Large deviations and importance sampling for a tandem network with slow-down

Paul Dupuis*, Kevin Leder†, and Hui Wang‡
Lefschetz Center for Dynamical Systems
Brown University
Providence, R.I. 02912, U.S.A.

Abstract

We consider a variant of the two node tandem Jackson network where the upstream server reduces its service rate when the downstream queue exceeds some prespecified threshold. The rare event of interest is the overflow of the downstream queue. Based on a game/subsolution approach, we rigorously identify the exponential decay rate of the rare event probabilities and construct asymptotically optimal importance sampling schemes.

1 Introduction

Consider a two-node tandem network where the downstream server is protected in such a way that the upstream server will drop its service rate whenever the downstream queue size exceeds a prespecified slow-down threshold. This type of queueing model was introduced in [12], and has potential applications in manufacturing and Ethernet design.

The present paper is interested in the overflow probability of the downstream queue during a busy cycle, assuming the system is stable. Little is known about the large deviation properties of this rare event and how to design efficient importance sampling schemes for simulation. To our best

*Research of this author supported in part by the National Science Foundation (NSF-DMS-0404806 and NSF-DMS-0706003) and the Army Research Office (W911NF-05-1-0289).

†Research of this author supported in part by the National Science Foundation (NSF-DMS-0404806 and NSF-DMS-0706003)

‡Research of this author supported in part by the National Science Foundation (NSF-DMS-0404806 and NSF-DMS-0706003).

knowledge, the only existing work is [10], where the authors proposed heuristically the exponential decay rate of the overflow probability and the most likely path leading to overflow. [10] also suggested an importance sampling scheme based on the most likely path, which turned out to be efficient only in certain cases.

The goal of the current paper is to provide a construction of asymptotically optimal importance sampling schemes for general parameter values and also a rigorous analysis of the variational problem for the large deviations rate. The analysis relies on a recently developed game/subsolution approach toward importance sampling [6, 4], and utilizes the techniques in [1, 3] to overcome the difficulties introduced by the discontinuous dynamics at the slow-down threshold. We should remark that the current paper analyzes only the most relevant case from [10], in which the second queue is the more likely bottleneck without slowdown, and the first queue is the more likely bottleneck with slowdown. However, the other cases can be dealt with in an analogous fashion. The approach can also be extended to networks with Markov modulated arrival/service rates [6].

The paper is organized as follows. In Sections 2 and 3 we describe the model and system dynamics. Section 4 states the large deviations result for the rare event probability of interest. Importance sampling schemes and their performance criterion are given in Section 5. Sections 6 through 8 are concerned with the Isaacs equation, construction of subsolutions, and the optimality of the corresponding dynamic importance sampling algorithms. The main result is stated in Section 9, numerical results are given in Section 10, and some technical proofs are collected in an appendix.

2 The model setup

We consider a variant of the standard two-node tandem Jackson network. Suppose that the arrival process is Poisson with rate λ, and the downstream service times are exponentially distributed with rate μ_2. The distribution of the upstream service times is as follows. Let $Q = \{(Q_1(t), Q_2(t)) : t \geq 0\}$ denote the system state, that is, $Q_1(t)$ is the length of the upstream queue at time t and $Q_2(t)$ is that of the downstream queue. Let n be the overflow level of the downstream queue, and θn the slow-down threshold [assuming $\theta \in (0, 1)$] for the upstream queue. Then the upstream service time distribution is exponential with rate μ_1 if $Q_2 < \theta n$ and exponential with a smaller rate ν_1 if $Q_2 \geq \theta n$. The probability of interest is the overflow

probability

$$p_n \doteq \mathbb{P}\left\{Q_2 = n \text{ before } Q = (0,0), \text{ after starting from } Q = (1,0)\right\}.$$

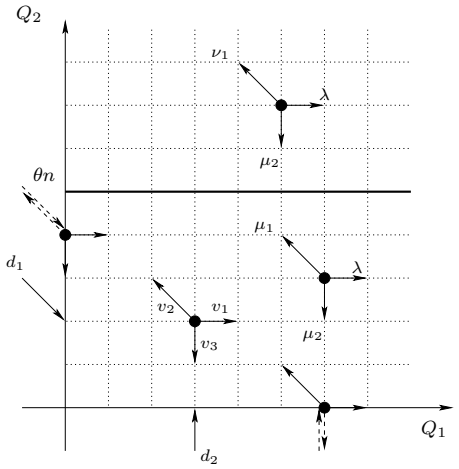

Figure 1: The system dynamics

To illustrate the main idea of game/subsolution approach, we restrict the analysis to the case where

$$\lambda < \nu_1 < \mu_2 \leq \mu_1. \tag{2.1}$$

Thus the downstream server is more likely to be the bottleneck when the system is below the slow-down threshold, while the upstream server becomes the bottleneck once the slow-down threshold is breached. When the slowdown mechanism is viewed as a control which tries to protect the downstream buffer from overflow, this is the most relevant case.

Remark 2.1. The paper [10] also considered two-node tandem Jackson networks without any server slow-down. For such a model the probability of interest p_n is a special case of one that appears in [4, Section 4.3.1] with $B_1 = \infty$ and $B_2 = 1$. One can use the subsolution there to build asymptotically optimal importance schemes for both $\mu_1 \leq \mu_2$ and $\mu_1 > \mu_2$.

3 The system dynamics

The state process Q is a continuous time pure jump Markov process defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In the interior of the state space the

possible jumps of Q belong to the set

$$\mathbb{V} = \{v_1 = (1,0),\ v_2 = (-1,1),\ v_3 = (0,-1)\}.$$

To describe the discontinuous dynamics on the boundary, we allow the process Q to make fictitious jumps of size v_{i+1} on the boundary $\{Q_i = 0\}$, but they have to be accounted for by pushing back the state along the direction of constraints

$$d_i = -v_{i+1},$$

so that the queue sizes remain non-negative [see Figure 1]. For every $x = (x_1, x_2) \in \mathbb{R}^2_+$ and $v \in \mathbb{V}$ let

$$\pi[x, v] \doteq \begin{cases} 0, & \text{if } x_i = 0 \text{ and } v = v_{i+1} \text{ for some } i = 1, 2, \\ v, & \text{otherwise.} \end{cases} \tag{3.1}$$

The required projection on the boundary of the state space will be handled by π. We also define the jump intensity function $r(x, v)$ by

$$r(x, v_1) = \lambda, \quad r(x, v_2) = \begin{cases} \mu_1, & \text{if } x_2 < \theta \\ \nu_1, & \text{if } x_2 \geq \theta \end{cases}, \quad r(x, v_3) = \mu_2,$$

and the total intensity function by

$$R(x) = \sum_{v \in \mathbb{V}} r(x, v).$$

Let $\{T_1, T_2, \ldots\}$ be the random jump times of the process Q with the convention $T_0 = 0$. The dynamics of Q are determined by the stochastic transition kernel $\Theta[\cdot|\cdot]$ on $\mathbb{R}_+ \times \mathbb{V}$ given $\mathbb{R}^2_+$, where

$$\begin{aligned} \Theta[dt, v|x] &\doteq \mathbb{P}\{T_{j+1} - T_j \in dt, Q(T_{j+1}) = nx + \pi[x, v] \,|\, Q(T_j) = nx\} \\ &= r(x, v) e^{-R(x)t} dt. \end{aligned} \tag{3.2}$$

In other words, the possible jumps of the process Q at state nx are $\{\pi[x, v] : v \in \mathbb{V}\}$ and the jump intensity from nx to $nx + \pi[x, v]$ is $r(x, v)$.

Notation. We collect here some useful notation [see Figure 2].

$$\begin{aligned} S &= \{(x_1, x_2) : x_1 > 0, \theta < x_2 < 1\}. \\ D &= \{(x_1, x_2) : x_1 > 0, 0 < x_2 < \theta\}. \\ \partial_1 &= \{(x_1, x_2) : x_1 = 0, x_2 > 0\}. \\ \partial_2 &= \{(x_1, x_2) : x_1 > 0, x_2 = 0\}. \\ \partial_e &= \{(x_1, x_2) : x_2 = 1\}. \end{aligned}$$

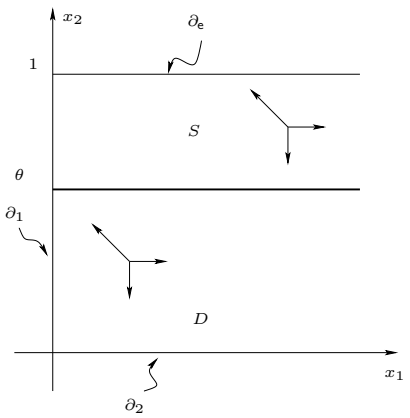

Figure 2: Notation for the State Space Partition

4 A representation for the exponential decay rate

The approach to be developed below is based on relations between large deviation properties and certain nonlinear partial differential equations (PDE). It will turn out that we will not need *solutions* to these PDE, but only *subsolutions*, and the PDE are simple enough that these subsolutions can be explicitly constructed. The PDE are described in terms of so-called Hamiltonians, with one for each region of different statistical behavior. As noted in the remark that follows, one can heuristically derive each Hamiltonian from the corresponding dynamics of the queueing process.

