Rare Event Simulation with Vanishing Error for Small Noise Diffusions

Eric Vanden-Eijnden[*], Jonathan Weare[*]

August 7, 2009

Abstract

We construct an importance sampling method for certain rare event problems involving small noise diffusions. Standard Monte Carlo schemes for these problems behave exponentially poorly in the small noise limit. Previous work in rare event simulation has focused on developing, in very specific situations, estimators with optimal exponential variance decay rates. This criterion still allows for exponential growth of the statistical relative error. We show that an estimator related to a deterministic control problem not only has an optimal variance decay rate, but can even have vanishingly small statistical relative error in the small noise limit. The method can be seen as the limit of a well known zero variance importance sampling scheme for diffusions which requires the solution of a second order partial differential equation. We test the scheme on several simple examples.

1 Introduction

The simulation of unlikely events and the approximation of their (small) probabilities are important problems which present several difficult computational and mathematical challenges. These problems arise for example in mathematical finance (see e.g. [1, 2]) as well as in computational statistical physics (see e.g. [3]) and in reliability testing during the design of medical or electronic devices (see e.g. [4]), among many other application areas. Unfortunately standard sampling techniques result in statistical errors that explode as the events under consideration become more and more rare.

Previous work in rare event simulation has focused mainly on developing estimators with optimal exponential variance decay rates (see in particular [5, 6]). Indeed in some settings this may be the best possible result. By tradition any estimator with the optimal exponential variance decay rate is called a *log-efficient* estimator. Unfortunately, log-efficient importance sampling estimators are difficult to identify in general settings. Moreover, even log-efficient estimators can have statistical relative error which grows exponentially in the small noise limit.

[1]Courant Institute, NYU, 251 Mercer St, New York, 10012

Estimators with bounded or vanishing relative errors are rare in the literature and do not seem to have been proposed previously in the small noise diffusion setting.

In a series of papers ([6, 7, 8, 9]) Dupuis and Wang have introduced a general framework for analyzing importance sampling estimators in several settings. They suggest several adaptive importance sampling techniques (including an analogue of the one presented here) motivated by the form a certain Hamilton–Jacobi equation. In [7] Dupuis and Wang establish a relationship between smooth subsolutions of this Hamilton–Jacobi equation and the rate of variance decay of associated importance sampling estimators. In this paper we show that a log-efficient estimator for a specific class of expectations (see (1) below) can be explicitly constructed from certain non-smooth (viscosity) solutions of an analogous Hamilton–Jacobi Equation. Moreover we show that this estimator can have vanishing relative error. Our estimator is associated with the solution to a deterministic optimal control problem (and the associated first order Hamilton–Jacobi equation) and is, in some sense, the limit of a family of zero variance estimators. This deterministic optimal control problem can, in principle, be solved on-the-fly without finding the global solution of any partial differential equation.

More precisely, the work in this paper concerns the estimation of quantities of the form

$$\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon)}\right] \tag{1}$$

where g is a suitable functional on $\mathcal{C}\left([0,T]:\mathbb{R}^d\right)$ and X^ϵ is the solution of the stochastic differential equation

$$\begin{aligned} dX^\epsilon(s) &= b(X^\epsilon(s))\,ds + \sqrt{\epsilon}\,\sigma(X^\epsilon(s))\,dW(s), \quad s \in [0,T] \\ X^\epsilon(0) &= x_0 \end{aligned} \tag{2}$$

with W a d-dimensional Brownian motion on some probability space $(\Omega, \mathcal{G}, P)$. An important special case of (1) corresponds to the choice, $g = 0$ if $X^\epsilon \in A \subset \mathcal{C}\left([0,T]:\mathbb{R}^d\right)$ and $g = \infty$ otherwise. In this case (1) becomes

$$P\left(X^\epsilon \in A\right). \tag{3}$$

In (1), (2), and (3), $\epsilon > 0$ is a parameter and we will be interested mainly in situations where $\epsilon \ll 1$, i.e. when the noise amplitude in (2) is small and the functional $e^{-\frac{1}{\epsilon}g(X^\epsilon)}$ is rapidly varying in X^ϵ.

To understand the difficulties presented by the problem, consider the following simple estimator for the expectation in (1)

$$\delta(\epsilon) = \frac{1}{M}\sum_{j=1}^{M} e^{-\frac{1}{\epsilon}g\left(X_j^\epsilon\right)} \tag{4}$$

where each X_j^ϵ is an independent sample of X^ϵ. This estimator is unbiased, i.e.

$$\mathbf{E}\left[\delta(\epsilon)\right] = \mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon)}\right],$$

and its variance is

$$\mathrm{var}\left(\delta(\epsilon)\right) = \frac{1}{M}\left(\mathbf{E}\left[e^{-\frac{2}{\epsilon}g(X^\epsilon)}\right] - \mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon)}\right]^2\right).$$

Therefore, the relative error of this estimator, defined as

$$\rho\left(\delta(\epsilon)\right) = \frac{\sqrt{\mathrm{var}\left(\delta(\epsilon)\right)}}{\mathbf{E}\left[\delta(\epsilon)\right]}, \tag{5}$$

is given by

$$\rho\left(\delta(\epsilon)\right) = \frac{1}{\sqrt{M}}\sqrt{\frac{\mathbf{E}\left[e^{-\frac{2}{\epsilon}g(X^\epsilon)}\right]}{\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon)}\right]^2} - 1}. \tag{6}$$

As one would hope, the error of the estimator decreases with increasing sample size (M). However, Varadhan's lemma (see e.g. [10, 11, 12]) indicates that, under suitable assumptions, we have

$$\lim_{\epsilon\to 0}\epsilon\log\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon)}\right] = -\inf_{\substack{\varphi\in\mathcal{AC}([0,T]),\\ \varphi(0)=x_0}}\left\{I\left(\varphi\right) + g\left(\varphi\right)\right\} := -\gamma_1 \tag{7}$$

and

$$\lim_{\epsilon\to 0}\epsilon\log\mathbf{E}\left[e^{-\frac{2}{\epsilon}g(X^\epsilon)}\right] = -\inf_{\substack{\varphi\in\mathcal{AC}([0,T]),\\ \varphi(0)=x_0}}\left\{I\left(\varphi\right) + 2g\left(\varphi\right)\right\} := -\gamma_2$$

where $\mathcal{AC}\left([0,T]\right)$ is the set of all absolutely continuous functions from $[0,T]$ into $\mathbb{R}^d$. Here $I(\varphi)$ is the rate functional for the process X^ϵ defined as

$$I(\varphi) = \inf_{\substack{u\in L^2([0,T]),\\ \dot\varphi = b+\sigma\, u}}\int_0^T \frac{1}{2}|u(s)|^2 ds \tag{8}$$

for $\varphi\in\mathcal{AC}\left([0,T]\right)$. From Jensen's inequality it is always the case that

$$\gamma_2 \le 2\,\gamma_1. \tag{9}$$

When $\gamma_2 = 2\,\gamma_1$ the simple estimator in (4) is called log-efficient. We will discuss the concept of log-efficiency more below. Often the inequality in (9) is strict and (6) implies that the relative error $\rho\left(\delta(\epsilon)\right)$ increases exponentially for fixed M as $\epsilon\to 0$. In fact, from expression (6),

$$\rho\left(\delta(\epsilon)\right) = \frac{1}{\sqrt{M}}\sqrt{e^{\frac{2\,\gamma_1-\gamma_2+o(1)}{\epsilon}} - 1}.$$

and the relative error can blow up even when $\gamma_2 = 2\,\gamma_1$. In order to control the relative error of the estimator $\delta(\epsilon)$ one must increase the number of samples exponentially as ϵ is decreased.

Importance sampling is a standard technique by which one can attempt to improve the efficiency of Monte Carlo simulations. In the context of diffusions importance sampling is carried out as follows. Suppose that $\mathcal{F}$ is the complete filtration induced by W and consider the distribution Q given by the change of measure

$$\frac{dQ}{dP} = \exp\left(\frac{1}{\sqrt{\epsilon}}\int_0^T \langle U(t), dW(t)\rangle - \frac{1}{2\epsilon}\int_0^T |U(t)|^2 dt\right)$$

where $U(s)$ is an $\mathcal{F}$-progressively measurable process such that the right hand side of this expression is a true martingale. By Girsanov's Theorem, we know that the above changes of measure completely describe the family of distributions which are absolutely continuous with respect to P. If we focus on choices of U which can be written $v(t, X^\epsilon(t))$ for some function v on $[0,T]\times\mathbb{R}^d$, then instead of (4) we can use the following estimator for the expectation (1)

$$\hat{\delta}(\epsilon) = \frac{1}{M}\sum_{j=1}^{M} e^{-\frac{1}{\epsilon}g\left(\hat{X}_j^\epsilon\right)} Z_j \tag{10}$$

where

$$Z_j = \exp\left(-\frac{1}{\sqrt{\epsilon}}\int_0^T \langle v(t, \hat{X}_j^\epsilon(t)), d\hat{W}_j(t)\rangle + \frac{1}{2\epsilon}\int_0^T |v(t, \hat{X}_j^\epsilon(t))|^2 dt\right) \tag{11}$$

and each pair $\left(\hat{W}_j, \hat{X}_j^\epsilon\right)$ is an independent sample of the pair (W, X^ϵ) generated according to the distribution Q with

$$\frac{dQ}{dP} = \exp\left(\frac{1}{\sqrt{\epsilon}}\int_0^T \langle v(t, X^\epsilon(t)), dW(t)\rangle - \frac{1}{2\epsilon}\int_0^T |v(t, X^\epsilon(t))|^2 dt\right)$$

instead of P. Notice that Z_j is the realization of $\frac{dP}{dQ}$ corresponding to the pair $\left(\hat{W}_j, \hat{X}_j^\epsilon\right)$. As we will see in the next section (see Remark 1) generating the required samples is not difficult.

For any reasonable choice of v expression (10) defines an unbiased estimator for (1). The goal of any importance sampling implementation is to replace the estimator (4) by one with smaller variance by making an intelligent choice of v. This is not a trivial task. To see what this entails, note that the relative error of the estimator defined in (10) and (11) is

$$\rho(\hat{\delta}(\epsilon)) = \frac{\sqrt{\mathrm{var}\left(\hat{\delta}(\epsilon)\right)}}{\mathbf{E}\left[\hat{\delta}(\epsilon)\right]} = \frac{1}{\sqrt{M}}\sqrt{\frac{\mathbf{E}_Q\left[e^{-\frac{2}{\epsilon}g(X^\epsilon)}\left(\frac{dP}{dQ}\right)^2\right]}{\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon)}\right]^2} - 1} \tag{12}$$

where $\mathbf{E}_Q$ represents expectation with respect to the measure Q. Therefore, to

control the relative error of $\hat{\delta}(\epsilon)$ one must control the ratio

$$R(\hat{\delta}(\epsilon)) = \frac{\mathbf{E}_Q\left[e^{-\frac{2}{\epsilon}g(X^\epsilon)}\left(\frac{dP}{dQ}\right)^2\right]}{\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon)}\right]^2}. \tag{13}$$

Just as in (9), Jensen's inequality implies that $R(\hat{\delta}(\epsilon)) \geq 1$ and

$$\begin{aligned}\limsup_{\epsilon\to 0} -\epsilon \log \mathbf{E}_Q\left[e^{-\frac{2}{\epsilon}g(X^\epsilon)}\left(\frac{dP}{dQ}\right)^2\right] &\leq 2\lim_{\epsilon\to 0} -\epsilon\log \mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon)}\right] \\ &= 2\,\gamma_1\end{aligned}$$

where γ_1 is defined in expression (7). $2\,\gamma_1$ therefore represents the slowest possible rate of growth of the ratio $R(\hat{\delta}(\epsilon))$ (and of $\rho(\hat{\delta}(\epsilon))$ by (12)). Our first goal is to choose a function v in expressions (10) and (11) so that the resulting importance sampling estimator achieves this minimal rate of error growth. For this reason we use the following standard definition.

Definition 1. *An importance sampling estimator of the form* (10) *is* log-efficient *if*

$$\lim_{\epsilon\to 0} -\epsilon \log R(\hat{\delta}(\epsilon)) = 0. \tag{14}$$

The criterion in (14) is also variously refered to as efficiency and asymptotic efficiency or optimality. Log-efficiency is difficult to establish in all but very specific settings. Moreover, log-efficiency is far from the best result that one might hope for in an estimator. In particular it only implies that

$$\rho(\hat{\delta}(\epsilon)) = \frac{1}{\sqrt{M}}\, e^{o(1)/\epsilon},$$

and does not rule out, for example, that the relative error increases exponentially in $\epsilon^{-\alpha}$ for some $\alpha \in (0,1)$.

Inspection of formula (7) suggests that it may be beneficial to bias paths of X^ϵ to follow the trajectory $\hat{\varphi}$ where

$$I(\hat{\varphi}) + g(\hat{\varphi}) = \gamma_1.$$

If we re-weight the likelihood that X^ϵ follows any particular trajectory $\varphi \in \mathcal{AC}([0,T])$ by the weight $e^{-\frac{1}{\epsilon}g(\varphi)}$ then one can think of $\hat{\varphi}$ as the trajectory that X^ϵ is most likely to follow (when ϵ is small). One might hope then that replacing samples of X^ϵ by samples of, for example, the process Y^ϵ which solves the stochastic differential equation

$$dY^\epsilon(t) = \dot{\hat{\varphi}}(t)\,dt + \sqrt{\epsilon}\,\sigma(Y^\epsilon(t))\,dW(t), \quad \hat{X}^\epsilon(0) = x_0 \tag{15}$$

and reweighting each sample appropriately might produce an estimator with favorable error properties. The resulting importance sampling estimator would correspond to the choice

$$v(t,x) = \sigma^{-1}(t,x)\,(\dot{\hat{\varphi}}(t) - b(t,x)).$$

We will refer to this estimator as the Cramér Tranformation estimator. Unfortunately, as we will see in Section 3, the Cramér Tranformation estimator is not, in general, log-efficient.

As we discuss in the next section, in some cases it is actually possible to find a function $v = v^\epsilon$ depending on ϵ, such that

$$\rho(\hat{\delta}(\epsilon)) = 0$$

for each $\epsilon > 0$, i.e.

$$R(\hat{\delta}(\epsilon)) = 1.$$

The estimator corresponding to v^ϵ is not at all practical, but it gives hope that one might be able to find a more practical choice for v which is independent of ϵ and such that

$$\lim_{\epsilon \to 0} \rho(\hat{\delta}(\epsilon)) = 0$$

which is equivalent to the vanishing error criterion given by the following definition.

Definition 2. *Any estimator satisfying*

$$\lim_{\epsilon \to 0} \rho(\hat{\delta}(\epsilon)) = 0$$

will be called a vanishing error *estimator.*

The theoretical results in this paper will focus on establishing the vanishing error criterion under appropriate conditions. We will also provide numerical evidence that, under more general conditions, our estimator has the following slightly weaker property.

