# Subsolutions of an Isaacs equation and efficient schemes for importance sampling

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

May 12, 2006

### Abstract

It was established in [6, 7] that importance sampling algorithms for estimating rare-event probabilities are intimately connected with two-person zero-sum differential games and the associated Isaacs equation. This game interpretation shows that *dynamic* or state-dependent schemes are needed in order to attain asymptotic optimality in a general setting. The purpose of the present paper is to show that classical *subsolutions* of the Isaacs equation can be used as a basic and flexible tool for the construction and analysis of efficient dynamic importance sampling schemes. There are two main contributions. The first is a basic theoretical result characterizing the asymptotic performance of importance sampling estimators based on subsolutions. The second is an explicit method for constructing classical subsolutions as a mollification of piecewise affine functions. Numerical examples are included for illustration and to demonstrate that simple, nearly asymptotically optimal importance sampling schemes can be obtained for a variety of problems via the subsolution approach.

[*] 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).

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

# 1 Introduction

It was established in [6, 7] that importance sampling algorithms for estimating rare-event probabilities, or functionals that are largely determined by rare events, are closely related to deterministic differential games. More precisely, the asymptotic optimal performance of importance sampling schemes can be characterized by the value function of a two-person zero-sum differential game, which can in turn be characterized by the solution to the Isaacs equation (a nonlinear PDE) associated with the game. It was also discussed in [6, 7] that one can construct asymptotically optimal importance sampling algorithms based on this solution.

The purpose of the present paper is to explore this connection in further depth. A main new feature is that it is possible to construct efficient importance sampling schemes based on classical *subsolutions* of the Isaacs equation. This leads to a more general class of schemes and lends great flexibility to the designer. We will see that one can often construct subsolutions that are structurally much simpler than the actual *solution*, but which correspond to asymptotically optimal, or at least nearly asymptotically optimal, importance sampling schemes that reflect this simpler structure. This simplicity is important, since one is interested in properties other than just asymptotic optimality, e.g., ease of construction and ease of implementation.

The main theoretical contribution of the paper is a basic result on the asymptotic performance of subsolution-based importance sampling schemes. It characterizes the performance in terms of 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. Another contribution of the current paper is a method for the systematic construction of classical subsolutions that lead to simple yet efficient importance sampling schemes. More precisely, we show that in many cases one can build such subsolutions as a mollification of the minimum of finitely many affine functions.

We wish to point out that the potential application of the subsolution approach 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 expected values involving path dependent events. For the purpose of illustration, we include a few numerical examples that do not fit directly into the theoretical framework of the current paper, and yet for which efficient importance sampling algorithms can still be built via subsolutions.

The paper is organized as follows. Section 2 gives a brief account of importance sampling and asymptotic optimality. 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 3 in the setting of sums of iid random variables. Sections 4 through 8 are the theoretical part of the paper. In Section 4 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 5 reviews the properties of these eigenfunctions. Sections 6 and 7 introduce the concept of generalized subsolution/control and describe the associated importance sampling algorithms. The main theoretical result, which characterizes the asymptotic performance of such schemes, is stated in Section 8. Section 9 discusses methods for the construction of subsolutions in great detail, starting with the simplest possible examples and then extending to more complicated situations. Section 10 is devoted to numerical examples. To illustrate the broad application of the subsolution approach, the latter part of Section 10 considers several classes of problems that are not covered by the main theoretical result, including multi-dimensional level crossing problems, probabilities and expected values that involve path-dependent events, and buffer overflow in a mixed open/closed queueing network. In each case the application of the Isaacs equation and subsolution approach follows the pattern laid out for the simple case of sums of random variables. To streamline the presentation, technical proofs are collected in appendices.

**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 $\sigma$-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. Then $[\mu]_1$ will denote the first marginal of $\mu$, 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 $\mu(dy_1)q(dy_2|y_1) \in \mathcal{P}(S_1 \times S_2)$.

**Relative Entropy.** Given a Polish space $S$ and two probability measures

$\gamma, \mu \in \mathcal{P}(S)$, the *relative entropy* is defined by

$$R(\gamma\|\mu) \doteq \begin{cases} \displaystyle\int_S \log \frac{d\gamma}{d\mu} d\gamma & ; \quad \text{if } \gamma \ll \mu \\ \infty & ; \quad \text{otherwise} \end{cases} .$$

The relative entropy is always non-negative, and is a convex, lower semicontinuous function of $(\gamma, \mu) \in \mathcal{P}(S) \times \mathcal{P}(S)$ [4, Section 1.4].

## 2 An overview of importance sampling

### 2.1 Basics of importance sampling

Importance sampling a variance reduction technique widely used in Monte Carlo simulation. The basic idea of importance sampling is "change of measure." In other words, the system is simulated under a different probability distribution and the outcomes are multiplied by appropriate Radon-Nikodým derivatives to form unbiased estimators.

Suppose we wish to estimate the expected value of a random variable $X$ with distribution $\theta$. Consider an alternative sampling distribution $\mu$ such that $\theta \ll \mu$. Let $f(x) \doteq (d\theta/d\mu)(x)$ denote the Radon-Nikodým derivative. Then importance sampling considers independent copies of a random variable $\bar{X}$ with distribution $\mu$, and forms an estimator by averaging the independent copies of

$$Z \doteq \bar{X} f(\bar{X}).$$

This estimator is unbiased since

$$E[\bar{X} f(\bar{X})] = \int x f(x) \mu(dx) = \int x \theta(dx) = EX,$$

and its rate of convergence is determined by the variance $Z$ — the smaller the variance, the faster the convergence. One typically seeks to minimize the variance of $Z$, or equivalently the second moment of $Z$, within a parametric family of alternative sampling distributions.

**Remark 2.1** In the analysis it is often convenient to write the second moment of $Z$ in terms of the original random variable $X$, that is,

$$EZ^2 = \int x^2 f^2(x) \mu(dx) = \int x^2 f(x) \theta(dx) = E[X^2 f(X)].$$

### 2.2 Asymptotic optimality

The following asymptotic optimality criterion is often adopted when one is interested in a family of random variables $\{X_n\}$ that satisfy

$$\lim_n -\frac{1}{n} \log EX_n = \gamma$$

for some constant $\gamma > 0$. Let $Z_n$ be an unbiased importance sampling estimator for $E[X_n]$. Recall that a major concern of importance sampling is to minimize the second moment of $Z_n$. However, by Jensen's inequality,

$$EZ_n^2 \geq (EZ_n)^2 = (EX_n)^2 .$$

Therefore,

$$\limsup_n -\frac{1}{n} \log EZ_n^2 \leq 2\gamma.$$

We say the importance sampling estimator is *asymptotically optimal* if the upper bound is achieved, i.e., if

$$\liminf_n -\frac{1}{n} \log EZ_n^2 \geq 2\gamma.$$

Sometimes $2\gamma$ is simply referred to as the "optimal decay rate."

## 3 An introduction to the role of subsolutions

This section describes how an Isaacs equation arises in importance sampling, and the implications for the performance of schemes based on subsolutions. Since it is an overview, we do not give all details and will not be precise regarding all necessary assumptions.

### 3.1 Problem formulation for sums of iid random variables

Consider a sequence of iid random variables $\{Y_i, i \in \mathbb{N}\}$ with distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$, and define

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

Let $H$ be the log-moment generating function, that is, for $\alpha \in \mathbb{R}^d$,

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

Assume $H$ is finite for each $\alpha$. Denote by $L$ the Legendre transform

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

Note that both $H$ and $L$ are convex functions.

Suppose one is interested in the importance sampling estimation of

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

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 $\bar\alpha(x,t) : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ recursively define the following quantities. Let $\bar X_0^n = 0$, and assume that $\bar X_j^n, \bar Y_j^n, j = 1,\ldots,i$ have been defined. Let $\bar Y_{i+1}^n$, conditioned on $\bar X_j^n, \bar Y_j^n, j = 1,\ldots,i$, 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 $i = 1,\ldots,n$, the importance sampling estimator is given by

$$Z^n \doteq e^{-nF(\bar X_n^n)} \prod_{i=0}^{n-1} e^{H(\bar\alpha(\bar X_i^n,i/n)) - \left\langle \bar\alpha(\bar X_i^n, i/n), \bar Y_{i+1}^n\right\rangle}.$$

The importance sampling algorithm then takes the sample average of independent replications of $Z^n$ as the estimate. Using a conditioning argument, it is not difficult to check that $Z^n$ is unbiased, and therefore to minimize the variance, it suffices to minimize the second moment of $Z^n$.

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 to 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 $\bar X_j^n, j = i,\ldots,n-1$ as above except $\bar X_i^n = x$, and then define

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

It will be more convenient to express this in terms of the original random variables as in Remark 2.1:

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

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 $\liminf_{n\to\infty} W^n(0,0)$, with larger values indicating better performance.

### 3.2 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

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

A variational formula [4, Section 1.4] shows how to represent exponential integrals in terms of relative entropy. For any bounded and continuous function $f:\mathbb{R}^d\to\mathbb{R}$,

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

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

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

To formally relate $W^n(x,i)$ to the solution of a PDE, suppose that for a smooth function $W:\mathbb{R}^d\times[0,1]\to\mathbb{R}$, $W^n(x,i)\approx W(x,i/n)$. We also use the following relationship [4, Section C.5] between relative entropy and the function $L$ defined previously as the Legendre transform of $H$. For any $\beta\in\mathbb{R}^d$

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

Let $W_t$ denote the partial derivative with respect to $t$, $DW$ the gradient in $x$, and

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

for $s, \alpha, \beta \in \mathbb{R}^d$. We bring $W^n(x,i) \approx W(x,i/n)$ to the right side of the Isaacs equation, expand via Taylor series, insert the expression for $L$, multiply by $n$ and send $n \to \infty$ to get

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

Note that one also expects the terminal condition $W(x,1) = 2F(x)$ to hold.

This PDE, which is also known as an Isaacs equation, was identified in [6, 7] and its solution was used there to construct asymptotically optimal importance sampling schemes.

### 3.3 Subsolutions and importance sampling

The purpose of the present paper is to show that it is only the subsolution property that is essential in the context of importance sampling. The definition of a subsolution simply replaces the equalities that appear in the Isaacs equation and terminal condition with inequalities. More precisely, by a classical subsolution, we mean a continuously differentiable function $\bar{W} : \mathbb{R}^d \times [0,1] \to \mathbb{R}$ 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 $\bar{W}(x,1) \leq 2F(x)$.

The sufficiency of the subsolution property can be understood intuitively as follows. Recall that we are only interested in bounding the quantity $W^n(0,0)$ from *below*, since an upper bound is automatic from Jensen's inequality (see Section 2.2). The inequalities in the definition of a subsolution are those which give lower bounds when the subsolution is combined with a verification argument to estimate the performance.

In a more general context than the one used in this overview, we will show how subsolutions naturally suggest importance sampling schemes. The main theoretical result of this paper can then be roughly stated as follows: If $Z^n$ is the importance sampling estimator constructed according to a subsolution $\bar{W}$, then

$$\liminf_{n \to \infty} -\frac{1}{n} \log E(Z^n)^2 \geq \bar{W}(0,0). \tag{3.1}$$

With result (3.1), the design problem becomes clear: Construct a subsolution whose associated importance sampling scheme can be implemented with reasonable effort and for which $\bar{W}(0,0)$ is equal or close to the optimal decay rate $2\gamma$.

**Remark 3.1** Because subsolutions deal with inequalities (rather than equalities), there is not a unique importance sampling scheme associated with each subsolution. We will turn this flexibility to our advantage, but it requires that the control for the importance sampling player be specified as part of the definition. As a consequence, the notion of a *generalized subsolution/control* will be introduced, which specifies a set of differential inequalities for the given pair. The lower bound (3.1) still holds for importance sampling estimators based on generalized subsolution/controls [see Section 7].

# 4 The general setup

The broader collection of importance sampling problems we wish to analyze includes sums of 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

$$X_n \doteq \frac{1}{n}\sum_{i=1}^{n} b_i\left(Y_i\right). \tag{4.1}$$

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 $Y$. Note that $(Y_n, nX_n)$ forms a Markov additive process.

**Remark 4.1** 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 (4.1) of $X_n$, is that it significantly simplifies our notation in later analysis. However, there is no loss of generality in that all the results in this paper hold if we replace definition (4.1) by the standard one where the summation is taken from $i = 0$ to $n - 1$.

The following conditions are assumed throughout the paper.

**Condition 4.1** *1. There is a reference probability measure $\lambda$ on $S$, a positive integer $m_0$, and $\delta \in (0,1)$, such that*

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

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

2. *The transition kernel $p\left(y_1, dy_2\right)$ satisfies the Feller property, i.e., the mapping $y_1 \mapsto p\left(y_1, dy_2\right)$ 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 4.1 automatically hold when $Y$ is an irreducible, aperiodic, finite state Markov chain.

Under Condition 4.1, $\{X_n, n \in \mathbb{N}\}$ satisfies a large deviation principle with the rate function

$$\begin{aligned} L(\beta) \quad = \quad \inf\Bigg\{ & R\left(\mu\left\|\theta\otimes p\right.\right) + R\left(\theta\otimes\nu\left\|\theta\otimes m\right.\right) \\ & : [\mu]_1 = [\mu]_2 = \theta, \int_S\int_{\mathbb{R}^d} z\nu\left(dz\left|y\right.\right)\theta\left(dy\right) = \beta\Bigg\}. \end{aligned} \tag{4.2}$$

Here $\mu$ is a probability measure on $S \times S$ and $\nu$ is a stochastic kernel on $\mathbb{R}^d$ given $S$. The fact that a large deviation principle holds is proved in [13], 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 [4] which automatically gives this more concrete form of the rate function (4.2). See, in particular, the analogous prelimit representation formula in [4, Section 4.4].

**Remark 4.2** Condition 4.1, especially parts 1 and 4, are strong. However, while the results of the paper hold under weaker conditions, assuming Condition 4.1 helps keep the technicalities to a minimum. The uniform recurrency assumption (part 1 of Condition 4.1) is imposed in order to ensure that the eigenfunction $r(\cdot;\cdot)$ defined in the next section is uniformly positive and uniformly bounded. A technique developed in [2] that combines the "split-chain" technique and modified likelihood-ratios can perhaps be employed to relax this assumption.

# 5 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 $r(\cdot;\alpha) : S \to [0,\infty)$ such that

$$\int_S \int_{\mathbb{R}^d} e^{\langle \alpha, z\rangle} r(y;\alpha) m\left(dz \left| y\right.\right) p(x,dy) = e^{G(\alpha)} r(x;\alpha). \tag{5.1}$$

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

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

and is again a convex function.

The needed properties of the solution to this problem are summarized in the following lemma [13, Section 3].

**Lemma 5.1** *Assume Condition 4.1. 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 each $y \in S$ there is $M < \infty$ such that $|r(y;\alpha_1) - r(y;\alpha_2)| \leq M|\alpha_1 - \alpha_2|$ for all $\alpha_1, \alpha_2 \in K$ and $y \in S$.*

For each $\alpha \in \mathbb{R}^d$ and each $y_1 \in S$, it follows from equation (5.1), the strict positiveness of $r(\cdot;\alpha)$, and $G(\alpha) = H(\alpha)$, that

$$P\left(y_1, dy_2, dz; \alpha\right) \doteq e^{\langle \alpha, z\rangle - H(\alpha)} \cdot \frac{r(y_2;\alpha)}{r(y_1;\alpha)} \cdot p(y_1, dy_2) \cdot m\left(dz \left| y_2\right.\right) \tag{5.2}$$

defines a probability measure on $S \times \mathbb{R}^d$.

**Remark 5.1** When $S$ is a single point, $X_n$ reduces to the average of iid random variables with distribution $m(dz) \equiv m(dz|y)$. In this case $r(\cdot;\alpha) \equiv 1$ and the change of measure (5.2) reduces to the "exponential tilt"

$$P(dz;\alpha) = e^{\langle \alpha, z\rangle - H(\alpha)} m\,(dz)\,,$$

where $H$ is the log-moment generating function for $m(dz)$.

