On Lyapunov Inequalities and Subsolutions for Efficient Importance Sampling

By Jose Blanchet, Peter Glynn and Kevin Leder
Columbia University, Stanford University and Columbia University

December 21, 2009

Abstract

In this paper we explain some connections between Lyapunov methods and subsolutions of an associated Isaacs equation for the design of efficient importance sampling schemes. As we shall see, subsolutions can be derived by taking an appropriate limit of an associated Lyapunov inequality. They have been recently proposed (see e.g. [13, 15]) and applied to address several important problems in rare-event simulation ([16, 12]). Lyapunov inequalities have been used for testing the efficiency of state-dependent importance sampling schemes in heavy-tailed or discrete settings ([6, 7, 4]). While subsolutions provide an analytic criterion for the construction of efficient samplers, Lyapunov inequalities are useful for finding more precise information, in the form of bounds, for the behavior of the coefficient of variation of the associated importance sampling estimator in the prelimit. In addition, Lyapunov inequalities allow to gain insight into the various mollification procedures that are often required in constructing associated subsolutions. Our aim is to demonstrate that applying Lyapunov inequalities for verification of efficiency can help both, guide the selection of various mollification parameters and sharpen the information on the efficiency gain induced by the sampler.

1 Introduction

In recent years, there has been a substantial amount of research related to the efficient design of state-dependent importance sampling estimators. This research has been motivated by a series of examples and counter-examples (see [18] and [19]) related to the use of large deviations principles in the design of efficient importance sampling estimators for light-tailed systems, and also by the development of efficient importance sampling estimators for heavy-tailed systems, which as explained in [3] must often be state-dependent (see also [2] for more on the challenges that arise in the context of efficient rare-event simulation for heavy-tailed systems).

In order to systematically address the construction of efficient importance sampling estimators that can be rigorously shown to be efficient in large deviations environments for light-tailed systems P. Dupuis, H. Wang and their students have developed a method based

on control theory and the use of subsolutions of an associated Isaacs equation for constructing and testing asymptotically optimal importance sampling estimators ([13, 15, 12, 11]). A related approach, based on the construction of Lyapunov inequalities has been used in ([6, 8, 7]) for the construction and analysis of state-dependent importance sampling estimators for heavy-tailed systems. The Lyapunov method has also been used for counting bipartite graphs with a given degree sequence in [4]).

In this paper we discuss connections between subsolutions of associated Isaacs equations and Lyapunov inequalities. We focus on situations involving light-tailed systems because the Isaacs equation is proposed by a limiting procedure which makes sense only in light-tailed settings under an appropriate large deviations scaling. The Isaacs equation arises in the context of zero-sum games. As pointed out by [13], such equation comes up in the analysis of importance sampling estimators by the introduction of an artificial player arising from a variational representation obtained by working with the logarithm of the value function of an associated control problem.

In Section 3 we derive the form of a basic Lyapunov inequality for the analysis of state-dependent importance sampling algorithms for estimating first-passage time probabilities. We will first derive the form of the associated Isaacs equation as an approximation to the zero-variance change-of-measure in Section 4, we refer the reader to [13] for the original derivation involving control theory techniques and for the game theoretic interpretation. Then, as we shall see, a subsolution of the Isaacs equation can be obtained as a limiting process starting from a suitable Lyapunov inequality by fixing the family of importance sampling distributions as that of exponential changes-of-measure. The limiting process, which connects the verification procedure of efficiency of a given importance sampling estimator with the fluid scales at which the large deviations rates are obtained, is necessary for the derivation of the subsolutions. Since the limiting process requires the existence of certain derivatives, then the associated subsolutions must satisfy certain smoothness properties such as continuous differentiability. Often, when the associated dynamics are of random walk type, the subsolutions are obtained as the infimum of finitely many affine functions (see [13]) and therefore are not continuously differentiable. The approach, suggested in ([13]), is to mollify the subsolution in order to obtain the required differentiability. The mollification parameters, as we shall see, seem to play a substantial role in the efficiency of the algorithm. It is important to stress that subsolutions provide only (asymptotic) upper bounds for the second moment of the underlying estimator. The efficiency (also known as asymptotic optimality) is enforced by a judicious construction of the subsolution and the boundary conditions.

As we shall see in Section 5, the direct application of Lyapunov inequalities, without going through the process of taking the logarithm to work with the rate function, often yields a verification approach that is free of a limiting procedure. As a consequence, no additional mollification parameter is required, which in particular implies that more information on the performance of the importance sampling estimator can be obtained. The construction of Lyapunov inequalities often requires some ingenuity. However, we believe it is often the case that subsolutions to the associated Isaacs equation can be used to construct an appropriate Lyapunov function which, as we indicated before, can provide more precise information about the performance of the algorithm.

In the second part of Section 5 we illustrate how, by avoiding the use of a mollification

parameter and working directly with Lyapunov inequalities, one can construct strongly efficient estimators for a large class of multidimensional first-passage time problems for Markov random walks.

Finally, in order to illustrate how the direct use of Lyapunov inequalities can provide useful information on the performance of state-dependent importance sampling estimators for light-tailed systems, in Section 6 we consider the class of Markovian tandem networks. We revisit the estimators proposed by [12]. Our analysis provides a refinement on the performance analysis of such estimators and guides the selection of the various mollification parameters involved in the construction of their sampler. Our analysis in Section 6 allows us to conclude that, in common circumstances, a suitably mollified version of the estimator from [12] has a computational complexity that is polynomial with a rate that is not worse than solving the associated linear system of equations corresponding to the exact solution of the overflow probability in question. In contrast, the analysis presented in [12] just concludes subexponential complexity of their proposed family of importance sampling schemes.

The relationship that we see between Lyapunov inequalities and the subsolution of associated Isaacs equation is a symbiotic one, each approach benefits from the other. The technique of Lyapunov inequalities is a very general (and powerful) approach for establishing stability of dynamic systems (stochastic and deterministic). If one can find the appropriate Lyapunov function then the analysis of importance sampling estimators is easily done with a Lyapunov inequality. Sometimes, as in the heavy-tailed settings illustrated in [8, 7], one can obtain a reasonable guess the from the form of the Lyapunov function by using fluid heuristics analysis. In the light-tailed case, however, it can be quite difficult to find the appropriate function. Fortunately, the work of Dupuis et. al. on subsolutions to an associated Isaacs equation can help resolve this issue in light tailed systems because the integral equations that arise in the corresponding Lyapunov inequalities are replaced by differential inequalities which are often easier to solve using large deviations fluid analysis. We refer the reader to the cited work of Dupuis et. al. for the illustration via examples in queueing networks of how the fluid analysis of large deviations paths suggests, via the evaluation of the optimal-path gradients, which are often piecewise constant, the construction of an appropriate weak sense subsolution. In the examples we present in this paper, suitable Lyapunov functions are constructed based on the form suggested by the subsolution approach of Dupuis et al. Our focus is on constructing Lyapunov functions that improve the convergence rates of the estimators in the prelimit and help guide the selection of various importance sampling parameters.

The rest of the paper is organized as follows...

2 Basic Notions on Efficiency of Rare-event Simulation Estimators

Here we describe what is strong efficiency, weak efficiency and other notions of complexity. We use the notation $\alpha_n = P(A_n)$ and assume that $\alpha_n \longrightarrow 0$ as $n \nearrow \infty$. We concentrate on unbiased estimators. An estimator R_n is said to be strongly efficient if $ER_n = \alpha_n$ and

$$ER_n^2 = O\left(\alpha_n^2\right)$$

as $n \nearrow \infty$. The estimator is *weakly efficient* or *asymptotically optimal* if for each $\varepsilon > 0$ we have that

$$ER_n^2 = O\left(\alpha_n^{2-\varepsilon}\right)$$

as $n \nearrow \infty$. Finally, we say that the estimator has *polynomial complexity* of order at most $l \geq 0$ if

$$ER_n^2 = O\left(\alpha_n^2 \log\left(1/\alpha_n\right)^{2l}\right), \tag{1}$$

in other words, if its relative mean squared error (or coefficient of variation) grows at most at a polynomial rate with degree l in $\log\left(1/\alpha_n\right)$. The estimator is said to be *strongly efficient* if $l = 0$ in (1).

In most cases the analysis of importance sampling estimators (specially state-dependent samplers in light-tailed cases) concludes only weak efficiency of the estimators. As we shall see in future sections, the use of Lyapunov inequalities often allows to lift such weak efficiency results to more precise statements involving polynomial complexity.

Suppose that an estimator R_n has polynomial complexity of order at most l and consider $Z_{n,m} = \sum_{i=1}^{m} R_n\left(i\right)/m$ where the $R_n\left(i\right)$'s are independent replications of R_n. It follows from Chebyshev's inequality that at most $m = O\left(\varepsilon^{-2}\delta^{-1}\log\left(1/\alpha_n\right)^{2l}\right)$ replications are required to conclude that $Z_{n,m}$ is ε-close to α_n in relative terms with at least $1-\delta$ confidence.

3 State-dependent Importance Sampling and Lyapunov Inequalities

Let $W = \left(W_n : n \geq 0\right)$ be a Markov chain taking values on some state space $\mathcal{S}$. Given disjoint sets $A, B \subseteq \mathcal{S}$ define $T = \inf\{n \geq 0 : W_n \in A \cup B\}$ and non-negative (measurable) function $f\left(\cdot\right)$. Suppose we are interested in developing bounds for the variance of a given state-dependent importance sampling estimator for the expectation

$$u\left(w\right) = E_w\left(f\left(W_T\right)\mathbb{I}\left(T < \infty\right)\right),$$

where $E_w\left(\cdot\right)$ denotes the expectation operator in path space associated to the Markov process W given that $W_0 = w$.

Let us write $K\left(x, dy\right)$ to denote the transition kernel associated to W; in other words, $P_w\left(W_1 \in A\right) = \int_A K\left(w, dy\right)$. A state-dependent importance sampling change-of-measure is specified by a Markov transition kernel $\widetilde{K}\left(x, dy\right)$ defined via

$$\widetilde{K}\left(x, dy\right) = \frac{1}{r\left(x, y\right)} K\left(x, dy\right). \tag{2}$$

for some positive function $r\left(\cdot\right)$ such that $\int r\left(x, y\right)^{-1} K\left(x, dy\right) = 1$ for each $x \in \mathcal{S}$. We use $\widetilde{P}_w\left(\cdot\right)$ to define the path-measure induced by the kernel $\widetilde{K}\left(\cdot\right)$ given the initial condition w.

A state-dependent importance sampling estimator for $g\left(w\right)$ takes the form

$$Y = \prod_{k=0}^{T-1} r\left(\widetilde{W}_k, \widetilde{W}_{k+1}\right) f\left(\widetilde{W}_T\right) I\left(T < \infty\right),$$

where the notation $\widetilde{W} = \left(\widetilde{W}_k : k \geq 0\right)$ has been introduced to emphasize that the process follows the law induced by the Markov kernel $\widetilde{K}(\cdot)$ assuming $\widetilde{W}_0 = w$. It is clear that $u(w) = \widetilde{E}_w(Y)$, so Y is unbiased. The performance of the estimator Y can then be measured in terms of its mean squared error, which in turn is bounded by its second moment, namely,

$$\begin{aligned} s_r(w) &= \widetilde{E}_w\left(Y^2\right) \\ &= \widetilde{E}_w\left(\prod_{k=0}^{T-1} r\left(\widetilde{W}_k, \widetilde{W}_{k+1}\right)^2 f\left(\widetilde{W}_T\right)^2 \mathbb{I}(T<\infty)\right) \\ &= E_w\left(\prod_{k=0}^{T-1} r(W_k, W_{k+1})\, f(W_T)^2\, \mathbb{I}(T<\infty)\right). \end{aligned}$$

The following inequality allows to find a bound for $s_r(\cdot)$.

