Moderate Deviations and Subsolution-Based Importance Sampling for Recursive Stochastic Algorithms

by

Dane Johnson

B.A., Hamilton College; Clinton, NY, 2007

A dissertation submitted in partial fulfillment of the

requirements for the degree of Doctor of Philosophy

in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2015

This dissertation by Dane Johnson is accepted in its present form
by The Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date___________________ ____

Paul Dupuis, Ph.D., Advisor

Recommended to the Graduate Council

Date___________________ ____

Amarjit Budhiraja, Ph.D., Reader

Date___________________ ____

Kavita Ramanan, Ph.D., Reader

Approved by the Graduate Council

Date___________________ ____

Peter M. Weber, Dean of the Graduate School

Vitae

Dane Johnson was born on February 16, 1985 in Red Bank, New Jersey. He received a B.A. in Mathematics and Economics from Hamilton College in May 2007. The following September he entered the Ph.D. program in the Department of Statistical Science at Duke University. After a year in this program he decided he wanted exposure to other areas of mathematics, and in September of 2008 he entered the Ph.D. program in the Division of Applied Mathematics at Brown University under the supervision of Paul Dupuis. In August 2014 he will begin a one year position as a postdoctoral research associate under Paul Dupuis.

Acknowledgements

I am extremely grateful to Paul Dupuis for his support and encouragement over the years. He taught me the importance of investigating concepts until they are intuitive and comfortable. While there still are and always will be plenty of topics in math that I don't understand, our conversations have made many concepts which I initially found confusing and elusive now seem obvious. I have greatly enjoyed learning from him and working with him, and I feel very fortunate to have had the opportunity.

I am also in debt to Amarjit Budhiraja and Kavita Ramanan for being on my defense comittee and for giving me feedback on my thesis. Boris Rozovsky, Hui Wang, Justin Holmer, Constantine Dafermos, Markus Fischer, and J. Thomas Beale have all kindly contributed to my understanding of analysis . I would also like to thank Jim Gubernatis for allowing me to spend two weeks at Los Alamos with him and for having the patience to discuss physics with me despite my very weak knowledge of the subject. I would like to thank my undergraduate math advisor, Robert Kantrowitz, for giving me a love of mathematics despite the fact that I had no interest in it prior to taking his class. I am also grateful to my undergraduate economics advisor, Christopher Georges, for going out of his way to help me explore my interests and encouraging me to go to graduate school.

Of course I owe a great deal to my family for supporting me over the years. My wife has always been encouraging and patient throughout this whole process. My parents have been supportive as well, and have always generously made sacrifices to benefit my education.

Abstract of “ Moderate Deviations and Subsolution-Based Importance Sampling for Recursive Stochastic Algorithms ” by Dane Johnson, Ph.D., Brown University, May 2015

We prove a moderate deviations principle for the continuous time linear interpolation of discrete time recursive stochastic processes, and then investigate importance sampling schemes based on subsolutions to the Hamilton-Jacobi-Bellman equation associated with the moderate deviations structure. Both the proof of the moderate deviations principle itself as well as the proof of the asymptotic performance of importance sampling schemes based on moderate deviations rely on proving tightness of the empirical measures of the conditional means of the controlled noises as well as the controlled processes themselves. This is more complex than what is needed in the large deviation setting primarily because of the moderate deviations scaling which amplifies the noise and the weaker assumptions imposed on the moment generating function. The main tools used are the relative entropy representation of exponential integrals and the weak convergence of probability measures.

The resulting moderate deviations structure is essentially a linear approximation of the large deviations dynamics and a quadratic approximation of the large deviations costs, both centered around the law of large numbers limit. Consequently importance sampling schemes based on moderate deviations subsolutions generally don’t perform as well as their large deviations counterparts, but the subsolutions themselves are typically easier to find. For this reason we recommend using importance sampling based on moderate deviations when finding large deviation subsolutions is prohibitively difficult, which can occur even in simple situations, or when considering events that are rare but not too rare so that the moderate deviation approximation centered around the law of large numbers limit captures the distributional properties which are important for determining the probability.

Contents

Vitae **iv**

Acknowledgements **v**

1 Introduction **1**

2 Preliminaries **7**

3 Moderate Deviations Principle **14**

3.1 Statement of Result 14

3.2 Laplace Upper Bound 15

3.3 Laplace Lower Bound 17

4 Importance Sampling Schemes **24**

4.1 Preliminaries and Statement of Result 24

4.2 Proof of Theorem 4.7 32

5 Invariance of Embedding **43**

6 Constructing Subsolutions **51**

7 Numerical Examples **58**

7.1 One-Dimension, Bernoulli Noise and Linear Drift 58

7.2 One-Dimension, Exponential Noise and Linear Driftl 61

7.3 Two-Dimensions with Linear Driftl 63

7.4 One-Dimension, State Dependent Noise and Nonlinear Driftl 66

A Tightness Proof **72**

B Terminal Position Cost Proof **89**

List of Tables

1 7.1 Numerical Results 61

2 7.2 Numerical Results 63

3 7.3 Numerical Results 65

4 7.4 Numerical Results 71

5 7.4 Numerical Results 71

6 7.4 Numerical Results 71

List of Figures

1 Terminal Condition Example ... 56

2 7.1 Terminal Condition ... 60

3 7.4 "Min" Terminal Condition ... 68

4 7.4 "MaxMin" Terminal Condition ... 70

1 INTRODUCTION

In this thesis we prove a moderate deviations principle and then study the effectiveness (including asymptotic optimality) of importance sampling schemes based on moderate deviations. The importance sampling results presented here involve $\mathbb{R}^d$-valued discrete time processes of the form

$$Z_{i+1} = Z_i + \epsilon\hat{b}(Z_i) + \epsilon v_i(Z_i),\ Z_0 = z_0, \tag{1.1}$$

where $\{v_i(\cdot)\}_{i\in\mathbb{N}_0}$ are zero mean random independent and identically distributed (iid) vector fields, and their continuous time piecewise linear interpolations $\{Z(t)\}_{0\le t\le T}$ with $Z(\epsilon i) = Z_i$ (see (2.7) for the precise definition). Here ϵ is some fixed small, positive real number. This could easily be built into the drift $\hat{b}(Z_i)$ and noise $v_i(Z_i)$, we include it to indicate that we are interested in processes which are close to $Z^0(t)$ given by the ordinary differential equation (ODE)

$$Z^0_t(t) = \hat{b}(Z^0(t)),\ Z_0 = z_0,$$

which would be the law of large numbers limit if we sent $\epsilon \to 0$. We specify that $Z(t)$ be close to its law of large numbers limit because our importance sampling changes of measure are based on the moderate deviations asymptotics of sequences of processes

$$X^n_{i+1} = X^n_i + \frac{1}{n}b(X^n_i) + \frac{1}{n}u_i(X^n_i),\ X^n_0 = x_0$$

for $n \in \mathbb{N}$, where the $\{u_i(\cdot)\}_{i\in\mathbb{N}_0}$ are zero mean iid vector fields. In order to do this we must embed the general process Z_i in such a sequence and use the resulting moderate deviations information for the sequence to suggest importance sampling schemes for the original process Z_i.

The large deviations of the continuous time linear interpolations $X^n \in C([0,T] : \mathbb{R}^d)$ have been studied extensively under various conditions (see, e.g., [1, 10, 12, 21, 23]). Here we introduce a scaling $a(n)$ satisfying $a(n) \to 0$ and $a(n)\sqrt{n} \to \infty$, and study the amplified difference between X^n and its noiseless version X^0:

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

We demonstrate, under weaker conditions on the noise $u_i(\cdot)$ than are necessary when considering X^n, that Y^n satisfies the large deviation principle on $C([0,T] : \mathbb{R}^d)$ with a "Gaussian" type rate function. As is customary for this type of scaling, we refer to this as moderate deviations.

To demonstrate this result we prove the equivalent Laplace principle, which involves evaluating limits of quantities of the form

$$a(n)^2 \log E\left[\exp\left\{-\frac{1}{a(n)^2}F(Y^n)\right\}\right]$$

when F is bounded and continuous. This is done by representing each of these quantities in terms of a stochastic control problem, and then using weak convergence methods as in [12]. Key results needed in this approach are establishing tightness of controls and controlled processes, and identifying their limits.

While one might expect the proof of this moderate deviations result to be similar to the corresponding large deviations result, there are important differences. For example, the tightness proof is significantly more complicated in the case of moderate deviations than it is in the case of large deviations. For large deviations one is able to establish an a priori bound on certain relative entropy costs associated with any sequence of nearly minimizing controls, and under this boundedness of the relative entropy costs, the empirical measures of the controlled driving noises as well as the controlled processes are tight. However, owing to the scaling in moderate deviations, even with the information that the analogous relative entropy costs decay like $O(1/a(n)^2 n)$, tightness of the empirical measures of the noises does not hold. Instead, one must consider empirical measures of the conditional means of the noises, and additional effort is required for the law of large numbers type result that shows that the conditional means are adequate to determine the limit. This extra difficulty arises for moderate deviations (even with the vanishing relative entropy costs), because the noise itself is being amplified by $a(n)\sqrt{n}$.

A second way in which the proofs for large and moderate deviations differ is in their treatment of degenerate noise, i.e., problems where the support of $u_i(\cdot)$ is not all of $\mathbb{R}^d$. This leads to significant difficulties in the proof of the large deviation lower bound, and requires a delicate and involved mollification argument. In contrast, the proof in the setting of moderate deviations, though more involved than the nondegenerate case, is much more straightforward.

The existing literature on moderate deviations considers various settings. Baldi [2] considers the same scaling used here but with no state dependence. For the empirical measure of a Markov chain, de Acosta [8] and de Acosta and Chen [7] prove lower and upper bounds, respectively. Guillin [25] considers inhomogeneous functionals of a "fast" continuous time ergodic Markov chain, and in [26] this is extended to a small noise diffusion whose coefficients depend on the "fast" Markov chain. There are also results for martingale differences such as Dembo [9], Gao [22], and Djellout [11]. For various reasons, the issues previously mentioned regarding the difficulties in the proof of the upper bound and the simplification in the lower bound for degenerate noise do not play a role in these papers. For instance, proving tightness in a moderate deviations setting for continuous time processes is typically much easier. This is because measures on path space that have bounded relative entropy with respect to a continuous time process generally have significantly less variability than those with bounded relative entropy with respect to a discrete time process. This is illustrated by the convenient alternative formulations of the relative entropy repesentation for some continuous time processes (see [5] for Brownian motion and [6] for Poisson random measures). The present situation essentially deals with central limit theorem and moderate deviations issues simultaneously. Although all proofs are given for the time interval $[0, 1]$, they extend with only notational differences to $[0, T]$ for any $T \in (0, \infty)$.

We then apply the moderate deviations information to the problem of importance sampling. Accurately estimating the probability of an event using Monte Carlo can be computationally demanding if the event is sufficiently rare. Importance sampling can reduce the number of samples needed by changing the underlying measure of the dynamics and then adjusting the estimate using the likelihood ratio. Its first use in the context of rare events appears in [28]. While a well chosen change of measure can improve the accuracy substantially, a poor choice can make the importance sampling scheme worse than standard Monte Carlo. It was generally expected that if the process could be embedded in a sequence satisfying a large deviations principle information contained in the large deviations rate function could be used to suggest effective changes of measure. For a discussion on some nonrigorous and in some sense misleading early approaches to the problem of algorithm design see [24]. It turns of that a rigorous and proper approach to design can be based on subsolutions to a Hamilton-Jacobi-Bellman (HJB) equation associated with the large deviation rate function [19] (see also the related notion of Lyapunov inequality [3]). This approach has been further developed in various ways and for many different process models [14, 13, 18, 19, 20, 17, 16, 15]. In all of these papers some form of "asymptotic optimality" of the scheme is considered as the large deviation parameter tends to its limit.

The use of asymptotic methods introduces a more simplistic approximate description, i.e., the large deviation approximation. As demonstrated in the references, in many cases this simplification can be exploited. Schemes can be explicitly constructed based on the asymptotic approximation, which are then shown to yield good importance sampling schemes for the original process. Of course a process model such as (1.1) can be embedded as an element of many different sequences with possibly different asymptotic behavoirs. The large deviation limit mentioned previously is one such sequence, and indeed a very natural one when the event of interest (i.e., the events that largely determine $Ee^{-\hat{F}(Z_N)}$) are rare. There are, however, other useful sequences into which (1.1) can be embedded which may suggest different importance sampling schemes. There are, of course, tradeoffs. The performance of differing schemes may be better or worse depending on how well the asymptotics capture the essential aspects of the distribution of (1.1). The schemes may also be simpler or more complicated to implement, depending on the complexity of the approximation.

In this thesis we investigate the use of moderate deviations as a tool for importance sampling scheme design. Thus we will embed (1.1) into a sequence of processes satisfying a moderate deviations principle, and choose a change of measure based on moderate deviation asymptotics. This will naturally limit the range of functionals $\hat{F}$ to those for which a moderate deviations approximation captures the distributionals properties of $\{Z_i\}$ that are important in determining $Ee^{-\hat{F}(Z_N)}$. Intuitively, these should be functionals and events determined by trajectories falling somewhere between a law of large numbers approximation and the full large deviation approximation (i.e., rare but not too rare). In exchange for this restriction on the class of functionals, one gets to work with an approximation (i.e., a rate function) that is typically more tractable than the large deviation rate.

One may ask if in this situation of "rare but not too rare" events it is really worth the bother to use importance sampling, and if instead one should use standard Monte Carlo. That will depend on the difficulty and cost of generating samples. One can imagine at least three scenarios where it is worth the bother. One is when the generation of even a single trajectory of $\{Z_i\}$ is time consuming, e.g., when $\{Z_i\}$ corresponds to a discrete time approximation to a stochastic partial differential equation. The second is when a quantity such as $Ee^{-\hat{F}(Z_N)}$ must be computed many times, such as when the output is used as part of an optimization scheme. A third is when the event of interest is only marginally within the moderate deviations regime, but schemes based on the large deviation rate are simply intractable, and the improvement may be sufficient to make the numerical problem feasible. However, the moderate deviation approximation is not a panacea, and there are many situations (e.g., in the analysis of metastability issues) for which it is simply inadequate.

In both the large deviations and moderate deviations approaches the construction of an importance sampling scheme is based on the construction of a subsolutions to a related nonlinear partial differential equation (PDE) associated with the rate function (see [19] for the case of large deviations). However, fewer restrictions are needed on the process dynamics to embed Z_i in a moderate deviations sequence, and in fact for some choices of the stochastic kernel $\hat{\mu}(\cdot\,|\,z)$ the moderate deviations scaling may be all that is available. In general one expects that the moderate deviations rate function to involve a linear approximation of the dynamics and a quadratic approximation of the costs found in the large deviations rate function, centered around the law of large numbers limit. This corresponds to a "local Gaussian" approximation, and the PDEs are those associated with the famous "linear-quadratic-regulator," whose solution can be constructed in terms of the solution to an appropriate Ricatti equation.

In an analagous manner to the proof of the moderate deviations principle, proving the asymptotic properties of importance sampling schemes in the moderate deviations setting involves challenges not found in the large deviations setting. Once again proving tightness of the empirical measures of the noises is not possible due to the moderate deviations scaling, and we must instead settle for tightness of the empirical measures of the conditional means of the noises. In addition, proving the large deviations principle for these types of processes requires assuming that the noise has a finite moment generating function everywhere. In the moderate deviations setting we make only the weaker assumption of a finite moment generating function in a neighborhood of the origin, and this makes the tightness proof more complex.

2 PRELIMINARIES

Let

$$Z_{i+1} = Z_i + \epsilon\hat{b}(Z_i) + \epsilon v_i(Z_i),\ Z_0 = z_0 \tag{2.1}$$

where the $\{v_i(\cdot)\}_{i\in\mathbb{N}_0}$ are zero mean iid vector fields with distribution given by the stochastic kernel $\hat{\mu}(\cdot|\, z)$. Let $\mathcal{B}(\mathbb{R}^d)$ be the Borel σ-algebra on $\mathbb{R}^d$. Then $\hat{\mu}(\cdot|\, z)$ is a probability measure on $\mathcal{B}(\mathbb{R}^d)$ for all $z \in \mathbb{R}^d$, $P(v_i(z) \in B) = \hat{\mu}(B|\, z)$ for all $z \in \mathbb{R}^d$, $B \in \mathcal{B}(\mathbb{R}^d)$ and $i \in \mathbb{N}_0$, and $z \to \hat{\mu}(B|\, z)$ is Borel measurable for all $B \in \mathcal{B}(\mathbb{R}^d)$. We will also consider sequences of processes

$$X_{i+1}^n = X_i^n + \frac{1}{n}b(X_i^n) + \frac{1}{n}u_i(X_i^n),\ X_0^n = x_0 \tag{2.2}$$

where the $\{u_i(\cdot)\}_{i\in\mathbb{N}_0}$ are zero mean iid vector fields with distribution given by the stochastic kernel $\mu(\cdot|\, x)$. We will make the following assumptions on $\mu(\cdot|\, x)$ and $b(x)$ in the case of $\{X_i^n\}_{n=1}^\infty$ and $\hat{\mu}(\cdot|\, z)$ and $\hat{b}(z)$ in the case of Z_i, although we will only state them for $\mu(\cdot|\, x)$ and $b(x)$ for brevity. Define

$$H_c(x,\alpha) \doteq \log\left(\int_{\mathbb{R}^d} e^{\langle u,\alpha\rangle}\mu(du|\, x)\right)$$

for $\alpha \in \mathbb{R}^d$. The subscript c reflects the fact that this log moment generating function uses the centered distribution $\mu(\cdot|\, x)$, rather than the usual $H(x,\alpha) = H_c(x,\alpha) + \langle\alpha, b(x)\rangle$. We will use the following.

Condition 2.1 • *There exists $\lambda > 0$ and $K_{mgf} < \infty$ such that*

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

- *$x \to \mu(\cdot|\, x)$ is continuous with respect to the topology of weak convergence.*

- *$b(x)$ is continuously differentiable, and the norms of both $b(x)$ and its derivative are uniformly bounded by some constant $K_b < \infty$.*

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

$$A_{ij}(x) \doteq \int_{\mathbb{R}^d} u_i u_j \mu(du|\, x),$$

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

$$\frac{\partial H_c(x, 0)}{\partial \alpha_i} = \int_{\mathbb{R}^d} u_i \mu(du|\, x) = 0$$

and

$$\frac{\partial^2 H_c(x, 0)}{\partial \alpha_i \partial \alpha_j} = \int_{\mathbb{R}^d} u_i u_j \mu(du|\, x) = A_{ij}(x)$$

for all $i, j \in \{1, \ldots, d\}$ and $x \in \mathbb{R}^d$, and that $A(x)$ is nonnegative-definite and symmetric. For $x \in \mathbb{R}^d$ we can therefore write

$$A(x) = Q(x)\Lambda(x)Q^T(x), \tag{2.4}$$

where $Q(x)$ is an orthogonal matrix whose columns are the eigenvectors of $A(x)$ and $\Lambda(x)$ is the diagonal matrix consisting of the eigenvalues of $A(x)$ in descending order.

Definition 2.2 *For a symmetric, nonnegative definite matrix A, $\|\alpha\|_{A^{-1}}^2$ will mean the value ∞ for $\alpha \in \mathbb{R}^d$ not in the linear span of the eigenvectors corresponding to the positive eigenvalues, and $\langle \alpha, A^{-1}\alpha \rangle$ for vectors $\alpha \in \mathbb{R}^d$ in that linear span where $A^{-1}\alpha$ is defined.*

Assumption (2.3) implies there exists some $K_{DA} < \infty$ and $\lambda_{DA} \in (0, \lambda]$ (independent of x) such that

$$\sup_{x \in \mathbb{R}^d} \sup_{\|\alpha\| \le \lambda_{DA}} \max_{i,j,k} \left| \frac{\partial^3 H_c(x,\alpha)}{\partial \alpha_i \partial \alpha_j \partial \alpha_k} \right| \le \frac{K_{DA}}{d^3}, \tag{2.5}$$

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

$$\frac{1}{2} \|\alpha\|^2_{A(x)} - \|\alpha\|^3 K_{DA} \le H_c(x,\alpha) \le \frac{1}{2} \|\alpha\|^2_{A(x)} + \|\alpha\|^3 K_{DA}. \tag{2.6}$$

These assumptions will allow us to prove a moderate deviations principle $\{X_i^n\}_{n=1}^{\infty}$, and will allow us to embed Z_i in a sequence satisfying a moderate deviations principle so that we can use that information to choose importance sampling schemes.

Define the continuous time linear interpolation of X_i^n by $X^n(i/n) = X_i^n$ for $i = 0, ..., n$ and

$$X^n(t) = (i+1-nt)X_i^n + (nt-i)X_{i+1}^n \tag{2.7}$$

for $t \in (i/n, i/n + 1/n)$. In addition, define

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

and let $X^{n,0}(t)$ be the analogously defined continuous time linear interpolation. Clearly $X^{n,0}(t) \to X^0(t)$ in $C([0,1] : \mathbb{R}^d)$, where

$$X^0(t) = \int_0^t b(X^0(s))ds + x_0.$$

Since $Eu_i(x) = 0$ for all $x \in \mathbb{R}^d$, we know that $X^n(t) \to X^0(t)$ in $C([0,1] : \mathbb{R}^d)$ in probability. One can estimate probabilities for events involving paths outside the law of large numbers limit X^0 by proving a large deviation principle and finding the corresponding rate function.

Definition 2.3 *Let $\{Z^n, \, n \in \mathbb{N}\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ and taking values in a Polish space $\mathcal{Z}$. A function $I : \mathcal{Z} \to [0, \infty]$ is called a rate function if for any $M < \infty$ the set $\{x : I(x) \le M\}$ is compact in $\mathcal{Z}$. The sequence $\{Z^n\}$ satisfies the large deviation principle on $\mathcal{Z}$ with rate function I and sequence $r(n)$ if the following two conditions hold.*

- *Large Deviation Upper Bound: for each closed subset F of $\mathcal{Z}$*

$$\limsup_{n\to\infty} r(n) \log P(Z^n \in F) \leq -\inf_{z\in F} I(z).$$

- *Large Deviation Lower Bound: for each open subset G of $\mathcal{Z}$*

$$\liminf_{n\to\infty} r(n) \log P(Z^n \in G) \geq -\inf_{z\in G} I(z).$$

Under significantly stronger assumptions, including the assumption

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

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

$$I_L(\phi) = \inf \left\{ \int_0^1 L_c(\phi(s), u(s))ds : \phi(t) = x_0 + \int_0^t b(\phi(s))ds \right.$$
$$\left. + \int_0^t u(s)ds, t \in [0,1] \right\}.$$

where

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

is the Legendre transform of $H_c(x,\alpha)$ [12, 29, 30, 31, 32].

Assume $a(n)$ satisfies

$$a(n) \to 0 \text{ and } a(n)\sqrt{n} \to \infty. \tag{2.8}$$

We define the rescaled difference

$$Y^n(t) = a(n)\sqrt{n}(X^n(t) - X^{n,0}(t)). \tag{2.9}$$

Note that

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

Due to Condition 2.1

$$\sup_{t\in[0,T]} \left\|X^{n,0}(t) - X^0(t)\right\| = O(1/n)$$

and it is straightforward to demonstrate that the results we prove for Y^n also hold for

$$Y_i^{l,n} = a(n)\sqrt{n}(X_i^n - X^0(i/n)).$$

For $\eta, \mu \in \mathcal{P}(\mathbb{R}^d)$ [the set of probability measures on $\mathcal{B}(\mathbb{R}^d)$] , the relative entropy of η with respect to μ is defined by

$$R(\eta\|\,\mu) \doteq \int_{\mathbb{R}^d} \log\left(\frac{d\eta}{d\mu}(x)\right)\eta(dx) \in [0,\infty]$$

if η is absolutely continuous with respect to μ, and $R(\eta\|\,\mu) \doteq \infty$ otherwise. In addition, for a general bounded, measurable function $G : \mathbb{R}^d \to \mathbb{R}$ we have the variational formula

$$-\log\left(\int_{\mathbb{R}^d} e^{-G(x)}\mu(dx)\right) = \inf_{\eta\in\mathcal{P}(\mathbb{R}^d)}\left\{\int_{\mathbb{R}^d} G(x)\eta(dx) + R(\eta\|\,\mu)\right\} \tag{2.10}$$

For a proof of this or for general properties of relative entropy we refer to [12, Section 1.4]. Throughout the paper we will use the following notation when we want a new process with the same structure as Y^n but driven by a new measure.

Construction 2.4 Given a probability measure $\eta \in \mathcal{P}((R^d)^n)$, let $(\bar{u}_0^\eta, ..., \bar{u}_{n-1}^\eta)$ have distribution η. Define

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

and the scaled difference

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

Define also the continuous time linear interpolations $\bar{X}^{n,\eta}(t)$ and $\bar{Y}^{n,\eta}(t)$ as in (2.7). Let η_i denote the conditional distribution of $\bar{u}_i^\eta$ given $(\bar{u}_0^\eta, ...\bar{u}_{i-1}^\eta)$, with the understanding that we generally supress the dependence on $(\bar{u}_0^\eta, ...\bar{u}_{i-1}^\eta)$ in the notation. Define the conditional means of the noises

$$w^{n,\eta}(t) \doteq \int_{\mathbb{R}^d} u\eta_i^n(du)\ for\ t \in \left[\frac{i}{n}, \frac{i+1}{n}\right),$$

the amplified conditional means

$$\hat{w}^{n,\eta}(t) \doteq a(n)\sqrt{n}w^{n,\eta}(t),$$

and random measures on $\mathbb{R}^d \otimes [0,1]$

$$\varsigma^{n,\eta}(dw \otimes dt) \doteq \delta_{\hat{w}^{n,\eta}(t)}(dw)dt = \delta_{a(n)\sqrt{n}w^{n,\eta}(t)}(dw)dt.$$

The variational formula (2.10) and chain rule [12, Theorem C.3.1] imply that

$$-a(n)^2 \log E\left[e^{-\frac{1}{a(n)^2}F(Y^n)}\right] = \inf_{\eta} E\left[\sum_{i=0}^{n-1} a(n)^2 R(\eta_i \| \, \mu(\cdot \,|\, \bar{X}_i^{n,\eta})) + F(\bar{Y}^{n,\eta})\right]$$

for any bounded, measurable $F : C([0,1] : \mathbb{R}^d) \to \mathbb{R}$. Both the proof of the moderate deviation principle as well as the asymptotic performance of a class of importance sampling changes of measure rely heavily on the following tightness result. The proof of this is rather long and consequnetly has been moved to Appendix A for the convenience of the reader.

Theorem 2.5 *Let $\{\eta^n\}$ be a sequence of measures, each $\eta^n \in \mathcal{P}((\mathbb{R}^d)^n)$, and define the corresponding random variables as in Construction 2.4. Assume that for some $K_E < \infty$*

$$\sup_{n\in\mathbb{N}} \left\{ a(n)^2 n E\left[\frac{1}{n}\sum_{i=0}^{n-1} R(\eta_i^n \| \, \mu(\cdot \,|\, \bar{X}_i^{n,\eta^n}))\right]\right\} \leq K_E. \tag{2.11}$$

Then the $\{\varsigma^{n,\eta^n}\}$ are uniformly integrable in the sense that

$$\lim_{C\to\infty} \limsup_{n\to\infty} E\left[\int_{\{w:\|w\|\geq C\}} \|w\| \, \varsigma^{n,\eta^n}(dw \otimes dt)\right] = 0 \tag{2.12}$$

and $\{(\varsigma^{n,\eta^n}, \bar{Y}^{n,\eta^n})\}_{n\in\mathbb{N}}$ is tight in $\mathcal{P}(\mathbb{R}^d \otimes [0,1]) \otimes C([0,1] : \mathbb{R}^d)$. Consider a subsequence (keeping the index n for convenience) such that $\{(\varsigma^{n,\eta^n}, \bar{Y}^{n,\eta^n})\}$ converges weakly to some $(\varsigma, \hat{Y})$. Then with *probability 1* $\varsigma_2(dt)$ *is Lebesgue measure and*

$$\hat{Y}(t) = \int_0^t Db(X^0(s))\hat{Y}(s)ds + \int_0^t \hat{w}(s)ds, \tag{2.13}$$

where

$$\hat{w}(t) = \int_{\mathbb{R}^d} w \varsigma_{1|2}(dw\,|t).$$

In addition,

$$\liminf_{n\to\infty} a(n)^2 nE\left[\frac{1}{n}\sum_{i=0}^{n-1} R(\eta_i^n \| \mu(\cdot | \bar{X}_i^{n,\eta^n}))\right] \geq E\left[\int_0^1 \frac{1}{2}\|\hat{w}(s)\|^2_{A^{-1}(X^0(s))}ds\right]. \tag{2.14}$$

Proof. The proof is in Appendix A. ■

3 MODERATE DEVIATIONS PRINCIPLE

3.1 Statement of Result

As noted in the introduction, the result stated below also holds with the interval $[0,1]$ replaced by $[0,T]$, $T \in (0,\infty)$. Let

$$\phi^u(t) = \int_0^t Db(X^0(s))\phi^u(s)ds + \int_0^t u(s)ds. \tag{3.1}$$

and let D denote the gradient operator.

