Queueing Syst (2009) 62: 113–157
DOI 10.1007/s11134-009-9124-y

Importance sampling for Jackson networks

Paul Dupuis · Hui Wang

Received: 18 March 2008 / Revised: 29 May 2009 / Published online: 17 June 2009
© Springer Science+Business Media, LLC 2009

Abstract Rare event simulation in the context of queueing networks has been an active area of research for more than two decades. A commonly used technique to increase the efficiency of Monte Carlo simulation is importance sampling. However, there are few rigorous results on the design of efficient or asymptotically optimal importance sampling schemes for queueing networks. Using a recently developed game/subsolution approach, we construct simple and efficient state-dependent importance sampling schemes for simulating buffer overflows in stable open Jackson networks. The sampling distributions do not depend on the particular event of interest, and hence overflow probabilities for different events can be estimated simultaneously. A by-product of the analysis is the identification of the minimizing trajectory for the calculus of variation problem that is associated with the sample-path large deviation rate function.

Keywords Importance sampling · Jackson networks · Isaacs equations · Subsolutions · Asymptotic optimality · Rare ever simulation

Mathematics Subject Classification (2000) Primary 60K25 · 60F10 · Secondary 65C05 · 68M20

1 Introduction

Rare event simulation in the context of queueing networks has been an active area of research for more than two decades. One of the most commonly used techniques

Research of P. Dupuis 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 H. Wang supported in part by the National Science Foundation (NSF-DMS-0404806 and NSF-DMS-0706003).

P. Dupuis · H. Wang (✉)
Lefschetz Center for Dynamical Systems, Brown University, Providence, RI 02912, USA
e-mail: huiwang@cfm.brown.edu

to increase the efficiency of Monte Carlo simulation is importance sampling, which amounts to simulating the system under a different probability distribution and then correcting for any induced bias by multiplying by the likelihood ratio. The literature on importance sampling for queueing networks is large; for a quick introduction and survey, see [18, 20].

Despite the considerable effort put into this problem, little is known regarding the design of efficient or asymptotically optimal importance sampling schemes for networks of queues. Most of the literature bases the construction on heuristics, and the resulting importance sampling schemes are supported only by limited numerical evidence. Such an approach is perilous in rare event simulation, since changes of measure suggested by seemingly reasonable heuristics can be very inefficient, sometimes even more so than standard Monte Carlo. For instance, for total population overflow in a simple two-node tandem Jackson network, [23] suggested a change of measure based on large deviation analysis that amounted to exchanging the arrival rate and the smallest service rate. This importance sampling scheme was later on shown to be inefficient in general [15] and in some cases will lead to an estimator with infinite variance [5]. A more subtle and serious danger of heuristic approaches is that a badly designed importance sampling scheme can be mistakenly identified as a good one after moderate (but not extensive) numerical experimentation. In particular, it may yield estimates that are far off from the true values, yet with very small *empirical* variance [7, 16].

The few exceptions in the literature that offer rigorous treatments of state-independent change of measures are only applicable to special classes of queueing networks and/or buffer overflow probabilities. For example, [1, 24] study efficient importance sampling algorithms for simulating large buildups of a single/multiple server queue; [15] shows asymptotic optimality of the change of measure proposed by [23] for total population overflow in tandem Jackson networks, but only under special assumptions on the arrival and service rates; [19] considers the buffer overflow of a single queue within a Jackson network where arrival/service rates and the routing matrix satisfy certain conditions.

It is by now clear that state-independent changes of measure will not, in general, be asymptotically optimal for networks of queues [5], and one must seek among the class of dynamic (i.e., state-dependent) importance sampling schemes in order to attain optimality. The main difficulty in the design problem is the discontinuity of the state dynamics on the boundaries where one or more queues are empty. However, under the game/subsolution framework for importance sampling [7, 8], the construction of an efficient dynamic scheme reduces to the construction of a classical *subsolution* to the associated Isaacs equation, a nonlinear partial differential equation (PDE) naturally connected to the network via large deviation asymptotics. The subsolution is required to satisfy appropriate boundary conditions that reflect the discontinuities of the state dynamics on the boundaries. The approach has already been successfully applied to tandem Jackson networks [11], two-node tandem networks with server slowdown or Markov modulated arrival/service processes [10, 26], and Jackson networks with tree topology [25].

