_i^n),\xi_i^{n,k}(Y_i^n)\rangle \\
&\qquad\qquad - H_c(X_i^n,\xi_i^{n,k}(Y_i^n))\}\Bigg)^{-1}\Bigg] \\
&\le \mathbb{E}\Bigg[\exp\Bigg\{-\frac{1}{a(n)^2}2G(Y^n(T))\Bigg\}\prod_{i=0}^{\lfloor Tn\rfloor}\exp\Bigg\{\sum_{k=1}^{K}\rho_k\Bigg(\frac{i}{n},Y_i^n\Bigg)(\langle u_i(X_i^n),-\xi_i^{n,k}(Y_i^n)\rangle \\
&\qquad\qquad + H_c(X_i^n,\xi_i^{n,k}(Y_i^n)))\Bigg\}\Bigg] \\
&= \mathbb{E}\Bigg[\exp\Bigg\{-\frac{1}{a(n)^2}2G(\tilde{Y}_i^n(T))\Bigg\} \\
&\qquad \times\exp\Bigg\{\sum_{i=0}^{\lfloor Tn\rfloor}H_c(\tilde{X}_i^n,-\bar{\xi}_i^n(\tilde{Y}_i^n))+\sum_{k=1}^{K}\rho_k\Bigg(\frac{i}{n},\tilde{Y}_i^n\Bigg)H_c(\tilde{X}_i^n,\xi_i^{n,k}(\tilde{Y}_i^n))\Bigg\}\Bigg]. \qquad (6.7)
\end{aligned}$$

We use the following variational formula for exponential integrals. For a proof, see [6, Proposition 4.5.1].

Lemma 6.1. *Let $\mathcal{X}$ be a Polish space and let $\mathcal{P}(\mathcal{X})$ be the collection of probability measures on $\mathcal{X}$ with the Borel σ-algebra. Let g be a Borel measurable function that is bounded from below. Then, for any $\mu\in\mathcal{P}(\mathcal{X})$,*

$$-\log\int_{\mathcal{X}}\mathrm{e}^{-g(x)}\mu(\mathrm{d}x)=\inf_{\eta\in\mathcal{P}(\mathcal{X})}\Bigg\{R(\eta\|\mu)+\int_{\mathcal{X}}g(x)\eta(\mathrm{d}x)\Bigg\}.$$

Recall that G is assumed to be bounded from below and that, due to (4.3), we have $\|DW^k\|_\infty<\infty$ for all k. Thus, by (6.1) and (2.3),

$$-\sum_{i=0}^{\lfloor Tn\rfloor}\sum_{k=1}^{K}\rho_k\Bigg(\frac{i}{n},\tilde{Y}_i^n\Bigg)H_c(\tilde{X}_i^n,\xi_i^{n,k}(\tilde{Y}_i^n))-H_c(\tilde{X}_i^n,-\bar{\xi}_i^n(\tilde{Y}_i^n))$$

is bounded for sufficiently large n. Consequently, (6.7) and Lemma 6.1 yield

$$\begin{aligned}
&-a(n)^2\log\mathbb{E}[(r^n)^2] \\
&\quad\ge\inf_{\eta}\Bigg\{a(n)^2R(\eta\|\theta^n) \\
&\qquad+\mathbb{E}\Bigg[2G(\bar{Y}^{n,\eta}(T))-a(n)^2\sum_{i=0}^{\lfloor Tn\rfloor}\sum_{k=1}^{K}\rho_k\Bigg(\frac{i}{n},\bar{Y}_i^{n,\eta}\Bigg)H_c(\bar{X}_i^{n,\eta},\xi_i^{n,k}(\bar{Y}_i^{n,\eta})) \\
&\qquad\qquad-a(n)^2\sum_{i=0}^{\lfloor Tn\rfloor}H_c(\bar{X}_i^{n,\eta},-\bar{\xi}_i^n(\bar{Y}_i^{n,\eta}))\Bigg]\Bigg\}, \qquad (6.8)
\end{aligned}$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