Lemma 1 *Suppose that there exists a non-negative function $v(\cdot)$, a constant $\rho > 0$ and $\delta \geq 0$ such that*

$$v(w)\exp(\delta) \geq E_w r(w, W_1)\, v(W_1)$$

for $w \notin A \cup B$ and $v\left(\widetilde{W}_T\right)\mathbb{I}(T<\infty) \geq \rho f\left(\widetilde{W}_T\right)^2 \mathbb{I}(T<\infty)$. Then,

$$\frac{v(w)}{\rho} \geq \widetilde{E}_w\left(\exp(-\delta T)\prod_{k=0}^{T-1} r\left(\widetilde{W}_k, \widetilde{W}_{k+1}\right)^2 f\left(\widetilde{W}_T\right)^2 \mathbb{I}(T<\infty)\right).$$

In addition, if $T \leq m/\delta$ for some $m \in (0,\infty)$, then

$$v(w)\exp(m)/\rho \geq s_r(w).$$

Proof. The proof proceeds along the same lines as [6], we provide the details in order to make the discussion self-contained. First, define

$$M_n = v(W_{T\wedge n}) \prod_{k=0}^{T\wedge n-1} \left(e^{-\delta} r(W_k, W_{k+1})\right).$$

We note that M_n is a non-negative supermartingale adapted to the filtration $(\mathcal{F}_n : n \geq 0)$ generated by W. To see this, observe that

$$\begin{aligned} &E_w\left(M_{n+1}\middle|\mathcal{F}_n\right)\mathbb{I}(T>n) \\ &= \prod_{k=0}^{n-1}\left(e^{-\delta} r(W_k, W_{k+1})\right) E_w\left(v(W_{n+1})\, e^{-\delta} r(W_n, W_{n+1})\middle|\mathcal{F}_n\right)\mathbb{I}(T>n) \\ &\leq v(W_n)\prod_{k=0}^{n-1} e^{-\delta} r(W_k, W_{k+1})\,\mathbb{I}(T>n) = M_n \mathbb{I}(T>n). \end{aligned}$$

Therefore,

$$E_w\left(M_{n+1}\middle|\mathcal{F}_n\right) = E_w\left(M_{n+1}\middle|\mathcal{F}_n\right)\mathbb{I}\left(T\leq n\right) + E_w\left(M_{n+1}\middle|\mathcal{F}_n\right)\mathbb{I}\left(T>n\right)$$
$$\leq M_n\mathbb{I}\left(T\leq n\right) + M_n\mathbb{I}\left(T>n\right) = M_n.$$

We conclude that

$$v\left(w\right) = M_0 \geq E_w M_n \geq E_w v\left(W_T\right)\prod_{k=0}^{T-1}\left(e^{-\delta}r\left(W_k,W_{k+1}\right)\right)\mathbb{I}\left(T\leq n\right)$$
$$\geq E_w \rho f\left(W_T\right)^2\exp\left(-\delta T\right)\prod_{k=0}^{T-1} r\left(W_k,W_{k+1}\right)\mathbb{I}\left(T\leq n\right).$$

Sending $n\nearrow\infty$ and using monotone convergence we obtain that

$$v\left(w\right)/\rho \geq E_w f\left(W_T\right)^2\exp\left(-\delta T\right)\prod_{k=0}^{T-1} r\left(W_k,W_{k+1}\right)\mathbb{I}\left(T<\infty\right)\left(w\right).$$

The rest of the lemma follows easily from the previous expression. ■

Remark 1 *Note that we have introduced a parameter $\delta\geq 0$ in the statement of the previous Lyapunov inequality. Ideally, we would like to construct $v\left(\cdot\right)$ based on the selection $\delta=0$; however, as we shall see below, it is often the case that introducing the parameter $\delta>0$ facilitates the construction of such a Lyapunov function $v\left(\cdot\right)$. Then, one has to deduce the inequality for $s_r\left(\cdot\right)$ using the problem structure, for instance, that $T\leq m/\delta$, other examples are discussed in future sections.*

Remark 2 *It is important to note that we can select $r\left(w_0,w_1\right)^{-1} = u\left(w_1\right)/u\left(w_0\right)$, in which case it follows that $v\left(w\right)=u\left(w\right)^2$ satisfies the Lyapunov inequality from the previous lemma with strict equality (selecting $\delta=0$ and $\rho=1$) and we have that*

$$s_r\left(w\right) = E_w\left(\frac{u\left(w\right)}{u\left(W_T\right)}f\left(W_T\right)^2\mathbb{I}\left(T<\infty\right)\right)$$
$$= u\left(w\right)E_w\left(f\left(W_T\right)\mathbb{I}\left(T<\infty\right)\right) = u\left(w\right)^2.$$

As a consequence selecting $r\left(w_0,w_1\right)^{-1} = u\left(w_1\right)/u\left(w_0\right)$ describes the zero-variance change-of-measure.

If one suspects that a good family of importance sampling schemes can be constructed from a specific parametric family of transition kernels, say $\{\widetilde{K}_\theta\left(x,dy\right) : x,y\in\mathcal{S}\}$ for θ in some set $\mathcal{A}$. Then, it is natural to consider the question of performing the construction by minimizing the second moment of the associated importance sampling estimator. More precisely, following the notation introduced in (2), let us write

$$\widetilde{K}_\theta\left(x,dy\right) = \frac{1}{r_\theta\left(x,y\right)}K\left(x,dy\right),$$

for an appropriate function $r_\theta(\cdot)$. Then, the Hamilton-Jacobi-Bellman equation associated to the control problem of minimizing the second moment of the estimator takes the form

$$U(x) = \min_{\theta \in \mathcal{A}} E_x[r_\theta(x, W_1) U(W_1)], \tag{3}$$

subject to the boundary condition that $U(x) = f(x)^2$ for $x \in A \cup B$. Observe, that given any policy $(\widetilde{\theta}(x) : x \in \mathcal{S})$ (optimal or not) Lemma 1 provides means for computing bound on the the second moment of the associated importance sampling estimator. If the problem has a suitable asymptotic structure, as we shall see occurs for instance in the light-tailed large deviations settings that we shall discuss in future sections, then one might try to attempt solving the asymptotic control problem and use the asymptotic solution to guide the construction of the importance sampling estimator. The performance of the asymptotically optimal policy can then be tested in the prelimit using Lemma 1 by constructing an appropriate Lyapunov functions. This is precisely what we shall illustrate in future sections.

4 Isaacs Equation, Subsolutions and Lyapunov Inequalities

We shall consider the setting of state-dependent random walks. Consider a family of systems $Y^{(\Delta)} = (Y_t : t \in \{0, \Delta, 2\Delta, ...\})$, indexed by the parameter $\Delta > 0$, and taking values in a subset $\mathcal{S}$ of $\mathbb{R}^d$ such that

$$Y_{t+\Delta} = Y_t + \Delta V_{t+\Delta}(Y_t),$$

we assume that the $V_t(y)$'s are i.i.d. r.v.'s (independent and identically distributed random variables) depending on the current position y. We have suppressed the explicit dependence of the Y_t's on the parameter Δ in order to simplify the notation. The constant $\Delta > 0$ is a scaling parameter which we shall send to zero in order to obtain corresponding large deviations estimates.

Define the log-moment generating function

$$\psi(\theta, y) = \log E \exp\left(\theta^T V_t(y)\right)$$

and assume that for each $y \in S$, $\psi(\theta, y) < \infty$ for all $\theta \in \mathbb{R}^d$ (this condition can be relaxed under special structure, for instance, if the $V_t(y)$'s do not depend on y). Settings in which the log-moment generating function is assumed to be finite, as in our case, are said to be light-tailed.

Under suitable regularity conditions, for instance if $\mu(\cdot) = EV_0(\cdot)$ is a Lipschitz function with at most linear growth (see e.g. [10]), it turns out that as $\Delta \searrow 0$, we have that $Y^\Delta \longrightarrow \overline{y}$ in probability over compact time intervals, where $\overline{y} = (\overline{y}(t) : t > 0)$ satisfies

$$\frac{d\overline{y}(t)}{dt} = EV_0(\overline{y}(t)), \quad \overline{y}(0) = y_0.$$

The large deviations behavior as $\Delta \searrow 0$ (i.e. probabilities associated to sample paths that deviate from $\overline{y}$ as $\Delta \searrow 0$) is dictated by the so-called rate function of the system, which is computed via

$$J(w) = \int_0^\infty I(w(s), \dot{w}(s))\, ds,$$

where

$$I(w(s), \dot{w}(s)) = \max_\theta [\theta \dot{w}(s) - \psi(\theta, w(s))].$$

In order to illustrate the use of large deviations analysis and its connections to importance sampling estimators and Lyapunov inequalities let us concentrate on a canonical example. Consider the problem of estimating

$$u_\Delta(y) = P_y^{(\Delta)}(T_A < T_B, T_{A\cup B} < \infty) = P_y^{(\Delta)}(Y_T \in A, T < \infty), \tag{4}$$

where, for any set C, $T_C = \inf\{t \geq 0 : Y_t \in C\}$. We put $T = T_{A\cup B}$ and note that the explicit dependence on Δ for the times T, T_A and T_B has been suppressed in order to simplify the notation. Assume that both A and B are open open sets for which $(\overline{y}(t) : t \geq 0) \cap A = \varnothing$ and $(\overline{y}(t) : t \geq 0) \cap B \neq \varnothing$. We then have, under suitable assumptions, that (see e.g. [17])

$$-\Delta \log u_\Delta(y) \longrightarrow h(y) \tag{5}$$

as $\Delta \searrow 0$ where

$$h(y) = \inf_{w(\cdot)\in C} J(w),$$

where the infimum is taken over the set C of absolutely continuous functions such that $w(0) = y$, $w(t) \in A$ for some $t < \infty$ and $w(s) \notin A \cup B$ for $s < t$.

Our goal is to find a family of state-dependent importance sampling estimators to calculate $u_\Delta(x)$. Note that the problem of estimating $u_\Delta(y)$ fits the framework described in the previous section, to see this, just let $f(w) = I(w \in A)$.

A family of importance sampling distributions that is often used in the context of light-tailed problems (i.e. in the case in which $\psi(\theta, y) < \infty$ as in our case) is given by exponential changes-of-measure. In other words, given that $Y_t = y$, the increment $V_{t+\Delta}(y)$ is sampled according to the distribution

$$P_{\theta(y)}(V_{t+\Delta}(y) \in v + dv) = \exp\left(\theta(y)^T v - \psi(\theta(y), y)\right) P(V_{t+\Delta}(y) \in v + dv).$$

Note that the parameter $\theta(y)$ is allowed to depend on the current state y of the process. A larger class of importance sampling distributions, which are quite useful, as we shall indicate below, is given by state-dependent mixtures of exponential changes-of-measure. In other words,

$$\frac{\widetilde{P}(V_{t+\Delta}(y) \in v + dv)}{P(V_{t+\Delta}(y) \in v + dv)} = \sum_{j=1}^{l} r_j(y) \exp\left(\theta_j(y)^T v - \psi(\theta_j(y), y)\right), \tag{6}$$

where $r_j(y) \geq 0$ and $\sum_{j=1}^{l} r_j(y) = 1$. For the moment, we shall concentrate on the use of standard exponential changes-of-measure but we shall return to the use of mixtures later in the paper.