In this paper, we consider the construction of efficient dynamic importance sampling schemes and subsolutions for general buffer overflows in stable open Jackson

networks of arbitrary dimension. These schemes are of interest on their own, and among other uses have applications in buffer sizing problems and estimation of functionals related to stationary distributions [17, 18]. The key question in the construction is to identify the appropriate gradients for the subsolutions on various boundaries of the state space. Surprisingly, it turns out that all the gradients, and hence the subsolution, can be determined by simply solving $2^d - 1$ systems of *linear* equations, where d is the dimension of the network. This greatly facilitates the implementation of the algorithm. Moreover, the resulting change of measure is *intrinsic* in the sense that it only depends on the system parameters, and is *independent* of the type of buffer overflow problem under consideration (e.g., individual buffer overflow versus total population overflow). Thus the change of measure for a particular network provides a *universal* importance sampling scheme, in that it can be used to estimate the probability of different sets, and in fact probabilities of families of events can be estimated simultaneously using a common set of simulated trajectories. One other feature of note is that while our primary interest is in obtaining an optimal decay rate for the second moment of the estimator, the schemes also have optimal decay rates for all moments of order strictly greater than one. This is discussed in Remark 4.1 and is of particular use in the analysis of the sample variance.

Given any particular event a variety of subsolutions, all of which lead to asymptotically optimal schemes, can be constructed using different approaches and motivations. The existence of universal schemes as described above is related to the fact that the negative of the well-known *quasipotential* function [13, Chap. 4.2] from large deviation theory formally defines a subsolution. We say "formally" because the negative of the quasipotential fails as a *classical sense* subsolution, in that certain boundary conditions hold only in a weak (viscosity) sense [4, 12] and must be suitably modified to produce a classical sense subsolution. For Jackson networks, the quasipotential is simply an affine function (reflecting the product form nature of the invariant distribution), and thus the main work is to identify the boundaries that require modification and identify appropriate gradients to use along these boundaries. This task is simplified by the explicit identification of the minimizing trajectories in the definition of the quasipotential (i.e., minimizers for the sample-path large deviations rate function that connect the origin to a given point). While our study of the minimizing trajectories heavily exploits the related PDE, optimal trajectories have also been explicitly identified in [21] using a very different time-reversal argument.

While the existence of a universal asymptotically optimal scheme is certainly attractive, it is not a panacea for problems of importance sampling for Jackson networks. For example, for the estimation of any particular event, the practical performance of a scheme designed specifically for that event may well be superior, since the particular parts of the boundary that require modification will depend on the configuration of the escape region. The universal scheme will likely be more complex than is necessary for this single event, since it must account for all escape regions. Furthermore, this universal scheme requires the system's initial condition to be near the stable point. On the other hand, we now have useful and provably stable schemes where there were none before, and the constructive methods developed in this paper may also find use in more specialized analyses.

We also remark that due to the universality of the proposed scheme, for a given buffer overflow problem, the sample paths generated from the scheme in general will

not follow the asymptotically most likely trajectory that leads to buffer overflow. This is not surprising since following the optimal trajectory may not necessarily lead to an optimal or even good scheme [15, 16], and optimal schemes can be constructed where the sample paths only follow the most likely path with a positive probability [14].

In this paper we restrict our study to Jackson networks. Analogous models for stochastic networks that possess a product form stationary distribution include skew-symmetric and related forms of reflecting Brownian motion and Kelly networks. Whether the current analysis can be extended to these more general systems is a topic for further research.