Theorem 3.1 *Assume Condition 2.1. Then $\{Y^n\}_{n\in\mathbb{N}}$ satisfies the large deviation principle on $C([0,1]:\mathbb{R}^d)$ with sequence $a(n)^2$ and rate function*

$$I_M(\phi) = \inf\left\{\frac{1}{2}\int_0^1 \|u(t)\|^2 dt : \phi(t) = \int_0^t Db(X^0(s))\phi(s)ds \right.$$
$$\left. + \int_0^t A^{1/2}(X^0(s))u(s)ds, t \in [0,1]\right\}.$$

I_M is essentially the same as what one would obtain by using a linear approximation around the law of large numbers limit X^0 of the dynamics and a quadratic approximation of the costs in

I_L. To prove the LDP, it suffices to show the Laplace principle [12, Theorem 1.2.3]

$$\begin{aligned} \lim_{n\to\infty} -a(n)^2 \log E\left[e^{-\frac{1}{a(n)^2}F(Y^n)}\right] \\ = \inf_{u\in L^2([0,1]:\mathbb{R}^d)} \left\{\frac{1}{2}\int_0^1 \|u(s)\|^2\, ds + F\left(\phi^{A^{1/2}(X^0)u}\right)\right\} \end{aligned} \tag{3.2}$$

where $\phi^{A^{1/2}(X^0)u}$ is given by (3.1). From (2.10) we have

$$-a(n)^2 \log E\left[e^{-\frac{1}{a(n)^2}F(Y^n)}\right] = \inf_\eta E\left[\sum_{i=0}^{n-1} a(n)^2 R(\eta_i \| \, \mu(\cdot\,| \, \bar{X}_i^{n,\eta})) + F(\bar{Y}^{n,\eta})\right] \tag{3.3}$$

for any bounded, continuous $F : C([0,1] : \mathbb{R}^d) \to \mathbb{R}$, where η_i, $\bar{X}^{n,\eta}$ and $\bar{Y}^{n,\eta}$ are given by Construction 2.4 in terms of η.

We will prove (3.2) by proving the lower bound

$$\begin{aligned} \liminf_{n\to\infty} -a(n)^2 \log E\left[e^{-\frac{1}{a(n)^2}F(Y^n)}\right] \\ \geq \inf_{u\in L^2([0,1]:\mathbb{R}^d)} \left\{\frac{1}{2}\int_0^1 \|u(s)\|^2\, ds + F\left(\phi^{A^{1/2}(X^0)u}\right)\right\} \end{aligned} \tag{3.4}$$

and the upper bound

$$\begin{aligned} \limsup_{n\to\infty} -a(n)^2 \log E\left[e^{-\frac{1}{a(n)^2}F(Y^n)}\right] \\ \leq \inf_{u\in L^2([0,1]:\mathbb{R}^d)} \left\{\frac{1}{2}\int_0^1 \|u(s)\|^2\, ds + F\left(\phi^{A^{1/2}(X^0)u}\right)\right\}. \end{aligned} \tag{3.5}$$

The results of Theorem 2.5 will be used in the proofs of both of these bounds.

3.2 Laplace Upper Bound

The goal of this section is to prove (3.4), which due to the minus sign corresponds to the Laplace upper bound. Suppose for each n that η^n comes within ε of achieving the infimum in (3.3), so that

$$\begin{aligned} \liminf_{n\to\infty} -a(n)^2 \log E\left[e^{-\frac{1}{a(n)^2}F(Y^n)}\right] + \varepsilon \\ \geq \liminf_{n\to\infty} E\left[\sum_{i=0}^{n-1} a(n)^2 R(\eta_i^n \| \, \mu(\cdot\,| \, \bar{X}_i^{n,\eta^n})) + F(\bar{Y}^{n,\eta^n})\right]. \end{aligned} \tag{3.6}$$

Since $\sup_{x\in\mathbb{R}^d} |F(x)| \leq K_F$ for some $K_F < \infty$, we also have

$$\sup_n a(n)^2 n E\left[\sum_{i=0}^{n-1} \frac{1}{n} R(\eta_i^n \| \mu(\cdot | \bar{X}_i^{n,\eta^n}))\right] \leq 2K_F + \varepsilon.$$

Consequently we can choose a subsequence of $\{\eta^n\}$ (we retain n as the index for convenience) along which the conclusions of Theorem 2.5 hold. Combining this with (3.6) gives

$$\begin{aligned}
&\liminf_{n\to\infty} -a(n)^2 \log E\left[e^{-\frac{1}{a(n)^2}F(Y^n)}\right] + \varepsilon \\
&\quad \geq \liminf_{n\to\infty} E\left[\sum_{i=0}^{n-1} a(n)^2 R(\eta_i^n \| \mu(\cdot | \bar{X}_i^{n,\eta^n})) + F(\bar{Y}^{n,\eta^n})\right] \\
&\quad \geq E\left[\int_0^1 \frac{1}{2} \|\hat{w}(s)\|^2_{A^{-1}(X^0(s))} ds + F(\hat{Y})\right].
\end{aligned}$$

Recalling

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

it follows that

$$\begin{aligned}
&E\left[\int_0^1 \frac{1}{2} \|\hat{w}(s)\|^2_{A^{-1}(X^0(s))} ds + F(\hat{Y})\right] \\
&\quad \geq \inf_{u\in L^2([0,1]:\mathbb{R}^d)} \left\{\int_0^1 \frac{1}{2} \|u(s)\|^2_{A^{-1}(X^0(s))} ds + F(\phi^u)\right\} \\
&\quad = \inf_{u\in L^2([0,1]:\mathbb{R}^d)} \left\{\int_0^1 \frac{1}{2} \|u(s)\|^2 ds + F\left(\phi^{A^{1/2}(X^0)u}\right)\right\},
\end{aligned}$$

with ϕ^u defined as in (3.1). Since $\varepsilon > 0$ is arbitrary, we have the lower bound (3.4).

3.3 Laplace Lower Bound

The goal of this section is to prove (3.5). Note that for $u, v \in L^2([0,1] : \mathbb{R}^d)$ with $\phi^{A^{1/2}(X^0)u}$ and $\phi^{A^{1/2}(X^0)v}$ given by (3.1)

$$\begin{aligned}
&\phi^{A^{1/2}(X^0)u}(t) - \phi^{A^{1/2}(X^0)v}(t) \\
&\quad = \int_0^t Db(X^0(s)) \left(\phi^{A^{1/2}(X^0)u}(s) - \phi^{A^{1/2}(X^0)v}(s)\right) ds \\
&\qquad + \int_0^t (A^{1/2}(X^0(s)u(s) - A^{1/2}(X^0(s)v(s))ds.
\end{aligned}$$

Thus by Gronwall's inequality

$$\begin{aligned}
&\sup_{t\in[0,1]} \left\|\phi^{A^{1/2}(X^0)u}(t) - \phi^{A^{1/2}(X^0)v}(t)\right\| \\
&\quad \leq (1 + K_b e^{K_b}) \int_0^1 \left\|A^{1/2}(X^0(s)u(s) - A^{1/2}(X^0(s)v(s)\right\| ds \\
&\quad \leq (1 + K_b e^{K_b}) K_A^{1/2} \left(\int_0^1 \|u(s) - v(s)\|^2 ds\right)^{\frac{1}{2}}.
\end{aligned} \tag{3.7}$$

Since $C([0,1] : \mathbb{R}^d)$ is dense in $L^2([0,1] : \mathbb{R}^d)$, the proof of the Laplace lower bound is reduced to showing that for an arbitrary $u \in C([0,1] : \mathbb{R}^d)$

$$\limsup_{n\to\infty} -a(n)^2 \log E\left[e^{-\frac{1}{a(n)^2}F(Y^n)}\right] \leq \frac{1}{2}\int_0^1 \|u(s)\|^2 ds + F\left(\phi^{A^{1/2}(X^0)u}\right). \tag{3.8}$$

The main difficulty is to deal with the possible degeneracy of the noise. Recall the orthogonal decomposition of $A(x)$ (2.4). Define

$$A_K^{-1}(x) = Q(x)\Lambda_K^{-1}(x)Q^T(x)$$

where $\Lambda_K^{-1}(x)$ is the diagonal matrix such that $\Lambda_{ii,K}^{-1}(x) = \Lambda_{ii}^{-1}(x)$ when $\Lambda_{ii}^{-1}(x) \leq K^2$ and $\Lambda_{ii,K}^{-1}(x) = K^2$ when $\Lambda_{ii}^{-1}(x) > K^2$. Note that by [27, Theorem 6.2.37] $A^{1/2}(x)$, $A_K^{-1}(x)$ and $A_K^{1/2}(x)$ are continuous functions of $A(x)$, and consequently they are also continuous functions of

$x \in \mathbb{R}^d$. In addition define

$$u_K(s) = \begin{cases} u(s) & \text{for } \|u(s)\| \leq K \\ \frac{Ku(s)}{\|u(s)\|} & \text{for } \|u(s)\| > K \end{cases}.$$

Let $\phi^{u,K}(t) = \phi^{A(X^0)A_K^{-1/2}(X^0)u_K}(t)$ recalling (3.1), and note that $\phi^{u,K}$ solves

$$\phi^{u,K}(t) = \int_0^t Db(X^0(s))\phi^{u,K}(s)ds + \int_0^t A(X^0(s))A_K^{-1/2}(X^0(s))u_K(s)ds. \tag{3.9}$$

To simplify notation we define $s_i^n \doteq i/n$ and $s^n(t) = \lfloor nt \rfloor /n$, where $\lfloor a \rfloor$ is the integer part of a. Note that $s^n(t) - t \to 0$ uniformly for $t \in [0,1]$ as $n \to \infty$. For n sufficiently large

$$\max_{0 \leq i \leq n-1} \left\{ \frac{1}{a(n)\sqrt{n}} \left\| A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right) \right\| \right\} \leq \frac{1}{a(n)\sqrt{n}} K^2 \leq \lambda_{DA}$$

and we can define the sequence $\{(\bar{X}^{n,\eta^{n,u,K}}, \bar{Y}^{n,\eta^{n,u,K}}, \eta^{n,u,K}, \varsigma^{n,\eta^{n,u,K}})\}$ as in Construction 2.4 where $\eta^{n,u,K}$ is defined in terms of the conditional distributions

$$\begin{aligned} &\eta_i^{n,u,K}(dv) \\ &\quad = \exp\left\{ \left\langle v, \frac{1}{a(n)\sqrt{n}} A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right) \right\rangle \right. \\ &\qquad \left. - H_c\left(\bar{X}_i^{n,\eta^{n,u,K}}, \frac{1}{a(n)\sqrt{n}} A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right) \right) \right\} \mu(dv|\, \bar{X}_i^{n,\eta^{n,u,K}}). \end{aligned}$$

Using (2.5) and the fact that

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

we have for $\|\alpha\| \leq \lambda_{DA}$

$$\left\| \int_{\mathbb{R}^d} u \exp\{\langle u, \alpha \rangle - H_c(x,\alpha)\} \mu(du|\, x) - A(x)\alpha \right\| \leq K_{DA} \|\alpha\|^2. \tag{3.10}$$

The next result identifies the limit in probability of the controlled processes and an asymptotic bound for the relative entropies.

Theorem 3.2 *Let $u \in C([0,1]:\mathbb{R}^d)$ and $K \in (0,\infty)$ be given, construct $\{(\bar{X}^{n,\eta^{n,u,K}}, \bar{Y}^{n,\eta^{n,u,K}}, \eta^{n,u,K}, \varsigma^{n,\eta^{n,u,K}})\}$ as in this section and define $\phi^{u,K}$ by (3.9). Then*

$$\bar{Y}^{n,\eta^{n,u,K}} \to \phi^{u,K} \tag{3.11}$$

in $C([0,1]:\mathbb{R}^d)$ in probability, and

$$\begin{aligned}\limsup_{n\to\infty} a^2(n) n E\left[\frac{1}{n}\sum_{i=0}^{n-1} R\left(\eta_i^{n,u,K} \middle\| \mu(\cdot| \bar{X}_i^{n,\eta^{n,u,K}})\right)\right] \\ \le \frac{1}{2}\int_0^1 \left\| A_K^{-1/2}(X^0(s))u_K(s)\right\|^2_{A(X^0(s))} ds.\end{aligned} \tag{3.12}$$

Proof. From (2.6) and (3.10) we have for all n sufficiently large that $\frac{1}{a(n)\sqrt{n}}K^2 \le \lambda_{DA}$

$$\begin{aligned}
&R\left(\eta_i^{n,u,K} \middle\| \mu(\cdot| \bar{X}_i^{n,\eta^{n,u,K}})\right) \\
&\quad = \int_{\mathbb{R}^d}\left\langle y, \frac{1}{a(n)\sqrt{n}} A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right)\right\rangle \eta_i^{n,u,K}(dy) \\
&\qquad - H_c\left(\bar{X}_i^{n,\eta^{n,u,K}}, \frac{1}{a(n)\sqrt{n}} A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right)\right) \\
&\quad \le \left\langle \frac{1}{a(n)\sqrt{n}} A\left(\bar{X}_i^{n,\eta^{n,u,K}}\right) A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right), \right. \\
&\qquad\qquad \left. \frac{1}{a(n)\sqrt{n}} A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right)\right\rangle \\
&\qquad - \frac{1}{2}\left\langle \frac{1}{a(n)\sqrt{n}} A\left(\bar{X}_i^{n,\eta^{n,u,K}}\right) A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right), \right. \\
&\qquad\qquad \left. \frac{1}{a(n)\sqrt{n}} A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right)\right\rangle + \frac{2}{a(n)^3 n^{3/2}} K_{DA} K^6 \\
&\quad = \frac{1}{2a(n)^2 n}\left\| A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right)\right\|^2_{A(\bar{X}_i^{n,\eta^{n,u,K}})} + \frac{2}{a(n)^3 n^{3/2}} K_{DA} K^6.
\end{aligned}$$

Consequently

$$\begin{aligned}\limsup_{n\to\infty} a^2(n) n E\left[\frac{1}{n}\sum_{i=0}^{n-1} R\left(\eta_i^{n,u,K} \middle\| \mu(\cdot| \bar{X}_i^{n,\eta^{n,u,K}})\right)\right] \\ \le \limsup_{n\to\infty} \frac{1}{2} E\left[\frac{1}{n}\sum_{i=0}^{n-1} \left\| A_K^{-1/2}\left(X^0\left(s_i^n\right)\right) u_K\left(s_i^n\right)\right\|^2_{A(\bar{X}_i^{n,\eta^{n,u,K}})}\right],\end{aligned} \tag{3.13}$$

where in fact

$$\limsup_{n\to\infty}\frac{1}{2}E\left[\frac{1}{n}\sum_{i=0}^{n-1}\left\|A_K^{-1/2}\left(X^0\left(s_i^n\right)\right)u_K\left(s_i^n\right)\right\|^2_{A(\bar{X}_i^{n,\eta^{n,u,K}})}\right]\leq\frac{1}{2}K^4K_A$$

(recall (2.3)). Therefore (2.11) is satisfied by $\{\eta^{n,u,K}\}$, so we can apply Theorem 2.5 and choose a subsequence (keeping n as the index for convenience) along which $\{(\varsigma^{n,\eta^{n,u,K}},\bar{Y}^{n,\eta^{n,u,K}})\}$ converges weakly to some limit $(\varsigma^{u,K},\hat{Y}^{u,K})$, where $\varsigma_2^{u,K}$ is Lebesgue measure and

$$\hat{Y}^{u,K}(t)=\int_0^t Db(X^0(s))\hat{Y}^{u,K}(s)ds+\int_0^t\int_{\mathbb{R}^d}w\varsigma_{1|2}^{u,K}(dw\,|s)ds.$$

This implies

$$\sup_{t\in[0,1]}\left\|\bar{X}^{n,\eta^{n,u,K}}(t)-X^0(t)\right\|\to 0$$

in probability. Because of this, the uniform bound on $A^{1/2}(x)$ and the continuity of $A^{1/2}(x)$, we have (recall that $s^n(t)\doteq\lfloor nt\rfloor/n$)

$$\sup_{t\in[0,1]}\left\|A^{1/2}(\bar{X}^{n,\eta^{n,u,K}}(s^n(t)))-A^{1/2}(X^0(s^n(t)))\right\|\to 0$$

in probability. However, the continuity of $A^{1/2}(X^0)A_K^{-1/2}(X^0)u_K$ gives

$$\begin{aligned}\sup_{t\in[0,1]}\Big\|&A^{1/2}(X^0(s^n(t)))A_K^{-1/2}(X^0(s^n(t)))u_K(s^n(t))\\&-A^{1/2}(X^0(t))A_K^{-1/2}(X^0(t))u_K(t)\Big\|\to 0.\end{aligned}$$

Combining these limits, and using the fact that $A_K^{-1/2}(X^0)u_K$ is uniformly bounded, shows that

$$\begin{aligned}\sup_{t\in[0,1]}\Big\|&A^{1/2}(\bar{X}^{n,\eta^{n,u,K}}(s^n(t)))A_K^{-1/2}(X^0(s^n(t)))u_K(s^n(t))\\&-A^{1/2}(X^0(t))A_K^{-1/2}(X^0(t))u_K(t)\Big\|\to 0\end{aligned}\tag{3.14}$$

in probability. This combined with the uniform bounds allows us to use dominated convergence to get

$$\limsup_{n\to\infty} E\left[\frac{1}{2}\int_0^1 \left\| A_K^{-1/2}(X^0(s^n(t)))u_K(s^n(t))\right\|^2_{A(\bar{X}^{n,\eta^{n,u,K}}(s^n(t)))} dt\right]$$
$$= \frac{1}{2}\int_0^1 \left\| A_K^{-1/2}(X^0(t))u_K(t)\right\|^2_{A(X^0(t))} dt.$$

Combining this with (3.13) shows (3.12).

To prove (3.11) we will show that in fact

$$\varsigma^{u,K}(dv\otimes dt) = \delta_{A(X^0(t))A_K^{-1/2}(X^0(t))u_K(t)}(dv)dt.$$

For all $\sigma > 0$ let

$$G_\sigma = \left\{(z,t)\in\mathbb{R}^d\times[0,1] : \left\| z - A(X^0(t))A_K^{-1/2}(X^0(t))u_K(t)\right\| \le \sigma\right\},$$

and note that by weak convergence $\limsup_{n\to\infty} E[\varsigma^{n,\eta^{n,u,K}}(G_\sigma)] \le E[\varsigma^{u,K}(G_\sigma)]$. Note also that

$$E[\varsigma^{n,\eta^{n,u,K}}(G_\sigma)]$$
$$\ge P\left[\sup_{t\in[0,1]}\left\| a(n)\sqrt{n}\int_{\mathbb{R}^d} v\eta^{n,u,K}_{\lfloor nt\rfloor}(dv) - A(X^0(t))A_K^{-1/2}(X^0(t))u_K(t)\right\| \le \sigma\right].$$

However, by (3.10) we can choose n large enough to make

$$\sup_{t\in[0,1]}\left\| a(n)\sqrt{n}\int_{\mathbb{R}^d} v\eta^{n,u,K}_{\lfloor nt\rfloor}(dv)\right.$$
$$\left. -A\left(\bar{X}^{n,\eta^{n,u,K}}(s^n(t))\right)A_K^{-1/2}\left(X^0\left(s^n(t)\right)\right)u_K\left(s^n(t)\right)\right\|$$

arbitrarily small, and the proof that

$$\sup_{t\in[0,1]}\left\| A(\bar{X}^{n,\eta^{n,u,K}}(s^n(t)))A_K^{-1/2}(X^0(s^n(t)))u_K(s^n(t))\right.$$
$$\left. -A(X^0(t))A_K^{-1/2}(X^0(t))u_K(t)\right\| \to 0$$

in probability is identical to the proof of (3.14). Therefore $\limsup_{n\to\infty} E[\varsigma^{n,\eta^{n,u,K}}(G_\sigma)] = 1$ for all

$\sigma > 0$, and so $E[\varsigma^{u,K}(\cap_{n\in\mathbb{N}} G_{1/n})] = 1$. This implies that with probability 1

$$\varsigma^{u,K}_{1|2}(\,dy|\,t) = \delta_{A(X^0(t))A_K^{-1/2}(X^0(t))u_K}(dy)$$

for a.e. t. It follows that

$$\hat{Y}^{u,K}(t) = \int_0^t Db(X^0(s))\hat{Y}^{u,K}(s)ds + \int_0^t A(X^0(s))A_K^{-1/2}(X^0(s))u_K(s)ds,$$

and therefore $\bar{Y}^{n,\eta^{n,u,K}} \to \phi^{u,K}$ weakly. This implies (3.11) and completes the proof. ■

The second theorem in this section allows us to approximate $F(\phi^{A^{1/2}(X^0)u})$ by $F(\phi^{u,K})$ and $\frac{1}{2}\int_0^1 \|u(s)\|^2\,ds$ by

$$\frac{1}{2}\int_0^1 \left\| A_K^{-1/2}(X^0(s))u_K(s)\right\|^2_{A(X^0(s))} ds.$$

Theorem 3.3 *Let $u \in C([0,1]:\mathbb{R}^d)$ and define $\phi^{A^{1/2}(X^0)u}$ by (3.1) and $\phi^{u,K}$ by (3.9). Then*

$$\phi^{u,K} \to \phi^{A^{1/2}(X^0)u}$$

in $C([0,1]:\mathbb{R}^d)$ and

$$\sup_{K\in(0,\infty)} \frac{1}{2}\int_0^1 \left\| A_K^{-1/2}(X^0(s))u_K(s)\right\|^2_{A(X^0(s))} ds \le \frac{1}{2}\int_0^1 \|u(s)\|^2\,ds.$$

Proof. Note that

$$\left\| A^{1/2}(X^0(s))A_K^{-1/2}(X^0(s))u_K(s)\right\| \le \|u(s)\|$$

for all $s \in [0,1]$ and $K \in (0,\infty)$ so

$$\sup_{K\in(0,\infty)} \frac{1}{2}\int_0^1 \left\| A_K^{-1/2}(X^0(s))u_K(s)\right\|^2_{A(X^0(s))} ds \le \frac{1}{2}\int_0^1 \|u(s)\|^2\,ds. \tag{3.15}$$

In addition,

$$A^{1/2}(X^0(s))A^{1/2}(X^0(s))A_K^{-1/2}(X^0(s))u_K(s) \to A^{1/2}(X^0(s))u(s)$$

and

$$\left\| A^{1/2}(X^0(s))A^{1/2}(X^0(s))A_K^{-1/2}(X^0(s))u_K(s)\right\| \le \left\| A^{1/2}(X^0(s))u(s)\right\|$$

for all $s \in [0,1]$ so dominated convergence gives

$$A^{1/2}(X^0)A^{1/2}(X^0)A_K^{-1/2}(X^0)u_K \to A^{1/2}(X^0)u$$

in $L^1([0,1]:\mathbb{R}^d)$. Combining this with the second line of (3.7) and using the definition of $\phi^{u,K}$ in (3.9) shows that

$$\phi^{u,K} \to \phi^{A^{1/2}(X^0)u}$$

in $C([0,1]:\mathbb{R}^d)$. ■

Using (3.3) and the fact that any given control is suboptimal,

$$\begin{aligned}
&- a(n)^2 \log E\left[e^{-\frac{1}{a(n)^2}F(Y^n)}\right] \\
&\leq E\left[\sum_{i=0}^{n-1} a(n)^2 R\left(\eta_i^{n,u,K} \middle\| \mu(\cdot\,|\,\bar{X}_i^{n,\eta^{n,u,K}})\right) + F(\bar{Y}^{n,\eta^{n,u,K}})\right].
\end{aligned}$$

Using Theorem 3.2, this implies

$$\begin{aligned}
&\limsup_{n\to\infty} -a(n)^2 \log E\left[e^{-\frac{1}{a(n)^2}F(Y^n)}\right] \\
&\leq \frac{1}{2}\int_0^1 \left\| A_K^{-1/2}(X^0(s))u_K(s)\right\|_{A(X^0(s))}^2 ds + F(\phi^{u,K}).
\end{aligned}$$

Sending $K \to \infty$ and using Theorem 3.3 gives (3.8), and hence completes the proof of the lower bound (3.5).

4 IMPORTANCE SAMPLING SCHEMES

4.1 Preliminaries and Statement of Result

Recall from Sections 1 and 2 that in order to choose an importance sampling scheme for the process Z_i given by (2.1) we study the asymptotic properties of schemes acting on sequences of processes $Y^n(t)$ given by (2.9) which satisfy the moderate deviations principle given in Chapter 3. The Laplace principle (3.2) indicates that for any bounded, continuous function $G : \mathbb{R}^d \to \mathbb{R}$

$$\begin{aligned}
&\lim_{n\to\infty} -a(n)^2 \log E e^{-\frac{1}{a(n)^2} G(Y^n(T))} \\
&\quad = \inf \left\{ \frac{1}{2} \int_0^T \|u(t)\|^2 \, dt + G(\phi(T)) : \phi(t) = \int_0^t Db(X^0(s))\phi(s) ds \right. \\
&\qquad\quad \left. + \int_0^t A^{1/2}(X^0(s)) u(s) ds, t \in [0,T] \right\}.
\end{aligned}$$

However, we are interested in the larger class of functions which are lower semicontinuous and bounded from below. Let F be one such function. Then the Laplace principle upper bound can be extended using approximation by uniformly bounded, Lipshitz continuous functions [12, Theorem 1.3.6] to give

$$
\begin{aligned}
&\liminf_{n\to\infty} -a(n)^2 \log E e^{-\frac{1}{a(n)^2}F(Y^n(T))} \qquad (4.1)\\
&\quad\geq \inf\left\{\frac{1}{2}\int_0^T \|u(t)\|^2\, dt + F(\phi(T)) : \phi(t) = \int_0^t Db(X^0(s))\phi(s)ds\right.\\
&\qquad\qquad \left. + \int_0^t A^{1/2}(X^0(s))u(s)ds, t\in[0,T]\right\}.
\end{aligned}
$$

Define

$$
\begin{aligned}
V(y,t) = \inf &\left\{\frac{1}{2}\int_t^T \|u(r)\|^2\, dr + F(\phi(T)) : \phi(s) = y + \int_t^s Db(X^0(r))\phi(r)ds\right.\\
&\left. + \int_t^s A^{1/2}(X^0(r))u(r)dr, s\in[t,T]\right\} \qquad (4.2)
\end{aligned}
$$

and note that (4.1) gives

$$
\liminf_{n\to\infty} -a(n)^2 \log E e^{-\frac{1}{a(n)^2}F(Y^n(T))} \geq V(0,0).
$$

We will use various families of changes of measure for importance sampling on the sequence $\{Y^n\}_{n=1}^\infty$. As in [19], each change of measure will be obtained from a subsolution to an associated HJB equation. The subsolutions are often constructed as minimums or maximums of some finite collection of smooth (classical sense) subsolutions. When a single smooth subsolution can be used, the change of measure can be identified as an "exponential tilt" of the underlying true distributions, with the tilt determined by the gradient of the subsolution. When the minimum of smooth subsolutions is used, one molifies it to get another smooth subsolution. The mollification procedure is outlined below, and the implementation of this mollification can be done in either the traditional or a randomized way which will be explained later (see also [19]). We will first state the partial differential equation (PDE) dynamics and give the definition of a classical sense subsolution to those dynamics. A classical sense solution to the PDE dynamics is smooth and satisfies

$$
W_t(y,t) = \frac{1}{2}\|DW(y,t)\|^2_{A(X^0(t))} - \left\langle DW(y,t), Db(X^0(t))y\right\rangle \qquad (4.3)
$$

and a classical sense subsolution to this PDE is smooth and satisfies

$$W_t\left(y,t\right) \geq \frac{1}{2}\left\|DW\left(y,t\right)\right\|^2_{A(X^0(t))} - \left\langle DW\left(y,t\right), Db(X^0(t))y\right\rangle . \tag{4.4}$$

For $\alpha \in C(\mathbb{R}^d \times [0,T] : \mathbb{R}^d)$ let

$$\left\|\alpha\right\|_\infty \doteq \sup_{(y,t)\in\mathbb{R}^d\times[0,T]} \left\|\alpha\left(y,t\right)\right\| . \tag{4.5}$$

The following construction defines the class of smooth subsolutions we use for importance sampling changes of measure.

Construction 4.1 Let

$$\mathcal{S}_0 \doteq \left\{W \in C^1(\mathbb{R}^d \times [0,T] : \mathbb{R}) : W \textit{ satisfies (4.4)and } \left\|DW\right\|_\infty < \infty\right\}.$$

For $W \in \mathcal{S}_0$ define the corresponding importance sampling random variables $\{r^n\}_{n=1}^\infty$ according to

$$r^n \doteq e^{-\frac{1}{a(n)^2}F(\bar{Y}^{n,\gamma^n}(T))} \prod_{i=0}^{\lfloor Tn \rfloor} \left(\frac{d\gamma_i^n\left(\cdot \middle| \bar{X}_i^{n,\gamma^n}\right)}{d\mu(\cdot| \bar{X}_i^{n,\gamma^n})}(\bar{u}_i^{\gamma^n}) \right)^{-1},$$

where the distribution of $\bar{u}_i^{\gamma^n}$, given $\bar{u}_j^{\gamma^n}$, $j = 0, \ldots, i-1$, is defined by

$$\begin{aligned}\gamma_i^n\left(du\middle| \bar{X}_i^{n,\gamma^n}\right) = &\exp\left\{\left\langle u, -\frac{1}{a(n)\sqrt{n}}DW\left(\bar{Y}_i^{n,\gamma^n}, \frac{i}{n}\right)\right\rangle\right. \\ &\left. -H_c\left(\bar{X}_i^{n,\gamma^n}, -\frac{1}{a(n)\sqrt{n}}DW\left(\bar{Y}_i^{n,\gamma^n}, \frac{i}{n}\right)\right)\right\}\mu\left(du\middle| \bar{X}_i^{n,\gamma^n}\right),\end{aligned}$$

and $\left(\bar{u}^{\gamma^n}, \bar{X}^{n,\gamma^n}, \bar{Y}^{n,\gamma^n}\right)$ are defined in terms of γ^n as in Construction 2.4, with γ^n defined implicitly in terms of the conditionals γ_i^n.

