

Subsolutions of an Isaacs equation and efficient schemes for importance sampling: Convergence analysis

Paul Dupuis* and Hui Wang†
Lefschetz Center for Dynamical Systems
Brown University
Providence, R.I. 02912
USA

August 8, 2005

Abstract

The papers [2, 3] establish the connection between importance sampling algorithms for estimating rare-event probabilities, two-person zero-sum differential games, and the associated Isaacs equation. In order to construct nearly optimal schemes in a general setting, one must consider *dynamic* schemes, i.e., changes of measure that, in the course of a single simulation, can depend on the outcome of the simulation up till that time. The present paper and a companion paper [4] show that classical sense *subsolutions* of the Isaacs equation provide a basic and flexible tool for the construction and analysis of nearly optimal schemes. Asymptotic analysis is the topic of the present paper, while [4] focuses on explicit methods for the construction of subsolutions, implementation aspects and numerical results.

*Research of this author supported in part by the National Science Foundation (NSF-DMS-0306070 and NSF-DMS-0404806) and the Army Research Office (DAAD19-02-1-0425 and W911NF-05-1-0289).

†Research of this author supported in part by the National Science Foundation (NSF-DMS-0103669 and NSF-DMS-0404806).

1 Introduction

In a pair of recent papers [2, 3], we discuss how one can characterize the optimal achievable performance of importance sampling schemes in the large deviation limit in terms of a deterministic differential game. The value function of the game can, in turn, be characterized as the solution to a certain nonlinear partial differential equation (PDE) known as an Isaacs equation. Asymptotically optimal importance sampling schemes are then constructed based on this solution.

The purpose of the present paper and a companion paper is to explore this connection in further depth. More precisely, we show how one can construct importance sampling schemes based on *subsolutions* of the Isaacs equation. Since a *solution* is always a subsolution, this leads to a more general class of schemes. The main result of the paper is a basic result on the asymptotic performance of importance sampling schemes that are based on a given subsolution. The performance is in fact characterized by the value of the subsolution at a particular point. The proof is carried out in a general setting that contains as special cases sums of independent identically distributed (iid) random variables and the empirical measure of a finite-state Markov chain. However, its potential application is much broader, and includes systems with state dependencies and small noise effects, solutions to stochastic differential equations, systems with constrained dynamics (e.g., queuing networks), and to different forms of the expected value (e.g., probabilities of path dependent events). Some of these developments will be reported elsewhere.

One is often interested in properties other than just asymptotic optimality (e.g., ease of construction, ease of implementation). It turns out that one can often construct subsolutions that have a much simpler structure than the actual solution, and which induce schemes that are asymptotically optimal. This is important since the simplicity of the subsolution is usually reflected in the schemes they generate. It is therefore important to develop flexible techniques for the construction of subsolutions. That is the topic of the companion paper [4], which also presents some numerical results for the broader class of applications mentioned in the last paragraph.

The paper is organized as follows. Since the underlying game and Isaacs equation are not yet widely exposed in the importance sampling literature, we give some heuristics and a formal overview in Section 2 in the setting of sums of iid random variables. In particular, we formally derive the Isaacs equation, and indicate why subsolutions to this equation both suggest schemes and serve as a basic tool in their analysis. In Section 3

the general model and assumptions are stated. Importance sampling for Markov chains uses a collection of eigenfunctions that are related to the transition kernel of the chain, and Section 4 reviews the properties of these eigenfunctions. Section 5 identifies the Isaacs equation appropriate for the class of importance sampling problems introduced in Section 3, and Section 6 constructs the importance sampling schemes that are associated with particular subsolutions. The main result of the paper analyzes the asymptotic variance of a scheme associated with a given subsolution, and characterizes the performance of the scheme in terms of the value of the subsolution at a particular point. This result is stated and proved in Section 7. Finally, a tightness result needed for the asymptotic analysis is proved in the appendix.

Notation. For a Polish space S , $\mathcal{P}(S)$ denotes the collection of all probability measures on $(S, \mathcal{B}(S))$, where $\mathcal{B}(S)$ is the Borel σ -algebra. There will be many instances in this paper where we decompose measures on a product space as the product of a marginal distribution and a stochastic kernel. The following notation will be used. Suppose that $\mu \in \mathcal{P}(S_1 \times S_2)$ (with each S_i a Polish space) is such a probability measure. Then $[\mu]_1$ will denote the first marginal of μ , and $\mu(dy_2|y_1)$ will denote the stochastic kernel on S_2 given S_1 such that $\mu(dy_1 \times dy_2) = [\mu]_1(dy_1)\mu(dy_2|y_1)$. Quantities such as $[\mu]_2$, $\mu(dy_1|y_2)$, and the extension to products of more than two Polish spaces are all defined in the analogous fashion. Given $\mu \in \mathcal{P}(S_1)$ and a stochastic kernel q on S_2 given S_1 , we let $\mu \otimes q$ denote $\mu(dy_1)q(dy_2|y_1) \in \mathcal{P}(S_1 \times S_2)$.

2 An Introduction to the Role of Subsolutions

This section describes how an Isaacs equation arises in importance sampling, how subsolutions to that equation can be constructed, how they induce importance sampling schemes, and the implications for performance of the schemes. Since it is an overview, we will not give all details and will not be precise regarding all necessary assumptions. The overview is provided in the simplest possible setting: sums of iid random variables. The rest of the paper and the companion paper will consider more elaborate models.

2.1 Problem formulation for sums of iid random variables

Consider

$$X_n \doteq \frac{1}{n} \sum_{i=1}^n Y_i,$$

where the $\{Y_i, i \in \mathbb{N}\}$ are iid with distribution μ . For $\alpha \in \mathbb{R}^d$ let

$$H(\alpha) \doteq \log \int_{\mathbb{R}^d} e^{\langle \alpha, y \rangle} \mu(dy),$$

where we assume $\int_{\mathbb{R}^d} \exp \langle \alpha, y \rangle \mu(dy) < \infty$ for each such α . Consider also the Legendre transform

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

Importance sampling is a Monte Carlo method for the estimation of expected values. One samples from a distribution that may differ from the true distribution, and in order to guarantee that the resulting estimate is unbiased one multiplies each sample by the appropriate Radon-Nikodým derivative. The goal is then to choose the sampling distribution so that this estimate has low variance. Suppose the functional of interest is

$$E \exp \{-nF(X_n)\}.$$

In the context of sums of iid random variables, one typically uses the following parametric family of exponential changes of measure to generate the replacements for the Y_i :

$$\mu_\alpha(dy) \doteq e^{\langle \alpha, y \rangle - H(\alpha)} \mu(dy).$$

In constructing the replacement for X_n we use a *dynamic* change of measure. For a function $\bar{\alpha}(x, t) : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$ recursively define the following quantities. Let $\bar{X}_0^n = 0$, and assume that $\bar{X}_j^n, \bar{Y}_j^n, j = 1, \dots, i$ have been defined. Let \bar{Y}_{i+1}^n , conditioned on $\bar{X}_j^n, \bar{Y}_j^n, j = 1, \dots, i$, have distribution $\mu_{\bar{\alpha}(\bar{X}_i^n, i/n)}$, and then set $\bar{X}_{i+1}^n \doteq \bar{X}_i^n + \bar{Y}_{i+1}^n/n$. When \bar{X}_i^n, \bar{Y}_i^n have been defined for all $i = 1, \dots, n$, set

$$Z^n \doteq e^{-nF(\bar{X}_n^n)} \prod_{i=0}^{n-1} e^{H(\bar{\alpha}(\bar{X}_i^n, i/n)) - \langle \bar{\alpha}(\bar{X}_i^n, i/n), \bar{Y}_{i+1}^n \rangle}.$$

It is easy to check that $EZ^n = Ee^{-nF(X_n)}$, and so the average of K independent samples of Z^n converges almost surely to $Ee^{-nF(X_n)}$ as $K \rightarrow \infty$. Since the estimator is unbiased, to minimize the variance one can minimize the second moment, and to do this it is enough to minimize the second moment of the single sample Z^n .

We consider the problem of minimizing the second moment as a control problem, with $\bar{\alpha}$ the control. It is here that the problem connects naturally

with a PDE. To make the connection we must extend the problem slightly. For $i \in \mathbb{N} \cup \{0\}$ and $x \in \mathbb{R}^d$, define $\bar{X}_j^n, j = i, \dots, n-1$ as above save $\bar{X}_i^n = x$, and then define

$$V^n(x, i) \doteq \inf_{\bar{\alpha}} E \left[e^{-nF(\bar{X}_n^n)} \prod_{j=i}^{n-1} e^{H(\bar{\alpha}(\bar{X}_j^n, j/n)) - \langle \bar{\alpha}(\bar{X}_j^n, j/n), \bar{Y}_{j+1}^n \rangle} \right]^2.$$

It will be more convenient to express this in terms of the original random variables:

$$V^n(x, i) \doteq \inf_{\bar{\alpha}} E \left[e^{-n2F(X_n^n)} \prod_{j=i}^{n-1} e^{H(\bar{\alpha}(X_j^n, j/n)) - \langle \bar{\alpha}(X_j^n, j/n), Y_{j+1}^n \rangle} \right].$$

Owing to the exponential scaling in n , one gets a simple asymptotic problem by considering the logarithmic transform

$$W^n(x, i) = -\frac{1}{n} \log V^n(x, i).$$

The performance of the scheme corresponding to $\bar{\alpha}$ can then be characterized in terms of $\liminf_{n \rightarrow \infty} W^n(0, 0)$, with larger values indicating better performance.