Given an arbitrary $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$, the relevant Hamiltonians are defined as follows.

1. Below the slow-down threshold [region D],

$$H(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \mu_1(e^{\alpha_2 - \alpha_1} - 1) + \mu_2(e^{-\alpha_2} - 1).$$

2. Above the slow-down threshold [region S],

$$H_s(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \nu_1(e^{\alpha_2 - \alpha_1} - 1) + \mu_2(e^{-\alpha_2} - 1).$$

3. On the boundary ∂_1,

$$H_{\partial_1}(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \mu_2(e^{-\alpha_2} - 1).$$

4. On the boundary ∂_2,

$$H_{\partial_2}(\alpha) \doteq \lambda(e^{\alpha_1} - 1) + \mu_1(e^{\alpha_2 - \alpha_1} - 1).$$

Remark 4.1. We will (heuristically) show how the function H occurs in a PDE characterization of p_n. In a similar manner one can relate the Hamiltonians to the optimal possible performance in importance sampling (see [5]). We introduce the scaled process $X^n(t) = Q(nt)/n$, and for each $x \in D$ define

$$p_n(x) = \mathbb{P}_x \left\{ X_2^n \text{ exceeds 1 before } X^n = 0 \right\}.$$

By conditioning on the value of the first jump (given that the process starts at x), it follows that

$$(\lambda + \mu_1 + \mu_2) p_n(x) = \lambda p_n \left(x + \frac{v_1}{n} \right) + \mu_1 p_n \left(x + \frac{v_2}{n} \right) + \mu_2 p_n \left(x + \frac{v_3}{n} \right).$$

If the probabilities $p_n(x)$ decay at an exponential rate then one would expect $W_n(x) = -\frac{1}{n} \log p_n(x)$ to converge to some function $W(x)$. By rewriting the previous display in terms of W_n we see

$$\begin{aligned} 0 \quad = \quad & \lambda \left[\exp \left[-n \left(W_n(x + v_1/n) - W_n(x) \right) \right] - 1 \right] \\ & + \mu_1 \left[\exp \left[-n \left(W_n(x + v_2/n) - W_n(x) \right) \right] - 1 \right] \\ & + \mu_2 \left[\exp \left[-n \left(W_n(x + v_3/n) - W_n(x) \right) \right] - 1 \right]. \end{aligned}$$

For a smooth function W let DW denote its gradient. If each term of the form $n \left(W_n(x + v_i/n) - W_n(x) \right)$ also converges to $\langle DW(x), v_i \rangle$, then W should satisfy the PDE

$$\lambda \left(e^{-\langle DW, v_1 \rangle} - 1 \right) + \mu_1 \left(e^{-\langle DW, v_2 \rangle} - 1 \right) + \mu_2 \left(e^{-\langle DW, v_3 \rangle} - 1 \right) = 0.$$

This is just $H(-DW(x)) = 0$, and the other Hamiltonians can be derived similarly. The definition in a form similar to a moment generating function is motivated by standard notation in large deviation theory, and is responsible for the minus sign. ■

Denote by L, L_s, L_{∂_1}, and L_{∂_2} the Legendre transforms of H, H_s, H_{∂_1}, and H_{∂_2}, respectively. That is, for example,

$$L(\beta) = \sup_{\alpha \in \mathbb{R}^2} \left[\langle \alpha, \beta \rangle - H(\alpha) \right].$$

We continue by defining the so-called *local rate function* associated with the scaled processes $X^n(t) = Q(nt)/n, t \geq 0$. To ease exposition, we use the notation $L_1 \oplus \cdots \oplus L_d$ to denote the inf-convolution of convex functions $\{L_1, L_2, \ldots, L_d\}$, that is, for any $\beta \in \mathbb{R}^2$

$$[L_1 \oplus \cdots \oplus L_d](\beta) \doteq \inf \left\{ \sum_{i=1}^d \rho_i L_i(\beta_i) : \rho_i \geq 0, \sum_{i=1}^d \rho_i = 1, \sum_{i=1}^d \rho_i \beta_i = \beta \right\}.$$

The local rate function, denoted by $L(x, \beta)$, is defined as follows.

1. For $x \in S$, $L(x, \beta) \doteq L_s(\beta)$, and for $x \in D$, $L(x, \beta) \doteq L(\beta)$.
2. For $x \in \partial_2$, $L(x, \beta) = [L \oplus L_{\partial_2}](\beta)$ if $\beta_2 \geq 0$ and ∞ if $\beta_2 < 0$.
3. For $x \in \partial_1$,

$$L(x, \beta) \doteq \begin{cases} [L_s \oplus L_{\partial_1}](\beta), & \text{if } x_2 > \theta, \beta_1 \geq 0, \\ [L \oplus L_s \oplus L_{\partial_1}](\beta), & \text{if } x_2 = \theta, \beta_1 \geq 0, \\ [L \oplus L_{\partial_1}](\beta), & \text{if } x_2 < \theta, \beta_1 \geq 0, \\ \infty, & \text{if } \beta_1 < 0. \end{cases}$$

4. For $x = (0, 0)$, $L(x, \beta) = [L \oplus L_{\partial_1} \oplus L_{\partial_2}](\beta)$ if $\beta_1, \beta_2 \geq 0$ and ∞ otherwise.
5. For $x \in \{(x_1, \theta) : x_1 > 0\}$, $L(x, \beta) \doteq [L \oplus L_s](\beta)$.

Theorem 4.2.

$$\lim_n -\frac{1}{n} \log p_n = \inf \int_0^\tau L(\phi(t), \dot{\phi}(t))\, dt,$$

where the infimum is taken over all absolutely continuous functions $\phi : [0, \infty) \to \mathbb{R}^2_+$ *such that*

$$\phi(0) = 0, \quad \tau \doteq \inf \{t \geq 0 : [\phi(t)]_2 = 1\} < \infty.$$

The main difficulties in the proof of Theorem 4.2 are due to the discontinuous dynamics of the prelimit process X^n. The discontinuities come in two forms: those along the interface between the region D and the slowdown region S, and those along the boundary of the state space due to non-negativity constraints on queue length. The first type of discontinuity is well understood in this particular setting (two regions of continuous behavior separated by a single smooth interface). See for example [1, Chapter 7] where the model discussed is a discrete time random walk, and [3, Theorem 3.1] for an application in a continuous time setting. In a very general setting (as in [1, Chapter 7]), the local rate function appears to be more complicated than it does here, since additional "stability-about-the-interface" constraints not present in the inf-convolution must be added. However, owing to the structure of the server slowdown dynamics these constraints are implicitly contained in the inf-convolution formula.

The discontinuities along the boundary can be dealt with by considering an unconstrained process which is mapped to the constrained process via a

Skorokhod mapping Γ. It is easily shown that the mapping Γ is Lipschitz [2]. With the continuity of Γ the proof is now a simple application of the Contraction Principle.

5 The basics of importance sampling

Consider a family of events $\{A_n\}$ in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that

$$\lim_n -\frac{1}{n} \log \mathbb{P}(A_n) = \gamma$$

for some positive constant γ. In order to estimate $\mathbb{P}(A_n)$, importance sampling generates samples under a different probability distribution $\mathbb{Q}$ (i.e., change of measure) such that $\mathbb{P} \ll \mathbb{Q}$, and forms an estimator by averaging independent replications of

$$\hat{p}_n \doteq 1_{A_n} \frac{d\mathbb{P}}{d\mathbb{Q}},$$

where $d\mathbb{P}/d\mathbb{Q}$ is the Radon-Nikodým derivative or likelihood ratio. It is easy to check that $\hat{p}_n$ is unbiased.