The paper is organized as follows. Section 2 introduces the setup. The buffer overflow probabilities of interest and their large deviations asymptotics are given in Sect. 3. Section 4 gives a brief discussion on the asymptotic optimality of importance sampling estimators. The mathematical model for system dynamics is established in Sect. 5, and the dynamic importance sampling scheme is defined in Sect. 6. The various Hamiltonians are introduced in Sect. 7. In Sect. 8 we identify the related Isaacs equation. The intuition and motivation of our construction of the subsolution is spelled out in Sect. 9. The roots of the Hamiltonians, which are essential to the construction of the subsolution, are analyzed in Sect. 10. Subsolutions and the corresponding importance sampling schemes are constructed in Sects. 11 and 12. In Sect. 13 we explicitly solve the calculus of variation problem associated with the sample-path large deviation rate function. Numerical results are presented in Sect. 14. We note that many technical proofs are postponed to the appendix in order to ease exposition.

2 Model setup

Consider a classical d-node open Jackson network with arrival rates $\lambda = (\lambda_1, \ldots, \lambda_d)'$ and service rates $\mu = (\mu_1, \ldots, \mu_d)'$. Departures from node i join node j with probability P_{ij} and leave the system with probability

$$P_{i0} \doteq 1 - \sum_{j=1}^{d} P_{ij}. \tag{2.1}$$

Define $P = [P_{ij}]_{1 \le i,j \le d}$, which is a substochastic matrix. The state process is denoted by $\{Q(t) = (Q_1(t), \ldots, Q_d(t))' : t \ge 0\}$, where $Q_i(t)$ is the queue length at node i and at time t. The process Q is *constrained* since the queue lengths have to be nonnegative.

Assumptions throughout the paper

[a] For each i, either $\lambda_i > 0$ or $\lambda_{j_1} P_{j_1 j_2} \cdots P_{j_k i} > 0$ for some $j_1, \ldots, j_k$.
[b] For each i, either $P_{i0} > 0$ or $P_{i j_1} P_{j_1 j_2} \cdots P_{j_k 0} > 0$ for some $j_1, \ldots, j_k$.
[c] The network is stable.

Assumptions [a, b, c] amount to saying that each node in the network will receive external input (possibly through other nodes), and each job will eventually leave the system. Under these assumptions, $P = [P_{ij}]_{1\le i,j\le d}$ has spectral radius less than one [2], and the *utilization parameters* $\rho = (\rho_1, \dots, \rho_d)'$ given by

$$\rho_i = \frac{[\lambda'(I-P)^{-1}]_i}{\mu_i} \tag{2.2}$$

satisfy $\rho_i \in (0,1)$ for all $i = 1, \dots, d$.

3 Buffer overflow probabilities

Consider a set $\Gamma \subset \mathbb{R}_+^d$ such that $0 \notin \bar{\Gamma}$, where $\bar{\Gamma}$ denotes the closure of Γ. The probability of interest is

$$\begin{aligned} p_n \doteq \mathbb{P}\{&\text{the state process } Q \text{ reaches } n\Gamma \text{ before coming} \\ &\text{back to 0, starting from 0}\} \end{aligned} \tag{3.1}$$

for large n. Under the stability assumption on the network, p_n is the probability of a rare event. The goal of the current paper is to design a simple and asymptotically optimal importance sampling algorithm for the Monte Carlo estimation of p_n.

The following result, which characterizes the exponential decay rate of $\{p_n\}$, will be useful. Its proof is analogous to that of [15, Theorem 2.3] and thus omitted. Let Γ° denote the interior of Γ.

Proposition 3.1 *We have*

$$-\inf_{x\in\Gamma^\circ}\langle -r^*, x\rangle \le \liminf_n \frac{1}{n}\log p_n \le \limsup_n \frac{1}{n}\log p_n \le -\inf_{x\in\Gamma}\langle -r^*, x\rangle,$$

where

$$r^* \doteq (\log\rho_1, \dots, \log\rho_d)'. \tag{3.2}$$

Assumptions throughout the paper (continued)

[d]

$$\gamma \doteq \inf_{x\in\Gamma^\circ}\langle -r^*, x\rangle = \inf_{x\in\Gamma}\langle -r^*, x\rangle. \tag{3.3}$$

It follows from Proposition 3.1 that γ is the exponential decay rate of p_n, or

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

4 Basics of importance sampling

In this section, we give a brief discussion on the asymptotic optimality of importance sampling algorithms. Consider a family of events $\{A_n\}$ with