where $(\bar{u}^{n,\eta}, \bar{X}^{n,\eta}, \bar{Y}^{n,\eta})$ is defined in terms of η as in Construction 2.1. Let

$$K_\xi \doteq \max\{\|D_y W^1\|_\infty, \ldots, \|D_y W^K\|_\infty\} + K_T \tag{6.9}$$

and note that, for large enough n,

$$a(n)\sqrt{n} \geq \max\left\{1, \frac{2K_\xi}{\lambda_{DA}}\right\}, \tag{6.10}$$

we have

$$\|\xi_i^{n,k}(y)\| \leq \frac{\lambda_{DA}}{2} \quad \text{for all } i, k \text{ and } y \in \mathbb{R}^d. \tag{6.11}$$

Given any $\varepsilon > 0$, let $\{\eta^n\}$ come within ε of achieving the infimum in (6.8), and let $(\hat{u}^n, \hat{X}^n, \hat{Y}^n)$ be associated with η^n through Construction 2.1. Then

$$\begin{aligned}-a(n)^2 \log \mathbb{E}[(r^n)^2] + \varepsilon \geq a(n)^2 R(\eta^n \| \theta^n) \\ + \mathbb{E}\Bigg[2G(\hat{Y}^n(T)) - a(n)^2 \sum_{i=0}^{\lfloor Tn \rfloor} H_c(\hat{X}_i^n, -\bar{\xi}_i^n(\hat{Y}_i^n)) \\ - a(n)^2 \sum_{i=0}^{\lfloor Tn \rfloor} \sum_{k=1}^{K} \rho_k\left(\frac{i}{n}, \hat{Y}_i^n\right) H_c(\hat{X}_i^n, \xi_i^{n,k}(\hat{Y}_i^n))\Bigg],\end{aligned}$$

and, for all n satisfying (6.10) (recall (2.4) and (6.3)),

$$-a(n)^2 \log \mathbb{E}[(r^n)^2] + \varepsilon + 2T K_\xi^2 \left(\frac{K_A}{2} + \lambda_{DA} K_{DA}\right) - 2 \inf_{y \in \mathbb{R}^d} G(y) \geq a(n)^2 R(\eta^n \| \theta^n).$$

Combining this with (6.4), we can choose a subsequence of $\{\eta^n, r^n\}$, retaining n as the index for convenience, such that

$$\limsup_{n\to\infty} a(n)^2 R(\eta^n \| \theta^n) < \infty \quad \text{and} \quad \lim_{n\to\infty} -a(n)^2 \log \mathbb{E}[(r^n)^2] = L.$$

The subsequence satisfies

$$\begin{aligned}L + \varepsilon \geq \limsup_{n\to\infty} \Bigg(a(n)^2 R(\eta^n \| \theta^n) + \mathbb{E}[2G(\hat{Y}^n(T))] \\ - \mathbb{E}\Bigg[a(n)^2 \sum_{i=0}^{\lfloor Tn \rfloor} H_c(\hat{X}_i^n, -\bar{\xi}_i^n(\hat{Y}_i^n))\Bigg] \\ - \mathbb{E}\Bigg[a(n)^2 \sum_{i=0}^{\lfloor Tn \rfloor} \sum_{k=1}^{K} \rho_k\left(\frac{i}{n}, \hat{Y}_i^n\right) H_c\left(\hat{X}_i^n, \xi_i^{n,k}\left(\hat{Y}_i^n, \frac{i}{n}\right)\right)\Bigg]\Bigg).\end{aligned} \tag{6.12}$$

Note that we must keep the linear combination of the $H_c(x, \xi_i^{n,k})$ with weights ρ_k instead of replacing it with $H_c(x, \bar{\xi}_i^n)$, because $H_c(x, \cdot)$ is convex rather than concave. Using the chain rule for relative entropy [6, Theorem C.3.1], we can write

$$a(n)^2 R(\eta^n \| \theta^n) = \mathbb{E}\left[\frac{1}{n} \sum_{i=0}^{\lfloor Tn \rfloor} a(n)^2 n R(\eta_i^n \| \theta_i^n(\cdot \mid \hat{X}_i^n))\right],$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

where η_i^n is the conditional distribution on $\hat{u}_i^n$ given $\hat{u}_j^n$ for $j < i$ under η^n as in Construction 2.1 and θ_i^n is defined in (6.5).