Given the previously indicated large deviations results, it is natural to pursue the strategy outlined at the end of the previous section, namely, asymptotically solving the problem (3). In particular, let $-\Delta \log U_\Delta(x) = H_\Delta(x)$ and note that (3) takes the form

$$\exp\left(-H_\Delta(y)/\Delta\right) = \inf_{\theta(y)\in\mathbb{R}^d} E_y\{\exp[-\theta(y)^T V(y) + \psi(\theta(y), y)]\exp[-H_\Delta(y + \Delta V(y))/\Delta]\}, \tag{7}$$

subject to $H_\Delta(x) = \infty I(x \notin A)$ for $x \in A \cup B$ (with the convention that $0\cdot\infty = 0$). The large deviations theory mentioned before suggests that $H_\Delta(x) \approx H(x)$ for some $H(\cdot)$ as $\Delta \nearrow 0$ and that, consequently, there is a corresponding asymptotic Hamilton-Jacoby-Bellman as $\Delta \longrightarrow 0$ for $H(x)$. In particular, formally applying these considerations into (7) we obtain

$$\begin{aligned}
1 &= \inf_{\theta(y)\in\mathbb{R}^d} E_y\{\exp[-\theta(y)^T V(y) + \psi(\theta(y), y)]\exp[-\{H_\Delta(y + \Delta V(y)) - H_\Delta(y)\}/\Delta]\} \\
&\approx \inf_{\theta(y)\in\mathbb{R}^d} E_y\{\exp[-\theta(y)^T V(y) + \psi(\theta(y), y)]\exp[-\{H(y + \Delta V(y)) - H(y)\}/\Delta]\} \\
&\approx \inf_{\theta(y)\in\mathbb{R}^d} E_y\{\exp[-\theta(y)^T V(y) + \psi(\theta(y), y)]\exp[-\partial_y H(y) V(y)]\} \\
&= \inf_{\theta(y)\in\mathbb{R}^d} \exp[\psi(-\theta(y) - \partial_y H(y), y) + \psi(\theta(y), y)].
\end{aligned}$$

Applying first order optimality conditions and solving for $\theta(y)$ we conclude that $2\theta(y) = -\partial_y H(y)$ and therefore

$$2\psi(-\partial_y H(y)/2, y) = 0, \tag{8}$$

subject to the boundary conditions that $H(x) = \infty I(x \notin A)$ for $x \in A \cup B$; this equation precisely corresponds to the Isaacs equation introduced by Dupuis and Wang. A natural importance sampling strategy then proceeds by applying the exponential tilting with tilting parameter $\theta(y) = \partial_y H(y)/2$. Existence of smooth classical solutions to the previous equation is a delicate issue, our intention is here to provide a formal explanation of the ideas behind the development of the Isaacs equation approach of Dupuis and Wang, at the end, the performance of the estimators discussed in our future examples will be rigorously tested using Lyapunov inequalities, so we decided to keep the discussion formal for pedagogical purposes here.

Now, a natural question that comes to mind is, how good we might expect the previous strategy to be. Let us now connect the Isaacs equation introduced earlier to the zero-variance change-of-measure. As we indicated in Remark 2 following Lemma 1, the zero-variance change-of-measure is selected by choosing $r(y_0, y_1)^{-1} = u_\Delta(y_1)/u_\Delta(y_0)$. We shall derive re-derive the associated Isaacs equation, which, as mentioned before dictates the optimal exponential change-of-measure as $\Delta \searrow 0$, as a natural approximation to the zero-variance change-of-measure. First, write

$$v_\Delta(x) = -\Delta^{-1}\log u_\Delta(x).$$

We shall use $K^*(x, y)$ to denote the zero-variance change-of-measure, that is

$$K^*(x, y) = P(x, y)\frac{u_\Delta(y)}{u_\Delta(x)},$$

which is a well defined transition kernel because $u_\Delta(x) = Eu_\Delta(x + \Delta V(x))$. Or, in terms of $\upsilon_\Delta(\cdot)$, we obtain

$$1 = E\exp[\Delta^{-1}(\upsilon_\Delta(x) - \upsilon_\Delta(x + \Delta V(x)))].$$

Assuming that $\upsilon_\Delta(x + \Delta v) = \upsilon_\Delta(x) + \Delta\partial_x\upsilon_\Delta(x) \cdot v + o(\Delta)$ and that $\upsilon_\Delta(x)$ and $\partial_x\upsilon_\Delta(x)$ converge in an appropriate sense to $\upsilon(x)$ and $\partial_x\upsilon(x)$, respectively we obtain that

$$1 = E\exp(-\partial_x\upsilon(x)V(x) + o(1)),$$

which yields

$$\psi(-\partial_x\upsilon(x), x) = 0. \tag{9}$$

Since $\upsilon_\Delta(x) = \infty I(x \notin A)$ on $A \cup B$, we also must have that $\upsilon_\Delta(x) = \infty I(x \notin A)$ on $A \cup B$. By letting $2\upsilon(x) = h(x)$ we see that (9) together with the associated boundary conditions correspond to the Isaacs equation (8) and, not surprisingly given the large deviations result (5), it is satisfied (in a weak sense) by setting $\upsilon(x) = h(x)$.

Now, let us connect the solutions of the Isaacs equation with its corresponding subsolutions via Lyapunov inequalities. The connection will allow us to see how to obtain importance sampling estimators for which bounds on their variance performance can be guaranteed by means of a suitable Lyapunov functions.

Suppose we have a function, say $(g(x) : x \in S)$ that is continuously differentiable. We wish to apply importance sampling according to exponential tiltings given by the gradient of $g(\cdot)$. In view of Lemma 1, we need to find a Lyapunov function $v(x)$ such that

$$E_x v(x + \Delta V(x))\exp(-\partial_x g(x)V(x) + \psi(-\partial_x g(x), x)) \leq v(x).$$

We select $v(x) = \exp(\Delta^{-1}g(x))$ and obtain that

$$E_x\exp\left(\Delta^{-1}(g(x + \Delta V(x)) - g(x)) - \partial_x g(x)V(x) + \psi(-\partial_x g(x), x)\right) \leq 1.$$

If $g(\cdot)$ is continuously differentiable, then we arrive at

$$\begin{aligned}
&\left(\Delta^{-1}(g(x + \Delta V(x)) - g(x)) - \partial_x g(x)V(x) + \psi(-\partial_x g(x), x)\right) \\
&= \partial_x g(x)V(x) + o(1) - \partial_x g(x)V(x) + \psi(-\partial_x g(x), x) \\
&= \psi(-\partial_x g(x), x) + o(1).
\end{aligned} \tag{10}$$

So, we require that $\psi(-\partial_x g(x), x) \leq 0$, which is the subsolution property described in [15]. Following Lemma 1, the corresponding boundary condition is $v(x) \geq I(x \in A)$ for $x \in A \cup B$, which is equivalent to $v(x) \geq \rho$ for $x \in A$ or $g(x) \geq \Delta\log(\rho)$ for $x \in A$.

Neglecting the contribution of order $o(1)$ is handled using a relaxed Lyapunov inequality introducing a parameter $\delta \geq 0$ as in Lemma 1. At the end one needs to deal with the fact that the strict Lyapunov inequality is not satisfied but it may be violated by an amount of order $o(1)$ as $\Delta \searrow 0$. Often, as we shall see in future sections, this can be done if we ensure that $o(1)$ is actually $o(\Delta)$ or even $O(\Delta)$ as $\Delta \searrow 0$.

It is important to note that the derivation above requires $g(\cdot)$ to be continuously differentiable, otherwise we cannot guarantee the estimate $o(1)$ above. Sometimes it is difficult or

might be even impossible to find a continuously differentiable subsolution to the Isaacs equation. However, we might consider a smooth sequence of functions that violate the Lyapunov bound by an amount of order $o(\Delta)$ as indicated above. These types of situations arise often in the context of queueing networks, as we shall discuss in future sections.

Remark 3 *The mixtures samplers of the form (6) are applied in the implementation of the subsolution approach as a smoothing technique, see Section 9.3 in [15]. However, as we shall see, by working directly with the mixtures themselves one can directly construct a Lyapunov function that provides a bound on the second moment of the estimator in the prelimit and that is often free from limiting procedures.*

5 Markov Random Walks

For simplicity we shall work with an irreducible and aperiodic finite state-space Markov chain $Z = (Z_j : j \geq 0)$ taking values on the space $\mathcal{S}$. The discussion that follows can be easily extended to more general Markov chains under appropriate geometric ergodicity assumptions. However, our objective here is to illustrate how to construct associated Lyapunov functions in order to test efficiency therefore, for pedagogical reasons, we prefer to keep the technical conditions at a minimum.

We introduce a sequence $(\xi_j : j \geq 1)$ of i.i.d. random variables and put $X_j = H(\xi_j, Z_j)$ (for a given measurable, function $H(\cdot)$ taking values on $\mathbb{R}^d$). Let us define

$$\phi(\theta, w) = E\left(\exp\left(\theta^T X_j\right) \middle| Z_j = w\right)$$

and assume that there exists $\delta > 0$ such that for all $w \in \mathcal{S}$, $\phi(\theta, z) < \infty$ for $\|\theta\| \leq \delta$. Moreover, let $(p(z_0, z_1) : z_0, z_1 \in \mathcal{S})$ be the probability transition matrix associated to Z and define $Q_\theta(z_0, z_1) = p(z_0, z_1)\phi(\theta, z_1)$. The Perron-Frobenius theorem (see [1]) indicates that there exists left and right strictly positive eigenvectors, π_θ and u_θ respectively, associated to a positive eigenvalue $\beta_\theta = \exp(\psi(\theta))$ such that

$$\pi_\theta Q_\theta = \beta_\theta \pi_\theta, \quad Q_\theta u_\theta = \beta_\theta u_\theta,$$

and for each $z \in \mathcal{S}$

$$\frac{1}{n}\log\left(Q_\theta^n \cdot \mathbf{1}\right)(z) = \frac{1}{n}\log E_z \exp\left(\theta^T S_n\right) \longrightarrow \psi(\theta)$$

as $n \nearrow \infty$. We shall assume that $\partial\psi(0) = 0$ and that $\psi(\cdot)$ is *steep* in the sense that for each $y \in \mathbb{R}^d$ there exists θ_y such that $\partial\psi(\theta_y) = y$ and we will put $I(y) = \theta_y^T y - \psi(\theta_y)$. Note that when X_j's are i.i.d. one can simply take $u_\theta = \mathbf{1}$ (a vector of ones) and $\psi(\theta) = \log E \exp\left(\theta^T X_1\right)$.

5.1 A Classical Time In-homogeneous Example

We first consider a classical problem in rare-event simulation, namely, evaluating

$$u_n(z_0) = P_{z_0}\left(S_n/n \in A_1 \cup ... \cup A_m\right),$$

where $A_1, A_2, ..., A_m$ are closed convex sets with non-empty interior that do not contain the origin. We put $I_{A_i} = \inf_{y \in A_i} I(y)$ and set $I = \min_{1 \leq i \leq m} I_{A_i}$. The i.i.d. case was first considered in [24]. The next result follows by adapting the analysis leading to (3.4) in [22] taking advantage of the approximations developed in [20].