Definition 3. *Any estimator satisfying*

$$\lim_{\epsilon \to 0} \rho(\hat{\delta}(\epsilon)) < \infty$$

will be called a bounded error *estimator.*

For a scheme to have vanishing or bounded error it must not only have the optimal rate of variance decay, but it must also match (or nearly match) the large deviations prefactor, i.e. the $e^{-o(1)/\epsilon}$ term in the large deviations expression

$$\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon)}\right] = e^{-(\gamma_1 + o(1))/\epsilon} \tag{16}$$

which is equivalent to (7). Vanishing and bounded error estimators are the exception in rare event importance sampling schemes. However, recent studies have established bounded error for an analogue of the estimator suggested here in some specific settings (see [13]). Instead of requiring more and more samples as $\epsilon \to 0$ which is the case for most log-efficient estimators (estimators satisfying (14)), these schemes will actually require a constant number or even fewer samples as $\epsilon \to 0$.

The remainder of this paper is organized as follows. In Section 2 we motivate the construction of our estimator and discuss our main results. Our main theorems, Theorem 1 and Theorem 2, which establish the log-efficiency and vanishing error characteristics of our estimator are also first stated in Section 2. Section 3 contains numerical studies on two simple problems. Sections 4 and 5 contain more detailed descriptions of the log-efficiency result and the vanishing error result along with their proofs.

Before concluding this introduction we record the following definitions and assumptions that will be used throughout the paper.

Definition 4. *For any $x \in \mathbb{R}^d$ define the function,*

$$a(x) = \sigma(x)\sigma(x)^T.$$

Definition 5. *For any $x \in \mathbb{R}^d$ the norm $\|\cdot\|_x$ is defined by*

$$\|\beta\|_x = \sqrt{\langle \beta, a(x)^{-1}\,\beta \rangle} \qquad \text{for} \qquad \beta \in \mathbb{R}^d.$$

Assumption 1. *$b(x)$ and $\sigma(x)$ are smooth and bounded and have bounded first derivatives.*

Assumption 2. *There exists an $\eta > 0$ such that for all $x \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^d$,*

$$\langle \beta, a(x)\,\beta \rangle \geq \eta\,|\beta|^2.$$

2 Discussion and statements of main results

As mentioned in the introduction, while log-efficiency is often the best available result in rare event sampling, it still allows for very poor behavior of the relative error as $\epsilon \to 0$. Moreover, log-efficiency is difficult to verify in all but very specific situations. In this section we introduce an estimator which, under appropriate conditions, is not only log-efficient on very general problems but also satisfies the remarkable property that

$$\rho\left(\delta^0(\epsilon)\right) \to 0$$

as $\epsilon \to 0$ (see Theorems 1 and 2). In the interest of clarity, we restrict the discussion in this section and in the rest of the paper to the estimation of expectations of the form $\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon(T))}\right]$ which includes probabilities of the form $P\left(X^\epsilon(T) \notin D\right)$ for an open domain with $x_0 \in D$. However, the results extend (in a modified form) to more general cases such as path dependent functionals of the form

$$g\left(X^\epsilon\right) = \int_0^{\tau_D \wedge T} l(t, X^\epsilon(t))\,dt + \psi(X^\epsilon(\tau_D \wedge T)) \tag{17}$$

where τ_D is the first escape time of the process from some domain D.

2.1 Motivation and description of the method

We begin by establishing a formal connection with a zero variance estimator. Let $X^\epsilon_{t,x}$ be the solution of the stochastic differential equation

$$dX^\epsilon_{t,x}(s) = b(X^\epsilon_{t,x}(s))\,ds + \sqrt{\epsilon}\,\sigma(X^\epsilon_{t,x}(s))\,dW(s), \qquad X^\epsilon_{t,x}(t) = x \tag{18}$$

(note that the subscript t denotes the initial time and not the value of the time parameter as is sometimes the case in the stochastic process literature). Define the function

$$\Phi^\epsilon(t,x) = \mathbf{E}\left[e^{-\frac{1}{\epsilon}g\left(X^\epsilon_{t,x}(T)\right)}\right].$$

Our goal as stated in the introduction is to construct an importance sampling estimator of $\Phi^\epsilon(0,x_0)$. It will prove useful to expand this goal to any possible initial condition, i.e. to the estimation of $\Phi^\epsilon(t,x)$ for any particular $(t,x) \in [0,T]\times\mathbb{R}^d$. It is well known that, for each $\epsilon > 0$, a zero variance estimator of Φ is available. Indeed a simple application of Ito's formula shows that if

$$v^\epsilon := -\epsilon\frac{\sigma^T\, D_x\Phi^\epsilon}{\Phi^\epsilon} \tag{19}$$

then $Q^\epsilon_{t,x}$-almost surely we have

$$e^{-\frac{1}{\epsilon}g\left(X^\epsilon_{t,x}(T)\right)}\frac{dP}{dQ^\epsilon_{t,x}} = \Phi^\epsilon(t,x)$$

where

$$\frac{dQ^\epsilon_{t,x}}{dP} = \exp\left(-\frac{1}{\sqrt{\epsilon}}\int_t^T \langle v^\epsilon(s,X^\epsilon_{t,x}(s)), dW(s)\rangle + \frac{1}{2\epsilon}\int_t^T |v^\epsilon(s,X^\epsilon_{t,x}(s))|^2 ds\right). \tag{20}$$

This change of measure is sometimes called the Doob h-transform. An obvious strategy is to construct an estimator of the form (10) using $v = v^\epsilon$, i.e. to construct the estimator

$$\delta_{t,x}(\epsilon) = \frac{1}{M}\sum_{j=1}^M e^{-\frac{1}{\epsilon}g\left(\hat{X}^\epsilon_j\right)} Z_j \tag{21}$$

where

$$Z_j = \exp\left(-\frac{1}{\sqrt{\epsilon}}\int_t^T \langle v^\epsilon(s,\hat{X}^\epsilon_j(s)), d\hat{W}_j(s)\rangle + \frac{1}{2\epsilon}\int_t^T |v^\epsilon(s,\hat{X}^\epsilon_j(s))|^2 ds\right) \tag{22}$$

and each pair $\left(\hat{W}_j, \hat{X}^\epsilon_j\right)$ is an independent sample of the pair $\left(W, X^\epsilon_{t,x}\right)$ generated according to the distribution $Q^\epsilon_{t,x}$. From (20) each sample is almost surely equal to $\Phi^\epsilon(t,x)$ and the resulting estimator has zero variance and therefore zero relative error.

The function Φ^ϵ satisfies the backward Kolmogorov equation

$$\partial_t \Phi^\epsilon + \langle b, D_x \Phi^\epsilon \rangle + \frac{\epsilon}{2} \operatorname{tr} a \, D_x^2 \Phi^\epsilon = 0. \tag{23}$$

One could, therefore, attempt to discretize and solve the partial differential equation (23) and use the estimator of the form above to approximate $\mathbf{E}\left[e^{-\frac{1}{\epsilon}g\left(X^\epsilon_{t,x}(T)\right)}\right]$. While in low dimensions these steps can be carried out, this strategy is not practical since if we can solve for Φ^ϵ we have our answer and Monte Carlo is not necessary.

These ideas, however, can be modified and put to use in high dimensions. Notice that for each $\epsilon > 0$ the function

$$G^\epsilon = -\epsilon \log \Phi^\epsilon \tag{24}$$

solves the second order Hamilton–Jacobi equation

$$-\partial_t G^\epsilon + H(x, D_x G^\epsilon(t,x)) - \frac{\epsilon}{2} \operatorname{tr} a \, D_x^2 G^\epsilon = 0 \tag{25}$$

with terminal condition $G^\epsilon(T,x) = g(x)$ where

$$H(x,p) = -\langle b(x), p \rangle + \frac{1}{2} |\sigma(x)^T \, p|^2. \tag{26}$$

In terms of G^ϵ, the zero variance estimator above corresponds to the choice

$$v^\epsilon = -\sigma^T \, D_x G^\epsilon. \tag{27}$$

It is natural to replace G^ϵ by its zero viscosity approximation G, i.e. by the viscosity solution to the first order Hamilton–Jacobi equation,

$$-\partial_t G + H(x, D_x G(t,x)) = 0 \tag{28}$$

with terminal condition $G(T,x) = g(x)$, and to use the function

$$v^0 := -\sigma^T \, D_x G \tag{29}$$

in place of v^ϵ in (21) to define an importance sampling estimator. For any particular initial point $(t,x) \in [0,T] \times \mathbb{R}^d$ we will call the resulting estimator $\delta^0_{t,x}(\epsilon)$. At times we will need to consider the properties of our estimator over sets of initial conditions. For this reason we will often refer to our estimator as $\delta^0(\epsilon)$ without arguments in the subscript.

The expression, (29), for v^0 immediately raises an important practical issue. Evaluating the function v^0 in principle requires solving the partial differential equation (28) to find G and then differentiating to find $D_x G$. In more than a few dimensions, finding the global solution of (28) is no more practical than is finding G^ϵ the solution of (25). This difficulty can be avoided by taking advantage of an optimal control representation of G. Indeed, equations of the form (28) are

typically associated with an appropriate optimal control problem. Under very general conditions the unique uniformly continuous viscosity solution of (28) is

$$G(t,x) = \inf_{\substack{\varphi\in\mathcal{AC}([t,T]):\\ \varphi(t)=x}} \left\{ \int_t^T \frac{1}{2}\|\dot\varphi(s) - b(\varphi(s))\|^2_{\varphi(s)}\, ds + g\left(\varphi(T)\right) \right\}. \tag{30}$$

Indeed if the expression on the right hand side of (30) is continuous and if a minimizer $\hat\varphi_{t,x} \in ([t,T])$ exists for every $(t,x) \in [0,T]\times\mathbb{R}^d$, then by Theorem II.7.2 in [14] relation (30) holds. The following two definitions will be useful.

Definition 6. *We will call a function* $\hat\varphi_{t,x} \in \mathcal{AC}([t,T])$ *an* optimal control trajectory *at the point* (t,x) *if*

$$\hat\varphi_{t,x} \in \arg\inf_{\substack{\varphi\in\mathcal{AC}([t,T]):\\ \varphi(t)=x}} \left\{ \int_t^T \frac{1}{2}\|\dot\varphi(s) - b(\varphi(s))\|^2_{\varphi(s)}\, ds + g\left(\varphi(T)\right) \right\} \tag{31}$$

i.e.

$$G(t,x) = \int_t^T \frac{1}{2}\|\dot{\hat\varphi}_{t,x}(s) - b(\hat\varphi_{t,x}(s))\|^2_{\hat\varphi_{t,x}(s)}\, ds + g\left(\hat\varphi_{t,x}(T)\right). \tag{32}$$

Definition 7. *A point* (t,x) *is called a* regular point *if there is a unique* $\hat\varphi_{t,x} \in \mathcal{AC}([t,T])$ *satisfying* (31).

At any regular point there is a convenient representation of v^0 in terms of the unique optimal control. The statement of the relevant proposition is as follows.

Proposition 1. *If* $(t,x) \in [0,T]\times\mathbb{R}^d$ *is a regular point then the unique optimal control trajectory at* (t,x), $\hat\varphi_{t,x}$, *solves the ordinary differential equation*

$$\dot{\hat\varphi}_{t,x}(s) = b(\hat\varphi_{t,x}(s)) + \sigma(\hat\varphi_{t,x}(s))\, v^0(s,\hat\varphi_{t,x}(s)), \qquad \textit{for a.e. } s\in[t,T].$$

The proof of Proposition 1 uses a standard control theory argument and is therefore not included. When the optimal control $\hat\varphi_{t,x}$ has a continuous derivative Proposition 1 implies that

$$v^0(t,x) = \sigma(x)^{-1}\left(\dot{\hat\varphi}_{t,x}(t) - b(x)\right).$$

This alternative description allows one to evaluate v^0 "on-the-fly" only at those points where it is needed. As we evolve each sample we will need to know $v^0(s,\hat X^\epsilon_j(s))$. Instead of precomputing G everywhere we solve an optimization problem at the single point $(s,\hat X^\epsilon_j(s))$ to find the $\hat\varphi_{s,\hat X^\epsilon_j(s)} \in \mathcal{AC}([s,T])$ satisfying (31). An analogous computational strategy was suggested, in a different setting, in [6] and [8]. In our setting, the approximation of the optimal trajectories can be accomplished as in [15]. While this procedure has to be carried out for each $s\in[0,T]$, if the optimal control trajectory varies slowly along the path of $(s,\hat X^\epsilon_j(s))$ the computation can be accelerated by continuation. We

will investigate computational issues in future work. In this paper we focus on the theoretical aspects of the resulting estimator. Before we move on to a discussion of those theoretical issues we briefly mention how our importance sampling estimator is implemented.

Remark 1. *If W is a Brownian motion under P then the process $\hat{W} = W - \frac{1}{\sqrt{\epsilon}} \int_t^{\cdot} v^0(s, X^\epsilon_{t,x}(s))\, ds$ is a Brownian motion under $Q^0_{t,x}$, where*

$$\frac{dQ^0_{t,x}}{dP} = \exp\left(-\frac{1}{\sqrt{\epsilon}} \int_t^T \langle v^0(s, X^\epsilon_{t,x}(s)), dW(s)\rangle + \frac{1}{2\epsilon} \int_t^T |v^0(s, X^\epsilon_{t,x}(s))|^2 ds \right). \tag{33}$$

In terms of $\hat{W}$, the process $X^\epsilon_{t,x}$ solves the stochastic differential equation

$$dX^\epsilon_{t,x}(s) = (b(X^\epsilon_{t,x}(s)) + \sigma(X^\epsilon_{t,x}(s))\, v^0(s, X^\epsilon_{t,x}(s)))\, ds + \sqrt{\epsilon}\, \sigma(X^\epsilon_{t,x}(s))\, d\hat{W}(s). \tag{34}$$

This implies that the pair $(W, X^\epsilon_{t,x})$ has the same distribution under $Q^0_{t,x}$ as has the pair $\left(W + \frac{1}{\sqrt{\epsilon}} \int_t^{\cdot} v^0(s, \hat{X}^\epsilon_{t,x}(s))\, ds, \hat{X}^\epsilon_{t,x}\right)$ under P where $\hat{X}^\epsilon_{t,x}$ is the strong solution of the equation

$$d\hat{X}^\epsilon_{t,x}(s) = (b(\hat{X}^\epsilon_{t,x}(s)) + \sigma(\hat{X}^\epsilon_{t,x}(s))\, v^0(s, \hat{X}^\epsilon(s)))\, ds + \sqrt{\epsilon}\, \sigma(\hat{X}^\epsilon_{t,x}(s))\, dW(s). \tag{35}$$

Thus we can define an estimator with the same distribution as $\delta^0_{t,x}(\epsilon)$ by replacing each sample $(\hat{W}_j, \hat{X}_j)$ of $(W, X^\epsilon_{t,x})$ generated under $Q^0_{t,x}$ by an independent sample of $\left(W + \frac{1}{\sqrt{\epsilon}} \int_t^{\cdot} v^0(s, \hat{X}^\epsilon_{t,x}(s))\, ds, \hat{X}^\epsilon_{t,x}\right)$ generated under P. In fact we will define $\delta^0_{t,x}(\epsilon)$ by

$$\delta^0_{t,x}(\epsilon) = \frac{1}{M} \sum_{j=1}^M e^{-\frac{1}{\epsilon} g\left(\hat{X}^\epsilon_j\right)} Z_j \tag{36}$$

where now

$$Z_j = \exp\left(-\frac{1}{\sqrt{\epsilon}} \int_t^T \langle v^0(s, \hat{X}^\epsilon_j(s)), d\hat{W}_j(s)\rangle - \frac{1}{2\epsilon} \int_t^T |v^0(s, \hat{X}^\epsilon_j(s))|^2 ds \right) \tag{37}$$

and each pair $\left(\hat{W}_j, \hat{X}^\epsilon_j\right)$ is an independent sample of the pair $\left(W, X^\epsilon_{t,x}\right)$ generated according to the distribution P.