We would like to obtain a bound on

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R(\eta_i^n \| \mu(\cdot \mid \hat{X}_i^n))\right]$$

from one on

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R(\eta_i^n \| \theta_i^n(\cdot \mid \hat{X}_i^n))\right].$$

This new bound will allow us to invoke Theorem 6.1. We first need the following, whose proof is practically identical to that of [8, Lemma 3.1] (which is the same result except it is assumed in [8, Lemma 3.1] that the noise has zero mean) and is consequently omitted.

Lemma 6.2. *Given $\mu \in \mathcal{P}(\mathbb{R}^d)$, let*

$$H(\alpha) = \log\left(\int_{\mathbb{R}^d} \mathrm{e}^{\langle \alpha, y\rangle}\mu(\mathrm{d}y)\right) \quad \text{and} \quad L(\beta) = \sup_{\alpha\in\mathbb{R}^d}\{\langle \alpha, \beta\rangle - H(\alpha)\}.$$

If there exists $\lambda > 0$ such that $\sup_{\|\alpha\|\le\lambda} H(\alpha) < \infty$ then, for any $\eta \in \mathcal{P}(\mathbb{R}^d)$,

$$R(\eta\|\mu) \ge L\left(\int_{\mathbb{R}^d} u\eta(\mathrm{d}u)\right).$$

Lemma 6.3. *Let $\{\eta^n\}$ be a sequence of measures with $\eta^n \in \mathcal{P}((\mathbb{R}^d)^{\lceil Tn\rceil})$ and define the corresponding random variables based on these measures as in Construction 2.1. Define the measures θ^n using (6.5). If*

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R(\eta_i^n \| \theta_i^n(\cdot \mid \hat{X}_i^n))\right] < \infty$$

then

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R(\eta_i^n \| \mu(\cdot \mid \hat{X}_i^n))\right] < \infty. \tag{6.13}$$

Proof. From (6.5), we have

$$\begin{aligned}
R(\eta_i^n\|\mu(\cdot \mid \hat{X}_i^n)) &= \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\eta_i^n}{\mathrm{d}\theta_i^n(\cdot \mid \hat{X}_i^n)}(u)\frac{\mathrm{d}\theta_i^n(\cdot \mid \hat{X}_i^n)}{\mathrm{d}\mu(\cdot \mid \hat{X}_i^n)}(u)\right)\eta_i^n(\mathrm{d}u)\\
&= \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\eta_i^n}{\mathrm{d}\theta_i^n(\cdot \mid \hat{X}_i^n)}(u)\right)\eta_i^n(\mathrm{d}u) + \int_{\mathbb{R}^d} \log\left(\frac{\mathrm{d}\theta_i^n(\cdot \mid \hat{X}_i^n)}{\mathrm{d}\mu(\cdot \mid \hat{X}_i^n)}(u)\right)\eta_i^n(\mathrm{d}u)\\
&= R(\eta_i^n\|\theta_i^n(\cdot \mid \hat{X}_i^n)) + \int_{\mathbb{R}^d} \langle u, -\bar{\xi}_i^n(\hat{Y}_i^n)\rangle \eta_i^n(\mathrm{d}u) - H_c(\hat{X}_i^n, -\bar{\xi}_i^n(\hat{Y}_i^n))\\
&\le R(\eta_i^n\|\theta_i^n(\cdot \mid \hat{X}_i^n)) + \frac{1}{a(n)\sqrt{n}} K_\xi \left\| \int_{\mathbb{R}^d} u\eta_i^n(\mathrm{d}u)\right\|
\end{aligned}$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