# 6 The Isaacs equation and subsolutions

Suppose that one wishes to estimate the expected values of certain functionals of $X_n$ for large $n$, using importance sampling schemes based on changes of measure of the form (5.2). Analogous to Section 2, the PDE associated with this problem is the Isaacs equation

$$W_t + \sup_{\alpha\in\mathbb{R}^d} \inf_{\beta\in\mathbb{R}^d} \mathbb{H}(DW;\alpha,\beta) = 0 \tag{6.1}$$

with suitable terminal conditions (which depend on the functionals of interest). Here $W : \mathbb{R}^d \times [0,1] \to \mathbb{R}$, $W_t$ denotes the partial derivative with respect to $t$, $DW$ the gradient in $x$, and

$$\mathbb{H}(s;\alpha,\beta) \doteq \langle s,\beta\rangle + L(\beta) + \langle \alpha,\beta\rangle - H(\alpha) \tag{6.2}$$

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

In order to construct good importance sampling schemes, one does not need the solution to the Isaacs equation. It turns out that finding a good subsolution is often sufficient. Indeed, we will introduce a slightly more complicated notion of *generalized subsolution/control*, which is very convenient for the study of importance sampling algorithms. Its connection with classical subsolution will be discussed once we give the definition.

**Definition 6.1** *Given $K \in \mathbb{N}$, consider functions $\bar{W} : \mathbb{R}^d \times [0,1] \to \mathbb{R}$, $\rho_k : \mathbb{R}^d \times [0,1] \to \mathbb{R}$, $\bar{\alpha}_k : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$, $1 \le k \le K$. We say the collection $(\bar{W}, \rho_k, \bar{\alpha}_k)$ is a* generalized subsolution/control *to the Isaacs equation (6.1) if the following conditions hold. $\{\rho_k\}$ is a partition of unity, i.e., $\rho_k \ge 0$ and*

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

*for all $(x,t) \in \mathbb{R}^d \times [0,1]$. $\bar{W}_t$ and $D\bar{W}$ have representations*

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

*and for each $k = 1, \ldots, K$*

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

*The functions $(r_k, s_k, \rho_k, \bar{\alpha}_k)$ are uniformly bounded and Lipschitz continuous.*

For any generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$, it is not difficult to show that

$$\bar{W}_t + \sup_{\alpha\in\mathbb{R}^d} \inf_{\beta\in\mathbb{R}^d} \mathbb{H}(D\bar{W}; \alpha, \beta) \geq 0. \tag{6.3}$$

In other words, $\bar{W}$ is a classical subsolution to the Isaacs equation (6.1). It will turn out that *only* the $(\rho_k, \bar{\alpha}_k)$-component will be used to define the change of measure used in importance sampling (see the next section), and so we use terminology "subsolution/control." We will also use the term *subsolution* to refer to the $\bar{W}$ component alone.

**Remark 6.1** For the special case where $K = 1$, and with notation $\bar{\alpha} = \bar{\alpha}_1$, we simply write $(\bar{W}, \bar{\alpha})$ and call it a *subsolution/control pair*.

**Remark 6.2** Suppose that $\bar{W}$ is a classical subsolution to the Isaacs equation, that is, $\bar{W}$ satisfies inequality (6.3). Let $\alpha^*(x,t)$ be the supremizer [indeed, one can easily identify $\alpha^*(x,t) = -D\bar{W}(x,t)/2$]. Then $(\bar{W}, \alpha^*)$ is a subsolution/control pair, provided that $\alpha^*$ is uniformly bounded and Lipschitz continuous.

# 7 Importance sampling based on subsolutions

In this section, we describe the importance sampling algorithms associated with a given generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$. We recall $P(y_1, dy_2, dz; \alpha)$ as in (5.2) defines a probability measure on $S \times \mathbb{R}^d$ for each $\alpha \in \mathbb{R}^d$ and $y_1 \in S$. These probability measures, the weights $\rho_k$, and the functions $\bar{\alpha}_k$ will be used to construct the importance sampling scheme.

For fixed $n$, define for $j = 0, 1, \ldots, n$

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

Processes $\bar{X}^n_j, \bar{Y}^n_j$, and $\bar{b}^n_j$, analogous to $X_j, Y_j$, and $b_j(Y_j)$, are constructed recursively as follows. Let $\bar{X}^n_0 = 0$ and $\bar{Y}_0 = Y_0 = y$. Suppose that $\bar{X}^n_j = x$

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

$$\sum_{k=1}^{K} \rho_{k,j}^n(x) P\left(y_1, dy_2, dz; \bar{\alpha}_{k,j}^n(x)\right). \tag{7.1}$$

In other words, one first simulates a multinomial random variable $I$ taking values in $\{1, 2, \ldots, K\}$ such that $P\{I = k\} = \rho_{k,j}^n(x)$, and then simulates $(\bar{Y}_{j+1}^n, \bar{b}_{j+1}^n)$ from the distribution $P(y_1, dy_2, dz; \bar{\alpha}_{I,j}^n(x))$. Finally, update the state dynamics by letting

$$\bar{X}_{j+1}^n \doteq \bar{X}_j^n + \bar{b}_{j+1}^n/n.$$

An unbiased importance sampling estimator can then be obtained by multiplying the functional of interest with the Radon-Nikodým derivative (i.e., likelihood ratio). For example, suppose that we are interested in estimating $E_y \exp\{-nF(X_n)\}$ for some function $F$. Then the unbiased importance sampling estimator is

$$\begin{aligned} Z^n \quad \doteq \quad & e^{-nF(\bar{X}_n^n)} \prod_{j=0}^{n-1} \left[ \sum_{k=1}^{K} \rho_{k,j}^n(\bar{X}_j^n) \cdot e^{\left\langle \bar{\alpha}_{k,j}^n(\bar{X}_j^n), \bar{b}_{j+1}^n \right\rangle - H\left(\bar{\alpha}_{k,j}^n(\bar{X}_j^n)\right)} \right. \\ & \qquad\qquad \left. \cdot \frac{r(\bar{Y}_{j+1}^n; \bar{\alpha}_{k,j}^n(\bar{X}_j^n))}{r(\bar{Y}_j^n; \bar{\alpha}_{k,j}^n(\bar{X}_j^n))} \right]^{-1}. \end{aligned} \tag{7.2}$$

**Remark 7.1** In most of the applications considered in this paper, one can construct a generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$ where the $\bar{\alpha}_k$ are all constants and with $K$ of moderate size. This has a distinct advantage in numerical implementation. For example, to compute a change of measure one often needs to numerically solve the eigenvalue/eigenvector problem (5.1). If $\bar{\alpha}_k$ is not a constant, one needs to solve eigenvalue/eigenvector problems over and over again, depending on the current state of the simulation. This could become computationally demanding.

# 8 Statement of the main result

In this section we present the main theoretical result, which is an asymptotic bound on the second moment for the 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.

Suppose that we wish to numerically approximate the quantity

$$E \exp \{-nF(X_n)\} \tag{8.1}$$

for a Borel measurable function $F : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Given a generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$, the corresponding importance sampling estimator $Z^n$ is given by (7.2).

**Theorem 8.1** *Assume that Condition 4.1 holds and that*

$$\bar{W}(x, 1) \leq 2F(x) \tag{8.2}$$

*for every $x \in \mathbb{R}^d$. Then*

$$\liminf_{n\to\infty} -\frac{1}{n} \log E\left[(Z^n)^2\right] \geq \bar{W}(0,0).$$

The proof of this theorem is a combination of weak convergence and a verification argument, and is deferred to Appendix A.

Under various sets of regularity conditions on $F$, one has the large deviation asymptotic approximation [16, 4]

$$\gamma \doteq \lim_{n} -\frac{1}{n} \log E \exp\{-nF(X_n)\} = \inf_{\beta\in\mathbb{R}^d} [F(\beta) + L(\beta)]. \tag{8.3}$$

Thanks to the discussion in Section 2.2, for a generalized subsolution/control that satisfies the terminal condition (8.2), the corresponding importance sampling scheme is asymptotically optimal or nearly asymptotically optimal if $\bar{W}(0,0)$ is equal or close to the optimal decay rate $2\gamma$.

**Remark 8.1** As noted in the Introduction, the use of subsolutions is applicable in much broader settings including, for example, path-dependent events and systems with constrained dynamics (e.g., queuing networks [5]). In these cases, depending on the class of changes of measure used and the dynamics of the system, the Isaacs equation may take different forms. Moreover, it may be required that the subsolution satisfy certain boundary conditions besides terminal conditions such as (8.2). However, it is always the case that the use of subsolutions is critical to the construction of importance sampling schemes and asymptotic results analogous to Theorem 8.1.

# 9 Construction of generalized subsolution/controls

To illustrate how one constructs generalized subsolution/controls, we assume that the quantity of interest is (8.1) and the large deviation limit (8.3) holds.

There are several concerns in the construction. To begin, the terminal condition (8.2), or $\bar{W}(x,1) \leq 2F(x)$, must be satisfied in order for Theorem 8.1 to be valid. Secondly, for optimality or near optimality, one wishes $\bar{W}(0,0)$ to be equal or close to the optimal decay rate $2\gamma$. Finally, one would like the controls $(\rho_k, \bar{\alpha}_k)$ to take simple forms, since this leads to importance sampling algorithms that are simpler and easier to implement. Our construction can be roughly divided into the following steps.

1. *Identification of affine subsolution/control pairs as the building block.* We identify a family of particularly simple subsolution/control pairs to the Isaacs equation (6.1). For each of these pairs, say $(\bar{W}, \bar{\alpha})$, $\bar{W}$ is affine in $(x,t)$ and $\bar{\alpha}$ takes a constant value. This family, denoted from now on by $\mathbb{A}$, contains the building blocks for our construction. It is appropriate for the class of problems under consideration, where the Hamiltonian $\mathbb{H}$ does not depend on state $x$.

2. *Construction of piecewise affine subsolutions.* It is often possible to take a finite collection of pairs in $\mathbb{A}$, say $\{(\bar{W}_k, \bar{\alpha}_k), k = 1, 2, \ldots, K\}$, so that $\wedge_{k=1}^{K} \bar{W}_k$, the pointwise minimum of $\{\bar{W}_k\}$, satisfies the terminal condition (8.2) and $\wedge_{k=1}^{K} \bar{W}_k(0,0)$ is equal or close to $2\gamma$. Since each $\bar{W}_k$ is a classical subsolution, $\wedge_{k=1}^{K} \bar{W}_k$ is a weak sense subsolution, but not a classical subsolution (except when $K = 1$).

3. *Obtaining generalized subsolution/controls through mollification.* A generalized subsolution/control can be obtained as a simple and easily implemented mollification of $\wedge_{k=1}^{K} \bar{W}_k$ when $K \geq 2$.

## 9.1 Affine solutions to the Isaacs equation

In this section we identify $\mathbb{A}$, the collection of affine subsolution/control pairs to the Isaacs equation (6.1).

For any given $\bar{\alpha} \in \mathbb{R}^d$ and $\bar{c} \in \mathbb{R}$, consider the affine function

$$\bar{W}(x,t) = -2\langle \bar{\alpha}, x \rangle + \bar{c} - 2(1-t)H(\bar{\alpha}).$$

We claim that $(\bar{W}, \bar{\alpha})$ is a subsolution/control pair. Indeed, thanks to the convex conjugacy between $H$ and $L$,

$$\begin{aligned}
&\bar{W}_t + \inf_{\beta\in\mathbb{R}^d} \mathbb{H}(D\bar{W};\bar{\alpha},\beta) \\
&\quad = \bar{W}_t + \inf_{\beta\in\mathbb{R}^d} \left[\left\langle D\bar{W},\beta\right\rangle + L(\beta) + \langle\bar{\alpha},\beta\rangle - H(\bar{\alpha})\right] \\
&\quad = H(\bar{\alpha}) + \inf_{\beta\in\mathbb{R}^d} [L(\beta) - \langle\bar{\alpha},\beta\rangle] \\
&\quad = 0.
\end{aligned}$$

Let $\mathbb{A}$ be the collection of all such pairs, that is

$$\mathbb{A} \doteq \left\{(\bar{W},\bar{\alpha}) : \bar{W} = -2\langle\bar{\alpha},x\rangle + \bar{c} - 2(1-t)H(\bar{\alpha}), \bar{\alpha}\in\mathbb{R}^d, \bar{c}\in\mathbb{R}\right\}.$$

**Remark 9.1** It is not difficult to show that for every $(\bar{W},\bar{\alpha})\in\mathbb{A}$, the affine function $\bar{W}$ is indeed a *solution* to the Isaacs equation (6.1).

## 9.2 Piecewise affine viscosity subsolutions

As mentioned above, the technique used to construct a generalized subsolution/control requires finding a finite collection of affine subsolution/control pairs in $\mathbb{A}$ such that their minimum satisfies the appropriate terminal condition and takes a large value at $(0,0)$ [preferably $2\gamma$, the optimal decay rate]. We begin in this section with the simplest examples.

### 9.2.1 Example: Estimating $P\{X_n \in A\}$

Consider the special case where one wishes to estimate $P\{X_n\in A\}$ for some Borel set $A\subset\mathbb{R}^d$. This is obtained by letting $F(x)=0$ if $x\in A$ and $F(x)=\infty$ if $x\notin A$. Therefore the terminal condition (8.2) amounts to

$$\bar{W}(x,1)\leq 0 \qquad \text{for } \ x\in A. \tag{9.1}$$

Throughout this section we assume

$$\inf_{\beta\in A^\circ} L(\beta) = \inf_{\beta\in\bar{A}} L(\beta)\in(0,\infty),$$

where $A^\circ$, $\bar{A}$ denote the interior and the closure of $A$, respectively. It follows that

$$\gamma = \inf_{\beta\in A} L(\beta) = \inf_{\beta\in A^\circ} L(\beta) = \inf_{\beta\in\bar{A}} L(\beta).$$

Let $\beta^* \in \bar{A}$ be a minimizer of $L$ over $\bar{A}$. Denote by $\alpha^*$ the convex conjugate of $\beta^*$, that is,

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

Case 1. Consider the simplest case where $A$ is convex. Thanks to the convexity of $\bar{A}$, the vector $-\alpha^*$ defines an outward normal of $\bar{A}$, whence

$$A \subset \{x \in \mathbb{R}^d : \langle x, \alpha^* \rangle \geq \langle \beta^*, \alpha^* \rangle\}. \tag{9.2}$$

Consider the element of $\mathbb{A}$ with $\bar{\alpha} = \alpha^*$ and $\bar{c} = 2\langle \beta^*, \alpha^* \rangle$, i.e.,

$$\bar{W}(x,t) = -2\langle \alpha^*, x \rangle + 2\langle \beta^*, \alpha^* \rangle - 2(1-t)H(\alpha^*).$$

It is easy to check $\bar{W}(x,1) \leq 0$ for each $x \in A$, thanks to (9.2). Therefore, when $A$ is convex, $(\bar{W}, \alpha^*)$ provides a simple subsolution/control pair that satisfies the terminal condition (9.1). The value at $(0,0)$ is

$$\bar{W}(0,0) = 2\langle \beta^*, \alpha^* \rangle - 2H(\alpha^*) = 2L(\beta^*) = 2\gamma,$$

the optimal decay rate. Note that the analysis holds if we replace the convexity assumption on $A$ by the assumption that (9.2) holds.

Case 2. More generally, suppose that for some $K \in \mathbb{N}$,

$$A \subset \cup_{k=1}^{K} \{x : \langle x, \alpha_k \rangle \geq \langle \beta_k, \alpha_k \rangle\} \tag{9.3}$$

where $\beta_k$ and $\alpha_k$ are convex conjugates, and that $L(\beta_k) \geq \gamma$ for each $k$. A necessary and sufficient condition for these two assumptions to hold is that $A$ should be contained in the union of a finite number of half-spaces, and that the infimum of $L$ on each of these half spaces is at least $\gamma$. In this case $\beta_k$ can be taken as the point on the $k$th half space that minimizes $L$, and we have $\gamma = L(\beta_k)$ for some $k$. Several of the numerical examples in the paper will fall into this category.