2.2 The associated Isaacs equation

V^n is the value function of a discrete time stochastic control problem, and as such, satisfies the dynamic programming equation

$$V^n(x, i) = \inf_{\alpha} \left[\int_{\mathbb{R}^d} e^{H(\alpha) - \langle \alpha, y \rangle} V^n(x + y/n, i + 1) \mu(dy) \right].$$

A variational formula involving relative entropy (see [1, Section 1.4] and below) shows how to represent exponential integrals in terms of relative entropy. For $\gamma \in \mathcal{P}(\mathbb{R}^d)$ with $\gamma \ll \mu$ and $\log(d\gamma/d\mu)$ integrable with respect to γ set

$$R(\gamma \|\mu) = \int_{\mathbb{R}^d} \log \left(\frac{d\gamma}{d\mu} \right) d\gamma,$$

and otherwise set $R(\gamma \|\mu) = \infty$. Then for any bounded and continuous function $f : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$-\log \int_{\mathbb{R}^d} e^{-f(y)} \mu(dy) = \inf_{\gamma \in \mathcal{P}(\mathbb{R}^d)} \left[R(\gamma \|\mu) + \int_{\mathbb{R}^d} f(y) \gamma(dy) \right].$$

Applying this to the dynamic programming equation and using the definition of W^n gives the following discrete time Isaacs equation:

$$W^n(x, i) = \sup_{\alpha \in \mathbb{R}^d} \inf_{\gamma \in \mathcal{P}(\mathbb{R}^d)} \left[\int_{\mathbb{R}^d} W^n \left(x + \frac{y}{n}, i + 1 \right) \gamma(dy) + \frac{1}{n} \left(R(\gamma \parallel \mu) + \int_{\mathbb{R}^d} \langle \alpha, y \rangle \gamma(dy) - H(\alpha) \right) \right].$$

To formally relate $W^n(x, i)$ to the solution of a PDE, suppose that for a smooth function $W : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}$,

$$W^n(x, i) \approx W(x, i/n).$$

We also use the following relationship between relative entropy and the function L defined as the Legendre transform of H (see [1, Section C.5]). For any $\beta \in \mathbb{R}^d$

$$\inf \left[R(\gamma \parallel \mu) : \int_{\mathbb{R}^d} y \gamma(dy) = \beta \right] = L(\beta). \quad (2.1)$$

(It in fact turns out that the infimizing γ is of the form μ_α for the point α that is conjugate to β in the sense of convex duality. It is for this reason that the class of “exponential tilts” is asymptotically optimal.) We then bring $W^n(x, i) \approx W(x, i/n)$ to the right side of the Isaacs equation, expand via Taylor series, insert the relation above and then multiply by n and send $n \rightarrow \infty$ to get

$$W_t(x, t) + \sup_{\alpha \in \mathbb{R}^d} \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(DW(x, t); \alpha, \beta) = 0.$$

Here W_t denotes the partial derivative with respect to t , DW the gradient in x , and

$$\mathbb{H}(s; \alpha, \beta) \doteq \langle s, \beta \rangle + L(\beta) + \langle \alpha, \beta \rangle - H(\alpha) \quad (2.2)$$

for $s, \alpha, \beta \in \mathbb{R}^d$. Note that also one expects the terminal condition $W(x, 1) = 2F(x)$ to hold.

This PDE, which is also known as an Isaacs equation, was identified in [2] and used there to study the performance of certain importance sampling schemes. However, the purpose of the present paper is to show that it is only the subsolution property that is essential. By a classical sense subsolution, we mean a function $\bar{W} : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}$ with a smooth extension to an open neighborhood of $\mathbb{R}^d \times [0, 1]$ such that

$$\bar{W}_t(x, t) + \sup_{\alpha \in \mathbb{R}^d} \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(D\bar{W}(x, t); \alpha, \beta) \geq 0$$

for all (x, t) and $\bar{W}(x, 1) \leq 2F(x)$. We also consider a priori fixed change of measure controls $\bar{\alpha}(x, t)$, and call $(\bar{W}, \bar{\alpha})$ a subsolution/control pair if

$$\bar{W}_t(x, t) + \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(D\bar{W}(x, t); \bar{\alpha}(x, t), \beta) \geq 0$$

for all (x, t) and $\bar{W}(x, 1) \leq 2F(x)$. The definition of a subsolution simply replaces the equality that appears in the Isaacs equation and terminal condition with inequalities. However, we are only interested in bounding the quantity $W^n(0, 0)$ from *below*, with an upper bound being automatic from the fact that the best possible performance is bounded. The inequalities in the definition are those which give lower bounds when the smooth subsolution is combined with a verification argument to estimate the performance.

Remark 2.1 The supremum and infimum in the Isaacs equation can be evaluated to give

$$W_t - 2H(-DW/2) = 0.$$

This equation immediately suggests the form of certain simple but important solutions to the Isaacs equation—see the next section and the discussion in Section 3.1 of [4]. However, the analysis of a specific proposed importance sampling scheme requires the equation and definition given above for a subsolution/control pair.

In the remainder of this motivational section we give several simple examples of subsolutions and discuss how the Isaacs equation gives bounds on the second moment of the associated schemes.

2.3 Two simple examples

Example 1. Let F be convex and bounded from below. Interchanging the supremum and infimum in the Isaacs equation and evaluating the supremum on α gives

$$W_t(x, t) + \inf_{\beta \in \mathbb{R}^d} [\langle DW(x, t), \beta \rangle + 2L(\beta)] = 0.$$

The viscosity solution to this PDE and terminal condition is well known, and indeed

$$W(x, t) = \inf_{\beta \in \mathbb{R}^d} [2(1-t)L(\beta) + 2F(x + (1-t)\beta)].$$

(Strictly speaking, this solution need not be smooth. We will not concern ourselves with such issues in this overview, but note that all the subsolutions we work with will be classical sense, smooth subsolutions.)

Although it is easy in this example to construct the exact solution, one may wish to obtain a subsolution that will generate simpler importance sampling schemes with the same asymptotic performance. As we will see, the property that is needed so that the asymptotic performance of a subsolution \bar{W} is optimal is $\bar{W}(0,0) = W(0,0)$. Let β^* achieve the infimum in the definition of $W(0,0)$, and let α^* satisfy

$$L(\beta^*) = \langle \alpha^*, \beta^* \rangle - H(\alpha^*)$$

(i.e., α^* is conjugate to β^*). Let

$$\bar{W}(x, t) \doteq -2 \langle \alpha^*, x \rangle + 2tH(\alpha^*) + 2[L(\beta^*) + F(\beta^*)].$$

Then $\bar{W}_t(x, t) = 2H(\alpha^*)$ and $D\bar{W}(x, t) = -2\alpha^*$. Since

$$\begin{aligned} & \bar{W}_t(x, t) + \sup_{\alpha \in \mathbb{R}^d} \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(D\bar{W}(x, t); \alpha, \beta) \\ &= 2H(\alpha^*) + \inf_{\beta \in \mathbb{R}^d} \sup_{\alpha \in \mathbb{R}^d} [\langle -2\alpha^*, \beta \rangle + L(\beta) + \langle \alpha, \beta \rangle - H(\alpha)] \\ &= 2H(\alpha^*) + \inf_{\beta \in \mathbb{R}^d} [\langle -2\alpha^*, \beta \rangle + 2L(\beta)] \\ &= 2H(\alpha^*) - 2 \sup_{\beta \in \mathbb{R}^d} [\langle \alpha^*, \beta \rangle - L(\beta)] \\ &= 2H(\alpha^*) - 2H(\alpha^*) \\ &= 0, \end{aligned}$$

we have only to check the terminal condition. To simplify we assume L is differentiable at β^* , a very mild condition. Then one can verify that $\bar{W}(x, 1)$ is a supporting hyperplane to $2F$ at β^* , and so $\bar{W}(x, 1) \leq 2F(x)$. Note also that $\bar{W}(0, 0) = 2[L(\beta^*) + F(\beta^*)] = W(0, 0)$. Thus \bar{W} achieves the maximum possible value among all subsolutions at $(0, 0)$, with a much simpler structure than the true solution.

Evaluating the infimum on β first, we find that in the case of the exact solution

$$\begin{aligned} & \sup_{\alpha \in \mathbb{R}^d} \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(DW(x, t); \alpha, \beta) \\ &= \sup_{\alpha \in \mathbb{R}^d} - \left(\sup_{\beta \in \mathbb{R}^d} [-L(\beta) + \langle -DW(x, t) - \alpha, \beta \rangle + H(\alpha)] \right) \\ &= - \inf_{\alpha \in \mathbb{R}^d} [H(-DW(x, t) - \alpha) + H(\alpha)]. \end{aligned}$$

By convexity the supremum on α is achieved at $\bar{\alpha}(x, t) = -DW(x, t)/2$. This is the importance sampling control that is naturally suggested by the

exact solution. For the affine subsolution the supremum is at $\bar{\alpha}(x, t) = -D\bar{W}(x, t)/2 = \alpha^*$. The corresponding very simple importance sampling scheme is well known in the literature.

Example 2. Here we take $F(x) = F_1(x) \wedge F_2(x)$, where each F_i is convex and bounded from below. Although the exact solution takes the same form as in Example 1, there may not be an affine subsolution. However, it is natural to consider the minimum of the affine subsolutions associated with each of the F_i . Thus let

$$[L(\beta_i^*) + F_i(\beta_i^*)] = \inf_{\beta \in \mathbb{R}^d} [L(\beta) + F_i(\beta)],$$

let α_i^* be the convex conjugate of β_i^* , and set

$$\bar{W}_i(x, t) = -2 \langle \alpha_i^*, x \rangle + 2tH(\alpha_i^*) + 2[L(\beta_i^*) + F_i(\beta_i^*)].$$

Let $\bar{W}(x, t) \doteq \bar{W}_1(x, t) \wedge \bar{W}_2(x, t)$. Ignoring the issue of what is meant at points where \bar{W} is not differentiable, this provides a subsolution. One can check that

$$\bar{W}(0, 0) = 2 \wedge_{i=1}^2 [L(\beta_i^*) + F_i(\beta_i^*)] = W(0, 0).$$

To produce a smooth subsolution, it turns out that one can simply mollify $\bar{W}(x, t)$ [4]. Since \bar{W} is identified as the pointwise minimum of smooth (affine) functions, we use the standard approximation which we will call *exponential weighting*. Let δ be a small positive number, and

$$\bar{W}^\delta(x, t) \doteq -\delta \log \left(\sum_{i=1}^2 e^{-\frac{1}{\delta} \bar{W}_i(x, t)} \right).$$

Define the probability vector $(\rho_1(x, t), \rho_2(x, t))$ by

$$\rho_i(x, t) \doteq e^{-\frac{1}{\delta} \bar{W}_i(x, t)} / \left(e^{-\frac{1}{\delta} \bar{W}_1(x, t)} + e^{-\frac{1}{\delta} \bar{W}_2(x, t)} \right).$$

It follows that

$$\bar{W}_t^\delta(x, t) = \rho_1(x, t)2H(\alpha_1^*) + \rho_2(x, t)2H(\alpha_2^*), \quad (2.3)$$

$$D\bar{W}^\delta(x, t) = -\rho_1(x, t)2\alpha_1^* - \rho_2(x, t)2\alpha_2^*. \quad (2.4)$$

One can also easily verify that

$$\bar{W}(x, t) \geq \bar{W}^\delta(x, t) \geq \bar{W}(x, t) - \delta \log 2.$$

Hence the value $\bar{W}^\delta(0,0)$ may be slightly smaller than $\bar{W}(0,0)$, but this difference can be made arbitrarily small.