for all n satisfying (6.10), where K_ξ is given by (6.9). Therefore,

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R(\eta_i^n \| \mu(\cdot \mid \hat{X}_i^n))\right] \\
&\quad \le \limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R(\eta_i^n \| \theta_i^n(\cdot \mid \hat{X}_i^n))\right] \\
&\qquad + K_\xi \limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)\sqrt{n}\left\| \int_{\mathbb{R}^d} u\eta_i^n(\mathrm{d}u)\right\|\right]
\end{aligned}$$

and (6.13) follows if

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)\sqrt{n}\left\| \int_{\mathbb{R}^d} u\eta_i^n(\mathrm{d}u)\right\|\right] < \infty. \tag{6.14}$$

Using $\lambda_{DA} \le \lambda$ with λ as in (2.3), the fact that the centering implies that $H_c \ge 0$ and also (6.5), we have, for all n satisfying (6.10) (and, hence, also (6.11)),

$$\begin{aligned}
K_{\mathrm{mgf}} &\ge \max_{0\le i\le\lfloor Tn\rfloor} \sup_{x\in\mathbb{R}^d} \sup_{\|\alpha\|\le\lambda_{DA}/2} \log\left(\int_{\mathbb{R}^d} \mathrm{e}^{\langle \alpha - \bar{\xi}_i^n(\hat{Y}_i^n), u\rangle} \mu(\mathrm{d}u \mid \hat{X}_i^n)\right) \\
&\ge \max_{0\le i\le\lfloor Tn\rfloor} \sup_{x\in\mathbb{R}^d} \sup_{\|\alpha\|\le\lambda_{DA}/2} \log\left(\int_{\mathbb{R}^d} \mathrm{e}^{\langle \alpha - \bar{\xi}_i^n(\hat{Y}_i^n), u\rangle - H_c(\hat{X}_i^n, -\bar{\xi}_i^n(\hat{Y}_i^n))} \mu(\mathrm{d}u \mid \hat{X}_i^n)\right) \\
&= \max_{0\le i\le\lfloor Tn\rfloor} \sup_{x\in\mathbb{R}^d} \sup_{\|\alpha\|\le\lambda_{DA}/2} \log\left(\int_{\mathbb{R}^d} \mathrm{e}^{\langle \alpha, u\rangle} \theta_i^n(\mathrm{d}u \mid \hat{X}_i^n)\right).
\end{aligned}$$

When combined with Lemma 6.2, the last display yields

$$\begin{aligned}
R(\eta_i^n \| \theta_i^n(\cdot \mid \hat{X}_i^n)) &\ge \sup_{\alpha\in\mathbb{R}^d} \left\{\left\langle \alpha, \int_{\mathbb{R}^d} u\eta_i^n(\mathrm{d}u)\right\rangle - \log\left(\int_{\mathbb{R}^d} \mathrm{e}^{\langle \alpha, u\rangle} \theta_i^n(\mathrm{d}u \mid \hat{X}_i^n)\right)\right\} \\
&\ge \sup_{\alpha\in\mathbb{R}^d} \left\{\left\langle \alpha, \int_{\mathbb{R}^d} u\eta_i^n(\mathrm{d}u)\right\rangle - H_c(\hat{X}_i^n, \alpha - \bar{\xi}_i^n(\hat{Y}_i^n))\right\}.
\end{aligned} \tag{6.15}$$