Note that for n sufficiently large $\frac{1}{a(n)\sqrt{n}}\left\|DW\right\|_\infty < \lambda$ (recall (2.3)) so this is well defined. Clearly the estimates are unbiased, meaning

$$E[r^n] = E\left[e^{-\frac{1}{a(n)^2}F(Y^n(T))}\right].$$

For this reason minimizing the variance is equivalent to minimizing the second moment. Consequently we evaluate the effectiveness of these schemes by evaluating

$$\liminf -a(n)^2 \log E[(r^n)^2] \tag{4.6}$$

and naturally prefer higher values since that means a faster rate of decay. Note that we have not yet specified the terminal conditions for the subsolutions $W \in \mathcal{S}_0$. Specifying that the subsolution is less than or equal to the function F,

$$W(y,T) \leq F(y) \text{ for all } y \in \mathbb{R}^d, \tag{4.7}$$

allows us to get a lower bound on (4.6). Consequently we will also define the following class of subsolutions

Definition 4.2 Let

$$\mathcal{S} \doteq \left\{ W \in \mathcal{S}_0 : W(y,T) \leq F(y) \textit{ for all } y \in \mathbb{R}^d \right\}.$$

In order to find a subsolution in $\mathcal{S}_0$ which satisfies this terminal condition it may be useful to take mollified minimums and maximums of subsolultions in $\mathcal{S}_0$ as long as these mollifications remain in $\mathcal{S}_0$. We introduce the following mollifications which accomplish this.

Definition 4.3 *Given K functions $f^k : \mathbb{R}^d \otimes [0,T] \to \mathbb{R}$ define the smooth approximation to* $\min(\{f^k\}_{k=1}^K)$

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

and the smooth approximation to $\max(\{f^k\}_{k=1}^K)$

$$U^{\delta,\max}(\{f^k\}_{k=1}^K)(y,t) = \delta \log \left\{ \sum_{k=1}^K \frac{1}{K} e^{\frac{1}{\delta} f^k(y,t)} \right\}.$$

In addition, define the weightings

$$\rho_k^{\delta,\min}(\{f^k\}_{k=1}^K)(y,t) = e^{-\frac{1}{\delta}f^k(y,t)} \Big/ \sum_{k=1}^K e^{-\frac{1}{\delta}f^k(y,t)}$$

and

$$\rho_k^{\delta,\max}(\{f^k\}_{k=1}^K)(y,t) = e^{\frac{1}{\delta}f^k(y,t)} \Big/ \sum_{k=1}^K e^{\frac{1}{\delta}f^k(y,t)} .$$

Note that

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

and

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

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

$$DU(y,t) = \sum_{k=1}^K \rho_k(y,t) DW^k(y,t) \tag{4.8}$$

and

$$U_t(y,t) = \sum_{k=1}^K \rho_k(y,t) W_t^k(y,t) \tag{4.9}$$

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

$$\begin{aligned} &\mathcal{W}(\mathbb{R}^d \times [0,T]) \\ &\quad \doteq \left\{ \rho \in C(\mathbb{R}^d \times [0,T] : [0,1]^K), \sum_{k=1}^K \rho_k(y,t) = 1 \text{ for all } (y,t) \in \mathbb{R}^d \times [0,T] \right\} \end{aligned} \tag{4.10}$$

be the class of continuous weightings and note that $\rho^{\delta,\min}(\{W^k\}_{k=1}^K)(y,t), \rho^{\delta,\max}(\{W^k\}_{k=1}^K)(y,t) \in \mathcal{W}(\mathbb{R}^d \times [0,T])$. The next lemma states that the class $\mathcal{S}_0$ is closed under these types of mollifications.

Lemma 4.4 *If $W^k \in \mathcal{S}_0$ for $k = 1, ..., K$ then $U^{\delta,\min}(\{W^k\}_{k=1}^K), U^{\delta,\max}(\{W^k\}_{k=1}^K) \in \mathcal{S}_0$.*

Proof. For convenience denote $\left(U^{\delta,\min}(\{W^k\}_{k=1}^K), \rho^{\delta,\min}(\{W^k\}_{k=1}^K), \{W^k\}_{k=1}^K\right)$ or

$\left(U^{\delta,\max}(\{W^k\}_{k=1}^K), \rho^{\delta,\max}(\{W^k\}_{k=1}^K), \{W^k\}_{k=1}^K\right)$ by $(U, \rho, \{W^k\}_{k=1}^K)$. Recall that $(U, \rho, \{W^k\}_{k=1}^K)$ satisfies (4.8) and (4.9). Clearly $\rho_k \in C(\mathbb{R}^d \times [0,T] : \mathbb{R})$ for $k = 1, ..., K$ and by assumption $W^k \in C^1(\mathbb{R}^d \times [0,T] : \mathbb{R})$ for $k = 1, ..., K$ so $U \in C^1(\mathbb{R}^d \times [0,T] : \mathbb{R})$. In addition $\|DU\|_\infty < \infty$ follows from the fact that $\left\|DW^k\right\|_\infty < \infty$ for $k = 1, ..., K$ and $\rho \in \mathcal{W}(\mathbb{R}^d \times [0,T])$. Also U automatically also satisfies (4.4) since the convexity of $\|\alpha\|_A^2$ implies

$$
\begin{aligned}
&U_t(y,t) \\
&\quad = \sum_{k=1}^K \rho_k(y,t) W_t^k(y,t) \\
&\quad \geq \sum_{k=1}^K \rho_k(y,t) \left(\frac{1}{2} \left\| DW^k(y,t) \right\|_{A(X^0(t))}^2 - \left\langle DW^k(y,t), Db(X^0(t))y \right\rangle \right) \\
&\quad \geq \frac{1}{2} \left\| \sum_{k=1}^K \rho_k(y,t) DW^k(y,t) \right\|_{A(X^0(t))}^2 - \left\langle \sum_{k=1}^K \rho_k(y,t) DW^k(y,t), Db(X^0(t))y \right\rangle \\
&\quad = \frac{1}{2} \|DU(y,t)\|_{A(X^0(t))}^2 - \left\langle DU(y,t), Db(X^0(t))y \right\rangle.
\end{aligned}
$$

This completes the proof. ■

Because the first derivatives of $U^{\delta,\max}(\{W^k\}_{k=1}^K)$ and $U^{\delta,\max}(\{W^k\}_{k=1}^K)$ are weighted combinations of the first derivatives of $W^1, ..., W^K$, with weights given by $\rho_k^{\delta,\min}(\{W^k\}_{k=1}^K)$ and $\rho_k^{\delta,\max}(\{W^k\}_{k=1}^K)$, respectively, the exponential tilt change of measure given by Construction 4.1 for one of these mollifications is a weighted combination of the exponential tilts given by the individual subsolutions. Motivated by this property of the mollifications we introduce the next class of changes of measure to provide an alternative implementation of importance sampling schemes based on mollifications of subsolutions which can be easier to implement for complex problems. The following is referred to as the randomized implementation of a mollification of subsolutions.

Construction 4.5 Let

$$
\begin{aligned}
\mathcal{MS}_0 \quad \doteq \quad & \Big\{ (U, \rho, \{W^k\}_{k=1}^K) : \rho \in \mathcal{W}(\mathbb{R}^d \times [0,T]),\ U \in \mathcal{S}_0,\ W^k \in \mathcal{S}_0 \textit{ for } k = 1, ..., K, \\
& \textit{and } (U, \rho, \{W^k\}_{k=1}^K) \textit{ satisfies (4.8) and (4.9)} \Big\}.
\end{aligned}
$$

For $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}_0$ define the corresponding importance sampling random variables according to

$$r^n \doteq e^{-\frac{1}{a(n)^2}F(\bar{Y}^{n,\gamma^n}(T))} \prod_{i=0}^{\lfloor Tn \rfloor} \left(\frac{d\gamma_i^n \left(\cdot | \, \bar{X}_i^{n,\gamma^n} \right)}{d\mu(\cdot | \, \bar{X}_i^{n,\gamma^n})} (\bar{u}_i^{\gamma^n}) \right)^{-1},$$

where the distribution of $\bar{u}_i^{\gamma^n}$, given $\bar{u}_j^{\gamma^n}, j = 0, \ldots, i-1$, is defined by

$$\gamma_i^n \left(du | \, \bar{X}_i^{n,\gamma^n} \right) = \sum_{k=1}^K \rho_k \left(\bar{Y}_i^{n,\gamma^n}, \frac{i}{n} \right) \exp \left\{ \left\langle u, -\frac{1}{a(n)\sqrt{n}} DW^k \left(\bar{Y}_i^{n,\gamma^n}, \frac{i}{n} \right) \right\rangle \right.$$
$$\left. -H_c \left(\bar{X}_i^{n,\gamma^n}, -\frac{1}{a(n)\sqrt{n}} DW^k \left(\bar{Y}_i^{n,\gamma^n}, \frac{i}{n} \right) \right) \right\} \mu \left(du | \, \bar{X}_i^{n,\gamma^n} \right),$$

and $\left(\bar{u}^{\gamma^n}, \bar{X}^{n,\gamma^n}, \bar{Y}^{n,\gamma^n} \right)$ are defined in terms of γ^n as in Construction 2.4.

Note that given any $W \in \mathcal{S}_0$ we have $(W, 1, W) \in \mathcal{MS}_0$, and the two implementations, Construction 4.1 for $W \in \mathcal{S}_0$ and Construction 4.5 for $(W, 1, W) \in \mathcal{MS}_0$, give the same importance sampling scheme. In addition, given any $\{W^k\}_{k=1}^K$ with $W^k \in \mathcal{S}_0$ for $k = 1, ..., K$ we have (recalling that
$\left(U^{\delta,\min}(\{W^k\}_{k=1}^K), \rho^{\delta,\min}(\{W^k\}_{k=1}^K), \{W^k\}_{k=1}^K \right)$ satisfies (4.8) and (4.9))
$\left(U^{\delta,\min}(\{W^k\}_{k=1}^K), \rho^{\delta,\min}(\{W^k\}_{k=1}^K), \{W^k\}_{k=1}^K \right) \in \mathcal{MS}_0$. Analogously
$\left(U^{\delta,\max}(\{W^k\}_{k=1}^K), \rho^{\delta,\max}(\{W^k\}_{k=1}^K), \{W^k\}_{k=1}^K \right) \in \mathcal{MS}_0$. Therefore the two implementations of importance sampling schemes based on $U^{\delta,\min}(\{W^k\}_{k=1}^K)$ are Construction 4.1 using $U^{\delta,\min}(\{W^k\}_{k=1}^K) \in \mathcal{S}_0$ and Construction 4.5 using $\left(U^{\delta,\min}(\{W^k\}_{k=1}^K), \rho^{\delta,\min}(\{W^k\}_{k=1}^K), \{W^k\}_{k=1}^K \right) \in \mathcal{MS}_0$. The same holds analogously for $U^{\delta,\max}(\{W^k\}_{k=1}^K)$. We now introduce notation for mollified subsolutions which satisfy the terminal condition.

Definition 4.6 Let

$$\mathcal{MS} \doteq \left\{ (U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}_0 : W(y,T) \le F(y) \textit{ for all } y \in \mathbb{R}^d \right\}.$$

The following theorem gives a lower bound on (4.6) based on the value of the subsolution at the origin.

Theorem 4.7 *For an arbitrary* $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ *let* $\{r^n\}_{n=1}^{\infty}$ *be the importance sampling random variables given by Construction 4.5. Then*

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

Proof. See Section 4.2 ■

This gives us a measure of the asymptotic performance of an importance sampling scheme given by the mollification of subsolutions $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ using the randomized implementation given in Construction 4.5. The following corollary gives the same bound for an importance sampling scheme based on a subsolution $W \in \mathcal{S}$ using the standard implementation given by Construction 4.1.

Corollary 4.8 *For an arbitrary* $W \in \mathcal{S}$ *let* $\{r^n\}_{n=1}^{\infty}$ *be the importance sampling random variables given by Construction 4.1. Then*

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

Proof. This follows easily from Theorem 4.7 and the fact that the importance sampling random variables given by $W \in \mathcal{S}$ under the standard implementation in Construction 4.1 and the importance sampling random variables given by $(W, 1, W) \in \mathcal{MS}$ under the randomized implemenation in Construction 4.5 are identical. ■

Note that given any $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ we also have $U \in \mathcal{S}$, and Constructions 4.5 and 4.1 give distinct (assuming $K > 1$) changes of measure whose estimators satisfy the same asymptotic lower bound on their second moments. We include the randomized implementation in Construction 4.5 because for complex problems it is more convenient to work with, so it is useful to know that the asymptotic bound the second moment holds for this implemenation as well. Because, as was stated earlier, it is a generalization of the standard implementation given in Construction 4.1 it suffices to investigate the randomized implementation.

4.2 Proof of Theorem 4.7

The proof of the asymptotic bound on the second moment of importance sampling estimators based on subsolutions in the moderate deviations setting differs from the proof of it in the large deviations setting found in [19] because of tightness complications. Similar to the proof of the moderate deviations principle, we can only get tightness of the occupation measure of the conditional means of the controlled noises, rather than tightness of the controlled noises themselves (which exists in the large deviations setting). In order to acquire this tightness we use the result of Theorem 2.5 which requires a bound on the relative entropy cost of the control measure with respect to the original measure. This bound is also required for the tightness proved in [19], however it can be attained more easily there. This is done by employing a change of measure to equate the expectation of the second moment of the estimator to an expectation under the measure of the original process. An extension of the variational formula (2.10) is then employed, and it can be shown that the relative entropies of near minimizing measures in the variational formulas are uniformly bounded. However, doing so uses the fact that the moment generating function of the noise $u_i(x)$ is finite everywhere, an assumption made in the large deviation setting but not in the moderate deviations setting. Without this assumption the approach used in [19] becomes signfiicantly more complicated, and to avoid this we employ a second change of measure to the second moment of the estimator. This gives us an expecation of a bounded function which is more convenient, but it is no longer under the original measure (see [16]). More work is then needed to get a bound on the relative entropy costs of the near minimizing measures in the variational formula with respect to the measures of the original processes in order to employ Theorem 2.5.

Recall that r^n and $\left(\bar{u}^{\gamma^n}, \bar{X}^{n,\gamma^n}, \bar{Y}^{n,\gamma^n}\right)$ are given by $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS} \subset \mathcal{MS}_0$ as in Construction 4.5. First of all if

$$\liminf_{n\to\infty} -a(n)^2 \log E[(r^n)^2] = \infty$$

then the result is true trivially, so we assume

$$\liminf_{n\to\infty} -a(n)^2 \log E[(r^n)^2] = L < \infty. \tag{4.11}$$

Since removing one likelihood ratio amounts to replacing the processes under γ^n by the original ones, it follows that

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

This is the first change of measure which equates the second moment with an expectation under the measure of the original process. In contrast to the approach taken in [19] (which involves a large deviations scaling as opposed to moderate deviations) we find it convenient to make a second change of measure to remove the exponential tilt term. Note that Jensen's inequality gives

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

This allows us to employ a second change of measure. Define $\left(\bar{u}^{\theta^n},\bar{X}^{n,\theta^n},\bar{Y}^{n,\theta^n}\right)$ in terms of θ^n as in Construction 2.4 where the conditional distribution of $\bar{u}_i^{\theta^n}$ given $\bar{u}_j^{\theta^n}$, $j<i$, is given by

$$
\begin{aligned}
\theta_i^n(du|\,\bar{X}_i^{n,\theta^n}) = \exp&\left\{\left\langle u,\frac{1}{a(n)\sqrt{n}}\alpha\left(\bar{Y}_i^{n,\theta^n},\frac{i}{n}\right)\right\rangle\right. \\
&\left. -H_c\left(\bar{X}_i^{n,\theta^n},\frac{1}{a(n)\sqrt{n}}\alpha\left(\bar{Y}_i^{n,\theta^n},\frac{i}{n}\right)\right)\right\}\mu(du|\,\bar{X}_i^{n,\theta^n})
\end{aligned}
\tag{4.12}
$$

and

$$
\alpha\left(y,t\right)=\sum_{k=1}^{K}\rho_k\left(y,t\right)DW^k\left(y,t\right). \tag{4.13}
$$

Note that this measure is well defined for n sufficiently large because $\left\|DW^k\right\|_\infty$ for $k=1,...,K$.

Then

$$E\left[e^{-\frac{1}{a(n)^2}2F(Y^n(T))}\prod_{i=0}^{\lfloor Tn\rfloor}\left(\sum_{k=1}^{K}\rho_k\left(Y_i^n,\frac{i}{n}\right)\exp\left\{\left\langle u_i(X_i^n),-\frac{1}{a(n)\sqrt{n}}DW^k\left(Y_i^n,\frac{i}{n}\right)\right\rangle\right.\right.\right.$$
$$\left.\left.\left.-H_c\left(X_i^n,-\frac{1}{a(n)\sqrt{n}}DW^k\left(Y_i^n,\frac{i}{n}\right)\right)\right\}\right)^{-1}\right]$$
$$\leq E\left[e^{-\frac{1}{a(n)^2}2F(Y^n(T))}\prod_{i=0}^{\lfloor Tn\rfloor}\exp\left\{\sum_{k=1}^{K}\rho_k\left(Y_i^n,\frac{i}{n}\right)\left(\left\langle u_i(X_i^n),\frac{1}{a(n)\sqrt{n}}DW^k\left(Y_i^{n,},\frac{i}{n}\right)\right\rangle\right.\right.\right.$$
$$\left.\left.\left.+H_c\left(X_i^n,-\frac{1}{a(n)\sqrt{n}}DW^k\left(Y_i^n,\frac{i}{n}\right)\right)\right)\right\}\right]$$
$$= E\left[e^{-\frac{1}{a(n)^2}2F(\bar{Y}^{n,\theta^n}(T))}\exp\left\{\sum_{i=0}^{\lfloor Tn\rfloor}\sum_{k=1}^{K}\rho_k\left(\bar{Y}_i^{n,\theta^n},\frac{i}{n}\right)H_c\left(\bar{X}_i^{n,\theta^n},-\frac{1}{a(n)\sqrt{n}}DW^k\left(\bar{Y}_i^{n,\theta^n},\frac{i}{n}\right)\right)\right.\right.$$
$$\left.\left.+H_c\left(\bar{X}_i^{n,\theta^n},\frac{1}{a(n)\sqrt{n}}\alpha\left(\bar{Y}_i^{n,\theta^n},\frac{i}{n}\right)\right)\right\}\right]. \tag{4.14}$$

The variational formula (2.10) can be extended in the following manner. For a proof see [12, Proposition 4.5.1].

Lemma 4.9 *Let $\mathcal{X}$ be a Polish space and let $\mathcal{P}(\mathcal{X})$ be the collection of probability measures on $\mathcal{X}$ with the Borel σ-algebra. Let g be a Borel measurable function that is bounded from below. Then for any $\mu \in \mathcal{P}(\mathcal{X})$*

$$-\log\int_{\mathcal{X}} e^{-g(x)}\mu(dx) = \inf_{\eta\in\mathcal{P}(\mathcal{X})}\left\{R(\eta\,\|\mu) + \int_{\mathcal{X}} g(x)\eta(dx)\right\}.$$

Using (4.14) and Lemma 4.9 we can write

$$-a(n)^2\log E[(r^n)^2] \tag{4.15}$$
$$\geq \inf_{\eta}\left\{a(n)^2R(\eta\,\|\theta^n) + E\left[2F(\bar{Y}^{n,\eta}(T))\right.\right.$$
$$-a(n)^2\sum_{i=0}^{\lfloor Tn\rfloor}\sum_{k=1}^{K}\rho_k\left(\bar{Y}_i^{n,\eta},\frac{i}{n}\right)H_c\left(\bar{X}_i^{n,\eta},-\frac{1}{a(n)\sqrt{n}}DW^k\left(\bar{Y}_i^{n,\eta},\frac{i}{n}\right)\right)$$
$$\left.\left.-a(n)^2\sum_{i=0}^{\lfloor Tn\rfloor}H_c\left(\bar{X}_i^{n,\eta},\frac{1}{a(n)\sqrt{n}}\alpha\left(\bar{Y}_i^{n,\eta},\frac{i}{n}\right)\right)\right]\right\}$$

where $\left(\bar{u}^{\eta^n}, \bar{X}^{n,\eta^n}, \bar{Y}^{n,\eta^n}\right)$ is defined in terms of of η^n as in Construction 2.4. Let

$$\left\|\{DW^k\}_{k=1}^K\right\|_\infty = \max\{\left\|DW^1\right\|_\infty, ..., \left\|DW^K\right\|_\infty\} \tag{4.16}$$

where $\|\cdot\|_\infty$ is given by (4.5). Given any $\varepsilon > 0$ let $\{\eta^n\}$ come within ε of achieving the infimum in (4.15). Then for all n

$$\begin{aligned}
&- a(n)^2 \log E[(r^n)^2] + \varepsilon \geq a(n)^2 R(\eta^n \,\|\theta^n\,) \\
&\quad + E\left[2F(\bar{Y}^{n,\eta^n}(T)) - a(n)^2 \sum_{i=0}^{\lfloor Tn \rfloor} H_c\left(\bar{X}_i^{n,\eta^n}, \frac{1}{a(n)\sqrt{n}}\alpha\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right)\right. \\
&\qquad \left. -a(n)^2 \sum_{i=0}^{\lfloor Tn \rfloor} \sum_{k=1}^{K} \rho_k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right) H_c\left(\bar{X}_i^{n,\eta^n}, -\frac{1}{a(n)\sqrt{n}} DW^k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right)\right],
\end{aligned}$$

and in particular due to (2.3) and (2.6) for all n sufficiently large that $\frac{1}{a(n)\sqrt{n}}\left\|\{W^k\}_{k=1}^K\right\|_\infty < \lambda_{DA}$ (recall (4.13))

$$\begin{aligned}
&- a(n)^2 \log E[(r^n)^2] + \varepsilon + \left(T\frac{K_A}{2} + T\lambda_{DA}K_{DA}\right)\left\|\{DW^k\}_{k=1}^K\right\|_\infty^2 - 2\inf_{y\in\mathbb{R}^d} F(y) \\
&\geq a(n)^2 R(\eta^n \,\|\theta^n\,).
\end{aligned}$$

Combining this with (4.11) implies we can choose a subsequence of $\{\eta^n, r^n\}$, retaining n as the index for convenience, such that

$$\limsup_{n\to\infty} a(n)^2 R(\eta^n \,\|\theta^n\,) < \infty$$

and

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

Note that the subsequence satisfies

$$
\begin{aligned}
&L+\varepsilon && (4.17)\\
&\geq \limsup_{n\to\infty}\Bigg(a(n)^2 R(\eta^n \,\|\theta^n\,) + E\left[2F(\bar{Y}^{n,\eta^n}(T))\right] \\
&- E\left[a(n)^2 \sum_{i=0}^{\lfloor Tn\rfloor} H_c\left(\bar{X}_i^{n,\eta^n}, \frac{1}{a(n)\sqrt{n}}\alpha\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right)\right] \\
&-E\left[a(n)^2 \sum_{i=0}^{\lfloor Tn\rfloor}\sum_{k=1}^{K} \rho_k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right) H_c\left(\bar{X}_i^{n,\eta^n}, -\frac{1}{a(n)\sqrt{n}} DW^k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right)\right]\Bigg).
\end{aligned}
$$

Note that we have to keep the linear combination of the $H_c(x, DW^k)$ with weights ρ_k instead of replacing it with $H_c(x,\alpha)$ because H_c is convex rather than concave. Using the chain rule for relative entropy [12, Theorem C.3.1] we can write

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

where η_i^n is the conditional distribution on $\bar{u}_i^{\eta^n}$ given $\bar{u}_j^{\eta^n}$ for $j < i$ under η^n as in Construction 2.4. The following lemma allows us to use the bound on $E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 nR\left(\eta_i^n \,\middle\|\, \theta_i^n(\cdot\,|\,\bar{X}_i^{n,\eta^n})\right)\right]$ to get a bound on $E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 nR\left(\eta_i^n \,\middle\|\, \mu(\cdot\,|\,\bar{X}_i^{n,\eta^n})\right)\right]$. This new bound will allow us to invoke Theorem 2.5.

Lemma 4.10 *Let $\{\eta^n\}$ be a sequence of measures with $\eta^n \in \mathcal{P}((R^d)^{\lceil Tn\rceil})$ and define the corresponding random variables based on these measures as in Construction 2.4. Given some $(U,\rho,\{W^k\}_{k=1}^K) \in \mathcal{MS}_0$ as in Construction 4.5 define the measures θ^n using (4.12). If*

$$
\limsup_{n\to\infty} E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 nR\left(\eta_i^n \,\middle\|\, \theta_i^n(\cdot\,|\,\bar{X}_i^{n,\eta^n})\right)\right] < \infty
$$

then

$$
\limsup_{n\to\infty} E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 nR\left(\eta_i^n \,\middle\|\, \mu(\cdot\,|\,\bar{X}_i^{n,\eta^n})\right)\right] < \infty. \tag{4.18}
$$

Proof. We have from (4.12)

$$\begin{aligned}
R\left(\eta_i^n \left\| \mu(\cdot | \bar{X}_i^{n,\eta^n})\right.\right) &= \int_{\mathbb{R}^d} \log\left(\frac{d\eta_i^n}{d\theta_i^n(\cdot | \bar{X}_i^{n,\eta^n})}(u)\frac{d\theta_i^n(\cdot | \bar{X}_i^{n,\eta^n})}{d\mu(\cdot | \bar{X}_i^{n,\eta^n})}(u)\right)\eta_i^n(du) \\
&= \int_{\mathbb{R}^d} \log\left(\frac{d\eta_i^n}{d\theta_i^n(\cdot | \bar{X}_i^{n,\eta^n})}(u)\right)\eta_i^n(du) \\
&\quad + \int_{\mathbb{R}^d} \log\left(\frac{d\theta_i^n(\cdot | \bar{X}_i^{n,\eta^n})}{d\mu(\cdot | \bar{X}_i^{n,\eta^n})}(u)\right)\eta_i^n(du) \\
&= R\left(\eta_i^n \left\| \theta_i^n(\cdot | \bar{X}_i^{n,\eta^n})\right.\right) + \int_{\mathbb{R}^d}\left\langle u, \frac{1}{a(n)\sqrt{n}}\alpha\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right\rangle \eta_i^n(du) \\
&\quad - H_c\left(\bar{X}_i^{n,\eta^n}, \frac{1}{a(n)\sqrt{n}}\alpha\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right) \\
&\leq R\left(\eta_i^n \left\| \theta_i^n(\cdot | \bar{X}_i^{n,\eta^n})\right.\right) + \frac{1}{a(n)\sqrt{n}}\left\|\{DW^k\}_{k=1}^K\right\|_\infty \left\|\int_{\mathbb{R}^d} u\eta_i^n(du)\right\|
\end{aligned}$$

because $H_c \geq 0$, where $\left\|\{DW^k\}_{k=1}^K\right\|_\infty$ is given by (4.16). Therefore

$$\begin{aligned}
&\limsup_{n\to\infty} E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R\left(\eta_i^n \| \mu(\cdot | \bar{X}_i^{n,\eta^n})\right)\right] \\
&\leq \limsup_{n\to\infty} E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R\left(\eta_i^n \left\| \theta_i^n(\cdot | \bar{X}_i^{n,\eta^n})\right.\right)\right] \\
&\quad + \left\|\{DW^k\}_{k=1}^K\right\|_\infty \limsup_{n\to\infty} E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)\sqrt{n}\left\|\int_{\mathbb{R}^d} u\eta_i^n(du)\right\|\right]
\end{aligned}$$

and the result (4.18) follows if we can show

$$\limsup_{n\to\infty} E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)\sqrt{n}\left\|\int_{\mathbb{R}^d} u\eta_i^n(du)\right\|\right] < \infty. \tag{4.19}$$

There exists $N_{\lambda_{DA}} < \infty$ such that $\frac{1}{a(n)\sqrt{n}}\left\|\{DW^k\}_{k=1}^K\right\|_\infty \leq \frac{\lambda_{DA}}{2}$ for all $n \geq N_{\lambda_{DA}}$, where λ_{DA} is given by (2.5). Then using (2.3) and (4.12) we have

$$\max_{0\leq i\leq \lfloor Tn\rfloor} \sup_{x\in\mathbb{R}^d} \sup_{\|\hat{\alpha}\|\leq \lambda_{DA}/2} \log\left(\int_{\mathbb{R}^d} e^{\langle\hat{\alpha},u\rangle}\theta_i^n(du | x)\right) \leq 2K_{\text{mgf}}$$

which combined with Lemma A.1 gives

$$R(\eta_i^n \| \theta_i^n(\cdot | \bar{X}_i^{n,\eta^n})) \geq \sup_{\xi \in \mathbb{R}^d} \left\{ \left\langle \xi, \int_{\mathbb{R}^d} u\eta_i^n(du) \right\rangle - \log \left(\int_{\mathbb{R}^d} e^{\langle \xi, u \rangle} \theta_i^n(du | \bar{X}_i^{n,\eta^n}) \right) \right\} \tag{4.20}$$
$$= \sup_{\xi \in \mathbb{R}^d} \left\{ \left\langle \xi, \int_{\mathbb{R}^d} u\eta_i^n(du) \right\rangle - H_c \left(\bar{X}_i^{n,\eta^n}, \xi - \frac{1}{a(n)\sqrt{n}} \alpha \left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n} \right) \right) \right\}$$