The rate of convergence of the importance sampling estimator is determined by the variance, or equivalently the second moment, of $\hat{p}_n$. The smaller the second moment, the faster the convergence. However, by Jensen's inequality

$$\limsup_n -\frac{1}{n} \log E^{\mathbb{Q}}[\hat{p}_n^2] \leq \limsup_n -\frac{2}{n} \log E^{\mathbb{Q}}[\hat{p}_n] = 2\gamma.$$

We say the importance sampling estimator $\hat{p}_n$ or the change of measure $\mathbb{Q}$ is *asymptotically optimal* if the upper bound is achieved, i.e., if

$$\liminf_n -\frac{1}{n} \log E^{\mathbb{Q}}[\hat{p}_n^2] \geq 2\gamma.$$

Sometimes 2γ is referred to simply as the "optimal decay rate."

Remark 5.1. The requirement that $\mathbb{P}$ be absolutely continuous with respect to $\mathbb{Q}$ is more stringent than necessary. It is sufficient that $\mathbb{P}$ be absolutely continuous with respect to $\mathbb{Q}$ on a sub-σ-algebra that contains A_n, in which case the likelihood ratio is defined as the Radon-Nikodym derivative of $\mathbb{P}$ and $\mathbb{Q}$ when they are restricted on this sub-σ-algebra. This fact is needed in the present paper, since we consider events on a potentially unbounded time interval.

6 Dynamic importance sampling schemes

Dynamic, or state-dependent, importance sampling schemes can be characterized by an alternative stochastic transition kernel $\bar{\Theta}^n[\cdot|\cdot]$ on $\mathbb{R}_+ \times \mathbb{V}$ given $\mathbb{R}^2_+$. That is, the dynamics of Q under the sampling probability measure, say $\mathbb{Q}$, are determined by

$$\mathbb{Q}\{T_{j+1} - T_j \in dt, Q(T_{j+1}) = nx + \pi[x, v] \,|\, Q(T_j) = nx\} = \bar{\Theta}^n[dt, v|x],$$

where $\{T_1, T_2, \ldots\}$ are jump times for the process Q with the convention $T_0 = 0$. The corresponding importance sampling estimator is as follows. Let $s_j \doteq T_j - T_{j-1}$ (sojourn times) and $Q(T_j) - Q(T_{j-1}) \doteq \pi[Q(T_{j-1}), \triangle_j]$ (jump sizes). Here $\triangle_j$ is the direction of the jth jump. Note that the effect of this jump may be negated by the boundary condition [see definition of mapping π in (3.1)]. Define

$$N \doteq \inf\{k \geq 1 : Q_2(T_k) = n \text{ or } Q(T_k) = (0, 0)\}. \tag{6.1}$$

Then the importance sampling estimator is

$$\hat{p}_n \doteq 1_{\{Q_2(T_N)=n\}} \prod_{j=1}^{N} \frac{\Theta[ds_j, \triangle_j|Q(T_{j-1})/n]}{\bar{\Theta}^n[ds_j, \triangle_j|Q(T_{j-1})/n]}. \tag{6.2}$$

We should point out that $\hat{p}_n$ will be used for the analysis of the asymptotic optimality. However, the implementation will use a different version of $\hat{p}_n$. See Remark 6.2 for more details.

Remark 6.1. From the point of view of implementation, a convenient choice is to use stochastic transition kernels determined by an alternative jump intensity function $\bar{r}(x, v)$, i.e.,

$$\bar{\Theta}^n[dt, v|x] \equiv \bar{r}(x, v)e^{-\bar{R}(x)t}dt, \qquad \text{where } \bar{R}(x) = \sum_{v \in \mathbb{V}} \bar{r}(x, v). \tag{6.3}$$

The modified rates will depend on the large deviation parameter n, but for ease of notation this dependence is omitted. As we will see, one can construct asymptotically optimal importance sampling schemes by considering simple mixtures of such transition kernels.

Remark 6.2. The likelihood ratio in the definition of $\hat{p}_n$ is with respect to continuous-time sample paths. However, most importance sampling literature uses the likelihood ratio defined on the embedded discrete-time sample

paths, or more precisely, a conditional expectation of the continuous-time likelihood ratio. To illustrate, consider the case where $\bar{\Theta}^n[dt, v|x]$ is defined as in Remark 6.1. Under this transition kernel, the process Q is again a continuous time Markov process with $\bar{r}(x, v)$ as the jump intensity from state nx to $nx + \pi[x, v]$. Then the importance sampling estimator based on the embedded discrete time Markov chain $Z = \{Z(i) = Q(T_i) : i \geq 0\}$ is

$$\bar{p}_n \doteq 1_{\{Z_2(N)=n\}} \prod_{j=0}^{N-1} \frac{r[Z(j)/n, \triangle_{j+1}]/R(Z(j)/n)}{\bar{r}[Z(j)/n, \triangle_{j+1}]/\bar{R}(Z(j)/n)}.$$

Recalling the definitions of Θ and $\bar{\Theta}$ in equations (3.2) and (6.3)

$$E^{\mathbb{Q}}[\hat{p}_n|Q(T_1), \ldots, Q(T_N)] = \bar{p}_n. \tag{6.4}$$

This implies that $\bar{p}_n$ is unbiased and its second moment is at most that of $\hat{p}_n$. It is important to distinguish these two importance sampling estimators. $\hat{p}_n$ is more suitable for analysis, and $\bar{p}_n$ is easier for implementation. We should analyze the change of measure using $\hat{p}_n$, and use its discrete-time counterpart $\bar{p}_n$ for implementation. Note that if one can establish the asymptotic optimality of $\hat{p}_n$, then $\bar{p}_n$ is automatically asymptotically optimal.

7 The Isaacs equation

It has been established that importance sampling is closely related to differential games and that subsolutions to the corresponding Isaacs equation (a nonlinear PDE) can be used for constructing efficient importance sampling schemes [6, 4]. The Isaacs equation can be stated in terms of the Hamiltonians introduced in Section 4.

Define the function ℓ by

$$\ell(x) = \begin{cases} x \log x - x + 1, & \text{if } x \geq 0, \\ \infty \quad , & \text{if } x < 0. \end{cases} \tag{7.1}$$

Let $\mathcal{R}$ be the collection of functions from $\mathbb{V}$ to $(0, \infty)$, and for every $\hat{r} \in \mathcal{R}$ define

$$\mathbb{F}(\hat{r}) \doteq \sum_{v \in \mathbb{V}} \hat{r}[v] \cdot v.$$

For each $\alpha = (\alpha_1, \alpha_2) \in \mathbb{R}^2$, the direct continuous time analogue of the discrete time Hamiltonian used in [4] is

$$\sup_{\bar{r} \in \mathcal{R}} \inf_{\hat{r} \in \mathcal{R}} \left[\langle \alpha, \mathbb{F}(\hat{r}) \rangle + \sum_{v \in \mathbb{V}} \left(\hat{r}[v] \log \left(\frac{\bar{r}[v]}{r(x, v)} \right) + r(x, v) - \bar{r}[v] \right) \right.$$

$$+\sum_{v\in\mathbb{V}} r(x,v)\ell\left(\frac{\hat{r}[v]}{r(x,v)}\right)\Bigg].$$

We will find it convenient to rewrite the Hamiltonian in the form

$$\mathbb{H}(x,\alpha) \doteq \sup_{\bar{r}\in\mathcal{R}}\inf_{\hat{r}\in\mathcal{R}}\left[\langle\alpha,\mathbb{F}(\hat{r})\rangle + 2\sum_{v\in\mathbb{V}} r(x,v)\ell\left(\frac{\hat{r}[v]}{r(x,v)}\right) - \sum_{v\in\mathbb{V}}\bar{r}[v]\ell\left(\frac{\hat{r}[v]}{\bar{r}[v]}\right)\right].$$

Recall that DW denotes the gradient for a smooth function W. A (classical) solution to the Isaacs equation is a continuously differentiable function $W : \mathbb{R}_+ \times [0,1] \to \mathbb{R}$ satisfying

1. $\mathbb{H}(x, DW(x)) = 0$ for $x \in (0,\infty)\times(0,1)$,
2. $\langle DW(x), d_i\rangle = 0$ for $x \in \partial_i$,
3. $W(x) = 0$ for $x \in \partial_e$.

A formal derivation analogous to the one given in Section 4 shows that this PDE should characterize the optimal decay rate among all importance sampling schemes. In general one would not expect classical solutions to exist, and thus one should work with a weaker notion of solution. However, we will not need or use solutions, but rather only certain subsolutions whose regularity properties will be specified when they are introduced.