Define an affine subsolution/control pair $(\bar{W}_k, \alpha_k)$ by

$$\bar{W}_k(x,t) \doteq -2\langle \alpha_k, x \rangle + 2\langle \alpha_k, \beta_k \rangle - 2(1-t)H(\alpha_k)$$

for each $k = 1, \ldots, K$. Consider the pointwise minimum

$$\bar{W}(x,t) \doteq \wedge_{k=1}^{K} \bar{W}_k(x,t).$$

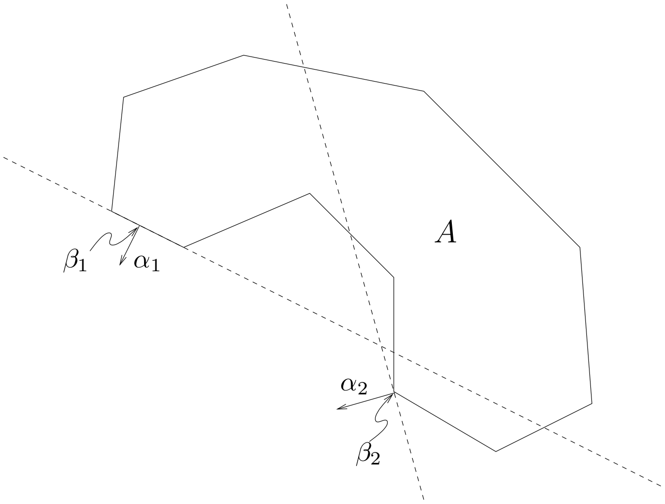

Figure 1: Example of a non-convex $A$ with a two-piece subsolution.

As we have pointed out, $\bar{W}$ defines a weak sense subsolution to the Isaacs equation. The terminal condition (9.1) is satisfied, since for each $x \in A$

$$\bar{W}(x,1) = \wedge_{k=1}^{K} \bar{W}_k(x,1) = \wedge_{k=1}^{K} \left[-2\left\langle x,\alpha_k\right\rangle + 2\left\langle \beta_k,\alpha_k\right\rangle\right] \leq 0$$

by (9.3). Finally, we observe that

$$\bar{W}(0,0) = \wedge_{k=1}^{K} \left[2\left\langle \beta_k,\alpha_k\right\rangle - 2H(\alpha_k)\right] = 2\wedge_{k=1}^{K} L(\beta_k) = 2\gamma.$$

The last equality holds since $L(\beta_k) \geq \gamma$ for each $k$ and with equality for some $k$.

### 9.2.2 Example: Estimating $E\exp\{-nF(X_n)\}$

The development here parallels the probability case as described in the previous section. For simplicity, we assume that there exists $\beta^*$ that minimizes $L(\beta)+F(\beta)$ over $\beta \in \mathbb{R}^d$, and let $\alpha^*$ be its convex conjugate. To avoid technicalities, we assume that $L$ is differentiable at $\beta^*$, and thus $\alpha^* = DL(\beta^*)$.

Case 1. We first consider the simplest case where $F$ is convex. Consider an affine subsolution/control pair $(\bar{W},\alpha^*)$ where

$$\bar{W}(x,t) \doteq -2\left\langle \alpha^*,x\right\rangle + 2[F(\beta^*) + \left\langle \alpha^*,\beta^*\right\rangle] - 2(1-t)H(\alpha^*).$$

Since $\beta^*$ is a minimizer of $L(\beta) + F(\beta)$, we have $0 \in \partial(L + F)(\beta^*)$, where $\partial$ denotes the set of subdifferentials. Therefore

$$-\alpha^* = -DL(\beta^*) \in \partial F(\beta^*).$$

It follows that the affine function $\bar{W}(x, 1)$ is a supporting hyperplane to $2F$ at $\beta^*$, and hence

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

for every $x$, i.e., the terminal condition (8.2) holds. Also observe that

$$\bar{W}(0, 0) = 2F(\beta^*) + 2\langle \alpha^*, \beta^* \rangle - 2H(\alpha^*) = 2F(\beta^*) + 2L(\beta^*) = 2\gamma.$$

Case 2. Next suppose that $F$ is no longer convex. If there exists a convex function $G$ such that $G \leq F$, $G(\beta^*) = F(\beta^*)$, and

$$\inf_{\beta \in \mathbb{R}^d} [L(\beta) + G(\beta)] = \inf_{\beta \in \mathbb{R}^d} [L(\beta) + F(\beta)] = \gamma,$$

then we reduce to the previous case. More generally, suppose there exist convex functions $G_k, k = 1, \ldots, K$, such that

$$\wedge_{k=1}^{K} G_k \leq F, \tag{9.4}$$

and for each $k$,

$$\inf_{\beta \in \mathbb{R}^d} [L(\beta) + G_k(\beta)] \geq \inf_{\beta \in \mathbb{R}^d} [L(\beta) + F(\beta)]. \tag{9.5}$$

If each $G_k$ is bounded from below and lower semicontinuous then a minimizer $\beta_k$ of $L(\beta) + G_k(\beta)$ will exist, and we can define the weak sense subsolution

$$\bar{W}(x, t) \doteq \wedge_{k=1}^{K} \bar{W}_k(x, t),$$

where $(\bar{W}_k, \alpha_k)$ is the affine subsolution/control pair with

$$\bar{W}_k(x, t) \doteq -2\langle \alpha_k, x \rangle + 2[G_k(\beta_k) + \langle \alpha_k, \beta_k \rangle] - 2(1 - t)H(\alpha_k).$$

The same argument as in Case 1 shows that the terminal condition $\bar{W}(x, 1) \leq 2F(x)$ is satisfied, and we have

$$\bar{W}(0, 0) = \wedge_{k=1}^{K} [L(\beta_k) + G(\beta_k)] \geq 2\gamma.$$

Actually, the equality holds, since (9.4) implies (9.5) must hold as an equality for some $k$.

### 9.3 Mollification

As discussed previously, once a weak sense subsolution is identified as the pointwise minimum of a collection of affine subsolution/control pairs, mollification is used to produce generalized subsolution/controls.

Let $(\bar{W}_k, \alpha_k) \in \mathbb{A}, k = 1, 2 \ldots, K$, be affine subsolution/control pairs. We use a standard numerical approximation which we call *exponential weighting* for $\bar{W}(x,t) = \wedge_{k=1}^{K} \bar{W}_k(x,t)$. Let $\delta$ be a small positive number, and define

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

For $1 \leq i \leq K$, let

$$\rho_i^{\delta}(x,t) \doteq \frac{e^{-\frac{1}{\delta}\bar{W}_i(x,t)}}{\sum_{k=1}^{K} e^{-\frac{1}{\delta}\bar{W}_k(x,t)}}.$$

Then one can easily verify

$$D\bar{W}^{\delta}(x,t) = \sum_{k=1}^{K} \rho_k^{\delta}(x,t) D\bar{W}_k(x,t)$$

and

$$\bar{W}_t^{\delta}(x,t) = \sum_{k=1}^{K} \rho_k^{\delta}(x,t) \left(\bar{W}_k\right)_t (x,t),$$

and so $\bar{W}^{\delta}$ takes the form prescribed for a generalized subsolution/control with $\bar{\alpha}_k(x,t) \equiv \alpha_k$. It is obvious that $s_k(x,t) \doteq D\bar{W}_k(x,t) = -2\alpha_k$, $r_k(x,t) \doteq \left(\bar{W}_k\right)_t (x,t) = 2H(\alpha_k)$, and $\bar{\alpha}_k(x,t)$ are all uniformly bounded and Lipschitz continuous, and it is easy to check that the same is true with regard to $\rho_k^{\delta}(x,t)$ for each fixed $\delta > 0$. Therefore, $(\bar{W}^{\delta}, \rho_k^{\delta}, \bar{\alpha}_k)$ is a generalized subsolution/control.

For a fixed $(x,t)$, since $\bar{W}(x,t) = \wedge_{k=1}^{K} \bar{W}_k(x,t)$, it follows easily that

$$e^{-\frac{1}{\delta}\bar{W}(x,t)} \leq \sum_{k=1}^{K} e^{-\frac{1}{\delta}\bar{W}_k(x,t)} \leq K e^{-\frac{1}{\delta}\bar{W}(x,t)},$$

which implies

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

Thus if $\bar{W}$ satisfies a given terminal condition then so will $\bar{W}^{\delta}$, though there may be a small reduction of the value at $(0,0)$.

It is important to observe is that the subsolution $\bar{W}^{\delta}$ itself does not play any explicit role in the computation of the change of measure and the algorithm is completely determined by $(\rho_k^{\delta}, \bar{\alpha}_k)$. However, the function $\bar{W}^{\delta}$ characterizes the performance of the corresponding importance sampling algorithm through results such as Theorem 8.1.

**Remark 9.2** Recall that mollification will possibly result in a small reduction in the value at $(0,0)$, which will lead to the strict inequality $\bar{W}^{\delta}(0,0) < 2\gamma$. It is worth noting that one can construct a sequence of schemes indexed by $n$ which achieves the theoretical bound $2\gamma$ on performance if one lets $\delta \to 0$ as $n \to \infty$ in an appropriate way. We will not pursue this issue here, since our computational experience suggests it is not needed to obtain good performance. However, the interested reader can consult [5] for the precise statement and further details in the context of stochastic networks.

**Remark 9.3** Exponential weighting is not the only way to achieve mollification. For example, one can mollify $\bar{W}(x,t) = \wedge_{k=1}^{K} \bar{W}_k(x,t)$ by integration against a smooth convolution kernel, for which a standard choice is

$$\eta(x) \doteq \begin{cases} C\exp\{1/(\|x\|^2 - 1)\}, & \text{if } \|x\| < 1, \\ \qquad 0 \qquad\qquad\quad , & \text{if } \|x\| \geq 1, \end{cases}$$

where $C$ is the normalizing constant so that the integral of $\eta$ over $\mathbb{R}^d$ is one [9, Section 7.2]. However, we do not recommend this method since in this case the weights $\{\rho_k^{\delta}(x,t)\}$ involve integrations that can be computationally demanding. In contrast, the weights are very easy to compute when using the exponential weighting mollification. Furthermore, the resulting importance sampling schemes based on these two mollifications will yield very similar estimates and standard errors (for the same sample size) even though the scheme based on exponential weighting is much faster.

### 9.4 Discussion

The construction of generalized subsolution/controls can be extended in many ways and to many other situations. For example, when computing escape probabilities of a stochastic network one is particularly interested in combining subsolutions so as to satisfy appropriate boundary conditions [5]. In other problems, one may wish to expand $\mathbb{A}$ so that it also includes $(\bar{W}, \bar{\alpha})$ where $\bar{W}$ is a strict subsolution to the Isaacs equation, or even a non-affine subsolution of some specific form. However, the basic structure of construction remains the same: We identify a class of subsolution/control pairs which correspond to changes of measures of simple form, and use these pairs as the building blocks for generalized subsolution/controls.

## 10 Examples of importance sampling algorithms

In this section we give examples of importance sampling algorithms based on subsolutions. Some of these examples are not covered by the theoretical framework for Theorem 8.1 and are included to demonstrate the broad applicability of the subsolution approach. Unless specified otherwise, the importance sampling algorithm based on a generalized subsolution/control will follow the description in Section 7 with the sampling distribution determined by (7.1) and (5.2) [see also Remark 5.1]. To ease exposition, when applying (7.1) we drop the superscript $n$.

Two points on the performance should be made. The first is that when mollification was used, no special tuning of the mollification parameter $\delta$ was needed. The second is that very good performance across a range of problems formulations and functionals was obtained with $20,000$ samples.

### 10.1 Example: Estimating $P\{X_n \in A\}$ for convex $A$

Assume that $\{Y_1, Y_2, \ldots\}$ is a sequence of iid 2-dimensional $N(0, I_2)$ random variables, where $I_2$ is the $2 \times 2$ identity matrix. Let

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

and consider the estimation of $P\{X_n \in A\}$ for a convex set $A$ of the form

$$A = \left\{x \in \mathbb{R}^2 : (x-a)^2 + y^2 \leq r^2\right\},$$

with $0 < r < a$. For this model, $H(\alpha) = \|\alpha\|^2/2$, $L(\beta) = \|\beta\|^2/2$, and

$$\gamma = \lim_{n\to\infty} -\frac{1}{n}\log P\{X_n \in A\} = \inf_{\beta \in A} L(\beta) = \frac{1}{2}(a-r)^2,$$

with the minimizing $\beta^* = (a-r, 0)$. The convex conjugate of $\beta^*$ is just $\alpha^* = (a-r, 0)$.

As discussed in Section 9.2.1, the affine subsolution/control pair $(\bar{W}, \alpha^*)$

$$\bar{W}(x,t) = -2\langle \alpha^*, x\rangle + 2\langle \beta^*, \alpha^*\rangle - 2(1-t)H(\alpha^*)$$

satisfies the terminal condition (9.1) and $\bar{W}(0,0) = 2L(\beta^*) = 2\gamma$. It is not difficult to check that $(\bar{W}, \alpha^*)$ induces a very simple importance sampling scheme under which $\{\bar{Y}_i\}$ are iid with distribution $N(\alpha^*, I_2)$. By Theorem 8.1 this scheme is asymptotically optimal.

The table below gives numerical results. We take $a = 2$, $r = 1$, and run simulations for the cases $n = 25, 50, 100$. Each estimate consists of 20,000 replications. The theoretical value (which is available from the standard statistics software S-plus) is presented for comparison. The standard error is also a numerical estimate, and C.I. stands for "confidence interval," though this is only formal since we make no assertion regarding normality of errors.

| | $n = 25$ | $n = 50$ | $n = 100$ |
|---|---|---|---|
| Theoretical value | $1.99 \times 10^{-7}$ | $5.39 \times 10^{-13}$ | $5.36 \times 10^{-24}$ |
| Estimate | $2.00 \times 10^{-7}$ | $5.48 \times 10^{-13}$ | $5.44 \times 10^{-24}$ |
| Standard Error | $0.04 \times 10^{-7}$ | $0.12 \times 10^{-13}$ | $0.14 \times 10^{-24}$ |
| 95% C.I. | $[1.92, 2.08] \times 10^{-7}$ | $[5.24, 5.72] \times 10^{-13}$ | $[5.16, 5.72] \times 10^{-24}$ |

Table 1. Estimating $P\{X_n \in A\}$ for convex $A$.

**Remark 10.1** This algorithm based on $(\bar{W}, \alpha^*)$ coincides with the importance sampling based on what one might call the "standard heuristic," which states that the change of measure used in the analysis of the large deviation lower bound is a good choice for importance sampling. As demonstrated in [11, 12, 6, 7], the standard heuristic importance sampling is efficient only in very special situations.

## 10.2 Example: Estimating $P\{X_n \in A\}$ for non-convex $A$

In this section, we give numerical estimates of $P\{X_n \in A\}$ when $A \subset \mathbb{R}^d$ takes the form

$$A \subset \{x : \langle x, \alpha_1 \rangle \geq \langle \beta_1, \alpha_1 \rangle\} \cup \{x : \langle x, \alpha_2 \rangle \geq \langle \beta_2, \alpha_2 \rangle\}, \tag{10.1}$$

with $\alpha_k \in \mathbb{R}^d$ and $\beta_k \in \mathbb{R}^d$ convex conjugates for $k = 1, 2$.

As discussed in Section 9.2.1, construct affine subsolution/control pairs $(\bar{W}_k, \alpha_k)$ with

$$\bar{W}_k(x, t) \doteq -2\langle \alpha_k, x \rangle + 2\langle \alpha_k, \beta_k \rangle - 2(1 - t)H(\alpha_k)$$

and let $\bar{W}(x, t) \doteq \bar{W}_1(x, t) \wedge \bar{W}_2(x, t)$. Then $\bar{W}$ is a weak sense subsolution which satisfies the terminal condition (9.1). Furthermore, $\bar{W}(0, 0) = 2\gamma$ if $L(\beta_k) \geq \gamma$ for each $k$ (which indeed is the case for both numerical examples in this section). A generalized subsolution/control can then be obtained via exponential weighting mollification.