Theorem 1 *In addition to the steepness assumption on $\psi(\cdot)$ indicated above, suppose that there exists z such that $H(\xi_j, z)$ is a continuous random variable. Then, there exists $c_-(z_0), c_+(z_0), n_0 > 0$ such that for all $n \geq n_0$*

$$c_-(z_0)/n^{d/2} \leq u_n(z_0) \exp(nI) \leq c_+(z_0)/n^{1/2}.$$

Our goal is to design a rare-event simulation algorithm whose coefficient of variation can be shown to grow at a polynomial rate in n. These types of problems have been considered by [14, 9]. A simple mixture based sampler can also be developed using the ideas explained in [21]. Our goal here is to make the connection to the use of subsolutions.

The idea is first to select a family of changes-of-measure based on mixtures. In particular, given $Z_{k-1} = z_{k-1}$, the next increment $X_k = H(\xi_k, Z_k)$ is simulated according to a finite mixture of transition laws of the form

$$\begin{aligned} &M_\theta(z_{k-1}, z_k, x + dx) \\ &= \frac{u(\theta, z_k)}{u(\theta, z_{k-1})} \exp(\theta x - \psi(\theta)) P(X_k \in x + dx | Z_k = z_k) p(z_{k-1}, z_k), \end{aligned}$$

where $u(\theta, z_k)$ is the z_k entry of the eigenvector u_θ. Note that such probabilistic law is properly defined as long as $\psi(\theta) < \infty$ in the sense that

$$\begin{aligned} &\sum_{z_k \in \mathcal{S}} \int \frac{u(\theta, z_k) p(z_{k-1}, z_k) \exp\left(\theta^T x - \psi(\theta)\right) P(X_k \in x + dx | Z_k = z_k)}{u(\theta, z_{k-1})} \\ &= \sum_{z_k \in \mathcal{S}} \frac{u(\theta, z_k) p(z_{k-1}, z_k) \exp(-\psi(\theta))}{u(\theta, z_{k-1})} \phi(\theta, z_k) \\ &= \frac{\exp(-\psi(\theta))}{u(\theta, z_{k-1})} \sum_{z_k \in \mathcal{S}} Q_\theta(z_{k-1}, z_k) u(\theta, z_k) = 1. \end{aligned}$$

More precisely, at time $k-1$, given $Z_{k-1} = z_{k-1}$ and $S_{k-1} = s_{k-1}$, we propose to sample the pair (Z_k, X_k) according to the transition law

$$\widetilde{P}(z_{k-1}, s_{k-1}, z_k, x + dx) = \sum_{j=1}^{m} r_k^{(j)}(z_{k-1}, s_{k-1}) M_{\theta_j}(z_{k-1}, z_k, x + dx),$$

where $r_j(\cdot) > 0$ and the θ_j's are appropriately selected in order to satisfy an appropriate Lyapunov inequality. Note that we postulate the use of m functions $(r_j(\cdot), 1 \leq j \leq m)$. Each of these functions and the θ_j's will correspond to each of the A_j's. In particular, because of convexity, there exists a unique y_j such that $I_{A_j} = I(y_j)$, we then put $\theta_j = \theta_{y_j}$. The selection of r_j's proceeds as follows.

The corresponding Lyapunov inequality can be obtained by means of Lemma 1 by defining $W_k = (k, Z_k, S_k/n,)$, $A = \{n\} \times \mathcal{S} \times A_1 \cup ... \cup A_m$ and $B = \{n\} \times \mathcal{S} \times (A_1 \cup ... \cup A_m)^c$. Applying Lemma 1 yields that we must find a non-negative function $v_k(z, s)$ such that

$$E_z \frac{v_{k+1}(Z_1, s + X_1/n)}{v_k(z, s)} \frac{1}{\sum_{j=1}^m r_k^{(j)}(z, s)\, u(\theta_j, Z_1) \exp\left(\theta_j^T X_1 - \psi(\theta_j)\right) / u(\theta_j, z)} \leq 1$$

subject to $v_n(z, s) \geq \rho > 0$ for each $z \in \mathcal{S}$ and $s \in A_1 \cup ... \cup A_m$.

We first introduce non-negative functions $w_k^{(j)}(z, s)$ for $1 \leq j \leq m$ (to be specified later) and put

$$r_k^{(j)}(z, s) = \frac{w_k^{(j)}(z, s)}{\sum_{i=1}^m w_k^{(i)}(z, s)}.$$

We propose $v_k(z, s) = \left(\sum_{j=1}^m w_k^{(j)}(z, s)\right)^2$ so that

$$\begin{aligned} &E_z \frac{v_{k+1}(Z_1, s + X_1)}{v_k(z, s)} \frac{1}{\sum_{j=1}^m r_k^{(j)}(z, s)\, u(\theta_j, Z_1) \exp\left(\theta_j^T X_1 - \psi(\theta)\right) / u(\theta_j, z)} \\ &= E_z \frac{1}{\sum_{j=1}^m w_k^{(j)}(z, s)} \frac{\left(\sum_{j=1}^m w_{k+1}^{(j)}(Z_1, s + X_1)\right)^2}{\sum_{j=1}^m w_k^{(j)}(z, s)\, u(\theta_j, Z_1) \exp\left(\theta_j^T X_1 - \psi(\theta_j)\right) / u(\theta_j, z)}. \end{aligned} \tag{11}$$

Then, select $w_k^{(j)}(z, s)$ such that

$$w_{k+1}^{(j)}(Z_1, s + X_1/n) = w_k^{(j)}(z, s) \frac{u(\theta_j, Z_1)}{u(\theta_j, z)} \exp\left(\theta_j^T X_1 - \psi(\theta_j)\right). \tag{12}$$

In particular, note that

$$w_k^{(j)}(z, s) = \exp\left(n\theta_j^T s - k\psi(\theta_j) + a_j(n)\right) u(\theta_j, z) \tag{13}$$

satisfies (12) for any $a_j(n) \in \mathbb{R}$. The constants $(a_j(n) : 1 \leq j \leq m)$ are selected depending on n in order to both, satisfy the boundary conditions and minimize the bound on the second moment implied by the Lyapunov inequality (i.e. the value of $v_0(z, 0)$). Note that selecting $w_k^{(j)}$ as indicated in (13) yields that (11) equals

$$\begin{aligned} \sum_{j=1}^m E_z \frac{w_{k+1}^{(j)}(Z_1, s + X_1)}{\sum_{j=1}^m w_k^{(j)}(z, s)} &= \sum_{j=1}^m r_k^{(j)}(z, s)\, E_z \frac{u(\theta_j, Z_1)}{u(\theta_j, z)} \exp\left(\theta_j^T X_1 - \psi(\theta_j)\right) \\ &= \sum_{j=1}^m r_k^{(j)}(z, s) = 1 \end{aligned}$$

and therefore the Lyapunov inequality is satisfied assuming that the boundary condition holds given an appropriate selection for the $a_j(n)$'s. In order to appropriately select the $a_j(n)$'s we note that, by definition of y_j,

$$0 \leq \theta_j^T y_j - \psi(\theta_j) \leq \theta_j^T s - \psi(\theta_j)$$

for all $s \in A_j$. Therefore, selecting $a_j(n) = -n\left(\theta_j^T y_j - \psi(\theta_j)\right)$, yields

$$\begin{aligned} w_k^{(j)}(z,s) &= \exp\left(n\theta_j^T s - k\psi(\theta_j) - n\left(\theta_j^T y_j - \psi(\theta_j)\right)\right) u(\theta_j, z) \\ &= \exp\left(n\theta_j^T (s - y_j) + (n-k)\psi(\theta_j)\right) u(\theta_j, z). \end{aligned}$$

with $w_n^{(j)}(z,s) \geq u(\theta_j, z)$ for each $s \in A_j$. This implies that if $s \in A_1 \cup ... \cup A_m$, then

$$v_n(z,s) \geq \min_{1 \leq j \leq m, z \in \mathcal{S}} u(\theta_j, z)^2 = \rho.$$

In particular, Lemma 1 yields that the second moment of the corresponding importance sampling estimator, namely,

$$R = \prod_{k=1}^{n} \frac{\mathbb{I}(S_n/n \in A_1 \cup ... \cup A_m)}{\sum_{j=1}^m r_k^{(j)}(Z_k, S_k)\, u(\theta_j, Z_k) \exp\left(\theta_j^T X_k - \psi(\theta_j)\right)/u(\theta_j, Z_{k-1})}$$

satisfies

$$\widetilde{E}_z R^2 \leq \frac{v_0(z,0)}{\rho} = \frac{\left(\sum_{j=1}^m \exp\left(-nI_{A_j}\right) u(\theta_j, z)\right)^2}{\rho}.$$

Theorem 1 combined with the previous analysis then yields the following result.

Corollary 2 *Under the assumptions of Theorem 1, R^2 has polynomial complexity. In particular,*

$$\frac{\widetilde{E}_z R^2}{u_n(z)^2} = O\left(n^d\right)$$

as $n \nearrow \infty$.

Connection to the subsolution approach of Dupuis and Wang: The change of measure given by $\widetilde{P}(\cdot)$ basically corresponds to a change of measure suggested by the subsolution of Dupuis and Wang. Indeed, in order to estimate $u_n(z_0)$, the subsolution approach requires finding a *smooth* function $W(\cdot)$ satisfying

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

and the boundary condition

$$W(x,1) \leq 0 \text{ for } x \in A_1 \cup \ldots \cup A_m.$$

The Hamiltonian $\mathbb{H}$ is defined as

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

where L is Legendre transform of ψ, i.e.,

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

Remark 4 *We note that we use the notation $\langle x, y\rangle$ instead for the inner product $x^T y$ we have been using so far. We do this to help the interested reader familiar with the development of Dupuis and Wang follow the connections with our development here.*

The function L is also the large deviations rate function for the process S_n for further information on this function see [15], and in the setting of Markov modulated queueing networks see [25].

It turns out that if we define for each $1 \leq k \leq m$,

$$W_k(x,t) = -2\langle \theta_k, x\rangle + 2\langle \theta_k, y_k\rangle - 2(1-t)\psi(\theta_k),$$

then the pointwise minimum

$$W(x,t) = \min\{W_k(x,t); 1 \leq k \leq m\}$$

is a viscosity (also known as weak sense) subsolution with maximal value $2I$ at the origin. In order to induce smoothness, one of the method suggested in Section 9.3 of [15] involves introducing a so-called *mollification parameter*, $\epsilon_\triangle$, and considering the function

$$W^\epsilon(x,t) = -\epsilon \log\left(\sum_{k=1}^m \exp\left(-\frac{1}{\epsilon} W_k(x,t)\right)\right).$$

The mollified function is smooth but the value of the mollified function at the origin is no longer the maximal value, which implies that the algorithm is no longer asymptotically optimal. Fortunately

$$W^\epsilon(0,0) \geq 2I - \epsilon \log m$$

and thus one can control the size of this violation. In the verification procedure of Dupuis and Wang, the relaxation (which is controlled by $\epsilon_\triangle$) must vanish as $\triangle \to 0$ in order to obtain asymptotic optimality. For this reason, [12] require choosing a sequence $\epsilon_\triangle$ satisfying $\epsilon_\triangle/\triangle \to \infty$ as $\triangle \to 0$. Using the Lyapunov inequality as a direct verification procedure (rather than taking the logarithms as in the verification approach studied by Dupuis and Wang when using subsolutions) allows us to see that, in fact, we can choose $\epsilon_\triangle = \triangle$.