We recall the functions H and H_s introduced in Section 4. Abusing the notation, we define for every α

$$\mathbb{H}_s(\alpha) \doteq -2H_s(-\alpha/2), \quad \mathbb{H}(\alpha) \doteq -2H(-\alpha/2).$$

We have the following result, whose proof is straightforward and thus omitted.

Proposition 7.1. *Fix an arbitrary $\alpha = (\alpha_1,\alpha_2) \in \mathbb{R}^2$. Then for $x = (x_1,x_2)$ the Hamiltonian $\mathbb{H}(x,\alpha)$ satisfies the following properties.*

1. *If $x_2 \geq \theta$, $\mathbb{H}(x,\alpha) = \mathbb{H}_s(\alpha)$ with the saddle point*

$$\bar{r}_s^*(\alpha) = \hat{r}_s^*(\alpha) = \left(\lambda e^{-\alpha_1/2}, \nu_1 e^{(\alpha_1-\alpha_2)/2}, \mu_2 e^{\alpha_2/2}\right).$$

2. *If $x_2 < \theta$, $\mathbb{H}(x,\alpha) = \mathbb{H}(\alpha)$ with the saddle point*

$$\bar{r}^*(\alpha) = \hat{r}^*(\alpha) = \left(\lambda e^{-\alpha_1/2}, \mu_1 e^{(\alpha_1-\alpha_2)/2}, \mu_2 e^{\alpha_2/2}\right).$$

Remark 7.2. The Isaacs equation is associated with a differential game where the $\bar{r}$-player represents the choice of change of measure and the $\hat{r}$-player is introduced through a representation formula for the large deviation rate of decay. The explicit formula for the $\bar{r}$-component of the saddle point given in Proposition 7.1 is particularly useful in the construction of importance sampling schemes. Roughly speaking, for a given subsolution $\bar{W}$, depending on the state of the process, we use $\bar{r}_s^*(D\bar{W})$ or $\bar{r}^*(D\bar{W})$ as the alternative jump intensity function for simulation.

Remark 7.3. In general one should be more careful with the definition of the Isaacs equation on the interface $\{(x_1, x_2) : x_2 = \theta\}$, and in fact if we were considering a more general class of rare event estimation problems one would use $(\mathbb{H}_s \wedge \mathbb{H})(DW) = 0$ on the interface. This observation is crucial for the construction of subsolutions that lead to asymptotically optimal importance sampling schemes, but only when the most likely path (i.e. a minimizer to the variational problem in Theorem 4.2) leading to the rare event will spend non-trivial time on the interface (i.e., the Lebesgue measure of the time the minimizer is on the interface is positive). In the present setting, however, it will turn out that the most likely path does not spend positive time on the interface.

8 Subsolutions and importance sampling schemes

A *classical subsolution* to the Isaacs equation is a continuously differentiable function $\bar{W} : \mathbb{R}_+ \times [0,1] \to \mathbb{R}$ such that

1. $\mathbb{H}(x, D\bar{W}(x)) \geq 0$ for all $x \in (0,\infty) \times (0,1)$,
2. $\langle D\bar{W}(x), d_i \rangle \geq 0$ for all $x \in \partial_i$,
3. $\bar{W}(x) \leq 0$ for all $x \in \partial_e$.

Classical subsolutions can often be constructed from a mollification of piecewise affine subsolutions, see [4, 6] and the many examples therein.

We will associate importance sampling schemes to subsolutions. The performance of the scheme will be measured by the value of the subsolution at 0, with larger values of $\bar{W}(0)$ indicating better performance. Note that for any subsolution $\bar{W}(0)$ is bounded above by the value of the solution at the origin.

8.1 Piecewise affine subsolutions

The goal of this section is to construct a piecewise affine subsolution whose value at the origin is maximal. Consider the variational problem introduced in the statement of Theorem 4.2, but with a general initial condition $x \in [0,1] \times \mathbb{R}_+$:

$$V(x) = \inf \left\{ \int_0^\tau L(\phi(t), \dot{\phi}(t)) dt : \phi(0) = x, [\phi(\tau)]_2 = 1, \tau \geq 0 \right\}. \qquad (8.1)$$

It can be shown formally by a dynamic programming argument that $\mathbb{H}(x, 2DV(x)) = 0$, which suggests that the large deviations most likely path might be useful in constructing subsolutions. In [10, Proposition 10] the authors propose that the most likely path to overflow will travel along the boundary ∂_1 from $(0,0)$ to $(0,\theta)$ pushing against the boundary during this leg of the journey. What is meant by "pushing against the boundary" is that in the absence of non-negativity constraints on queue length, the (large deviation) change of jump rates would lead to a trajectory whose velocity has a negative first component. This velocity is projected back along the direction d_1 at no cost. From $(0,\theta)$ to $(0,1)$ the path will also travel along ∂_1 but this time gliding along the boundary. What is meant here is that the change of rates produces exactly velocity zero in the first component, and no projection is needed. If the conjecture of [10] is correct, these statements give important information on the gradient of the solution in a neighborhood of the optimal trajectory. If the subsolution we construct is to be close to the solution at $x = 0$, then it must be close all along the optimal trajectory. Hence we obtain necessary conditions on the gradient of the subsolution along the optimal trajectory.

We next formally derive constraints on the gradient of the function V along the proposed optimal trajectory. We say "formal" since it is not known if V is sufficiently smooth to justify all the calculations. However, our goal is to simply motivate the construction of a particular subsolution. As noted above, in region D the optimal trajectory pushes into the boundary and is returned along the direction d_1 at no cost. This implies that the value function V should be constant along this direction. We thus obtain two constraints on $\alpha^{[1]} = 2DV(x)$:

$$\mathbb{H}(\alpha^{[1]}) = 0, \quad \langle d_1, \alpha^{[1]} \rangle = 0.$$

In region S $2DV(x)$ should satisfy the corresponding part of the PDE that applies on the interior of the domain. At the same time, the linking of the optimal velocity with the gradient via dynamic programming implies that

$\alpha^{[0]} = 2DV(x)$ should be dual (in the sense of convex duality) to a point whose first component is zero. Thus we obtain the constraints

$$\mathbb{H}_s(\alpha^{[0]}) = 0, \quad \frac{\partial \mathbb{H}_s}{\partial \alpha_1}(\alpha^{[0]}) = 0. \tag{8.2}$$

There are two roots with this property, and the correct (i.e., useful) root also satisfies $\alpha_1^{[0]}, \alpha_2^{[0]} < 0$ [see Lemma 8.1].

This identifies constraints on the gradient of a subsolution in the neighborhood of the optimal trajectory. In order to satisfy the boundary condition along the boundary ∂_2 we need the gradients $\alpha^{[2]} = 2\log(\mu_2/\lambda)(-1, 0)$, and $\alpha^{[3]} = (0, 0)$. As we will see, these choices allow the construction of a function which satisfies the subsolution property at all points in the domain and with a nearly optimal value at zero. Figure 3 shows how these vectors relate to the Hamiltonians.

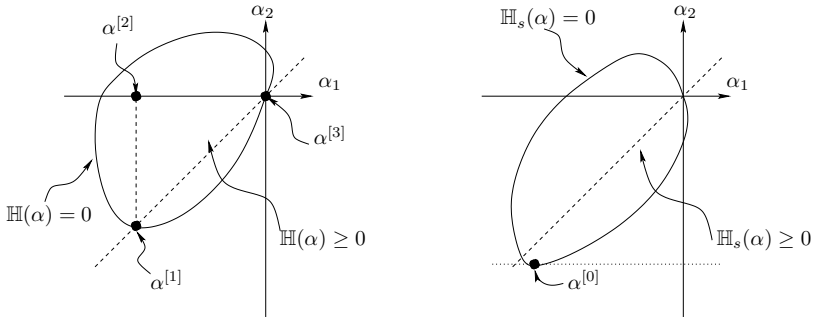

Figure 3: The gradients of the subsolution

In terms of the problem data we find $\alpha^{[1]} = 2\log(\mu_2/\lambda)(-1, -1)$, $\alpha^{[2]} = 2\log(\mu_2/\lambda)(-1, 0)$, and $\alpha^{[3]} = (0, 0)$. The existence of $\alpha^{[0]}$ is dealt with in the following lemma.