The optimizing $\bar{\alpha}$ can be found as in Example 1:

$$\bar{\alpha}(x,t) = \rho_1(x,t)\alpha_1^* + \rho_2(x,t)\alpha_2^*.$$

There are (at least) two ways that this control can be implemented. The first is the one given at the beginning of this overview, i.e., $\mu_{\bar{\alpha}(\bar{X}_i^n, i/n)}$ chooses the distribution of \bar{Y}_{i+1}^n . The second implementation leads to the notion of *generalized subsolution/control*, which will be discussed further in Section 5. With this implementation, one chooses between the indices 1 and 2, conditioned on all the past data, with weights $\rho_1(\bar{X}_i^n, i/n)$ and $\rho_2(\bar{X}_i^n, i/n)$, and then depending on the outcome generates \bar{Y}_{i+1}^n according to $\mu_{\alpha_1^*}$ or $\mu_{\alpha_2^*}$. The latter implementation has some advantages when the underlying process is more complicated than an iid sequence (i.e., a functional of a Markov chain). Of course with this implementation the Radon-Nikodým derivative takes a different form than the one given at the beginning of this section. See Section 6.

Many more examples and a more systematic approach to the construction of subsolutions appears in [4].

Remark 2.2 There are other methods of mollification to produce smooth subsolutions. For example, one can integrate \bar{W} against a smooth convolution kernel with support in the ball of radius δ around 0. It turns out that the resulting approximation (abusing notation) \bar{W}^δ also satisfies equations (2.3) and (2.4) for some probability vector $(\rho_1(x,t), \rho_2(x,t))$. However, the computation of $\rho_i(x,t)$ involves numerical integration and can be computationally demanding. In contrast, for the exponential weighting mollification we use, the $\{\rho_i(x,t)\}$ are easy to compute.

2.4 Performance of the schemes

Finally we remark on the performance of the importance sampling schemes so constructed. The main result of this paper, which will be proved for a more complex process model and which can be generalized considerably, is the following: If Z^n is constructed according to the subsolution/control pair $(\bar{W}, \bar{\alpha})$ and

$$W^n \doteq -\frac{1}{n} \log E(Z^n)^2,$$

then

$$\liminf_{n \rightarrow \infty} W^n \geq \bar{W}(0,0). \tag{2.5}$$

Moreover, the same result is true for generalized subsolutions and controls.

An elementary calculation based on Jensen’s inequality shows that for *any* importance sampling schemes the best performance possible is $W(0, 0)$, where W is the exact solution. Hence the design problem becomes clear. Construct a subsolution/control pair which can be implemented with reasonable effort and for which $\bar{W}(0, 0)$ is acceptably close to $W(0, 0)$.

The remainder of this paper is devoted to the precise statement and proof of (2.5) is a more general setting.

3 The General Setup

The broader collection of importance sampling problems we wish to analyze includes sums of independent and identically distributed (iid) random variables and sums of functionals of a finite state Markov chain. The following general model includes both as special cases. Let $Y \doteq \{Y_i, i \in \mathbb{N}_0\}$ denote a Markov chain with state space S . Assume that S is a Polish space, and let $p(y, dz)$ denote the probability transition kernel. Let $\{b_i(\cdot), i \in \mathbb{N}_0\}$ be a sequence of iid random vector fields on S that is independent of the Markov chain Y . For each $y \in S$, $b_i(y)$ is distributed according to a probability measure, say $m(\cdot|y)$, on \mathbb{R}^d . Our interest is in sums of the form

$$X_n \doteq \frac{1}{n} \sum_{i=1}^n b_i(Y_i). \tag{3.1}$$

By choosing S to be a single point we recover the case of sums of iid random variables, whereas taking $b_i(y)$ to be deterministic [i.e., $m(\cdot|y)$ is a single atom for each $y \in S$] produces the case of functionals of a Markov chain. The general case is also of interest, and occurs when the distribution of the summand b_i is modulated by the “exogenous” process Y .

Remark 3.1 In the literature on importance sampling for Markov chains it is standard to include the initial state $Y_0 = y$ in the sample mean. The sole reason to consider the sum from $i = 1$ to n , as in the definition (3.1) of X_n , is that it significantly simplifies our notation in later analysis. We point out, however, that there is no loss of generality, in that all the results in this paper hold if we replace definition (3.1) by the standard one where the summation is taken from $i = 0$ to $i = n - 1$.

Condition 3.1 *The following conditions are assumed throughout the paper.*

1. There is a reference probability measure λ on S , a positive integer m_0 , and $\delta \in (0, 1)$, such that

$$\delta\lambda(dy_2) \leq p^{(m_0)}(y_1, dy_2) \leq \frac{1}{\delta}\lambda(dy_2)$$

for all $y_1 \in S$. Here $p^{(m_0)}$ is the m_0 -step transition kernel corresponding to p .

2. The transition kernel $p(y_1, dy_2)$ satisfies the Feller property, i.e., the mapping $y_1 \mapsto p(y_1, dy_2)$ is continuous in the topology of weak convergence of probability measures on S .
3. The mapping $y \mapsto m(dz|y)$ is continuous in the topology of weak convergence of probability measures on \mathbb{R}^d .
4. For each $\alpha \in \mathbb{R}^d$,

$$\sup_{y \in S} \int_{\mathbb{R}^d} e^{\langle \alpha, z \rangle} m(dz|y) < \infty.$$

Note that parts 1, 2, and 3 of Condition 3.1 automatically hold when Y is an irreducible finite state Markov chain.

For a pair of probability measures $\gamma, \mu \in \mathcal{P}(S)$, we recall that the relative entropy of γ with respect to μ was defined as

$$R(\gamma||\mu) \doteq \int_S \log \frac{d\gamma}{d\mu} d\gamma$$

if $\gamma \ll \mu$ and $R(\gamma||\mu) \doteq \infty$ otherwise. The relative entropy $R(\gamma||\mu)$ is always non-negative, and is a convex, lower semicontinuous function of $(\gamma, \mu) \in \mathcal{P}(S) \times \mathcal{P}(S)$. We refer the reader to [1, Section 1.4] for the proof and other properties of relative entropy.

Under Condition 3.1, $\{X_n, n \in \mathbb{N}\}$ satisfies a large deviation principle with the rate function

$$L(\beta) = \inf \left\{ R(\tau || \theta \otimes p) + R(\theta \otimes \nu || \theta \otimes m) \right. \\ \left. : [\tau]_1 = [\tau]_2 = \theta, \int_S \int_{\mathbb{R}^d} z \nu(dz|y) \theta(dy) = \beta \right\}. \quad (3.2)$$

Here τ is a probability measure on $S \times S$ and ν is a stochastic kernel on \mathbb{R}^d given S . The fact that a large deviation principle holds is proved in [7],

although they do not identify the rate function in this form but rather in terms of a Legendre transform. One can give a direct proof of the large deviation result as in [1] which automatically gives this more concrete form of the rate function (3.2), which is analogous to (2.1). See, in particular, the analogous prelimit representation formula in [1, Section 4.4].

For a Borel measurable function $F : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{\infty\}$, we wish to numerically approximate the quantity

$$E_y \exp \{-nF(X_n)\}, \quad (3.3)$$

where E_y denotes expected value given initial state $Y_0 = y$. The special case of $P_y \{X_n \in A\}$ is obtained by letting $F(x) = 0$ for $x \in A$ and $F(x) = \infty$ for $x \notin A$. Under various sets of regularity conditions on F , one has the large deviation asymptotic approximation [8, 1]

$$-\frac{1}{n} \log E_y \exp \{-nF(X_n)\} \rightarrow \inf_{\beta \in \mathbb{R}^d} [F(\beta) + L(\beta)]. \quad (3.4)$$

4 Properties of the Relevant Eigenfunctions

It is well known that certain eigenfunctions are needed to construct good importance sampling schemes for functionals of a Markov chain. These eigenfunctions are used to essentially “cancel off” the effect of conditioning on the transition kernel. The eigenvalue/eigenfunction problem is to find, for each $\alpha \in \mathbb{R}^d$, a real number $G(\alpha)$ and a function $r(\cdot; \alpha) : S \rightarrow [0, \infty)$ such that

$$\int_S \int_{\mathbb{R}^d} e^{\langle \alpha, z \rangle} r(y; \alpha) m(dz | y) p(x, dy) = e^{G(\alpha)} r(x; \alpha).$$

A key fact is that the eigenvalues may be defined in terms of the Legendre transform of L . This is defined for $\alpha \in \mathbb{R}^d$ by

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

and is again a convex function.

The needed properties of the solution to this problem are summarized in the following lemma [7, Section 3].

Lemma 4.1 *Assume Condition 3.1. The following conclusions hold.*

1. For each $\alpha \in \mathbb{R}^d$, there exists a solution $(G(\alpha), r(\cdot; \alpha))$ to the eigenvalue/eigenfunction problem, with $G(\alpha) = H(\alpha)$.

2. Let a compact set $K \subset \mathbb{R}^d$ be given. Then there is $\delta \in (0, 1)$ such that $\delta < r(y; \alpha) < 1/\delta$ for all $y \in S$ and $\alpha \in K$.
3. Let a compact set $K \subset \mathbb{R}^d$ be given. Then the map $\alpha \mapsto r(y; \alpha)$ is uniformly Lipschitz continuous over $\alpha \in K$ for each $y \in S$.

5 The Isaacs Equation and Subsolutions

As in Section 2, the partial differential equation associated with this importance sampling problem is the Isaacs equation

$$W_t(x, t) + \sup_{\alpha \in \mathbb{R}^d} \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(DW(x, t); \alpha, \beta) = 0,$$

where \mathbb{H} is as defined in (2.2). We recall that $W : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}$, W_t denotes the partial derivative with respect to t , and DW the gradient in x , and that $(\bar{W}, \bar{\alpha})$ is a *subsolution/control* pair if \bar{W} is continuously differentiable on $\mathbb{R}^d \times (0, 1)$ with a uniformly bounded and uniformly Lipschitz continuous derivative,

$$\bar{W}_t(x, t) + \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(D\bar{W}(x, t); \bar{\alpha}(x, t), \beta) \geq 0,$$

and

$$\bar{W}(x, 1) \leq 2F(x). \tag{5.1}$$

We will also use the term *subsolution* to refer to the \bar{W} component alone, and will sometimes use the phrase even if it is not certain that the terminal condition holds. With each subsolution/control pair, one can associate an importance sampling scheme. The construction and analysis of this scheme are carried out in detail in the next two sections.