5.2 A Classical Time Homogeneous Example

Now suppose that $\partial\psi(0) = \mu \neq 0$ and let $\lambda_1, ..., \lambda_m$ be linearly independent vectors in $\mathbb{R}^d$ with unit norm and $c_i \in (0,\infty)$ for $0 \leq i \leq m$. Assume also that $\lambda_i^T \mu < 0$ for all $1 \leq i \leq m$. Let $b > 0$, define $T_b^{(i)} = \inf\{n : \lambda_i^T S_n \geq c_i b\}$ and put $T_b = \min_{1\leq i\leq m} T_b^{(i)}$.

Consider the problem of estimating

$$u_b(z) = P_z(T_b < \infty)$$

efficiently as $b \nearrow \infty$. The previous formulation is a natural generalization of the problem of estimating the tail of the maximum of a Markov random walk with negative drift (whose

i.i.d. counterpart was first studied by Siegmund in his classical 1976 paper, [26]). In the setting of stochastic networks this problem was recently studied in [25].

Define $X_k^{(i)} = \lambda_i^T X_k / c_i$ for all $k \geq 1$ and $1 \leq i \leq m$ and set $\psi_i(\eta) = \psi\left(\eta \lambda_i^T / c_i\right)$. Using this notation we are ready to state the following Theorem which follows easily from results in [1].

Theorem 3 *In addition to steepness of $\psi(\cdot)$ assume that there exists z such that $H(\xi_j, z)$ is a continuous random variable. Then, there exist constants $c_-(z), c_+(z) \in (0, \infty)$ such that*

$$c_-(z) \sum_{j=1}^{m} \exp(-\eta_i b) \leq u_b(z) \leq c_+(z) \sum_{j=1}^{m} \exp(-\eta_i b).$$

as $b \nearrow \infty$, where η_i is the only positive solution to the root $\psi_i(\eta_i) = 0$ for $1 \leq i \leq m$.

The description of the importance sampling strategy is similar to that given in the previous subsection. We first define $\theta_i = \eta_i \lambda_i^T / c_i$ for $1 \leq i \leq m$. Then, given $Z_{k-1} = z_k$, $S_{k-1} = s_{k-1}$ and that $T_b \geq k$ sample the pair (Z_k, X_k) according to the transition law

$$\widetilde{P}(z_{k-1}, s_{k-1}, z_k, x + dx) = \sum_{j=1}^{m} r^{(j)}(z_{k-1}, s_{k-1}) M_{\theta_j}(z_{k-1}, z_k, x + dx), \tag{14}$$

where

$$\begin{aligned} &M_{\theta_j}(z_{k-1}, z_k, x + dx) \\ &= \frac{u(\theta_j, z_k)}{u(\theta_j, z_{k-1})} \exp(\theta_j x) P(X_k \in x + dx | Z_k = z_k) p(z_{k-1}, z_k). \end{aligned}$$

(Since θ_i is now a column vector the notation $\theta_j x$ is interpreted as inner product.) Once again, as in the previous subsection, we set

$$r^{(j)}(z, s) = \frac{w^{(j)}(z, s)}{\sum_{j=1}^{m} w^{(j)}(z, s)},$$

for appropriate functions $w^{(j)}(\cdot)$. According to Lemma 1 (letting $W_j = (Z_j, S_j)$, $T_A = T_b$ and $B = \{\infty\}$), we obtain that we must find a non-negative function $v(z, s)$ such that

$$E_z \left(\frac{v(Z_1, s + X_1)}{v(z, s)} \frac{1}{\sum_{j=1}^{m} r^{(j)}(z, s) u(\theta_j, Z_1) \exp\left(\theta_j^T X_1\right) / u(\theta_j, z)} \right) \leq 1 \tag{15}$$

subject to the boundary condition $v(z, s) \geq \rho > 0$ if $\lambda_i^T s \geq c_i b$ for some $i \in \{1, ...m\}$ and each $z \in \mathcal{S}$. We postulate

$$v(z, s) = \left(\sum_{j=1}^{m} w^{(j)}(z, s) \right)^2.$$

Substituting $v(z,s)$ and the $r^{(j)}$ in terms of the $w^{(j)}$'s we obtain that

$$E_z\left(\frac{v(Z_1,s+X_1)}{v(z,s)}\frac{1}{\sum_{j=1}^m r^{(j)}(z,s)\,u(\theta_j,Z_1)\exp\left(\theta_j^TX_1\right)/u(\theta_j,z)}\right)$$
$$=E_z\left(\frac{\sum_{j=1}^m w^{(j)}(Z_1,s+X_1)}{\sum_{j=1}^m w^{(j)}(z,s)}\frac{\sum_{j=1}^m w^{(j)}(Z_1,s+X_1)}{\sum_{j=1}^m w^{(j)}(z,s)\,u(\theta_j,Z_1)\exp\left(\theta_j^TX_1\right)/u(\theta_j,z)}\right).$$

Selecting

$$w^{(j)}(z,s)=\exp\left(\theta_j^Ts+a_j(b)\right)u(\theta_j,z)$$

for any $a_j(b)\in\mathbb{R}$ yields

$$\frac{\sum_{j=1}^m w^{(j)}(Z_1,s+X_1)}{\sum_{j=1}^m w^{(j)}(z,s)\,u(\theta_j,Z_1)\exp\left(\theta_j^TX_1\right)/u(\theta_j,z)}=1$$

and therefore,

$$E_z\left(\frac{\sum_{j=1}^m w^{(j)}(Z_1,s+X_1)}{\sum_{j=1}^m w^{(j)}(z,s)}\right)$$
$$=\sum_{j=1}^m E_z\left(\frac{w^{(j)}(z,s)\exp\left(\theta_j^TX_1\right)u(\theta_j,Z_1)}{\sum_{j=1}^m w^{(j)}(z,s)\,u(\theta_j,z)}\right)$$
$$=\sum_{j=1}^m r^{(j)}(z,s)\,E_z\left(\frac{\exp\left(\theta_j^TX_1\right)u(\theta_j,Z_1)}{u(\theta_j,z)}\right)=1,$$

which implies that the Lyapunov inequality (15) is satisfied assuming that the $a_j(b)$'s are selected so that the boundary condition holds and the value of $h(z,0)$ is minimized (recall that we wish to obtain an upper bound).

In order to select the $a_j(b)$'s note that $w_j(z,0)=\exp(a_j(b))\,u(\theta_j,z)$, a natural candidate consists in selecting $a_j(b)=-\eta_jb$. In turn, note that

$$\theta_j^Ts-\eta_jb=\eta_j\left(\lambda_j^Ts-c_jb\right)/c_j.$$

Therefore, we conclude that if $\lambda_j^Ts\geq c_jb$ for some $1\leq j\leq m$, then we must have that

$$h(z,s)\geq\min_{1\leq j\leq m,z\in\mathcal{S}}u(\theta_j,z)^2=\rho$$

and therefore defining

$$R=\prod_{k=1}^T\frac{1}{\sum_{j=1}^m r^{(j)}(Z_{k-1},S_{k-1})\,u(\theta_j,Z_k)\exp\left(\theta_j^TX_k\right)/u(\theta_j,Z_{k-1})},$$

we obtain $\widetilde{E}_zR^2\leq\left(\sum_{j=1}^m\exp(-\eta_jb)\,u(\theta_j,z)\right)^2/\rho$, which implies, together with Theorem 3 the following result.

Corollary 4 *Under the assumptions of Theorem 3 the estimator R is strongly efficient in the sense that*

$$\frac{\widetilde{E}_z R^2}{u_b(z)} = O(1)$$

as $b \nearrow \infty$.

Connection to the subsolution approach of Dupuis and Wang: Similar to the finite horizon problem discussed in the previous section, we use a change of measure suggested by the subsolution approach. Indeed, in this setting,

$$W(x) = \min\{-2\langle\theta_k, x\rangle + 2\eta_k; 1 \le k \le m\}$$

is a weak sense subsolution. As in the finite horizon problem, a mollification parameter, $\epsilon_\triangle$, is introduced to induce smoothness, obtaining

$$W^\epsilon(x) = -\epsilon \log\left(\sum_{k=1}^m \exp\left[-\frac{1}{\epsilon}(-2\langle\theta_k, x\rangle + 2\eta_k)\right]\right).$$

The use of the subsolution directly in order to verify the asymptotic optimality of the sampler requires choosing $\epsilon_\triangle/\triangle \to \infty$ as $\triangle \to 0$. However, as in the finite horizon problem, applying the Lyapunov inequality directly allows us to gain insight into the selection of the mollification parameter. In particular, note that the selection $\epsilon_\triangle = \Delta$, which corresponds to the sampler of Corollary 4 is optimal because it allows us to conclude strong efficiency.

6 Tandem Networks

We consider a tandem network with d stations, indexed from 1 to d in such a way that the the i-th station's output goes to the $(i+1)$-th station. The arrival rate of the process is $\mu_0 > 0$ and the service rate at station i is $\mu_i > 0$. The process is assumed to be stable so $\rho_i = \mu_0/\mu_i \in (0,1)$. Without loss of generality we assume that $\mu_0 + \mu_1 + ... + \mu_d = 1$.

Put $\mu_* = \min_{1\le i\le d}\mu_i$, $\rho_* = \mu_0/\mu_*$, $\gamma = -\log\rho_*$ and let β be the number of bottleneck stations, that is, $\beta = |\{1 \le i \le d : \rho_i = \rho_*\}|$.

We are interested in estimating overflow probabilities at level $1/\Delta$ when $\Delta \approx 0$ for the whole population of the network within a busy period. In order to describe the overflow probability of interest in a suitable scale depending on $\Delta > 0$ let us introduce an appropriate family of processes corresponding to the embedded discrete time Markov chain appropriately scaled by the factor Δ.

Following the notation introduced in Section 4, let $Y^{(\Delta)} = (Y_t : t \in \{0, \Delta, 2\Delta, ...\})$, indexed by Δ, defined via

$$Y_{t+\Delta} = Y_t + \Delta V_{t+\Delta}(Y_t),$$

where the $V_t(y)$'s are independent random variables whose distribution is described as follows. First consider the power set $\mathcal{A} = 2^{\{1,2,...,d\}}$ (i.e. the family of subsets of $\{1,2,...,d\}$). Given $A \in \mathcal{A}$ we define

$$\mathcal{S}_A = \{y \in \Delta\mathbb{Z}_+^d : y_i = 0 \text{ for } i \in A \text{ and } y_j > 0 \text{ for } j \notin A\}.$$

In other words, $\mathcal{S}_A$ contains states describing situations where only the queues whose index belong to A are empty. The set $\mathcal{S}_\varnothing$ is called the interior of the space and each of the remaining $\mathcal{S}_A$'s describe the boundaries. Now, define, for $1 \leq i \leq d-1$

$$\begin{aligned}
\mathbf{e}_i &= (0,...,0,\underbrace{-1}_{i\text{-th element}},\underbrace{1}_{(i+1)\text{-th element}},0,...,0)^T,\\
\mathbf{e}_0 &= (\underbrace{1}_{\text{1st element}},...,0,0,0,0,...,0)^T,\\
\mathbf{e}_d &= (0,...,0,0,0,0,...,\underbrace{-1}_{d\text{-th element}})^T.
\end{aligned}$$

Then, we have that if $y \in \mathcal{S}_\varnothing$, then

$$P(V(y) = \mathbf{e}_i) = \mu_i \quad 0 \leq i \leq d.$$

More generally, if $y \in \mathcal{S}_A$ then,

$$P(V(y) = 0) = \sum_{i \in A} \mu_i \text{ if } i \in A \text{ and } P(V(y) = \mathbf{e}_i) = \mu_i \text{ if } i \in A^c,$$

where, by convention, the operation A^c is taken relative to the set $\{0,1,...,d\}$. The log-moment generating function at state y is defined as

$$\psi(\theta, y) = \log E \exp\left(\theta^T V(y)\right).$$

Lastly define the d dimensional vector of all ones as $\mathbf{1}$.