We will present two numerical examples: one for one-dimensional iid normal random variables, the other for a two-dimensional finite state Markov chain. Both examples have already appeared [6, 7]. The difference is that in these papers the importance sampling algorithm was based on the *solution* of the Isaacs equation, and whence the state dependence of the change of measure was much more involved.

### 10.2.1 IID normal random variables

Assume that $\{Y_1, Y_2, \ldots\}$ is a sequence of iid $N(0,1)$ random variables, and

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

Suppose we are interested in estimating $P\{X_n \in A\}$ for the non-convex set

$$A = (-\infty, a] \cup [b, \infty)$$

with $a < 0 < b$. One can write $A$ in the form of (10.1) by taking $\alpha_1 = \beta_1 = a$ and $\alpha_2 = \beta_2 = b$.

The importance sampling algorithm based on a generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \alpha_k)$ is as follows. Let $\bar{X}_0 = 0$ and

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{j} \bar{Y}_j.$$

The sequence $\{\bar{Y}_1, \bar{Y}_2, \ldots\}$ is simulated recursively so that the conditional distribution of $\bar{Y}_{j+1}$ given $\bar{X}_j = x$ is

$$\sum_{k=1}^{2} \rho_k^\delta(x, j/n) \frac{1}{\sqrt{2\pi}} e^{(y-\alpha_k)^2/2} dy.$$

For numerical experimentation, we take $a = -0.25$, $b = 0.2$, and run simulations for $n = 100, 200, 500$. The mollification parameter $\delta$ is set as 0.02. Each estimate consists of 20,000 simulations.

| | $n = 100$ | $n = 200$ | $n = 500$ |
|---|---|---|---|
| Theoretical value | $2.90 \times 10^{-2}$ | $2.54 \times 10^{-3}$ | $3.88 \times 10^{-6}$ |
| Estimate | $2.87 \times 10^{-2}$ | $2.50 \times 10^{-3}$ | $3.92 \times 10^{-6}$ |
| Standard Error | $0.03 \times 10^{-2}$ | $0.04 \times 10^{-3}$ | $0.08 \times 10^{-6}$ |
| 95% C.I. | $[2.81, 2.93] \times 10^{-2}$ | $[2.42, 2.58] \times 10^{-3}$ | $[3.76, 4.08] \times 10^{-6}$ |

Table 2. $P\{X_n \in A\}$ with one-dimensional, non-convex $A$.

### 10.2.2 A finite state Markov chain

Consider a two-node tandem Jackson network with arrival rate $\lambda$ and consecutive service rates $\mu_1, \mu_2$. The system has finite buffers of size $B_1$ and $B_2$, respectively. The embedded time-homogeneous discrete-time Markov chain is $Y = \{Y_i = (Y_i^1, Y_i^2), i \in \mathbb{N}_0\}$, representing the queue lengths of the nodes at the epochs of transitions in the network. This process has a finite state space $S \doteq \{(y_1, y_2) : y_i = 0, 1, \ldots, B_i, i = 1, 2\}$. It is assumed that the system is initially empty, i.e., $Y_0 = (0, 0)$. The transition probability matrix of $Y$ is denoted by $P$.

We are interested in estimating a class of probabilities associated with buffer overflow. More precisely, define $g : S \to \{0, 1\}^2$ by

$$g(y) \doteq \left(1_{\{y_1=B_1\}}, 1_{\{y_2=B_2\}}\right)$$

for every $y = (y_1, y_2) \in S$. Let

$$X_n \doteq \frac{1}{n} \sum_{i=0}^{n-1} g(Y_i).$$

We wish to estimate $P\{X_n \in A\}$, where $A$ takes the form

$$A = \{(x_1, x_2) : x_1 \geq \varepsilon_1 \text{ or } x_2 \geq \varepsilon_2\}$$

for some $0 \leq \varepsilon_1, \varepsilon_2 \leq 1$. We assume $\lambda < \mu_1 \wedge \mu_2$ so that $\{X_n \in A\}$ is a rare event for large $n$.

For each $k = 1, 2$, let $\beta_k \in \mathbb{R}^2$ be the minimizer of $L(\beta)$ over the half space

$$H_k \doteq \{(x_1, x_2) : x_k \geq \varepsilon_k\},$$

and $\alpha_k$ the convex conjugate of $\beta_k$. We can rewrite $A$ in the form of (10.1), that is,

$$A = \cup_{k=1}^2 H_k = \cup_{k=1}^2 \left\{x : \langle x, \alpha_k \rangle \geq \langle \beta_k, \alpha_k \rangle \right\}.$$

The importance sampling algorithm based on a generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \alpha_k)$ is as follows. We simulate $\{\bar{Y}_0, \bar{Y}_1, \ldots\}$ recursively, with initial state $\bar{Y}_0 = (0, 0)$. Let

$$\bar{X}_j = \frac{1}{n} \sum_{i=0}^{j} g(\bar{Y}_i).$$

Thanks to (7.1) and (5.2) with $m(dz|y) = \delta_{g(y)}$, the conditional distribution of $\bar{Y}_{j+1}$, given $\bar{X}_j = x$ and $\bar{Y}_j = y$, is the mixture

$$\sum_{k=1}^{2} \rho_k^\delta(x, j/n) P_k(y, \cdot),$$

where $P_k$ is a transition probability matrix defined by

$$P_k(y, \bar{y}) = e^{\langle \alpha_k, g(\bar{y}) \rangle - H(\alpha_k)} \frac{r(\bar{y}; \alpha_k)}{r(y; \alpha_k)} P(y, \bar{y}), \quad y, \bar{y} \in S.$$

In the numerical simulation we take $B_1 = B_2 = 6$, and $\lambda = 0.2, \mu_1 = \mu_2 = 0.4$, and set $\varepsilon_1 = 0.3$, $\varepsilon_2 = 0.4$. The mollification parameter is chosen as $\delta = 0.1$. We run simulations for $n = 50, 80, 110$, and each estimate consists of 20,000 replications. The theoretical values are obtained using a recursive algorithm and presented for comparison.

| | $n = 50$ | $n = 80$ | $n = 110$ |
|---|---|---|---|
| Theoretical $p_n$ | $5.15 \times 10^{-9}$ | $3.47 \times 10^{-12}$ | $1.83 \times 10^{-15}$ |
| Estimate | $5.05 \times 10^{-9}$ | $3.39 \times 10^{-12}$ | $1.87 \times 10^{-15}$ |
| Standard Error | $0.21 \times 10^{-9}$ | $0.13 \times 10^{-12}$ | $0.08 \times 10^{-15}$ |
| 95% C.I. | $[4.63, 5.47] \times 10^{-9}$ | $[3.13, 3.65] \times 10^{-12}$ | $[1.71, 2.03] \times 10^{-15}$ |

Table 3. $P\{X_n \in A\}$ with two-dimensional, non-convex $A$.

### 10.3 Example: Estimating an expectation $E \exp\{-nF(X_n)\}$

Consider a sequence of iid $N(0,1)$ random variables $\{Y_1, Y_2, \ldots\}$ and let

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

We wish to estimate $E \exp\{-nF(X_n)\}$ where $F \doteq G_1 \wedge G_2 \wedge G_3$, with

$$G_1(x) = (ax + a + 1)^+, \quad G_2(x) = 1, \quad G_3(x) = (bx - b - 1)^-.$$

Under these assumptions, we have $H(\alpha) = \alpha^2/2$, $L(\beta) = \beta^2/2$, and

$$\gamma = \lim_{n \to \infty} -\frac{1}{n} \log E \exp\{-nF(X_n)\} = \inf_{\beta \in \mathbb{R}} [F(\beta) + L(\beta)].$$

For each $k$, let $\beta_k$ be the minimizer of $L(\beta) + G_k(\beta)$ over $\beta \in \mathbb{R}^d$, and $\alpha_k$ its convex conjugate, whence $\alpha_k = \beta_k$. Then equations (9.4) and (9.5) hold, and one can follow the general construction detailed in Section 9.2.2. Let $(\bar{W}_k, \alpha_k)$ be the subsolution/control pair

$$\bar{W}_k(x,t) \doteq -2\langle \alpha_k, x\rangle + 2[G_k(\beta_k) + \langle \alpha_k, \beta_k\rangle] - 2(1-t)H(\alpha_k).$$

Then $\bar{W} = \bar{W}_1 \wedge \bar{W}_2 \wedge \bar{W}_3$ is a weak sense subsolution satisfying the terminal condition $\bar{W}(x,1) \leq 2F(x)$, and $\bar{W}(0,0)$ equals the optimal decay rate $2\gamma$.

A generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \alpha_k)$ obtained as in Section 9.3 induces the following importance sampling scheme. Let $\bar{X}_0 = 0$, and

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{j} \bar{Y}_i.$$

The sequence $\{\bar{Y}_1, \bar{Y}_2, \ldots\}$ is simulated recursively so that the conditional distribution of $\bar{Y}_{j+1}$, given $\bar{X}_j = x$, is the mixture of normal distributions

$$\sum_{k=1}^{3} \rho_k^\delta(x, j/n)\frac{1}{\sqrt{2\pi}}e^{(y-\alpha_k)^2/2}.$$

In the numerical simulation, we take $a = 3/2$, $b = 4$, and it is easy to check that $\alpha_1 = \beta_1 = -3/2$, $\alpha_2 = \beta_2 = 0$, and $\alpha_3 = \beta_3 = 5/4$. The mollification parameter $\delta$ is set to 0.1. We run simulations for $n = 10, 20, 30$, with 20,000 simulations for each estimate. The theoretical value is obtained by direct computation, which expresses $E\exp\{-nF(X_n)\}$ in terms of the cumulative distribution function of $N(0,1)$.

| | $n = 10$ | $n = 20$ | $n = 30$ |
|---|---|---|---|
| Theoretical value | $1.03 \times 10^{-4}$ | $1.87 \times 10^{-8}$ | $5.63 \times 10^{-12}$ |
| Estimate | $1.02 \times 10^{-4}$ | $1.86 \times 10^{-8}$ | $5.73 \times 10^{-12}$ |
| Standard Error | $0.01 \times 10^{-4}$ | $0.03 \times 10^{-8}$ | $0.09 \times 10^{-12}$ |
| 95% C.I. | $[1.00, 1.04] \times 10^{-4}$ | $[1.80, 1.92] \times 10^{-8}$ | $[5.58, 5.91] \times 10^{-12}$ |

Table 4. $E\exp\{-nF(X_n)\}$ for one-dimensional, non-convex $F$.

## 10.4 Example: Level crossing

In this section we consider importance sampling estimates for level crossing probabilities. To illustrate the main idea, we specialize to the following setup. Let $\{Y_1, Y_2, \ldots\}$ be a sequence of iid random vectors taking values in $\mathbb{R}^d$ with common distribution $\mu$, and $A \subset \mathbb{R}^d$ a Borel set. Define the partial sum $S_n = Y_1 + \cdots + Y_n$ with $S_0 = 0$, and for every real number $z > 0$ let

$$T_z = \inf\{n \geq 0 : S_n \in zA\}.$$

Under certain conditions, $\{T_z < \infty\}$ is a rare event for large $z$ and its probability will decay exponentially in the sense that

$$\lim_{z\to\infty} -\frac{1}{z}\log P\{T_z < \infty\} = \gamma$$

for some $\gamma > 0$. The simplest example is when $\{Y_i\}$ are iid, one dimensional random variables with a negative expectation and $A = (1, \infty)$. Naturally, the question is how to estimate $P\{T_z < \infty\}$ for large $z$.

The theoretical result Theorem 8.1 does not cover this case – for example the time horizon is now infinite (see Remark 10.2 for more information). However, the use of subsolutions still carries over and leads to simple and efficient importance sampling algorithms, and we will outline their use in the next few paragraphs.

Let $H$ be the log-moment generating function for $Y_1$, and let $L$ be the Legendre transform of $H$. It is not hard to argue that the Isaacs equation associated with level crossing problems is of the same form as (6.1), except that there is no time dependence. In other words, the Isaacs equation is

$$\sup_{\alpha\in\mathbb{R}^d}\inf_{\beta\in\mathbb{R}^d}\mathbb{H}(DW;\alpha,\beta) = 0 \tag{10.2}$$

and with boundary condition $W(x) = 0$ for $x \in A$. Here $W : \mathbb{R}^d \to \mathbb{R}$ and $DW$ is its gradient, and as before

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

for $s, \alpha, \beta \in \mathbb{R}^d$. A generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$ is defined in a completely analogous fashion to Definition 6.1.

For a given generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$, the corresponding importance sampling algorithm is as follows. Fix the parameter $z$. Let $\bar{X}_0 = 0$. Given $\bar{X}_j = x$, we simulate $\bar{Y}_{j+1}$ under the distribution

$$\sum_{k=1}^{K}\rho_k(x)P(dy;\bar{\alpha}_k(x)),$$

where $P(dy;\alpha)$ is the exponential twist

$$P(dy;\alpha) = e^{\langle\alpha,z\rangle - H(\alpha)}\mu(dz).$$

Finally, we update the dynamics and let

$$\bar{X}_{j+1} = \bar{X}_j + \frac{1}{z}\bar{Y}_{j+1}.$$

The simulation will be stopped at time

$$\bar{T}_z \doteq \inf\left\{n \geq 0 : \bar{X}_n \in A\right\} = \inf\left\{n \geq 0 : \bar{S}_n \in zA\right\},$$

and one forms the importance sampling estimator

$$Z^z \doteq 1_{\{\bar{T}_z < \infty\}} \cdot \prod_{j=0}^{\bar{T}_z - 1} \left[\sum_{k=1}^{K} \rho_k(\bar{X}_j) \cdot e^{\left\langle \bar{\alpha}_k(\bar{X}_j), \bar{Y}_{j+1}\right\rangle - H\left(\bar{\alpha}_k(\bar{X}_j)\right)}\right]^{-1}.$$

The performance of this importance sampling algorithm can be characterized by a result analogous to Theorem 8.1, except that the terminal condition (8.2) should be replaced by the boundary condition

$$\bar{W}(x) \leq 0, \quad \text{for all } x \in A. \tag{10.3}$$

Therefore, the goal is to construct a generalized subsolution/control of a simple form that satisfies the boundary condition (10.3) and such that its value at 0 is equal or close to the optimal decay rate $2\gamma$.

The construction follows the same path, that is, one first identifies a class of affine subsolution/control pairs and then builds a generalized subsolution/control by mollifying the minimum of such affine subsolution/control pairs. The class of affine subsolution/control pairs that serve as building block, again denoted by $\mathbb{A}$, takes a different form in this setting:

$$\mathbb{A} \doteq \left\{(\bar{W}, \bar{\alpha}) : \bar{W}(x) = -2\langle\bar{\alpha}, x\rangle + \bar{c},\ \bar{\alpha} \in \mathbb{R}^d, H(\bar{\alpha}) \leq 0, \bar{c} \in \mathbb{R}\right\}.$$

As in Section 9.1, it is not difficult to check that every $(\bar{W}, \bar{\alpha}) \in \mathbb{A}$ is indeed is a subsolution/control pair.

Analogous to the discussion in Section 9.2, suppose, for example, that there exists $K \in \mathbb{N}$ so that

$$A \subset \cup_{k=1}^{K} \{x : \langle x, \bar{\alpha}_k\rangle \geq \bar{c}_k\}$$

for some $\bar{c}_k \in \mathbb{R}$ and $\bar{\alpha}_k \in \mathbb{R}^d$ such that $H(\bar{\alpha}_k) \leq 0$. Then for each $k$, one can define an affine subsolution/control pair $(\bar{W}_k, \bar{\alpha}_k) \in \mathbb{A}$ with

$$\bar{W}_k(x) \doteq -2\langle\bar{\alpha}_k, x\rangle + 2\bar{c}_k.$$

The minimum of $\{\bar{W}_k\}$ is a weak sense subsolution to the Isaacs equation (10.2) and it satisfies the boundary condition (10.3). One then mollifies it in order to obtain a generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \bar{\alpha}_k)$.