Let e_i denote the ith unit vector and note that if $n \geq N_{\lambda_{DA}}$ then for all $x, y, \beta \in \mathbb{R}^d$ and $t \in [0, T]$

$$\begin{aligned}
&a(n)^2 n \sup_{\xi \in \mathbb{R}^d} \left\{ \langle \xi, \beta \rangle - H_c \left(x, \xi - \frac{1}{a(n)\sqrt{n}} \alpha(y, t) \right) \right\} \\
&\geq a(n)^2 n \left\{ \left\langle \pm \frac{\left\|\{DW^k\}_{k=1}^K\right\|_\infty}{a(n)\sqrt{n}} e_i, \beta \right\rangle - H_c \left(x, \pm \frac{\left\|\{DW^k\}_{k=1}^K\right\|_\infty}{a(n)\sqrt{n}} e_i - \frac{1}{a(n)\sqrt{n}} \alpha(y, t) \right) \right\} \\
&= \pm \left\| \{DW^k\}_{k=1}^K \right\|_\infty a(n)\sqrt{n}\beta_i - a(n)^2 n H_c \left(x, \pm \frac{\left\|\{DW^k\}_{k=1}^K\right\|_\infty}{a(n)\sqrt{n}} e_i - \frac{1}{a(n)\sqrt{n}} \alpha(y, t) \right) \\
&\geq \pm \left\| \{DW^k\}_{k=1}^K \right\|_\infty a(n)\sqrt{n}\beta_i - 2K_A \left\| \{DW^k\}_{k=1}^K \right\|_\infty^2 - 4\lambda_{DA} K_{DA} \left\| \{DW^k\}_{k=1}^K \right\|_\infty^2,
\end{aligned}$$

where for the last inequality we used (2.3) and (2.6). Therefore

$$\begin{aligned}
&a(n)^2 n \frac{d}{\left\|\{DW^k\}_{k=1}^K\right\|_\infty} \sup_{\xi \in \mathbb{R}^d} \left\{ \langle \xi, \beta \rangle - H_c \left(x, \xi - \frac{1}{a(n)\sqrt{n}} \alpha(y, t) \right) \right\} \\
&\quad + 2K_A \left\| \{DW^k\}_{k=1}^K \right\|_\infty d + 4\lambda_{DA} K_{DA} \left\| \{DW^k\}_{k=1}^K \right\|_\infty d \geq a(n)\sqrt{n} \|\beta\|
\end{aligned}$$

and combining this with (4.20) gives gives

$$\begin{aligned}
&\frac{d}{\left\|\{DW^k\}_{k=1}^K\right\|_\infty} E \left[\frac{1}{n} \sum_{i=0}^{\lfloor Tn \rfloor} a(n)^2 n R \left(\eta_i^n \left\| \theta_i^n(\cdot | \bar{X}_i^{n,\eta^n}) \right. \right) \right] \\
&+ 2K_A \left\| \{DW^k\}_{k=1}^K \right\|_\infty d + 4\lambda_{DA} K_{DA} \left\| \{DW^k\}_{k=1}^K \right\|_\infty d \\
&\geq E \left[\frac{1}{n} \sum_{i=0}^{\lfloor Tn \rfloor} a(n)\sqrt{n} \left\| \int_{\mathbb{R}^d} u\eta_i^n(du) \right\| \right]
\end{aligned}$$

whenever $n \geq N_{\lambda_{DA}}$. Hence (4.19) and also (4.18) are valid, which concludes the proof. ■

Recall the definitions of $w^{n,\eta^n}(t)$, $\hat{w}^{n,\eta^n}(t)$, and ς^{n,η^n} given by Construction 2.4 in terms of the measures η^n. Lemma 4.10 allows us to apply Theorem 2.5 and choose a (further) subsequence of

$\{(\varsigma^{n,\eta^n}, \bar{Y}^{n,\eta^n})\}$ (we retain n as the index for convenience) along which $\{(\varsigma^{n,\eta^n}, \bar{Y}^{n,\eta^n})\}$ converges weakly to some limit $(\varsigma, \hat{Y})$ in $\mathcal{P}(\mathbb{R}^d \otimes [0,T]) \otimes C([0,T] : \mathbb{R}^d)$, where $\varsigma_2(dt)$ is Lebesgue measure with probability 1,

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

with

$$\hat{w}(t) = \int_{\mathbb{R}^d} w\varsigma_{1|2}(dw|\, t),$$

and the $\{\varsigma^{n,\eta^n}\}$ are uniformly integrable in the sense that

$$\lim_{C\to\infty} \limsup_{n\to\infty} E\left[\int_{\{w:\|w\|\geq C\}} \|w\|\, \varsigma^{n,\eta^n}(dw \otimes dt)\right] = 0.$$

Because (4.12) gives

$$\begin{aligned}
&a(n)^2 nR\left(\eta_i^n \left\| \theta_i^n(\cdot|\,\bar{X}_i^{n,\eta^n})\right.\right) \\
&\quad = a(n)^2 nR\left(\eta_i^n \left\| \mu(\cdot|\,\bar{X}_i^{n,\eta^n})\right.\right) - \left\langle a(n)\sqrt{n}\int_{\mathbb{R}^d} u\eta_i^n(du), \alpha\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right\rangle \\
&\qquad + a(n)^2 nH_c\left(\bar{X}_i^{n,\eta^n}, \frac{1}{a(n)\sqrt{n}}\alpha\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right)
\end{aligned}$$

we have

$$\begin{aligned}
&E\left[2F(\bar{Y}^{n,\eta^n}(T)) + \frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor}\left(a(n)^2 nR\left(\eta_i^n \left\| \theta_i^n(\cdot|\,\bar{X}_i^{n,\eta^n})\right.\right)\right.\right. \\
&\quad - a(n)^2 nH_c\left(\bar{X}_i^{n,\eta^n}, \frac{1}{a(n)\sqrt{n}}\alpha\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right) \\
&\quad \left.\left. -a(n)^2 n\sum_{k=1}^K \rho_k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right) H_c\left(\bar{X}_i^{n,\eta^n}, -\frac{1}{a(n)\sqrt{n}}DW^k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right)\right)\right] \\
&= E\left[F(\bar{Y}^{n,\eta^n}(T)) + \frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 nR\left(\eta_i^n \left\| \mu(\cdot|\,\bar{X}_i^{n,\eta^n})\right.\right)\right] + E[F(\bar{Y}^{n,\eta^n}(T))] \\
&\quad - E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor}\left\langle a(n)\sqrt{n}\int u\eta_i^n(du), \alpha\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right\rangle\right] \\
&\quad - E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n\sum_{k=1}^K \rho_k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right) H_c\left(\bar{X}_i^{n,\eta^n}, -\frac{1}{a(n)\sqrt{n}}DW^k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right)\right].
\end{aligned} \tag{4.21}$$

Note that Lemma 4.9 combined with the chain rule for relative entropy [12, Theorem C.3.1] gives

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

which combined with (4.1) and the fact that r^n is unbiased implies

$$\begin{aligned}
&\liminf_{n\to\infty} E\left[F(\bar{Y}^{n,\eta^n}(T)) + \frac{1}{n} \sum_{i=0}^{\lfloor Tn \rfloor} a(n)^2 n R\left(\eta_i^n \left\| \mu(\cdot | \bar{X}_i^{n,\eta^n}) \right.\right) \right] \\
&\quad \geq \liminf_{n\to\infty} -a(n)^2 \log E[r^n] \\
&\quad = \liminf_{n\to\infty} -a(n)^2 \log E e^{-\frac{1}{a(n)^2} F(Y^n(T))} \\
&\quad \geq V(0,0).
\end{aligned}$$

where the last inequality comes from (4.2). Because $\bar{Y}^{n,\eta^n} \to \hat{Y}$ weakly and F is bounded from below and lower-semicontinuous we can apply a weak convergence version of Fatou's lemma to get

$$\liminf_{n\to\infty} E[F(\bar{Y}^{n,\eta^n}(T))] \geq E[F(\hat{Y}(T))].$$

In addition

$$\begin{aligned}
&E\left[\frac{1}{n} \sum_{i=0}^{\lfloor Tn \rfloor} \left\langle a(n)\sqrt{n} \int_{\mathbb{R}^d} y \eta_i^n(dy), \alpha\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n} \right) \right\rangle \right] \\
&\quad = E\left[\int_{[0,T]\otimes\mathbb{R}^d} \left\langle w, \alpha\left(\bar{Y}^{n,\eta^n}\left(\frac{\lfloor nt \rfloor}{n} \right), \frac{\lfloor nt \rfloor}{n} \right) \right\rangle \varsigma^{n,\eta^n}(dw \otimes dt) \right]
\end{aligned}$$

and using (2.12), the continuity and boundedness of α, and the fact that $(\varsigma^{n,\eta^n}, \bar{Y}^{n,\eta^n}) \to (\varsigma, \hat{Y})$ weakly it follows that

$$\begin{aligned}
&\lim_{n\to\infty} E\left[\int_{[0,T]\otimes\mathbb{R}^d} \left\langle w, \alpha\left(\bar{Y}^{n,\eta^n}\left(\frac{\lfloor nt \rfloor}{n} \right), \frac{\lfloor nt \rfloor}{n} \right) \right\rangle \varsigma^{n,\eta^n}(dw \otimes dt) \right] \\
&\quad = E\left[\int_{[0,T]\otimes\mathbb{R}^d} \left\langle w, \alpha\left(\hat{Y}(t), t \right) \right\rangle \varsigma(dw \otimes dt) \right] \\
&\quad = E\left[\int_0^T \left\langle \hat{w}(t), \alpha\left(\hat{Y}(t), t \right) \right\rangle dt \right].
\end{aligned}$$

Finally (2.6), the continuity and boundedness of DW^k, and the weak convergence of $\bar{Y}^{n,\eta^n}$ to $\hat{Y}$ imply

$$\begin{aligned}
&\lim_{n\to\to\infty} E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n \sum_{k=1}^{K} \rho_k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right) H_c\left(\bar{X}_i^{n,\eta^n}, -\frac{1}{a(n)\sqrt{n}} DW^k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right)\right] \\
&\quad = E\left[\sum_{k=1}^{K}\int_0^T \rho_k\left(\hat{Y}(t), t\right) \frac{1}{2}\left\| DW^k\left(\hat{Y}(t), t\right)\right\|^2_{A(X^0(t))} dt\right].
\end{aligned}$$

Consequently along this subsequence (recall (4.13))

$$\begin{aligned}
&\liminf_{n\to\infty}\left(E\left[F(\bar{Y}^{n,\eta^n}(T)) + \frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n R\left(\eta_i^n \left\| \mu(\cdot \,|\, \bar{X}_i^{n,\eta^n})\right.\right)\right] + E[F(\bar{Y}^{n,\eta^n}(T))]\right. \\
&\qquad - E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor}\left\langle a(n)\sqrt{n}\int y\eta_i^n(dy), \alpha\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right\rangle\right] \\
&\qquad \left. - E\left[\frac{1}{n}\sum_{i=0}^{\lfloor Tn\rfloor} a(n)^2 n\sum_{k=1}^{K}\rho_k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right) H_c\left(\bar{X}_i^{n,\eta^n}, -\frac{1}{a(n)\sqrt{n}} DW^k\left(\bar{Y}_i^{n,\eta^n}, \frac{i}{n}\right)\right)\right]\right) \\
&\quad \geq V(0,0) + E\left[F(\hat{Y}(T)) - \sum_{k=1}^{K}\int_0^T \rho_k\left(\hat{Y}(t), t\right)\left(\left\langle \hat{w}(t), DW^k\left(\hat{Y}(t), t\right)\right\rangle\right.\right. \\
&\qquad\qquad \left.\left. + \frac{1}{2}\left\| DW^k\left(\hat{Y}(t), t\right)\right\|^2_{A(X^0(t))}\right) dt\right]
\end{aligned}$$

which combined with (4.17) and (4.21) gives (recall (4.11))

$$\begin{aligned}
&\liminf_{n\to\infty} -a(n)^2 \log E[(r^n)^2] + \varepsilon \\
&\quad \geq V(0,0) + E\left[F(\hat{Y}(T)) - \sum_{k=1}^{K}\int_0^T \rho_k\left(\hat{Y}(t), t\right)\left(\left\langle \hat{w}(t), DW^k\left(\hat{Y}(t), t\right)\right\rangle\right.\right. \\
&\qquad\qquad \left.\left. + \frac{1}{2}\left\| DW^k\left(\hat{Y}(t), t\right)\right\|^2_{A(X^0(t))}\right) dt\right].
\end{aligned}$$

Finally, (2.13) and $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ with $\mathcal{MS}$ given by Definition 4.6 (recall this means the

W^k satisfy (4.4)) imply

$$
\begin{aligned}
& E\left[F(\hat{Y}(T)) - \sum_{k=1}^{K} \int_0^T \rho_k\left(\hat{Y}(t), t\right)\left(\left\langle \hat{w}(t), DW^k\left(\hat{Y}(t), t\right)\right\rangle + \frac{1}{2}\left\|DW^k\left(\hat{Y}(t), t\right)\right\|_{A(X^0(t))}^2\right) dt\right] \\
&\quad \geq E\left[F(\hat{Y}(T)) - \int_0^T \left\langle \hat{w}(t), \sum_{k=1}^{K} \rho_k\left(\hat{Y}(t), t\right) DW^k\left(\hat{Y}(t), t\right)\right\rangle dt\right. \\
&\qquad - \int_0^T \left\langle \sum_{k=1}^{K} \rho_k\left(\hat{Y}(t), t\right) DW^k\left(\hat{Y}(t), t\right), Db(X^0(t))\hat{Y}(t)\right\rangle dt \\
&\qquad \left. - \int_0^T \sum_{k=1}^{K} \rho_k\left(\hat{Y}(t), t\right) W_t^k\left(\hat{Y}(t), t\right) dt\right] \\
&\quad \geq E\left[F(\hat{Y}(T)) - \int_0^T \frac{d}{dt} U\left(\hat{Y}(t), t\right) dt\right] \\
&\quad \geq U(0,0).
\end{aligned}
$$

Therefore

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

and because $\varepsilon > 0$ was arbitrary this completes the proof of Theorem 4.7.

5 INVARIANCE OF EMBEDDING

Recall that we want to estimate

$$Ee^{-\hat{F}(Z(T))}$$

where $\hat{F}$ is an arbitrary lower semi-continuous function bounded from below, Z_i is given by (2.1), and $Z(t)$ is the continuous time linear interpolation of Z_i given by $Z(\epsilon i) = Z_i$ for $i = 0, ..., T/\epsilon$ and

$$Z(t) = ((i+1)/\epsilon - t)Z_i + (t - i/\epsilon)Z_{i+1} \tag{5.1}$$

for $t \in [i/\epsilon, (i+1)/\epsilon]$. In order to use the results of Theorem 4.7 to suggest a change of measure on the original process Z_i we must embed it in a sequence of processes $\{X^n\}$ as in (2.2) which satisfy Condition 2.1. In order to do this we assume that $\hat{b}(z)$ and $\hat{\mu}(\cdot\,|\, z)$ satisfy Condition 2.1. We consider the following class of embeddings.

Construction 5.1 *Let $\{X^n\}$ given by (2.2) with $x_0 = z_0$ be such that there exists $n_\epsilon \in \mathbb{N}$ satisfying*

$$\frac{1}{n_\epsilon} b(x) = \epsilon \hat{b}(x) \tag{5.2}$$

and

$$\frac{1}{n_\epsilon} u_i(\cdot) = \epsilon v_i(\cdot) \tag{5.3}$$

in distribution. Then for some arbitrary $a(n)$ satisfying (2.8) let

$$F(y) = a(n_\epsilon)^2 \hat{F}(X^{n_\epsilon,0}(T_\epsilon) + y/a(n_\epsilon)\sqrt{n_\epsilon})$$

where F is used in (4.7) with

$$T_\epsilon = T/(\epsilon n_\epsilon).$$

Then

$$Z = X^{n_\epsilon} = X^0 + \frac{1}{a(n_\epsilon)\sqrt{n_\epsilon}} Y^{n_\epsilon}$$

and the quantity of interest is

$$Ee^{-\hat{F}(Z(T))} = Ee^{-\frac{1}{a(n_\epsilon)^2} F(Y^{n_\epsilon}(T_\epsilon))}.$$

Note that due to (5.2) and (5.3) $b(x)$ and $\mu(\cdot|\, x)$ satisfy Condition 2.1 because $\hat{b}(z)$ and $\hat{\mu}(\cdot|\, z)$ do, so the results of Theorem 4.7 hold. For some $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ the change of measure given by Construction 4.5 is given by the conditional distributions

$$\begin{aligned} \gamma_i^{n_\epsilon}\left(du|\, \bar{X}_i^{n_\epsilon,\gamma^{n_\epsilon}}\right) = &\sum_{k=1}^K \rho_k \left(\bar{Y}_i^{n_\epsilon,\gamma^{n_\epsilon}}, \frac{i}{n_\epsilon}\right) \exp\left\{\left\langle u, -\frac{1}{a(n_\epsilon)\sqrt{n_\epsilon}} DW^k\left(\bar{Y}_i^{n_\epsilon,\gamma^{n_\epsilon}}, \frac{i}{n_\epsilon}\right)\right\rangle\right. \\ &\left. - H_c\left(\bar{X}_i^{n_\epsilon,\gamma^{n_\epsilon}}, -\frac{1}{a(n_\epsilon)\sqrt{n_\epsilon}} DW^k\left(\bar{Y}_i^{n_\epsilon,\gamma^{n_\epsilon}}, \frac{i}{n_\epsilon}\right)\right)\right\} \mu\left(du|\, \bar{X}_i^{n_\epsilon,\gamma^{n_\epsilon}}\right) \end{aligned}$$

as in Construction 2.4 and the importance sampling estimate is

$$r^{n_\epsilon} \doteq e^{-\frac{1}{a(n_\epsilon)^2} F(\bar{Y}^{n_\epsilon,\gamma^{n_\epsilon}}(T_\epsilon))} \prod_{i=0}^{\lfloor T_\epsilon n_\epsilon \rfloor} \left(\frac{d\gamma_i^{n_\epsilon}\left(\cdot|\, \bar{X}_i^{n_\epsilon,\gamma^{n_\epsilon}}\right)}{\mu(\cdot|\, \bar{X}_i^{n_\epsilon,\gamma^{n_\epsilon}})}(\bar{u}_i^{\gamma^{n_\epsilon}})\right)^{-1}.$$

Our goal in this section is to show that the choice of embedding does not impact the changes of measure suggested by this subsolution approach to importance sampling. Recall in (2.9) that by working with Y^n instead of X^n we are amplifiying the noise by $a(n)\sqrt{n}$, and the limit results in a first order approximation of the drift. In order to account for this a $1/a(n)\sqrt{n}$ term is present in the exponential tilt, essentially reducing the conditional mean of the controlled noise at the same rate that the noise is being amplified. Because of this effective cancellation the particular embedding doesn't make a difference, we always end up with a first order approximation of the dynamics and a quadratic approximation of the cost in the control problem (4.2) determining our PDE. We now work with the original process Z_i in order to demonstrate this.

Define $\hat{H}_c(z,\alpha)$ and $\hat{A}(z)$ in terms of $\hat{\mu}(\cdot|\, z)$ analogously to how $H_c(x,\alpha)$ and $A(x)$ were defined in terms of $\mu(\cdot|\, x)$ in the Introduction. We introduce the following processes to get a class of subsolution changes of measure analogous to $\mathcal{MS}$ but in terms of our original process Z. Let

$$Z_{i+1}^{0,\epsilon} = Z_i^{0,\epsilon} + \epsilon\hat{b}(Z_i^{0,\epsilon}), Z_0^{0,\epsilon} = z_0$$

and let $Z^{0,\epsilon}(t)$ be the continuous time linear interpolation constructed in the same manner as (5.1). Let $Z^0(t)$ be given by the ODE

$$Z_t^0(t) = \hat{b}(Z^0(t)), Z^0(0) = z_0$$

and consider the partial differential inequality

$$W_t(y,t) \geq \frac{1}{2}\|DW(y,t)\|^2_{\hat{A}(Z^0(t))} - \left\langle DW(y,t), D\hat{b}(Z^0(t))y\right\rangle. \tag{5.4}$$

Note that $Z_i^{0,\epsilon} = X_i^{0,n_\epsilon}$ and $Z^0(t) = X^0(t/\epsilon n_\epsilon)$. We now define subsolutions in terms of Z.

Construction 5.2 *Recall the definitions of $\mathcal{W}(\mathbb{R}^d\times[0,T])$ in (4.10) and $\|\cdot\|_\infty$ in (4.5). Let*

$$\begin{aligned}\mathcal{MS}^Z \doteq\ & \Big\{(U,\rho,\{W^k\}_{k=1}^K) : \rho\in\mathcal{W}(\mathbb{R}^d\times[0,T]), W^k\in C^1(\mathbb{R}^d\times[0,T]:\mathbb{R}),\\ & \left\|DW^k\right\|_\infty < \infty,\ \text{and } W^k \text{ satisfies (5.4) for } k=1,...,K,\\ & (U,\rho,\{W^k\}_{k=1}^K) \text{ satisfies (4.8) and (4.9)}\\ & \text{and } U(y,T)\leq\epsilon\hat{F}(Z^{0,\epsilon}(T)+y) \text{ for all } y\in\mathbb{R}^d\Big\}.\end{aligned}$$

For $(U,\rho,\{W^k\}_{k=1}^K)\in\mathcal{MS}^Z$ define the corresponding importance sampling estimate

$$r_Z \doteq e^{\hat{F}(\bar{Z}^{\gamma^Z}(T))}\prod_{i=0}^{\lfloor T/\epsilon\rfloor}\left(\frac{d\gamma_i^Z\left(\cdot|\,\bar{Z}_i^{\gamma^Z}\right)}{d\mu(\cdot|\,\bar{Z}_i^{\gamma^Z})}(\bar{v}_i^{\gamma^Z})\right)^{-1}$$

where

$$\bar{Z}^{\gamma^Z}_{i+1} = \bar{Z}^{\gamma^Z}_i + \epsilon\hat{b}(\bar{Z}^{\gamma^Z}_i) + \epsilon\bar{v}^{\gamma^Z}_i, \bar{Z}^{\gamma^Z}_0 = z_0$$

and the conditional distribution of $\bar{v}^{\gamma^Z}_i$, given $\bar{v}^{\gamma^Z}_j$ for $j < i$, is

$$\begin{aligned}\gamma^Z_i\left(dv|\,\bar{Z}^{\gamma^Z}_i\right) &= \sum_{k=1}^{K}\rho_k\left(\bar{Z}^{\gamma^Z}_i - Z^{0,\epsilon}_i, \epsilon i\right)\exp\left\{\left\langle v, -DW^k\left(\bar{Z}^{\gamma^Z}_i - Z^{0,\epsilon}_i, \epsilon i\right)\right\rangle\right.\\ &\quad \left.-\hat{H}_c\left(\bar{Z}^{\gamma^Z}_i, -DW^k\left(\bar{Z}^{\gamma^Z}_i - Z^{0,\epsilon}_i, \epsilon i\right)\right)\right\}\hat{\mu}\left(dv|\,\bar{Z}^{\gamma^Z}_i\right)\end{aligned}$$

We now introduce definitions which will help us show that the choice of embedding does not matter. The following will be called the $\mathcal{MS}^Z$ version of a subsolution in $\mathcal{MS}$.

Definition 5.3 *For an arbitrary embedding in Construction 5.1 and given any $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ define*

$$U^Z(y,t) = \frac{\epsilon}{a(n_\epsilon)^2}U(a(n_\epsilon)\sqrt{n_\epsilon}y, t/\epsilon n_\epsilon),$$

$$\rho^Z(y,t) = \rho(a(n_\epsilon)\sqrt{n_\epsilon}y, t/\epsilon n_\epsilon),$$

and

$$W^{Z,k}(y,t) = \frac{\epsilon}{a(n_\epsilon)^2}W^k(a(n_\epsilon)\sqrt{n_\epsilon}y, t/\epsilon n_\epsilon).$$

The following will be called the $\mathcal{MS}$ version of a subsolution in $\mathcal{MS}^Z$.

Definition 5.4 *For an arbitrary embedding in Construction 5.1 given any $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}^Z$ define*

$$U^X(y,t) = \frac{a(n_\epsilon)^2}{\epsilon}U(y/a(n_\epsilon)\sqrt{n_\epsilon}, \epsilon n_\epsilon t),$$

$$\rho^X(y,t) = \rho(y/a(n_\epsilon)\sqrt{n_\epsilon}, \epsilon n_\epsilon t),$$

and

$$W^{X,k}(y,t) = \frac{a(n_\epsilon)^2}{\epsilon}W^k(y/a(n_\epsilon)\sqrt{n_\epsilon}, \epsilon n_\epsilon t).$$

The following theorem demonstrates that all embeddings given in Construction 5.1 suggest the same importance sampling schemes.

Theorem 5.5 *Consider an arbitrary embedding given by Construction 5.1. Given any* $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ *define* $(U^Z, \rho^Z, \{W^{Z,k}\}_{k=1}^K)$ *according to Definition 5.3. Then* $(U^Z, \rho^Z, \{W^{Z,k}\}_{k=1}^K) \in \mathcal{MS}^Z$ *and the importance sampling scheme given by* $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ *in Construction 4.5 is identical to the importance sampling scheme given by* $(U^Z, \rho^Z, \{W^{Z,k}\}_{k=1}^K) \in \mathcal{MS}^Z$ *in Construction 5.2. Similarly, given any* $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}^Z$ *define* $(U^X, \rho^X, \{W^{X,k}\}_{k=1}^K)$ *according to Definition 5.4. Then* $(U^X, \rho^X, \{W^{X,k}\}_{k=1}^K) \in \mathcal{MS}$ *and the importance sampling scheme given by* $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}^Z$ *in Construction 5.2 is identical to the importance sampling scheme given by* $(U^X, \rho^X, \{W^{X,k}\}_{k=1}^K) \in \mathcal{MS}$ *in Construction 4.5.*

Proof. First let $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ be arbitrary and let $(U^Z, \rho^Z, \{W^{Z,k}\}_{k=1}^K)$ be given by Definition 5.3. Then

$$\begin{aligned}
DU^Z(y,t) &= D(\frac{\epsilon}{a(n_\epsilon)^2} U(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon)) \\
&= \frac{\epsilon\sqrt{n_\epsilon}}{a(n_\epsilon)} DU(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon) \\
&= \sum_{k=1}^K \rho_k(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon) \frac{\epsilon\sqrt{n_\epsilon}}{a(n_\epsilon)} DW^k(a(n_\epsilon)\sqrt{\epsilon n_\epsilon} y, t/\epsilon n_\epsilon) \\
&= \sum_{k=1}^K \rho_k^Z(y,t) DW^{Z,k}(y,t)
\end{aligned}$$

and

$$\begin{aligned}
U_t^Z(y,t) &= \frac{d}{dt}(\frac{\epsilon}{a(n_\epsilon)^2} U(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon)) \\
&= \frac{1}{a(n_\epsilon)^2 n_\epsilon} U_t(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon) \\
&= \sum_{k=1}^K \rho_k(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon) \frac{1}{a(n_\epsilon)^2 n_\epsilon} W_t^k(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon) \\
&= \sum_{k=1}^K \rho_k^Z y, t) W^{Z,k}(y,t)
\end{aligned}$$

so $(U^Z, \rho^Z, \{W^{Z,k}\}_{k=1}^K)$ satisfies (4.8) and (4.9). In addition,

$$\begin{aligned} U^Z(y,T) &= \frac{\epsilon}{a(n_\epsilon)^2} U(a(n_\epsilon)\sqrt{n_\epsilon} y, T/\epsilon n_\epsilon) \\ &= \frac{\epsilon}{a(n_\epsilon)^2} U(a(n_\epsilon)\sqrt{n_\epsilon} y, T_\epsilon) \\ &\leq \epsilon \hat{F}(X^{n_\epsilon,0}(T_\epsilon) + y) \\ &= \epsilon \hat{F}(Z^{0,\epsilon}(T) + y) \end{aligned}$$

for all $y \in \mathbb{R}^d$. Clearly $\rho^Z \in \mathcal{W}(\mathbb{R}^d \times [0,T])$ because $\rho \in \mathcal{W}(\mathbb{R}^d \times [0,T_\epsilon])$ and $W^{Z,k} \in C^1(\mathbb{R}^d \times [0,T] : \mathbb{R})$ and $\|DW^{Z,k}\|_\infty < \infty$ because $W^k \in \mathcal{S}_0$ for $k = 1, ..., K$. In order to prove that $(U^Z, \rho^Z, \{W^{Z,k}\}_{k=1}^K) \in \mathcal{MS}^Z$ it remains to show that the $\{W^{Z,k}\}_{k=1}^K$ satisfy (5.4). Note that Construction 5.1 gives

$$n_\epsilon^2 \epsilon^2 \hat{A}(x) = A(x)$$

and

$$n_\epsilon \epsilon D\hat{b}(x) = Db(x)$$

so

$$W_t^{Z,k}(y,t) = \frac{1}{a(n_\epsilon)^2 n_\epsilon} W_t^k(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon)),$$

$$\frac{1}{2}\left\| DW^{Z,k}(y,t) \right\|^2_{\hat{A}(Z^0(t))} = \frac{1}{a(n_\epsilon)^2 n_\epsilon 2}\left\| DW^k(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon)\right\|^2_{A(X^0(t/\epsilon n_\epsilon))},$$

and

$$\left\langle DW^{Z,k}(y,t), D\hat{b}(Z^0(t))y \right\rangle = \frac{1}{a(n_\epsilon)^2 n_\epsilon}\left\langle DW^k(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon), Db(X^0(t/\epsilon n_\epsilon))a(n_\epsilon)\sqrt{n_\epsilon} y \right\rangle.$$

Then because $W^k \in \mathcal{S}_0$ we have

$$\begin{aligned} &\frac{1}{a(n_\epsilon)^2 n_\epsilon} W_t^k(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon) \\ &\quad \geq \frac{1}{a(n_\epsilon)^2 n_\epsilon 2}\left\| DW^k(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon)\right\|^2_{A(X^0(t/\epsilon n_\epsilon))} \\ &\qquad - \frac{1}{a(n_\epsilon)^2 n_\epsilon}\left\langle DW^k(a(n_\epsilon)\sqrt{n_\epsilon} y, t/\epsilon n_\epsilon), Db(X^0(t/\epsilon n_\epsilon))a(n_\epsilon)\sqrt{n_\epsilon} y \right\rangle \end{aligned}$$

or equivalently

$$W_t^{Z,k}(y,t) \geq \frac{1}{2}\left\|DW^{Z,k}(y,t)\right\|^2_{\hat{A}(Z^0(t))} - \left\langle DW^{Z,k}(y,t), D\hat{b}(Z^0(t))y\right\rangle.$$