Note that the process $\Delta^{-1}Y_{k\Delta}$ is the state of the embedded discrete time Markov chain associated to the underlying tandem network evaluated at time k. Define $T_0 = \inf\{k \geq 1 : Y_{k\Delta} = 0\}$ and $T_1 = \inf\{k \geq 1 : Y_{k\Delta}^T \mathbf{1} \geq 1\}$. We are interested in analyzing a class efficient state-dependent importance sampling estimators for $u(\Delta) = P_0(T_0 < T_1)$ as $\Delta \searrow \infty$. The class of samplers that we consider are those proposed by [12]. Our objective is to use Lyapunov inequalities in order to find further insights into the various mollification parameters introduced by Dupuis et. al. and also obtain more information on the behavior of the second moment of the proposed estimator, beyond the asymptotic behavior of its (optimal) exponential rate of convergence. The main result of this section is the following.

Theorem 5 *There exists a suitable selection of mollification parameters in the subsolution-based change-of-measure of [12] so that the corresponding estimator, $R(y)$, satisfies*

$$\frac{\widetilde{E}\left(R(y)^2\right)}{u(\Delta)^2} = O\left(\Delta^{-2(d+1-\beta)}\right) \tag{16}$$

as $\Delta \searrow 0$.

The rest of the section is dedicated to the analysis of Theorem 5, however, before continuing with such analysis, let us point out the implications behind (16) in terms of the computational cost associated to the importance sampling estimator underlying Theorem 5. As we indicated in Section 2, Chebyshev's inequality implies that $O\left(\Delta^{-2(d+1-\beta)}\right)$ replications are required to obtain an estimator with a prescribed relative accuracy. On the other hand solving the associated linear system of equations to compute $P_y\left(T_1 < T_0\right)$, as a function of $y \in \Delta\mathbb{Z}_+^d$, using say Gaussian elimination will typically take at least $c_0\Delta^{-2d}$ operations for some $c_0 > 0$. As a consequence, Theorem 5 guarantees that the corresponding importance sampling scheme has no worse complexity rate (as a function of Δ) than solving the associated linear system of equations by Gaussian elimination.

We now are ready to move into the development behind Theorem 5. We start our discussion by stating the following simple lower bound on $u\left(\Delta\right)$. The proof is given in [5].

Proposition 1 *There exists a constant $c_- \in (0,\infty)$ such that*

$$u\left(\Delta\right) \geq c_- \exp\left(-\gamma/\boldsymbol{\Delta}\right)/\Delta^{\beta-1}.$$

We now proceed to the construction of associated Lyapunov inequalities. In order to describe the issues that arise when constructing appropriate Lyapunov functions let us start our discussion with the case of a two node tandem network studied by Parekh-Walrand,[23].

6.1 A Two Node Tandem Network

Let us first consider the case $d = 2$ and assume that $\mu_0 < \mu_2 \leq \mu_1$. As in Section 5 it is natural to consider mixtures of exponential changes-of-measure. In particular,

$$\frac{\widetilde{P}\left(V\left(y\right) \in dz\right)}{P\left(V\left(y\right) \in dz\right)} = L\left(y\right) = \sum_{j=0}^{k} r_j\left(y\right) \exp\left(\theta_j^T z - \psi\left(\theta_j, y\right)\right), \tag{17}$$

for an appropriate value $k > 0$, some θ_j's and given functions $r_j\left(\cdot\right)$'s. Once again, motivate by the analysis in Section 5 we propose setting

$$r_j\left(y\right) = \frac{w_j\left(y\right)}{\sum_{j=0}^{k} w_j\left(y\right)}$$

with

$$w_j\left(y\right) = \exp\left(\theta_j^T y/\Delta + a_j\left(\Delta\right)\right).$$

We must find an appropriate Lyapunov function, which, once again following the arguments in Section 5, we propose

$$v\left(y\right) = \left(\sum_{j=0}^{k} w_j\left(y\right)\right)^2.$$

Verifying the Lyapunov inequality would entail showing

$$E\frac{v\left(y+\Delta V\left(y\right)\right)}{v\left(y\right)}\frac{1}{\sum_{j=0}^{k}r_{j}\left(y\right)\exp\left(\theta_{j}^{T}V\left(y\right)-\psi\left(\theta_{j},y\right)\right)}$$
$$=E\frac{\sum_{j=0}^{k}w_{j}\left(y\right)\exp\left(\theta_{j}^{T}V\left(y\right)\right)}{\sum_{j=0}^{k}w_{j}\left(y\right)}\frac{\sum_{j=0}^{k}w_{j}\left(y\right)\exp\left(\theta_{j}^{T}V\left(y\right)\right)}{\sum_{j=0}^{k}w_{j}\left(y\right)\exp\left(\theta_{j}^{T}V\left(y\right)-\psi\left(\theta_{j},y\right)\right)}\leq1. \tag{18}$$

An interesting twist arises in the current setting, which did not occur in Section 5. The issue is that we may not be able to guarantee that $\psi\left(\theta_{j},y\right)\leq0$ for all y, j and therefore (18) does not follow immediately as we saw in Section 5. However, an observation that follows by the analysis in [12] is that one can select the constants $a_{j}\left(\Delta\right)$'s so that

$$E\frac{\left(\sum_{j=0}^{k}w_{j}\left(y\right)\exp\left(\theta_{j}^{T}V\left(y\right)\right)\right)^{2}}{\sum_{j=0}^{k}w_{j}(y)\sum_{j=0}^{k}w_{j}(y)\exp\left(\theta_{j}^{T}V\left(y\right)-\psi\left(\theta_{j},y\right)\right)}=1+\chi\left(\Delta\right),$$

for some error $\chi\left(\Delta\right)$ that can be appropriately controlled. In particular, taking advantage of the associated subsolution to the Isaacs equation introduced by, [12] one takes $k=2$ and

$$\theta_{0}=\gamma\cdot\left(1,1\right)^{T},\quad\theta_{1}=\gamma\cdot\left(1,0\right)^{T},\quad\theta_{2}=\left(0,0\right)^{T},$$
$$a_{j}\left(\Delta\right)=-\gamma/\Delta+j\delta/\Delta,\quad0\leq j\leq2.$$

The parameter $\delta>0$ will be selected as a function of Δ in order to appropriately control the error $\chi\left(\Delta\right)$. As we shall see, the Lyapunov inequality suggests a suitable selection that minimizes the growth of the upper bound on the second moment of the estimator, namely, $\delta=\Delta\log\left(1/\Delta\right)$ (see also Remark 3.7 in [12])).

The overall estimator then takes the form

$$R\left(y\right)=\prod_{j=0}^{T_{1}-1}L\left(Y_{j\Delta}\right)^{-1}I\left(T_{1}<T_{0}\right). \tag{19}$$

In order to discuss the properties implied by the previous selection of parameters recall

$$\mathcal{S}_{\emptyset}=\{(y_{1},y_{2})\in\Delta\mathbb{Z}_{+}^{2}:y_{1},y_{2}>0\},$$
$$\mathcal{S}_{1}=\{(y_{1},y_{2})\in\Delta\mathbb{Z}_{+}^{2}:y_{1}=0,y_{2}>0\},$$
$$\mathcal{S}_{2}=\{(y_{1},y_{2})\in\Delta\mathbb{Z}_{+}^{2}:y_{1}>0,y_{2}=0\}.$$

It is easy to verify that:

1. If $y\in\mathcal{S}_{\emptyset}$ then $\psi\left(\theta_{j},y\right)\leq0$ for $j=0,1,2$.
2. If $y\in\mathcal{S}_{1}$ then $\psi\left(\theta_{j},y\right)=0$ for $j=0,2$ but $\psi\left(\theta_{1},y\right)$ can be positive.
3. If $y\in\mathcal{S}_{2}$ then $\psi\left(\theta_{j},y\right)\leq0$ for $j=1,2$ but $\psi\left(\theta_{0},y\right)$ can be positive.

Given the previous observations, it follows immediately that (18) holds in the interior, namely in the set $\mathcal{S}_{\emptyset}$. The cases $y\in\mathcal{S}_{1}$ and $y\in\mathcal{S}_{2}$ are similar, so let us just explain what

occurs if $y \in \mathcal{S}_1$. Define

$$G(y) = \frac{\left(\sum_{j=0}^{2} w_j(y + \Delta V(y))\right)^2}{\left(\sum_{j=0}^{2} w_j(y)\right)^2 \sum_{j=0}^{2} r_j(y + \Delta V(y)) \exp(-\psi(\theta_j, y))}$$
$$= \frac{\left(\sum_{j=0}^{2} r_j(y) \exp\left(\theta_j^T V(y)\right)\right)^2}{\sum_{j=0}^{2} r_j(y) \exp\left(\theta_j^T V(y) - \psi(\theta_j, y)\right)}.$$

We note that on the set $\mathcal{S}_1$ we have that (18) can be violated simply because of $w_1(y)$, however, we note that on $y \in \mathcal{S}_1$, $w_1(y)/(w_0(y) + w_2(y)) \leq \exp(-\delta/\Delta)$. Therefore if $y \in \mathcal{S}_1$,

$$G(y) = \frac{\left(\sum_{j=0}^{2} r_j(y) \exp\left(\theta_j^T V(y)\right)\right)^2}{\sum_{j=0}^{2} r_j(y) \exp\left(\theta_j^T V(y) - \psi(\theta_j, y)\right)}$$
$$\leq \frac{\left(\sum_{j=0}^{2} r_j(y) \exp\left(\theta_j^T V(y)\right)\right)^2}{\sum_{j=0,2} r_j(y) \exp\left(\theta_j^T y - \psi(\theta_j, y)\right)}$$
$$\leq \frac{\left(\sum_{j=0,2} r_j(y) \exp\left(\theta_j^T V(y)\right) + r_1(y) e^{\theta_j^T V(y)}\right)^2}{\sum_{j=0,2} r_j(y) e^{\theta_j^T V(y)}}$$
$$\leq \sum_{j=0,2} r_j(y) e^{\theta_j^T V(y)} + 2me^{-\delta/\triangle},$$

for some constant $m \in (0, \infty)$ which can be suitably chosen. Note that for $y \in \mathcal{S}_1$,

$$E \sum_{j=0,2} r_j(y) e^{\theta_j^T V(y)} \leq 1.$$

A similar argument follows for the case $y \in \mathcal{S}_2$. A key observation in such case is that $w_0(y)/(w_1(y) + w_2(y)) \leq \exp(-\delta/\Delta)$; the argument then follows just as before.