We next present two examples. The first example is a one dimensional level crossing problem, which has been studied extensively [15, 14, 1], and we will see how the subsolution approach leads to the commonly used change of measure. The second example was studied in [12], where it was used as a counterexample to illustrate the danger of blindly following the standard heuristic approach to importance sampling. Level crossing for general Markov additive process and non-convex target sets is also studied in [3].

For each example the numerical experiment considers exponential random variables. There are two reasons for choosing the exponential distribution. One is that an assumption we used very often to facilitate the analysis (e.g., in [6, 7, 8]) is that the log moment generating function $H$ is finite everywhere. This is not true for exponential distributions, and as we will see in fact it is not necessary. The second reason is that for exponential distributions the level crossing probabilities can be explicitly computed, and these theoretical values can be used for comparison.

**Remark 10.2** In the theoretical analysis one needs to "bound" the infinite time horizon in a certain way – essentially to justify an approximation by a finite time problem – and then apply a verification argument similar to the proof of Theorem 8.1. More details can be found in [5], where analysis of this type is carried out for the problem of estimating buffer overflow probabilities in queueing networks.

### 10.4.1 One dimensional level crossing

Assume that $\{Y_1, Y_2, \ldots\}$ are iid random variables with common distribution $\mu$ and that $EY_i < 0$. Let $S_n = Y_1 + \cdots + Y_n$ be the partial sum. Let $A = (1, \infty)$ and

$$T_z \doteq \inf\{n \geq 0 : S_n \in zA\} = \inf\{n \geq 0 : S_n > z\}.$$

Denote by $H$ the log-moment generating function and $\bar{\alpha}$ the unique positive solution to $H(\alpha) = 0$. It is well known that, under mild conditions,

$$\gamma \doteq \lim_{z\to\infty} -\frac{1}{z} \log P\{T_z < \infty\} = \bar{\alpha}.$$

It is obvious that $A \subset \{x : x \cdot \bar{\alpha} \geq \bar{\alpha}\}$. Therefore

$$\bar{W}(x) = -2\bar{\alpha}x + 2\bar{\alpha},$$

and $\bar{\alpha}$ are a subsolution/control pair which also satisfies the boundary condition $\bar{W}(x) \leq 0$ for $x \in A$. Note that $\bar{W}(0) = 2\bar{\alpha} = 2\gamma$, whence the corresponding importance sampling algorithm is asymptotically optimal. This subsolution/control pair induces a change of measure that coincides with the classical choice, that is, the algorithm simulates iid $\{\bar{Y}_i\}$ with common distribution

$$P(dy; \bar{\alpha}) = e^{\bar{\alpha} y - H(\bar{\alpha})} \mu(dy).$$

For numerical experimentation, we consider the special case where for some constant $\theta > 0$, $Y_i + \theta$ is exponentially distributed with parameter $\lambda$. The assumption of $EY_i < 0$ is equivalent to $\theta\lambda > 1$. A bit of algebra yields that $\bar{\alpha}$ is the unique positive root to the equation

$$0 = H(\alpha) = -\alpha\theta + \log\lambda - \log(\lambda - \alpha), \tag{10.4}$$

and that $\{\bar{Y}_i + \theta\}$ are iid exponentially distributed with parameter $\lambda - \bar{\alpha}$. It is not difficult to show that $E\bar{Y}_i > 0$, and thus $\bar{T}_z$ is finite with probability one.

We take $\lambda = 1$, $\theta = 2$, and run simulations for $m = 10, 20, 30$. Each estimate uses 20,000 simulations. The theoretical values are obtained by explicitly solving an integral equation (we omit the details), and

$$P\{T_z < \infty\} = \frac{\lambda - \bar{\alpha}}{\lambda} e^{-\bar{\alpha} z}. \tag{10.5}$$

The value of $\bar{\alpha}$ is obtained by numerically solving equation (10.4) using the bisection method, and $\bar{\alpha} \approx 0.80$.

| | $m = 10$ | $m = 20$ | $m = 30$ |
|---|---|---|---|
| Theoretical value | $7.04 \times 10^{-5}$ | $2.44 \times 10^{-8}$ | $8.44 \times 10^{-12}$ |
| Estimate | $7.11 \times 10^{-5}$ | $2.44 \times 10^{-8}$ | $8.30 \times 10^{-12}$ |
| Standard Error | $0.07 \times 10^{-5}$ | $0.02 \times 10^{-8}$ | $0.08 \times 10^{-12}$ |
| 95% C.I. | $[6.97, 7.28] \times 10^{-5}$ | $[2.40, 2.48] \times 10^{-8}$ | $[8.14, 8.46] \times 10^{-12}$ |

Table 5. Estimating probability of level crossing for 1-dim random walk.

### 10.4.2 Two dimensional level crossing

Let $\{Y_n = (Y_n^1, Y_n^2), n \in \mathbb{N}\}$ be a sequence of iid random vectors with $EY_i^1 < 0$, $EY_i^2 < 0$, and common distribution $\mu$. As before, denote the partial sum by $S_n = (S_n^1, S_n^2) = Y_1 + \cdots + Y_n$ with $S_0 = (0, 0)$. Let

$$A \doteq \{x = (x_1, x_2) : x_1 > 1 \text{ or } x_2 > 1\}.$$

It follows that

$$T_z = \inf\{n \geq 0 : S_n \in zA\} = \inf\left\{n \geq 0 : S_n^1 > z \text{ or } S_n^2 > z\right\}.$$

Denote by $H$ the log-moment generating function. Let $\bar{\alpha}_1 \doteq (\gamma_1, 0)$ and $\bar{\alpha}_2 \doteq (0, \gamma_2)$, where $\gamma_k$ is the unique positive number such that $H(\bar{\alpha}_k) = 0$, $k = 1, 2$. Under mild conditions, we have

$$\gamma \doteq \lim_{z\to\infty} -\frac{1}{z} \log P\{T_z < \infty\} = \gamma_1 \wedge \gamma_2.$$

It is not difficult to check

$$A \subset \{x : \langle x, \bar{\alpha}_1\rangle \geq \gamma_1\} \cup \{x : \langle x, \bar{\alpha}_2\rangle \geq \gamma_2\}.$$

For each $k = 1, 2$, an affine subsolution/control pair is $(\bar{W}_k, \bar{\alpha}_k)$ with

$$\bar{W}_k(x) \doteq -2\langle \bar{\alpha}_k, x\rangle + 2\gamma_k.$$

The minimum $\bar{W} \doteq \bar{W}_1 \wedge \bar{W}_2$ is a weak sense subsolution that satisfies the boundary condition (10.3), and $\bar{W}(0) = 2(\gamma_1 \wedge \gamma_2) = 2\gamma$, the optimal decay rate. A generalized subsolution/control $(\bar{W}^\delta, \rho_k^\delta, \bar{\alpha}_k)$ can then be obtained by mollification, and the corresponding importance sampling algorithm is as follows. Let $\bar{S}_n = \bar{Y}_1 + \cdots + \bar{Y}_n$. We recursively simulate $\{\bar{Y}_1, \bar{Y}_2, \ldots\}$ such that given $\bar{S}_n/z = x$, the conditional distribution of $\bar{Y}_{n+1}$ is

$$\sum_{k=1}^{2} \rho_k^\delta(x) P(dy; \bar{\alpha}_k),$$

where

$$P(dy; \bar{\alpha}_k) = e^{\langle \bar{\alpha}_k, y\rangle - H(\bar{\alpha}_k)} \mu(dy).$$

We stop the simulation once the process $\{\bar{S}_n/z\}$ reaches the set $A$.

For the purpose of numerical experimentation, we consider the special case where for some constants $\theta_k > 0$, the distribution of $Y_1^k + \theta_k$ is exponential with parameter $\lambda_k$, $k = 1, 2$. Assume $\theta_k\lambda_k > 1$, or equivalently $E[Y_1^k] < 0$, for every $k = 1, 2$. We also assume that $\{Y_i^1\}$ and $\{Y_i^2\}$ are independent sequences. It is not difficult to check that, for each $k$, $\gamma_k > 0$ is uniquely determined and satisfies the equation

$$0 = -\gamma_k\theta_k + \log \lambda_k - \log(\lambda_k - \gamma_k). \tag{10.6}$$

Furthermore, $P(dy; \bar{\alpha}_k)$ is the joint distribution of two independent random variables, say $(\bar{Y}^1, \bar{Y}^2)$, such that $\bar{Y}^i + \theta_i$ is exponentially distributed with parameter $\lambda_i - (\bar{\alpha}_k)_i$.

Below is a numerical result. We take $\lambda_1 = \lambda_2 = 1$, and $\theta_1 = 2$, $\theta_2 = 3$. We run simulations for $m = 10, 20, 30$, and each estimate consists of 20,000 simulations. The theoretical values can be easily obtained from the one-dimensional formula (10.5). Again, the value of $\gamma_k$ is obtained by numerically solving equation (10.6) using the bisection method, with $\gamma_1 \approx 0.80$ and $\gamma_2 \approx 0.86$. It is worth pointing out that the importance sampling estimate suggested by the standard heuristic, which is to simulate iid $\{\bar{Y}_i\}$ with distribution $P(dy; \bar{\alpha}_1)$, will have unbounded variance as $z$ tends to infinity [12, Theorem 2(i)]. The mollification parameter $\delta$ is taken as 0.1.

| | $m = 10$ | $m = 20$ | $m = 30$ |
|---|---|---|---|
| Theoretical value | $9.51 \times 10^{-5}$ | $2.88 \times 10^{-8}$ | $9.24 \times 10^{-12}$ |
| Estimate | $9.56 \times 10^{-5}$ | $2.87 \times 10^{-8}$ | $9.31 \times 10^{-12}$ |
| Standard Error | $0.10 \times 10^{-5}$ | $0.03 \times 10^{-8}$ | $0.09 \times 10^{-12}$ |
| 95% C.I. | $[9.36, 9.76] \times 10^{-5}$ | $[2.81, 2.93] \times 10^{-8}$ | $[9.13, 9.49] \times 10^{-12}$ |

Table 6. Probability of level crossing for 2-dim random walk.

### 10.5 Example: A path-dependent event

Let $\{Y_1, Y_2, \ldots\}$ be a sequence of iid random variables with common distribution $\mu$ and $E[Y_i] = 0$. As before, let $H$ be the log-moment generating function and $L$ its convex conjugate. Fix $n \in \mathbb{N}$, and for $1 \leq i \leq n$ define

$$X_i \doteq \frac{1}{n} \sum_{j=1}^{i} Y_j,$$

with $X_0 \doteq 0$. We are interested in estimating

$$E_n \doteq E\left[e^{-nF(X_n)} 1_{\{\max_{0 \leq i \leq n} X_i \geq h\}}\right]$$

where $h > 0$ is a given constant. Let AC$[0,1]$ denote the collection of all absolutely continuous functions on $[0,1]$. Assume that the large deviation limit

$$\lim_{n \to \infty} -\frac{1}{n} \log E_n = \gamma$$

holds, with $\gamma$ the solution of the following variational problem:

$$\gamma = \inf\left\{ \int_0^1 L(\dot{\phi}(t))\, dt + F(\phi(1)) : \phi \in \mathrm{AC}[0,1], \max_{0 \leq t \leq 1} \phi(t) \geq h, \phi(0) = 0 \right\}.$$

To write down the Isaacs equation associated with this estimation problem, we need to expand the state space to accommodate the path-dependence of the event. More precisely, the state process will be $(X_i, B_i)$, where

$$B_i \doteq 1_{\left\{\max_{0 \leq j \leq i} X_j \geq h\right\}}$$

is the indicator of whether or not the "barrier" $h$ has been breached by step $i$. This problem can then be thought of as a combination of the level crossing problem of Section 10.4 and the finite time problem of Section 9.2. First consider the problem conditioned on $B_i = 1$. In this case the large deviation problem takes exactly the same form as in Section 9.2, and the subsolutions of interest are characterized by

$$\bar{W}_t(1, x, t) + \sup_{\alpha \in \mathbb{R}} \inf_{\beta \in \mathbb{R}} \mathbb{H}(D_x \bar{W}(1, x, t); \alpha, \beta) \geq 0$$

with $\mathbb{H}$ be as defined in (6.2), together with the terminal condition

$$\bar{W}(1, x, t) \leq 2F(x). \tag{10.7}$$

An appropriate subsolution will give us a good importance sampling scheme for use at all times *after* the threshold is crossed. The question then is to identify the importance sampling scheme to use before the threshold is crossed. Let us suppose that as soon as the threshold is crossed we switch to the scheme associated with $\bar{W}(1, x, t)$, so that $\bar{W}(1, x, t)$ identifies an upper bound on the performance after this time. We are therefore back in the setting of the level crossing problem, though there is now an *exit cost* $\bar{W}(1, x, t)$ depending on the (scaled) time that the barrier is crossed. Hence the subsolution for times prior to crossing the barrier is also

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

(where the time derivative is needed because the exit cost depends on time), together with the boundary condition

$$\bar{W}(0, x, t) \leq \bar{W}(1, x, t) \tag{10.8}$$

for $x \geq h, t \in [0, 1]$.

Generalized subsolution/controls to these equations are defined as in Sections 10.4 and 9.2, and the corresponding importance sampling algorithm is as follows. Let $\bar{X}_0 = 0$ and $\bar{B}_0 = 0$. Given $\bar{X}_j = x$ and $\bar{B}_j = b$, we simulate $\bar{Y}_{j+1}$ under the distribution

$$\sum_{k=1}^{K} \rho_k(b, x, j/n) P(dy; \bar{\alpha}_k(b, x, j/n))$$

where $P(dy; \alpha)$ is the exponential twist

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

Then we update the dynamics by

$$\bar{X}_{j+1} = \bar{X}_j + \frac{1}{n}\bar{Y}_{j+1}, \quad \bar{B}_{j+1} = 1_{\{\max_{0\leq i\leq j+1} \bar{X}_i \geq h\}}.$$

The performance of this importance sampling algorithm can be characterized by a result analogous to Theorem 8.1: If $V^n$ is the second moment of the importance sampling estimator corresponding to $\bar{W}$, then

$$\liminf_{n\to\infty} -\frac{1}{n}\log V^n \geq \bar{W}(0,0,0).$$

Therefore, the goal is to find a structurally simple generalized subsolution/control $(\bar{W}, \rho_k, \bar{\alpha}_k)$ satisfying (10.8) and (10.7) with $\bar{W}(0,0,0)$ as large as possible, preferably equal to the optimal decay rate $2\gamma$.