This completes the proof that $(U^Z, \rho^Z, \{W^{Z,k}\}_{k=1}^K) \in \mathcal{MS}^Z$. To show that the scheme given by $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ in Construction 4.5 is identical to the scheme given by $(U^Z, \rho^Z, \{W^{Z,k}\}_{k=1}^K) \in \mathcal{MS}^Z$ in Construction 5.2 recall that for all $x \in \mathbb{R}^d$ $\epsilon v_i(x) = \frac{1}{n_\epsilon}u_i(x)$ in distribution where the distribution of $v_i(x)$ is given by $\hat{\mu}(du|\,x)$ and the distribution of $u_i(x)$ is given by $\mu(du|\,x)$. Then given any $\{\alpha_k\}_{k=1}^K, \alpha_k \in \mathbb{R}^d$ and $\{\rho_k\}_{k=1}^K, \rho_k \in [0,1], \sum_{k=1}^K \rho_k = 1$ we have $\epsilon\bar{v}_i(x) = \frac{1}{n_\epsilon}\bar{u}_i(x)$ in distribution where $\bar{v}_i(x)$ has distribution

$$\sum_{k=1}^K \rho_k \exp\{\langle \epsilon n_\epsilon \alpha_k, v\rangle - \hat{H}_c(x,\alpha_k)\}\hat{\mu}(dv|\,x).$$

and $\bar{u}_i(x)$ has distribution

$$\sum_{k=1}^K \rho_k \exp\left\{\left\langle \frac{1}{\epsilon n_\epsilon}\alpha_k, u\right\rangle - H_c(x, \frac{1}{\epsilon n_\epsilon}\alpha_k)\right\}\mu(du|\,x).$$

However, the conditional distribution of $\bar{u}_i^{\gamma^{n_\epsilon}}$ under $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ in Construction 4.5 given $\bar{X}_i^{n_\epsilon,\gamma^{n_\epsilon}} = x$ is

$$\begin{aligned}
&\gamma_i^{n_\epsilon}(du|\,x) \\
&= \sum_{k=1}^K \rho_k\left(a(n_\epsilon)\sqrt{n_\epsilon}(x - X_i^{0,n_\epsilon}), \frac{i}{n_\epsilon}\right)\exp\left\{\left\langle u, -\frac{1}{a(n_\epsilon)\sqrt{n_\epsilon}}DW^k\left(a(n_\epsilon)\sqrt{n_\epsilon}(x - X_i^{0,n_\epsilon}), \frac{i}{n_\epsilon}\right)\right\rangle\right. \\
&\quad \left. -H_c\left(x, -\frac{1}{a(n_\epsilon)\sqrt{n_\epsilon}}DW^k\left(a(n_\epsilon)\sqrt{n_\epsilon}(x - X_i^{0,n_\epsilon}), \frac{i}{n_\epsilon}\right)\right)\right\}\mu(du|\,x) \\
&= \sum_{k=1}^K \rho_k^Z\left(x - Z_i^{0,\epsilon}, \epsilon i\right)\exp\left\{\left\langle u, -\frac{1}{\epsilon n_\epsilon}DW^{Z,k}\left(x - Z_i^{0,\epsilon}, \epsilon i\right)\right\rangle\right. \\
&\quad \left. -H_c\left(x, -\frac{1}{\epsilon n_\epsilon}DW^{Z,k}\left(x - Z_i^{0,\epsilon}, \epsilon i\right)\right)\right\}\mu(du|\,x)
\end{aligned}$$

and the conditional distribution of $\bar{v}_i^{\gamma^Z}$ under $(U^Z, \rho^Z, \{W^{Z,k}\}_{k=1}^K) \in \mathcal{MS}^Z$ in Construction 5.2 given $\bar{Z}_i^{\gamma^Z} = x$ is

$$\begin{aligned}\gamma_i^Z(dv|\, x) &= \sum_{k=1}^K \rho_k^Z\left(x - Z_i^{0,\epsilon}, \epsilon i\right) \exp\left\{\left\langle v, -DW^{Z,k}\left(x - Z_i^{0,\epsilon}, \epsilon i\right)\right\rangle\right. \\ &\quad \left. -\hat{H}_c\left(x, -DW^{Z,k}\left(\bar{Z}_i^{\gamma^Z} - Z_i^{0,\epsilon}, \epsilon i\right)\right)\right\} \hat{\mu}(dv|\, x).\end{aligned}$$

Combining this fact with the definition of X_i^n in terms of Z_i given by Construction 5.1 and the definitions of r^{n_ϵ} and r_Z in terms of Constructions 4.5 and 5.2 demonstrates that the scheme given by $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ in Construction 4.5 is identical to the scheme given by $(U^Z, \rho^Z, \{W^{Z,k}\}_{k=1}^K) \in \mathcal{MS}^Z$ in Construction 5.2. The proof in the other direction, that given an arbitrary $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}^Z$ and $(U^X, \rho^X, \{W^{X,k}\}_{k=1}^K)$ as in Definition 5.4 $(U^X, \rho^X, \{W^{X,k}\}_{k=1}^K) \in \mathcal{MS}$ and the scheme given by $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}^Z$ in Construction 5.2 is identical to the scheme given by $(U^X, \rho^X, \{W^{X,k}\}_{k=1}^K) \in \mathcal{MS}$ in Construction 4.5, follows in the same manner. ■

In addition, Definitions 5.3 and 5.4 demonstrate that a ranking of subsolutions based on their value at a particular position and time of the original processs $(Z(t), t)$ (such as at $(z_0, 0)$ which is motivated by Theorem 4.7) is preserved under a change of embeddings. Consequently the choice of embedding can be determined by convenience, they all suggesst the same changes of measure on the original process Z_i in the same order of preference.

6 CONSTRUCTING SUBSOLUTIONS

In order to use the result of Theorem 4.7 we need a subsolution $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ where $\mathcal{MS}$ is given by Definition 4.6. Because the resulting estimates r^n given by Construction 4.5 satisfy the following asymptotic decay rate of their second moments

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

we prefer U with large values at the origin $(0,0)$ (such a U is necessarily less than V).

Our approach is to find relatively simple class of elements in $\mathcal{S}_0$ which is broad enough to allow us to build desirable subsolutions $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ using (possibly compositions of) minimum and maximum mollifications found in Definition 4.3. This is similar to pevious work done in a large deviations setting, such as [19]. The difference in the moderate deviations control problem is that the drift is linear in the position and the control cost is quadratic (and independent of the position). This results in the linear-quadratic-regulator PDE which is relatively easy to work with. We can find classical sense solutions to (4.3) with quadratic terminal conditions, and for arbitrary terminal conditions we can write the control problem (2.9) as an infimum over the terminal condition and a quadratic final position cost (this is described in Theorem 6.5). Here we propose using classical sense solutions to (4.3) with affine terminal conditions (as opposed to quadratic) because many functions can be approximated well from below with an arbitrary number of minimums and maximums of affine functions, and the numerical work needed for solutions with affine terminal values does not depend on the affine parameters. Consequently once it is done you have the solution for any affine terminal condition. However, there may be situations where a quadratic terminal condition is superior.

Definition 6.1 *Let $\Omega(s,t)$ be the time dependent matrix which satisfies*

$$\frac{d}{dt}\exp\{\Omega(s,t)\} = Db(X^0(t))\exp\{\Omega(s,t)\}$$

for $s < t$ and

$$\Omega(s,s) = 0.$$

It can be easily verified that this implies

$$\frac{d}{dt}\exp\{\Omega(t,T)^T\} = -Db(X^0(t))^T\exp\{\Omega(t,T)^T\}. \tag{6.1}$$

Note that $\Omega(s,t)$ can be easily approximated numerically [4].

Definition 6.2 *Given arbitrary $c \in \mathbb{R}$ and $\xi \in \mathbb{R}^d$ let*

$$W^{c,\xi}(y,t) = \left\langle \exp\{\Omega(t,T)^T\}\xi, y\right\rangle - \frac{1}{2}\xi^T\int_t^T \exp\{\Omega(s,T)\}A\left(X^0(s)\right)\exp\{\Omega(s,T)^T\}ds\xi + c$$

The following theorem asserts that the functions given in Defintition 6.2 are in fact the classical sense solutions we seek.

Theorem 6.3 *$W^{c,\xi}(y,t)$ is a solution to (4.3) with terminal condition*

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

Proof. Clearly

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

from Definition 6.1. It remains to show that

$$W_t^{c,\xi}(y,t) = -\left\langle DW^{c,\xi}(y,t), Db(X^0(t))y\right\rangle + \frac{1}{2}\left\|DW^{c,\xi}(y,t)\right\|^2_{A(X^0(t))}.$$

We have

$$W_t^{c,\xi}(y,t) = y^T \frac{d}{dt} \exp\{\Omega(t,T)^T\}\xi + \frac{1}{2}\xi^T \exp\{\Omega(t,T)\} A\left(X^0(t)\right) \exp\{\Omega(t,T)^T\}\xi$$

which combined with (6.1) gives

$$\begin{aligned} W_t^{c,\xi}(y,t) &= -\left\langle \exp\{\Omega(t,T)^T\}\xi, Db(X^0(t))y \right\rangle + \frac{1}{2}\left\|\exp\{\Omega(t,T)^T\}\xi\right\|^2_{A(X^0(t))} \\ &= -\left\langle DW^{c,\xi}(y,t), Db(X^0(t))y \right\rangle + \frac{1}{2}\left\|DW^{c,\xi}(y,t)\right\|^2_{A(X^0(t))} \end{aligned}$$

and proves the result. ■

Note that

$$\sup_{(y,t)\in\mathbb{R}^d\otimes[0,T]} \left\|DW^{c,\xi}(y,t)\right\| \leq \|\xi\| \sup_{t\in[0,T]} \|\exp\{\Omega(s,T)\}\| < \infty$$

so $W^{c,\xi} \in \mathcal{S}_0$. The two quantities that need to be approximated numerically to compute $W^{c,\xi}$ from Definition 6.2,

$$\exp\{\Omega(t,T)^T\}$$

and

$$\int_t^T \exp\{\Omega(s,T)\} A\left(X^0(s)\right) \exp\{\Omega(s,T)^T\} ds,$$

do not depend on the choice of $c \in \mathbb{R}$ and $\xi \in \mathbb{R}^d$. Consequently we can use a large number of these solutions in our creation of $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{B}$ without substantially increasing the numerical work.

Given any $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ we have for all $(y,t) \in \mathbb{R}^d \otimes [0,T]$

$$U(y,t) \leq V(y,t)$$

where V is given by (4.2). We want to use available information about V to assist us in constructing a subsolution $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ that is large (or analogously close to V) around the origin. The following quadratic "costs" will allow us to write V as an infimum of the sum of two functions over final positions $z \in \mathbb{R}^d$.

Definition 6.4 *Define*

$$
\begin{aligned}
&C(y,t,z,r)\\
&\quad = (z-\exp\{\Omega(t,r)\}y)^T \left(\int_t^r \exp\left\{\Omega(q,r)\right\} A(X^0(q)) \exp\left\{\Omega(q,r)^T\right\} dq \right)^{-1} (z-\exp\{\Omega(t,r)\}y)
\end{aligned}
$$

for $0 \le t < r \le T$ *and* $y, z \in \mathbb{R}^d$ *using Definition 2.2 for the inverse of a symmetric, nonnegative definite matrix. In addition, use the convention that*

$$C(z,r) = C(0,0,z,r)$$

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

We now show how V can be written as a minimization over the quadtratic "position costs" defined above and terminal condition F.

Theorem 6.5 *Let* $C(y,t,z,r)$ *be given by Definition 6.4 and let* $V(y,t)$ *be given by (4.2). Then for any* $y \in \mathbb{R}^d$ *and* $0 \le t < T$

$$V(y,t) = \inf_{z \in \mathbb{R}^d} \{C(y,t,z,T) + F(z)\}.$$

In particular,

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

Proof. The proof is in Appendix B. ■

Recall that our method of building subsolutions involves approximating F from below by a function G which is given by minimums and maximums of affine functions, and then using the mollifications given in Definition 4.3 to get a classical sense subsolution U for the PDE (4.3) with terminal condition G instead of F. If we let V^G be the true solution to the PDE (4.3) with terminal condition G we have

$$V(y,t) \ge V^G(y,t) \ge U(y,t)$$

for all $(y,t) \in \mathbb{R}^d \otimes [0,T]$. We want U to be as large as possible at the origin, and we can make $U(y,t)$ arbitrarily close to $V^G(y,t)$ by making the mollification parameter δ used in Definition 4.3 smaller. However, in most problems we are interested in the origin is an attractor (because Y^n is a scaled version of X^n centered around its law of large numbers limit). Subsolutions with large second derivatives at the attractor (as we send $\delta \to 0$ the mollification approaches the nonsmooth function, either min or max, that it is approximating) can have large second moments for some finite noise values despite performing well asymptotically in accordinance with the result of Theorem 4.7. This is discussed thoroughly in the setting of one-dimensional diffusion processes in [17]. Although our present situation isn't the same, these issues may arise here and we recommend not making δ too small because you might hurt performance rather than enhance it (for instance in [17] $\delta = 2\epsilon$ was shown to be effective where ϵ was the variance of the noise). Consequently we turn our attention to decreasing $V(y,t) - V^G(y,t)$ with our choice of G. Theorem 6.5 gives

$$\begin{aligned} &V(y,t) - V^G(y,t) \\ &\quad = \inf_{z\in\mathbb{R}^d}\{C(y,t,z,T) + F(z)\} - \inf_{z\in\mathbb{R}^d}\{C(y,t,z,T) + G(z)\} \end{aligned}$$

and in particular

$$\begin{aligned} &V(0,0) - V^G(0,0) \\ &\quad = \inf_{z\in\mathbb{R}^d}\{C(z,T) + F(z)\} - \inf_{z\in\mathbb{R}^d}\{C(z,T) + G(z)\}. \end{aligned}$$

Consequently, we can ensure that

$$\inf_{z\in\mathbb{R}^d}\{C(z,T) + G(z)\} \tag{6.2}$$

is close to

$$\inf_{z\in\mathbb{R}^d}\{C(z,T) + F(z)\} \tag{6.3}$$

by ensuring that G approximates F well at the near minimizing $z \in \mathbb{R}^d$ in (6.3) and that the approximation does not degrade substantially enough away from these points to allow new minimizers in (6.2) that significantly reduce the infimum. Recall that with the method proposed here G is constructed by taking minimums and maximums of affine functions (however quadratic func-

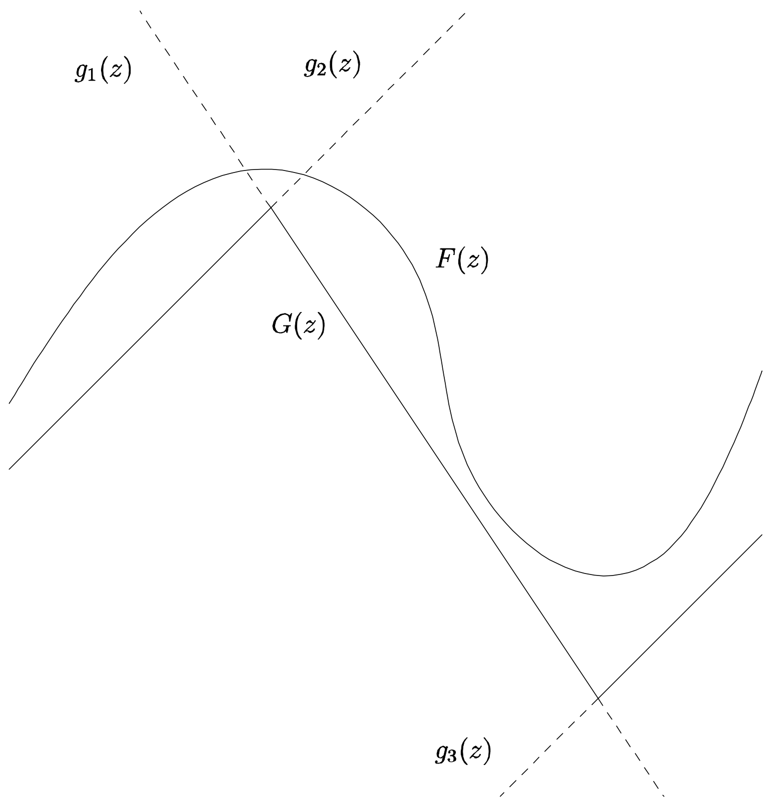

Figure 1: $G(z) = \max\{g_3(z), \min\{g_1(z), g_2(z)\}\}$ is an approximation of $F(z)$ from below

tions could also be used). This method will be demonstrated in the examples, but an illustration is provided in Figure 1. Here the affine functions are $g_1(x), g_2(x)$, and $g_3(x)$ and we are using $G(z) = \max\{g_3(z), \min\{g_1(z), g_2(z)\}\}$ to approximate $F(z)$ from below.

An additional issue to consider for subsolution creation is that we only specified the moment generating function be finite on a neighborhood around the origin, not necessarily on all of $\mathbb{R}^d$. For this reason we must restrict ourselves to subsolutions whose negative gradients stay within this neighborhood so that the exponential tilts they suggest can be implemented. Recall that if $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ then $U \in \mathcal{S}$. The randomized implementation usinig $(U, \rho, \{W^k\}_{k=1}^K) \in \mathcal{MS}$ in Construction 4.5 involves randomly selecting a $k \in \{1, ..., K\}$ at each time step using the weights given by ρ and using $-DW^k$ as our exponential tilt. The standard implementation using $U \in \mathcal{S}$ according to Construction 4.1 involves implementing an exponential tilt which is a linear combination of $\{-DW^k\}_{k=1}^K$ according to the weights ρ at each time step. We have observed comparable performance by both implementations, but for complex problems the randomized implementation may be easier to work with.

7 NUMERICAL EXAMPLES

In this section we test the performance of importance sampling schemes based on moderate deviations subsolutions in a variety of problems. When possible we compare them to their large deviations counterparts, and we investigate where they break down and why. In the last example (Section 7.4) we consider a problem with dynamics complex enough to make finding near-optimal large deviations subsolutions challenging, and we investigate the benefit of adding additional subsolutions to the mollifications in Definition 4.3.

7.1 One-Dimension, Bernoulli Noise and Linear Drift

We first consider a one-dimensional example where $\{\xi_i\}_{i=1}^{\infty}$ are iid with

$$P(\xi_i = 1) = P(\xi_i = -1) = .5$$

and

$$Z_{i+1} = Z_i - \epsilon Z_i + \epsilon \xi_i,\ Z_0 = 5.$$

Consequently

$$\hat{b}(Z_i) = -Z_i$$

and

$$v_i(Z_i) = \xi_i.$$

We will use the embedding in Construction 5.2. The law of large numbers limit is

$$Z^0(t) = 5e^{-t} \text{ for } t \in [0,1]$$

and we are interested in estimating the probability

$$\begin{aligned} p_\epsilon^\alpha &= P(Z(1) \notin (5e^{-1} - \alpha, 5e^{-1} + \alpha)) \\ &= P(Z(1) - Z^0(1) \notin (-\alpha, \alpha)) \end{aligned}$$

for some $\alpha > 0$. Based on Theorem 6.5

$$\begin{aligned} V(0,0) &= \inf_{z\in\mathbb{R}} \{\infty I_{(-\alpha,\alpha)}(z) + C(z,1)\} \\ &= \inf_{z\in\mathbb{R}} \{\infty I_{(-\alpha,\alpha)}(z) + z^2(1-e^{-2})^{-1}\} \end{aligned}$$

and clearly the minimizers are $z = \pm\alpha$. In order to approximate $\infty I_{(-\alpha,\alpha)}(z)$ from below with the minimum of affine functions G in such a way that

$$\inf_{z\in\mathbb{R}} \{\infty I_{(-\alpha,\alpha)}(z) + C(z,1)\} = \inf_{z\in\mathbb{R}} \{G(z) + C(z,1)\},$$

as was suggested in Section 6, we exploit the convexity of the terminal position cost $C(z,1)$ by choosing an affine function (which is a line here since $d = 1$) for each minimizer $\pm\alpha$ which is 0 at the minimizer and has slope equal to $-C_z(z,N)$ at the minimizer. We define classical solutions to (4.3) $W^{c_1,\beta_1}(y,t)$ and $W^{c_2,\beta_2}(y,t)$ as in Definition 6.2 with

$$\beta_1 = 2\alpha(1-e^{-2})^{-1}, c_1 = \alpha\beta_1$$

$$\beta_2 = -2\alpha(1-e^{-2})^{-1}, c_2 = -\alpha\beta_2$$

which have terminal values

$$W^{c_1,\beta_1}(z,1) = c_1 + \beta_1 z$$

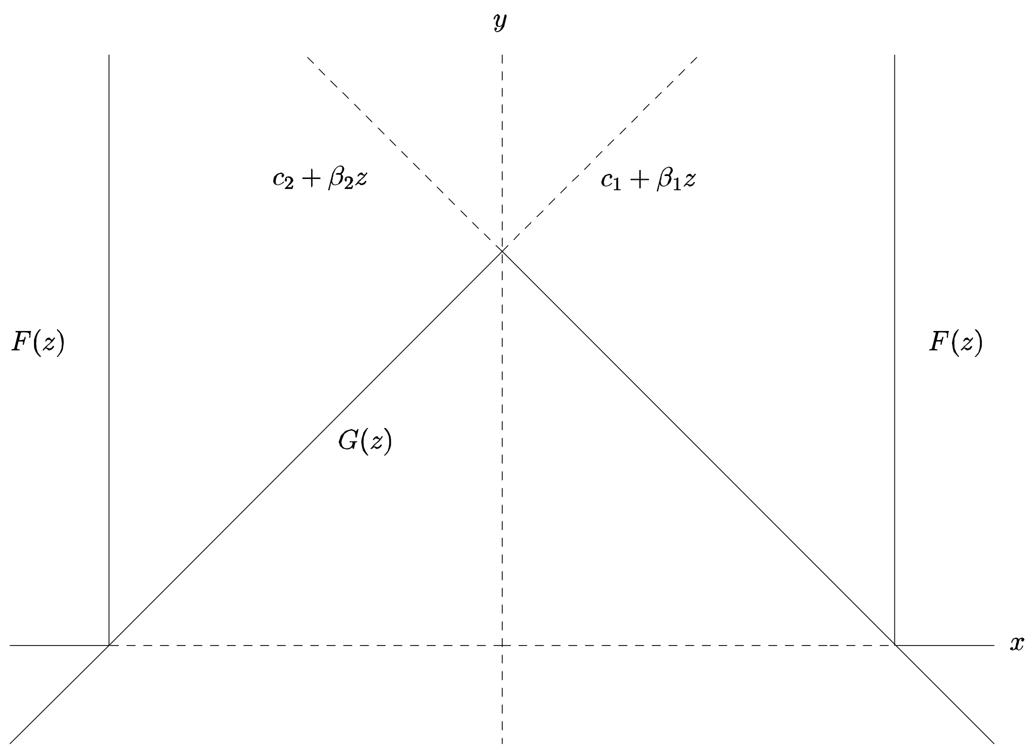

Figure 2: $F(z) = \infty I_{(-\alpha,\alpha)}(z)$ and $G(z) = \min\{c_1 + \beta_1 z, c_2 + \beta_2 z\}$

and

$$W^{c_2,\beta_2}(z,1) = c_2 + \beta_2 z.$$

Note that

$$G(z) = \min\{W^{c_1,\beta_1}(z,1), W^{c_2,\beta_2}(z,1)\} \leq \infty I_{(-\alpha,\alpha)}(z)$$

for all $z \in \mathbb{R}$ and

$$\begin{aligned}
&\inf_{z\in\mathbb{R}}\{G(z) + z^2(1-e^{-2})^{-1}\} \\
&\quad = \inf_{z\in\mathbb{R}}\{\infty I_{(-\alpha,\alpha)}(z) + z^2(1-e^{-2})^{-1}\} \\
&\quad = V(0,0)
\end{aligned}$$

again with minimizers $z = \pm\alpha$. Consequently we used a change of measure based on $U^{\delta,\min}(W^{c_1,\beta_1}, W^{c_2,\beta_2})$ given by Definition 4.3. We also implemented a change of measure based on a subsolution to the corresponding large deviation PDE as in [19]. This large deviations subsolution was created in analagous way to what was done above. It is the mollification of the minimum of two exact solutions based on affine terminal conditions which have the same value at the origin as the true solution. It is worth noting that more effort was needed to do this using the large deviation framework and in more complex situations the effort required could be prohibitively demanding. Due to concerns about a large second derivative at the attractor 0 mentioned in section 6 we choose $\delta = 2\epsilon$ (see [17] for a detailed discussion of this in the case of a one dimensional diffusion process). What is provided in the table below is an estimate of the standard deviation of the importance sampling estimator divided by an estimate of the probability of interest. Because these quantities are unknown we used 100,000 simulations to estimate the standard deviations and an independent 100,000 simulations using the large deviations subsolution to estimate the probability of interest. We used the randomized implementation given by Construction 4.5.

$\epsilon = 1/100$	$\alpha = .4$	$\alpha = .5$	$\alpha = .6$
Mod Std/Prob	4.45	11.08	197.15
Large Std/Prob	3.39	3.84	4.01
Prob	$3.84x10^{-10}$	$5.98x10^{-16}$	$6.80x10^{-25}$

Table 1: Comparing Moderate and Large Deviations IS for Bernoulli Noise

Note that because the drift is linear nothing is lost in the moderate deviations approximation of it (it is a linear approximation). The error comes from the quadratic approximation of the cost. As expected the importance sampling schemes based on moderate deviations perform comparably to those based on large deviations when estimating probabilities of events near the law of large numbers limit (small α) but are less effective when estimating probabilities of events farther away from this limit where the appoximation is less accurate (large α).

7.2 One-Dimension, Exponential Noise and Linear Drift

Recall that both the moderate deviations rate function and the result on asymptotic efficiency of importance sampling schemes given by subsolutions only require that the moment generating function of the noise be finite in a neighborhood of the origin, unlike in the large deviations setting where it must be finite everywhere. We explore one such example here. Although the large deviations results don't apply here, we can still choose a scheme based on them formally and compare its performance to a scheme based on moderate deviations. Let $\{\theta_i\}_{i=1}^{\infty}$ iid exponential random variables with rate parameter 1 and let

$$\xi_i = \theta_i - 1.$$

Let

$$Z_{i+1} = Z_i - \epsilon Z_i + \epsilon \xi_i,\ Z_0 = 5$$

so the drift is the same as in the previous example (Section 7.1). Despite the fact that the noise is different, the variance is still 1 so the moderate deviations approximations for these problems are identical. Consequently we use the same subsolutions and importance sampling scheme given in Section 7.1. Because the moderate deviations appromixation essentially assumes that the noise is normally distributed, for values of α larger than $(1 - e^{-2})/2$ moderate deviations importance sampling suggests exponential tilts which are outside the region where the moment generating function of ξ_i is finite (consequently they cannot be implemented). This suggests that as we approach this value of α the moderate deviations approximation of the noise is inadequate which may result in poor performance.

We created a large deviations subsolution (see [19]) in an analogous way to what we did for moderate deviations in that it is the mollification of the minimum of two exact solutions based on affine terminal conditions which have the same value at the origin as the true solution. It should be noted that unlike the large deviations subsolution in the first example (Section 7.1) where the noise was symmetric, or the moderate deviations subsolution which assumes that the noise is symmetric (in one dimension) due to its "Gaussian approximation," this large deviations subsolution takes into account that the noise behaves differently in the positive and negative directions and adjusts the scheme accordingly. Once again we choose $\delta = 2\epsilon$ due to concerns about the rest point at 0 (see Sections 6 and 7.1 and [17]), and we consider an estimate of the standard deviation of the importance sampling estimator divided by an estimate of the probability of interest. Because these quantities are unknown we used 10,000 simulations to estimate the standard deviation and an independent 10,000 simulations using the large deviations subsolution to estimate the probability of interest. We used the randomized implementation given by Construction 4.5.

$\epsilon = 1/200$	$\alpha = .2$	$\alpha = .3$	$\alpha = .4$
Mod Std/Prob	4.88	114.26	$1.89x10^{-22}$
Large Std/Prob	2.20	2.72	3.21
Prob	$3.43x10^{-5}$	$6.91x10^{-9}$	$2.09x10^{-13}$

Table 2: Comparing Moderate and Large Deviations IS for Exponential Noise

Once again because the drift is linear the only error in the moderate deviations "approximation" of the large deviations dynamics comes from the quadratic approximation of the cost. As expected, importance sampling based on moderate deviations performs well for smaller values of α where the approximation is accurate and degrades as α gets larger. Note that for $\alpha = .4$ the ratio in Table 2 is small but I am almost certain that this is only because the performance of the scheme has degraded to the point where 10,000 samples is no longer adequate to estimate the standard deviation. This explanation is supported by the fact that the estimate of the probability was of the order of 10^{-35} which very far off from the true value. Recall that $\alpha = .4$ is close to the region where the exponential tilts suggested by moderate deviations cannot even be implemented because the approximation of the dynamics is very inaccurate. It is worth noting that the importance sampling scheme based on large deviations performs very well despite the fact that the analysis (the rate function and asymptotic optimality) do not pertain to this problem.