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

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

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

where $d\mathbb{P}/d\mathbb{Q}_n$ 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 of $\hat{p}_n$. Since the estimate is unbiased, minimizing the variance is equivalent to minimizing the second moment, and the smaller the second moment, the faster the convergence. However, by Jensen's inequality

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

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

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

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

Remark 4.1 It is sometimes of interest to study other moments of $\hat{p}_n$. One reason is that when an estimate is obtained, its standard error will be estimated from the empirical data. It would be nice to know if the empirical standard error is close to the theoretical standard error. This clearly connects with the fourth moments of $\hat{p}_n$. In general, when the bth moment ($b > 1$) of $\hat{p}_n$ is of interest, we have the analogous inequality

$$\limsup_n -\frac{1}{n}\log E^{\mathbb{Q}_n}\left[\hat{p}_n^b\right] \le \limsup_n -\frac{b}{n}\log E^{\mathbb{Q}_n}[\hat{p}_n] = b\gamma$$

and say that $\hat{p}_n$ is *asymptotically optimal in the bth moment* if

$$\liminf_n -\frac{1}{n}\log E^{\mathbb{Q}_n}\left[\hat{p}_n^b\right] \ge b\gamma.$$

The classical definition of asymptotical optimality is just the special case with $b = 2$. As we will see, the importance sampling estimators we construct will be asymptotically optimal in the bth moment for all $b > 1$.

5 The system dynamics

To ease exposition, we consider the embedded discrete-time Markov chain $Z = \{Z(k) : k \geq 0\}$, defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, that represents the queue lengths at the transition epochs of the Jackson network, and assume without loss of generality that

$$\sum_{i=1}^{d} [\lambda_i + \mu_i] = 1. \tag{5.1}$$

We introduce the following notation. Denote by e_i the unit vector with the ith component one and zero elsewhere, and let

$$\mathbb{V} \doteq \{e_i, -e_i + e_j, -e_i : i, j = 1, 2, \ldots, d\}.$$

Note that e_i represents an arrival at node i, $-e_i + e_j$ a departure from node i that joins node j, and $-e_i$ a departure from node i that leaves the system.

Convention We distinguish between $-e_i + e_i$ (a service at station i that immediately returns to i) and $-e_j + e_j$ when $i \neq j$. We also distinguish them from 0, which by our convention is not an element of $\mathbb{V}$.

Since the state process Z is constrained in the nonnegative orthant by nature, its dynamics are discontinuous on the boundaries where one or more queues are empty. These boundaries are indexed by subsets $E \subset \{1, 2, \ldots, d\}$, where

$$\partial_E \doteq \left\{x \in \mathbb{R}_+^d : \ x_i = 0 \text{ for } i \in E \text{ and } x_i > 0 \text{ for } i \notin E\right\},$$

with the convention

$$\partial_\emptyset \doteq \left\{x = (x_1, \ldots, x_d) \in \mathbb{R}_+^d : x_i > 0 \text{ for all } i\right\}$$

for the interior of the state space.

For each node $i = 1, \ldots, d$, all the departures from node i are denoted by

$$\mathbb{V}[-e_i] \doteq \{v : v = -e_i \text{ or } v = -e_i + e_j \text{ for some } j = 1, \ldots, d\}. \tag{5.2}$$

Consider an arbitrary $E \subset \{1, 2, \ldots, d\}$. The nonnegativity constraint mandates that on ∂_E the state process Z cannot make physical jumps of size v if v belongs to the set $\bigcup_{i \in E} \mathbb{V}[-e_i]$. This motivates the definition of the constraining mapping π such that for $x \in \partial_E$,

$$\pi[x, v] \doteq \begin{cases} 0 & \text{if } v \in \cup_{i \in E} \mathbb{V}[-e_i], \\ v & \text{otherwise.} \end{cases} \tag{5.3}$$

The evolution of the Markov chain Z can then be modeled by the equation

$$Z(k+1) = Z(k) + \pi\big[Z(k), Y(k+1)\big], \tag{5.4}$$

where $Y = \{Y(k) : k \geq 1\}$ is a sequence of random variables taking values in $\mathbb{V}$. In other words, on ∂_E, the process Z is still allowed to make "fictitious" jumps of sizes

that are in $\bigcup_{i\in E}\mathbb{V}[-e_i]$, but it will be pushed back so that it stays in the positive orthant.