The relationship given by Proposition 1, between the optimal control $\hat{\varphi}_{t,x}$ defined in (31) and the function $v^0 = -\sigma^T D_x G$, implies that equation (35) can be written as

$$d\hat{X}^\epsilon_{t,x}(s) = \dot{\hat{\varphi}}_{s, \hat{X}^\epsilon_{t,x}(s)}(s)\, ds + \sqrt{\epsilon}\, \sigma(\hat{X}^\epsilon_{t,x}(s))\, dW(s). \tag{38}$$

This equation should be compared with expression (15) corresponding to the Cramér Tranformation estimator. Equation (38) requires finding the optimal control trajectory at each point along a path of $\hat{X}^{\epsilon}_{t,x}$, while equation (15) only requires finding one initial optimal control trajectory. As we will see in Section 3, this appearant computational advantage of the Cramér Tranformation estimator can be more than negated by its poor performance on some generic problems.

2.2 Main results.

As mentioned in the introduction, previous work on importance sampling schemes for rare event problems has focused on developing log-efficient schemes in specific cases. In Theorems 1 and 2 below we will give conditions under which our estimator is not only log-efficient, but also has vanishing error. In general we cannot expect the viscosity solution of (28) to be continuously differentiable so our control function $v^0 = -\sigma^T D_x G$ may not be continuous. It is important, therefore, that our conditions include some cases in which v^0 is not continuous. The non-smooth behavior of v^0 is the major difficulty that has to be overcome in the proofs of both Theorems 1 and 2 below. The statement of Theorem 1 reflects this difficulty in that we are forced to be very specific about the possible discontinuities allowed for v^0 (see Assumption 3 in Section 4). Extending Theorem 1 to allow more general discontinuities would have to be accomplished on a case by case basis since the many possible discontinuities all require slightly different treatment. To reflect our belief that log-efficiency holds under much weaker conditions, Assumption 3 is replaced in the statement of Theorem 2 by the more general requirement that our estimator satisfies a form of log-efficiency (see Definition 8).

In slightly more generality than in (7), the Laplace Principle for $X^{\epsilon}_{t,x}$ is

$$\lim_{\epsilon \to 0} -\epsilon \log \mathbf{E}\left[e^{-\frac{1}{\epsilon} g(X^{\epsilon}_{t,x})}\right] = \inf_{\substack{\varphi \in \mathcal{AC}([t,T]) \\ \varphi(t)=x}} \left\{I_t(\varphi) + g(\varphi(T))\right\}$$

where we have extended the definition of the rate function in (8) to

$$I_t(\varphi) = \inf_{\substack{u \in L^2([t,T]): \\ \dot{\varphi} = b + \sigma u}} \int_t^T \frac{1}{2}|u(s)|^2 ds \tag{39}$$

for any $t \in [0,T]$. Under our assumptions,

$$I_t(\varphi) = \frac{1}{2}\int_t^T \|\dot{\varphi}(s) - b(\varphi(s))\|^2_{\varphi(s)}\, ds$$

and so

$$\inf_{\varphi \in \mathcal{AC}([t,T])} \left\{I_t(\varphi) + g(\varphi(T))\right\} = G(t,x).$$

Thus the Laplace Principle for $X^{\epsilon}_{t,x}$ can be written,

$$\lim_{\epsilon \to 0} -\epsilon \log \mathbf{E}\left[e^{-\frac{1}{\epsilon} g(X^{\epsilon}_{t,x})}\right] = G(t,x). \tag{40}$$

Recalling Definition 1, the importance sampling estimator corresponding to the change of measure

$$\frac{dQ}{dP} = \exp\left(-\frac{1}{\sqrt{\epsilon}}\int_t^T \langle v(s, X^\epsilon_{t,x}(s)), dW(s)\rangle + \frac{1}{2\epsilon}\int_t^T |v(s, X^\epsilon_{t,x}(s))|^2 ds\right). \quad (41)$$

is log-efficient at the point $(t,x) \in [0,T]\times\mathbb{R}^d$ when

$$\liminf_{\epsilon\to 0} V^\epsilon(t,x) = 2\,G(t,x) \quad (42)$$

where we have defined the function

$$V^\epsilon(t,x) = -\epsilon \log \mathbf{E}_Q\left[e^{-\frac{2}{\epsilon}g\left(X^\epsilon_{t,x}(T)\right)}\left(\frac{dP}{dQ}\right)^2\right]. \quad (43)$$

In Theorem 2 below we will show that when $\delta^0(\epsilon)$ satisfies a slightly stronger form of log-efficiency, uniform log-efficiency, then the relative error can actually vanish as $\epsilon \to 0$. Uniform log-efficiency is defined as follows.

Definition 8. *The importance sampling estimator corresponding to the change of measure in* (41) *is* uniformly log-efficient *if*

$$\lim_{\epsilon\to 0}\sup_{(t,x)\in K} |V^\epsilon(t,x) - 2\,G(t,x)| = 0 \quad (44)$$

for any compact set $K \subset [0,T]\times\mathbb{R}^d$.

In other words an importance sampling scheme is uniformly log-efficient if the function V^ϵ defined in (43) converges to $2\,G$ uniformly on compact subsets of $[0,T]\times\mathbb{R}^d$. If an estimator is uniformly log-efficient then it is log-efficient at any point $(t,x) \in [0,T]\times\mathbb{R}^d$. An arbitrary choice of the function v in (41) will not result in a uniformly efficient or even efficient importance sampling scheme. The next theorem asserts that if $v = v^0$ then the resulting scheme is uniformly efficient. As already mentioned, in the statement of Theorem 1 we make specific assumptions about the form of the possible discontinuities of $v^0 = -\sigma^T D_x G$. The precise statement of the result is as follows.

Theorem 1. *Assume that* $v^0 = -\sigma^T D_x G$ *satisfies Assumption 3 in Section 4. Then the estimator* $\delta^0(\epsilon)$ *corresponding to the choice* $v = v^0$ *in* (41) *is uniformly log-efficient.*

The proof of Theorem 1 proceeds in two steps. First we identify a function $V(t,x)$ (see equation (66)) which is the uniform limit on compact sets of V^ϵ and then we show that $V = 2\,G$. The identification of V is essentially a Laplace Principle type result and is the subject of Section 4.1. In our case we will see that what is needed is a Laplace Principle for a discontinuous exponential functional. Results establishing Laplace Principles with discontinuous functionals are rare in the literature. We will apply the weak convergence arguments used to prove

the Laplace Principle for diffusions with discontinuous drift in [16]. Section 4.2 contains the proof that $V = 2\,G$, the second step in our proof of Theorem 1. The function V is defined by an optimal control type variational problem with a discontinuous running cost. This discontinuity requires a slight modification of the standard verification argument.

With uniform log-efficiency established by Theorem 1 it is relatively easy to prove that we can expect much more of the estimator $\delta^0(\epsilon)$ at least in special regions. In Theorem 2 we show that our estimator has vanishing relative error in regions where G is sufficiently smooth. As mentioned after Definitions 2 and 3 this result establishes that our estimator accurately captures the $e^{-o(1)/\epsilon}$ prefactor in the expression (equivalent to the Laplace Principle in (40))

$$\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^{\epsilon}_{t,x}(T))}\right] = e^{-(G(t,x)+o(1))/\epsilon}$$

The precise statement of the theorem is as follows.

Theorem 2. *Let (t,x) be contained in a region of strong regularity $N \subset [0,T]\times\mathbb{R}^d$. Suppose that G is smooth on $\bar{N}$ and that $\delta^0(\epsilon)$ is uniformly log-efficient. Then the relative error $\rho(\delta^0_{t,x}(\epsilon))$ defined in (12) satisfies*

$$\lim_{\epsilon\to 0}\rho(\delta^0_{t,x}(\epsilon)) = 0.$$

The concept of a region of strong regularity will be defined in Section 5. In particular, on any region of strong regularity G solves the Hamilton–Jacobi equation (28) in the classical sense. As established by the following proposition, when g is sufficiently smooth these sets comprise almost all of space and the requirement in the statement of Theorem 2 that (t,x) be contained in a region of strong regularity is hardly restrictive.

Proposition 2. *Suppose that g is smooth. Then the set of points which are contained in a region of strong regularity has full Hausdorff Dimension.*

Even with Proposition 2 in hand, it is comforting to know that while there may be points where our scheme does not have vanishing error it will be log-efficient everywhere (at least under the assumptions of Theorem 1).

Theorem 2 may seem natural given that, at least informally, the estimator $\delta^0(\epsilon)$ is the limit of the zero variance estimation scheme described in the beginning of this section. However, v^0 is, at least formally, the limit of the log transformation $-\epsilon\log\Phi^\epsilon$ of Φ^ϵ. One might expect, therefore, that all subexponential information about Φ^ϵ should be lost in this limit. It is somewhat surprising that the estimator can be more than just log-efficient. It is also highly unusual to find a generally applicable log-efficient scheme, let alone one that has vanishing or bounded relative error. As mentioned earlier, even a log-efficient scheme can require more and more samples as $\epsilon\to 0$ our estimator can actually requires fewer samples (under appropriate conditions) as $\epsilon\to 0$. In the next section we will see that even some educated choices of importance sampling measures (Q) can lead to disastrous results.

The restriction on the possible discontinuities of v^0 in Theorem 1 (in particular Assumption 3) excludes many interesting cases. Once uniform log-efficiency has been established as assumed in Theorem 2 those discontinuities become much less of an issue. What remains an issue, however, is the boundedness of v^0. For many important problems the function $v^0(t,x) = -\sigma^T(t,x)\, D_x G(t,x)$ will blow up as t approaches the terminal time T. In particular our requirement that v^0 be bounded essentially constricts us (with some exceptions) to cases in which

$$\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon_{t,x}(T))}\right] = e^{-\frac{1}{\epsilon}G(t,x)}(C + o(1)) \tag{45}$$

for some constant C, i.e. to leading order the Laplace Principle prefactor does not depend on ϵ (see the discussion around expression (16)).

The behavior in (45) is in contrast, for example, to the case that $g = \infty$ on D and $g = 0$ on D^c for some domain D. Since estimating probabilities of the form

$$P\left(X^\epsilon_{t,x}(T) \notin D\right) \tag{46}$$

is an important problem, it useful to try to extend Theorem 2 to cover this case. Even for probabilities of the form in (46) there are many possible prefactors depending on the geometric properties of the set D. As for the possible discontinuities of v^0, each of the different cases requires slightly different treatment. Furthermore, examination of the proof and numerical experiments suggest that vanishing error may be too much to hope for here. However there is numerical and theoretical evidence that, at least when the boundary of D is smooth, our estimator may have asymptotically bounded relative error, i.e.

$$\limsup_{\epsilon\to 0} \rho\left(\delta^0_{t,x}(\epsilon)\right) < \infty. \tag{47}$$

We will consider this extension in future work.

3 Illustrative examples

In this section we consider a couple of simple example problems to illustrate some interesting features of our results. Let X^ϵ be the solution of the simple stochastic differential equation

$$dX^\epsilon(t) = \sqrt{\epsilon}\, dW(t), \qquad X^\epsilon(0) = 0.1, \tag{48}$$

for $0 \le t \le 1$ ($T = 1$), where W is a Brownian motion in $\mathbb{R}$. In the first of the following subsections we estimate $\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon(1))}\right]$ for a continuous function g. We will find that, consistent with the conclusion of Theorem 2, our estimator appears to have vanishing error. We will also discuss our assumptions further in connection with the example. In the second subsection we will consider the probability $P\left(X^\epsilon(1) \notin D\right)$ that $X^\epsilon(1)$ is outside of a domain D given that $X^\epsilon(0) \in D$. As already mentioned, while this case is not contained in Theorem 2 there is some theoretical evidence that our estimator might have bounded

error. We will provide numerical evidence supporting this assertion. We will also compare our estimator with the standard Monte Carlo estimator $\delta(\epsilon)$ described in the introduction as well as an estimator suggested by the proof of the Large Deviations Principle (the Cramér Tranformation estimator) which has been suggested and studied in several publications.

As described in the previous section, the estimator presented in this paper requires finding the minimizing trajectory for the variational problem in (30) at each point along any sample trajectory (see Remark 1 and equation (38)). In both of the examples below the function g for which we wish to approximate $\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon(1))}\right]$ is chosen so that whenever $x = 0$, there are two distinct minimizers in (30). As any sample trajectory $X^\epsilon(s)$ passes across $\{x = 0\}$, the control $v^0(s, X^\epsilon(s))$ changes discontinuously from one of the two minimizers to the other. If instead one were to use a local minimizer (not the global minimizer) when evaluating v^0, the resulting method would not only fail to be log-efficient, but would have exploding variance. As described below, the Cramér Tranformation estimator uses the initial optimal control (computed at $t = 0$ and $x = 0$) at all times resulting in dramatically poor performance.

3.1 Estimating $\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon(1))}\right]$.

Our first example is chosen so that the control function $v^0 = -\sigma^T D_x G$ is smooth everywhere away from a discontinuity across the set $[0, T] \times \{x = 0\}$. We will check that this discontinuity is consistent with the assumptions of Theorems 1 and 2.