For illustration, we will consider the simple setting where

$$E_n = P\left\{\max_{0\leq i\leq n} X_i \geq h, X_n \leq l\right\},$$

for some constant $0 < l < h$. In other words, $F(x) = \infty$ for $x > l$ and 0 otherwise. As in the cases studied previously, a subsolution can be identified in terms of the solution to the large deviation variational problem. Using convexity and Jensen's inequality, this problem can be written in the form

$$\inf\left\{\rho_0 L\left(\frac{h}{\rho_0}\right) + \rho_1 L\left(\frac{l-h}{\rho_1}\right) : \rho_i \geq 0, i = 0,1, \rho_0 + \rho_1 = 1\right\}.$$

Since the mean of $\mu$ is zero, $L(\beta) = 0$ if and only if $\beta = 0$, and thus the infimum is achieved at $\rho_i^*$ with $\rho_i^* > 0$ for $i = 1,2$. Let $\beta_0^* = h/\rho_0^*$ and $\beta_1^* = (l-h)/\rho_1^*$, and let $\alpha_0^*$ and $\alpha_1^*$ be the convex conjugates, respectively. It is not difficult to see that $\alpha_1^* < 0 < \alpha_0^*$. We claim that $H(\alpha_0^*) = H(\alpha_1^*)$. Indeed, the necessary condition for a minimizer gives

$$L\left(\beta_0^*\right) - L'\left(\beta_0^*\right)\beta_0^* - L\left(\beta_1^*\right) + L'\left(\beta_1^*\right)\beta_1^* = 0$$

Using the characterization $\alpha_i^* = L'\left(\beta_i^*\right)$, we have

$$H(\alpha_0^*) = L'\left(\beta_0^*\right)\beta_0^* - L\left(\beta_0^*\right) = L'\left(\beta_1^*\right)\beta_1^* - L\left(\beta_1^*\right) = H(\alpha_1^*).$$

Using the interpretation of $\bar{W}(1,x,t)$ as the solution to the finite time problem with the given terminal condition, we know from Section 9.2 that a subsolution is given by

$$\bar{W}(1,x,t) = -2\alpha_1^* x + 2l\alpha_1^* - 2(1-t)H(\alpha_1^*).$$

As subsolution for the times prior to exceeding $h$ we use the form

$$\bar{W}(0,x,t) = -2\alpha_0^* x + c_0 - 2(1-t)H(\alpha_0^*).$$

Since $H(\alpha_0^*) = H(\alpha_1^*)$ and $\alpha_1^* < 0 < \alpha_0^*$, to satisfy the boundary condition (10.8) for $x \geq h$ we need $-2\alpha_0^* h + c_0 \leq -2\alpha_1^* h + 2l\alpha_1^*$, and to obtain the largest value for $\bar{W}(0,0,0)$ we take $c_0 = 2\alpha_0^* h + 2(l-h)\alpha_1^*$, so that the two functions agree on $x = h$. It follows that

$$\begin{aligned} \bar{W}(0,0,0) &= 2\alpha_0^* h + 2(l-h)\alpha_1^* - 2H(\alpha_0^*) \\ &= 2\rho_0^* \left[\alpha_0^* \beta_0^* - 2H(\alpha_0^*)\right] + 2\rho_1^* \left[\alpha_1^* \beta_1^* - 2H(\alpha_1^*)\right] \\ &= 2\gamma. \end{aligned}$$

In other words, the corresponding scheme is asymptotically optimal.

For a numerical example we take $Y_i \sim N(0,1)$. The corresponding importance sampling algorithm takes a very simple form. Let $\bar{X}_0 = 0$ and $\bar{B}_0 = 0$. If $\bar{B}_j = 0$, that is, the sample path maximum has not yet surpassed barrier $h$, we simulate $\bar{Y}_{j+1}$ under the distribution $N(2h-l,1)$. If $\bar{B}_j = 1$, that is, the barrier $h$ has already been reached, we simulate $\bar{Y}_{j+1}$ under the distribution $N(l-2h,1)$. Then we update the dynamics by

$$\bar{X}_{j+1} = \bar{X}_j + \frac{1}{n}\bar{Y}_{j+1}, \quad \bar{B}_{j+1} = 1_{\{\max_{0 \leq i \leq j+1} \bar{X}_i \geq h\}}.$$

In the numerical experiment below, $h = 1$ and $l = 0.8$. Simulations were run for $n = 10, 20, 30$, and each estimate consists of 20,000 samples. What we call the "theoretical value" is an estimate based on 1 billion samples of the importance sampling scheme.

| | $n = 10$ | $n = 20$ | $n = 30$ |
|---|---|---|---|
| Theoretical value | $1.68 \times 10^{-5}$ | $9.66 \times 10^{-9}$ | $6.09 \times 10^{-12}$ |
| Estimate | $1.74 \times 10^{-5}$ | $9.58 \times 10^{-9}$ | $6.26 \times 10^{-12}$ |
| Standard Error | $0.04 \times 10^{-5}$ | $0.27 \times 10^{-9}$ | $0.19 \times 10^{-12}$ |
| 95% C.I. | $[1.66, 1.82] \times 10^{-5}$ | $[9.04, 10.12] \times 10^{-9}$ | $[5.88, 6.64] \times 10^{-12}$ |

Table 7. Estimating a path-dependent probability.

## 10.6 Example: A mixed open/closed queueing network

Consider the mixed open/closed queueing network as shown in Figure 2. The open jobs arrive at server 1 according to a Poisson process with rate $\lambda$. There is one closed job that circulates between server 1 and server 2, and it has pre-emptive priority over open jobs at server 1. All services rates

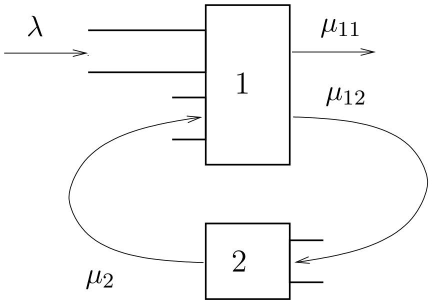

Figure 2: An open/closed queueing network.

are exponentially distributed. The service rates at server 1 are $\mu_{11}$ for open jobs and $\mu_{12}$ for the closed job, and the service rate at server 2 is $\mu_2$. Note that this system is equivalent to an $M/M/1$ queue with server breakdowns or vacations.

The state of the system is described by process $\{(Y_t, Z_t) : t \geq 0\}$, where $Y_t$ and $Z_t$ are the numbers of open jobs and closed jobs at server 1 at time $t$. We wish to estimate $p_n$, the probability that the number of open jobs reaches $n$ before the system returns to state $(0, 0)$, given that the system starts in $(0, 0)$. Assuming that the system is stable, that is,

$$\frac{\lambda}{\mu_{11}} + \frac{\mu_2}{\mu_2 + \mu_{12}} < 1, \tag{10.9}$$

$p_n$ is a rare-event probability when $n$ gets large.

The associated large deviation asymptotics can be characterized by an ordinary differential equation (ODE). More precisely, let $y \in \{0, 1, \ldots, n\}$, $z \in \{0, 1\}$, and $V_n(y, z)$ the probability that the number of open jobs reaches $n$ before the system returns to state $(0, 0)$, given that the system starts in $(y, z)$ [whence $p_n = V_n(0, 0)$ by definition]. Given any $x \in [0, 1]$ and $z \in \{0, 1\}$, we have

$$\lim_{n\to\infty} -\frac{1}{n} \log V_n(\lfloor nx \rfloor, z) = v(x)$$

where $v$ is a viscosity solution to an ODE. To describe this ODE, we define the convex function

$$\ell(s) \doteq \left\{ \begin{array}{cc} s \log s - s + 1; & s \geq 0 \\ +\infty \quad ; & s < 0 \end{array} \right. .$$

Using notation $\rho = (\rho_0, \rho_1), \hat{\Theta} = (\hat{\lambda}_0, \hat{\lambda}_1, \hat{\mu}_{11}, \hat{\mu}_{12}, \hat{\mu}_2)$, we define for $\beta \in \mathbb{R}$

$$L(\beta) \doteq \inf_{\rho, \hat{\Theta}} G(\rho, \hat{\Theta}), \tag{10.10}$$

where

$$G(\rho, \hat{\Theta}) \doteq \rho_0 \left[ \lambda \ell \left( \frac{\hat{\lambda}_0}{\lambda} \right) + \mu_{12} \ell \left( \frac{\hat{\mu}_{12}}{\mu_{12}} \right) \right] \\ + \rho_1 \left[ \lambda \ell \left( \frac{\hat{\lambda}_1}{\lambda} \right) + \mu_{11} \ell \left( \frac{\hat{\mu}_{11}}{\mu_{11}} \right) + \mu_2 \ell \left( \frac{\hat{\mu}_2}{\mu_2} \right) \right]$$

and with the infimum in (10.10) taken over all $(\rho, \hat{\Theta})$ such that

$$\rho_0 \geq 0,\ \rho_1 \geq 0,\ \rho_0 + \rho_1 = 1,\ \rho_0 \hat{\mu}_{12} = \rho_1 \hat{\mu}_2,\ \beta = \rho_0 \hat{\lambda}_0 + \rho_1 \hat{\lambda}_1 - \rho_1 \hat{\mu}_{11}. \quad (10.11)$$

One can show that function $L$ is indeed convex and explicitly calculate the Legendre transform $H$ of $L$. Indeed, we have (see Appendix C)

$$H(\alpha) = \inf_q [H_0(\alpha, q) \vee H_1(\alpha, q)] \quad (10.12)$$

where

$$\begin{aligned} H_0(\alpha, q) &= \lambda(e^\alpha - 1) + \mu_{12}(e^q - 1), \\ H_1(\alpha, q) &= \lambda(e^\alpha - 1) + \mu_{11}(e^{-\alpha} - 1) + \mu_2(e^{-q} - 1). \end{aligned}$$

Then $v$ satisfies the ODE

$$0 = \inf_{\beta \in \mathbb{R}} \left[ v'(x) \cdot \beta + L(\beta) \right] = -H(-v'(x)),$$

with the boundary condition $v(1) = 0$. Solving this ODE (see Appendix C), we obtain

$$v(x) = \gamma(1 - x), \quad (10.13)$$

where

$$\gamma = -\log \frac{(\lambda + \mu_{11} + \mu_{12} + \mu_2) + \sqrt{(\lambda + \mu_{11} + \mu_{12} + \mu_2)^2 - 4\mu_{11}(\lambda + \mu_{12})}}{2\mu_{11}(1 + \lambda^{-1}\mu_{12})}$$

is a positive number. In particular,

$$\lim_{n \to \infty} -\frac{1}{n} \log p_n = v(0) = \gamma.$$

Moreover, the minimizing $\hat{\Theta}$ for (10.10) is

$$\Theta^* = (\lambda_0^*, \lambda_1^*, \mu_{11}^*, \mu_{12}^*, \mu_2^*) = (e^\gamma \lambda, e^\gamma \lambda, e^{-\gamma} \mu_{11}, e^{q^*} \mu_{12}, e^{-q^*} \mu_2), \quad (10.14)$$

where $q^*$ is the minimizer in equation (10.12) with $\alpha = \gamma$, or

$$q^* = \log \frac{\lambda + \mu_{12} - \lambda e^{\gamma}}{\mu_{12}}. \tag{10.15}$$

Now let us consider the construction of importance sampling algorithms. We first describe the associated Isaacs equation. Let $\bar{\Theta} \doteq (\bar{\lambda}_0, \bar{\lambda}_1, \bar{\mu}_{11}, \bar{\mu}_{12}, \bar{\mu}_2)$. Define

$$\bar{L}(\beta, \bar{\Theta}) \doteq \inf_{\rho, \hat{\Theta}} \left[ 2G(\rho, \hat{\Theta}) - \bar{G}(\rho, \hat{\Theta}, \bar{\Theta}) \right] \tag{10.16}$$

where

$$\bar{G}(\rho, \hat{\Theta}, \bar{\Theta}) \doteq \rho_0 \left[ \bar{\lambda}_0 \ell \left( \frac{\hat{\lambda}_0}{\bar{\lambda}_0} \right) + \bar{\mu}_{12} \ell \left( \frac{\hat{\mu}_{12}}{\bar{\mu}_{12}} \right) \right] \\ + \rho_1 \left[ \bar{\lambda}_1 \ell \left( \frac{\hat{\lambda}_1}{\bar{\lambda}_1} \right) + \bar{\mu}_{11} \ell \left( \frac{\hat{\mu}_{11}}{\bar{\mu}_{11}} \right) + \bar{\mu}_2 \ell \left( \frac{\hat{\mu}_2}{\bar{\mu}_2} \right) \right]$$

and the infimum in (10.16) is taken over all $(\rho, \hat{\Theta})$ satisfying the constraints (10.11). The Isaacs equation associated with importance sampling can then be written as

$$\sup_{\bar{\Theta}} \inf_{\beta} \left[ W'(x) \cdot \beta + \bar{L}(\beta, \bar{\Theta}) \right] = 0,$$

with boundary condition $W(1) = 0$. Let $\bar{W} = 2v$. It is not difficult to show [see Appendix C] that $(\bar{W}, \Theta^*)$ defines an affine subsolution/control pair to the Isaacs equation, and it satisfies the terminal condition $\bar{W}(1) = 2v(1) = 0$. Furthermore, $\bar{W}(0) = 2v(0) = 2\gamma$, the optimal decay rate.

The importance sampling algorithm corresponding to this affine subsolution/control pair $(\bar{W}, \Theta^*)$ is very simple. When $z = 1$, we simulate the system under the alternative probability measure such that the open job arrival rate is $\lambda_1^*$ and the service rate for the closed job at server 1 is $\mu_{12}^*$. When $z = 0$, the simulation distribution is such that the open job arrival rate is $\lambda_0^*$ and the service rate for the open job is $\mu_{11}^*$ and the service rate for the closed job at server 2 is $\mu_2^*$. Note that, for this special network, since $\lambda_0^* = \lambda_1^* \doteq \lambda^*$, the above change of measure is equivalent to simulation under the alternative rates $(\lambda^*, \mu_{11}^*, \mu_{12}^*, \mu_2^*)$.

In the numerical example below we take $\lambda = 1, \mu_{11} = 4, \mu_{12} = 2, \mu_2 = 0.5$. It is easy to check that the stability condition (10.9) holds. We run simulations for $n = 20, 40, 80$, and each estimate consists of 20,000 simulations. The theoretical value can be found in [10].

|  | $n = 20$ | $n = 40$ | $n = 80$ |
|---|---|---|---|
| Theoretical value | $3.91 \times 10^{-8}$ | $2.02 \times 10^{-15}$ | $5.40 \times 10^{-30}$ |
| Estimate | $3.93 \times 10^{-8}$ | $2.01 \times 10^{-15}$ | $5.45 \times 10^{-30}$ |
| Standard Error | $0.03 \times 10^{-8}$ | $0.02 \times 10^{-15}$ | $0.04 \times 10^{-30}$ |
| 95% C.I. | $[3.87, 3.99] \times 10^{-8}$ | $[1.97, 2.05] \times 10^{-15}$ | $[5.37, 5.53] \times 10^{-30}$ |

Table 8. Overflow probabilities of a mixed open-closed queueing network.

**Remark 10.3** It is worth mentioning that our approach can easily extend to the general case where the system has multiple closed jobs, with the sole difference being that the computation of $\gamma$ becomes more involved.

## 10.7 Example: A "general purpose" importance sampling scheme

For all the examples we have discussed, finitely many affine subsolution/control pairs are used to construct a generalized subsolution/control. In this section, we present an example where infinitely many subsolution/control pairs are used for this purpose. The corresponding importance sampling scheme has some interesting features, which are further discussed in Remark 10.4.