Under the original measure $\mathbb{P}$, $Y=\{Y(k):k\geq 1\}$ are independent and identically distributed (iid) with common distribution

$$\Theta[v]=\begin{cases}\lambda_i & \text{if } v=e_i,\\ \mu_i P_{ij} & \text{if } v=-e_i+e_j,\\ \mu_i P_{i0} & \text{if } v=-e_i.\end{cases} \tag{5.5}$$

In importance sampling algorithms, the distribution of Y will be altered by change of measure. We should restrict our attention to those probability distributions that are equivalent to Θ, namely,

$$\mathcal{P}^+(\mathbb{V})\doteq\left\{\theta:0\leq\theta[v]\leq 1,\ \sum_{v\in\mathbb{V}}\theta[v]=1,\ \theta[v]=0 \text{ iff } \Theta[v]=0\right\}. \tag{5.6}$$

Remark 5.1 Note that since with our convention $-e_i+e_i\neq -e_j+e_j$, $\mathbb{V}[-e_i]\cap\mathbb{V}[-e_j]=\emptyset$ whenever $i\neq j$.

6 Dynamic importance sampling schemes

Given the large deviation parameter n, it is convenient to study the scaled state process $X^n=\{X^n(k)\doteq Z(k)/n:k\geq 0\}$. Note that, by (5.3) and (5.4), X^n satisfies

$$X^n(k+1)=X^n(k)+\frac{1}{n}\pi\big[X^n(k),Y(k+1)\big]. \tag{6.1}$$

With this notation, the probability of interest p_n equals the probability of the scaled process X^n reaching Γ before coming back to the origin, after starting from the origin [i.e., $X^n(0)=0$].

Dynamic, or state-dependent, importance sampling schemes can be characterized by alternative stochastic kernels $\bar{\Theta}^n[\cdot|\cdot]$ on $\mathbb{V}$ given $\mathbb{R}^d_+$ that satisfy $\bar{\Theta}^n[\cdot|x]\in\mathcal{P}^+(V)$ for every $x\in\mathbb{R}^d_+$. More precisely, the conditional probability of $Y(k+1)=v$ given the history $\{Y(j):j=1,\ldots,k\}$, is just $\bar{\Theta}^n[v|X^n(k)]$ for every $v\in\mathbb{V}$ and every k. Define two hitting times

$$T_n\doteq\inf\big\{k\geq 1:X^n(k)\in\Gamma\big\},\qquad T_0\doteq\inf\big\{k\geq 1:X^n(k)=0\big\}. \tag{6.2}$$

Then the corresponding unbiased importance sampling estimator is

$$\hat{p}_n\doteq 1_{\{T_n<T_0\}}\prod_{k=0}^{T_n-1}\frac{\Theta[Y(k+1)]}{\bar{\Theta}^n[Y(k+1)|X^n(k)]}. \tag{6.3}$$

7 The Hamiltonians

The construction of efficient importance sampling schemes is intimately connected with differential games, Isaacs equations, and subsolutions [8]. We start by considering various Hamiltonians for the state dynamics. Knowledge about the roots of these Hamiltonians is essential for constructing good subsolutions.

Given any $x \in \mathbb{R}_+^d$ and $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^d$, the Hamiltonian at x is defined by

$$H(x, \alpha) \doteq \log\left(\sum_{v \in \mathbb{V}} \Theta[v] \cdot e^{\langle \alpha, \pi[x,v] \rangle}\right). \tag{7.1}$$

Note that $H(x, \cdot)$ is just the log-moment generating function of the increments of the state process at state x. As such, it is intimately connected with the large deviation properties of the queueing process. It is worth noting that for $x \in \partial_E$ with $E \subset \{1, \ldots, d\}$, $H(x, \alpha) = H_E(\alpha)$, where

$$H_E(\alpha) \doteq \log\left(\sum_{v \notin \cup_{i \in E} \mathbb{V}[-e_i]} \Theta[v] \cdot e^{\langle \alpha, v \rangle} + \sum_{v \in \cup_{i \in E} \mathbb{V}[-e_i]} \Theta[v]\right). \tag{7.2}$$

We refer to $H_\emptyset$ as the *interior* Hamiltonian and H_E as a *boundary* Hamiltonian when E is nonempty. A standard calculation using Hölder's inequality shows that $H(x, \cdot)$ and $H_E(\cdot)$ are convex.