Lemma 8.1. *There exists a unique $z > 1$ that satisfies the equation*

$$2\sqrt{\lambda\nu_1 z} + \frac{\mu_2}{z} = \lambda + \nu_1 + \mu_2. \tag{8.3}$$

Furthermore, the unique solution $\alpha^{[0]}$ to equation (8.2) with $\alpha_1^{[0]}, \alpha_2^{[0]} < 0$ can be expressed as

$$\alpha^{[0]} = (\alpha_1^{[0]}, \alpha_2^{[0]}) = (-\log(\nu_1 z/\lambda), -2\log z)\,. \tag{8.4}$$

We also have

$$\mathbb{H}_s(\alpha^{[1]}) = 0, \qquad \alpha_2^{[0]} < \alpha_1^{[1]} < \alpha_1^{[0]}, \qquad \langle \alpha^{[0]}, d_1 \rangle \geq 0.$$

The proof is deferred to an appendix. Note that the equation (8.3) for z is the same as equation (31) in [10] with z replaced by z^{-1}. We also let

$$\gamma = (1-\theta)\log z + \theta \log(\mu_2/\lambda). \tag{8.5}$$

The first term is the cost of the optimal trajectory between $(0,\theta)$ and $(0,1)$, the second term is the cost of the trajectory traveling between $(0,0)$ and $(0,\theta)$. Fix an arbitrary small $\delta > 0$, and define the affine functions

$$\begin{aligned}
\bar{W}_0^\delta(x) &= \langle x, \alpha^{[0]} \rangle + 2\log z, \\
\bar{W}_1^\delta(x) &= \langle x, \alpha^{[1]} \rangle + 2\gamma, \\
\bar{W}_2^\delta(x) &= \langle x, \alpha^{[2]} \rangle + 2\gamma - \delta, \\
\bar{W}_3^\delta(x) &= \langle x, \alpha^{[3]} \rangle + 2\gamma - 2\delta.
\end{aligned}$$

Let $\bar{W}^\delta \doteq \bar{W}_0^\delta \wedge \bar{W}_1^\delta \wedge \bar{W}_2^\delta \wedge \bar{W}_3^\delta$. Figure 4 depicts the gradients of $\bar{W}^\delta$ in different regions of the state space, with $\bar{W}^\delta = \bar{W}_i^\delta$ in region G_i.

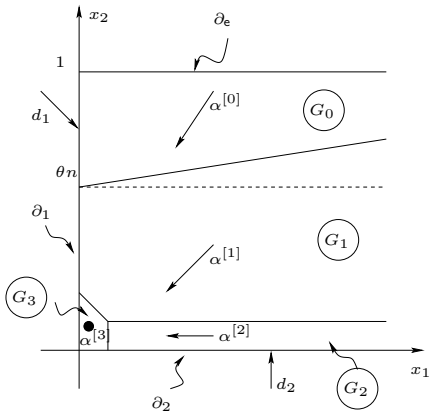

Figure 4: The subsolution

Lemma 8.1 and a simple computation show that $\bar{W}^\delta$ indeed defines a piecewise affine subsolution, that is, $\bar{W}^\delta$ satisfies the definition of classical subsolution in the regions where it is smooth.

Remark 8.2. In Remark 7.3 it was mentioned that the subsolution would be more complicated if the optimal trajectory spent significant amount of time on the interface. In particular this would require that an additional affine function $\bar{W}_4^\delta$ be constructed such that $(\mathbb{H} \wedge \mathbb{H}_s)(D\bar{W}_4^\delta) \geq 0$. In addition it would be necessary that $\bar{W}^\delta = \bar{W}_4^\delta$ for x in an $O(\delta)$ strip around the interface.

8.2 The mollification

In the proof of asymptotic optimality of the estimator $\hat{p}_n$ it is essential that the subsolutions we work with have a bounded second derivative. As far as the authors know this boundedness condition is indispensable. Since the function $\bar{W}^\delta$ is not smooth, this problem is resolved by mollifying the function. As we will see, the implementation of the algorithm corresponding to the mollified function is only slightly more computationally demanding than it would be for the corresponding un-mollified version.

A natural mollification technique for this problem is *exponential weighting*. See [6, 4] for a discussion of the benefits of this method and a comparison with other standard mollifications. Let ε be a small positive number, and define

$$W^{\varepsilon,\delta}(x) \doteq -\varepsilon \log \sum_{i=0}^{3} \exp\left\{-\frac{1}{\varepsilon}\bar{W}_i^\delta(x)\right\}.$$

The function $W^{\varepsilon,\delta}$ is continuously differentiable, and

$$DW^{\varepsilon,\delta}(x) = \sum_{i=0}^{3} \rho_i^{\varepsilon,\delta}(x)\alpha^{[i]}, \quad \rho_i^{\varepsilon,\delta}(x) \doteq \frac{\exp\left\{-\bar{W}_i^\delta(x)/\varepsilon\right\}}{\sum_{k=0}^{3}\exp\left\{-\bar{W}_k^\delta(x)/\varepsilon\right\}}.$$

Note that for every x, $\{\rho_i^{\varepsilon,\delta}(x) : i = 0, 1, 2, 3\}$ forms a probability vector since

$$\rho_i^{\varepsilon,\delta}(x) > 0, \qquad \sum_{i=0}^{3} \rho_i^{\varepsilon,\delta}(x) = 1.$$

8.3 The importance sampling algorithm

For each $i = 0, 1, 2, 3$, let

$$\bar{r}_s^{[i]} \doteq \bar{r}_s^*(\alpha^{[i]}), \quad \bar{r}^{[i]} \doteq \bar{r}^*(\alpha^{[i]}). \tag{8.6}$$

Explicit formulas for $\{\bar{r}_s^{[i]}, \bar{r}^{[i]}\}$ can be found in Remark 8.3. For each i, also define the jump intensity functions $\{\bar{r}_i(x, v) : v \in \mathbb{V}\}$ and total intensity function $\bar{R}_i(x)$ by

$$\bar{r}_i(x, v_k) \doteq \begin{cases} \left[\bar{r}_s^{[i]}\right]_k & \text{if } x_2 \geq \theta \\ \left[\bar{r}^{[i]}\right]_k & \text{if } x_2 < \theta \end{cases}, \qquad \bar{R}_i(x) \doteq \sum_{k=0}^{3} \bar{r}_i(x, v_k).$$

The corresponding stochastic transition kernel on $\mathbb{R}_+ \times \mathbb{V}$ given $\mathbb{R}_+ \times [0, 1]$, denoted by $\bar{\Theta}^{[i]}$, is

$$\bar{\Theta}^{[i]}[dt, v|x] \doteq \bar{r}_i(x, v)e^{-\bar{R}_i(x)t}dt.$$

That is, the stochastic transition kernel $\bar{\Theta}^{[i]}$ corresponds to a continuous time Markov process whose jump intensity from nx to $nx + \pi[x, v]$ is $\bar{r}_i(x, v)$.

The dynamic importance sampling scheme corresponding to $W^{\varepsilon,\delta}$ uses a mixture of $\{\bar{\Theta}^{[i]}\}$:

$$\bar{\Theta}^n[\cdot|x] = \bar{\Theta}^{\varepsilon,\delta}[\cdot|x] = \sum_{i=0}^{3} \rho_i^{\varepsilon,\delta}(x)\bar{\Theta}^{[i]}[\cdot|x],$$

and the corresponding importance sampling estimator $\hat{p}_n$ is just as defined in (6.2). In other words, the importance sampling simulates the process Q in the following fashion. Given that the current state of Q is nx, the algorithm selects a random index I taking values in $\{0, 1, 2, 3\}$ according to the weights $\{\rho_i^{\varepsilon,\delta}(x) : i = 0, 1, 2, 3\}$. Then the algorithm simulates the sojourn time according to the exponential distribution with rate $\bar{R}_I(x)$ and the jump size $\pi[x, v]$ with probability $\bar{r}_I(x, v)/\bar{R}_I(x)$.

In general one can allow ε, δ to depend on n, and denote them by ε_n, δ_n. The corresponding stochastic transition kernel will be $\bar{\Theta}^n[\cdot|x] = \bar{\Theta}^{\varepsilon_n,\delta_n}[\cdot|x]$, and the corresponding importance sampling estimator will still be denoted by $\hat{p}_n$ when no confusion is incurred.