A companion paper [4] describes in detail how to construct subsolution/control pairs that satisfy the terminal condition (5.1). As we saw in Section 2, it is often the case that one can work with simple subsolutions (e.g., functions that are affine in x and t) for certain functionals F , and then use the pointwise minimum of such subsolutions to handle more complex F . Suppose we label the individual smooth subsolution/control pairs $(\bar{W}_k, \bar{\alpha}_k), k = 1, \dots, K$. Let $W(x, t) \doteq \wedge_{k=1}^K \bar{W}_k(x, t)$. Since W is not smooth, we mollify and use convexity to obtain a smooth subsolution denoted by \bar{W} . If $\bar{\alpha}(x, t)$ is defined as a saddle point in the min/max problem

$$\sup_{\alpha \in \mathbb{R}^d} \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(D\bar{W}(x, t); \alpha, \beta),$$

then $(\bar{W}, \bar{\alpha})$ is a subsolution/control pair.

For the general model of this paper the individual subsolutions \bar{W}_k are often affine functions in (x, t) , and each $\bar{\alpha}_k$ is a constant. In this case, the change of measure associated with each subsolution/control pair $(\bar{W}_k, \bar{\alpha}_k)$ is fairly simple. For example, in the case of functionals of a Markov process we need only compute the corresponding eigenfunction at the single point $\bar{\alpha}_k$. However, the subsolution/control pair $(\bar{W}, \bar{\alpha})$, where $\bar{\alpha}$ is defined by the saddle point property, produces a state and time dependent $\bar{\alpha}$, and thus a scheme that could be significantly more complicated.

As discussed in Section 2, a scheme which preserves the simplicity of the affine subsolutions can be found by appropriately randomizing between the individual importance sampling schemes associated with $(\bar{W}_k, \bar{\alpha}_k)$ according to the mollification weights. Schemes of this sort require the following more complicated notion of a subsolution/control pair, which subsumes the previous special case. As suggested by the examples given in Section 2, it is typically the case that $r_k = (\bar{W}_k)_t$ and $s_k = D\bar{W}_k$ in the definition below.

Definition 5.1 *The collection $(\bar{W}, \rho_k, \bar{\alpha}_k)$ will be called a generalized subsolution/control if the following conditions hold. $\rho_k : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}, k = 1, \dots, K$ is a partition of unity, i.e., each ρ_k is non-negative, and*

$$\sum_{k=1}^K \rho_k(x, t) = 1$$

for all $(x, t) \in \mathbb{R}^d \times [0, 1]$. The functions ρ_k and $\bar{\alpha}_k, k = 1, \dots, K$, are uniformly bounded and Lipschitz continuous. \bar{W}_t and $D\bar{W}$ have representations

$$\bar{W}_t(x, t) = \sum_{k=1}^K \rho_k(x, t)r_k(x, t), \quad D\bar{W}(x, t) = \sum_{k=1}^K \rho_k(x, t)s_k(x, t),$$

where each r_k and s_k is uniformly bounded and Lipschitz continuous, and for each $k = 1, \dots, K$

$$r_k(x, t) + \inf_{\beta \in \mathbb{R}^d} \mathbb{H}(s_k(x, t); \bar{\alpha}_k(x, t), \beta) \geq 0.$$

It is only the $(\rho_k, \bar{\alpha}_k)$ part of this collection that will be used to define the importance sampling scheme (see the next section). As noted in Section 2, a key measure of efficiency of any importance sampling scheme associated with the collection $(\bar{W}, \rho_k, \bar{\alpha}_k)$ is $\bar{W}(0, 0)$, with larger values of $\bar{W}(0, 0)$ corresponding to greater variance reduction. The design problem is to maximize $\bar{W}(0, 0)$, subject to the constraints that $(\bar{W}, \rho_k, \bar{\alpha}_k)$ be a generalized

subsolution/control, and that the terminal condition (5.1) hold. There is considerable flexibility, and an appreciation of all the possibilities for even simple problems requires some experience. These issues are explored at length in [4].

6 Importance Sampling Based on Subsolutions

As discussed in Section 2 and many other places, the idea behind importance sampling is to simulate the system of interest under an alternative distribution, multiply the sample by the inverse of the Radon-Nikodým derivative, and then consider the sample average. The main issue is how to choose the new distribution so that the variance of the estimate is as small as possible. It is also by now well known that seemingly reasonable schemes can perform very poorly [5, 6], whence the development of usable tools for the analysis of variance is essential if the method is ever to be used with any confidence. As we now show, the subsolution property gives strong quantitative control on the second moments of the estimates under the scheme.

Let $(\bar{W}, \rho_k, \bar{\alpha}_k)$ be a generalized subsolution/control. We recall the eigenvalue/eigenfunction relation

$$\int_S \int_{\mathbb{R}^d} e^{\langle \alpha, z \rangle} r(y; \alpha) m(dz | y) p(x, dy) = e^{H(\alpha)} r(x; \alpha).$$

It follows that for each $\alpha \in \mathbb{R}^d$

$$P(y_1, dy_2, dz; \alpha) = e^{\langle \alpha, z \rangle - H(\alpha)} \cdot \frac{r(y_2; \alpha)}{r(y_1; \alpha)} \cdot p(y_1, dy_2) \cdot m(dz | y_2) \quad (6.1)$$

defines a probability measure on $S \times \mathbb{R}^d$. These probability measures, the weights $\rho_k(x, t)$, and the functions $\bar{\alpha}_k(x, t)$ will be used to construct the importance sampling scheme. To this end, let

$$\bar{\alpha}_{k,j}^n(x) \doteq \bar{\alpha}_k(x, j/n), \quad \rho_{k,j}^n(x) \doteq \rho_k(x, j/n).$$

Processes \bar{X}_j^n, \bar{Y}_j^n , and \bar{b}_j^n , analogous to X_j, Y_j , and $b_j(Y_j)$, are constructed recursively as follows. Let $\bar{X}_0^n = 0$ and $\bar{Y}_0 = Y_0 = y$. Suppose that $\bar{X}_j^n = x$ and $\bar{Y}_j^n = y_1$ are given. We then simulate $(\bar{Y}_{j+1}^n, \bar{b}_{j+1}^n)$ under the distribution

$$\sum_{k=1}^K \rho_{k,j}^n(x) P(y_1, dy_2, dz; \bar{\alpha}_{k,j}^n(x)),$$

which can be thought of as a randomized version of $P(y_1, dy_2, dz; \bar{\alpha}_{k,j}^n(x))$, with the weights given by $\rho_{k,j}^n(x)$. Finally, $\bar{X}_{j+1}^n \doteq \bar{X}_j^n + \bar{b}_{j+1}^n/n$. An unbiased estimate for $E \exp\{-nF(X_n)\}$ is then obtained by averaging replications of

$$Z^n \doteq e^{-nF(\bar{X}_n^n)} \prod_{j=0}^{n-1} \left[\sum_{k=1}^K \rho_{k,j}^n(\bar{X}_j^n) \cdot e^{\langle \bar{\alpha}_{k,j}^n(\bar{X}_j^n), \bar{b}_{j+1}^n \rangle - H(\bar{\alpha}_{k,j}^n(\bar{X}_j^n))} \cdot \frac{r(\bar{Y}_{j+1}^n; \bar{\alpha}_{k,j}^n(\bar{X}_j^n))}{r(\bar{Y}_j^n; \bar{\alpha}_{k,j}^n(\bar{X}_j^n))} \right]^{-1}. \quad (6.2)$$

As noted previously, the numerical estimate in importance sampling is the sample average of independent replications of Z^n . Since the goal is to control the sample variance, it is enough to bound the second moment of a single replication.

7 Statement and Proof of the Main Result

In this section we present the main result, which is an asymptotic bound on the second moment for importance sampling estimator associated with a given subsolution. Although both the quantity being approximated and the importance sampling scheme depend on the initial state $Y_0 = y$, to simplify the exposition, the dependence of expected values on y is not explicitly denoted.

Theorem 7.1 *Assume Condition 3.1. Let $(\bar{W}, \rho_k, \bar{\alpha}_k)$ be a generalized subsolution/control such that $2F(x) \geq \bar{W}(x, 1)$ for every $x \in \mathbb{R}^d$. Let V^n be the second moment of a single replication used in the corresponding importance sampling scheme for the estimation of $E \exp\{-nF(X_n)\}$, that is,*

$$V^n \doteq E [(Z^n)^2],$$

where Z^n is defined in (6.2). Let

$$W^n \doteq -\frac{1}{n} \log V^n.$$

Then

$$\liminf_{n \rightarrow \infty} W^n \geq \bar{W}(0, 0).$$

Outline of the Proof of Theorem 7.1. By expressing the second moment in terms of the original random variables, we can write

$$V^n = E e^{-2nF(X_n)} \prod_{j=0}^{n-1} \left[\sum_{k=1}^K \rho_{k,j}^n(X_j) \cdot e^{\langle \bar{\alpha}_{k,j}^n(X_j), b_{j+1}(Y_{j+1}) \rangle - H(\bar{\alpha}_{k,j}^n(X_j))} \cdot \frac{r(Y_{j+1}; \bar{\alpha}_{k,j}^n(X_j))}{r(Y_j; \bar{\alpha}_{k,j}^n(X_j))} \right]^{-1}.$$

The proof is divided into 5 parts.

1. *Representation.* We replace V^n by an upper bound, and then derive a stochastic control representation for the normalized logarithm of this quantity. This produces a lower bound for W^n .
2. *Tightness.* Associate certain stochastic processes and measure valued processes to the representation. Under the assumption that the costs in the representation are bounded as $n \rightarrow \infty$, show that these processes are tight.
3. *Identification of limits.* Derive characterizations and relations between the limit processes.
4. *Analysis of the cost.* Go back to the representation, and analyze the asymptotics of the cost using weak convergence.
5. *Verification.* Finally, use the Isaacs equation and a classical verification argument to show that the proper asymptotic bound holds for the representation.

The chain rule for relative entropy (see, e.g., [1, Theorem C.3.1]) will be used several times in the proof. If S_1 and S_2 are Polish spaces and $\mu, \nu \in P(S_1 \times S_2)$, then

$$R(\mu \parallel \nu) = R([\mu]_1 \parallel [\nu]_1) + \int_{S_1} R(\mu(\cdot|y_1) \parallel \nu(\cdot|y_1)) [\mu]_1(dy_1) \quad (7.1)$$

7.1 Representation.

Using convexity of e^x and the definition $G(x) \doteq \bar{W}(x, 1) \leq 2F(x)$, the second moment V^n is bounded above by

$$\tilde{V}^n \doteq E e^{-nG(X_n)} \prod_{j=0}^{n-1} \exp \left\{ - \sum_{k=1}^K \rho_{k,j}^n(X_j) \left[\langle \bar{\alpha}_{k,j}^n(X_j), b_{j+1}(Y_{j+1}) \rangle \right. \right.$$

$$\left. - H(\bar{\alpha}_{k,j}^n(X_j)) + \log \frac{r(Y_{j+1}; \bar{\alpha}_{k,j}^n(X_j))}{r(Y_j; \bar{\alpha}_{k,j}^n(X_j))} \right\}.$$

Define

$$\tilde{W}^n \doteq -\frac{1}{n} \log \tilde{V}^n.$$

Clearly $\tilde{W}^n \leq W^n$. Therefore, it suffices to show

$$\liminf_{n \rightarrow \infty} \tilde{W}^n \geq \bar{W}(0, 0). \quad (7.2)$$

We would like to use the variational representation for exponential integrals to derive a stochastic control representation for \tilde{W}^n . Because of the unbounded terms $\langle \bar{\alpha}_{k,j}^n(X_j), b_{j+1}(Y_{j+1}) \rangle$ and $G(X_n)$, an extension of this representation is required.