7.3 Two-Dimensions with Linear Drift

In this subsection we combine the first two numerical examples given in Sections 7.1 and 7.2 into a two-dimensional problem. We have

$$Z_{i+1} = Z_i - \epsilon Z_i + \epsilon \xi_i,\ Z_0 = (5, 5)$$

where now $Z_i \in \mathbb{R}^2$ and the $\{\xi_i\}_{i=1}^{\infty}$ are iid with

$$\xi_i = (\theta_{i,1}, \theta_{i,2} - 1)$$

where

$$P(\theta_{i,1} = 1) = P(\theta_{i,1} = -1) = .5$$

and $\theta_{i,2}$ is exponentially distributed with rate parameter 1 and is independent of $\theta_{i,1}$. The dynamics in the two dimensions are independent, with the first dimension using the dynamics from the example in Section 7.1 and the second dimension using the dynamics from the example in Section 7.2. Consequently

$$\hat{b}(Z_i) = -Z_i$$

and

$$v_i(Z_i) = \xi_i.$$

We will use the embedding in Construction 5.2. The law of large numbers limit is

$$Z^0(t) = (5e^{-t}, 5e^{-t}) \text{ for } t \in [0, 1]$$

and we are interested in estimating the probability

$$\begin{aligned} p_\epsilon^\alpha &= P(Z(1) \notin (5e^{-1} - \alpha, 5e^{-1} + \alpha) \times (5e^{-1} - \alpha, 5e^{-1} + \alpha)) \\ &= P(Z(1) - Z^0(1) \notin (-\alpha, \alpha) \times (-\alpha, \alpha)) \end{aligned}$$

for some $\alpha > 0$. Based on Theorem 6.5

$$\begin{aligned} V(0,0) &= \inf_{z\in\mathbb{R}^2} \{\infty I_{(-\alpha,\alpha)\times(-\alpha,\alpha)}(z) + C(z, 1)\} \\ &= \inf_{z\in\mathbb{R}^2} \{\infty I_{(-\alpha,\alpha)\times(-\alpha,\alpha)}(z) + \|z\|^2 (1 - e^{-2})^{-1}\}. \end{aligned}$$

Consequently we can use the subsolutions from Section 7.1 to handle the first dimension and the subsolutions from Section 7.2 to handle the second dimension. Let $V^{1,1} = W^{c_1,\beta_1}$ and $V^{1,2} = W^{c_2,\beta_2}$ where W^{c_1,β_1} and W^{c_2,β_2} are given by Section 7.1 and let $V^{2,1} = W^{c_1,\beta_1}$ and $V^{2,2} = W^{c_2,\beta_2}$ where W^{c_1,β_1} and W^{c_2,β_2} are now given by Section 7.2 (in fact the definitions of W^{c_1,β_1} and W^{c_2,β_2} given in Sections 7.1 and 7.2 are identical, but we specify the dependence on the section because for large deviations the subsolutions in the two sections are not the same). If we write $z = (z_1, z_2)$ then

$V^{1,1}(z_1,t), V^{1,2}(z_1,t), V^{2,1}(z_2,t), V^{2,2}(z_2,t)$ are all subsolutions,

$$G(z) = \min\{V^{i,j}(z_i,1)\} \leq \infty I_{(-\alpha,\alpha)}(z)$$

for all $z \in \mathbb{R}$ and

$$\begin{aligned}
&\inf_{z\in\mathbb{R}}\{G(z) + \|z\|^2 (1-e^{-2})^{-1}\} \\
&\quad = \inf_{z\in\mathbb{R}}\{\infty I_{(-\alpha,\alpha)}(z) + \|z\|^2 (1-e^{-2})^{-1}\} \\
&\quad = V(0,0).
\end{aligned}$$

Consequently we used a change of measure based on $U^{\delta,\min}(V^{1,1}, V^{1,2}, V^{2,1}, V^{2,2})$ given by Definition 4.3. We create a subsolution to the large deviation problem in the same manner. We use the two large deviations subsolutions from Section 7.1 to handle the first dimension and the two large deviations subsolutions from Section 7.2 to handle the second dimension (note that unlike the moderate deviations situation the subsolutions in these two sections are not the same). We then get our large deviations subsolution from a minimum mollification of these four subsolutions as in Definition 4.3.

In contrast to the modereate deviations subsolutions the large deviations subsolutions in the second dimension (the one with exponential noise) take into account the lack of symmetry in the dynamics. We choose $\delta = 2\epsilon$ due to concerns about the rest point at $(0,0)$ (see Sections 6 and 7.1 and [17]), and we consider an estimate of the standard deviation of the importance sampling estimator divided by an estimate of the probability of interest. Because these quantities are unknown we used 10,000 simulations to estimate the standard deviation and an independent 10,000 simulations using the large deviations subsolution to estimate the probability of interest. We used the randomized implementation given by Construction 4.5.

$\epsilon = 1/200$	$\alpha = .2$	$\alpha = .3$	$\alpha = .4$
Mod Std/Prob	7.32	37.01	$1.89x10^{-5}$
Large Std/Prob	2.23	2.78	3.10
Prob	$3.37x10^{-5}$	$6.70x10^{-9}$	$2.28x10^{-13}$

Table 3: Comparing Moderate and Large Deviations IS in 2-dimensions

Once again the drift is linear so the error in the moderate deviations approximation comes from the quadratic approximation of the cost rather than the linear approximation of the drift. The results are similar to the previous example. Recall in Section 7.2 that the moderate deviations "Gaussian approximation" of the exponential noise is inaccurate for large α and suggests exponential tilts which in the positive direction amplify the controlled noise excessively. Here the moderate deviations algorithm still struggles for $\alpha = .4$ because of the exponential noise in the second dimension (the ratio in Table 3 is small due to very poor performance of the algorithm rather than success, the estimate of the probability itself was very far off). However, for smaller values of α moderate deviations importance sampling schemes perform comparably to their large deviations counterparts.

7.4 One-Dimension, State Dependent Noise and Nonlinear Drift

Our last numerical example considers more complex dynamics but in one-dimension. The drift is quadratic and the noise is Bernoulli but the variance depends on the state. The dynamics are

$$Z_{i+1} = Z_i - \epsilon Z_i^2 + \epsilon \xi_i(Z_i),\ Z_0 = 5$$

and

$$P(\xi_i(Z_i) = Z_i) = P(\xi_i(Z_i) = -Z_i) = .5.$$

Note that the process cannot escape the compact set $[-5, 5]$ so Condition 2.1 holds. We have

$$\hat{b}(Z_i) = -Z_i^2$$

and

$$v_i(Z_i) = \xi_i(Z_i).$$

We will use the embedding in Construction 5.2. The law of large numbers limit is

$$Z^0(t) = 1/(.2 + t) \text{ for } t \in [0, 1]$$

and we are interested in estimating the probability

$$\begin{aligned}
p^{\alpha}_{\epsilon,\eta} &= P(Z(1) \in [1/(.2+t) - \alpha - \eta, 1/(.2+t) - \alpha] \cup [1/(.2+t) + \alpha, 1/(.2+t) + \alpha + \eta]) \\
&= P(Z(1) - Z^0(1) \in [-\alpha - \eta, -\alpha] \cup [\alpha, \alpha + \eta])
\end{aligned}$$

for some $\alpha, \eta > 0$.

The purpose of this example is to investigate how well moderate deviations performs under these more complex dynamics and to compare two schemes based on different moderate deviations subsolutions. One of them takes into account that the set of interest is $[-\alpha - \eta, -\alpha] \cup [\alpha, \alpha + \eta]$ rather than simply $(-\infty, -\alpha] \cup [\alpha, \infty)$ by using a maximum mollification in addition to a minimum mollification, and the other does not. We have

$$C(z,1) = z^2 3(1.2^4)/(1.2^3 - .2^3),$$

$$C_z(z,1) = z6(1.2^4)/(1.2^3 - .2^3),$$

$$\exp\{\Omega(t,T)\} = \left(\frac{.2+t}{1.2}\right)^2,$$

$$A(X^0(t)) = \left(\frac{1}{.2+t}\right)^2,$$

and

$$\frac{1}{2}\int_t^T \exp\{\Omega(s,T)\} A\left(X^0(s)\right) \exp\{\Omega(s,T)^T\} ds = (1.2^3 - (.2+t)^3)/(6(1.2^4)).$$

Using Theorem 6.5

$$\begin{aligned}
V(0,0) &= \inf_{z \in \mathbb{R}} \{\infty I_{(-\infty,-\alpha-\eta)\cup(-\alpha,\alpha)\cup(\alpha+\eta,\infty)}(z) + C(z,1)\} \\
&= \inf_{z \in \mathbb{R}} \{\infty I_{(-\infty,-\alpha-\eta)\cup(-\alpha,\alpha)\cup(\alpha+\eta,\infty)}(z) + z^2 3(1.2^4)/(1.2^3 - .2^3)\}
\end{aligned}$$

and clearly the minimizers are $z = \pm\alpha$. Analogously to what was done in Section 7.1 we choose

$$\beta_1 = \alpha 6(1.2^4)/(1.2^3 - .2^3), c_1 = \alpha\beta_1$$

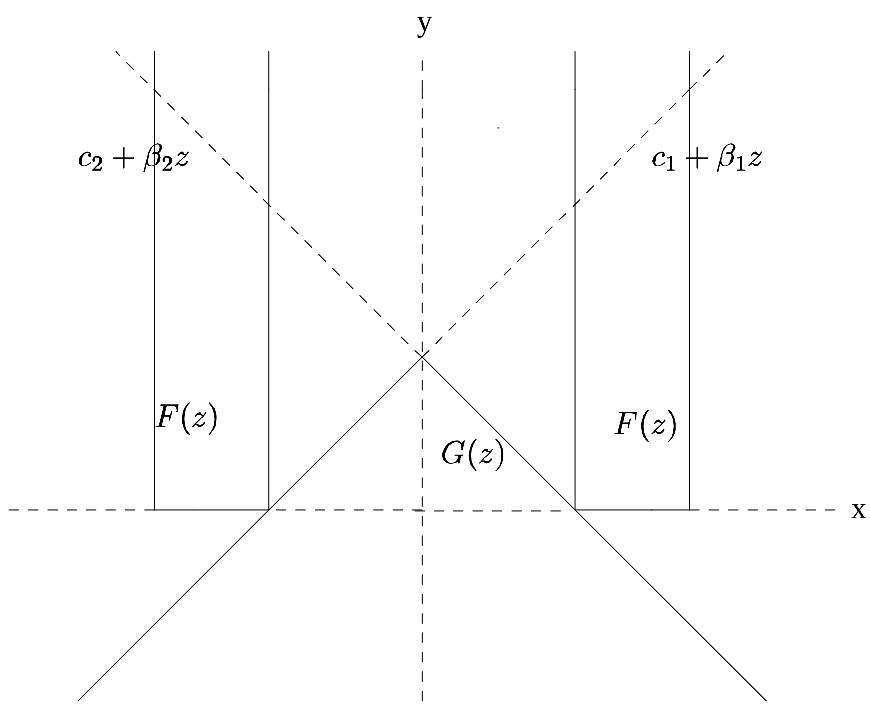

Figure 3: $F(z) = \infty I_{(-\infty,-\alpha-\eta)\cup(-\alpha,\alpha)\cup(\alpha+\eta,\infty)}(z)$ and $G(z) = \min\{c_1 + \beta_1 z, c_2 + \beta_2 z\}$

$$\beta_2 = -\alpha 6(1.2^4)/(1.2^3 - .2^3), c_2 = -\alpha\beta_2$$

and define $W^{c_1,\beta_1}(y,t)$ and $W^{c_2,\beta_2}(y,t)$ as in Definition 6.2. The terminal values are

$$W^{c_1,\beta_1}(z,1) = c_1 + \beta_1 z$$

and

$$W^{c_2,\beta_2}(z,1) = c_2 + \beta_2 z.$$

Note that

$$\begin{aligned} G(z) &= \min\{W^{c_1,\beta_1}(z,1), W^{c_2,\beta_2}(z,1)\} \\ &\leq \infty I_{(-\infty,-\alpha-\eta)\cup(-\alpha,\alpha)\cup(\alpha+\eta,\infty)}(z) \end{aligned}$$

for all $z \in \mathbb{R}$ and

$$\begin{aligned}
&\inf_{z\in\mathbb{R}}\{G(z) + z^2 3(1.2^4)/(1.2^3 - .2^3)\} \\
&\quad = \inf_{z\in\mathbb{R}}\{\infty I_{(-\infty,-\alpha-\eta)\cup(-\alpha,\alpha)\cup(\alpha+\eta,\infty)}(z) + z^2 3(1.2^4)/(1.2^3 - .2^3)\} \\
&\quad = V(0,0)
\end{aligned}$$

again with minimizers $z = \pm\alpha$. Consequently we create the subsolution given by $U^{\delta,\min}(W^{c_1,\beta_1}, W^{c_2,\beta_2})$ using Definition 4.3. In what follows we will refer to importance sampling based on this subsolution as the "Min" scheme. This is the same subsolution that we would use for the terminal condition

$$F(z) = \infty I_{(-\alpha,\alpha)}(z)$$

(or equivalently entry into the set $(-\infty,-\alpha]\cup[\alpha,\infty)$). Although the "Min" scheme performs well asymptotically according to Theorem 4.7, it continues to push the process to ∞ once it exceeds $\alpha+\eta$ and to $-\infty$ once it falls below $-\alpha-\eta$. To change this we introduce two additional subsolutions with affine terminal conditions. Let

$$\beta_3 = -\alpha 6(1.2^4)/(1.2^3 - .2^3), c_3 = (\alpha+\eta)\beta_3,$$

$$\beta_4 = \alpha 6(1.2^4)/(1.2^3 - .2^3), c_4 = -(\alpha+\eta)\beta_4$$

and define $W^{c_3,\beta_3}(y,t)$ and $W^{c_4,\beta_4}(y,t)$ as in Definition 6.2. Note that

$$\begin{aligned}
G(z) &= \max\{\min\{W^{c_1,\beta_1}(z,1), W^{c_2,\beta_2}(z,1)\}, W^{c_3,\beta_3}(z,1), W^{c_4,\beta_4}(z,1)\} \\
&\le \infty I_{(-\infty,-\alpha-\eta)\cup(-\alpha,\alpha)\cup(\alpha+\eta,\infty)}(z)
\end{aligned}$$

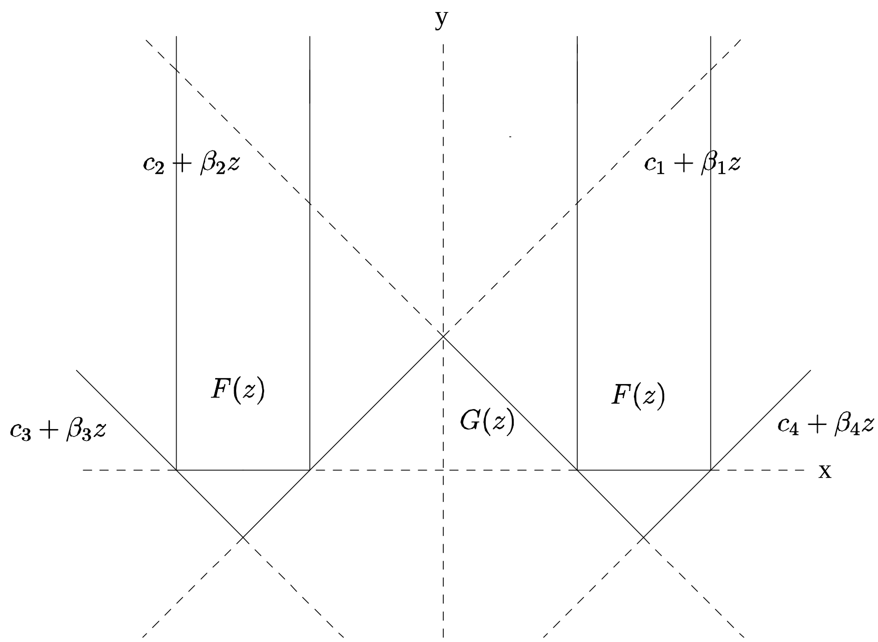

Figure 4: $F(z) = \infty I_{(-\infty,-\alpha-\eta)\cup(-\alpha,\alpha)\cup(\alpha+\eta,\infty)}(z)$ and $G(z) = \max\{\min\{c_1 + \beta_1 z, c_2 + \beta_2 z\}, c_3 + \beta_3 z, c_4 + \beta_4 z\}$

for all $z \in \mathbb{R}$ and

$$\begin{aligned}
&\inf_{z\in\mathbb{R}}\{G(z) + z^2 3(1.2^4)/(1.2^3 - .2^3)\} \\
&= \inf_{z\in\mathbb{R}}\{\infty I_{(-\infty,-\alpha-\eta)\cup(-\alpha,\alpha)\cup(\alpha+\eta,\infty)}(z) + z^2 3(1.2^4)/(1.2^3 - .2^3)\} \\
&= V(0,0)
\end{aligned}$$

again with minimizers $z = \pm\alpha$. We also create the subsolution given by $U^{\delta,\max}(U^{\delta,\min}(W^{c_1,\beta_1}, W^{c_2,\beta_2}), W^{c_3,\beta_3}, W^{c_4,\beta_4})$ using Definition 4.3 and refer to importance sampling based on this subsolution as the "MaxMin" scheme. The idea behind this scheme is to discourage the process from exceeding $\alpha + \eta$ or falling below $-\alpha - \eta$. However, additonal subsolutions and mollifications require more work. They could also result in second derivative problems like those discussed in [17] (we describe them briefly in Sections 6 and 7.1). In an attempt to account for those second derivative problems we continue to use mollification parameter $\delta = 2\epsilon$ in these simulations. Due to the complexity of this problem we did not try to a find large deviations subsolution. The tables below contain estimates of the standard deviation of the importance sampling estimators based on the "Min" and "MaxMin" schemes, as well as their estimates of the probability of interest. These are all based on 100,000 simulations. We used the randomized implementation given by Construction 4.5.

$\boldsymbol{\epsilon = 1/400, \alpha = .1}$	$\eta = .0015$	$\eta = .001$	$\eta = .0005$
Min Est	$1.29x10^{-4}$	$2.07x10^{-5}$	$4.70x10^{-6}$
Min Std Est	$3.17x10^{-2}$	$2.80x10^{-3}$	$4.05x10^{-4}$
MaxMin Est	$2.44x10^{-5}$	$1.69x10^{-5}$	$6.37x10^{-6}$
MaxMin Std Est	$1.32x10^{-3}$	$1.39x10^{-3}$	$4.29x10^{-4}$

Table 4: Comparing "Min" and "MaxMin' Schemes for $\alpha = .1$

$\boldsymbol{\epsilon = 1/400, \alpha = .11}$	$\eta = .0015$	$\eta = .001$	$\eta = .0005$
Min Est	$1.61x10^{-6}$	$5.24x10^{-7}$	$6.52x10^{-7}$
Min Std Est	$1.21x10^{-4}$	$6.80x10^{-5}$	$8.44x10^{-5}$
MaxMin Est	$4.33x10^{-6}$	$5.62x10^{-6}$	$1.90x10^{-6}$
MaxMin Std Est	$5.24x10^{-4}$	$1.09x10^{-3}$	$2.68x10^{-4}$

Table 5: Comparing "Min" and "MaxMin' Schemes for $\alpha = .11$

$\boldsymbol{\epsilon = 1/400, \alpha = .12}$	$\eta = .0015$	$\eta = .001$	$\eta = .0005$
Min Est	$3.40x10^{-7}$	$3.49x10^{-7}$	$1.81x10^{-7}$
Min Std Est	$2.68x10^{-5}$	$3.19x10^{-5}$	$2.78x^{-5}$
MaxMin Est	$3.78x10^{-7}$	$4.00x10^{-7}$	$1.10x10^{-7}$
MaxMin Std Est	$3.40x10^{-5}$	$4.09x10^{-5}$	$8.56x10^{-6}$

Table 6: Comparing "Min" and "MaxMin' Schemes for $\alpha = .12$

Although the results are comparable for both schemes, it does seem like the "MaxMin" scheme, which attempts to discourage the process from going too far away from the law of large numbers limit by employing a maximum mollification of additional subsolutions, performs slightly better for small η as expected. Overall both moderate deviations based schemes performed fairly well despite the complexity of this problem and the relative ease of finding moderate deviations subsolutions. More investigation is necessary to determine near-optimal mollification parameters, when it is worthwhile to add additional subsolutions, and if subsolutions based on quadratic terminal conditions can be useful.

A TIGHTNESS PROOF

Assume that the bound (2.11) holds. We will show tightness of the $\{\varsigma^{n,\eta^n}\}$ measures using the following lemma.

Lemma A.1 *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be arbitrary, define*

$$H(\alpha) = \log\left(\int_{\mathbb{R}^d} e^{\langle u,\alpha\rangle}\mu(du)\right)$$

and assume there exists $\lambda > 0$, $K_{mgf} < \infty$ such that

$$\sup_{\|\alpha\|\le\lambda} |H(\alpha)| < K_{mgf}. \tag{A.1}$$

Then if

$$L(\beta) = \sup_{\alpha\in\mathbb{R}^d}\{\langle\alpha,\beta\rangle - H(\alpha)\} \tag{A.2}$$

is the Legendre transform of $H(\cdot)$ we have for any $\eta \in \mathcal{P}(\mathbb{R}^d)$

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

Proof. While the result is likely known we could not locate a proof (see [12, Lemma 6.2.3(f)] for a proof when $H(\alpha)$ is finite for all $\alpha \in \mathbb{R}^d$), and so for completeness provide the details. If $R(\eta\|\,\mu) =$

∞ the lemma is automatically true, so we assume $R(\eta\| \mu) < \infty$. Define $\ell(b) \doteq b\log b - b + 1$ and note that for $a, b \geq 0$

$$ab \leq e^a + \ell(b).$$

From (A.1) we have

$$\int_{\mathbb{R}^d} e^{\frac{\lambda}{\sqrt{d}}\|u\|}\mu(du) \leq 2^d e^{K_{\mathrm{mgf}}} < \infty.$$

Therefore

$$\begin{aligned}
&\int_{\mathbb{R}^d} \frac{\lambda}{\sqrt{d}} \|u\| \frac{d\eta}{d\mu}(u)\mu(du) \\
&\quad\leq \int_{\mathbb{R}^d} e^{\frac{\lambda}{\sqrt{d}}\|u\|}\mu(du) + \int_{\mathbb{R}^d} \ell\left(\frac{d\eta}{d\mu}(u)\right)\mu(du) \\
&\quad\leq 2^d e^{K_{\mathrm{mgf}}} + R(\eta\| \mu),
\end{aligned}$$

and consequently for any $\alpha \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} \|\alpha\| \|u\| \frac{d\eta}{d\mu}(u)\mu(du) \leq \frac{\sqrt{d}\,\|\alpha\|}{\lambda}\left(2^d e^{K_{\mathrm{mgf}}} + R(\eta\| \mu)\right) < \infty. \tag{A.3}$$

Define the bounded, continuous function

$$F_K(u,\alpha) = \begin{cases} \langle \alpha, u\rangle & \text{if } |\langle \alpha, u\rangle| \leq K \\ \frac{K\langle \alpha, u\rangle}{|\langle \alpha, u\rangle|} & \text{otherwise,} \end{cases}$$

and note that (A.3) and dominated convergence give

$$\lim_{K\to\infty} \int_{\mathbb{R}^d} F_K(u,\alpha)\eta(du) = \left\langle \alpha, \int_{\mathbb{R}^d} u\eta(du)\right\rangle.$$

In addition, dominated convergence gives

$$\lim_{K\to\infty} \int_{\{u:\langle\alpha,u\rangle<0\}} e^{F_K(u,\alpha)}\mu(du) = \int_{\{u:\langle\alpha,u\rangle<0\}} e^{\langle\alpha,u\rangle}\mu(du)$$

and monotone convergence gives

$$\lim_{K\to\infty}\int_{\{u:\langle\alpha,u\rangle\geq 0\}} e^{F_K(u,\alpha)}\mu(du)=\int_{\{u:\langle\alpha,u\rangle\geq 0\}} e^{\langle\alpha,u\rangle}\mu(du),$$

so

$$\lim_{K\to\infty}\log\left(\int_{\mathbb{R}^d} e^{F_K(u,\alpha)}\mu(du)\right)=H(\alpha).$$

By the Donsker-Varadhan variational formula [12, Lemma 1.4.3(a)]

$$R(\eta\|\,\mu)\geq\int_{\mathbb{R}^d}F_K(u,\alpha)\eta(du)-\log\left(\int_{\mathbb{R}^d} e^{F_K(u,\alpha)}\mu(du)\right)$$

for all $K<\infty$ and $\alpha\in\mathbb{R}^d$, and so

$$R(\eta\|\,\mu)\geq\sup_{\alpha\in\mathbb{R}^d}\left\{\left\langle\alpha,\int_{\mathbb{R}^d}u\eta(du)\right\rangle-H(\alpha)\right\}=L\left(\int_{\mathbb{R}^d}u\eta(du)\right),$$

which completes the proof of the lemma. ■

The lemma implies the following theorem, which in turn will give tightness of $\{\hat{\eta}^n\}$.

Theorem A.2 *Assume Condition 2.1 and (2.11). For the processes $\{w^{n,\eta^n}\}$ obtained in Construction 2.4*

$$\sup_{n\in\mathbb{N}}E\left[\int_0^1 a(n)\sqrt{n}\left\|w^{n,\eta^n}(s)\right\|ds\right]<\infty.$$

In addition, $\{a(n)\sqrt{n}w^{n,\eta^n}(\cdot)\}_{n\in\mathbb{N}}$ is uniformly integrable in the sense that

$$\lim_{C\to\infty}\limsup_{n\to\infty}E\left[\int_0^1 1_{\{a(n)\sqrt{n}\|w^{n,\eta^n}(s)\|>C\}}a(n)\sqrt{n}\left\|w^{n,\eta^n}(s)\right\|ds\right]=0.$$

Proof. We use the following inequality. Let $G>0$ satisfy $\lambda_{DA}\min_{n\in\mathbb{N}}\{a(n)\sqrt{n}\}=\sqrt{G}$ [recall (2.8)] so that $\lambda_{DA}\geq\sqrt{G}/a(n)\sqrt{n}$ for all n. Define L_c by (A.2). Let $\bar{K}\doteq\lambda_{DA}K_{DA}+K_A/2$. Then with e_i denoting the standard unit vectors

$$\begin{aligned}
&a(n)^2 n L_c(x,\beta) \\
&\quad = \sup_{\alpha\in\mathbb{R}^d} \left[a(n)\sqrt{n}\left\langle \alpha, a(n)\sqrt{n}\beta \right\rangle - a(n)^2 n H_c(x,\alpha) \right] \\
&\quad \geq \pm a(n)\sqrt{n}\left\langle \frac{\sqrt{G}}{a(n)\sqrt{n}} e_i, a(n)\sqrt{n}\beta \right\rangle - a(n)^2 n H_c\left(x, \pm\frac{\sqrt{G}}{a(n)\sqrt{n}} e_i \right) \\
&\quad \geq \pm\sqrt{G} a(n)\sqrt{n}\beta_i - \frac{1}{2} G \left\| A(x) \right\| - G\lambda_{DA} K_{DA} \\
&\quad \geq \pm\sqrt{G} a(n)\sqrt{n}\beta_i - G\bar{K},
\end{aligned}$$

where the first inequality follows from making a specific choice of α and the second uses (2.6). Therefore

$$da(n)^2 n L_c(x,\beta) + dG\bar{K} \geq \sqrt{G} a(n)\sqrt{n} \left\| \beta \right\|. \tag{A.4}$$

Using the bound on L_c from Lemma A.1 together with (2.11),

$$\begin{aligned}
&d\left(\frac{K_E}{\sqrt{G}} + \sqrt{G}\bar{K} \right) \\
&\quad \geq \frac{da(n)^2 n}{\sqrt{G}} E\left[\int_0^1 L_c\left(\bar{X}^{n,\eta^n}\left(\frac{\lfloor ns \rfloor}{n} \right), w^{n,\eta^n}(s) \right) ds \right] + d\sqrt{G}\bar{K} \\
&\quad \geq E\left[\int_0^1 a(n)\sqrt{n} \left\| w^{n,\eta^n}(s) \right\| ds \right].
\end{aligned} \tag{A.5}$$

For the uniform integrability, let $C \in (1,\infty)$ be arbitrary and consider n large enough that

$$\min\{\lambda_{DA}, 1\} \geq \frac{\sqrt{C}}{a(n)\sqrt{n}}.$$

Since $\lambda_{DA} \geq 1/a(n)\sqrt{n}$ (which corresponds to using the estimate above with $G = 1$) we have

$$E\left[\int_0^1 a(n)\sqrt{n} \left\| w^{n,\eta^n}(s) \right\| ds \right] \leq K^* \doteq d\left(K_E + \frac{1}{2} K_A + \lambda_{DA} K_{DA} \right).$$

Therefore

$$E\left[\int_0^1 1_{\{a(n)\sqrt{n}\|w^{n,\eta^n}(s)\| > C\}} ds \right] \leq \frac{K^*}{C},$$

which combined with the estimate (A.4) with G replaced by C and (A.5) gives

$$
\begin{aligned}
&\sqrt{C}E\left[\int_0^1 1_{\{a(n)\sqrt{n}\|w^{n,\eta^n}(s)\|>C\}}a(n)\sqrt{n}\left\|w^{n,\eta^n}(s)\right\|ds\right] \\
&\quad\leq E\left[d\int_0^1 1_{\{a(n)\sqrt{n}\|w^{n,\eta^n}(s)\|>C\}}\left(a(n)^2 nL_c\left(\bar{X}^{n,\eta^n}\left(\frac{\lfloor ns\rfloor}{n}\right),w^{n,\eta^n}(s)\right)+C\bar{K}\right)ds\right] \\
&\quad\leq da(n)^2 nE\left[\int_0^1 L_c\left(\bar{X}^{n,\eta^n}\left(\frac{\lfloor ns\rfloor}{n}\right),w^{n,\eta^n}(s)\right)ds\right] \\
&\qquad+Cd\bar{K}E\left[\int_0^1 1_{\{a(n)\sqrt{n}\|w^{n,\eta^n}(s)\|>C\}}ds\right] \\
&\quad\leq K^* d\left(1+\bar{K}\right).
\end{aligned}
$$

We conclude that

$$
\lim_{C\to\infty}\limsup_{n\to\infty}E\left[\int_0^1 1_{\{a(n)\sqrt{n}\|w^{n,\eta^n}(s)\|>C\}}a(n)\sqrt{n}\left\|w^{n,\eta^n}(s)\right\|ds\right]=0,
$$

which is the claimed uniform integrability. ■

This gives (2.12), and we now continue to prove the rest of Theorem 2.5. Note that $g(w,t)=\|w\|$ is a tightness function on $\mathbb{R}^d\otimes[0,1]$, so by [12, Theorem A.3.17]

$$
G(\eta)=\int_{\mathbb{R}^d\otimes[0,1]}\|w\|\,\eta(dw\otimes dt)
$$

is a tightness function on $\mathcal{P}(\mathbb{R}^d\otimes[0,1])$ and

$$
\bar{G}(\gamma)=\int_{\mathcal{P}\left(\mathbb{R}^d\otimes[0,1]\right)}\int_{\mathbb{R}^d\otimes[0,1]}\|w\|\,\eta(dw\otimes dt)\gamma(d\eta)
$$

is a tightness function on $\mathcal{P}(\mathcal{P}(\mathbb{R}^d\otimes[0,1]))$. Since

$$
\sup_{n\in\mathbb{N}}EG(\varsigma^{n,\eta^n})=\sup_{n\in\mathbb{N}}E\left[\int\|w\|\,\varsigma^{n,\eta^n}(dw\otimes dt)\right]=\sup_{n\in\mathbb{N}}E\left[\int_0^1 a(n)\sqrt{n}\left\|w^{n,\eta^n}(s)\right\|ds\right]<\infty,
$$

$\{\varsigma^{n,\eta^n}\}$ is tight and consequently there is a subsequence of $\{\varsigma^{n,\eta^n}\}$ which converges weakly. To simplify notation we retain n as the index of this convergent subsequence, and denote the weak limit of $\{\varsigma^{n,\eta^n}\}$ by ς. Note that for all n the second marginal of $\varsigma^{n,\eta^n}(dy\otimes dt)$, which we denote by $\varsigma_2^{n,\eta^n}(dt)$, is Lebesgue measure, and therefore $\varsigma_2(dt)$ is Lebesgue measure with probability 1.