Let e_i denote the ith unit vector. For all n satisfying (6.10) and for all $x, y \in \mathbb{R}^d$ and i,

$$\begin{aligned}
&a(n)^2 n \sup_{\alpha\in\mathbb{R}^d} \{\langle \alpha, \beta\rangle - H_c(x, \alpha - \bar{\xi}_i^n(y))\} \\
&\quad \ge a(n)^2 n \left\{\left\langle \pm\frac{K_\xi}{a(n)\sqrt{n}} e_i, \beta\right\rangle - H_c\left(x, \pm\frac{K_\xi}{a(n)\sqrt{n}} e_i - \bar{\xi}_i^n(y)\right)\right\} \\
&\quad = \pm K_\xi a(n)\sqrt{n}\beta_i - a(n)^2 n H_c\left(x, \pm\frac{K_\xi}{a(n)\sqrt{n}} e_i - \bar{\xi}_i^n(y)\right) \\
&\quad \ge \pm K_\xi a(n)\sqrt{n}\beta_i - K_A K_\xi^2 - \lambda_{DA} K_{DA} K_\xi^2,
\end{aligned}$$

where for the last inequality we used (2.3) and (6.3). Therefore,

$$a(n)^2 n \frac{d}{K_\xi} \sup_{\alpha\in\mathbb{R}^d} \{\langle \alpha, \beta\rangle - H_c(x, \alpha - \bar{\xi}_i^n(y))\} + K_A K_\xi d + \lambda_{DA} K_{DA} K_\xi d \ge a(n)\sqrt{n}\|\beta\|$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

and combining this with (6.15) yields

$$\frac{d}{K_\xi}\mathbb{E}\bigg[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 nR(\eta_i^n\|\theta_i^n(\cdot\mid\hat{X}_i^n))\bigg] + K_A K_\xi d + \lambda_{DA}K_{DA}K_\xi d$$
$$\geq \mathbb{E}\bigg[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)\sqrt{n}\bigg\|\int_{\mathbb{R}^d} u\eta_i^n(\mathrm{d}u)\bigg\|\bigg]$$

for all n satisfying (6.10). Hence, (6.14) and, thus, also (6.13) are valid, which concludes the proof of the lemma. □

6.2. Weak convergence

Let $\hat{\varsigma}^n$ and $\hat{Y}^n$ be associated as in Construction 2.1 with the measures η^n. Lemma 6.3 allows us to apply Theorem 6.1 and choose a (further) subsequence of $\{(\hat{\varsigma}^n, \hat{Y}^n)\}$ (we retain n as the index for convenience) along which $\{(\hat{\varsigma}^n, \hat{Y}^n)\}$ converges weakly to some limit $(\hat{\varsigma}, \hat{Y})$ in $\mathcal{P}([0,T]\times\mathbb{R}^d)\times C([0,T]\colon\mathbb{R}^d)$. We can express relative entropy with respect to the second change of measure (recall (6.5)) as

$$\begin{aligned}
&a(n)^2 nR(\eta_i^n\|\theta_i^n(\cdot\mid\hat{X}_i^n))\\
&\quad= a(n)^2 nR(\eta_i^n\|\mu(\cdot\mid\hat{X}_i^n)) + \bigg\langle a(n)\sqrt{n}\int_{\mathbb{R}^d} u\eta_i^n(\mathrm{d}u), a(n)\sqrt{n}\bar{\xi}_i^n(\hat{Y}_i^n)\bigg\rangle\\
&\qquad + a(n)^2 nH_c(\hat{X}_i^n, -\bar{\xi}_i^n(\hat{Y}_i^n)).
\end{aligned}$$

Thus, we can write the lower bound on the decay rate obtained in (6.12) once again in terms of the original distribution μ:

$$\begin{aligned}
&\mathbb{E}\bigg[2G(\hat{Y}^n(T)) + \frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor}\bigg[a(n)^2 nR(\eta_i^n\|\theta_i^n(\cdot\mid\hat{X}_i^n)) - a(n)^2 nH_c(\hat{X}_i^n, -\bar{\xi}_i^n(\hat{Y}_i^n))\\
&\qquad\qquad - a(n)^2 n\sum_{k=1}^K \rho_k\bigg(\frac{i}{n}, \hat{Y}_i^n\bigg)H_c(\hat{X}_i^n, \xi_i^{n,k}(\hat{Y}_i^n))\bigg]\bigg]\\
&\quad= \mathbb{E}\bigg[G(\hat{Y}^n(T)) + \frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 nR(\eta_i^n\|\mu(\cdot\mid\hat{X}_i^n))\bigg] + \mathbb{E}[G(\hat{Y}^n(T))]\\
&\qquad - \mathbb{E}\bigg[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor}\bigg\langle a(n)\sqrt{n}\int u\eta_i^n(\mathrm{d}u), -a(n)\sqrt{n}\bar{\xi}_i^n(\hat{Y}_i^n)\bigg\rangle\bigg]\\
&\qquad - \mathbb{E}\bigg[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n\sum_{k=1}^K \rho_k\bigg(\frac{i}{n}, \hat{Y}_i^n\bigg)H_c(\hat{X}_i^n, \xi_i^{n,k}(\hat{Y}_i^n))\bigg].
\end{aligned}\tag{6.16}$$