For illustration, we consider again the simple setting where $\{Y_1, Y_2, \ldots\}$ is a sequence of iid random variables taking values in $\mathbb{R}^d$ and let

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

with $X_0 \doteq 0$. We wish to estimate $P\{X_n \in A\}$ where $A \subset \mathbb{R}^d$ is a Borel set, and assume a large deviation limit holds, that is,

$$\lim_{n \to \infty} -\frac{1}{n} \log P\{X_n \in A\} = \inf_{\beta \in A} L(\beta) = \gamma.$$

A new way to construct a subsolution is as follows. Consider the level set of the rate function $L$

$$\Theta_\gamma \doteq \left\{ \beta \in \mathbb{R}^d : L(\beta) \geq \gamma \right\}.$$

It follows easily that $\Theta_\gamma$ is the complement of a convex set and that $A \subset \Theta_\gamma$. For each $\beta \in \partial\Theta_\gamma$, let $\alpha(\beta)$ be its convex conjugate (assuming its existence). Define $(\bar{W}_\beta, \alpha(\beta)) \in \mathbb{A}$ by

$$\bar{W}_\beta(x,t) \doteq -2\langle \alpha(\beta), x \rangle + 2\langle \alpha(\beta), \beta \rangle - 2(1-t)H[\alpha(\beta)].$$

Let

$$\bar{W}(x,t) \doteq \inf\left\{\bar{W}_\beta(x,t) : \beta \in \partial\Theta_\gamma\right\}.$$

Then $\bar{W}$ defines a (weak sense) subsolution to the Isaacs equation (6.1). Furthermore, thanks to the convexity of $\Theta_\gamma^c$, we have

$$\bar{W}(x,1) = \inf\left\{-2\langle\alpha(\beta), x-\beta\rangle : \beta \in \partial\Theta_\gamma\right\} \leq 0$$

for $x \in \Theta_\gamma$. In particular, $\bar{W}$ satisfies the terminal condition (9.1) since $A \subset \Theta_\gamma$. In many cases (e.g., when $L$ is finite), we have

$$\begin{aligned}\bar{W}(0,0) &= \inf\left\{2\langle\alpha(\beta),\beta\rangle - 2H[\alpha(\beta)] : \beta \in \partial\Theta_\gamma\right\}\\ &= \inf\left\{2L(\beta) : \beta \in \partial\Theta_\gamma\right\}\\ &= 2\gamma.\end{aligned}$$

Using Remark 6.2 and Theorem 8.1, if $\bar{W}$ were continuously differentiable then $(\bar{W}, \bar{\alpha})$ with $\alpha(x,t) = -D\bar{W}(x,t)/2$ would form a subsolution/control pair, and the corresponding importance sampling schemes would yield asymptotically optimal performance. Since $\bar{W}$ is not continuously differentiable, one possibility is to resort to mollification. However, an alternative that is possible in some cases is to show that for each $\varepsilon > 0$ one can find a smooth function $\bar{W}^\varepsilon$ such that $(\bar{W}^\varepsilon, \bar{\alpha})$ form a subsolution/control pair and $\bar{W}^\varepsilon \to \bar{W}$ as $\varepsilon \to 0$. This approach is used in the following example.

To give a concrete example, we will work out the details for iid $N(0, I_d)$ sequence $\{Y_1, Y_2, \ldots\}$, where $I_d$ denotes the identity matrix of dimension $d$. It follows that $H(\alpha) = \|\alpha\|^2/2$, $L(\beta) = \|\beta\|^2/2$, and whence

$$\Theta_\gamma = \left\{\beta \in \mathbb{R}^d : \|\beta\| \geq \sqrt{2\gamma}\right\}.$$

In this case $\alpha(\beta) = \beta$, and

$$\begin{aligned}\bar{W}(x,t) &\doteq \inf\left\{-2\langle\beta,x\rangle + 2\langle\beta,\beta\rangle - 2(1-t)H(\beta) : \|\beta\| \in \partial\Theta_\gamma\right\}\\ &= \inf\left\{-2\langle\beta,x\rangle + 2(1+t)\gamma : \|\beta\| = \sqrt{2\gamma}\right\}\\ &= -2\sqrt{2\gamma}\|x\| + 2(1+t)\gamma.\end{aligned}$$

It is not difficult to check that $\bar{W}$ satisfies the Isaacs equation (6.1) except at $\{x = 0\}$, $\bar{W}(x,1) \leq 0$ on $\Theta_\gamma$, and $\bar{W}(0,0) = 2\gamma$.

Even though $\bar{W}$ is not continuously differentiable at $\{x = 0\}$, it induces a control

$$\bar{\alpha}(x,t) \doteq -\frac{D\bar{W}(x,t)}{2} = \sqrt{2\gamma}\frac{x}{\|x\|}, \quad \text{if } x \neq 0.$$

For $x = 0$, we just define

$$\bar{\alpha}(x, 0) \doteq \sqrt{2\gamma}\theta, \tag{10.17}$$

where $\theta$ is an arbitrarily fixed unit vector. We claim that the importance sampling scheme corresponding to this control $\bar{\alpha}$ is asymptotically optimal. Indeed, consider the approximating sequence $\bar{W}^{\varepsilon}$ defined by

$$\bar{W}^{\varepsilon}(x, t) \doteq -2\sqrt{2\gamma}\sqrt{\|x\|^2 + \varepsilon} + 2(1 + t)\gamma.$$

It is not difficult to check that, for every $\varepsilon > 0$, $\bar{W}^{\varepsilon}$ is continuously differentiable and $(\bar{W}^{\varepsilon}, \bar{\alpha})$ is a subsolution/control pair. The asymptotic optimality follows if one applies Theorem 8.1 to this subsolution/control pair, and observes that $\lim_{\varepsilon \to 0} W^{\varepsilon}(0, 0) = 2\gamma$, the optimal decay rate.

For numerical experimentation, we take $d = 2$ and

$$A \doteq \left\{x = (x_1, x_2) : (x_1 + a)^2 + x_2^2 \geq R^2\right\}$$

for some constants $0 < a < R$, whence $\gamma = (R - a)^2/2$. Setting $\theta = (1, 0)$ in equation (10.17), the importance sampling scheme is as follows. We recursively simulate $\{\bar{Y}_1, \bar{Y}_2, \ldots\}$ and let

$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{j} \bar{Y}_i.$$

The conditional distribution of $\bar{Y}_{j+1}$ given $\bar{X}_j = x = (x_1, x_2)$ is

$$N\left(\left[\begin{array}{c} (R - a)x_1/\|x\| \\ (R - a)x_2/\|x\| \end{array}\right], I_2\right)$$

if $x \neq 0$, and

$$N\left(\left[\begin{array}{c} R - a \\ 0 \end{array}\right], I_2\right)$$

if $x = 0$.

For the table below we take $R = 0.5, a = 0.05$. We run simulations for $n = 40, 80, 120$, and each estimate consists of 20,000 simulations. The theoretical value can be obtained using standard software such as S-plus.

| | $n = 40$ | $n = 80$ | $n = 120$ |
|---|---|---|---|
| Theoretical value | $8.49 \times 10^{-3}$ | $1.00 \times 10^{-4}$ | $1.40 \times 10^{-6}$ |
| Estimate | $8.38 \times 10^{-3}$ | $1.04 \times 10^{-4}$ | $1.50 \times 10^{-6}$ |
| Standard Error | $0.18 \times 10^{-3}$ | $0.04 \times 10^{-4}$ | $0.07 \times 10^{-6}$ |
| 95% C.I. | $[8.02, 8.74] \times 10^{-3}$ | $[0.96, 1.12] \times 10^{-4}$ | $[1.36, 1.64] \times 10^{-6}$ |

Table 9. A “universal” scheme

**Remark 10.4** An important feature of this approach is that the construction does not need any information on the set $A$ other than $\gamma$, the infimum of $L(\beta)$ over $\beta \in A$. However, it is often the case that one knows more detailed properties of the target set $A$, which can be used to design more efficient schemes. Furthermore, there is a practical computational issue of obtaining $\bar{W}$ as the minimum of infinitely many subsolutions in the case of, say, sums of functionals of a Markov chain. However, here one may be willing to approximate $\Theta_\gamma$ by a finite number of points and then use exponential weighting for mollification.

## 11 Summary

We have shown that importance sampling schemes based on subsolutions can be applied in a wide variety of settings and deliver excellent performance. Besides being fast and accurate, the behavior of the schemes is very stable across a broad range of problem formulations. For example, in each setting and for each simulation we use the same number of samples (20,000) with remarkably similar performance. Moreover the asymptotic behavior of the schemes can be backed up by rigorous theoretical justification. Both of these properties stand in sharp contrast to the instability and poor performance exhibited by standard heuristic importance sampling schemes in many situations [12, 11, 6, 7].

Compared with the schemes based on subsolutions, those based on *solutions* can in some cases lead to better estimates with the same sample size (e.g., 2 or 3 times better standard error), but they are often much slower, especially when the computation involves solving nontrivial eigenvalue/eigenfunction problems. However, these comparisons are possible only for those problems for which solutions can be found (either theoretically or numerically), and a point that is even more important is that subsolutions can be constructed for a much wider class of problems.

An interesting question is whether one can forgo mollification and directly use the piecewise affine subsolution (a weak sense subsolution) to construct importance sampling schemes. This is tempting since a piecewise affine subsolution does not suffer any loss of performance from mollification. However, we conjecture that there are no results analogous to Theorem 8.1 except for special cases, and this is supported numerical experimentation. A final remark on this point is that [5] rigorously shows that, in the context of queueing networks, a weak sense subsolution (or even a weak sense solution) to the Isaacs equation can lead to inefficient importance sampling schemes.

**Acknowledgment.** We are indebted to the associate editor and two anonymous referee for their careful reading and many helpful suggestions.

# A Appendix. Proof of the main theorem

**Outline of the Proof of Theorem 8.1.** Let $V^n \doteq E\left[(Z^n)^2\right]$ be second moment of the importance sampling estimator and

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

By expressing the second moment in terms of the original random variables [see Remark 2.1], we can write

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

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. Show that these processes are tight under the assumption that the costs in the representation are bounded.

3. *Identification of limits.* Characterize 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, a classical verification argument to show that the proper asymptotic bound holds for the representation.

The chain rule for relative entropy [4, 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

$$R\left(\mu\,\|\nu\right) = R\left([\mu]_1\,\|[\nu]_1\right) + \int_{S_1} R\left(\mu(\cdot|y_1)\,\|\nu(\cdot|y_1)\right)[\mu]_1(dy_1). \tag{A.1}$$

## A.1 Representation.

Using the 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

$$\begin{aligned}\tilde{V}^n \quad \doteq \quad & Ee^{-nG(X_n)} \prod_{j=0}^{n-1} \exp\left\{ -\sum_{k=1}^{K} \rho_{k,j}^n\left(X_j\right) \left[ \left\langle \bar{\alpha}_{k,j}^n(X_j), b_{j+1}(Y_{j+1}) \right\rangle \right.\right. \\ & \left.\left. \qquad - H\left(\bar{\alpha}_{k,j}^n(X_j)\right) + \log \frac{r(Y_{j+1}; \bar{\alpha}_{k,j}^n(X_j))}{r(Y_j; \bar{\alpha}_{k,j}^n(X_j))} \right] \right\}.\end{aligned}$$

Define

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

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

$$\liminf_{n\to\infty} \tilde{W}^n \geq \bar{W}(0,0). \tag{A.2}$$

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 $G\left(X_n\right)$, an extension of this representation is required.

**Lemma A.1** *Let $\lambda$ be a probability measure on a measurable space $(\Omega, \mathcal{F})$, and $f : \Omega \to \mathbb{R}$ a measurable function. If $e^{-f}$ and $fe^{-f}$ are integrable with respect to $\lambda$, 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 $\gamma$ for which the sum at the right-hand-side is meaningful.*

The proof only involves minor changes to that of [4, 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 $b_i(y)$ and the assumed Lipschitz property of $\bar{W}$.

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 $\lambda$, 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 $\gamma$ be the probability measure introduced by the representation. Now factor both the original distribution $\lambda$ and the new probability measure $\gamma$ as a product of conditional distributions. For example, if $\lambda$ 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 (A.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., [4, Theorem B.2.2]), we simply state the final result. Consider a collection of stochastic kernels $\mu^n_j$ and $\nu^n_j$ that are defined recursively as follows. $\mu^n_j$ is allowed to depend in any measurable way on $\{\tilde{b}^n_i, 0 \leq i \leq j\}$ and $\{\tilde{Y}^n_i, 0 \leq i \leq j+1\}$, $\nu^n_j$ is allowed to depend in any measurable way on $\{\tilde{b}^n_i, 0 \leq i \leq j\}$ and $\{\tilde{Y}^n_i, 0 \leq i \leq j\}$, and $\mu^n_j$ and $\nu^n_j$ choose the conditional distributions of $\tilde{b}^n_{j+1}$ and $\tilde{Y}^n_{j+1}$, respectively. To simplify the notation the dependencies of $\mu^n_j$ and $\nu^n_j$ on the past will not be made explicit. Let

$$\begin{aligned} J(\mu^n_\cdot, \nu^n_\cdot) \doteq \tilde{E}\Bigg[ & \frac{1}{n}\sum_{j=0}^{n-1}\sum_{k=1}^{K} \rho^n_{k,j}(\tilde{X}^n_j) \Bigg[ R\left(\mu^n_j(\cdot) \| m(\cdot|\tilde{Y}^n_{j+1})\right) + R\left(\nu^n_j(\cdot)\| p(\tilde{Y}^n_j, \cdot)\right) \\ & + \left\langle \bar{\alpha}^n_{k,j}(\tilde{X}^n_j), \tilde{b}^n_{j+1} \right\rangle - H\left(\bar{\alpha}^n_{k,j}(\tilde{X}^n_j)\right) + \log \frac{r(\tilde{Y}^n_{j+1}; \bar{\alpha}^n_{k,j}(\tilde{X}^n_j))}{r(\tilde{Y}^n_j; \bar{\alpha}^n_{k,j}(\tilde{X}^n_j))} \Bigg] \\ & + G(\tilde{X}^n_n) \Bigg]. \end{aligned} \tag{A.3}$$

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

## A.2 Tightness

To analyze the asymptotics of $\tilde{W}^n$ we first establish the tightness of the processes that appear therein. For $j = 0, \ldots, n-1$ 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 $\theta^n$ is a random probability measure on space $(S \times S) \times [0, 1]$. Similarly define random probability measures $\nu^n, \mu^n, \gamma^n, \zeta^n$, and $\eta^n$.

**Lemma A.2** *Assume Condition 4.1. Let $(\bar{W}, \rho_k, \bar{\alpha}_k)$ be a generalized subsolution/control. Consider any subsequence and collection $\{(\mu_j^n, \nu_j^n), j = 0, 1, \ldots, n-1\}$ for which the expected cost $J(\mu_\cdot^n, \nu_\cdot^n)$ as defined in (A.3) is uniformly bounded from above. Then (with the supremum on $n$ restricted to elements of the subsequence)*

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

*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*

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

The proof of the lemma is given in Appendix B. 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 (A.2), all we need to show is

$$\liminf_{n\to\infty} J(\mu^n_\cdot, \nu^n_\cdot) \geq \bar{W}(0,0) \tag{A.4}$$

for any sequence $\{(\mu^n_j, \nu^n_j), j = 0, \ldots, n-1\}$. Abusing notation a bit, assume from now on that $\{(\mu^n_j, \nu^n_j), j = 0, \ldots, n-1\}$ is an arbitrary subsequence such that the cost $J(\mu^n_\cdot, \nu^n_\cdot)$ is uniformly bounded from above. Clearly, we only need to show inequality (A.4) along every such subsequence.