Lemma 7.2 *Let λ be a probability measure on a measurable space (Ω, \mathcal{F}) , and $f : \Omega \rightarrow \mathbb{R}$ a measurable function. If e^{-f} and $f e^{-f}$ are integrable with respect to λ , then*

$$-\log \int_{\Omega} e^{-f} d\lambda = \inf_{\gamma} \left\{ R(\gamma \| \lambda) + \int_{\Omega} f d\gamma \right\},$$

where the infimum is taken over all probability measures γ for which the sum on the right-hand-side is meaningful (i.e., not of the form $\infty - \infty$).

The proof only involves minor changes to that of [1, Proposition 1.4.2] and is thus omitted. It is easy to check that the condition for this representation, that is, the finiteness of the two integrals, holds in our case. This is due to the bound on the moment generating function of the $b_i(y)$ and the assumed Lipschitz property of $G(x) \doteq \bar{W}(x, 1)$.

Once one has this general relative entropy representation for exponential integrals, it is easy to extract a more useful form by a standard argument. Consider the total distribution, say λ , of the component random variables used to construct the process [here the Y_i and $b_i(Y_i)$], and write the expectation in terms of an exponential integral against this distribution. Apply the relative entropy representation to this exponential integral, and let γ be the probability measure introduced by the representation. Now factor both the original distribution λ and the new probability measure γ as a product of conditional distributions. For example, if λ were a distribution on S^3 it would be factored as $[\lambda]_1(dx_1)[\lambda]_2(dx_2|x_1)[\lambda]_3(dx_3|x_1, x_2)$. One then decomposes the relative entropy according to the chain rule (7.1), giving rise to

a relative entropy cost for the perturbation of the conditional distribution of each component random variable. Finally, for convenience one writes the right hand side of the relative entropy representation in terms of this decomposition and random variables distributed according to the new probability measure. Since the analogous elementary proof appears in many places (e.g., [1, Theorem B.2.2]), we simply state the final result. Consider a collection of stochastic kernels μ_j^n and ν_j^n , where μ_j^n is allowed to depend in any measurable way on $\{\tilde{b}_i^n, 0 \leq i \leq j\}$ and $\{\tilde{Y}_i^n, 0 \leq i \leq j+1\}$, ν_j^n is allowed to depend in any measurable way on $\{\tilde{b}_i^n, 0 \leq i \leq j\}$ and $\{\tilde{Y}_i^n, 0 \leq i \leq j\}$, and μ_j^n and ν_j^n choose the conditional distributions of \tilde{b}_{j+1}^n and \tilde{Y}_{j+1}^n , respectively. To simplify the notation the dependencies of μ_j^n and ν_j^n on the past will not be made explicit. Let

$$\begin{aligned}
J(\mu^n, \nu^n) &\doteq \tilde{E} \left[\frac{1}{n} \sum_{j=0}^{n-1} \sum_{k=1}^K \rho_{k,j}^n(\tilde{X}_j^n) \left[R\left(\mu_j^n(\cdot) \| m(\cdot | \tilde{Y}_{j+1}^n)\right) + R\left(\nu_j^n(\cdot) \| p(\tilde{Y}_j^n, \cdot)\right) \right. \right. \\
&\quad \left. \left. + \left\langle \bar{\alpha}_{k,j}^n(\tilde{X}_j^n), \tilde{b}_{j+1}^n \right\rangle - H\left(\bar{\alpha}_{k,j}^n(\tilde{X}_j^n)\right) + \log \frac{r(\tilde{Y}_{j+1}^n; \bar{\alpha}_{k,j}^n(\tilde{X}_j^n))}{r(\tilde{Y}_j^n; \bar{\alpha}_{k,j}^n(\tilde{X}_j^n))} \right] \right. \\
&\quad \left. + G(\tilde{X}_n^n) \right]. \tag{7.3}
\end{aligned}$$

Then $\tilde{W}^n \doteq \inf J(\mu^n, \nu^n)$, where the infimum is over all such collections.

7.2 Tightness

To analyze the asymptotics of \tilde{W}^n we first establish the tightness of the processes that appear therein. For $j = 0, \dots, n-1$ and $t \in [j/n, (j+1)/n)$ define

$$\begin{aligned}
\tilde{X}^n(t) &\doteq \tilde{X}_j^n \\
\nu^n(dy_2 | t) &\doteq \nu_j^n(dy_2) \\
\mu^n(dz | t) &\doteq \mu_j^n(dz) \\
\theta^n(dy_1 \times dy_2 | t) &\doteq \delta_{\tilde{Y}_j^n}(dy_1) \nu_j^n(dy_2) \\
\gamma^n(dy_1 \times dy_2 | t) &\doteq \delta_{\tilde{Y}_j^n}(dy_1) p(y_1, dy_2) \\
\zeta^n(dy \times dz | t) &\doteq \delta_{\tilde{Y}_{j+1}^n}(dy) \mu_j^n(dz) \\
\eta^n(dy \times dz | t) &\doteq \delta_{\tilde{Y}_{j+1}^n}(dy) m(dz | y),
\end{aligned}$$

and let left continuity define these processes at $t = 1$. We also set, for Borel subsets $A \subset S \times S$ and $B \subset [0, 1]$,

$$\theta^n(A \times B) \doteq \int_B \theta^n(A|t) dt.$$

Then θ^n is a random probability measure on space $(S \times S) \times [0, 1]$. Define in the analogous fashion random probability measures $\nu^n, \mu^n, \gamma^n, \zeta^n$, and η^n on spaces $S \times [0, 1], \mathbb{R}^d \times [0, 1], (S \times S) \times [0, 1], (S \times \mathbb{R}^d) \times [0, 1]$, and $(S \times \mathbb{R}^d) \times [0, 1]$, respectively.

Lemma 7.3 *Assume Condition 3.1 and let $(\bar{W}, \rho_k, \bar{\alpha}_k)$ be a generalized sub-solution/control. Consider any subsequence and collection $\{(\mu_j^n, \nu_j^n), j = 0, 1, \dots, n-1\}$ for which the expected cost $J(\mu^n, \nu^n)$ as defined in (7.3) is uniformly bounded from above. Then (with the supremum on n restricted to elements of the subsequence)*

$$\lim_{C \rightarrow \infty} \sup_n \tilde{E} \left[\frac{1}{n} \sum_{j=1}^n \left\| \tilde{b}_j^n \right\| \mathbf{1}_{\{\|\tilde{b}_j^n\| \geq C\}} \right] = 0,$$

the collection

$$\left\{ \left(\tilde{X}^n, \nu^n, \mu^n, \theta^n, \gamma^n, \zeta^n, \eta^n \right) \right\}$$

is tight, $\{\tilde{X}^n(1)\}$ is uniformly integrable, and $\{\mu^n\}$ is uniformly integrable in the sense that

$$\lim_{C \rightarrow \infty} \sup_n \tilde{E} \left[\int_{\mathbb{R}^d \times [0, 1]} \|y\| \mathbf{1}_{\{\|y\| \geq C\}} \mu^n(dy \times dt) \right] = 0.$$

The proof of the lemma is given in the appendix. However, it is worth noting that the first estimate is the key result, and that the tightness and uniform integrability follow easily from this.

In order to show the desired lower bound (7.2), all we need to show is

$$\liminf_{n \rightarrow \infty} J(\mu^n, \nu^n) \geq \bar{W}(0, 0). \quad (7.4)$$

for any sequence $\{(\mu_j^n, \nu_j^n), j = 0, \dots, n-1\}$. Abusing notation a bit, assume from now on that $\{(\mu_j^n, \nu_j^n), j = 0, \dots, n-1\}$ is an arbitrary subsequence such that the cost $J(\mu^n, \nu^n)$ is uniformly bounded from above. Clearly, we only need to show inequality (7.4) along every such subsequence.

Owing to the positivity, boundedness, and Lipschitz properties of the eigenfunctions and $\bar{\alpha}_k$ (see Lemma 4.1), there exists $M < \infty$ such that for all $y \in S$, $k = 1, \dots, K$, $n \in \mathbb{Z}_+$, $j \in \{1, \dots, n\}$, $x_1 \in \mathbb{R}^d$, and $x_2 \in \mathbb{R}^d$,

$$\left| \log \frac{r(y; \bar{\alpha}_{k,j-1}(x_1))}{r(y; \bar{\alpha}_{k,j}(x_2))} \right| = \left| \log \frac{r(y; \bar{\alpha}_k(x_1, (j-1)/n))}{r(y; \bar{\alpha}_k(x_2, j/n))} \right| \leq M(|x_1 - x_2| + 1/n).$$

Thanks to the first part of Lemma 7.3, for any $\delta > 0$ and along this subsequence with bounded cost,

$$\limsup_{n \rightarrow \infty} \tilde{E} \left[\frac{1}{n} \sum_{j=1}^n \|\tilde{b}_j^n\| \mathbf{1}_{\{\|\tilde{b}_j^n\| \geq n\delta\}} \right] = 0.$$

Therefore, the Lipschitz properties of the ρ_k that are part of the definition of a generalized subsolution and the definition $\tilde{X}_{j+1}^n = \tilde{X}_j^n + \tilde{b}_{j+1}^n/n$ imply