We next use the weak convergence to obtain lower bounds on all terms in (6.12) as now expressed in (6.16). Lemma 6.1 combined with the chain rule for relative entropy [6, Theorem C.3.1] yields (recall that $(\bar{u}^{n,\eta}, \bar{X}^{n,\eta}, \bar{Y}^{n,\eta})$ is defined in terms of η as in Construction 2.1)

$$-a(n)^2\log\mathbb{E}[r^n] = \inf_\eta\bigg\{\mathbb{E}G(\bar{Y}^{n,\eta}(T)) + \frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 nR(\eta_i\|\mu(\cdot\mid\bar{X}_i^{n,\eta}))\bigg\}.$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

For the particular choice $\eta = \eta^n$, this and the fact that r^n is an unbiased estimator yield

$$\liminf_{n\to\infty} \mathbb{E}\left[G(\hat{Y}^n(T)) + \frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R(\eta_i^n \| \mu(\cdot \mid \hat{X}_i^n)) \right] \geq \liminf_{n\to\infty} -a(n)^2 \log \mathbb{E}[r^n]$$
$$\geq V(0,0). \tag{6.17}$$

This gives a lower bound on the first expected value on the right-hand side of (6.16), and we now consider the remaining terms. Because $\hat{Y}^n \to \hat{Y}$ weakly and G is bounded from below and is lower semicontinuous, we can apply the weak convergence version of Fatou's lemma to obtain $\liminf_{n\to\infty} \mathbb{E}[G(\hat{Y}^n(T))] \geq \mathbb{E}[G(\hat{Y}(T))]$. In addition, (6.1) and (6.2) yield (recall also (6.6))

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} \left\langle a(n)\sqrt{n} \int_{\mathbb{R}^d} y \eta_i^n(\mathrm{d}y), -a(n)\sqrt{n}\bar{\xi}_i^n(\hat{Y}_i^n) \right\rangle \right]$$
$$- \mathbb{E}\left[\int_{[0,T]\times\mathbb{R}^d} \left\langle w, D_y U\left(\frac{\lfloor nt\rfloor}{n}, \hat{Y}^n\left(\frac{\lfloor nt\rfloor}{n} \right) \right) \right\rangle \hat{\varsigma}^n(\mathrm{d}w \times \mathrm{d}t) \right]$$
$$\to 0 \quad \text{as } n \to \infty,$$

where

$$D_y U(t,y) = \sum_{k=1}^K \rho_k(t,y) D_y W^k(t,y).$$

Using (6.2), the continuity and boundedness of $D_y U$, and the fact that $(\hat{\varsigma}^n, \hat{Y}^n) \to (\hat{\varsigma}, \hat{Y})$ weakly, it follows that

$$\lim_{n\to\infty} \mathbb{E}\left[\int_{[0,T]\times\mathbb{R}^d} \left\langle w, D_y U\left(\frac{\lfloor nt\rfloor}{n}, \hat{Y}^n\left(\frac{\lfloor nt\rfloor}{n} \right) \right) \right\rangle \hat{\varsigma}^n(\mathrm{d}w \times \mathrm{d}t) \right]$$
$$= \mathbb{E}\left[\int_{[0,T]\times\mathbb{R}^d} \langle w, D_y U(t, \hat{Y}(t)) \rangle \hat{\varsigma}(\mathrm{d}w \times \mathrm{d}t) \right]$$
$$= \mathbb{E}\left[\int_0^T \langle \hat{w}(t), D_y U(t, \hat{Y}(t)) \rangle \, \mathrm{d}t \right].$$