Remark 8.3. Using Proposition 7.1, Lemma 8.1 and (8.6) it follows that

$$\bar{r}_s^{[0]} = (\sqrt{\lambda\nu_1 z}, \sqrt{\lambda\nu_1 z}, \mu_2/z), \quad \bar{r}^{[0]} = (\sqrt{\lambda\nu_1 z}, \mu_1\sqrt{\lambda z/\nu_1}, \mu_2/z),$$

$$\bar{r}_s^{[1]} = (\mu_2, \nu_1, \lambda), \quad \bar{r}^{[1]} = (\mu_2, \mu_1, \lambda),$$

$$\bar{r}_s^{[2]} = (\mu_2, \lambda\nu_1/\mu_2, \mu_2), \quad \bar{r}^{[2]} = (\mu_2, \lambda\mu_1/\mu_2, \mu_2),$$

$$\bar{r}_s^{[3]} = (\lambda, \nu_1, \mu_2), \quad \bar{r}^{[3]} = (\lambda, \mu_1, \mu_2).$$

9 The main results

We recall the definition of γ in (8.5).

Theorem 9.1. *Suppose $\delta_n \to 0$, $\varepsilon_n/\delta_n \to 0$ and $n\varepsilon_n \to \infty$. Then the corresponding importance sampling estimator $\hat{p}_n$ satisfies*

$$\liminf_n -\frac{1}{n}\log[\text{2nd moment of } \hat{p}_n] \geq 2\gamma.$$

This theorem can be shown by a verification argument analogous to [4, Theorem 3.18]. Since by Lemma 8.1 $\mathbb{H}(\alpha^{[1]}) = \mathbb{H}_s(\alpha^{[1]}) = 0$, the subsolution inequality holds on both sides of the slow-down interface (see Figure 4).

Hence the only major distinction is that in [4] the domain is assumed to be compact so that the exit time can be shown to have bounded moment generating function in a small neighborhood of origin. This compactness condition turns out to be unnecessary, and a proof of this fact is given in Appendix B.

Theorem 9.2. *We have the large deviation result*

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

Proof. Since $E^{\mathbb{Q}}[\hat{p}_n^2] \geq (E^{\mathbb{Q}}[\hat{p}_n])^2 = p_n^2$, Theorem 9.1 implies the large deviations upper bound

$$\liminf_n -\frac{1}{n}\log p_n \geq \gamma. \tag{9.1}$$

Note that this upper bound can also be shown by applying a direct verification argument [7] to the control representation of the large deviations rate function [see Theorem 4.2], using the classical subsolution $W^{\varepsilon_n,\delta_n}/2$.

We now consider the lower bound. By Theorem 4.2, we only need to show that there exists an absolutely continuous function $\phi^* : \mathbb{R}_+ \to \mathbb{R}_+^2$ such that

$$\int_0^\tau L(\phi^*(t), \dot{\phi}^*(t))\, dt \leq \gamma, \qquad \text{where } \tau \doteq \inf\{t \geq 0 : [\phi^*(t)]_2 = 1\} < \infty.$$

Theorems 4.2 and 9.1 imply that if such a ϕ^* exists it will be an optimal path.

The construction of ϕ^* is based on the change of measure determined by $\bar{r}_s^{[0]}$ and $\bar{r}^{[1]}$ [see Remark 8.3]. In other words, let

$$\beta_s^* \doteq \sum_{i=1}^3 \left[\bar{r}_s^{[0]}\right]_i v_i = (0, \sqrt{\lambda\nu_1 z} - \mu_2/z),$$

and

$$\beta^* \doteq \sum_{i=1}^3 \left[\bar{r}^{[1]}\right]_i v_i = (\mu_2 - \mu_1, \mu_1 - \lambda).$$

Since $[\beta^*]_1 = \mu_2 - \mu_1 \leq 0$, we further define, abusing the notation, $\pi[\beta^*]$ to be the projection of β^* onto ∂_1 along the direction of constraint $d_1 = (1, -1)$. Thus

$$\pi[\beta^*] = \beta^* - [\beta^*]_1\, d_1 = (0, \mu_2 - \lambda).$$

To ease exposition, write $\beta_s^* \doteq (0, u_s)$ and $\pi[\beta^*] \doteq (0, u)$, where

$$u_s \doteq \sqrt{\lambda\nu_1 z} - \mu_2/z, \quad u \doteq \mu_2 - \lambda.$$

Note that $u_s > 0$ thanks to (A.1) and $u > 0$ since $\mu_2 > \lambda$. Define $\phi^*(t)$ such that

$$\dot{\phi}^*(t) = \begin{cases} \pi[\beta^*], & \text{if } 0 \le t < \theta u^{-1}, \\ \beta_s^*, & \text{if } \theta u^{-1} \le t \le \theta u^{-1} + (1-\theta)u_s^{-1}. \end{cases}$$

Then ϕ^* is a vertical path passing through $(0,0) \to (0,\theta) \to (0,1)$, and by the definition of L [see Section 4],

$$\int_0^\tau L(\phi^*(t), \dot{\phi}^*(t))\, dt = \theta u^{-1} \cdot [L \oplus L_{\partial_1}](\pi[\beta^*]) + (1-\theta)u_s^{-1} \cdot [L_s \oplus L_{\partial_1}](\beta_s^*).$$

However, note that L_s and H_s are conjugate functions and β_s^* is conjugate to $-\alpha^{[0]}/2$, since by direct computation

$$D_\alpha H_s(-\alpha^{[0]}/2) = (0, u_s) = \beta_s^*.$$

Therefore, observing $H_s(-\alpha^{[0]}/2) = -\mathbb{H}_s(\alpha^{[0]})/2 = 0$, we have

$$[L_s \oplus L_{\partial_1}](\beta_s^*) \le L_s(\beta_s^*) = \langle -\alpha^{[0]}/2, \beta_s^* \rangle - H_s(-\alpha^{[0]}/2) = u_s \log z.$$

The local rate functions L, L_s, L_{∂_1}, L_{∂_2}, and their inf-convolutions have standard representations in terms of the ℓ function as defined in (7.1) [3, 11]. In particular,

$$\begin{aligned} [L \oplus L_{\partial_1}](\beta) \;=\; \inf \Bigg\{ & \lambda\ell\left(\frac{\bar\lambda}{\lambda}\right) + \bar\rho\mu_1\ell\left(\frac{\bar\mu_1}{\mu_1}\right) + \mu_2\ell\left(\frac{\bar\mu_2}{\mu_2}\right) : \\ & \bar\rho \in [0,1],\ \bar\lambda v_1 + \bar\rho\bar\mu_1 v_2 + \bar\mu_2 v_3 = \beta \Bigg\}. \end{aligned}$$

Since

$$\pi[\beta^*] = \bar\lambda v_1 + \bar\rho\bar\mu_1 v_2 + \bar\mu_2 v_3$$

where

$$\bar\rho \doteq \mu_2/\mu_1 \in [0,1], \quad \bar\lambda = \mu_2, \quad \bar\mu_1 = \mu_1, \quad \bar\mu_2 = \lambda,$$

it follows that

$$\begin{aligned} [L \oplus L_{\partial_1}](\pi[\beta^*]) \;&\le\; \bar\rho\mu_1\ell\left(\frac{\bar\mu_1}{\mu_1}\right) + \lambda\ell\left(\frac{\bar\lambda}{\lambda}\right) + \mu_2\ell\left(\frac{\bar\mu_2}{\mu_2}\right) \\ &=\; \lambda\ell\left(\frac{\mu_2}{\lambda}\right) + \mu_2\ell\left(\frac{\lambda}{\mu_2}\right) \\ &=\; u\log(\mu_2/\lambda). \end{aligned}$$

Therefore,

$$\int_0^\tau L(\phi^*(t), \dot{\phi}^*(t))\, dt \le \theta \log(\mu_2/\lambda) + (1-\theta)\log z = \gamma.$$

This completes the proof. ■

Corollary 9.3. *Suppose $\delta_n \to 0$, $\varepsilon_n/\delta_n \to 0$ and $n\varepsilon_n \to \infty$. Then the corresponding importance sampling estimator $\hat{p}_n$ is asymptotically optimal.*

Remark 9.4. Analogous to [4, Theorem 3.6], a near asymptotic optimality result can be shown for the case where $\varepsilon_n \equiv \varepsilon$ and $\delta_n \equiv \delta$. It also suggests that a good strategy is to set $\delta_n = -\varepsilon_n \log \varepsilon_n$ [4, Remark 3.7].

Remark 9.5. Our results show that the heuristically derived exponential decay rate of p_n and the limit most likely path to overflow in [10] are indeed correct.