Lemma 7.1 *For all subsets $E, E_1, E_2 \subset \{1, \ldots, d\}$ such that $E_1 \cap E_2 = \emptyset$ and $E_1 \cup E_2 = E$,*

$$e^{H_\emptyset} + e^{H_E} = e^{H_{E_1}} + e^{H_{E_2}}.$$

The proof is straightforward from the definition (7.2) and Remark 5.1, and therefore omitted.

Remark 7.2 To ease exposition, we will denote H_E by H_i when $E = \{i\}$.

8 The Isaacs equation

In this section we introduce the Isaacs equation that is relevant to importance sampling for Jackson networks. A heuristic motivation, which in any case would not be used in the analysis, can be found in [11] for two-dimensional networks and in [7, 8] for other problem formulations. Subsolutions to the Isaacs equation, together with a verification argument (one of the oldest techniques in optimal control), will be used to obtain bounds on the performance of related IS schemes.

With the convention $0 \log 0/0 = 0$, for each $\alpha \in \mathbb{R}^d$ and $x \in \mathbb{R}_+^d$, we define

$$\mathbb{H}(x, \alpha) = \sup_{\bar{\Theta} \in \mathcal{P}^+(\mathbb{V})} \inf_{\theta \in \mathcal{P}^+(\mathbb{V})} \left[\left\langle \alpha, \mathbb{F}(x, \theta) \right\rangle + \sum_{v \in \mathbb{V}} \theta[v] \log \frac{\bar{\Theta}[v]}{\Theta[v]} + R(\theta \| \Theta)\right],$$

where

$$\mathbb{F}(x,\theta) \doteq \sum_{v\in\mathbb{V}} \theta[v]\cdot\pi[x,v],$$

and the relative entropy

$$R(\theta\|\Theta) \doteq \sum_{v\in\mathbb{V}} \theta[v]\log\frac{\theta[v]}{\Theta[v]}.$$

$\mathbb{H}$ is the Hamiltonian associated with a differential game where the $\bar{\Theta}$-player represents the choice of change of measure of an importance sampling scheme, whereas the θ-player is introduced from the relative entropy representation formula [6, Proposition 1.4.2] and characterizes certain large deviation properties. It is intimately connected with the Hamiltonian H as defined in (7.1). Their relation is described in the following lemma, whose proof can be found in the Appendix.

Lemma 8.1 *Consider any $E\subset\{1,\dots,d\}$ and $x\in\partial_E$. For every $\alpha\in\mathbb{R}^d$,*

$$\mathbb{H}(x,\alpha) = -2H(x,-\alpha/2) = -2H_E(-\alpha/2),$$

where H and H_E are defined in (7.1) and (7.2), respectively. The min/max condition holds in the definition of $\mathbb{H}$, and the saddle point is $(\bar{\Theta}^*(x,\alpha),\theta^*(x,\alpha)) = (\Theta_E^*(\alpha),\Theta_E^*(\alpha))\in\mathcal{P}^+(\mathbb{V})$, where

$$\begin{aligned}\Theta_E^*(\alpha)[v] &\doteq e^{-H_E(-\alpha/2)}\cdot\Theta[v]\cdot e^{-\langle\alpha,\pi[x,v]\rangle/2}\\ &= e^{-H_E(-\alpha/2)}\cdot\begin{cases}\Theta[v]\cdot e^{-\langle\alpha,v\rangle/2} & \text{if } v\notin\cup_{i\in E}\mathbb{V}[-e_i],\\ \Theta[v] & \text{if } v\in\cup_{i\in E}\mathbb{V}[-e_i].\end{cases}\end{aligned}$$

In particular, for every x, $\mathbb{H}(x,\cdot)$ is concave.