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \tilde{E} \left[\frac{1}{n} \sum_{j=0}^{n-1} \sum_{k=1}^K \rho_{k,j}^n(\tilde{X}_j^n) \left| \log \frac{r(\tilde{Y}_{j+1}^n; \bar{\alpha}_{k,j}^n(\tilde{X}_j^n))}{r(\tilde{Y}_j^n; \bar{\alpha}_{k,j}^n(\tilde{X}_j^n))} \right| \right] \\ & \leq \limsup_{n \rightarrow \infty} \tilde{E} \left[\frac{1}{n} \sum_{j=1}^n \sum_{k=1}^K \rho_{k,j}^n(\tilde{X}_j^n) \left| \log \frac{r(\tilde{Y}_j^n; \bar{\alpha}_{k,j-1}^n(\tilde{X}_{j-1}^n))}{r(\tilde{Y}_j^n; \bar{\alpha}_{k,j}^n(\tilde{X}_j^n))} \right| \right] \\ & \quad + \limsup_{n \rightarrow \infty} \tilde{E} \left[\frac{1}{n} \sum_{j=0}^{n-1} \sum_{k=1}^K \left| \rho_{k,j+1}^n(\tilde{X}_{j+1}^n) - \rho_{k,j}^n(\tilde{X}_j^n) \right| \right. \\ & \quad \quad \left. \cdot \left| \log r(\tilde{Y}_{j+1}^n; \bar{\alpha}_{k,j}^n(\tilde{X}_j^n)) \right| \right] \\ & = 0. \end{aligned}$$

Thus we need only prove the lower bound

$$\begin{aligned} & \liminf_{n \rightarrow \infty} \tilde{E} \left[\frac{1}{n} \sum_{j=0}^{n-1} \sum_{k=1}^K \rho_{k,j}^n(\tilde{X}_j^n) \left[R\left(\mu_j^n(\cdot) \left\| m(\cdot | \tilde{Y}_{j+1}^n)\right.\right) + R\left(\nu_j^n(\cdot) \left\| p(\tilde{Y}_j^n, \cdot)\right.\right) \right. \right. \\ & \quad \left. \left. + \left\langle \bar{\alpha}_{k,j}^n(\tilde{X}_j^n), \tilde{b}_{j+1}^n \right\rangle - H\left(\bar{\alpha}_{k,j}^n(\tilde{X}_j^n)\right) \right] + G(\tilde{X}_n^n) \right] \geq \bar{W}(0, 0). \end{aligned}$$

Note that the relative entropy terms do not depend on k , and so they can be moved past the corresponding sum. Thanks to the uniform boundedness and Lipschitz continuity of ρ_k and $\bar{\alpha}_k$, the uniform integrability of $\{\mu^n\}$

(Lemma 7.3), and the chain rule for relative entropy (7.1), all we need to show is the lower bound

$$\liminf_{n \rightarrow \infty} \bar{J}^n \geq \bar{W}(0, 0), \quad (7.5)$$

where

$$\begin{aligned} \bar{J}^n \doteq & \tilde{E} \left[R(\zeta^n \|\eta^n) + R(\theta^n \|\gamma^n) - \sum_{k=1}^K \int_0^1 \rho_k(\tilde{X}^n(t), t) H(\bar{\alpha}_k(\tilde{X}^n(t), t)) dt \right. \\ & \left. + \sum_{k=1}^K \int_{\mathbb{R}^d \times [0,1]} \rho_k(\tilde{X}^n(t), t) \langle \bar{\alpha}_k(\tilde{X}^n(t), t), z \rangle \mu^n(dz \times dt) + G(\tilde{X}^n(1)) \right]. \end{aligned}$$

In order to show (7.5), we need to identify limits of the involved processes. That is the goal of the next subsection.

7.3 Identification of the Limits.

Lemma 7.4 *Assume Condition 3.1, and consider any subsequence along which $J(\mu^n, \nu^n)$ is uniformly bounded from above and*

$$\left(\tilde{X}^n, \nu^n, \mu^n, \theta^n, \gamma^n, \zeta^n, \eta^n \right) \rightarrow \left(\tilde{X}, \nu, \mu, \theta, \gamma, \zeta, \eta \right)$$

in distribution. Then the following conclusions hold. Each of the measures $\nu, \mu, \theta, \gamma, \zeta, \eta$ (for example, ν) can be factored in the form $\nu(dy \times dt) = \nu(dy|t) dt$, where dt is Lebesgue measure. Furthermore, w.p.1

$$\tilde{X}(t) = \int_{[0,t]} \int_{\mathbb{R}^d} z \mu(dz|s) ds,$$

$$\begin{aligned} \gamma(dy_1 \times dy_2|t) &= \nu(dy_1|t) p(y_1, dy_2) \\ \eta(dy \times dz|t) &= \nu(dy|t) m(dz|y), \end{aligned}$$

and

$$\begin{aligned} [\theta]_1(dy|t) &= [\theta]_2(dy|t) = \nu(dy|t), \\ [\zeta]_1(dy|t) &= \nu(dy|t), \quad [\zeta]_2(dy|t) = \mu(dy|t). \end{aligned}$$

Proof. The fact that the t -marginal of the random measures is Lebesgue measure follows from the weak convergence and the fact that the same is true of the analogous prelimit measures. Also, the existence of the factored

form is standard, and follows from the same sort of arguments one uses to prove the existence of regular conditional distributions [1, Lemma 3.3.1].

We next consider the representation for \tilde{X} , and use an argument similar to that of [1, Theorem 5.3.5]. For any time t of the form j/n , $0 \leq j \leq n$, we can write

$$\begin{aligned}\tilde{X}^n(j/n) &= \frac{1}{n} \sum_{i=0}^{j-1} \int_{\mathbb{R}^d} z \mu_i^n(dz) + M^n(j/n) \\ &= \int_0^{j/n} \int_{\mathbb{R}^d} z \mu^n(dz \times dt) + M^n(j/n)\end{aligned}$$

where

$$M^n(j/n) \doteq \frac{1}{n} \sum_{i=0}^{j-1} \left(\tilde{b}_{i+1}^n - \int_{\mathbb{R}^d} z \mu_i^n(dz) \right)$$

is a martingale. Fix $\delta > 0$, and define random variables and random measures

$$c_j^n \doteq \tilde{b}_j^n 1_{\{\|\tilde{b}_j^n\| \geq n\delta\}}, \quad \lambda_j^n(dz) \doteq \mu_j^n(dz) 1_{\{\|z\| \geq n\delta\}} + \delta_0(dz) \mu_j^n(\{\|z\| < n\delta\}),$$

where $\delta_0(dz)$ is the probability measure with mass 1 at zero. It is not difficult to see that λ_j^n gives the conditional distribution of c_{j+1}^n , whence

$$N^n(j/n) \doteq \frac{1}{n} \sum_{i=0}^{j-1} \left(c_{i+1}^n - \int_{\mathbb{R}^d} z \lambda_i^n(dz) \right)$$

is also a martingale. By a standard submartingale inequality

$$\begin{aligned}& \tilde{P} \left\{ \max_{j=1, \dots, n} \|M^n(j/n) - N^n(j/n)\| \geq \varepsilon \right\} \\ & \leq \frac{1}{\varepsilon^2} \tilde{E} \left[\left\| \frac{1}{n} \sum_{j=0}^{n-1} \left(\tilde{b}_{j+1}^n 1_{\{\|\tilde{b}_{j+1}^n\| < n\delta\}} - \int_{\mathbb{R}^d} z \mu_j^n(dz) 1_{\{\|z\| < n\delta\}} \right) \right\|^2 \right] \\ & = \frac{1}{n^2 \varepsilon^2} \sum_{j=0}^{n-1} \tilde{E} \left[\left\| \tilde{b}_{j+1}^n 1_{\{\|\tilde{b}_{j+1}^n\| < n\delta\}} - \int_{\mathbb{R}^d} z \mu_j^n(dz) 1_{\{\|z\| < n\delta\}} \right\|^2 \right] \\ & \leq \frac{1}{n^2 \varepsilon^2} \sum_{j=1}^n \tilde{E} \left[\left\| \tilde{b}_j^n 1_{\{\|\tilde{b}_j^n\| < n\delta\}} \right\|^2 \right] \\ & \leq \frac{\delta}{n \varepsilon^2} \sum_{j=1}^n \tilde{E} \left[\left\| \tilde{b}_j^n 1_{\{\|\tilde{b}_j^n\| < n\delta\}} \right\| \right]\end{aligned}$$

$$\leq \frac{\delta}{\varepsilon^2} \sum_{j=1}^n \tilde{E} \left[\frac{1}{n} \left\| \tilde{b}_j^n \right\| \right].$$

By Chebyshev's inequality and a conditioning argument,

$$\begin{aligned} \tilde{P} \left\{ \max_{j=1, \dots, n} \|N^n(j/n)\| \geq \varepsilon \right\} &\leq \tilde{P} \left\{ \frac{1}{n} \sum_{j=1}^n \left\| \tilde{b}_j^n \right\| 1_{\{\|\tilde{b}_j^n\| \geq n\delta\}} \geq \varepsilon/2 \right\} \\ &\quad + \tilde{P} \left\{ \frac{1}{n} \sum_{j=0}^{n-1} \int_{\{\|z\| \geq n\delta\}} \|z\| \mu_j^n(dz) \geq \varepsilon/2 \right\} \\ &\leq \frac{2}{\varepsilon} \tilde{E} \left[\frac{1}{n} \sum_{j=1}^n \left\| \tilde{b}_j^n \right\| 1_{\{\|\tilde{b}_j^n\| \geq n\delta\}} \right] \\ &\quad + \frac{2}{\varepsilon} \tilde{E} \left[\frac{1}{n} \sum_{j=0}^{n-1} \int_{\{\|z\| \geq n\delta\}} \|z\| \mu_j^n(dz) \right] \\ &= \frac{4}{\varepsilon} \tilde{E} \left[\frac{1}{n} \sum_{j=1}^n \left\| \tilde{b}_j^n \right\| 1_{\{\|\tilde{b}_j^n\| \geq n\delta\}} \right]. \end{aligned}$$

The last quantity tends to zero as n tends to infinity for each fixed $\delta > 0$ by Lemma 7.3. Sending first $n \rightarrow \infty$ and then $\delta \rightarrow 0$, it follows that for each $\varepsilon > 0$

$$\tilde{P} \left\{ \max_{j=1, \dots, n} \|M^n(j/n)\| \geq 2\varepsilon \right\} \rightarrow 0$$

as $n \rightarrow \infty$. Thus

$$\tilde{X}^n(j/n) - \int_0^{j/n} \int_{\mathbb{R}^d} z \mu^n(dz \times dt) \rightarrow 0$$

uniformly in $j \in \{1, \dots, n\}$, in probability. Using the uniform integrability and weak convergence of μ^n we justify the limit

$$\tilde{X}(t) - \int_0^t \int_{\mathbb{R}^d} z \mu(dz \times ds) = 0$$

for all $t \in [0, 1]$, w.p.1. When combined with the factorization $\mu(dz \times ds) = \mu(dz|s) ds$, this proves the representation for \tilde{X} .