Let g be defined as follows

$$g(x) = \begin{cases} \frac{1}{2}(1-x)^2, & x > 0, \\ \frac{1}{2}(1+x)^2, & x < 0. \end{cases}$$

In this simple case we can exactly solve the variational problem (30) which defines G and see that

$$G(t,x) = \begin{cases} \frac{(1-x)^2}{2(2-t)}, & x > 0, \\ \frac{(1+x)^2}{2(2-t)}, & x < 0. \end{cases}$$

which implies that

$$v^0(t,x) = -\sigma(x)^T D_x G(t,x) = \begin{cases} \frac{1-x}{2-t}, & x > 0, \\ \frac{-1-x}{2-t}, & x < 0. \end{cases}$$

To tie this discussion to that in Section 4 we will use the notation

$$\mathbb{H}^+ = \{x \in \mathbb{R} : x > 0\}, \qquad \mathbb{H}^- = \{x \in \mathbb{R} : x < 0\},$$

and

$$\Lambda = \{0\}\,.$$

The function v^0 is smooth everywhere away from the set $[0,T]\times\Lambda$. It is also clear that Assumption 3, required in Theorem 1 is satisfied, i.e. the restrictions of v^0 to the sets $[0,1]\times\mathbb{H}^+$ and $[0,1]\times\mathbb{H}^-$ have the smooth extensions to all of $[0,1]\times\mathbb{R}^d$,

$$v^+(t,x) = \frac{1-x}{2-t}$$

and

$$v^-(t,x) = \frac{-1-x}{2-t}$$

respectively. The extentions v^+ and v^- are not bounded as required by Assumption 3, but smooth growth of the kind they exhibit is not a serious problem for the theory in Sections 4 and 5. Therefore Theorem 1 assures us that our estimator will be uniformly log-efficient.

Next we check the assumptions of Theorem 2 to see at which points we can expect to see vanishing error. It is easy to see that any set of the form $[0,1]\times\{x\in\mathbb{R}:\,x>\eta\}$ or $[0,1]\times\{x\in\mathbb{R}:\,x<\eta\}$ for $\eta>0$ is a region of strong regularity. Moreover G is smooth on these sets. Thus any point in either $[0,1]\times\mathbb{H}^+$ or $[0,1]\times\mathbb{H}^-$ satisfies the assumptions of Theorem 2 and our estimator $\delta^0(\epsilon)$ (the one corresponding to setting $v=v^0$ in (10) and (11)) is guaranteed to have vanishing error everywhere except on the set $[0,1]\times\Lambda$ where it will still be log-efficient.

To better appreciate the strength of these results we will consider the related Cramér Tranformation estimator. At the point $(0,X^\epsilon(0)) = (0,0.1)$, the unique minimizer in expression (30) is the function

$$\hat{\varphi}_{0,0.1}(s) = 0.1 + 0.45\,s.$$

The Large Deviations Theory essentially says that if the likelihood of each possible trajectory $\varphi\in\mathcal{AC}([0,1])$ is reweighted by $e^{-\frac{1}{\epsilon}g(\varphi(1))}$ then the control trajectory $\hat{\varphi}_{0,0.1}$ is the most likely to be close to the trajectory of X^ϵ. Thus one might hope that an importance sampling estimator in which one samples the process

$$Y^\epsilon(s) = \hat{\varphi}_{0,0.1}(s) + \sqrt{\epsilon}\,W(s) \tag{49}$$

would have small variance. The Cramér Tranformation estimator, which we denote by $\bar{\delta}_{0,0.1}(\epsilon)$, corresponds to setting

$$v \equiv \dot{\hat{\varphi}}_{0,0.1} = 0.45$$

in expressions (10) and (11).

Analogues of the Cramér Tranformation estimator have been has been studied by several authors (see [17],[2]). For more complicated problems in which one cannot find the minimizer in (30) exactly the estimator $\bar{\delta}_{0,0.1}(\epsilon)$ can be much cheaper to generate than our estimator because the optimal control $\hat{\varphi}_{0,0.1}$

ϵ	$R(\delta_{0,0.1}(\epsilon))$	$R(\bar{\delta}_{0,0.1}(\epsilon))$	$R(\delta^0_{0,0.1}(\epsilon))$	$\delta^0_{0,0.1}(\epsilon)$	exact value
1	1.0340	1.2564	1.1746	0.8368	0.8369
2^{-1}	1.0800	1.6636	1.3494	0.7225	0.7227
2^{-2}	1.3084	3.5982	1.6971	0.4848	0.4852
2^{-3}	2.2672	25.526	2.2903	0.1983	0.1986
2^{-4}	7.7807	977.66	2.5990	0.3316×10^{-1}	0.3323×10^{-1}
2^{-5}	81.266	–	1.5193	0.1127×10^{-2}	0.1129×10^{-2}
2^{-6}	6008.4	–	1.0200	0.1666×10^{-5}	0.1666×10^{-5}

Table 1: Comparison of ratios defined in (13) for each estimator. The ratio, and therefore the relative error, is much smaller for $\delta^0_{0,0.1}(\epsilon)$ than for the other two estimators. Notice also that the ratio $R(\delta^0_{0,0.1}(\epsilon))$ appears to converge to 1 as $\epsilon \to 0$ (and, therefore, the relative error appears to converges to 0).

can be precomputed once while generating $\delta^0_{0,0.1}(\epsilon)$ requires solving a separate optimization problem to evaluate v^0 at all points along each sample trajectory. However, the performance of $\bar{\delta}_{0,0.1}(\epsilon)$ can be catastrophic on generic problems while Theorems 1 and 2 establish the favorable performance of $\delta^0_{0,0.1}(\epsilon)$ in significant generality.

In Table 1 we compare the performance of our importance sampling estimator $\delta^0_{0,0.1}(\epsilon)$ with the unweighted estimator $\delta_{0,0.1}(\epsilon)$ (see (4)) and the Cramér Tranformation estimator, $\bar{\delta}_{0,0.1}(\epsilon)$, just described. The error results are reported in terms of the ratio R defined in (13) which has no explicit dependence on the number of samples M. Recall from formula (12) that in terms of R, the relative error is given by

$$\rho = \frac{\sqrt{R-1}}{\sqrt{M}}$$

so that when R is close to 1 the relative error is small. In all of our experiments in this subsection and the next the time step used to evolve $\hat{X}^\epsilon$ (see Remark 1) is 10^{-3}. This small time step is chosen to reduce discretization effects. All of our results have been checked with up to 10^9 samples to verify that the reported values of R are stable.

The performance of $\delta^0_{0,0.1}(\epsilon)$ far surpasses that of the other two estimators. Notice that despite its intuitive appeal $\bar{\delta}_{0,0.1}(\epsilon)$ behaves far worse than even the standard estimator $\delta_{0,0.1}(\epsilon)$. A simple calculation shows that

$$\text{var}\left(\bar{\delta}_{0,0.1}(\epsilon)\right) = e^{0.79/\epsilon + o(1)}.$$

This poor behavior is due to the large importance weights corresponding to samples of $Y^\epsilon(1)$ which are close to -1. Such samples are exponentially unlikely but their importance weights are large enough to cause the variance to grow exponentially. In contrast, by Remark 1 generating an estimator with the same distribution as $\delta^0_{0,0.1}(\epsilon)$ requires sampling the process $\hat{X}^\epsilon$ which solves

$$d\hat{X}^\epsilon(s) = v^0(s, \hat{X}^\epsilon(s))\, ds + \sqrt{\epsilon}\, dW(s), \quad \hat{X}^\epsilon(0) = 0.1.$$

ϵ	$R(\delta_{0,0}(\epsilon))$	$R(\bar{\delta}_{0,0}(\epsilon))$	$R(\delta^0_{0,0}(\epsilon))$	$\delta^0_{0,0}(\epsilon)$	exact value
1	1.0336	1.3034	1.1762	0.8372	0.8373
2^{-1}	1.0800	1.8775	1.3714	0.7212	0.7217
2^{-2}	1.3110	5.2088	1.7604	0.4805	0.4793
2^{-3}	2.2918	65.397	2.4711	0.1868	0.1870
2^{-4}	8.3531	8805.1	3.3540	0.2607×10^{-1}	0.2584×10^{-1}
2^{-5}	121.08	–	4.7635	0.4715×10^{-3}	0.4744×10^{-3}
2^{-6}	18596	–	5.6141	0.1574×10^{-6}	0.1591×10^{-6}

Table 2: Comparison of ratios defined in (13) for each estimator. The ratio, and therefore the relative error, is much smaller for $\delta^0_{0,0}(\epsilon)$ than for the other two estimators. Notice however that the ratio $R(\delta^0_{0,0}(\epsilon))$ continues to grow as $\epsilon \to 0$.

The drift v^0 pulls $\hat{X}^\epsilon$ in the direction of the closest minimum of g. Consequently samples of $\hat{X}^\epsilon(1)$ near -1 are much more likely than samples of $Y^\epsilon(1)$ near -1 and produce moderate importance weights $\hat{Z}$ in (11).

The initial point $(0, 0.1)$ is contained in a region of strong regularity which in accordance with the results reported in Table 1 implies that $\delta^0_{0,0.1}(\epsilon)$ will have vanishing error (by Theorem 2). As we have checked above, v^0 meets the requirements of Theorem 1 and we know that $\delta^0(\epsilon)$ is an log-efficient estimator at all points, including those not contained in a region of strong regularity. Therefore it is worth checking the performance of $\delta^0_{0,0}(\epsilon)$, the estimator corresponding to the initial condition $X(0) = 0$ which is not contained in any region of strong regularity. These results are reported in Table 2 and we can see that again $\delta^0_{0,0}(\epsilon)$ outperforms the other two estimators and again $\bar{\delta}_{0,0}(\epsilon)$ (where the optimal control $\hat{u}$ corresponding to the point (0,0.1) has been replaced by an optimal control corresponding to the point (0,0)) behaves extremely poorly.

3.2 Estimating $P\left(X^\epsilon(1) \notin D\right).$

The goal of this subsection is to further investigate our claim at the end of Subsection 2.2 that under more general assumptions on the behavior of v^0 the estimator $\delta^0(\epsilon)$ can still have bounded relative error, i.e.

$$\lim_{\epsilon \to 0} \rho(\delta^0_{t,x}(\epsilon)) < \infty$$

for appropriate $(t, x) \in [0, T] \times \mathbb{R}^d$. In particular we will attempt to estimate $P\left(X^\epsilon(1) \notin D\right)$ where $D \subset \mathbb{R}^2$ is the open interval $(-1, 1)$. X^ϵ is still defined as the solution of equation (48). The domain D is chosen so that away from $\{t = 1\}$ the discontinuity of v^0 is qualitatively similar to the one described in the previous subsection. As $\epsilon \to 0$

$$\sup_{0 \le t \le 1} |X^\epsilon(t) - 0.1| \to 0.$$

Therefore the event $\{X^\epsilon(1) \notin D\}$ becomes increasingly unlikely in this limit.

If we define the function g as

$$g(x) = \begin{cases} \infty, & x \in (-1,1), \\ 0, & x \notin (-1,1) \end{cases}$$

then

$$P\left(X^\epsilon(1) \notin D\right) = \mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon(1))}\right].$$

In this simple case we can exactly solve the variational problem (30) which defines G and see that

$$G(t,x) = \begin{cases} \dfrac{(1-x)^2}{2(1-t)}, & x \geq 0, \\ \dfrac{(1+x)^2}{2(1-t)}, & x < 0. \end{cases}$$

and therefore that

$$v^0(t,x) = -\sigma(x)^T D_x G(t,x) = \begin{cases} \dfrac{1-x}{1-t}, & x > 0, \\ \dfrac{-1-x}{1-t}, & x < 0. \end{cases}$$

Let $\mathbb{H}^+$ and $\mathbb{H}^-$ be as defined in the previous subsection. The restrictions of v^0 to the sets $[0,1) \times \mathbb{H}^+$ and $[0,1) \times \mathbb{H}^-$ have the smooth extensions to all of $[0,1) \times \mathbb{R}^d$,

$$v^+(t,x) = \frac{1-x}{1-t}$$

and

$$v^-(t,x) = \frac{-1-x}{1-t}$$

respectively. Unfortunately the blow up of v^+ and v^- as $t \to 1$ does seem to require non-trivial modifications of the arguments in Sections 4 and 5.

For this problem the Cramér Tranformation estimator $\bar{\delta}_{0,0.1}(\epsilon)$ is constructed as follows. At the point $(0, X^\epsilon(0)) = (0, 0.1)$, the unique minimizer in expression (30) is the function

$$\hat{\varphi}_{0,0.1}(s) = 0.1 + 0.9\, s.$$

and the estimator $\bar{\delta}_{0,0.1}(\epsilon)$ corresponds to setting

$$v \equiv \dot{\hat{\varphi}}_{0,0.1}(0) = 0.9$$

in expressions (10) and (11). According to Remark 1, generating samples of an estimator with the same distribution as $\bar{\delta}_{0,0.1}(\epsilon)$ requires sampling the process

$$Y^\epsilon(s) = \hat{\varphi}_{0,0.1}(s) + \sqrt{\epsilon}\, W(s) \tag{50}$$

Because of the behavior of v^+ and v^- at $\{t = 1\}$ neither Theorem 1 or Theorem 2 apply in this case. Nevertheless as reported in Table 3 the performance of $\delta^0_{0,0.1}(\epsilon)$ far surpasses that of the other two estimators. Again,

ϵ	$R(\delta_{0,0.1}(\epsilon))$	$R(\bar{\delta}_{0,0.1}(\epsilon))$	$R(\delta^0_{0,0.1}(\epsilon))$	$\delta^0_{0,0.1}(\epsilon)$	exact value
1	3.1259	10.065	2.2648	0.3186	0.3197
2^{-1}	6.1948	77.460	2.8884	0.1609	0.1614
2^{-2}	20.083	3360.8	4.2575	0.5002×10^{-1}	0.4983×10^{-1}
2^{-3}	156.57	–	3.0951	0.6363×10^{-2}	0.6386×10^{-2}
2^{-4}	6053.3	–	2.4925	0.1643×10^{-3}	0.1645×10^{-3}
2^{-5}	–	–	2.4483	0.1782×10^{-6}	0.1782×10^{-6}
2^{-6}	–	–	2.4861	0.3011×10^{-12}	0.3011×10^{-12}

Table 3: Comparison of ratios defined in (13) for each estimator. The ratio, and therefore the relative error, is much smaller for $\delta^0_{0,0.1}(\epsilon)$ than for the other two estimators. Notice also that the ratio $R(\delta^0_{0,0.1}(\epsilon))$ appears to remain bounded as $\epsilon \to 0$.

the Cramér Tranformation estimator, $\bar{\delta}_{0,0.1}(\epsilon)$, behaves far worse than even the standard Monte Carlo estimator $\delta_{0,0.1}(\epsilon)$. While $\delta^0_{0,0.1}(\epsilon)$ no longer has vanishing error ($\lim_{\epsilon\to 0} R(\delta^0_{0,0.1}(\epsilon)) = 1$) it does seem to have bounded error ($\limsup_{\epsilon\to 0} R(\delta^0_{0,0.1}(\epsilon)) < \infty$). This supports our suggestion that it should be possible to modify the arguments in Sections 4 and 5 to prove that for problems of the form $P(X^\epsilon(1) \notin D)$, our estimator has bounded error (in appropriate regions).