Finally, (6.3), (6.1), the continuity and boundedness of $D_y W^k$, and the weak convergence of $\hat{Y}^n$ to $\hat{Y}$ imply that

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n \sum_{k=1}^K \rho_k\left(\frac{i}{n}, \hat{Y}_i^n \right) H_c(\hat{X}_i^n, \xi_i^{n,k}(\hat{Y}_i^n)) \right]$$
$$= \mathbb{E}\left[\sum_{k=1}^K \int_0^T \rho_k(t, \hat{Y}(t)) \frac{1}{2} \| D_y W^k(t, \hat{Y}(t)) \|^2_{A(X^0(t))} \, \mathrm{d}t \right].$$

Consequently, along this subsequence (recall (6.6))

$$\liminf_{n\to\infty} \left(\mathbb{E}[G(\hat{Y}^n(T))] - \mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} \left\langle a(n)\sqrt{n} \int y \eta_i^n(\mathrm{d}y), a(n)\sqrt{n}\bar{\xi}_i^n(\hat{Y}_i^n) \right\rangle \right] \right.$$
$$\left. - \mathbb{E}\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n \sum_{k=1}^K \rho_k\left(\frac{i}{n}, \hat{Y}_i^n \right) H_c(\hat{X}_i^n, \xi_i^{n,k}(\hat{Y}_i^n)) \right] \right)$$

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

$$\geq \mathbb{E}\Bigg[G(\hat{Y}(T)) - \sum_{k=1}^{K}\int_0^T \rho_k(t,\hat{Y}(t))\Bigg[\langle \hat{w}(t), D_y W^k(t,\hat{Y}(t))\rangle + \frac{1}{2}\|D_y W^k(t,\hat{Y}(t))\|^2_{A(X^0(t))}\Bigg]\mathrm{d}t\Bigg].$$

Since $(U,\rho,\{W^k\}_{k=1}^K) \in \mathcal{MS}$ (recall (3.2)) and U satisfies (2.12), we can continue this inequality as

$$\begin{aligned}
&\geq \mathbb{E}\Bigg[G(\hat{Y}(T)) - \int_0^T \Bigg\langle \hat{w}(t), \sum_{k=1}^{K}\rho_k(t,\hat{Y}(t))D_y W^k(t,\hat{Y}(t))\Bigg\rangle \mathrm{d}t \\
&\qquad - \int_0^T \Bigg\langle \sum_{k=1}^{K}\rho_k(t,\hat{Y}(t))D_y W^k(t,\hat{Y}(t)),\, D_x b(X^0(t))\hat{Y}(t)\Bigg\rangle \mathrm{d}t \\
&\qquad - \int_0^T \sum_{k=1}^{K}\rho_k(t,\hat{Y}(t))W_t^k(t,\hat{Y}(t))\,\mathrm{d}t\Bigg] \\
&= \mathbb{E}\Bigg[G(\hat{Y}(T)) - \int_0^T \frac{\mathrm{d}}{\mathrm{d}t}U(t,\hat{Y}(t))\,\mathrm{d}t\Bigg] \\
&\geq U(0,0).
\end{aligned}$$

Together with (6.17) and (6.16), this yields

$$\liminf_{n\to\infty} -a(n)^2 \log \mathbb{E}[(r^n)^2] + \varepsilon \geq V(0,0) + U(0,0).$$

Since $\varepsilon > 0$ is arbitrary this completes the proof of Theorem 3.1. □

Acknowledgements

P. Dupuis was supported in part by the Department of Energy (grant number DE-SC0010539), the National Science Foundation (grant number DMS-1317199), and the Defense Advanced Research Projects Agency (grant number W911NF-15-2-0122). D. Johnson was supported in part by the Army Research Office (grant number W911NF-14-1-0331) and the Defense Advanced Research Projects Agency (grant number W911NF-15-2-0122).