In summary, defining $\chi(\Delta) = \exp(-\delta/\Delta)$, we obtain that

$$EG(y) = E\frac{v(y + \Delta V(y))}{v(y)} \frac{1}{\sum_{j=0}^{2} r_j(y) \exp\left(\theta_j^T V(y) - \psi(\theta_j, y)\right)}$$
$$\leq 1 + m\chi(\Delta), \tag{20}$$

evidently, we also have that $v(y) \geq 1$ whenever $y^T \mathbf{1} \geq 1$. Lemma 1 then implies, choosing δ so that $m\chi(\Delta) < 1/2$,

$$v(y) \geq E \exp(-m\chi(\Delta) T_1) R(y), \tag{21}$$

We still need to transform (21) into an inequality for the second moment of the estimator (i.e. $ER(y) = \widetilde{E}R(y)^2$). We will comeback to this issue later, after developing the corresponding

Lyapunov inequality for a d-node tandem network. The idea follows the same spirit as in the second part of Lemma 1 by arguing that $T_1 \leq \lambda/\Delta$ with very high probability and choosing $\chi(\Delta) = \Delta$ (i.e. $\delta = \Delta \log(1/\Delta)$).

Remark 5 *We note that we have deliberately avoided referring to δ as a mollification parameter. This is to be consistent with the terminology adopted by Dupuis and Wang. Typically, the term mollification is used to introduce smoothness. Strictly speaking, the role of δ in the subsolution approach is to satisfy the subsolution equation on the boundaries, but this is still in a weak sense (see for instance [12]). In the context of Lyapunov inequalities this means this corresponds to satisfying the Lyapunov bound at the boundaries (up to the error term as indicated above $\chi(\Delta)$).*

6.2 Relaxed Lyapunov Inequality for a Markovian Tandem Network

Once again, we depart from the subsolution to the Isaacs equation proposed by [12]. In particular, we define the i-th component of the vector θ_j (for $0 \leq j \leq k = d$ and $1 \leq i \leq d$)

$$(\theta_j)(i) = \begin{cases} \gamma, & 1 \leq i \leq d-j \\ 0, & otherwise \end{cases}$$

and put

$$w_j(y) = \exp\left(\theta_j^T y/\Delta - \gamma/\Delta + j\delta/\Delta\right).$$

The rest of the description of the sampler proceeds just as in the previous section, namely, we set $r_j(y) = w_j(y) / \sum_{j=0}^{d} w_j(y)$ and $\widetilde{P}(V(y) \in dv)$ as indicated in (17). We wish to test a Lyapunov inequality such as (20), therefore we define the Lyapunov function

$$v(y) = \left(\sum_{j=0}^{d} w_j(y)\right)^2. \tag{22}$$

Just as it occurred in the two dimensional case, introducing $\delta > 0$ allows us to handle states in which at least one queue might be empty (i.e. potential cases in which $\psi(\theta_j, y) > 0$ might occur). In order to see how one handles these states define, given $y \in \Delta\mathbb{Z}_+^d$,

$$B(y) = \{0 \leq j \leq d-1 : y_{d-j} = 0\}.$$

This collection of indices will represent the roots θ_j that should be avoided when the scaled queue length is at y, i.e. $\psi(\theta_j, y) > 0$. In particular we have the following lemma (the results of this lemma are very similar to those found in Lemma 4.4 of [12], however we include the details of the proof here to ease confusion).

Lemma 2 *If $j \notin B(y)$ then $\psi(\theta_j, y) \leq 0$. If $j \in B(y)$ then $\psi(\theta_j, y) > 0$ and*

$$r_j(y) \leq \frac{w_j(y)}{\sum_{k \notin B(y)} w_k(y)} \leq \exp(-\delta/\Delta). \tag{23}$$

Proof. First note that we only need to consider θ_j for $0 \leq j \leq d-1$, since by definition $d \notin B(y)$ and it is trivially seen that $\psi(\theta_d, y) = 0$ for all y. Thus consider $0 \leq j \leq d-1$, θ_j, and $1 \leq i \leq d$ and observe

$$\theta_j^T \mathbf{e}_i = \begin{cases} 0, & d-j \neq i \\ -\gamma, & d-j = i. \end{cases} \tag{24}$$

Next for $A \in \mathcal{A}$ and $y \in \mathcal{S}_A$ we have

$$\psi(\theta_j, y) = \log\left(\mu_0 e^{\theta_j^T \mathbf{e}_0} + \sum_{i \notin A} \mu_i e^{\theta_j^T \mathbf{e}_i} + \sum_{i \in A} \mu_i\right).$$

If $j \notin B(y)$ then by definition $d-j \notin A$ and therefore the previous display and (24) gives

$$\psi(\theta_j, y) = \log\left(\mu_* + \sum_{i \neq d-j} \mu_i + \mu_{d-j}\frac{\mu_0}{\mu_*}\right) \leq 0.$$

The details for the above inequality can be found in [12].

If $j \in B(y)$ then we see that

$$\psi(\theta_j, y) = \log\left(\mu_* + \sum_{i=1}^{d} \mu_i\right) > 0.$$

Thus it remains to prove (23).

For $j \in B(y)$ define

$$j' = \min\{k > j : y_{d-k} \neq 0\} \wedge d,$$

and note that by definition $j' \notin B(y)$. In addition we have the following

$$\begin{aligned} \Delta \log w_j(y) &= \theta_j^T y - \gamma + j\delta = \gamma \sum_{k=1}^{d-j} y_k - \gamma + j\delta = \gamma \sum_{k=1}^{d-j'} y_k - \gamma + j\delta \\ &= \Delta \log w_{j'}(y) - (j'-j)\delta. \end{aligned}$$

Using the previous display, (23) follows immediately. ■

We now are ready to prove the following Lyapunov inequality. Indeed, set

$$\begin{aligned} G(y) &= \frac{v(y + \Delta V(y))}{v(y)} \frac{1}{\sum_{j=0}^{d} r_j(y) \exp\left(\theta_j^T V(y) - \psi(\theta_j, y)\right)} \\ &= \frac{\left(\sum_{j=0}^{d} r_j(y) \exp\left(\theta_j^T V(y)\right)\right)^2}{\sum_{j=0}^{d} r_j(y) \exp\left(\theta_j^T V(y) - \psi(\theta_j, y)\right)}. \end{aligned}$$

It follows from Lemma 2 and simple algebraic manipulations entirely analogous to those performed in the case $d = 2$ that

$$\begin{aligned} G(y) &\leq \frac{\left(\sum_{j=0}^{d} r_j(y) \exp\left(\theta_j^T V(y)\right)\right)^2}{\sum_{j \notin B(y)} r_j(y) \exp\left(\theta_j^T V(y)\right)} \\ &\leq \frac{\left(\sum_{j \notin B(y)} r_j(y) \exp\left(\theta_j^T V(y)\right) + m_0 \exp\left(-\delta/\Delta\right)\right)^2}{\sum_{j \notin B(y)} r_j(y) \exp\left(\theta_j^T V(y)\right)} \\ &\leq \sum_{j \notin B(y)} r_j(y) \exp\left(\theta_j^T V(y)\right) + m_1 \exp\left(-\delta/\Delta\right). \end{aligned} \tag{25}$$

We then conclude, combining the previous analysis together with Lemma 2 and Lemma 1 to get the following result

Proposition 2 *If $\delta := \delta_\Delta = \Delta \log(1/\Delta)$ then, there exists $m_1 \in (0, \infty)$ such that*

$$E \frac{v(y + \Delta V(y))}{v(y)} \frac{1}{\sum_{j=0}^{d} r_j(y) \exp\left(\theta_j^T V(y) - \psi(\theta_j, y)\right)} \leq 1 + m_1 \Delta,$$

for $y \neq 0$ and $y^T \mathbf{1} \leq 1$. Since $v(y) \geq 1$ if $y^T \mathbf{1} \geq 1$ we obtain

$$v(y) \geq E \exp\left(-m_1 \Delta T_1\right) R(y),$$

where

$$R(y) = \prod_{j=0}^{T_1 - 1} L\left(Y_{\Delta j}\right)^{-1} I\left(T_1 < T_0\right).$$

The final ingredient in the proof of Theorem 5 consists in transforming the bound for $E \exp\left(-m_1 \Delta T_1\right) R(0)$ given in Proposition 2 into a bound for $ER(0)$.

First fix $\lambda \in (0, \infty)$ and note that

$$\begin{aligned} e^{-m_1 \lambda} E\left[R(0)\right] &= e^{-m_1 \lambda} \left(E\left[R(0); T_1 > \lambda/\Delta, T_1 < T_0\right] + E\left[R(0); T_1 \leq \lambda/\Delta, T_1 < T_0\right]\right) \\ &\leq e^{-m_1 \lambda} E\left[R(0); T_1 > \lambda/\Delta, T_1 < T_0\right] + E\left[e^{-m \Delta T_1} R(0); T_1 \leq \lambda/\Delta, T_1 < T_0\right]. \end{aligned}$$

We will argue that for every $\kappa \in (0, \infty)$ there exists $\lambda \in (0, \infty)$ such that

$$E\left[R(0); T_1 > \lambda/\Delta, T_1 < T_0\right] = o\left(\exp\left(-\kappa/\Delta\right)\right). \tag{26}$$

Then using the non-negativity of $R(0)$, the result of Proposition 2, and the previous display with $\kappa > 2\gamma$ we see that

$$\begin{aligned} E\left[R(0)\right] &\leq E\left[R(0); T_1 > \lambda/\Delta, T_1 < T_0\right] + e^{m_1 \lambda} E\left[e^{-m_1 \Delta T_1} R(0)\right] \\ &\leq o\left(\exp\left(-\kappa/\Delta\right)\right) + e^{m_1 \lambda} v(0) = O\left(\Delta^{-2d} \exp\left(-2\gamma/\Delta\right)\right). \end{aligned}$$

Therefore it remains to prove the result in (26). First an application of the Cauchy-Schwarz inequality gives

$$E\left[R(0); T_1 > \lambda/\Delta, T_1 < T_0\right] \leq \left(E\left[R(0)^2\right]\right)^{1/2} \left(P\left(T_1 > \lambda/\Delta, T_1 < T_0\right)\right)^{1/2}.$$

From Lemma B.1 of [11] it follows that there exists $c > 0$ such that

$$E\left[e^{cT_0}\right] < \infty. \tag{27}$$

Therefore we have that given $\kappa \in (0, \infty)$ there exists $\lambda > 0$ such that

$$P(T_0 > \lambda/\Delta) = o\left(\exp\left(-\kappa/\Delta\right)\right).$$

Thus display (26) follows if we can show that $E\left[R(0)^2\right]$ stays bounded as $\Delta \to 0$. Recalling the definition of $R(0)$ from equation (19) we have

$$E\left[R(0)^2\right] = E\left[I(T_1 < T_0)\left(\prod_{j=0}^{T_1-1} L(Y_{j\Delta})^{-1}\right)^2\right].$$

Recalling the Lyapunov function $v(y)$ from display (22) and using similar arguments as in display (25) observe the following inequality

$$\begin{aligned}
&\frac{v\left(y + \triangle V(y)\right)}{v(y)} \frac{1}{\left(\sum_{j=0}^{d} r_j(y) \exp\left[\theta_j^T V(y) - \psi(\theta_j, y)\right]\right)^2} \\
&= \frac{\left(\sum_{j=0}^{d} r_j(y) e^{\theta_j^T V(y)}\right)^2}{\left(\sum_{j=0}^{d} r_j(y) \exp\left[\theta_j^T V(y) - \psi(\theta_j, y)\right]\right)^2} \leq 1 + K e^{-\delta/\triangle},
\end{aligned}$$

where K is a constant depending only on the system parameters. Importantly note that this inequality holds pathwise, i.e., it is does not involve expected values. Therefore if we take $\delta = -\triangle \log \triangle$ we have the following

$$\begin{aligned}
E[R(0)^2] &= E\left[I(T_1 < T_0) \frac{v(Y_0)}{v(Y_{T_1})} \prod_{k=0}^{T_1-1} \frac{v(Y_{(k+1)\triangle})}{v(Y_{k\triangle})} \left(\sum_{j=0}^{d} r_j(Y_{k\triangle}) \exp\left[\theta_j^T V(Y_{k\triangle}) - \psi(\theta_j, Y_{k\triangle})\right]\right)^{-2}\right] \\
&\leq E[I(T_1 < T_0) \frac{v(Y_0)}{v(Y_{T_1})} (1 + K\triangle)^{T_1}] \leq 2E\left[I(T_1 < T_0) \exp(T_1 \log(1 + K\triangle))\right],
\end{aligned}$$

where the last inequality follows from the definition of the function $v(\cdot)$ and the stopping time T_1. The result then follows from equation (27). This completes the proof of Theorem 5.