To further investigate this possibility recall that The Large Deviations Theory tells us that

$$-\epsilon \log P(X^\epsilon(1) \notin D) \to G(0, 0.1) \qquad \text{as } \epsilon \to 0$$

but it does not provide a prefactor. In other words the Large Deviations Theory only reveals that

$$P(X^\epsilon(1) \notin D) = e^{-(G(0,0.1)+o(1))/\epsilon}.$$

It can be shown by an asymptotic expansion (see [18]) that, in this case,

$$P(X^\epsilon(1) \notin D) = \sqrt{\epsilon}\, e^{-\frac{1}{\epsilon}G(0,0.1)} \left(1 + \mathcal{O}(\sqrt{\epsilon})\right).$$

Any log-efficient estimator must satisfy

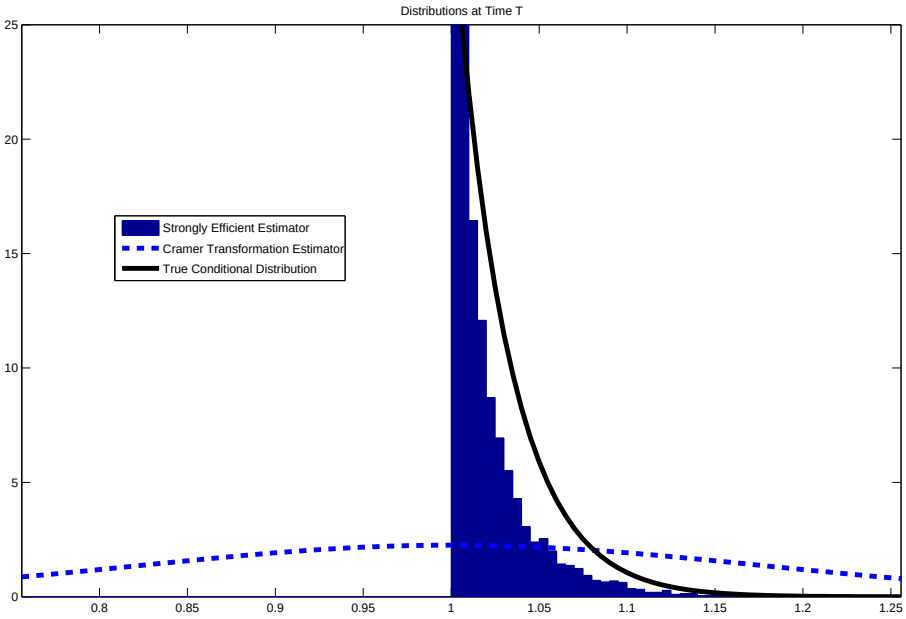
$$\lim_{\epsilon\to 0} -\epsilon \log \mathbf{E}_Q \left[e^{-\frac{2}{\epsilon}g(X^\epsilon)} \left(\frac{dP}{dQ}\right)^2 \right] = 2\, G(0, 0.1)$$

so that

$$\frac{\mathbf{E}_Q \left[e^{-\frac{2}{\epsilon}g(X^\epsilon)} \left(\frac{dP}{dQ}\right)^2 \right]}{P(X^\epsilon(1) \notin D)^2} \sim \frac{1}{\epsilon} e^{-2\, G(0,0.1)/\epsilon}\, \mathbf{E}_Q \left[e^{-\frac{2}{\epsilon}g(X^\epsilon)} \left(\frac{dP}{dQ}\right)^2 \right].$$

In this case, an estimator with bounded error must satisfy

$$\mathbf{E}_Q \left[e^{-\frac{2}{\epsilon}g(X^\epsilon)} \left(\frac{dP}{dQ}\right)^2 \right] = \mathcal{O}(\epsilon)\, e^{-2\, G(0,0.1)/\epsilon}. \tag{51}$$


Distributions at Time T
Strongly Efficient Estimator
Cramer Transformation Estimator
True Conditional Distribution
25
20
15
10
5
0
0.8
0.85
0.9
0.95
1
1.05
1.1
1.15
1.2
1.25

Careful inspection of the proof of Theorem 2 suggests that the validity of expression (51) is related to the behavior of $\hat{X}^\epsilon(1)$ near the boundary of D. In Figure 3.2 we compare the distribution of $\hat{X}^\epsilon(1)$ with the distribution of $Y^\epsilon(1)$ and the conditional distribution of $X^\epsilon(1)$ under P given that $X^\epsilon(1) \notin D$. The distribution of $\hat{X}^\epsilon(1)$ is a much better approximation of the conditional distribution of $X^\epsilon(1)$ under P given that $X^\epsilon(1) \notin D$.

4 log-efficiency.

This section and the next contain a more in depth discussion of the theoretical properties of the importance sampling estimator $\delta^0(\epsilon)$ introduced in Section 2. As in that section we will focus on the estimation of quantities of the form

$$\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon_{t,x}(T))}\right],$$

though the arguments are readily generalized to the case that g takes the form in (17).

In this section we identify (somewhat restrictive) conditions under which our estimator is uniformly log-efficient (see Assumption 3 below). Uniform log-efficiency is often difficult to verify and is required in the proof of our vanishing error result, Theorem 2. Repeating Definition 8, an importance sampling estimator is uniformly log-efficient when for any compact subset $K \subset [0,T] \times \mathbb{R}^d$,

$$\lim_{\epsilon \to 0} \sup_{(t,x)\in K} |V^\epsilon(t,x) - 2\,G(t,x)| = 0 \tag{52}$$

where the functions V^ϵ and G^ϵ are defined in expressions (43) and (30) respectively.

In this section we will prove the following theorem:

Theorem 1. *Assume that $v^0 = -\sigma^T\, D_x G$ satisfies Assumption 3 below. Then the estimator $\delta^0(\epsilon)$ corresponding to the choice $v = v^0$ in (41) is uniformly log-efficient.*

This result will be established in two steps:

1. First we identify a function V which is the uniform limit of V^ϵ, i.e.

$$V = \lim_{\epsilon \to 0} V^\epsilon$$

   and the convergence is uniform on compact sets.

2. Second we show that

$$V = 2\,G.$$

As we will see both of these steps are more difficult to prove when G is not continuously differentiable (so that v^0 is not continuous). The first step will be accomplished using existing techniques for proving the Laplace Principle for

diffusions with discontinuous drift coefficients and is the subject of the next subsection. The second step will be accomplished using a verification type argument where extra care is taken along the set of points at which $D_x G$ does not exist (see Section 4.2).

Before we tackle the proof of uniform log-efficiency in more generality we consider Steps 1 and 2 when v^0 is replaced by some arbitrary smooth and bounded function v. By a straightforward application of Girsanov's Theorem we can rewrite V^ϵ as

$$V^\epsilon(t,x) = -\epsilon \log \mathbf{E}\left[e^{-\frac{2}{\epsilon}g(\tilde{X}^\epsilon_{t,x}(T))+\frac{1}{\epsilon}\int_t^T |v(s,\tilde{X}^\epsilon_{t,x}(s))|^2\,ds}\right] \tag{53}$$

where $\tilde{X}^\epsilon_{t,x}$ is the unique strong solution of the stochastic differential equation

$$\begin{aligned} d\tilde{X}^\epsilon_{t,x}(s) = (b(\tilde{X}^\epsilon_{t,x}(s)) - \sigma(\tilde{X}^\epsilon_{t,x}(s))\, v(s,\tilde{X}^\epsilon_{t,x}(s)))\,dt \\ + \sqrt{\epsilon}\sigma(\tilde{X}^\epsilon_{t,x}(s))\,dW(s), \qquad \tilde{X}^\epsilon_{t,x}(t) = x \end{aligned} \tag{54}$$

(the reader should note that (54) differs from (35), the equation solved by $\hat{X}^\epsilon_{t,x}$). When v is a continuous bounded function the functional

$$F(X) = 2\,g(X(T)) - \int_t^T |v(s,X(s))|^2\,ds \tag{55}$$

is bounded and continuous on the Wiener space of continuous functions on $[t,T]$ with the topology of uniform convergence. We can therefore appeal to the Laplace Principle for $\tilde{X}^\epsilon_{t,x}$ to conclude that

$$\begin{aligned} \lim_{\epsilon\to 0} V^\epsilon(t,x) &= \lim_{\epsilon\to 0} -\epsilon\log\mathbf{E}\left[e^{-\frac{1}{\epsilon}F(\tilde{X}^\epsilon_{t,x})}\right] \\ &= \inf_{\substack{\varphi\in\mathcal{AC}([t,T]):\\ \varphi(t)=x}} \left\{ \int_t^T \frac{1}{2}\|\dot\varphi(s) - b(\varphi(s)) \right. \\ &\qquad \left. + \sigma(\varphi(s))\, v(s,\varphi(s))\|^2_{\varphi(s)}\,ds + F(\varphi)\right\}. \end{aligned}$$

This identity can be rewritten as

$$\lim_{\epsilon\to 0} V^\epsilon(t,x) = \inf_{\substack{\varphi\in\mathcal{AC}([t,T]):\\ \varphi(t)=x}} \left\{ \int_t^T L\left(\varphi(s), v(s,\varphi(s)), \dot\varphi(s)\right)\,ds + 2g(\varphi(T))\right\} \tag{56}$$

and the function L, which we will refer to as a running cost, is defined as

$$L(x,\alpha,\beta) = \|\beta - b(x)\|^2_x - \frac{1}{2}\|\beta - b(x) - \sigma(x)\,\alpha\|^2_x \tag{57}$$

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

If G (and therefore v^0) is smooth then we can apply the same argument with v^0 in place of v to establish Step 1. When G is smooth Step 2 is also straightforward for the choice $v = v^0$. One can use, for example, the method of characteristics to prove that the limit of V^ϵ is equal to $2\,G$.

Considers however, the smooth function

$$v(t,x) = \sigma^{-1}(t,x)\left(\dot{\hat{\varphi}}_{0,x_0}(t) - b(t,x)\right) \tag{58}$$

which corresponds to the Cramér transform estimator at the point $(0,x_0)$ which was introduced earlier. As mentioned in the introduction, this estimator is constructed using sample paths of the process

$$dY^\epsilon(s) = \dot{\hat{\varphi}}_{0,x_0}(s)\,ds + \sqrt{\epsilon}\,\sigma(s,Y^\epsilon(s))\,dW(s). \tag{59}$$

In both expressions (58) and (59) $\hat{\varphi}_{0,x_0}$ corresponds to the unique optimal control trajectory in (31) at the point $(0,x_0)$. For this choice of v expression (56) implies that

$$\lim_{\epsilon\to 0} V^\epsilon(0,x_0) = \inf_{\substack{\varphi\in\mathcal{AC}([0,T]):\\ \varphi(0)=x_0}} \Big\{ \int_0^T \|\dot\varphi(s) - b(s,\varphi(s))\|^2_{\varphi(s)}\,ds + 2g(\varphi(T)) \\ - \frac{1}{2}\int_t^T \|\dot\varphi(s) - \dot{\hat{\varphi}}_{0,x_0}(s)\|^2_{\varphi(s)}\,ds\Big\}. \tag{60}$$

In order for the Cramér transform estimator to be log-efficient at the point $(0,x_0)$ we must have that $\hat{\varphi}_{0,x_0}$ is the optimal control trajectory in (60) so that the second integral term in that expression vanishes.

Using expression (60) in the simple case that $b \equiv 0$ and $\sigma \equiv I$, it is not hard to see that if g is a convex function then the Cramér transform estimator is log-efficient. In the $b \equiv 0$, $\sigma \equiv I$ setting the convexity of g implies that G is actually a classical solution of (28). Note however that, even in this simple setting, one can construct (non-convex) g so that G is a classical solution of (28) yet the Cramér transform estimator is not log-efficient.

It is easier to understand why the Cramér fails to be log-efficient when G is not a classical solution of (28). By a well known result in control theory (see for example [24] or [14]) the viscosity solution of (28) fails to be differentiable at exactly those points where the optimal control problem (30) has multiple solutions, i.e. points at which there are multiple trajectories satisfying (31). Suppose, for example, that $\hat{\varphi}'_{0,x_0}$ is another optimal control trajectory at $(0,x_0)$ (besides $\hat{\varphi}_{0,x_0}$ which was used to define v in (58)). Then from (60) and the definition of G in (30), it is clear that

$$\lim_{\epsilon\to 0} V^\epsilon(0,x_0) \le G(0,x_0) - \frac{1}{2}\int_t^T \|\dot{\hat{\varphi}}'_{0,x_0}(s) - \dot{\hat{\varphi}}_{0,x_0}(s)\|^2_{\hat{\varphi}'_{0,x_0}(s)}\,ds$$

which shows that the Cramér transform estimator is not log-efficient. Similar problems can arise if there are multiple optimal controls at any point in $[0,T]\times$

$\mathbb{R}^d$, i.e. when G fails to be differentiable at a point away from $(0, x_0)$, because the control trajectories in (60) might find it favorable to pass through such a point. In Section 4.2 below we will show that when we choose $v = v^0$ these problems cannot arise. However, we will see that special care is required to deal with points at which G is not differentiable.

4.1 Step 1: a Laplace Principle when v^0 is discontinuous.

Assume now that v^0 is no longer continuous and consider the argument just used to establish Step 1. As a consequence of the discontinuity of v^0, the functional F defined in (55) is no longer continuous. Moreover, the diffusion $\tilde{X}^\epsilon_{t,x}$ has a discontinuous drift coefficient. We cannot, therefore, apply the Laplace Principle as in the previous discussion and relation (56) no longer holds. The validity of the large deviations principle for diffusions with discontinuous drift coefficient is an interesting problem and has been investigated in [19],[20], and [16] among other papers. In a nontrivial setting these authors were able to identify a rate function and prove a large deviations principle. In this subsection we will apply the techniques in [16] to identify (in a nontrivial setting) a uniform limit for V^ϵ. The proof of Theorem 1 is completed in the next subsection where we show that this uniform limit is equivalent to $2\,G$.

As already mentioned the limiting behavior of V^ϵ is more difficult to analyze when v^0 is not continuous. Under Assumption 2 the particular definition of v^0 on any set of measure zero does not effect V^ϵ. Thus when G is continuously differentiable almost everywhere v^0 can be redefined arbitrarily on the set of points where it is not continuous without changing V^ϵ. The same should be true of V, the limit of V^ϵ. This already suggests that the running cost L in the definition of V in (66) will have to be modified. The new running cost (see (65)) does not depend on the value of v^0 at its discontinuities, but it does depend on the geometry of the set of discontinuities of v^0. In fact, every possible form of discontinuity of v^0 requires different treatment, particularly in the proof of Lemma 1 below. We restrict our analysis to the specific form of discontinuity given in Assumption 3. This choice not only allows us to make use of the arguments in [16] but also to illustrate some of the nontrivial consequences of a nonsmooth v^0. A major difficulty in adapting our arguments to handle more general discontinuities is the identification of the new running cost in these cases (for discussion of a closely related issue see [11]). In fact numerical experiments suggest that Theorem 1 holds in much more generality.