References

[1] Anderson, B. D. O. and Moore, J. B. (2007). *Optimal Control: Linear Quadratic Methods*. Prentice Hall, Englewood Cliffs, NJ.

[2] Azencott, R. and Ruget, G. (1977). Mélanges d'équations différentielles et grands écarts à la loi des grands nombres. *Z. Wahrscheinlichkeitsth.* **38,** 1–54.

[3] Blanchet, J., Glynn, P. and Leder, K. (2012). On Lyapunov inequalities and subsolutions for efficient importance sampling. *ACM Trans. Model. Comput. Simul.* **22,** 13.

[4] Blanes, S., Casas, F., Oteo, J. A. and Ros, J. (2009). The Magnus expansion and some of its applications. *Phys. Rep.* **470,** 151–238.

[5] Dembo, A. and Zeitouni, O. (1993). *Large Deviations Techniques and Applications*. Jones and Bartlett, Boston, MA.

[6] Dupuis, P. and Ellis, R. S. (1997). *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley, New York.

[7] Dupuis, P. and James, M. R. (1998). Rates of convergence for approximation schemes in optimal control. *SIAM J. Control Optimization* **36,** 719–741.

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press

[8] Dupuis, P. and Johnson, D. (2015). Moderate deviations for recursive stochastic algorithms. *Stoch. Systems* **5,** 87–119.
[9] Dupuis, P. and Wang, H. (2004). Importance sampling, large deviations, and differential games. *Stoch. Stoch. Reports* **76,** 481–508.
[10] Dupuis, P. and Wang, H. (2007). Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Operat. Res.* **32,** 723–757.
[11] Dupuis, P., Leder, K. and Wang, H. (2007). Large deviations and importance sampling for a tandem network with slow-down. *Queuing Systems* **57,** 71–83.
[12] Dupuis, P., Sezer, A. D. and Wang, H. (2007). Dynamic importance sampling for queueing networks. *Ann. Appl. Prob.* **17,** 1306–1346.
[13] Dupuis, P., Spiliopoulos, K. and Wang, H. (2012). Importance sampling for multiscale diffusions. *Multiscale Model. Simul.* **10,** 1–27.
[14] Dupuis, P., Spiliopoulos, K. and Zhou, X. (2015). Escaping from an attractor: importance sampling and rest points I. *Ann. Appl. Prob.* **25,** 2909–2958.
[15] Freidlin, M. I. and Wentzell, A. D. (1984). *Random Perturbations of Dynamical Systems*. Springer, New York.
[16] Glasserman, P. and Wang, Y. (1997). Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Prob.* **7,** 731–746.
[17] Joffe, A. and Métivier, M. (1986). Weak convergence of sequences of semimartingales with applications to multitype branching processes. *Adv. Appl. Prob.* **18,** 20–65.
[18] Johnson, D. (2015). Moderate deviations and subsolution-based importance sampling for recursive stochastic algorithms. Doctoral Thesis, Brown University.
[19] Siegmund, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4,** 673–684.
[20] Ventsel', A. D. (1976). Rough limit theorems on large deviations for Markov stochastic processes. I. *Theory Prob. Appl.* **21,** 227–242.
[21] Ventsel', A. D. (1976). Rough limit theorems on large deviations for Markov stochastic processes. II. *Theory Prob. Appl.* **21,** 499–512.
[22] Ventsel', A. D. (1979). Rough limit theorems on large deviations for Markov stochastic processes. III. *Theory Prob. Appl.* **24,** 675–692.
[23] Ventsel', A. D. (1982). Rough limit theorems on large deviations for Markov stochastic processes. IV. *Theory Prob. Appl.* **27,** 215–234.

https://doi.org/10.1017/apr.2017.31 Published online by Cambridge University Press