10 Numerical results

As discussed in Remark 6.2, the numerical implementation in this section is carried out using the embedded discrete-time Markov chain $Z = \{Z(j) = Q(T_j) : j \ge 0\}$. That is, given that the current state of $Z(j)$ is nx, the algorithm selects a random index I taking values in $\{0,1,2,3\}$ according to the weights $\{\rho_i^{\varepsilon,\delta}(x) : i = 0,1,2,3\}$. Then the algorithm generates $Z(j+1) = Z(j) + \pi[x, \triangle_{j+1}]$ with $\mathbb{Q}(\triangle_{j+1} = v) = \bar{r}_I(x,v)/\bar{R}_I(x)$. The corresponding importance sampling estimator is

$$\bar{p}_n \doteq 1_{\{Z_2(N)=n\}} \prod_{j=0}^{N-1} \frac{r[Z(j)/n, \triangle_{j+1}]/R(Z(j)/n)}{\sum_{i=0}^3 \rho_i^{\varepsilon,\delta}(Z(j)/n)\bar{r}_i[Z(j)/n, \triangle_{j+1}]/\bar{R}_i(Z(j)/n)},$$

where

$$N \doteq \inf\{k \ge 0 : Z_2(k) = n \text{ or } Z(k) = 0\}.$$

As noted previously (6.4) holds, which says that $\bar{p}_n$ is a conditional expectation of $\hat{p}_n$. Therefore, if $\hat{p}_n$ is asymptotically optimal, so is $\bar{p}_n$.

We present simulations for the two cases $(\lambda, \mu_1, \mu_2, \nu_1) = (0.1, 0.7, 0.2, 0.15)$ and $(0.3, 0.36, 0.34, 0.32)$, with $n = 20, 50, 100$. These cases are distinguished in that, according to [10], a variant on the standard "open loop" approach to importance sampling can be expected to be efficient for the first but not the second. Numerical results are also presented in [10] using such a scheme. A comparison of the numerical data supports the statements of [10]. Indeed,

both algorithms perform well on the data of Table 1 with the state independent scheme producing slightly smaller confidence intervals, while only the state dependent algorithm appears to be asymptotically optimal for the data of Table 2.

For each case, we let the mollification parameter take the form $\varepsilon_n \doteq c/\sqrt{n}$ for a constant c, and set $\delta_n \doteq -\varepsilon_n \log \varepsilon_n$ as suggested by Remark 9.4. Asymptotic optimality of $\bar{p}_n$ follows from Corollary 9.3 and the discussion in the preceding paragraph.

The constant c is determined so that when $n = 20$, $\delta_n \approx 0.05\gamma$, where γ is the corresponding large deviation rate [see (8.5)]. It follows that the value of the piecewise affine subsolution $\bar{W}^{\delta_n}$ at the origin is $\bar{W}^{\delta_n}(0,0) = 2\gamma - 2\delta_n \approx 0.95 \cdot 2\gamma$ when $n = 20$. We wish to point out that the algorithm is fairly robust in that different small values of c yield similar simulation results. As in [10], a sample size of one million is used for each estimate, and the slowdown threshold $\theta = 0.8$.

The theoretical values were obtained by finding that for the related question of what is the probability that either queue 1 has n customers or queue 2 has m customers before emptying. The theoretical value for the finite state space problem can be found by a first step analysis, and then solving the resulting numerical system. This problem is then solved for increasing values of m, until the overflow probability given is constant for at least 4 significant figures.

	$n = 20$	$n = 50$	$n = 100$
Theoretical value	3.80×10^{-7}	1.27×10^{-16}	3.55×10^{-32}
Estimate	3.82×10^{-7}	1.27×10^{-16}	3.53×10^{-32}
Std. Err.	0.01×10^{-7}	0.01×10^{-16}	0.07×10^{-32}
95% C.I.	$[3.80, 3.84] \times 10^{-7}$	$[1.25, 1.29] \times 10^{-16}$	$[3.36, 3.67] \times 10^{-32}$

Table 1. $(\lambda, \mu_1, \mu_2, \nu_1) = (0.1, 0.7, 0.2, 0.15)$, $c = 0.03$. Sample size 1 million.

	$n = 20$	$n = 50$	$n = 100$
Theoretical value	5.63×10^{-2}	1.19×10^{-3}	1.63×10^{-6}
Estimate	5.62×10^{-2}	1.18×10^{-3}	1.61×10^{-6}
Std. Err.	0.03×10^{-2}	0.01×10^{-3}	0.02×10^{-6}
95% C.I.	$[5.56, 5.68] \times 10^{-2}$	$[1.16, 1.20] \times 10^{-3}$	$[1.57, 1.65] \times 10^{-6}$

Table 2. $(\lambda, \mu_1, \mu_2, \nu_1) = (0.3, 0.36, 0.34, 0.32)$, $c = 0.004$. Sample size 1 million.

A Appendix. Proof of Lemma 8.1

We first show the existence and uniqueness of z. Define $h : [1, \infty) \to \mathbb{R}$ by

$$h(y) \doteq 2\sqrt{\lambda\nu_1 y} + \frac{\mu_2}{y} - (\lambda + \nu_1 + \mu_2).$$

Let $y^* \doteq \left[\mu_2^2/(\lambda\nu_1)\right]^{1/3}$. Note that $y^* > 1$ thanks to (2.1). Since

$$h'(y) = \frac{1}{y^2}\left[y\sqrt{\lambda\nu_1 y} - \mu_2\right],$$

it follows that $h'(y) < 0$ for $1 \le y < y^*$ and $h'(y) > 0$ for $y > y^*$. Furthermore,

$$h(1) = 2\sqrt{\lambda\nu_1} - \lambda - \nu_1 = -(\sqrt{\lambda} - \sqrt{\nu_1})^2 < 0, \quad h(\infty) = \infty.$$

Therefore there exists a unique $z > 1$ such that $h(z) = 0$. Indeed, we must have

$$z > y^* = \left[\mu_2^2/(\lambda\nu_1)\right]^{1/3} \tag{A.1}$$

and that $h(y) < 0$ if and only if $1 \le y < z$. A discussion of the above material can also be found in [8]. We now consider equation (8.2). Since $\mathbb{H}_s(\alpha) = -2H_s(-\alpha/2)$,

$$\mathbb{H}_s(\alpha) = -2\left[\lambda(e^{-\alpha_1/2} - 1) + \nu_1(e^{(\alpha_1-\alpha_2)/2} - 1) + \mu_2(e^{\alpha_2/2} - 1)\right],$$

$$\frac{\partial \mathbb{H}_s}{\partial \alpha_1}(\alpha) = \lambda e^{-\alpha_1/2} - \nu_1 e^{(\alpha_1-\alpha_2)/2}.$$

Simple algebra yields that equation (8.2) amounts to

$$h(e^{-\alpha_2/2}) = 0, \qquad \alpha_1 = -\log(\nu_1 e^{-\alpha_2/2}/\lambda).$$

Equation (8.4) follows immediately.

Finally, $\mathbb{H}_s(\alpha^{[1]}) = 0$ is just simple calculation, and thus omitted. It remains to show $\alpha_2^{[0]} < \alpha_1^{[1]} < \alpha_1^{[0]}$, from which $\langle \alpha^{[0]}, d_1\rangle \ge 0$ follows immediately since $d_1 = (1, -1)$. Plugging the formulae for $\alpha^{[0]}$ and $\alpha^{[1]}$, the inequalities reduce to

$$\frac{\mu_2}{\lambda} < z < \frac{\mu_2^2}{\lambda\nu_1}.$$

But $h(\mu_2/\lambda) = -(\sqrt{\mu_2} - \sqrt{\nu_1})^2 < 0$ and

$$h\left(\frac{\mu_2^2}{\lambda\nu_1}\right) = \frac{1}{\mu_2}(\mu_2 - \nu_1)(\mu_2 - \lambda) > 0.$$

This completes the proof. ■

B Appendix. Exponential bounds on T_N

We recall the definition of N from (6.1). In the proof of Theorem 9.1 one needs to establish an exponential bound on the exit time T_N, namely, there exists a strictly positive constance c such that

$$\limsup_n \frac{1}{n} \log E_0[\exp\{cT_N\}] < \infty. \tag{B.1}$$

The proof in [4, Proposition A.1], which deals with a similar problem for the tandem queue networks, is not applicable here since it assumes that the domain is compact. However, this compactness assumption is not necessary and the goal of this appendix is to show a slightly stronger result, namely, there exists a strictly positive constant c such that

$$E_0[\exp\{cT_{N_0}\}] < \infty, \tag{B.2}$$

where T_{N_0} is the first time the process returns to origin, that is,

$$N_0 \doteq \inf\{k \geq 1 : Q(T_k) = 0\}.$$

Note that N_0 is independent of n and $T_N \leq T_{N_0}$. Therefore the exponential bound (B.1) is clearly implied by (B.2).