We will assume that G is Lipschitz continuous and that $v^0 = -\sigma^T D_x G$ is also continuous on the sets $[0,T] \times \mathbb{H}^+$ and $[0,T] \times \mathbb{H}^-$ where

$$\mathbb{H}^+ = \left\{x \in \mathbb{R}^d : x_1 > 0\right\} \quad \text{and} \quad \mathbb{H}^- = \left\{x \in \mathbb{R}^d : x_1 < 0\right\}$$

but has a discontinuity on the set $[0,T] \times \Lambda$ where

$$\Lambda = \left\{x \in \mathbb{R}^d : x_1 = 0\right\}.$$

Assumption 3. *There exist two bounded continuous functions v^+ and v^- defined on all of $[0,T]\times\mathbb{R}^d$ such that*

$$v^0 = v^+ \quad \text{on } \left\{x \in \mathbb{H}^+\right\}$$

and

$$v^0 = v^- \quad \text{on } \left\{x \in \mathbb{H}^-\right\}.$$

We will prove a result of the form in (56) with a slightly more complicated function in the place of the running cost L. Define the functions

$$L^+(t,x,\beta) = L(x, v^+(t,x), \beta), \qquad L^-(t,x,\beta) = L(x, v^-(t,x), \beta) \tag{61}$$

and let

$$L^0(t,x,\beta) = \inf\left\{\lambda\, L^+(t,x,\beta^+) + (1-\lambda)\, L^-(t,x,\beta^-)\right\} \tag{62}$$

where the infimum in (62) is taken over $\lambda \in [0,1]$, and $\beta^+, \beta^- \in \mathbb{R}^d$ such that

$$\beta_1^+ \le 0, \quad \beta_1^- \ge 0 \tag{63}$$

and

$$\lambda\,\beta^+ + (1-\lambda)\,\beta^- = \beta. \tag{64}$$

Finally combine L^+, L^-, and L^0 to form the function $\bar{L}$ given by

$$\bar{L}(t,x,\beta) = \begin{cases} L^+(t,x,\beta) & x \in \mathbb{H}^+ \\ L^0(t,x,\beta) & x \in \Lambda \\ L^-(t,x,\beta) & x \in \mathbb{H}^-. \end{cases} \tag{65}$$

The next lemma establishes that

$$\lim_{\epsilon\to 0} V^\epsilon = V$$

where V is now defined as

$$V(t,x) = \inf_{\substack{\varphi\in\mathcal{AC}([t,T]):\\ \varphi(t)=x}} \left\{ \int_t^T \bar{L}(s,\varphi(s),\dot{\varphi}(s))\, ds + 2g(\varphi(T)) \right\}. \tag{66}$$

In this definition of V the running cost L in (56) has been replace by $\bar{L}$.

Lemma 1. *Let g be any continuous and bounded function on $\mathbb{R}^d$ and suppose that v^0 satisfies Assumption 3. Then*

$$\lim_{\epsilon\to 0} V^\epsilon = V$$

and the convergence is uniform on compact subsets of $[0,T]\times\mathbb{R}^d$.

Proof. Recall from expression (53) that

$$V^\epsilon(t,x) = -\epsilon \log \mathbf{E}\left[e^{-\frac{1}{\epsilon}F(\tilde{X}^\epsilon_{t,x})}\right]$$

where F is defined in (55) and the process $\tilde{X}^\epsilon_{t,x}$ is the unique strong solution of the stochastic differential equation in expression (54).

The functional F is not continuous but it is bounded and measurable. Therefore we can apply Theorem 3.1 in [21] to obtain the stochastic control representation

$$V^\epsilon(t,x) = \inf_{U\in\mathcal{A}(t)} \mathbf{E}\left[\int_t^T \frac{1}{2}|U(s)|^2\,ds + F(\tilde{X}^{U,\epsilon}_{t,x})\right] \tag{67}$$

where $\mathcal{A}(t)$ is the space of all $\mathbb{R}^d$ valued $\mathcal{F}$ progressively measurable processes on $[t,T]$ with

$$\mathbf{E}\left[\int_t^T |U(s)|^2\,ds\right] < \infty \quad \text{for all} \quad U\in\mathcal{A}(t)$$

and $\tilde{X}^{U,\epsilon}_{t,x}$ is the unique strong solution[1] of the stochastic differential equation

$$\begin{aligned} d\tilde{X}^{U,\epsilon}_{t,x}(s) = (b(\tilde{X}^{U,\epsilon}_{t,x}(s)) + \sigma(\tilde{X}^{U,\epsilon}_{t,x}(s))\,(U(s) - v^0(s,\tilde{X}^{U,\epsilon}_{t,x}(s))))\,dt \\ + \sqrt{\epsilon}\sigma(\tilde{X}^{U,\epsilon}_{t,x}(s))\,dW(s), \qquad \tilde{X}^{U,\epsilon}_{t,x}(t) = x. \end{aligned} \tag{68}$$

Notice that by setting $U \leftarrow U - v^0(\cdot,\tilde{X}^{U,\epsilon}_{t,x}(\cdot))$ expression (67) can be rewritten as

$$V^\epsilon(t,x) = \inf_{U\in\mathcal{A}(t)} \mathbf{E}\left[\int_t^T |U(s)|^2 - \frac{1}{2}|U(s) - v^0(s,X^{U,\epsilon}_{t,x}(s))|^2\,ds + 2g(X^{U,\epsilon}_{t,x}(T))\right] \tag{69}$$

where for any $U\in\mathcal{A}(t)$, $X^{U,\epsilon}_{t,x}$ is the unique strong solution of the stochastic differential equation

$$\begin{aligned} dX^{U,\epsilon}_{t,x}(s) = (b(X^{U,\epsilon}_{t,x}(s)) + \sigma(X^{U,\epsilon}_{t,x}(s))\,U(s))\,dt \\ + \sqrt{\epsilon}\,\sigma(X^{U,\epsilon}_{t,x}(s))\,dW(s), \qquad X^{U,\epsilon}_{t,x}(t) = x. \end{aligned}$$

With expression (69) in hand the proof can be completed using the weak convergence arguments in [16] and [11].

□

Extentions of this result to certain classes of unbounded g and v^0 can be proved using standard techniques (see [11]).

[1] At several points we require that certain stochastic differential equations have unique strong solutions. General conditions under which this is true can be found in [22] and [23].

4.2 Step 2: A verification argument.

We now move on to establishing Step 2 in the procedure outlined on page 23. In this subsection, as in the previous subsection, Assumption 3 continues to apply. Notice that the value function V in expression (66) has a discontinuous running cost $\bar{L}$ and we know very little in advance about its behavior. For this reason it is easier to work with G and use the fact that it is the viscosity solution of (28) to show that it is also the optimal value of the control problem that defines V. This will require a verification type argument in which we take special care at the discontinuity of $D_x G$.

Where G is not continuously differentiable we will use the following generalized definition of the derivative of G (see [24]),

$$D^*G(t,x) = \left\{ \begin{array}{ll} (q,p) \in \mathbb{R} \times \mathbb{R}^d : & \lim_{n\to\infty}(q_n,p_n) = (q,p) \text{ for some sequence} \\ & (t_n,x_n) \to (t,x) \text{ such that for each } n, \\ & G \text{ is differentiable at } (t_n,x_n) \text{ and} \\ & (q_n,p_n) = (\partial_t G(t_n,x_n), D_x G(t_n,x_n)). \end{array} \right\} \tag{70}$$

If G is continuously differentiable at (t,x) then

$$D^*G(t,x) = \{(\partial_t G(t,x), D_x G(t,x))\}.$$

Assumption 3 implies that the restrictions of the derivative $D_x G$ to $\mathbb{H}^+$ and $\mathbb{H}^-$ are continuously extendable to all of $[0,T] \times \mathbb{R}^d$. We have also assumed that G is continuous (in fact Lipschitz continuous). It is easy to check that these facts imply that the derivatives $\partial_t G$ and $\frac{\partial G}{\partial x_j}$ for $j > 1$ must be continuous on $\mathbb{R}^d$. This implies that if $(q,p) \in D^*G(t,x)$ then

$$q = \partial_t G \quad \text{and} \quad p_{[1]} = D_x G_{[1]}. \tag{71}$$

where we have used the notation

$$x_{[1]} = (x_2, \ldots, x_d)$$

for $x \in \mathbb{R}^d$. Assumption 3 also implies that if $(q,p) \in D^*G(t,x)$ for $x \in \Lambda$ then either

$$p_1 = p_1^+ = \lim_{\substack{z \to x \\ y \in \mathbb{H}_+}} \frac{\partial G}{\partial x_1}(t,z) \tag{72}$$

or

$$p_1 = p_1^- = \lim_{\substack{z \to x \\ y \in \mathbb{H}_-}} \frac{\partial G}{\partial x_1}(t,z) \tag{73}$$

both of which exist. Thus $D^*G(t,x)$ is a particularly simple set consisting of at most two points which themselves only differ in 1 component.

The next lemma provides an instantaneous optimality condition that is needed in the proof of Lemma 3 and follows from the the fact that the functions v^+ and v^- in the definition of $\bar{L}$ are not arbitrary but are related to the continuous value function G.

Lemma 2. *Under Assumption 3 for* $(q,p) \in D^*G(t,x)$

1. *if* $x \notin \Lambda$ *then*

$$\sup_{\beta \in \mathbb{R}^d} \big\{ -\langle \beta, 2\,p\rangle - \bar{L}(t,x,\beta)\big\} \leq 2H(x,p). \tag{74}$$

2. *if* $x \in \Lambda$ *then*

$$\sup_{\substack{\beta \in \mathbb{R}^d \\ \beta_1 = 0}} \big\{ -\langle \beta, 2p\rangle - \bar{L}(t,x,\beta)\big\} \leq 2H(x,p). \tag{75}$$

Proof. Part 1 of this Lemma follows from the fact that $v^0 = -\sigma^T D_x G$ and a direct computation. To prove part 2 first recall that a point $(t,x) \in [0,T]\times\mathbb{R}^d$ is regular if there is a unique optimal trajectory, $\hat{\varphi}_{t,x} \in \mathcal{AC}([t,T])$ at (t,x) (see expression (31)). By a standard result in control theory G is differentiable at a point $(t,x) \in [0,T]\times\mathbb{R}^d$ if and only if (t,x) is a regular point. Thus we can assume that every point in $[0,T]\times\Lambda$ is not a regular point and every point in $[0,T]\times\Lambda^c$ is regular. If (t,x) is regular then every point in $\{(s,\hat{\varphi}_{t,x}(s)) : s \in [t,T)\}$ is also regular. By Proposition 1 $\hat{\varphi}_{t,x}$ satisfies the ordinary differential equation

$$\dot{\hat{\varphi}}_{t,x}(s) = b(s,\hat{\varphi}_{t,x}(s)) + \sigma(s,\hat{\varphi}_{t,x}(s))\, v^0(s,\hat{\varphi}_{t,x}(s)).$$

Thus we can assume that Λ is not an attracting set for the field $b+\sigma\, v^0$, i.e. for any $x \in \Lambda$ and all $t \in [0,T]$ we have that,

$$[b(t,x) + \sigma(t,x)\, v^+(t,x)]_1 \geq 0 \qquad \text{and} \qquad [b(t,x) + \sigma(t,x)\, v^-(t,x)]_1 \leq 0. \tag{76}$$

where $[x]_1$ represents the first component of a vector $x \in \mathbb{R}^d$.

Further, G is a viscosity solution of (28),

$$-\partial_t G + H(x, D_x G) = 0$$

where we recall from expression (26)

$$H(x,p) = -\langle b(x), p\rangle + \frac{1}{2}|\sigma(x)^T\, p|^2.$$

The fact that $\partial_t G$ is continuous on $[0,T]\times\mathbb{R}^d$ and that G solves (28) in the classical sense for $x \notin \Lambda$, therefore implies that the function

$$f(t,x) = H(x, D_x G(t,x)) \tag{77}$$

is also continuous on $[0,T]\times\mathbb{R}^d$.

Using (76) a direct computation reveals that for $x \in \Lambda$ and $\beta \in \mathbb{R}^d$ with $\beta_1 = 0$ we have that,

$$\bar{L}(t,x,\beta) = \inf_{\lambda\in[0,1]} \Big\{ \frac{1}{2}\|\beta - b(x) - \lambda\sigma(x)\, v^+(t,x) - (1-\lambda)\sigma(x)\, v^-(t,x)\|_x^2 \\ - \lambda|v^+(t,x)|^2 - (1-\lambda)|v^-(t,x)|^2 \Big\}. \tag{78}$$

We need to show that for $(q,p) \in D^*G(t,x)$ (and by expressions (71), (72), and (73) there are only two such points),

$$\sup_{\substack{\beta\in\mathbb{R}^d \\ \beta_1=0}} \left\{ -\langle \beta, p\rangle - \bar{L}(t,x,p) \right\} \leq 2H(x,p)$$

which can be rewritten as

$$\sup_{\substack{\beta\in\mathbb{R}^d \\ \beta_1=0}} \left\{ -\langle \beta, p\rangle - \bar{L}(t,x,p) \right\} \leq 2f(t,x) \tag{79}$$

where f was defined in (77).

Because the more general computation is somewhat tedious we will prove (79) only in the case that $b \equiv 0$ and σ is the $d \times d$ identity matrix. In other words, using expression (78), we will verify that for some $(q,p) \in D^*G(t,x)$

$$\sup_{\substack{\beta\in\mathbb{R}^d \\ \beta_1=0}} \Big\{ -\langle \beta, 2p\rangle - \inf_{\lambda\in[0,1]} \{ \frac{1}{2}|\beta - \lambda v^+(t,x) - (1-\lambda)\, v^-(t,x)|^2 \\ - \lambda|v^+(t,x)|^2 - (1-\lambda)|v^-(t,x)|^2 \} \Big\} \leq 2f(t,x) \tag{80}$$

where now

$$f(t,x) = \frac{1}{2}|D_x G(t,x)|^2$$

and $v^0 = -D_x G$ on Λ^c. Note that f is continuous on $[0,T]\times\mathbb{R}^d$ and $f = \frac{1}{2}|v^0|^2$.