Owing to the positivity, boundedness, and Lipschitz properties of the eigenfunctions and $\bar{\alpha}_k$ (see Lemma 5.1), there exists $M < \infty$ such that for all $y \in S$, $k = 1, \ldots, K$, $n \in \mathbb{Z}_+$, $j \in \{1, \ldots, n\}$, $x_1 \in \mathbb{R}^d$, and $x_2 \in \mathbb{R}^d$,

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

Thanks to the first part of Lemma A.2, for any $\delta > 0$ and along this subsequence with bounded cost,

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

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

$$\begin{aligned}
&\limsup_{n\to\infty} \tilde{E}\left[\frac{1}{n}\sum_{j=0}^{n-1}\sum_{k=1}^{K} \rho^n_{k,j}(\tilde{X}^n_j)\left|\log \frac{r(\tilde{Y}^n_{j+1}; \bar{\alpha}^n_{k,j}(\tilde{X}^n_j))}{r(\tilde{Y}^n_j; \bar{\alpha}^n_{k,j}(\tilde{X}^n_j))}\right|\right] \\
&\leq \limsup_{n\to\infty} \tilde{E}\left[\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{K} \rho^n_{k,j}(\tilde{X}^n_j)\left|\log \frac{r(\tilde{Y}^n_j; \bar{\alpha}^n_{k,j-1}(\tilde{X}^n_{j-1}))}{r(\tilde{Y}^n_j; \bar{\alpha}^n_{k,j}(\tilde{X}^n_j))}\right|\right] \\
&\quad + \limsup_{n\to\infty} \tilde{E}\left[\frac{1}{n}\sum_{j=0}^{n-1}\sum_{k=1}^{K} \left|\rho^n_{k,j+1}(\tilde{X}^n_{j+1}) - \rho^n_{k,j}(\tilde{X}^n_j)\right|\right. \\
&\qquad\qquad \left.\cdot \left|\log r(\tilde{Y}^n_{j+1}; \bar{\alpha}^n_{k,j}(\tilde{X}^n_j))\right|\right] \\
&= 0.
\end{aligned} \tag{A.5}$$

Thanks to (A.3) and (A.5), in order to show (A.4) it suffices to prove

$$\liminf_{n\to\infty} \tilde{E}\left[\frac{1}{n}\sum_{j=0}^{n-1}\sum_{k=1}^{K} \rho^n_{k,j}(\tilde{X}^n_j)\left[R\left(\mu^n_j(\cdot)\left\|m(\cdot|\tilde{Y}^n_{j+1})\right.\right) + R\left(\nu^n_j(\cdot)\left\|p(\tilde{Y}^n_j, \cdot)\right.\right)\right.\right.$$

$$+\left\langle \bar{\alpha}^n_{k,j}(\tilde{X}^n_j), \tilde{b}^n_{j+1}\right\rangle - H\left(\bar{\alpha}^n_{k,j}(\tilde{X}^n_j)\right)\Big] + G(\tilde{X}^n_n)\Bigg] \geq \bar{W}(0,0).$$

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 $\rho_k$ and $\bar{\alpha}_k$, the uniform integrability of $\{\mu^n\}$ (Lemma A.2), and the chain rule for relative entropy (A.1), all we need to show is the lower bound

$$\liminf_{n\to\infty} \bar{J}^n \geq \bar{W}(0,0), \tag{A.6}$$

where

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

## A.3 Identification of the Limits.

In order to show (A.6), we need to identify limits of the involved processes.

**Lemma A.3** *Assume Condition 4.1, and consider any subsequence along which $J(\mu^n_\cdot, \nu^n_\cdot)$ is uniformly bounded from above and*

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

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

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

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

*and*

$$[\theta]_1(dy|t) = [\theta]_2(dy|t) = \nu(dy|t),$$

$$[\zeta]_1(dy|t) = \nu(dy|t), \qquad [\zeta]_2(dy|t) = \mu(dy|t).$$

**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 [4, Lemma 3.3.1].

We next consider the representation for $\tilde{X}$, and use an argument similar to that of [4, Theorem 5.3.5]. For any $0 \leq j \leq n$, 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

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

where $\delta_0(dz)$ is the probability measure with mass 1 at zero. It is not difficult to see that $\lambda_j^n$ 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 Chebyshev's inequality and a conditioning argument,

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

The last quantity tends to zero as $n$ tends to infinity for each fixed $\delta > 0$ by Lemma A.2. Applying a standard submartingale inequality to $M^n - N^n$, we have

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

Sending first $n\to\infty$ and then $\delta\to 0$, it follows that for each $\varepsilon>0$

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

as $n\to\infty$. Thus

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

uniformly in $j\in\{1,\ldots,n\}$, in probability. Using the uniform integrability and weak convergence of $\mu^n$ 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 $\mu\left(dz\times ds\right)=\mu\left(dz|s\right)ds$, 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 $\theta^n$. 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

$$|I_n| = \frac{1}{n}\left|g(\tilde{Y}_0^n) - g(\tilde{Y}_n^n)\right| \leq \frac{2}{n}\|g\|_\infty$$

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 $\nu_j^n$ is the conditional distribution of $\tilde{Y}_{j+1}^n$, by Chebyshev's inequality and a conditioning argument, for $n \geq N_0$

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

Sending $n \to \infty$ and then $\varepsilon \to 0$, Fatou's Lemma and the arbitrariness of $g$ imply $[\theta]_{1,3} = [\theta]_{2,3}$ 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 $\gamma$, an analogous argument shows that, for any $\varepsilon > 0$ and bounded continuous functions $g_1, g_2$ on $S$, we have

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

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 $\gamma$ now follows from the weak convergence of $\gamma^n$ and $\nu^n$, Fatou's Lemma, the arbitrariness of $\varepsilon$, and the fact that product functions are convergence determining (see, for example, [4, Theorem A.3.14]).

The expressions for $\zeta$ and $\eta$ can be proved in the same way, and we omit the proof. ■

### A.4 Analysis of the cost.

We claim that $\liminf_{n\to\infty} \bar{J}^n$ [see equation (A.6)] is bounded below by

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

The bound for the first two relative entropy terms follows from the weak convergence, Fatou's Lemma, and the lower semicontinuity of relative entropy [4, Lemma 1.4.3]. The convergence of the next two terms follows from the weak convergence, the continuity and boundedness properties of the $\rho_k$ and $\bar{\alpha}_k$, and the Dominated Convergence Theorem. Lastly, we show that

$$\liminf_{n\to\infty}\tilde{E}\left[G(\tilde{X}^n(1))\right]\geq\tilde{E}\left[G(\tilde{X}(1))\right]. \tag{A.7}$$

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

$$G(x)=\bar{W}(x,1)\geq -C(\|x\|+1). \tag{A.8}$$

By Fatou's Lemma,

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

Since the uniform integrability of $\{\tilde{X}^n(1)\}$ proved in Lemma A.2 implies $\lim_{n\to\infty}\tilde{E}\|\tilde{X}^n(1)\|=\tilde{E}\|\tilde{X}(1)\|$, the inequality (A.7) follows.

Using the factorization properties of relative entropy (A.1), we now do some rewriting of the various terms. We have

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

However, by Lemma A.3, $[\theta]_1\left(\cdot\,|t\right)=[\theta]_2\left(\cdot\,|t\right)=\nu\left(\cdot\,|t\right)$, $\gamma(\cdot|t)=\nu(\cdot|t)\otimes p$, $\eta(\cdot|t)=\nu(\cdot|t)\otimes m$, and $\zeta(\cdot|t)=\nu(\cdot|t)\otimes q_t$ for some stochastic kernel $q_t$. It follows from the definition of $L$ in (4.2) that

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

where for each $t$,

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

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

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

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

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

## A.5 Verification.

We now do a classical verification argument to show $\Gamma \geq \bar{W}(0,0)$. By assumption (see Definition 6.1),

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

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

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

Since $G(x) = \bar{W}\,(x,1)$ we complete the proof of Theorem 8.1. ■

# B Appendix. Proof of Lemma A.2

The proof uses ideas from [4, Proposition 5.3.2]. We start by observing a few facts. Thanks to (A.8),

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

By Lemma 5.1, the eigenfunctions $r(y;\alpha)$ are bounded uniformly from above and below away from zero on $\{\alpha : \|\alpha\| \le C\}$, and $H(\alpha)$ is bounded from below on this set since $H$ is finite and convex (whence continuous). These and the non-negativity of relative entropy imply the existence of $C_1 < \infty$ and $C_2 < \infty$ such that

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

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

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

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

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

Under Condition 4.1, for each $\rho$ there is a finite and uniform bound $B(\rho)$ on $\int_{\mathbb{R}^d} e^{\rho\|z\|}m\left(dz\left|y\right.\right)$ for all $y \in S$. This allows us to continue the inequality as

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

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

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

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

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

Under Condition 4.1,

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

as $C \to \infty$. Thanks to the uniform bound (B.2), the first part of the lemma follows by first sending $C \to \infty$ and then $\rho \to \infty$.

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 $i/n$. It follows that if $\bar{X}^n$ converges in distribution in the sup norm to a limit $\tilde{X}$ then so does $\tilde{X}^n$, since the sup norm of $\tilde{X}^n - \bar{X}^n$ converges to 0 in probability. 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

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

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

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

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

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

and let $\delta \doteq (\varepsilon/2C) \wedge 1$. Write $f(t) = d\bar{X}^n(t)/dt$, then since $C\delta \leq \varepsilon/2$

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

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

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

This implies

$$\begin{aligned}
\|\tilde{X}^n(1)\|1_{\{\|\tilde{X}^n(1)\|\geq 2C\}} &\leq C1_{\{\|\tilde{X}^n(1)\|\geq 2C\}} + \frac{1}{n}\sum_{i=1}^n \left\|\tilde{b}_i^n\right\| 1_{\{\|\tilde{b}_i^n\|\geq C\}} \\
&\leq \frac{\|\tilde{X}^n(1)\|}{2}1_{\{\|\tilde{X}^n(1)\|\geq 2C\}} + \frac{1}{n}\sum_{i=1}^n \left\|\tilde{b}_i^n\right\| 1_{\{\|\tilde{b}_i^n\|\geq C\}},
\end{aligned}$$

or

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

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,

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

Uniform integrability holds since the last quantity tends to zero uniformly in $n$ as $C \to \infty$, and the tightness is a consequence of the uniform integrability [4, Theorem A.3.17]. ■

# C Appendix. The mixed open/closed network

Proof of Equation (10.12). For $\delta = (\delta_1, \delta_2) \in \mathbb{R}^2$ define

$$L_0(\delta) \doteq \lambda \ell \left( \frac{\delta_1}{\lambda} \right) + \mu_{12} \ell \left( \frac{\delta_2}{\mu_{12}} \right).$$

Similarly, for $\eta = (\eta_1, \eta_2) \in \mathbb{R}^2$ define

$$L_1(\eta) \doteq \inf \left\{ \lambda \ell \left( \frac{\hat{\lambda}_1}{\lambda} \right) + \mu_{11} \ell \left( \frac{\hat{\mu}_{11}}{\mu_{11}} \right) + \mu_2 \ell \left( \frac{\hat{\mu}_2}{\mu_2} \right) \right\},$$

where the infimum is taken over all $(\hat{\lambda}_1, \hat{\mu}_{11}, \hat{\mu}_2)$ such that

$$\hat{\lambda}_1 - \hat{\mu}_{11} = \eta_1, \quad - \mu_2 = \eta_2.$$

It is not difficult to show by direct computation that $L_i$ is the Legendre transform of $H_i$, for each $i = 0, 1$. Given $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2$, let

$$Q(\theta) \doteq \inf \left\{ \rho_0 L_0(\delta) + \rho_1 L_1(\eta) : \rho_0 \geq 0, \rho_1 \geq 0, \rho_0 + \rho_1 = 1, \rho_0 \delta + \rho_1 \eta = \theta \right\}.$$

Thanks to [4, Corollary D.4.3], $G$ is the Legendre transform of $H_0 \vee H_1$. But it is easy to see by definition that $L(\beta) = Q(\beta, 0)$, whence $L$ is the Legendre transform of $H$.

Proof of Equations (10.13) – (10.15). For every fixed $\alpha$, $H_0(\alpha, q)$ is a strictly increasing function of $q$ with $\lim_{q \to \infty} H_0(\alpha, q) = +\infty$ and $H_1(\alpha, q)$ is a strictly decreasing function of $q$ with $\lim_{q \to -\infty} H_1(\alpha, q) = +\infty$. It follows that, for each fixed $\alpha$, there exists a unique $q = q(\alpha)$ such that

$$H_0(\alpha, q(\alpha)) = H_1(\alpha, q(\alpha)) = H(\alpha).$$

Therefore, solving $H(\alpha) = 0$ is equivalent to finding $(\alpha, q)$ such that

$$H_0(\alpha, q) = H_1(\alpha, q) = 0,$$

which yields the $\gamma$ of (10.13) and the $q^*$ of (10.15). Thus $(\gamma, q^*)$ satisfies

$$H_0(\gamma, q^*) = H_1(\gamma, q^*) = 0. \tag{C.1}$$

(It turns out that $(0, 0)$ is also a solution, but it is elementary that this root does not characterize the relevant solution to the PDE.) The computation for $\Theta^*$ is straightforward and thus omitted.

Proof that $(\bar{W}, \Theta^*)$ is a subsolution/control pair. We only need to show that

$$\inf_{\beta} \left[ -2\gamma\beta + \bar{L}(\beta, \Theta^*) \right] \geq 0. \tag{C.2}$$

However, analogous to the proof of equation (10.12), one can show that $\bar{L}(\beta, \Theta^*)$ is convex with respect to $\beta$ and its Legendre transform is

$$\bar{H}(\alpha) = \inf_q \left[ \bar{H}_0(\alpha, q) \vee \bar{H}_1(\alpha, q) \right]$$

with

$$\begin{aligned} \bar{H}_0(\alpha, q) &= \left( \frac{\lambda^2}{\lambda_0^*} e^{\alpha} - 2\lambda + \lambda_0^* \right) + \left( \frac{\mu_{12}^2}{\mu_{12}^*} e^{q} - 2\mu_{12} + \mu_{12}^* \right) \\ \bar{H}_1(\alpha, q) &= \left( \frac{\lambda^2}{\lambda_1^*} e^{\alpha} - 2\lambda + \lambda_1^* \right) + \left( \frac{\mu_{11}^2}{\mu_{11}^*} e^{-\alpha} - 2\mu_{11} + \mu_{11}^* \right) + \left( \frac{\mu_2^2}{\mu_2^*} e^{-q} - 2\mu_2 + \mu_2^* \right). \end{aligned}$$

Then the inequality (C.2) reduces to $-\bar{H}(2\gamma) \geq 0$. Indeed, we claim that $\bar{H}(2\gamma) = 0$. To this end, note that by direction computation, equations (10.14) and (C.1), we have

$$\bar{H}_0(2\gamma, 2q^*) = 2\lambda(e^{\gamma} - 1) + \mu_{12}(e^{q^*} - 1) = 2H_0(\gamma, q^*) = 0,$$

and

$$\bar{H}_1(2\gamma, 2q^*) = 2\lambda(e^{\gamma} - 1) + 2\mu_{11}(e^{-\gamma} - 1) + 2\mu_2(e^{-q^*} - 1) = 2H_1(\gamma, q^*) = 0.$$

It is now very easy to argue that

$$\bar{H}(2\gamma) = \inf_q \left[ \bar{H}_0(2\gamma, q) \vee H_1(2\gamma, q) \right] = \bar{H}_0(2\gamma, 2q^*) \vee H_1(2\gamma, 2q^*) = 0,$$

which completes the proof. ■

# References

[1] S. Asmussen. *Ruin Probabilities.* World Scientific, Singapore, 2000.

[2] H.P. Chan and T.L. Lai. Efficient importance sampling for monte carlo evaluation of exceedance probabilities. *Preprint*, 2006.

[3] J. F. Collamore. Importance sampling techniques for the multidimensional ruin problem for general Markov additive sequences of random vectors. *Ann. Appl. Prob.*, 12:382–421, 2002.

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

[5] P. Dupuis, A. Sezer, and H. Wang. Dynamic importance sampling for queueing networks. *Preprint*, 2005.

[6] P. Dupuis and H. Wang. Importance sampling, large deviations, and differential games. *Stoch. and Stoch. Reports.*, 76:481–508, 2004.

[7] P. Dupuis and H. Wang. Dynamic importance sampling for uniformly recurrent Markov chains. *Ann. Appl. Prob.*, 15:1–38, 2005.

[8] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling: Convergence analysis. *Preprint*, 2005.

[9] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer–Verlag, Berlin, second edition, 1983.

[10] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. *Operations Research*, 47:585–600, 1999.

[11] P. Glasserman and S. Kou. Analysis of an importance sampling estimator for tandem queues. *ACM Trans. Modeling Comp. Simulation*, 4:22–42, 1995.

[12] P. Glasserman and Y. Wang. Counter examples in importance sampling for large deviations probabilities. *Ann. Appl. Prob.*, 7:731–746, 1997.

[13] I. Iscoe, P. Ney, and E. Nummelin. Large deviations of uniformly recurrent Markov additive processes. *Adv. Appl. Math.*, 6:373–412, 1985.

[14] T. Lehtonen and H. Nyhrinen. Simulating level crossing probabilities by importance sampling. *Adv. Appl. Probab.*, 24:858–874, 1992a.

[15] D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.*, 4:673–684, 1976.

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