Our aim is to show that $\bar{Y}^{n,\eta^n}(t) \to \hat{Y}(t)$ weakly in $C([0,1]:\mathbb{R}^d)$, where $\hat{Y}(t)$ is given by (2.13) in terms of the weak limit ς. To achieve this we introduce the following processes which serve as intermediate steps. Let $\check{Y}_0^{n,\eta^n} = 0$ and

$$\check{Y}_{i+1}^{n,\eta^n} = \check{Y}_i^{n,\eta^n} + \frac{a(n)}{\sqrt{n}}\left(b\left(X_i^{n,0} + \frac{1}{a(n)\sqrt{n}}\check{Y}_i^{n,\eta^n}\right) - b\left(X_i^{n,0}\right)\right) + \frac{a(n)}{\sqrt{n}}w^{n,\eta^n}\left(\frac{i}{n}\right),$$

together with its continuous time linear interpolation defined for $t \in [i/n, i/n+1/n]$ by

$$\check{Y}^{n,\eta^n}(t) = (i+1-nt)\check{Y}_i^{n,\eta^n} + (nt-i)\check{Y}_{i+1}^{n,\eta^n}.$$

Also let

$$\hat{Y}^{n,\eta^n}(t) = \int_0^t Db\left(X^0(s)\right)\hat{Y}^{n,\eta^n}(s)ds + \int_0^t \hat{w}^{n,\eta^n}(s)ds \tag{A.6}$$

where

$$\hat{w}^{n,\eta^n}(t) = \int_{\mathbb{R}^d} w\varsigma_{1|2}^{n,\eta^n}(dw\,|t)$$

as in Construction 2.4. These are both random variables taking values in $C([0,1]:\mathbb{R}^d)$. Note that $\bar{Y}^{n,\eta^n}$ differs from $\check{Y}^{n,\eta^n}$ because $\bar{Y}^{n,\eta^n}$ is driven by the actual noises and $\check{Y}^{n,\eta^n}$ is driven by their conditional means. While the driving terms of $\hat{Y}^{n,\eta^n}$ and $\check{Y}^{n,\eta^n}$ are the same [recall that $a(n)\sqrt{n}w^{n,\eta^n}(t) = \hat{w}^{n,\eta^n}(t)$], they differ in that $\check{Y}^{n,\eta^n}$ is still a linear interpolation of a discrete time process whereas $\hat{Y}^{n,\eta^n}$ satisfies an ODE. The goal is to show that along the subsequence where $\varsigma^{n,\eta^n} \to \varsigma$ weakly

$$\bar{Y}^{n,\eta^n} - \check{Y}^{n,\eta^n} \to 0, \check{Y}^{n,\eta^n} - \hat{Y}^{n,\eta^n} \to 0, \text{ and } \hat{Y}^{n,\eta^n} \to \hat{Y}$$

in $C([0,1]:\mathbb{R}^d)$, all in distribution. To show $\hat{Y}^{n,\eta^n} \to \hat{Y}$ we show that $\{\hat{Y}^{n,\eta^n}\}$ is tight in $C([0,1]:\mathbb{R}^d)$ and use the mapping defined by (A.6) from $\int_0^\cdot \hat{w}^{n,\eta^n}$ to $\hat{Y}^{n,\eta^n}$. Let

$$\phi^u(t) = \int_0^t Db(X^0(s))\phi^u(s)ds + \int_0^t u(s)ds. \tag{A.7}$$

Recall that $\sup_{x\in\mathbb{R}^d}\|Db(x)\| \le K_b$. The following lemma is an easy consequence of Gronwall's inequality.

Lemma A.3 *Let $u \in L^1([0,1]:\mathbb{R}^d)$ be arbitrary and ϕ^u be defined as in (A.7). Then for $0 \le s \le t \le 1$*

$$\|\phi^u(t) - \phi^u(s)\| \le (t-s)K_b e^{K_b} \int_0^1 \|u(r)\|\,dr + \int_s^t \|u(r)\|\,dr.$$

With this lemma and the uniform integrability of $\{\varsigma^{n,\eta^n}\}$ given in Theorem A.2, tightness follows.

Lemma A.4 *Assume Condition 2.1 and (2.11). The sequence $\{\hat{Y}^{n,\eta^n}\}$ defined in (A.6) in terms of the measures $\{\eta^n\}$ via Construction 2.4 is tight in $C([0,1]:\mathbb{R}^d)$, as is $\{\int_0^\cdot \hat{w}^{n,\eta^n} ds\}$.*

Proof. It suffices to show that for any $\varepsilon > 0$ there is $\delta > 0$ such that

$$\limsup_{n\to\infty} P\left(\sup_{|s-t|\le\delta} \left\|\hat{Y}^{n,\eta^n}(t) - \hat{Y}^{n,\eta^n}(s)\right\| > \varepsilon\right) < \varepsilon.$$

Define

$$\begin{aligned} T(C) &\doteq \limsup_{n\to\infty} E\left[\int_0^1 1_{\{\|\hat{w}^{n,\eta^n}(t)\|>C\}} \left\|\hat{w}^{n,\eta^n}(t)\right\| dt\right] \\ &= \limsup_{n\to\infty} E\left[\int_{\{\|w\|>C\}} \|w\|\, \varsigma^{n,\eta^n}(dw\otimes dt)\right]. \end{aligned}$$

By Theorem A.2 $T(C) \to 0$ as $C \to \infty$. Define also $K_\eta = \sup_{n\in\mathbb{N}} E\int_0^1 \left\|\hat{w}^{n,\eta^n}(t)\right\| dt$, which is finite by Theorem A.2. Let $\varepsilon > 0$ be arbitrary. Then for any $s < t$ satisfying $t - s \le \delta$ the previous lemma implies

$$\left\|\hat{Y}^{n,\eta^n}(t) - \hat{Y}^{n,\eta^n}(s)\right\| \le \delta K_b e^{K_b} \int_0^1 \left\|\hat{w}^{n,\eta^n}(r)\right\| dr + \int_s^t \left\|\hat{w}^{n,\eta^n}(r)\right\| dr.$$

Since

$$\int_s^t \left\|\hat{w}^{n,\eta^n}(r)\right\| dr \le C\delta + \int_0^1 1_{\{\|\hat{w}^{n,\eta^n}(r)\|>C\}} \left\|\hat{w}^{n,\eta^n}(r)\right\| dr,$$

it follows that

$$\begin{aligned} \left\|\hat{Y}^n(t) - \hat{Y}^n(s)\right\| \le \delta &\left(C + K_b e^{K_b} \int_0^1 \left\|\hat{w}^{n,\eta^n}(r)\right\| dr\right) \\ &+ \int_0^1 1_{\{\|\hat{w}^{n,\eta^n}(r)\|>C\}} \left\|\hat{w}^{n,\eta^n}(r)\right\| dr. \end{aligned}$$

Hence by Markov's inequality

$$
\begin{aligned}
&\limsup_{n\to\infty} P\left(\sup_{|s-t|\le\delta}\left\|\hat{Y}^{n,\eta^n}(t)-\hat{Y}^{n,\eta^n}(s)\right\|>\varepsilon\right)\\
&\quad\le \frac{\delta}{\varepsilon}\limsup_{n\to\infty} E\left[\left(C+K_b e^{K_b}\int_0^1\left\|\hat{w}^{n,\eta^n}(r)\right\|dr\right)\right]\\
&\quad\quad+\frac{1}{\varepsilon}\limsup_{n\to\infty} E\left[\int_0^1 1_{\{\|\hat{w}^{n,\eta^n}(r)\|>C\}}\left\|\hat{w}^{n,\eta^n}(r)\right\|dr\right]\\
&\quad\le \frac{\delta}{\varepsilon}(C+K_b e^{K_b}K_\eta)+\frac{1}{\varepsilon}T(C).
\end{aligned}
$$

Choose $C<\infty$ such that $T(C)<\varepsilon^2/2$ and then choose $\delta>0$ so that the $\delta(C+K_b e^{K_b}K_\eta)<\varepsilon^2/2$. This shows the tightness of $\{\hat{Y}^n\}$. The tightness of $\{\int_0^\cdot \hat{w}^{n,\eta^n}ds\}$ is simpler, and follows from the bound

$$
\limsup_{n\to\infty} P\left(\sup_{|s-t|\le\delta}\int_s^t\left\|\hat{w}^{n,\eta^n}(r)\right\|dr>\varepsilon\right)\le\delta\frac{C}{\varepsilon}+\frac{1}{\varepsilon}T(C).
$$

■

We still need to show that $\hat{Y}^{n,\eta^n}$ converges to $\hat{Y}$. This also relies on the uniform integrability given by Theorem A.2.

Lemma A.5 *Assume Condition 2.1 and (2.11). Let the sequence $\{\hat{Y}^{n,\eta^n}(t)\}$ be defined by (A.6), let $\hat{Y}(t)$ be defined by (2.13), and consider a convergent subsequence $\{(\hat{Y}^{n,\eta^n},\varsigma^{n,\eta^n})\}$ with limit $(\hat{Y}^*,\varsigma)$. Then w.p.1 $\hat{Y}^*=\hat{Y}$.*

Proof. We can write

$$
\hat{Y}^{n,\eta^n}(t)=\int_0^t Db(X^0(s))\hat{Y}^{n,\eta^n}(s)ds+\int_0^t\int_{\mathbb{R}^d} w\varsigma^{n,\eta^n}(dw\otimes ds).
$$

Using the uniform integrability proved in Theorem A.2 and that ς_2 is Lebesgue measure w.p.1, sending $n\to\infty$ and using the definition of $\hat{w}$ gives

$$
\begin{aligned}
\hat{Y}^*(t) &= \int_0^t Db(X^0(s))\hat{Y}^*(s)ds+\int_0^t\int_{\mathbb{R}^d} w\varsigma(dw\otimes ds)\\
&= \int_0^t Db(X^0(s))\hat{Y}^*(s)ds+\int_0^t \hat{w}(s)ds.
\end{aligned}
$$

By uniqueness of the solution, $\hat{Y}^*=\hat{Y}$ follows. ■

It remains to show $\bar{Y}^{n,\eta^n} - \check{Y}^{n,\eta^n} \to 0$ and $\check{Y}^{n,\eta^n} - \hat{Y}^{n,\eta^n} \to 0$. We begin with $\bar{Y}^{n,\eta^n} - \check{Y}^{n,\eta^n} \to 0$. Recall that the difference between $\bar{Y}^{n,\eta^n}$ and $\check{Y}^{n,\eta^n}$ is that the first is driven by the actual noises and the second is driven by their conditional means. The following theorem is a law of large numbers type result for the difference between the noises and their conditional means, and is the most complicated part of the analysis.

Theorem A.6 *Assume Condition 2.1 and (2.11). Consider the sequence $\{\bar{u}_i^{n,\eta^n}\}_{i=0,\ldots,n-1}$ of controlled noises and $\{w^{n,\eta^n}(i/n)\}_{i=0,\ldots,n-1}$ of means of the controlled noises as in Construction 2.4. For $i \in \{1,\ldots,n\}$ let*

$$W_i^n \doteq \frac{1}{n}\sum_{j=0}^{i-1} a(n)\sqrt{n}\left(\bar{u}_i^{n,\eta^n} - w^{n,\eta^n}(i/n)\right).$$

Then for any $\delta > 0$

$$\lim_{n\to\infty} P\left[\max_{i\in\{1,\ldots,n\}} \|W_i^n\| \geq \delta\right] = 0.$$

Proof. According to (2.11)

$$\frac{1}{n}\sum_{i=0}^{n-1} E[R(\eta_i^n \| \mu(\cdot|\bar{X}_i^{n,\eta^n}))] \leq \frac{K_E}{a^2(n)n}.$$

Because of this the (random) Radon-Nikodym derivatives

$$f_i^n(u) = \frac{d\eta_i^n}{d\mu(\cdot|\bar{X}_i^{n,\eta^n})}(u)$$

are well defined and can be selected in a measurable way. We will control the magnitude of the noise when the Radon-Nikodym derivative is large by bounding

$$\frac{1}{n}\sum_{i=0}^{n-1} E[1_{\{f_i^n(\bar{u}_i^{n,\eta^n})\geq r\}} \left\|\bar{u}_i^{n,\eta^n}\right\|]$$

for large r.

From the bound on the moment generating function (2.3),

$$\sup_{x\in\mathbb{R}^d} \int_{\mathbb{R}^d} e^{\frac{\lambda}{2^d}\|u\|}\mu(du|x) \leq 2^d e^{dK_{\mathrm{mgf}}}.$$

Let $\sigma=\min\{\lambda/2^{d+1},1\}$ and recall the definition $\ell(b)\doteq b\log b-b+1$. Then

$$\frac{1}{n}\sum_{i=0}^{n-1}E\left[1_{\{f_i^n(\bar{u}_i^{n,\eta^n})\geq r\}}\left\|\bar{u}_i^{n,\eta^n}\right\|\right]=\frac{1}{n}\sum_{i=0}^{n-1}E\left[\int_{\{u:f_i^n(u)\geq r\}}\|u\|\,f_i^n(u)\mu(du|\,\bar{X}_i^{n,\eta^n})\right]$$

and the bound $ab\leq e^a+\ell(b)$ for $a,b\geq 0$ with $a=\sigma\|u\|$ and $b=f_i^n(u)$ gives that for all i

$$\begin{aligned}&E\left[\int_{\{u:f_i^n(u)\geq r\}}\|u\|\,f_i^n(u)\mu(du|\,\bar{X}_i^{n,\eta^n})\right]\\&\quad\leq\frac{1}{\sigma}E\left[\int_{\{u:f_i^n(u)\geq r\}}e^{\sigma\|u\|}\mu(du|\,\bar{X}_i^{n,\eta^n})\right]+\frac{1}{\sigma}E\left[\int_{\{u:f_i^n(u)\geq r\}}\ell(f_i^n(u))\mu(du|\,\bar{X}_i^{n,\eta^n})\right].\end{aligned}$$

Since $\ell(b)\geq 0$ for all $b\geq 0$

$$\begin{aligned}E\left[\int_{\{u:f_i^n(u)\geq r\}}\ell\left(f_i^n(u)\right)\mu(du|\,\bar{X}_i^{n,\eta^n})\right]&\leq E\left[\int_{\mathbb{R}^d}\ell(f_i^n(u))\mu(du|\,\bar{X}_i^{n,\eta^n})\right]\\&=E[R(\eta_i^n\|\,\mu(\cdot|\,\bar{X}_i^{n,\eta^n}))],\end{aligned}$$

and by Hölder's inequality

$$\begin{aligned}&E\left[\int_{\{u:f_i^n(u)\geq r\}}e^{\sigma\|u\|}\mu(du|\,\bar{X}_i^{n,\eta^n})\right]\\&\quad\leq E\left[\left(\int_{\mathbb{R}^d}1_{\{u:f_i^n(u)\geq r\}}\mu(du|\,\bar{X}_i^{n,\eta^n})\right)^{\frac{1}{2}}\left(\int_{\mathbb{R}^d}e^{2\sigma\|u\|}\mu(du|\,\bar{X}_i^{n,\eta^n})\right)^{\frac{1}{2}}\right]\\&\quad=E\left[\mu(\{u:f_i^n(u)\geq r\}|\,\bar{X}_i^{n,\eta^n})^{\frac{1}{2}}\right]\left(2^de^{dK_{\mathrm{mgf}}}\right)^{\frac{1}{2}}.\end{aligned}$$

In addition Markov's inequality gives for $r\geq e^{-1}$

$$\mu(\{u:f_i^n(u)\geq r\}|\,\bar{X}_i^{n,\eta^n})\leq\frac{1}{r\log r}\int\log(f_i^n(u))f_i^n(u)\mu(du|\,\bar{X}_i^{n,\eta^n})=\frac{R(\eta_i^n\|\,\mu(\cdot|\,\bar{X}_i^{n,\eta^n}))}{r\log r}.$$

Therefore

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=0}^{n-1} E\left[\int_{\{u:f_i^n(u)\geq r\}} \|u\|\, f_i^n(u)\mu(\,du|\,\bar{X}_i^{n,\eta^n})\right] \\
&\quad\leq \frac{1}{\sigma}\left(2^d e^{dK_{\mathrm{mgf}}}\right)^{\frac{1}{2}} \frac{1}{n}\sum_{i=0}^{n-1} E\left[\left(\frac{R(\,\eta_i^n\|\,\mu(\cdot|\,\bar{X}_i^{n,\eta^n}))}{r\log r}\right)^{\frac{1}{2}}\right] + \frac{1}{\sigma}\frac{1}{n}\sum_{i=0}^{n-1} E[R(\,\eta_i^n\|\,\mu(\cdot|\,\bar{X}_i^{n,\eta^n}))].
\end{aligned}
$$

Since by Jensen's inequality

$$
\frac{1}{n}\sum_{i=0}^{n-1} E\left[\left(\frac{R(\,\eta_i^n\|\,\mu(\cdot|\,\bar{X}_i^{n,\eta^n}))}{r\log r}\right)^{\frac{1}{2}}\right] \leq \left(\frac{1}{r\log r}\right)^{\frac{1}{2}} \left(\frac{1}{n}\sum_{i=0}^{n-1} E[R(\,\eta_i^n\|\,\mu(\cdot|\,\bar{X}_i^{n,\eta^n}))]\right)^{\frac{1}{2}},
$$

we obtain the overall bound

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=0}^{n-1} E\left[1_{\{f_i^n(\bar{u}_i^n)\geq r\}}\left\|\bar{u}_i^{n,\eta^n}\right\|\right] \qquad\qquad (A.8)\\
&\quad\leq \frac{1}{\sigma}\left(2^d e^{dK_{\mathrm{mgf}}}\right)^{\frac{1}{2}} \left(\frac{1}{r\log r}\right)^{\frac{1}{2}} \left(\frac{1}{n}\sum_{i=0}^{n-1} E[R(\,\eta_i^n\|\,\mu(\cdot|\,\bar{X}_i^{n,\eta^n}))]\right)^{\frac{1}{2}} \\
&\qquad + \frac{1}{\sigma}\frac{1}{n}\sum_{i=0}^{n-1} E[R(\,\eta_i^n\|\,\mu(\cdot|\,\bar{X}_i^{n,\eta^n}))] \\
&\quad\leq \frac{1}{\sigma}\frac{K_E^{\frac{1}{2}}}{a(n)\sqrt{n}}\left(2^d e^{dK_{\mathrm{mgf}}}\right)^{\frac{1}{2}}\left(\frac{1}{r\log r}\right)^{\frac{1}{2}} + \frac{1}{\sigma}\frac{K_E}{a(n)^2 n}.
\end{aligned}
$$

Using this result we can complete the proof. Define

$$
\xi_i^{n,r} \doteq \begin{cases} \bar{u}_i^{n,\eta^n} & \text{if } f_i^n(\bar{u}_i^{n,\eta^n}) < r \\ 0 & \text{otherwise.} \end{cases}
$$

For any $\delta > 0$

$$
\begin{aligned}
&P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}a(n)\sqrt{n}\left(\bar{u}_i^{n,\eta^n}-w^{n,\eta^n}\left(\frac{i}{n}\right)\right)\right\|\geq 3\delta\right\}\\
&\quad\leq P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}a(n)\sqrt{n}(\bar{u}_i^{n,\eta^n}-\xi_i^{n,r})\right\|\geq\delta\right\}\\
&\qquad+P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}a(n)\sqrt{n}\left(\xi_i^{n,r}-\int_{\{u:f_i^n(u)<r\}}u\eta_i^n(du)\right)\right\|\geq\delta\right\}\\
&\qquad+P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}a(n)\sqrt{n}\left(w^{n,\eta^n}\left(\frac{i}{n}\right)-\int_{\{u:f_i^n(u)<r\}}u\eta_i^n(du)\right)\right\|\geq\delta\right\}.
\end{aligned}
$$

The first term satisfies

$$
\begin{aligned}
&P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}a(n)\sqrt{n}(\bar{u}_i^{n,\eta^n}-\xi_i^{n,r})\right\|\geq\delta\right\}\\
&\quad\leq\frac{1}{\delta}a(n)\sqrt{n}\frac{1}{n}\sum_{i=0}^{n-1}E\left[\left\|\bar{u}_i^{n,\eta^n}-\xi_i^{n,r}\right\|\right]\\
&\quad=\frac{1}{\delta}a(n)\sqrt{n}\frac{1}{n}\sum_{i=0}^{n-1}E\left[1_{\{f_i^n(\bar{u}_i^{n,\eta^n})\geq r\}}\left\|\bar{u}_i^{n,\eta^n}\right\|\right].
\end{aligned}
$$

The second term is a submartingale so by Doob's submartingale inequality

$$
\begin{aligned}
&P\left\{\max_{k=0,\ldots,n-1}\left\|\frac{1}{n}\sum_{i=0}^{k}a(n)\sqrt{n}\left(\xi_i^{n,r}-\int_{\{u:f_i^n(u)<r\}}u\eta_i^n(du)\right)\right\|\geq\delta\right\}\\
&\quad\leq\frac{1}{\delta^2}E\left[\left\|\frac{1}{n}\sum_{i=0}^{n-1}a(n)\sqrt{n}\left(\xi_i^{n,r}-\int_{\{u:f_i^n(u)<r\}}u\eta_i^n(du)\right)\right\|^2\right]\\
&\quad=\frac{1}{\delta^2}\frac{a(n)^2}{n}\sum_{i=0}^{n-1}E\left[\left\|\left(\xi_i^{n,k}-\int_{\{u:f_i^n(u)<r\}}u\eta_i^n(du)\right)\right\|^2\right]\\
&\quad\leq\frac{1}{\delta^2}\frac{a(n)^2}{n}\sum_{i=0}^{n-1}E\left[\left\|\xi_i^{n,k}\right\|^2\right]\\
&\quad=\frac{1}{\delta^2}\frac{a(n)^2}{n}\sum_{i=0}^{n-1}E\left[\int_{\{u:f_i^n(u)<r\}}\|u\|^2 f_i^n(u)\mu(du|\,\bar{X}_i^{n,\eta^n})\right]\\
&\quad\leq\frac{r}{\delta^2}\frac{a(n)^2}{n}\sum_{i=0}^{n-1}E\left[\int_{\mathbb{R}^d}\|u\|^2\mu(du|\,\bar{X}_i^{n,\eta^n})\right]\\
&\quad\leq\frac{r}{\delta^2}a(n)^2K_{\mu,2},
\end{aligned}
$$

where

$$K_{\mu,2} = \sup_{x\in\mathbb{R}^d} \int_{\mathbb{R}^d} \|u\|^2 \, \mu(\, du|\, x) < \infty,$$

and the finiteness is due to (2.3). We can use Jensen's inequality with the third term and get the same bound that was shown for the first. We have

$$\begin{aligned}
&P\left\{ \max_{k=0,\ldots,n-1} \left\| \frac{1}{n}\sum_{i=0}^{k} a(n)\sqrt{n}\left(w^{n,\eta^n}\left(\frac{i}{n}\right) - \int_{\{u:f_i^n(u)<r\}} u\eta_i^n(du)\right)\right\| \geq \delta \right\} \\
&\leq \frac{1}{\delta}a(n)\sqrt{n}\frac{1}{n}\sum_{i=0}^{n-1} E\left[\left\|\left(w^{n,\eta^n}\left(\frac{i}{n}\right) - \int_{\{u:f_i^n(u)<r\}} u\eta_i^n(du)\right)\right\|\right] \\
&= \frac{1}{\delta}a(n)\sqrt{n}\frac{1}{n}\sum_{i=0}^{n-1} E\left[\left\|\int_{\{u:f_i^n(u)\geq r\}} u\eta_i^n(du)\right\|\right] \\
&\leq \frac{1}{\delta}a(n)\sqrt{n}\frac{1}{n}\sum_{i=0}^{n-1} E\left[\int_{\{u:f_i^n(u)\geq r\}} \|u\|\, \eta_i^n(du)\right] \\
&= \frac{1}{\delta}a(n)\sqrt{n}\frac{1}{n}\sum_{i=0}^{n-1} E\left[1_{\{f_i^n(\bar{u}_i^{n,\eta^n})\geq r\}} \left\|\bar{u}_i^{n,\eta^n}\right\|\right].
\end{aligned}$$

Combining the bounds for these three terms with (A.8) gives

$$\begin{aligned}
&P\left\{ \max_{k=0,\ldots,n-1} \left\| \frac{1}{n}\sum_{i=0}^{k} a(n)\sqrt{n}\left(\bar{u}_i^{n,\eta^n} - w^{n,\eta^n}\left(\frac{i}{n}\right)\right)\right\| \geq 3\delta \right\} \\
&\leq \frac{2}{\delta}a(n)\sqrt{n}\frac{1}{n}\sum_{i=0}^{n-1} E\left[1_{\{f_i^n(\bar{u}_i^{n,\eta^n})\geq r]} \left\|\bar{u}_i^{n,\eta^n}\right\|\right] + \frac{r}{\delta^2}a(n)^2 K_{\mu,2} \\
&\leq \frac{2}{\sigma\delta}K_E^{\frac{1}{2}}\left(2^d e^{dK_{\mathrm{mgf}}}\right)^{\frac{1}{2}}\left(\frac{1}{r\log r}\right)^{\frac{1}{2}} + \frac{2}{\sigma\delta}\frac{K_E}{a(n)\sqrt{n}} + a(n)^2\frac{r}{\delta^2}K_{\mu,2}.
\end{aligned}$$

Choosing $r = 1/a(n)$ and using $a(n) \to 0, a(n)\sqrt{n} \to \infty$ gives

$$P\left\{ \max_{k=0,\ldots,n-1} \left\| \frac{1}{n}\sum_{i=0}^{k} a(n)\sqrt{n}\left(\bar{u}_i^{n,\eta^n} - w^{n,\eta^n}\left(\frac{i}{n}\right)\right)\right\| \geq 3\delta \right\} \to 0$$

as $n \to \infty$, which completes the proof. ■

This theorem, combined with the following discrete version of Gronwall's inequality, will allow us to prove $\bar{Y}^{n,\eta^n} - \check{Y}^{n,\eta^n} \to 0$.

Lemma A.7 *If $\{a_n\}$, $\{b_n\}$, and $\{c_n\}$ are nonnegative sequences defined for $n = 0, 1, \dots$ and satisfying*

$$a_n \leq c_n + \sum_{k=0}^{n-1} b_k a_k,$$

then

$$a_n \leq c_n + \sum_{k=0}^{n-1} b_k c_k \exp\left\{ \sum_{i=k+1}^{n-1} b_i \right\}.$$

Theorem A.8 *Under the conditions of Theorem A.6 $\check{Y}^{n,\eta^n} - \bar{Y}^{n,\eta^n} \to 0$ in probability.*

Proof. Recall that

$$\bar{Y}_k^{n,\eta^n} = \sum_{i=0}^{k-1} \frac{a(n)}{\sqrt{n}} \left(b\left(X_i^{n,0} + \frac{1}{a(n)\sqrt{n}} \bar{Y}_i^{n,\eta^n} \right) - b\left(X_i^{n,0} \right) \right) + \sum_{i=0}^{k-1} \frac{a(n)}{\sqrt{n}} \bar{u}_i^{n,\eta^n}$$

and

$$\check{Y}_k^{n,\eta^n} = \sum_{i=0}^{k-1} \frac{a(n)}{\sqrt{n}} \left(b\left(X_i^{n,0} + \frac{1}{a(n)\sqrt{n}} \check{Y}_i^{n,\eta^n} \right) - b\left(X_i^{n,0} \right) \right) + \sum_{i=0}^{k-1} \frac{a(n)}{\sqrt{n}} w^{n,\eta^n}\left(\frac{i}{n} \right),$$

so with W_k^n defined as in Theorem A.6

$$\left\| \bar{Y}_k^{n,\eta^n} - \check{Y}_k^{n,\eta^n} \right\| \leq \|W_k^n\| + \sum_{i=0}^{k-1} \frac{K_b}{n} \left\| \bar{Y}_i^{n,\eta^n} - \check{Y}_i^{n,\eta^n} \right\|.$$

Using Lemma A.7 gives

$$\begin{aligned}
\left\| \bar{Y}_k^{n,\eta^n} - \check{Y}_k^{n,\eta^n} \right\| &\leq \|W_k^n\| + \sum_{i=0}^{k-1} \|W_i^n\| \frac{K_b}{n} \exp\left\{ \frac{K_b}{n}(k-i-1) \right\} \\
&\leq (1 + K_b e^{K_b}) \max_{i \in \{1,\dots,k\}} \{\|W_i^n\|\}
\end{aligned}$$

so

$$\max_{i \in \{1,\dots,n\}} \left\{ \left\| \bar{Y}_i^{n,\eta^n} - \check{Y}_i^{n,\eta^n} \right\| \right\} \leq (1 + K_b e^{K_b}) \max_{i \in \{1,\dots,n\}} \{\|W_i^n\|\}.$$

Since $\max_{i \in \{1,\dots,n\}} \{\|W_i^n\|\} \to 0$ in probability

$$\max_{i \in \{1,\dots,n\}} \left\{ \left\| \bar{Y}_i^{n,\eta^n} - \check{Y}_i^{n,\eta^n} \right\| \right\} \to 0 \text{ and hence } \sup_{t \in [0,1]} \left\| \bar{Y}^{n,\eta^n}(t) - \check{Y}^{n,\eta^n}(t) \right\| \to 0$$

in probability. ■

To complete the proof of the convergence we need to show $\check{Y}^{n,\eta^n} - \hat{Y}^{n,\eta^n} \to 0$. Recall that these two processes have the same driving terms but different drifts, in that $\hat{Y}^{n,\eta^n}$ satisfies the ODE

$$\hat{Y}^{n,\eta^n}(t) = \int_0^t Db(X^0(s))\hat{Y}^{n,\eta^n}(s)ds + \int_0^t \hat{w}^{n,\eta^n}(s)ds$$

while $\check{Y}^{n,\eta^n}$ is the linear interpolation of the discrete time process defined by $\check{Y}_0^{n,\eta^n} = 0$ and

$$\check{Y}_{i+1}^{n,\eta^n} = \check{Y}_i^{n,\eta^n} + \frac{a(n)}{\sqrt{n}}\left(b\left(X_i^{n,0} + \frac{1}{a(n)\sqrt{n}}\check{Y}_i^{n,\eta^n}\right) - b\left(X_i^{n,0}\right)\right) + \frac{1}{n}\hat{w}^{n,\eta^n}\left(\frac{i}{n}\right).$$

However, essentially the same arguments as those used in Lemma A.4 to show tightness of $\{\hat{Y}^{n,\eta^n}\}$ can be used to prove tightness of $\{\check{Y}^{n,\eta^n}\}$, and then it easily follows as in Lemma A.5 that any limit will satisfy the same ODE (2.13) as the limit of $\{\hat{Y}^{n,\eta^n}\}$, and therefore $\check{Y}^{n,\eta^n} - \hat{Y}^{n,\eta^n} \to 0$ follows.