Recalling expression (71) we have that

$$v_j^+(t,x) = v_j^-(t,x) \quad \text{for} \quad j > 1,$$

i.e. the functions v_j^0 are continuous in all of $[0,T]\times\mathbb{R}^d$ for $j > 1$. We can rewrite the left hand side of inequality (80) as

$$\sup_{\substack{\beta\in\mathbb{R}^{d-1} \\ \lambda\in[0,1]}} \Big\{ -\langle \beta, 2p_{[1]}\rangle - \frac{1}{2}\left(\lambda v_1^+(t,x) - (1-\lambda)\, v_1^-(t,x)\right)^2 \\ + \lambda\left(v_1^+(t,x)\right)^2 + (1-\lambda)\left(v_1^-(t,x)\right)^2 - \frac{1}{2}|\beta - v_{[1]}^0(t,x)|^2 + |v_{[1]}^0(t,x)|^2 \Big\}.$$

This quantity is bounded above by

$$\max\left\{ \left(v_1^+(t,x)\right)^2, \left(v_1^-(t,x)\right)^2 \right\} \\ + \sup_{\beta\in\mathbb{R}^{d-1}} \left\{ -\langle \beta, 2p_{[1]}\rangle - \frac{1}{2}|\beta - v_{[1]}^0(t,x)|^2 + |v_{[1]}^0(t,x)|^2 \right\}. \tag{81}$$

Noting that $v^0_{[1]} = -p_{[1]}$ for $(q,p) \in D^*G(t,x)$, a straightforward computation reveals that

$$\sup_{\beta\in\mathbb{R}^{d-1}} \left\{ -\langle \beta, 2p_{[1]}\rangle - \frac{1}{2}|\beta - v^0_{[1]}(t,x)|^2 + |v^0_{[1]}(t,x)|^2 \right\} = |v^0_{[1]}(t,x)|^2.$$

Combining this with (81) we have shown that for $(q,p) \in D^*G(t,x)$,

$$\sup_{\substack{\beta\in\mathbb{R}^d \\ \beta_1=0}} \left\{ -\langle \beta, p\rangle - \bar{L}(t,x,p) \right\} \le \max\left\{ |v^+(t,x)|^2, |v^-(t,x)|^2 \right\}.$$

Since for $x \in \Lambda$ both of the functions in the maximum on the right are equal to $2f(t,x)$, the proof is complete.

□

We can now prove the main result of this section.

Lemma 3. *Under the Assumption 3,*

$$V = 2\,G.$$

Proof. For any $(t,x) \in [0,T] \times \mathbb{R}^d$ and $\varphi \in \mathcal{AC}([t,T])$ with $\varphi(t) = x$ define the function $h : [t,T] \to \mathbb{R}$ by

$$h(r) = \int_t^r \bar{L}(s,\varphi(s),\dot{\varphi}(s))\,ds + 2\,G(r,\varphi(r))$$

Notice that $h(t) = 2\,G(t,x)$ and

$$h(T) = \int_t^T \bar{L}(s,\varphi(s),\dot{\varphi}(s))\,ds + 2g(\varphi(T)).$$

We will prove that $h(t) \le h(T)$ by showing that $\dot{h} \ge 0$ almost everywhere in $[t,T]$. Since $\varphi \in \mathcal{AC}([t,T])$ is arbitrary this will imply that

$$\begin{aligned} 2\,G(t,x) &\le \inf_{\substack{\varphi\in\mathcal{AC}([t,T]):\\ \varphi(t)=x}} \left\{ \int_t^T \bar{L}(s,\varphi(s),\dot{\varphi}(s))\,ds + 2g(\varphi(T)) \right\} \\ &= V(t,x) \end{aligned}$$

and the proof will be complete.

Since φ is absolutely continuous we can restrict our attention to those $r \in [t,T]$ where φ is differentiable. The Lipschitz continuity of G and the absolute continuity of φ imply that h is also absolutely continuous and therefore differentiable almost everywhere. Where it exists, the derivative of h is given by

$$\dot{h}(r) = 2\frac{d}{dr}G(r,\varphi(r)) + \bar{L}(r,\varphi(r),\dot{\varphi}(r)). \tag{82}$$

We need to consider both of the cases $\varphi(r) \notin \Lambda$ and and $\varphi(r) \in \Lambda$. However, note that if $\varphi(r) \in \Lambda$ and $[\dot{\varphi}(r)]_1 \neq 0$ then r is contained in some neighborhood $U \in [t,T]$ such that $\varphi \notin \Lambda$ almost everywhere in U. We can therefore ignore such points and assume that whenever $\varphi(r) \in \Lambda$, we also have that $[\dot{\varphi}(r)]_1 = 0$.

If $\varphi(r) \notin \Lambda$ then

$$\frac{d}{dr}G(r,\varphi(r)) = \partial_t G(r,\varphi(r)) + \langle \dot{\varphi}(r), D_x G(r,\varphi(r))\rangle. \tag{83}$$

As in the discussion around expression (71) above, $D_x G_{[1]}$ is exists and is continuous so that if $\varphi(r) \in \Lambda$ with $\dot{\varphi}_{[1]}(r) = 0$ then

$$\frac{d}{dr}G(r,\varphi(r)) = \partial_t G(r,\varphi(r)) + \langle \dot{\varphi}_{[1]}(r), D_x G_{[1]}(r,\varphi(r))\rangle.$$

Since $p_{[1]} = D_x G_{[1]}(t,x)$ for $(q,p) \in D^* G(t,x)$, we have that for $\varphi(r) \in \Lambda$ with $\dot{\varphi}_{[1]}(r) = 0$

$$\frac{d}{dr}G(r,\varphi(r)) = \partial_t G(r,\varphi(r)) + \langle \dot{\varphi}(r), p\rangle. \tag{84}$$

Plugging expressions (83) and (84) into (82) we conclude that if $\varphi(r) \notin \Lambda$ or if $\varphi(r) \in \Lambda$ with $\dot{\varphi}_{[1]}(r) = 0$ then for $(q,p) \in D^* G(r,\varphi(r))$,

$$\dot{h}(r) = 2\partial_t G(r,\varphi(r)) + \langle \dot{\varphi}(r), 2p\rangle + \bar{L}(r,\varphi(r),\dot{\varphi}(r)).$$

Lemma 2 then implies that

$$\dot{h}(r) \geq 2\partial_t G(r,\varphi(r)) - 2\,H(\varphi(r),p).$$

By the continuity of $\partial_t G$ and H, and from the definition of $D^* G$, for some sequence $x_n \notin \Lambda$ with $x_n \to \varphi(r)$, we have that,

$$\dot{h}(r) \geq 2 \lim_{n\to\infty} \partial_t G(r,x_n) - H(x_n, D_x G(r,x_n)).$$

Since G solves (28) in the classical sense on Λ^c,

$$\partial_t G(r,x_n) - H(x_n, D_x G(r,x_n)) = 0$$

for all n so that

$$\dot{h}(r) \geq 0$$

and the proof is complete. □

Taken together Lemmas 1 and 3 prove Theorem 1.

5 Vanishing relative error.

In the previous section we provide conditions under which the estimator $\delta^0(\epsilon)$ is uniformly log-efficient. Uniform log-efficiency is difficult to establish as reflected

by need for Assumption 3 in that section. In contrast, once uniform log-efficiency has been established it is relatively easy to prove that for initial conditions in appropriate regions, $\delta^0(\epsilon)$ satisfies asymptotic error properties much more favorable than log-efficiency. This section's main result, Theorem 2 establishes that the relative error of $\delta^0(\epsilon)$ can decrease to zero as $\epsilon \to 0$. We will again focus our attention on the estimation problem for

$$\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^{\epsilon}_{t,x}(T))}\right].$$

In this section and in Theorem 2 we make no explicit restrictions on the form of the set of possible discontinuities of G (i.e. Assumption 3 is not required). Instead we assume that the estimator $\delta^0(\epsilon)$ is uniformly log-efficient and show that if (t,x) is contained in a region of sufficient regularity for G then the relative error of $\delta^0_{t,x}(\epsilon)$ satisfies

$$\lim_{\epsilon \to 0} \rho\left(\delta^0_{t,x}(\epsilon)\right) = 0$$

where $\rho\left(\delta^0_{t,x}(\epsilon)\right)$ is the relative error of the estimator $\delta^0_{t,x}(\epsilon)$ and is defined in (12).

These "region(s) of sufficient regularity for G" just mentioned are defined as follows (see [26] and [27]).

Definition 9. *Let N be any relatively open subset of $[0,T] \times \mathbb{R}^d$. We call N a* region of strong regularity *if*

1. *G is $\mathcal{C}^1$ on N and*
2. *For every $(t,x) \in N$, there exists a unique optimal control trajectory (see (31)) $\hat{\varphi}_{t,x} \in \mathcal{AC}([t,T])$ such that*

$$\{(s, \hat{\varphi}_{t,x}(s)) : (t,x) \in N \text{ and } t \leq s \leq T\} \subset N. \tag{85}$$

Both properties in Definition 9 are crucial in the proof of Theorem 2. They are also not very restrictive. The next proposition shows that, at least when g is smooth, there are actually very few points which are not contained in a region of strong regularity. The essential elements of the proof can be found in Section I.10 of [14].

Proposition 2. *Suppose that g is smooth. Then the set of points which are contained in a region of strong regularity has full Hausdorff Dimension.*

We now proceed to the statement and proof of our vanishing error result. First notice that it follows from Remark 1 that

$$V^{\epsilon}(t,x) = -\epsilon \log \mathbf{E}\left[e^{-\frac{2}{\epsilon}g(\hat{X}^{\epsilon}_{t,x}(T)) - \frac{2}{\sqrt{\epsilon}}\int_t^T \langle v^0(s,\hat{X}^{\epsilon}_{t,x}(s)), dW(s)\rangle - \frac{1}{\epsilon}\int_t^T |v^0(s,\hat{X}^{\epsilon}_{t,x}(s))|^2\, ds}\right] \tag{86}$$

where the process $\hat{X}^{\epsilon}_{t,x}$ is the unique strong solution of the stochastic differential equation in (35) and V^{ϵ} was originally defined in (43). We will use this alternative representation of V^{ϵ} in the proof of Theorem 2.

Before we give the statement of Theorem 2 the reader should recall from expression (12) that the relative error of our estimator $\delta^0_{t,x}(\epsilon)$ can written as

$$\begin{aligned}\rho\left(\delta^0_{t,x}(\epsilon)\right) &= \frac{1}{\sqrt{M}}\sqrt{\frac{\mathbf{E}_Q\left[e^{-\frac{2}{\epsilon}g(X^\epsilon)}\left(\frac{dP}{dQ}\right)^2\right]}{\mathbf{E}\left[e^{-\frac{1}{\epsilon}g(X^\epsilon)}\right]^2} - 1} \\ &= \frac{1}{\sqrt{M}}\sqrt{e^{-\frac{1}{\epsilon}(V^\epsilon(t,x) - 2\,G^\epsilon(t,x))} - 1}\end{aligned}$$

where the function G^ϵ is defined in expression (24).

Theorem 2. *Let (t,x) be contained in a region of strong regularity $N \subset [0,T]\times\mathbb{R}^d$. Suppose that G is smooth on $\bar{N}$ and that $\delta^0(\epsilon)$ is uniformly log-efficient. Then*

$$\lim_{\epsilon\to 0}\rho\left(\delta^0_{t,x}(\epsilon)\right) = 0.$$

Proof. This result essentially requires that we identify the first term in an asymptotic expansion of a large deviations limit. We follow the basic outline of the argument in [26]. We can assume without loss of generality that $\bar{N}$ is compact. We begin by defining the stopping time

$$\hat{\tau} = \inf\left\{s \in [t,T] : (s, \hat{X}^\epsilon_{t,x}(s)) \notin N\right\},$$

i.e. the first exit time of $(s, \hat{X}^\epsilon_{t,x}(s))$ from the set N.

Notice that, since $G \in \mathcal{C}^2(\bar{N})$, Ito's formula implies that,

$$\begin{aligned}\frac{G(\hat{\tau}, \hat{X}^\epsilon_{t,x}(\hat{\tau})) - G(t,x)}{\epsilon} &= \frac{1}{\epsilon}\int_0^{\hat{\tau}}\Big(\partial_t G(s, \hat{X}^\epsilon_{t,x}(s)) \\ &\quad + \left\langle b(\hat{X}^\epsilon_{t,x}(s)) + \sigma(\hat{X}^\epsilon_{t,x}(s))v^0(s, \hat{X}^\epsilon_{t,x}(s)), D_x G(s, \hat{X}^\epsilon_{t,x}(s))\right\rangle \\ &\quad + \frac{\epsilon}{2}\operatorname{tr} a(\hat{X}^\epsilon_{t,x}(s))\, D_x^2 G(s, \hat{X}^\epsilon_{t,x}(s))\Big)\, ds \\ &\quad + \frac{1}{\sqrt{\epsilon}}\int_0^{\hat{\tau}}\langle D_x G(s, \hat{X}^\epsilon_{t,x}(s)), \sigma(\hat{X}^\epsilon_{t,x}(s))\, dW(s)\rangle.\end{aligned}$$

On N, G is smooth and is a solution of the equation

$$\partial_t G + \langle b, D_x G\rangle = \frac{1}{2}|\sigma^T D_x G|^2$$

in the classical sense. Therefore, the last expression can be rewritten as

$$\frac{G(\hat{\tau},\hat{X}^{\epsilon}_{t,x}(\hat{\tau}))-G(t,x)}{\epsilon}=\frac{1}{\epsilon}\int_0^{\hat{\tau}}\left(\frac{1}{2}|\sigma(\hat{X}^{\epsilon}_{t,x}(s))^T\,D_xG(s,\hat{X}^{\epsilon}_{t,x}(s))|^2\right.$$
$$+\langle\sigma(\hat{X}^{\epsilon}_{t,x}(s))\,v^0(s,\hat{X}^{\epsilon}_{t,x}(s)),D_xG(s,\hat{X}^{\epsilon}_{t,x}(s))\rangle$$
$$\left.+\frac{\epsilon}{2}\,\mathrm{tr}\,a(\hat{X}^{\epsilon}_{t,x}(s))\,D_x^2G(s,\hat{X}^{\epsilon}_{t,x}(s))\right)ds$$
$$+\frac{1}{\sqrt{\epsilon}}\int_0^{\hat{\tau}}\langle D_xG(s,\hat{X}^{\epsilon}_{t,x}(s)),\sigma(\hat{X}^{\epsilon}_{t,x}(s))\,dW(s)\rangle.$$

Inserting the identity $v^0=-\sigma^T\,D_xG$ we obtain

$$\frac{G(\hat{\tau},\hat{X}^{\epsilon}_{t,x}(\hat{\tau}))-G(t,x)}{\epsilon}=\frac{1}{\epsilon}\int_0^{\hat{\tau}}\left(-\frac{1}{2}|v^0(s,\hat{X}^{\epsilon}_{t,x}(s))|^2\right.$$
$$\left.+\frac{\epsilon}{2}\,\mathrm{tr}\,a(\hat{X}^{\epsilon}_{t,x}(s))\,D_x^2G(s,\hat{X}^{\epsilon}_{t,x}(s))\right)ds-\frac{1}{\sqrt{\epsilon}}\int_0^{\hat{\tau}}\langle v^0(s,\hat{X}^{\epsilon}_{t,x}(s)),dW(s)\rangle.$$

or after rearranging and multiplying both sides by 2,

$$-\frac{2}{\sqrt{\epsilon}}\int_0^{\hat{\tau}}\langle v^0(s,\hat{X}^{\epsilon}_{t,x}(s)),dW(s)\rangle-\frac{1}{\epsilon}\int_0^{\hat{\tau}}|v^0(s,\hat{X}^{\epsilon}_{t,x}(s))|^2ds$$
$$=\frac{2\left(G(\hat{\tau},\hat{X}^{\epsilon}_{t,x}(\hat{\tau}))-G(t,x)\right)}{\epsilon}-\int_0^{\hat{\tau}}\mathrm{tr}\,a(\hat{X}^{\epsilon}_{t,x}(s))\,D_x^2D_x^2G(s,\hat{X}^{\epsilon}_{t,x}(s))\,ds. \quad (87)$$