Finally, we discuss the formulas for the limit measures. These all follow easily from analogous properties of the prelimit measures. For example,

consider the random probability measure θ^n . Let g be an arbitrary bounded continuous function on S . By definition,

$$\int_{S \times [0,1]} g(y) [\theta^n]_{1,3}(dy \times dt) = \frac{1}{n} \sum_{j=0}^{n-1} g(\tilde{Y}_j^n) = \frac{1}{n} \sum_{j=0}^{n-1} g(\tilde{Y}_{j+1}^n) + I_n,$$

where I_n is an error term with

$$|I_n| \leq \frac{2}{n} \|g\|_\infty$$

almost surely. Fix arbitrary $\varepsilon > 0$. Let $N_0 \in \mathbb{N}$ be such that $|I_n| \leq \varepsilon/2$ for all $n \geq N_0$. Since ν_j^n is the conditional distribution of \tilde{Y}_{j+1}^n , by Chebyshev's inequality and a conditioning argument, for $n \geq N_0$

$$\begin{aligned} & \tilde{P} \left\{ \left| \int_{S \times [0,1]} g(y) [\theta^n]_{1,3}(dy \times dt) - \int_{S \times [0,1]} g(y) [\theta^n]_{2,3}(dy \times dt) \right| \geq \varepsilon \right\} \\ & \leq \tilde{P} \left\{ \left| \frac{1}{n} \sum_{j=0}^{n-1} \left(g(\tilde{Y}_{j+1}^n) - \int_S g(y) \nu_j^n(dy) \right) \right| \geq \varepsilon/2 \right\} \\ & \leq \frac{4}{\varepsilon^2} \tilde{E} \left[\frac{1}{n^2} \sum_{j=0}^{n-1} \left(g(\tilde{Y}_{j+1}^n) - \int_S g(y) \nu_j^n(dy) \right)^2 \right] \\ & \leq \frac{16 \|g\|_\infty^2}{\varepsilon^2 n}. \end{aligned}$$

By Fatou's Lemma

$$\tilde{P} \left\{ \left| \int_{S \times [0,1]} g(y) [\theta]_{1,3}(dy \times dt) - \int_{S \times [0,1]} g(y) [\theta]_{2,3}(dy \times dt) \right| \geq \varepsilon \right\} = 0.$$

Thus $[\theta]_{1,3} = [\theta]_{2,3}$ almost surely. Since $[\theta^n]_{2,3} = \nu^n$,

$$[\theta]_{1,3}(dy \times dt) = [\theta]_{2,3}(dy \times dt) = \nu(dy \times dt) = \nu(dy|t)dt,$$

which proves $[\theta]_1(dy|t) = [\theta]_2(dy|t) = \nu(dy|t)$.

With regard to the decomposition of γ , an analogous argument shows that, for any $\varepsilon > 0$ and bounded continuous functions g_1, g_2 on S , we have

$$\begin{aligned} 0 & = \lim_{n \rightarrow \infty} \tilde{P} \left\{ \left| \int_{S^2 \times [0,1]} g_1(y_1) g_2(y_2) \gamma^n(dy_1 \times dy_2 \times dt) \right. \right. \\ & \quad \left. \left. - \int_{S^2 \times [0,1]} g_1(y_1) g_2(y_2) \nu^n(dy_1 \times dt) p(y_1, dy_2) \right| \geq \varepsilon \right\} \end{aligned}$$

However, by the Feller property the mapping $y_1 \mapsto \int_S g(y_2)p(y_1, dy_2)$ is bounded and continuous. The decomposition of γ now follows from the weak convergence of γ^n and ν^n , Fatou's Lemma, the arbitrariness of ε , and the fact that product functions are convergence determining (see, for example, [1, Theorem A.3.14]).

The expressions for ζ and η can be proved in the same way, and we omit the proof. \blacksquare

7.4 Analysis of the cost.

We claim that $\liminf_{n \rightarrow \infty} \bar{J}^n$ [see equation (7.5)] is bounded below by

$$\begin{aligned} \tilde{E} \left[R(\theta \parallel \gamma) + R(\zeta \parallel \eta) - \sum_{k=1}^K \int_{[0,1]} \rho_k(\tilde{X}(t), t) H(\bar{\alpha}_k(\tilde{X}(t), t)) dt \right. \\ \left. + \sum_{k=1}^K \int_{\mathbb{R}^d \times [0,1]} \rho_k(\tilde{X}(t), t) \langle \bar{\alpha}_k(\tilde{X}(t), t), z \rangle \mu(dz \times dt) + G(\tilde{X}(1)) \right]. \end{aligned}$$

The bound for the first two relative entropy terms follows from the weak convergence, Fatou's Lemma, and the lower semicontinuity of relative entropy [1, Lemma 1.4.3]. The convergence of the next two terms follows from the weak convergence, the continuity and boundedness properties of the ρ_k and $\bar{\alpha}_k$, and the Dominated Convergence Theorem. Lastly, we show that

$$\liminf_{n \rightarrow \infty} \tilde{E} \left[G(\tilde{X}^n(1)) \right] \geq \tilde{E} \left[G(\tilde{X}(1)) \right]. \quad (7.6)$$

Indeed, by the Lipschitz property of \bar{W} , there exists $C > 0$ such that

$$G(x) \geq -C(\|x\| + 1).$$

By Fatou's Lemma,

$$\liminf_{n \rightarrow \infty} \tilde{E} \left[G(\tilde{X}^n(1)) + C\|\tilde{X}^n(1)\| \right] \geq \tilde{E} \left[G(\tilde{X}(1)) + C\|\tilde{X}(1)\| \right].$$

Since the uniform integrability of $\{\tilde{X}^n(1)\}$ proved in Lemma 7.3 implies $\lim_{n \rightarrow \infty} \tilde{E}\|\tilde{X}^n(1)\| = \tilde{E}\|\tilde{X}(1)\|$, the inequality (7.6) follows.

Using the factorization properties of relative entropy (7.1), we now do some rewriting of the various terms. We have

$$\begin{aligned} R(\theta \parallel \gamma) &= \int_0^1 R(\theta(dy_1 \times dy_2 | t) \parallel \gamma(dy_1 \times dy_2 | t)) dt \\ R(\zeta \parallel \eta) &= \int_0^1 R(\zeta(dy \times dz | t) \parallel \eta(dy \times dz | t)) dt. \end{aligned}$$

However, by Lemma 7.4 $[\theta]_1(dy|t) = [\theta]_2(dy|t) = \nu(dy|t)$, $\gamma(dy_1 \times dy_2|t) = \nu(dy_1|t)p(y_1, dy_2)$, $\eta(dy \times dz|t) = \nu(dy|t)m(dz|y)$, and $\zeta(dy \times dz|t) = \nu(dy|t)q(dz|y, t)$ for some stochastic kernel q . Since

$$\begin{aligned} \int_S \int_{\mathbb{R}^d} zq(dz|y, t)\nu(dy|t) &= \int_{S \times \mathbb{R}^d} z\zeta(dy \times dz|t) \\ &= \int_{\mathbb{R}^d} z[\zeta]_2(dz|t) \\ &= \int_{\mathbb{R}^d} z\mu(dz|t) \\ &\doteq \beta(t), \end{aligned}$$

it follows from the definition of L in (3.2) that

$$R(\theta \|\gamma) + R(\zeta \|\eta) \geq \int_0^1 L(\beta(t)) dt,$$

Moreover, the definition of $\beta(t)$ gives

$$\int_{\mathbb{R}^d \times [0,1]} \langle \bar{\alpha}(\tilde{X}(t), t), z \rangle \mu(dz \times dt) = \int_{[0,1]} \langle \bar{\alpha}(\tilde{X}(t), t), \beta(t) \rangle dt.$$

We thus obtain a lower bound for $\liminf_{n \rightarrow \infty} \bar{J}^n$ in the form

$$\begin{aligned} \Gamma &\doteq \tilde{E} \left[\int_0^1 \sum_{k=1}^K \rho_k(\tilde{X}(t), t) \left[L(\beta(t)) - H(\bar{\alpha}_k(\tilde{X}(t), t)) \right. \right. \\ &\quad \left. \left. + \langle \bar{\alpha}_k(\tilde{X}(t), t), \beta(t) \rangle \right] dt + G(\tilde{X}(1)) \right]. \end{aligned}$$

7.5 Verification.

We now do a classical verification argument to show $\Gamma \geq \bar{W}(0, 0)$. By assumption (see Definition 5.1),

$$\begin{aligned} &\bar{W}_t(\tilde{X}(t), t) + \langle D\bar{W}(\tilde{X}(t), t), \beta(t) \rangle \\ &= \sum_{k=1}^K \rho_k(\tilde{X}(t), t) \left[r_k(\tilde{X}(t), t) + \langle s_k(\tilde{X}(t), t), \beta(t) \rangle \right] \\ &\geq \sum_{k=1}^K \rho_k(\tilde{X}(t), t) \left[L(\beta(t)) + \langle \bar{\alpha}_k(\tilde{X}(t), t), \beta(t) \rangle - H(\bar{\alpha}_k(\tilde{X}(t), t)) \right]. \end{aligned}$$

Integrating both sides from 0 to 1, and using the fact that $\beta(t) = d\tilde{X}(t)/dt$,

$$E \left[\int_0^1 \sum_{k=1}^K \rho_k(\tilde{X}(t), t) \left[L(\beta(t)) + \left\langle \bar{\alpha}_k(\tilde{X}(t), t), \beta(t) \right\rangle - H(\bar{\alpha}_k(\tilde{X}(t), t)) \right] dt \right] \\ \geq \bar{W}(0, 0) - E\bar{W}(\tilde{X}(1), 1)$$

Since $G(x) = \bar{W}(x, 1)$, upon bringing this term to the left hand side we obtain $\Gamma \geq \bar{W}(0, 0)$, thus completing the proof of Theorem 7.1. \blacksquare