6.3 Connection to the Subsolution Approach: Analysis with General Mollification Parameters

A noted difference between the importance sampling algorithm discussed in this work and the one considered in [12], is the latter algorithm used a general sequence of mollification parameters while the present algorithm simply uses Δ. In this section we would like to carry out our analysis under the condition of varying mollification parameters, ε.

In particular we consider a variant of the change of measure described by (17),

$$\frac{\widetilde{P}\left(V\left(y\right)\in dz\right)}{P\left(V\left(y\right)\in dz\right)}=L\left(y\right)=\sum_{j=0}^{k}r_{j}^{\varepsilon}\left(y\right)\exp\left(\theta_{j}z-\psi\left(\theta_{j},y\right)\right) \tag{28}$$

where

$$r_{j}^{\varepsilon}(y)=\frac{w_{j}^{\varepsilon}(y)}{\sum_{k=0}^{d}w_{j}^{\varepsilon}(y)}, \tag{29}$$

and

$$w_{j}^{\varepsilon}(y)=\exp\left[\left(\theta_{j}^{T}y-\gamma+j\delta\right)/\varepsilon\right]. \tag{30}$$

We can assume that ε and δ vary with Δ, but will suppress this notation. The Lyapunov function that we will use in this situation is

$$v^{\varepsilon}(y)=\left(\sum_{j=0}^{d}w_{j}^{\varepsilon}(y)\right)^{2\varepsilon/\Delta}.$$

Therefore in order to apply Lemma 1 we need to establish the following inequality

$$\begin{aligned}G(y)&=E\left\{\frac{v^{\varepsilon}(y+\Delta V(y))}{v^{\varepsilon}(y)}\frac{1}{\sum_{j=0}^{d}r_{j}^{\varepsilon}(y)\exp\left(\theta_{j}^{T}y-\psi(\theta_{j},y)\right)}\right\}\\&=E\left\{\frac{\left(\sum_{j=0}^{d}r_{j}^{\varepsilon}(y)\exp\left(\frac{\Delta}{\varepsilon}\theta_{j}^{T}y\right)\right)^{2\varepsilon/\Delta}}{\sum_{j=0}^{d}r_{j}^{\varepsilon}(y)\exp\left(\theta_{j}^{T}y-\psi(\theta_{j},y)\right)}\right\}\\&\leq 1.\end{aligned}$$

Under the assumption that $\varepsilon\geq\Delta$ for all $\Delta>0$ we can apply Jensen's inequality to the expression for $G(y)$ and get

$$G(y)\leq E\left\{\frac{\left(\sum_{j=0}^{d}r_{j}^{\varepsilon}(y)\exp\left(\theta_{j}^{T}y\right)\right)^{2}}{\sum_{j=0}^{d}r_{j}^{\varepsilon}(y)\exp\left(\theta_{j}^{T}y-\psi(\theta_{j},y)\right)}\right\}. \tag{31}$$

However the term in the previous display can be analyzed using the same techniques in the proof of Lemma 2, and we arrive at

$$G(y)\leq 1+e^{-\delta/\varepsilon}.$$

Therefore if we can ensure that $e^{-\delta/\varepsilon}$ goes to zero sufficiently fast with Δ we can say that the performance of the algorithm is determined by

$$v^{\varepsilon}(0) \approx e^{-2\gamma/\triangle} \left(\frac{1}{\triangle}\right)^{2\varepsilon(d-1)/\triangle}.$$

Based on this analysis it appears that little can be gained from introducing a mollification parameter ε that varies with Δ.

Remark 6 *There is a slight difference between the algorithm described in this section and the algorithm described in [12]. In particular both algorithms use a mixture of exponentials twists according to the vectors θ_j, but in Dupuis et. al. the weights are defined as*

$$\tilde{w}_j^{\varepsilon} = \exp\left[\left(2\theta_j^T y - 2\gamma + j\delta\right)/\varepsilon\right].$$

However the two algorithms can in fact be mapped to each other in a simple fashion. In particular for any δ and ε one uses in the algorithm of [12], one would get an identical algorithm by considering the algorithm in equations (28) to (30) with $\bar{\delta} = 2\delta$ and $\bar{\varepsilon} = 2\varepsilon$.

In [12] the parameter selection is made according to the rules

1. $\delta \to 0$ as $\Delta \to 0$,
2. $\varepsilon/\Delta \to \infty$ as $\Delta \to 0$,
3. $\delta/\varepsilon \to \infty$ as $\Delta \to 0$.

With these choices made it is possible to show that the importance sampling algorithm is asymptotically efficiently, i.e. the exponential decay rate of the second moment matches the exponential decay rate of the probability squared. Although the goal in [12] was just to provide an asymptotic analysis to show weak efficiency, one can extract explicit bounds from the analysis in [12] which in turn allow to obtain further information on the decay rate of the second moment of the importance sampling estimator. In the following we pursue this goal.

In particular if we look at the third display of page 39 from [12] we see a bound that is equivalent to the following

$$E\left[R(0)\right] \leq \exp\left[-\frac{1+\alpha(\Delta)}{\Delta\left(1+2\alpha(\Delta)\right)}W(0)\right] E\exp\left(\beta(\Delta)\frac{1+\alpha(\delta)}{2\alpha(\Delta)}T_0\right),$$

where

$$W(0) = -\varepsilon \log \sum_{k=0}^{d} \exp\left(\frac{k\delta}{\varepsilon} - 2\gamma/\varepsilon\right),$$

and

$$\beta(\Delta) \geq \frac{C\Delta}{\varepsilon} \text{ and } \alpha(\Delta) \geq K\beta(\Delta)$$

for a positive constant K.

In addition we have the following bound for positive α

$$\frac{1+\alpha}{1+2\alpha} \leq 1 - \frac{C\Delta}{\varepsilon}.$$

Putting this together with the assumptions (1)-(3) above we have we have the following bound (for positive constant M)

$$\begin{aligned} E[R(0)] &\leq M \exp\left[-\left(1 - \frac{C\Delta}{\varepsilon}\right)\left(\frac{2\gamma}{\Delta} - \frac{\delta d}{\Delta}\right)\right] \\ &\leq M \exp\left(-2\gamma/\Delta + \frac{\delta d}{\Delta} + \frac{2\gamma C}{\delta}\right). \end{aligned}$$

Minimizing the above exponent over δ, we see that the optimal choice of $\delta = C_0\Delta^{1/2}$ gives the following bound

$$E[R(0)] \leq M \exp\left(-2\gamma/\Delta + C_0\Delta^{-1/2}\right).$$

This is clearly a much looser bound than that obtained in the result of Theorem 5.

References

[1] S. Asmussen. *Applied Probability and Queues*. Springer-Verlag, New York, NY, USA, 2003.

[2] S. Asmussen, K. Binswanger, and B. Hojgaard. Rare events simulation for heavy-tailed distributions. *Bernoulli*, 6:303–322, 2000.

[3] A. Bassmaboo, S. Juneja, and A. Zeevi. On the inefficiency of state-independent importance sampling in the presence of heavy-tails. *Operations Research Letters*, 34:521–531, 2006.

[4] J. Blanchet. Efficient importance sampling for binary contingency tables. *Ann. of Appl. Prob., To appear*, 2008.

[5] J. Blanchet. Optimal sampling of overflow paths in jackson networks. *Preprint*, 2009.

[6] J. Blanchet and P. Glynn. Efficient rare-event simulation for the maximum of a heavy-tailed random walk. *Ann. of Appl. Probab.*, 18:1351–1378, 2008.

[7] J. Blanchet, P. Glynn, and J. C. Liu. Fluid heuristics, lyapunov bounds and efficient importance sampling for a heavy-tailed g/g/1 queue. *QUESTA*, 57:99–113, 2007.

[8] J. Blanchet and J. C. Liu. State-dependent importance sampling for regularly varying random walks. *Journal of Applied Probability. To Appear.*

[9] H. P. Chan and T. L. Lai. Efficient importance sampling for monte carlo evaluation of exceedance probabilities. *Ann. of Appl. Prob.*, 17:440–473, 2007.

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

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

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

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

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

[15] P. Dupuis and H. Wang. Subsolutions of an isaacs equation and efficient schemes of importance sampling. *Mathematics of Operations Research*, 32:723–757, 2007.

[16] P. Dupuis and H. Wang. Importance sampling for jackson networks. *Preprint*, 2008.

[17] M. Freidlin and A. Wentzell. *Random Perturbations of dynamical systems*. Springer-Verlag, New York, 1998.

[18] P. Glasserman and S. Kou. Analysis of an importance sampling estimator for tandem queues. *ACM TOMACS*, 5:22–42, 1995.

[19] P. Glasserman and Y. Wang. Counter examples in importance sampling for large deviations probabilities. *Annals of Applied Probability*, 7:731–746, 1997.

[20] M. Iltis. Sharp asymptotics of large deviations for general state-space Markov-additive chains in $\mathbb{R}^d$. *Statistics and Probability Letters*, 47:365–380, 2000.

[21] S. Juneja and P. Shahabuddin. Rare event simulation techniques: An introduction and recent advances. In S. G. Henderson and B. L. Nelson, editors, *Simulation*, Handbooks in Operations Research and Management Science. Elsevier, Amsterdam, The Netherlands, 2006.

[22] P. Ney. Dominating points and the asymptotics of large deviations for random walk on r^d. *The Annals of Probability*, 11:158–167, 1983.

[23] S. Parekh and J. Walrand. Quick simulation of rare events in networks. *IEEE Trans. Automat. Contr.*, 34:54–66, 1989.

[24] J. S. Sadowsky and J. A. Bucklew. On large deviations theory and asymptotically efficient monte carlo estimation. *IEEE Transactions on Information Theory*, 36:579–588, 1990.

[25] D. Sezer. Dynamic importance sampling for queueing networks with markov modulated arrivals and services. *Stochastic Processes and their Applications*, 119:491–517, 2009.

[26] D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Stat.*, 3:673–684, 1976.