Now notice that from expression (86) and the Strong Markov Property we have the following representation for V^{ϵ},

$$e^{-\frac{1}{\epsilon}V^{\epsilon}(t,x)}=\mathbf{E}\left[e^{-\frac{1}{\epsilon}V^{\epsilon}(\hat{\tau},\hat{X}^{\epsilon}_{t,x}(\hat{\tau}))-\frac{2}{\sqrt{\epsilon}}\int_0^{\hat{\tau}}\langle v^0(s,\hat{X}^{\epsilon}_{t,x}(s)),dW(s)\rangle-\frac{1}{\epsilon}\int_0^{\hat{\tau}}|v^0(s,\hat{X}^{\epsilon}_{t,x}(s))|^2\,ds}\right].$$

Inserting (87) into this expression and multiplying both sides by $e^{\frac{2}{\epsilon}G(t,x)}$ we obtain the representation

$$e^{-\frac{1}{\epsilon}(V^{\epsilon}(t,x)-2\,G(t,x))}=$$
$$\mathbf{E}\left[e^{-\frac{1}{\epsilon}(V^{\epsilon}(\hat{\tau},\hat{X}^{\epsilon}_{t,x}(\hat{\tau}))-2\,G(\hat{\tau},\hat{X}^{\epsilon}_{t,x}(\hat{\tau})))-\int_0^{\hat{\tau}}\mathrm{tr}\,a(\hat{X}^{\epsilon}_{t,x}(s))\,D_x^2G(s,\hat{X}^{\epsilon}_{t,x}(s))\,ds}\right]. \quad (88)$$

Now choose $\eta>0$ so small that

$$\mathrm{dist}\left(\{(s,\hat{\varphi}_{t,x}(s)):\ t\le s\le T\},N^c\right)>\eta$$

where the function $\hat{\varphi}_{t,x}$ is the unique minimizer in expression (30). That this is possible follows from the definition of a region of strong regularity, Definition 9. Then on the event

$$A_{\eta}=\left\{\sup_{0\le s\le T}|\hat{X}^{\epsilon}_{t,x}(s)-\hat{\varphi}_{t,x}(s)|<\eta\right\}$$

we have that $\hat{\tau} = T$. Since v^0 is smooth on $\bar{N}$ and $\bar{N}$ is compact, Proposition 1 and the Large Deviations Principle for the process $\hat{X}^\epsilon_{t,x}$ (see [28]) imply that there exists a constant $\eta > 0$ and an $\epsilon_1 > 0$ such that if $0 < \epsilon < \epsilon_1$,

$$\sup_{(s,y)\in\bar{N}} P\left(A^c_\eta\right) \leq e^{-\eta/\epsilon}.$$

Our uniform log-efficiency assumption implies that,

$$V^\epsilon - 2\,G \to 0$$

uniformly on $\bar{N}$. Therefore, there exists a constant C and an $\epsilon_0 \leq \epsilon_1$ such that if $0 < \epsilon < \epsilon_0$ then

$$|e^{-\frac{1}{\epsilon}(V^\epsilon - 2\,G)}| \leq C e^{\eta/2\epsilon}$$

on $\bar{N}$. This implies that

$$\begin{aligned}\mathbf{E}\left[e^{-\frac{1}{\epsilon}(V^\epsilon(\hat{\tau},\hat{X}^\epsilon_{t,x}(\hat{\tau}))-2\,G(\hat{\tau},\hat{X}^\epsilon_{t,x}(\hat{\tau})))-\int_0^{\hat{\tau}}\operatorname{tr} a(\hat{X}^\epsilon_{t,x}(s))\,D_x^2G(s,\hat{X}^\epsilon_{t,x}(s))\,ds}\,\mathbf{1}_{A^c_\eta}\right]&\\ \leq C e^{\eta/2\epsilon} P\left(A^c_\eta\right)&\\ \leq C e^{-\eta/2\epsilon}.&\end{aligned} \tag{89}$$

Notice that since $\hat{\tau} = T$ on A_η and

$$V^\epsilon(T,x) = 2\,G(T,x) = 2\,g(x),$$

we have that

$$e^{-\frac{1}{\epsilon}(V^\epsilon(\hat{\tau},\hat{X}^\epsilon_{t,x}(\hat{\tau}))-2\,G(\hat{\tau},\hat{X}^\epsilon_{t,x}(\hat{\tau})))} = 1$$

on A_η. Therefore,

$$\begin{aligned}\mathbf{E}\left[e^{-\frac{1}{\epsilon}(V^\epsilon(\hat{\tau},\hat{X}^\epsilon_{t,x}(\hat{\tau}))-2\,G(\hat{\tau},\hat{X}^\epsilon_{t,x}(\hat{\tau})))-\int_0^{\hat{\tau}}\operatorname{tr} a(\hat{X}^\epsilon_{t,x}(s))\,D_x^2G(s,\hat{X}^\epsilon_{t,x}(s))\,ds}\,\mathbf{1}_{A_\eta}\right]&\\ = \mathbf{E}\left[e^{-\int_0^{\hat{\tau}}\operatorname{tr} a(\hat{X}^\epsilon_{t,x}(s))\,D_x^2G(s,\hat{X}^\epsilon_{t,x}(s))\,ds}\,\mathbf{1}_{A_\eta}\right].&\end{aligned} \tag{90}$$

By the uniform Lipschitz continuity of $\operatorname{tr} a\,D_x^2G$ on $\bar{N}$, and since, $\hat{\tau} = T$ on A_η, there exists a constant C such that

$$\begin{aligned}\left|\int_0^{\hat{\tau}}\operatorname{tr} a(\hat{X}^\epsilon_{t,x}(s))\,D_x^2G(s,\hat{X}^\epsilon_{t,x}(s))\,ds - \int_0^T \operatorname{tr} a(\hat{\varphi}_{t,x}(s))\,D_x^2G(s,\hat{\varphi}_{t,x}(s))\,ds\right|&\\ \leq C\,\eta&\end{aligned}$$

on the event A_η. Therefore, since σ and D_x^2G are bounded on $\bar{N}$,

$$\begin{aligned}&\left|e^{-\int_0^{\hat{\tau}}\operatorname{tr} a(\hat{X}^\epsilon_{t,x}(s))\,D_x^2G(s,\hat{X}^\epsilon_{t,x}(s))\,ds} - e^{-\int_0^T\operatorname{tr} a(\hat{\varphi}_{t,x}(s))\,D_x^2G(s,\hat{\varphi}_{t,x}(s))\,ds}\right|\\ &\quad\leq C\left|\int_0^{\hat{\tau}}\operatorname{tr} a(\hat{X}^\epsilon_{t,x}(s))\,D_x^2G(s,\hat{X}^\epsilon_{t,x}(s))\,ds\right.\\ &\qquad\qquad\left.-\int_0^T\operatorname{tr} a(\hat{\varphi}_{t,x}(s))\,D_x^2G(s,\hat{\varphi}_{t,x}(s))\,ds\right|\\ &\quad\leq C\,\eta\end{aligned}$$

on the event A_η for some constant C. Applying this bound in expression (90) we obtain,

$$\left| \mathbf{E}\left[e^{-\frac{1}{\epsilon}(V^\epsilon(\hat{\tau},\hat{X}^\epsilon_{t,x}(\hat{\tau}))-2\,G(\hat{\tau},\hat{X}^\epsilon_{t,x}(\hat{\tau})))-\int_0^{\hat{\tau}} \operatorname{tr} a(\hat{X}^\epsilon_{t,x}(s))\,D_x^2 G(s,\hat{X}^\epsilon_{t,x}(s))\,ds}\,\mathbf{1}_{A_\eta}\right] \right.$$
$$\left. - e^{-\int_0^T \operatorname{tr} a(\hat{\varphi}^\epsilon_{t,x}(s))\,D_x^2 G(s,\hat{\varphi}^\epsilon_{t,x}(s))\,ds} \right| \leq C\,\eta \quad (91)$$

for some constant C.

Using the bounds in (89) and (91) in expression (88) we obtain

$$\left| e^{-\frac{1}{\epsilon}(V^\epsilon(t,x)-2\,G(t,x))} - e^{-\int_0^T \operatorname{tr} a(\hat{\varphi}_{t,x}(s))\,D_x^2 G(s,\hat{\varphi}_{t,x}(s))\,ds} \right| \leq C\left(e^{-\eta/2\epsilon}+\eta\right)$$

for some constant C. Since η is arbitrary this implies that

$$\lim_{\epsilon\to 0} e^{-\frac{1}{\epsilon}(V^\epsilon(t,x)-2\,G(t,x))} = e^{-\int_0^T \operatorname{tr} a(\hat{\varphi}_{t,x}(s))\,D_x^2 G(s,\hat{\varphi}_{t,x}(s))\,ds}. \quad (92)$$

In exactly the same way we can show that

$$\lim_{\epsilon\to 0} e^{-\frac{2}{\epsilon}(G^\epsilon(t,x)-G(t,x))} = e^{-\int_0^T \operatorname{tr} a(\hat{\varphi}_{t,x}(s))\,D_x^2 G(s,\hat{\varphi}_{t,x}(s))\,ds}. \quad (93)$$

Combining expressions (92) and (93) we obtain

$$\lim_{\epsilon\to 0} e^{-\frac{1}{\epsilon}(V^\epsilon(t,x)-2\,G^\epsilon(t,x))} = \lim_{\epsilon\to 0} \frac{e^{-\frac{1}{\epsilon}(V^\epsilon(t,x)-2\,G(t,x))}}{e^{-\frac{2}{\epsilon}(G^\epsilon(t,x)-G(t,x))}} = 1$$

which implies that

$$\lim_{\epsilon\to 0} \rho\left(\delta^0_{t,x}(\epsilon)\right) = 0$$

and completes the proof. □

6 Acknowledgments

We are grateful to Professors Paul Dupuis and Hui Wang for pointing out references [6, 7, 8, 9] which helped to guide important parts of this research. We would also like to thank Professors Gérard Ben Arous, Jonathan Goodman, Robert Kohn, and Toufic Suidan as well as Doctors Nawaf Bou-Rabee and Maria Cameron for many helpful conversations and useful suggestions. Jonathan Weare was supported in part by the Applied Mathematical Sciences Program of the U.S. Department of Energy under Contract DEFG0200ER25053.

References

[1] Glasserman, P. and Li, J. (2005). Importance sampling for portfolio credit risk. *Management Science.* **51**(11) 1643–1656.

[2] Guasoni, P. and Robertson, S. (2008). Optimal importance sampling with explicit formulas in continuous time. *Finance and Stochastics.* **12**(1) 1–19.

[3] Frenkel, D. and Smit, B. (1996). *Understanding Molecular Simulation.* Academic Press, San Diego.

[4] Shahabuddin, P. (1994). Importance sampling for the simulation of highly reliable Markovian systems. *Management Science.* **40** 333–352.

[5] Bucklew, J. (2004). *Introduction to Rare Event Simulation.* Springer–Verlag, New York.

[6] Dupuis, P. and Wang, H. (2004). Importance sampling, large deviations, and differential games. *Stochastics.* **76** 481–508.

[7] Dupuis, P. and Wang, H. (2007). Subsolutions of an isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.* **32**(3) 723–757.

[8] Dupuis, P. and Wang, H. (2005). Dynamic importance sampling for uniformly recurrent markov chains. *Ann. Appl. Probab.* **15**(1A) 1–38.

[9] Dupuis, P, Sezer, A, and Wang, H. (2007). Dynamic importance sampling for queueing networks. *Ann. Appl. Probab.* **17**(4) 1306–1346.

[10] Dembo, A. and Zeitouni, O. (1993). *Large Deviations Techniques and Applications.* Springer–Verlag, New York.

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

[12] Varadhan, S. (1985). *Large Deviations and Applications.* Society for Industrial and Applied Mathematics, Philadelphia.

[13] Blanchet, J. and Liu, J. (2008). State-dependent importance sampling for regularly varying random walks. *Adv. Appl. Prob.* **40** 1104–1128.

[14] Fleming, W. and Soner, H. (2006). *Controlled Markov Processes and Viscosity Solutions.* Springer–Verlag, New York.

[15] E, W., Ren, W. and Vanden-Eijnden, E. (2004). Minimum action method for the study of rare events. *Communications on Pure and Applied Mathematics* **57**(5) 637–656.

[16] Boué, M., Dupuis, P. and Ellis, R. (2000). Large deviations for small noise diffusions with discontinuous statistics. *Probability Theory and Related Fields.* **116**(1) 125–149.

[17] Glasserman, P. and Wang, Y. (1997). Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Probab.*, **7**(3) 731–746.

[18] Azencott, R. (1985). Petites perturbations aléatoires des systèmes dynamiques: développements asymptotiques. *Bull. Sci. Math. (2)* **109**(3) 253–308.

[19] Korostelëv, A. and Leonov, S. (1992). An action functional for a diffusion process with discontinuous drift. *Teor. Veroyatnost. i Primenen.* 37(3) 570–576.

[20] Chiang, T. and Sheu, S. (2000). Large deviation of diffusion processes with discontinuous drift and their occupation times. *Ann. Probab.* **28**(1) 140–165.

[21] Boué, M. and Dupuis, P. (1998). A variational representation for certain functionals of brownian motion. *Ann. Probab.* **26**(4) 1641-1659.

[22] Veretennikov, A. (1981). On strong solution and explicit formulas for solutions of stochastic integral equations. *Math. USSR Sb.* **39** 387–403.

[23] Gyongy, I. and Krylov, N. (1994). Existence of strong solutions for Ito's stochastic equations via approximations. *Probability Theory and Related Fields.* **105** 143–158.

[24] Bardi, M. and Capuzzo-Dolcetta, I. (2008). *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations.* Birkhuser, Boston.

[25] Clarke, F. (1983). *Optimization and Nonsmooth Analysis.* John Wiley & Sons, New York.

[26] Fleming, W. and James, M. (1992). Asymptotic series and exit time probabilities. *Ann. Probab.* **20**(3) 1369–1384.

[27] Fleming, W. and Souganidis, P. (1986). Asymptotic series and the method of vanishing viscosity. *Indiana Univ. Math. J.* **35**(2) 425–447.

[28] Freidlin, M. and Wentzell, A. (1984). *Random Perturbations of Dynamical Systems.* Springer-Verlag, New York.