In order to show (B.2), we first observe the following. Consider a standard tandem queue network with no server slowdown and with arrival rate λ and consecutive service rates ν_1 and μ_2. For this network, similarly to T_{N_0}, define τ_0 the first time the state process return to origin. It is not difficult to show, by a pathwise argument, that τ_0 *stochastically dominates* T_{N_0}, that is,

$$P(\tau_0 \geq t) \geq P(T_{N_0} \geq t) \tag{B.3}$$

for all $t \geq 0$. Indeed, consider a Poisson arrival process with arrival rate λ, and suppose that the i-th arrival is associated with a random vector $U_i \doteq (U_i^{(1)}, U_i^{(2)}, W_i^{(1)})$. The vectors $\{U_i\}$ are iid, and $U_i^{(1)}$, $U_i^{(2)}$, and $W_i^{(1)}$ are all exponentially distributed with respective rates ν_1, μ_2, and μ_1. Furthermore, $U_i^{(1)} \geq W_i^{(1)}$ and $U_i^{(2)}$ is independent of $(U_i^{(1)}, W_i^{(1)})$. Such vectors always exist since exponential distribution with rate ν_1 stochastically dominates exponential distribution with rate μ_1. Consider the following two scenarios: (1) The i-th arrival always uses $U_i^{(1)}$ as the service time at node 1; (2) The i-the arrival uses $U_i^{(1)}$ as the service time at node 1 if the second queue is larger than θn and uses $W_i^{(1)}$ otherwise. In both cases, the service time for the i-th arrival at node 2 is assumed to be $U_i^{(2)}$. Clearly, the first scenario

yields the standard tandem network with arrival λ and service rates ν_1 and μ_2, while the second scenario yields the tandem queue with server slowdown as studied in this paper. Since $U_i^{(1)} \geq W_i^{(1)}$ for every i, it is clear that the return time to the origin for the first scenario is pathwise bounded from below by that of the second scenario. This implies the stochastic dominance (B.3).

Thanks to the stochastic dominance (B.3), it suffices to show that there exists a strictly positive constant c such that $E_0[\exp\{c\tau_0\}] < \infty$. To this end, we consider the discrete time embedded Markov chain of the standard tandem network and (abusing notation) let $\{Z(k) \in \mathbb{Z}_+^2, k = 0, 1, \ldots\}$ denote the queue lengths at the transition epochs of the network. We also, without loss of generality, assume $\lambda + \nu_1 + \mu_2 = 1$. Let $\{Y(k)\}$ be an iid sequence of random vectors taking values in $\mathbb{V}$ with

$$P(Y(k) = v_1) = \lambda, \quad P(Y(k) = v_2) = \nu_1, \quad P(Y(k) = v_3) = \mu_2.$$

Then the dynamics of Z can be represented by the evolution

$$Z(k+1) = Z(k) + \pi[Z(k), Y(k+1)].$$

Define

$$\sigma_0 \doteq \inf\{k \geq 1 : Z(k) = 0\}.$$

Note that τ_0 can be written as

$$\tau_0 = \sum_{i=1}^{\sigma_0} s_i$$

where $\{s_i\}$ denote the time elapse between transition epochs in the continuous time model, and $\{s_i\}$ are iid exponential random variables with rate $\lambda + \nu_1 + \mu_2$, independent of σ_0. Therefore, it suffices to show that there exists a strictly positive constant c such that

$$E_0[\exp\{c\sigma_0\}] < \infty.$$

Indeed, we have a stronger result [9].

Lemma B.1. *There exist a constant $c > 0$ such that $E_z[\exp\{c\sigma_0\}]$ is finite for all $z \in \mathbb{Z}_+^2$.*

Proof. Define the stopping time $\hat{\sigma}_0 \doteq \inf\{k \geq 0 : Z(k) = 0\}$. Note that $\sigma_0 = \hat{\sigma}_0$ as long as $Z(0) \neq 0$ and $\hat{\sigma}_0 = 0$ if $Z(0) = 0$. Define $h(z) \doteq E_z[\hat{\sigma}_0]$

for each $z \in \mathbb{Z}_+^d$. Since the network is positive recurrent by assumption $\lambda < \nu_1 \leq \mu_2$, $h(z)$ is finite for every z. Define the process

$$S(k) \doteq \begin{cases} h(Z(k)) & \text{if } k \leq \hat{\sigma}_0, \\ \hat{\sigma}_0 - k & \text{if } k > \hat{\sigma}_0. \end{cases}$$

There are two key properties of S [4, Lemmas A.2, A.4, and A.5].

(i) Let $\{\mathcal{F}_k = \sigma(Z(0), Y(1), \ldots, Y(k))\}$ be the filtration. Then

$$E_z[S(k+1) - S(k)|\mathcal{F}_k] = -1$$

for all $z \in \mathbb{Z}_+^2$ and all $k \geq 0$.

(ii) The increments of the process S are *uniformly* bounded, say by $M < \infty$.

Note that there exists a $\varepsilon_0 > 0$ such that, for all $|x| \leq \varepsilon_0$,

$$e^x \leq 1 + x + x^2.$$

Define $\bar{\varepsilon} \doteq \min\left\{\varepsilon_0/M, 1/(2M^2)\right\}$. Then for every $k \geq 0$, thanks to property (ii) and that $\bar{\varepsilon}M \leq \varepsilon_0$,

$$\begin{aligned} e^{\bar{\varepsilon}(S(k+1)-S(k))} &\leq 1 + \bar{\varepsilon}(S(k+1) - S(k)) + \bar{\varepsilon}^2(S(k+1) - S(k))^2 \\ &\leq 1 + \bar{\varepsilon}(S(k+1) - S(k)) + \bar{\varepsilon}^2 M^2. \end{aligned}$$

But by property (i) and that $\bar{\varepsilon}^2 M^2 \leq \bar{\varepsilon}/2$, it follows that

$$E\left[e^{\bar{\varepsilon}(S(k+1)-S(k))}\middle|\mathcal{F}_k\right] \leq 1 - \bar{\varepsilon}/2.$$

This is equivalent to that the process $\{(1-\bar{\varepsilon}/2)^{-k}e^{\bar{\varepsilon}S(k)}, \mathcal{F}_k\}$ is a supermartingale. In particular, given $Z(0) = z \neq 0$, the Optional Sampling Theorem, $\hat{\sigma}_0 = \sigma_0$, and $h(0) = 0$ imply that

$$e^{\bar{\varepsilon}h(z)} = e^{\bar{\varepsilon}S(0)} \geq E_z\left[(1-\bar{\varepsilon}/2)^{-\sigma_0}\right].$$

Letting $c \doteq -\log(1-\bar{\varepsilon}/2) > 0$, we have

$$E_z\left[e^{c\sigma_0}\right] < \infty$$

for all $z \in \mathbb{Z}_+^2$ and $z \neq 0$. For $Z(0) = 0$, we only need to note that since the first jump away from 0 must be to e_1, there is $C < \infty$ such that

$$E_0\left[e^{c\sigma_0}\right] = CE_{e_1}\left[e^{c\sigma_0}\right] < \infty.$$

This completes the proof. ■

References

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

[2] P. Dupuis and H. Ishii. On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications. *Stochastics*, 35:31–62, 1991.

[3] P. Dupuis, K. Leder, and H. Wang. On the large deviations of the weighted-serve-the-longer-queue policy. *Preprint*.

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

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

[6] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Math. Oper. Res.*, 32:1–35, 2007.

[7] W. H. Fleming and R. Rishel. *Deterministic and Stochastic Optimal Control*. Springer, Berlin, 1975.

[8] D. Kroese, W. Scheinhardt, and P. Taylor. Spectral properties of the tandem jackson network, seen as quasi-birth-and-death process. *The Annals of Applied Probability*, 14:2057–2089, 2004.

[9] S.P. Meyn and D. Down. Stability of generalized Jackson networks. *Ann. Appl. Prob.*, 4:124–148, 1994.

[10] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. Efficient simulation of a tandem queue with server slow-down. *Simulation*, to appear, 2007.

[11] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis: Queues, Communication and Computing*. Chapman and Hall, New York, 1995.

[12] N.D. van Foreest, M.R.H. Mandjes, J.C.W. van Ommeren, and W.R.W. Scheinhardt. A tandem queue with server slow-down and blocking. *Stochastic Models*, 21:695–724, 2005.