8 Appendix

Proof of Lemma 7.3. The proof uses ideas from [1, Proposition 5.3.2]. We start by observing a few facts, namely, that

$$-2G(\tilde{X}_n^n) \leq 2C \left(\frac{1}{n} \sum_{i=1}^n \left\| \tilde{b}_i^n \right\| + 1 \right),$$

that the eigenfunctions $r(y; \alpha)$ are bounded uniformly from above and below away from zero on $\{\alpha : \|\alpha\| \leq C\}$, that $H(\alpha)$ is bounded from below on this set, and that relative entropy is non-negative. These imply the existence of $C_1 < \infty$ and $C_2 < \infty$ such that

$$\sup_n \tilde{E} \left[\frac{1}{n} \sum_{i=0}^{n-1} R\left(\mu_i^n(\cdot) \left\| m(\cdot | \tilde{Y}_{i+1}^n) \right\| \right) - C_1 \frac{1}{n} \sum_{i=1}^n \left\| \tilde{b}_i^n \right\| \right] \leq C_2, \quad (8.1)$$

where the supremum is over the same subsequence as in the statement of the lemma. It follows immediately that $\mu_i^n(\cdot) \ll m(\cdot | \tilde{Y}_{i+1}^n)$ for all $i = 0, \dots, n-1$, with probability one. We can find non-negative, measurable, random functions f_i^n such that f_i^n is a measurable version of $d\mu_i^n(\cdot) / dm(\cdot | \tilde{Y}_{i+1}^n)$. We use the fact that for all $a \geq 0, c \geq 0$, and $\rho \geq 1$,

$$ac \leq e^{\rho a} + \frac{1}{\rho} (c \log c - c + 1).$$

Since $c \log c - c + 1 \geq 0$, it follows that

$$\tilde{E} \left[\frac{1}{n} \sum_{i=1}^n \left\| \tilde{b}_i^n \right\| \right] \leq \tilde{E} \left[\frac{1}{n} \sum_{i=0}^{n-1} \int_{\mathbb{R}^d} \|z\| f_i^n(z) m(dz | \tilde{Y}_{i+1}^n) \right] \\ \leq \tilde{E} \left[\frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{\rho} \int_{\mathbb{R}^d} (f_i^n(z) \log f_i^n(z) - f_i^n(z) + 1) m(dz | \tilde{Y}_{i+1}^n) \right. \\ \left. + \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}^d} e^{\rho \|z\|} m(dz | \tilde{Y}_{i+1}^n) \right].$$

Under Condition 3.1, for each ρ there is a finite and uniform bound $B(\rho)$ on $\int_{\mathbb{R}^d} e^{\rho\|z\|} m(dz|y)$ for all $y \in S$. This allows us to continue the inequality as

$$\tilde{E} \left[\frac{1}{n} \sum_{i=1}^n \left\| \tilde{b}_i^n \right\| \right] \leq B(\rho) + \frac{1}{\rho} \tilde{E} \left[\frac{1}{n} \sum_{i=0}^{n-1} R \left(\mu_i^n(\cdot) \left\| m(\cdot | \tilde{Y}_{i+1}^n) \right\| \right) \right].$$

Choosing $1/\rho = 2C_1$ and rearranging (8.1),

$$\frac{1}{2} \sup_n \tilde{E} \left[\frac{1}{n} \sum_{i=0}^{n-1} R \left(\mu_i^n(\cdot) \left\| m(\cdot | \tilde{Y}_{i+1}^n) \right\| \right) \right] \leq C_2 + B \left(\frac{1}{2C_1} \right).$$

By a very similar argument to that just used, we find

$$\begin{aligned} \tilde{E} \left[\frac{1}{n} \sum_{i=1}^n \left\| \tilde{b}_i^n \right\| 1_{\{\|\tilde{b}_i^n\| \geq C\}} \right] &\leq \sup_{y \in S} \int_{\mathbb{R}^d} 1_{\{\|z\| \geq C\}} e^{\rho\|z\|} m(dz|y) \\ &\quad + \frac{1}{\rho} \tilde{E} \left[\frac{1}{n} \sum_{i=0}^{n-1} R \left(\mu_i^n(\cdot) \left\| m(\cdot | \tilde{Y}_{i+1}^n) \right\| \right) \right]. \end{aligned}$$

Under Condition 3.1,

$$\sup_{y \in S} \int_{\mathbb{R}^d} 1_{\{\|z\| \geq C\}} e^{\rho\|z\|} m(dz|y) \leq e^{-C} \sup_{y \in S} \int_{\mathbb{R}^d} e^{(\rho+1)\|z\|} m(dz|y) \rightarrow 0$$

as $C \rightarrow \infty$. Since we already have a uniform bound on

$$\tilde{E} \left[\frac{1}{n} \sum_{i=0}^{n-1} R \left(\mu_i^n(\cdot) \left\| m(\cdot | \tilde{Y}_{i+1}^n) \right\| \right) \right],$$

the first part of the lemma follows by first sending $C \rightarrow \infty$ and then $\rho \rightarrow \infty$.

We define a piecewise linear process \bar{X}^n by setting

$$\frac{d\bar{X}^n(t)}{dt} = \tilde{b}_i^n \text{ for } t \in \left(\frac{i-1}{n}, \frac{i}{n} \right).$$

Then \bar{X}^n is the piecewise linear interpolation that agrees with \tilde{X}^n at times of the form i/n , and hence if \tilde{X}^n converges in distribution in the sup norm to a limit \tilde{X} then so does \bar{X}^n , since

$$\sup_{0 \leq t \leq 1} \left\| \tilde{X}^n(t) - \bar{X}^n(t) \right\| \rightarrow 0$$

in probability as $n \rightarrow \infty$. Therefore, in order to show the tightness of $\{\tilde{X}^n\}$, it suffices to show that $\{\bar{X}^n\}$ is tight. To this end, define the modulus

$$w^n(\delta) \doteq \sup_{\{s,t \in [0,1]: |t-s| \leq \delta\}} \|\bar{X}^n(t) - \bar{X}^n(s)\|.$$

Tightness of $\{\bar{X}^n\}$ will hold if for each $\varepsilon > 0$ and $\eta > 0$ there is $\delta \in (0, 1)$ such that for all n

$$\tilde{P}\{w^n(\delta) \geq \varepsilon\} \leq \eta.$$

Choose $C < \infty$ such that for all n

$$\tilde{E} \left[\frac{1}{n} \sum_{i=1}^n \|\tilde{b}_i^n\| \mathbf{1}_{\{\|\tilde{b}_i^n\| \geq C\}} \right] \leq \eta\varepsilon/2,$$

and let $\delta \doteq (\varepsilon/2C) \wedge 1$. Then since $C\delta \leq \varepsilon/2$

$$\begin{aligned} \tilde{P}\{w^n(\delta) \geq \varepsilon\} &\leq \tilde{P} \left\{ \sup_{\{s,t \in [0,1]: |t-s| \leq \delta\}} \int_{s \wedge t}^{s \vee t} \left\| \frac{d\bar{X}^n(r)}{dr} \right\| dr \geq \varepsilon \right\} \\ &\leq \tilde{P} \left\{ \sup_{\{s,t \in [0,1]: |t-s| \leq \delta\}} \int_{s \wedge t}^{s \vee t} \left\| \frac{d\bar{X}^n(r)}{dr} \right\| \mathbf{1}_{\{\|\frac{d\bar{X}^n(r)}{dr}\| \geq C\}} dr \geq \varepsilon/2 \right\} \\ &\leq \tilde{P} \left\{ \int_0^1 \left\| \frac{d\bar{X}^n(r)}{dr} \right\| \mathbf{1}_{\{\|\frac{d\bar{X}^n(r)}{dr}\| \geq C\}} dr \geq \varepsilon/2 \right\} \\ &\leq \frac{2}{\varepsilon} \tilde{E} \left[\frac{1}{n} \sum_{i=1}^n \|\tilde{b}_i^n\| \mathbf{1}_{\{\|\tilde{b}_i^n\| \geq C\}} \right] \\ &\leq \eta. \end{aligned}$$

As for the uniform integrability of $\{\tilde{X}^n(1)\}$, observe that for every $C \geq 0$,

$$\|\tilde{X}^n(1)\| \leq \frac{1}{n} \sum_{i=1}^n \|\tilde{b}_i^n\| \leq C + \frac{1}{n} \sum_{i=1}^n \|\tilde{b}_i^n\| \mathbf{1}_{\{\|\tilde{b}_i^n\| \geq C\}}.$$

This implies

$$\begin{aligned} \|\tilde{X}^n(1)\| \mathbf{1}_{\{\|\tilde{X}^n(1)\| \geq 2C\}} &\leq C \mathbf{1}_{\{\|\tilde{X}^n(1)\| \geq 2C\}} + \frac{1}{n} \sum_{i=1}^n \|\tilde{b}_i^n\| \mathbf{1}_{\{\|\tilde{b}_i^n\| \geq C\}} \\ &\leq \frac{\|\tilde{X}^n(1)\|}{2} \mathbf{1}_{\{\|\tilde{X}^n(1)\| \geq 2C\}} + \frac{1}{n} \sum_{i=1}^n \|\tilde{b}_i^n\| \mathbf{1}_{\{\|\tilde{b}_i^n\| \geq C\}}, \end{aligned}$$

or

$$\|\tilde{X}^n(1)\|1_{\{\|\tilde{X}^n(1)\|\geq 2C\}} \leq \frac{2}{n} \sum_{i=1}^n \|\tilde{b}_i^n\| 1_{\{\|\tilde{b}_i^n\|\geq C\}},$$

which in turn implies the uniform integrability of $\{\tilde{X}^n(1)\}$.

The tightness and uniform integrability properties of the random measure $\{\mu^n(dy \times dt)\}$ is easy. Indeed,

$$\begin{aligned} \tilde{E} \left[\int_{\mathbb{R}^d \times [0,1]} \|y\| 1_{\{\|y\|\geq C\}} \mu^n(dy \times dt) \right] &= \tilde{E} \left[\sum_{j=0}^{n-1} \int_{\mathbb{R}^d} \|y\| 1_{\{\|y\|\geq C\}} \mu_j^n(dy) \right] \\ &= \tilde{E} \left[\frac{1}{n} \sum_{i=1}^n \|\tilde{b}_i^n\| 1_{\{\|\tilde{b}_i^n\|\geq C\}} \right]. \end{aligned}$$

Uniform integrability holds since the last quantity tends to zero uniformly in n as $C \rightarrow \infty$, and the tightness is a consequence of the uniform integrability [1, Theorem A.3.17]. ■

References

- [1] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, New York, 1997.
- [2] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. *Stoch. and Stoch. Reports.*, 76:481–508, 2004.
- [3] P. Dupuis and H. Wang. Dynamic importance sampling for uniformly recurrent Markov chains. *Annals of Applied Probab.*, 15:1–38, 2005.
- [4] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling: Examples and numerics. *Preprint*, 2005.
- [5] P. Glasserman and S. Kou. Analysis of an importance sampling estimator for tandem queues. *ACM Trans. Modeling Comp. Simulation*, 4:22–42, 1995.
- [6] P. Glasserman and Y. Wang. Counter examples in importance sampling for large deviations probabilities. *Ann. Appl. Probab.*, 7:731–746, 1997.
- [7] I. Iscoe, P. Ney, and E. Nummelin. Large deviations of uniformly recurrent Markov additive processes. *Adv. Appl. Math.*, 6:373–412, 1985.

- [8] S.R.S. Varadhan. *Large Deviations and Applications*. CBMS-NSF Regional Conference Series in Mathematics. SIAM, Philadelphia, 1984.