Combining $\bar{Y}^{n,\eta^n} - \check{Y}^{n,\eta^n} \to 0$, $\check{Y}^{n,\eta^n} - \hat{Y}^{n,\eta^n} \to 0$, and $\hat{Y}^{n,\eta^n} \to \hat{Y}$ demonstrates that along the subsequence where $\varsigma^{n,\eta^n} \to \varsigma$ weakly $\bar{Y}^{n,\eta^n} \to \hat{Y}$ in distribution, which implies that along this subsequence $(\varsigma^{n,\eta^n}, \bar{Y}^{n,\eta^n}) \to (\varsigma, \hat{Y})$ weakly. We have already shown that with probability 1 $\varsigma_2(dt)$ is Lebesgue measure and

$$\hat{Y}(t) = \int_0^t Db(X^0(s))\hat{Y}(s)ds + \int_0^t \int_{\mathbb{R}^d} w\varsigma_{1|2}(dw\,|s)ds,$$

so the proof of convergence (i.e., the first part of Theorem 2.5) is complete.

To finish Theorem 2.5 we must lastly show the bound (2.14). Note that the weak convergence of $\bar{Y}^{n,\eta^n}$ implies

$$\sup_{t\in[0,1]} \left\|\bar{X}^{n,\eta^n}(\lfloor nt\rfloor/n) - X^0(t)\right\| \to 0 \text{ in probability.} \tag{A.9}$$

Now define random measures on $\mathbb{R}^d \otimes \mathbb{R}^d \otimes [0,1]$ by

$$\gamma^n(dx \otimes dw \otimes dt) = \delta_{\bar{X}^{n,\eta^n}(\lfloor nt\rfloor/n)}(dx)\,\varsigma^{n,\eta^n}(dw \otimes dt).$$

Note that the tightness of $\{\gamma^n\}$ follows easily from (A.9) and from the tightness of $\{\varsigma^{n,\eta^n}\}$. Thus given any subsequence we can choose a further subsequence (again we will retain n as the index for simplicity) along which $\{\gamma^n\}$ converges weakly to some limit γ on $\mathcal{P}\left(\mathbb{R}^d \otimes \mathbb{R}^d \otimes [0,1]\right)$ with

$$\gamma_{2,3}\left(dw \otimes dt\right) = \varsigma\left(dw \otimes dt\right),$$

where $\gamma_{2,3}$ is the second and third marginal of γ. If we establish (2.14) for this subsequence it follows for the original one using a standard argument by contradiction. For $\sigma > 0$ let

$$G_\sigma^{X^0} = \left\{(x,w,t) : \left\|x - X^0\left(t\right)\right\| \leq \sigma\right\}$$

be closed sets centered around $X^0\left(t\right)$ in the x variable, and note that by (A.9) and weak convergence, for all $\sigma > 0$

$$1 = \limsup_{n\to\infty} E\left[\gamma^n\left(G_\sigma^{X^0}\right)\right] \leq E\left[\gamma\left(G_\sigma^{X^0}\right)\right].$$

Thus

$$E\left[\gamma\left(\cap_{n\in\mathbb{N}} G_{1/n}^{X^0}\right)\right] = 1,$$

so with probability 1 γ puts all its mass on $\left\{(x,w,t) : x = X^0\left(t\right)\right\}$. Therefore with probability 1, for a.e. (w,t) under $\gamma_{2,3}\left(dw \otimes dt\right)$,

$$\gamma_{1|2,3}\left(dx|\, w,t\right) = \delta_{X^0(t)}\left(dx\right).$$

Combined with the fact that the second marginal of $\varsigma\left(dw \otimes dt\right)$ is Lebesgue measure, this gives

$$\gamma\left(dx \otimes dw \otimes dt\right) = \delta_{X^0(t)}\left(dx\right)\varsigma\left(dw|\, t\right)dt. \tag{A.10}$$

Let

$$\bar{L}_K\left(x,\beta\right) = \sup_{\alpha\in\mathbb{R}^d}\left\{\langle\alpha,\beta\rangle - \frac{1}{2}\left\|\alpha\right\|_{A(x)}^2 - \frac{1}{2K}\left\|\alpha\right\|^2\right\}.$$

Then uniformly in x and compact subsets of β

$$\liminf_{n\to\infty} a(n)^2 n L_c\left(x, \frac{1}{a(n)\sqrt{n}}\beta\right) \geq \bar{L}_K\left(x,\beta\right),$$

and as $K \to \infty$

$$\bar{L}_K\left(x,\beta\right) \uparrow \frac{1}{2}\left\|\beta\right\|^2_{A^{-1}(x)}$$

for all $(x,\beta) \in \mathbb{R}^{2d}$. Combining this with Lemma A.1 and using Fatou's lemma for weak convergence,

$$\begin{aligned}
&\liminf_{n\to\infty} a(n)^2 nE\left[\sum_{i=0}^{n-1}\frac{1}{n}R(\eta_i^n \| \, \mu(\cdot|\,\bar{X}_i^{n,\eta^n}))\right] \\
&\geq \liminf_{n\to\infty} E\left[\int_{\mathbb{R}^d\otimes\mathbb{R}^d\otimes[0,1]} a(n)^2 nL_c\left(x, \frac{1}{a(n)\sqrt{n}}w\right)\gamma^n\left(dx\otimes dw\otimes dt\right)\right] \\
&\geq E\left[\int_{\mathbb{R}^d\otimes\mathbb{R}^d\otimes[0,1]} \bar{L}_K\left(x,w\right)\gamma\left(dx\otimes dw\otimes dt\right)\right]
\end{aligned}$$

for all K. Then using the monotone convergence theorem, the decomposition (A.10), and Jensen's inequality in that order shows that

$$\begin{aligned}
&\liminf_{n\to\infty} a(n)^2 nE\left[\sum_{i=0}^{n-1}\frac{1}{n}R(\eta_i^n \| \, \mu(\cdot|\,\bar{X}_i^{n,\eta^n}))\right] \\
&\geq \lim_{K\to\infty} E\left[\int_{\mathbb{R}^d\otimes\mathbb{R}^d\otimes[0,1]} \bar{L}_K\left(x,w\right)\gamma\left(dx\otimes dw\otimes dt\right)\right] \\
&= E\left[\int_{\mathbb{R}^d\otimes\mathbb{R}^d\otimes[0,1]} \frac{1}{2}\left\|w\right\|^2_{A^{-1}(x)}\gamma\left(dx\otimes dw\otimes dt\right)\right] \\
&= E\left[\int_0^1\int_{\mathbb{R}^d}\frac{1}{2}\left\|w\right\|^2_{A^{-1}(X^0(t))}\varsigma\left(dw|\,t\right)dt\right] \\
&\geq E\left[\frac{1}{2}\int_0^1\left\|\hat{w}(t)\right\|^2_{A^{-1}(X^0(t))}dt\right],
\end{aligned}$$

which is (2.14).

B TERMINAL POSITION COST PROOF

Here we provide the proof of Theorem 6.5. The goal is to show that the minimal running cost in (4.2) to go from (y,t) to (z,r) is given by $C(y,t,z,r)$. This would be a simple verification argument except for the fact due to the possible degeneracy of the noise certain locations (z,r) may not be attainable from certain (y,t). Let

$$\bar{C}(y,t,z,r) = \inf\left\{\frac{1}{2}\int_t^r \|u(q)\|^2\,dq : \phi(r) = z,\ \phi(s) = y + \int_t^s Db(X^0(q))\phi(q)ds \right.$$
$$\left. + \int_t^s A^{1/2}(X^0(q))u(q)dq, s \in [t,r]\right\}$$

and note that by (4.2)

$$V(y,t) = \inf_{z\in\mathbb{R}^d}\{\bar{C}(y,t,z,T) + F(z)\}$$

for all $y \in \mathbb{R}^d$ and $t \in [0,T]$. However if

$$\phi(s) = y + \int_t^s Db(X^0(q))\phi(q)dq + \int_t^s A^{1/2}(X^0(q))u(q)dq$$

for $s \in [t,r]$ then using Definition 6.1 it is clear that

$$\phi(s) = \exp\{\Omega(t,s)\}y + \int_t^s \exp\{\Omega(q,s)\}A^{1/2}(X^0(q))u(q)dq \tag{B.1}$$

for $s \in [t,r]$. Consequently we can write

$$\bar{C}(y,t,z,r) = \inf\left\{\frac{1}{2}\int_t^r \|u(q)\|^2 \, dq : z - \exp\{\Omega(t,r)\}y = \int_t^r \exp\{\Omega(q,r)\}A^{1/2}(X^0(q))u(q)dq\right\}.$$

In order to prove that

$$\bar{C}(y,t,z,r) = C(y,t,z,r)$$

for all $y, z \in \mathbb{R}^d$ and $0 \le t < r \le T$ we introduce the following functions. Let

$$A^\epsilon(x) = A(x) + \epsilon I.$$

and for any $0 \le t < r \le T$ define

$$B(t,r) = \int_t^r \exp\left\{\Omega(q,r)\right\} A(X^0(q)) \exp\left\{\Omega(q,r)^T\right\} dq,$$

$$B^\epsilon(t,r) = \int_t^r \exp\left\{\Omega(q,r)\right\} A^\epsilon(X^0(q)) \exp\left\{\Omega(q,r)^T\right\} dq, \tag{B.2}$$

and

$$D(t,r) = \int_t^r \exp\left\{\Omega(q,r)\right\} \exp\left\{\Omega(q,r)^T\right\} dq$$

and note that

$$B^\epsilon(t,r) = B(t,r) + \epsilon D(t,r).$$

In addition, for any $w \in \mathbb{R}^d$ and $0 \le t < r \le T$ define (using Definition 2.2)

$$R(w,t,r) = \frac{1}{2} w^T B(t,r)^{-1} w, \tag{B.3}$$

$$R^\epsilon(w,t,r) = \frac{1}{2} w^T B^\epsilon(t,r)^{-1} w, \tag{B.4}$$

and

$$\bar{R}(w,t,r) = \inf\left\{\frac{1}{2}\int_t^r \|u(q)\|^2 \, dq : w = \int_t^r \exp\{\Omega(q,r)\}A^{1/2}(X^0(q))u(q)dq\right\}. \tag{B.5}$$

Note that for all $y, z \in \mathbb{R}^d$ and $0 \le t < r \le T$

$$C(y, t, z, r) = R(z - \exp\{\Omega(t, r)\}y, t, r)$$

and

$$\bar{C}(y, t, z, r) = \bar{R}(z - \exp\{\Omega(t, r)\}y, t, r)$$

so the result of Theorem 6.5 will follow if we can show that for any $w \in \mathbb{R}^d$ and $0 \le t < r \le T$

$$R(w, t, r) = \bar{R}(w, t, r).$$

We first prove the following.

Theorem B.1 *Let $R(w, t, r)$, $R^\epsilon(w, t, r)$ and $\bar{R}(w, t, r)$ be defined by (B.3), (B.4) and (B.5). Then for all $w \in \mathbb{R}^d$ and $0 \le t < r \le T$*

$$R^\epsilon(w, t, r) \le \bar{R}(w, t, r) \le R(w, t, r).$$

Proof. Recall the definition of $B^\epsilon(t, r)$ in (B.2). Because $B^\epsilon(t, r)$ is positive definite and hence invertible we have

$$\begin{aligned} R^\epsilon_r(w, t, r) &= \frac{1}{2} w^T \frac{d}{dr}(B^\epsilon(t, r)^{-1})w \\ &= -\frac{1}{2} w^T B^\epsilon(t, r)^{-1}(B^\epsilon_r(t, r))B^\epsilon(t, r)^{-1} w \end{aligned}$$

and

$$B^\epsilon_r(t, r) = A^\epsilon(X^0(r)) + Db(X^0(r))B^\epsilon(t, r) + B^\epsilon(t, r)(Db(X^0(r)))^T$$

which gives

$$R^\epsilon_r(w, t, r) = -\frac{1}{2} w^T B^\epsilon(t, r)^{-1} A^\epsilon(X^0(r)) B^\epsilon(t, r)^{-1} w - w^T B^\epsilon(t, r)^{-1} Db(X^0(r)) w$$

In addition

$$DR^{\epsilon}(w,t,r) = B^{\epsilon}(t,r)^{-1}w.$$

Given any $w \in \mathbb{R}^d$ and $0 \leq t < r \leq T$ such that $\bar{R}(w,t,r) < \infty$ we can choose for any $\delta > 0$ a δ-near minimizing (u, ϕ_u) in (B.5), meaning that

$$\bar{R}(w,t,r) + \delta \geq \frac{1}{2}\int_t^r \|u(q)\|^2 \, dr,$$

$$w = \phi_u(r) = \int_t^r \exp\{\Omega(q,r)\} A^{1/2}(X^0(q))u(q)dq,$$

and

$$\phi_u(s) = \int_t^s \exp\{\Omega(q,s)\} A^{1/2}(X^0(q))u(q)dq$$

for $s \in [t,r]$. Because

$$\frac{1}{2}\int_t^r \|u(q)\|^2 \, dr < \infty$$

it follows that

$$\lim_{s \downarrow t} R^{\epsilon}\left(\phi_u(s), t, s\right) = 0.$$

Consequently

$$\begin{aligned}
R^{\epsilon}(w,t,r) &= R^{\epsilon}\left(\phi_u(r), t, r\right) \\
&= \int_t^r \frac{d}{ds}(R^{\epsilon}\left(\phi_u(s), t, s\right))ds \\
&= \int_t^r DR^{\epsilon}\left(\phi_u(s), t, s\right) A^{1/2}(X^0(s))u(s)ds \\
&\quad + \int_t^r DR^{\epsilon}\left(\phi_u(s), t, s\right) Db(X^0(s))\phi_u(s)ds \\
&\quad + \int_t^r R^{\epsilon}_s\left(\phi_u(s), t, s\right) ds \\
&= \int_t^r \phi_u(s)^T B^{\epsilon}(t,s)^{-1} A^{1/2}(X^0(s))u(s)ds \\
&\quad - \int_t^r \frac{1}{2}\phi_u(s)^T B^{\epsilon}(t,s)^{-1} A(X^0(s)) B^{\epsilon}(t,s)^{-1}\phi_u(s)ds \\
&\quad - \epsilon \int_t^r \frac{1}{2}\phi_u(s)^T B^{\epsilon}(t,r)^{-2}\phi_u(s)ds \\
&\leq \frac{1}{2}\int_t^r \|u(s)\|^2 \, ds.
\end{aligned}$$

Because δ was arbitrary we have for all $\epsilon > 0$, $w \in \mathbb{R}^d$ and $0 \le t < r \le T$

$$R^\epsilon(w,t,r) \le \bar{R}(w,t,r).$$

In addition, for any $w \in \mathbb{R}^d$ and $0 \le t < r \le T$ such that $R(w,t,r) < \infty$ let

$$u(s) = A^{\frac{1}{2}}(X^0(s)) \exp\left\{\Omega(s,r)^T\right\} \left(\int_t^r \exp\left\{\Omega(q,r)\right\} A(X^0(q)) \exp\left\{\Omega(q,r)^T\right\} dq\right)^{-1} w$$

and

$$\phi_u(s) = \int_t^s \exp\{\Omega(q,s)\} A^{1/2}(X^0(q)) u(q) dq$$

for $s \in [t,r]$. Then using (B.1) we have

$$\phi_u(r) = w$$

and clearly

$$\begin{aligned}\frac{1}{2}\int_t^r \|u(s)\|^2 \, ds &= \frac{1}{2} w^T \left(\int_t^r \exp\left\{\Omega(q,r)\right\} A(X^0(q)) \exp\left\{\Omega(q,r)^T\right\} ds\right)^{-1} w \\ &= R(w,t,r).\end{aligned}$$

Therefore for all $\epsilon > 0$, $w \in \mathbb{R}^d$ and $0 \le t < r \le T$ we have

$$R^\epsilon(w,t,r) \le \bar{R}(w,t,r) \le R(w,t,r).$$

■

This combined with the following theorem shows that for all $w \in \mathbb{R}^d$ and $0 \le t < r \le T$

$$\bar{R}(w,t,r) = R(w,t,r).$$

Theorem B.2 *Let $R^\epsilon(w,t,r)$ and $R(w,t,r)$ be defined by (B.4) and (B.3). Then for all $w \in \mathbb{R}^d$ and $0 \le t < r \le T$*

$$\lim_{\epsilon \to 0} R^\epsilon(w,t,r) = R(w,t,r).$$

Proof. Recall that

$$R(w,t,r) = \frac{1}{2} w^T (B(t,r))^{-1} w$$

and

$$R^\epsilon(w,t,r) = \frac{1}{2} w^T (B(t,r) + \epsilon D(t,r))^{-1} w$$

where $B(t,r)$ is symmetric and nonnegative definite and $D(t,r)$ is symmetric and positive definite with $\|D\| = \sup_{0 \le t < r \le T} \|D(t,r)\| < \infty$. Assume $\|w\| = 1$ and $B(t,r)w = 0$. Then

$$\begin{aligned}
1 &= \left\| w^T (B(t,r) + \epsilon D(t,r))(B(t,r) + \epsilon D(t,r))^{-1} w \right\| \\
&= \left\| w^T \epsilon D(t,r)(B(t,r) + \epsilon D(t,r))^{-1} w \right\| \\
&\le \epsilon \|D\| \left\| (B(t,r) + \epsilon D(t,r))^{-1} w \right\|
\end{aligned}$$

so

$$\left\| (B(t,r) + \epsilon D(t,r))^{-1} w \right\| \ge 1/\epsilon \|D\| .$$

Let k be the smallest eigenvalue of $D(t,r)$. Then

$$\frac{1}{\epsilon k} w^T (B(t,r) + \epsilon D(t,r))^{-1} w \ge \left\| (B(t,r) + \epsilon D(t,r))^{-1} w \right\|^2$$

so

$$w^T (B(t,r) + \epsilon D(t,r))^{-1} w \ge k/\epsilon \|D\|^2$$

and $w^T (B(t,r) + \epsilon D(t,r))^{-1} w \to \infty$ as $\epsilon \to 0$. For $w \in \mathbb{R}^d$ and $0 \le t < r \le T$ such that w is not orthogonal to the null space of $B(t,r)$ we can write $w = w_p + w_0$ where w_p is orthogonal to the null space and w_0 is in the null space. Then

$$\begin{aligned}
(w_0^T (B(t,r) + \epsilon D(t,r))^{-1} w_0)^{1/2} &\le (w_p^T (B(t,r) + \epsilon D(t,r))^{-1} w_p)^{1/2} \\
&\quad + (w^T (B(t,r) + \epsilon D(t,r))^{-1} w)^{1/2}
\end{aligned}$$

so

$$
\begin{aligned}
(w^T(B(t,r)+\epsilon D(t,r))^{-1}w)^{1/2} &\geq (w_0^T(B(t,r)+\epsilon D(t,r))^{-1}w_0)^{1/2} \\
&\quad - (w_p^T(B(t,r)+\epsilon D(t,r))^{-1}w_p)^{1/2} \\
&\geq (w_0^T(B(t,r)+\epsilon D(t,r))^{-1}w_0)^{1/2} \\
&\quad - (w_p^T(B(t,r))^{-1}w_p)^{1/2}
\end{aligned}
$$

where the last inequality comes from Theorem B.1. However,

$$w_p^T(B(t,r))^{-1}w_p < \infty$$

so

$$\lim_{\epsilon \to 0} w^T(B(t,r)+\epsilon D(t,r))^{-1}w = \infty.$$

For an arbitrary $w \in \mathbb{R}^d$ and $0 \leq t < r \leq T$ where w is orthogonal to the null space of $B(t,r)$ we can write $w = \sum_{i=1}^{I} c_i e_i$ where $\{e_i\}_{i=1}^{I}$ are orthonormal eigenvectors of $B(t,r)$ with positive eigenvalues $\{\lambda_i\}_{i=1}^{I}$. Again, letting k be the smallest eigenvalue of $D(t,r)$ we have

$$
\begin{aligned}
\left\|(B(t,r)+\epsilon D(t,r))^{-1}e_i\right\|^2 &\leq \frac{1}{\epsilon k} e_i^T(B(t,r)+\epsilon D(t,r))^{-1}e_i \\
&\leq \frac{1}{\epsilon k} e_i^T B(t,r)^{-1}e_i \\
&= \frac{1}{\epsilon k \lambda_i}
\end{aligned}
$$

where the second inequality comes from Theorem B.1. Then

$$\left\|(B(t,r)+\epsilon D(t,r))^{-1}e_i\right\| \leq \frac{1}{\sqrt{\epsilon k \lambda_i}}$$

for all $i \in \{1, ..., I\}$. Consequently

$$\begin{aligned}
\left|\frac{1}{\lambda_i} - e_i^T(B(t,r)+\epsilon D(t,r))^{-1}e_i\right| &= \frac{\epsilon}{\lambda_i}\left|e_i^T D(t,r)(B(t,r)+\epsilon D(t,r))^{-1}e_i\right| \\
&\leq \frac{\epsilon}{\lambda_i}\|D\|\left\|(B(t,r)+\epsilon D(t,r))^{-1}e_i\right\| \\
&\leq \sqrt{\epsilon}k^{-1/2}\lambda_i^{-3/2}\|D\|
\end{aligned}$$

and

$$\lim_{\epsilon\to 0}\left|\frac{1}{\lambda_i} - e_i^T(B(t,r)+\epsilon D(t,r))^{-1}e_i\right| = 0$$

for all $i \in \{1, ..., I\}$. In addition, for $i, j \in \{1, ..., I\}$ with $i \neq j$

$$\begin{aligned}
0 &= e_i^T I e_j \\
&= e_i^T(B(t,r)+\epsilon D(t,r))(B(t,r)+\epsilon D(t,r))^{-1}e_j \\
&= \lambda_i e_i^T(B(t,r)+\epsilon D(t,r))^{-1}e_j - \epsilon e_i^T D(t,r)(B(t,r)+\epsilon D(t,r))^{-1}e_j
\end{aligned}$$

so

$$\begin{aligned}
\left|e_i^T(B(t,r)+\epsilon D(t,r))^{-1}e_j\right| &\leq \epsilon\lambda_i^{-1}\|D\|\left\|(B(t,r)+\epsilon D(t,r))^{-1}e_j\right\| \\
&\leq \sqrt{\epsilon}\lambda_i^{-1}\lambda_j^{-1/2}k^{-1/2}\|D\|
\end{aligned}$$

and

$$\lim_{\epsilon\to 0}\left|e_i^T(B(t,r)+\epsilon D(t,r))^{-1}e_j\right| = 0.$$

Therefore

$$
\begin{aligned}
\lim_{\epsilon \to 0} w^T (B(t,r) + \epsilon D(t,r))^{-1} w &= \lim_{\epsilon \to 0} \left(\sum_{i=1}^{I} c_i e_i \right)^T (B(t,r) + \epsilon D(t,r))^{-1} \left(\sum_{i=1}^{I} c_i e_i \right) \\
&= \lim_{\epsilon \to 0} \sum_{i,j=1}^{I} c_i c_j e_i^T (B(t,r) + \epsilon D(t,r))^{-1} e_j \\
&= \sum_{i,j=1}^{I} c_i c_j \lim_{\epsilon \to 0} e_i^T (B(t,r) + \epsilon D(t,r))^{-1} e_j \\
&= \sum_{i=1}^{I} c_i^2 \lambda_i^{-1} \\
&= w^T (B(t,r))^{-1} w.
\end{aligned}
$$

Therefore for all $w \in \mathbb{R}^d$ and $0 \leq t < r \leq T$

$$
\lim_{\epsilon \to \infty} R^{\epsilon}(w,t,r) = R(w,t,r)
$$

and the proof is complete. ■

References

[1] R. Azencott and G. Ruget. Mélanges d'équations differentielles et grand écart à la loi des grandes nombres. *Z. Wahrsch. Verw. Gebiete*, 38:1–54, 1977.

[2] P. Baldi. Large deviations and stochastic homogenization. *Ann. Mat. Pura Appl.*, 151:161–177, 1988.

[3] J.H. Blanchet, P. Glynn, and K. Leder. On Lyapunov inequalities and subsolutions for efficient importance sampling. *TOMACS*, page to appear, 2012.

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

[5] M. Boué and P. Dupuis. A variational representation for certain functionals of Brownian motion. *The Annals of Prob.*, 26:1641–1659, 1998.

[6] A. Budhiraja, P. Dupuis, and V. Maroulas. Variational representations for continuous time processes. *Ann. de l'Inst. H. Poincaré, volume =.*

[7] X. Chen and A. de Acosta. Moderate deviations for empirical measures of Markov chains: Upper bounds. *J. Theor. Probab.*, 11:1075–1110, 1998.

[8] A. de Acosta. Moderate deviations for empirical measures of Markov chains: Lower bounds. *Ann. Probab.*, 25:259–284, 1997.

[9] A. Dembo. Moderate deviations for martingales with bounded jumps. *Elect. Comm. Probab.*, 1:11–17, 1996.

[10] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, Boston, 1993.

[11] H. Djellout. Moderate deviations for martingale differences and applications to ϕ-mixing sequences. *Stoch. Stoch. Rep.*, 73:37–63, 2002.

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

[13] P. Dupuis, K. Leder, and H. Wang. Large deviations and importance sampling for a tandem network with slow-down. *QUESTA*, 57:71–83, 2007.

[14] P. Dupuis, K. Leder, and H. Wang. On the large deviations properties of the weighted-serve-the-longest-queue policy. In V. Sidoravicius and M.E. Vares, editors, *In and Out of Equilibrium 2*. Birkhauser, New York, 2008.

[15] P. Dupuis, A. Sezer, and H. Wang. Dynamic importance sampling for queueing networks. *Ann. Appl. Prob.*, pages 1306–1346, 2007.

[16] P. Dupuis, K. Spiliopoulos, and H. Wang. Importance sampling for multiscale diffusions. *SIAM J. Multiscale Model. and Simul.*, 10:1–27, 2012.

[17] P. Dupuis, K. Spiliopoulos, and X. Zhou. Escaping from an attractor: Importance sampling and rest points. *preprint*, 2012.

[18] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. *Stoch. and Stoch. Reports.*, 76:481–508, 2004.

[19] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.*, 32:1–35, 2007.

[20] P. Dupuis and H. Wang. Importance sampling for Jackson networks. *Queueing Systems*, 62:113–157, 2009.

[21] M. I. Freidlin and A. D. Wentzell. *Random Perturbations of Dynamical Systems.* Springer-Verlag, New York, 1984.

[22] F.Q. Gao. Moderate deviations for martingales and mixing random processes. *Stoch. Proc. Appl.*, 61:263–275, 1996.

[23] J. Gärtner. On large deviations from the invariant measure. *Theory Probab. Appl.*, 22:24–39, 1977.

[24] P. Glasserman and Y. Wang. Counter examples in importance sampling for large deviations probabilities. *Ann. Appl. Prob.*, 7:731–746, 1997.

[25] A. Guillin. Moderate deviations of inhomogeneous functionals of Markov processes and application to averaging. *Stoch. Proc. Appl.*, 92:287–313, 2001.

[26] A. Guillin. Averaging principle of SDE with small diffusion: Moderate deviations. *Ann. Probab.*, 31:413–443, 2003.

[27] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, New York, New York, 1991.

[28] D. Siegmund. Importance sampling in the monte carlo study of sequential tests. *Ann. Stat.*, 4:673–684, 1976.

[29] A.D. Wentzell. Rough limit theorems on large deviations for Markov stochastic processes, I. *Theory Probab. Appl.*, 21:227–242, 1976.

[30] A.D. Wentzell. Rough limit theorems on large deviations for Markov stochastic processes, II. *Theory Probab. Appl.*, 21:499–512, 1976.

[31] A.D. Wentzell. Rough limit theorems on large deviations for Markov stochastic processes, III. *Theory Probab. Appl.*, 24:675–692, 1979.

[32] A.D. Wentzell. Rough limit theorems on large deviations for Markov stochastic processes, IV. *Theory Probab. Appl.*, 27:215–234, 1982.