For a function $W:\mathbb{R}_+^d\to\mathbb{R}$, denote by $DW(x)$ the gradient of W at x if it exists. The Isaacs equation is then

$$\mathbb{H}\big(x,DW(x)\big)=0,\quad x\in\mathbb{R}_+^d,$$

with the boundary condition

$$W(x)=0,\quad x\in\Gamma.$$

A continuously differentiable function $\bar{W}:\mathbb{R}_+^d\to\mathbb{R}$ is said to be a *classical subsolution* to this Isaacs equation if $\mathbb{H}(x,D\bar{W}(x))\geq 0$ for all $x\in\mathbb{R}_+^d$ and $\bar{W}(x)\leq 0$ for all $x\in\Gamma$. Then one importance sampling scheme corresponding to this subsolution is that which, if the system is at state nx, uses the change of measure determined by the $\bar{\Theta}$-component of the saddle point, with α replaced by $D\bar{W}(x)$. As we will see, it is also possible to use a *mixture* of saddle-point distributions when the gradient equals a convex combination of some fixed collection of vectors. This will be the case for the subsolutions we construct, and the use of mixtures will lead to schemes that are in some sense easier to implement.

9 Remarks on the construction of subsolutions

The key quantities in the construction of subsolutions are the appropriate roots of the various Hamiltonians H_E where $E \subset \{1, \ldots, d\}$. This section discusses the motivation and intuition on how these roots are determined. It can be safely skipped by readers interested only in algorithmic aspects of the construction. The discussion here is informal and assumes familiarity with the theory of large deviations.

It should be noted first that the Isaacs equation is closely related to the PDE associated with the corresponding large deviation problem, which takes the form

$$-H\big(x, -DU(x)\big) = 0, \quad U(x) = 0 \text{ for } x \in \partial\Gamma. \tag{9.1}$$

By Lemma 8.1, it suffices to construct a subsolution, say $\bar{U}$, to the large deviation PDE (9.1), and $\bar{W} \doteq 2\bar{U}$ will then be a subsolution to the Isaacs equation.

The central question in the construction of $\bar{U}$ is to determine the appropriate gradients in different regions of the state space. There are two considerations in the determination of these gradients: (a) the subsolution properties should be satisfied everywhere; (b) along the optimal path (i.e., least-cost path), the subsolution should satisfy the PDE with *equality*. The latter is required so that $\bar{U}(0) = \gamma$ or $\bar{W}(0) = 2\gamma$, which is indeed necessary for the optimality of the corresponding importance sampling estimator since $\bar{W}(0)$ will characterize the lower bound on the exponential decay rate of its second moment.

Here it is useful to recall that the goal in this paper is to establish importance sampling schemes that do not depend on the particular hitting set Γ. The distinction is crucial, in that it leads to a very different approach to the construction of subsolutions. When the hitting set is fixed, it is natural to think of the subsolution as solving a relaxed form of (9.1). Suppose that ϕ is the trajectory that minimizes the large deviation rate function over all paths that connect x to Γ and that ϕ first touches Γ at y. Then one can relax the condition $U(x) = 0$ for $x \in \partial\Gamma$ to $\bar{U}(x) \le 0$ on $\partial\Gamma$, so long as one keeps $\bar{U}(0) = U(0)$. It is often the case, especially for queueing type problems and for sets Γ with a simple shape, that there are simple functions which satisfy the relaxed problem, even though $U(x)$ itself has no explicit representation. See the examples in [11] and Sect. 14.

On the other hand, a natural starting point for schemes that are not directly tied to the hitting set Γ is the *quasipotential function* $V(x)$ from large deviations theory (see Sect. 13). The quasipotential is defined as the minimum of the large deviation rate over all trajectories that connect 0 to x and is known to characterize the large deviation properties of invariant distributions [13, Chap. 6]. Observe that with the quasipotential the free variable is the terminal location rather than the initial location. Hence it does not have a direct interpretation as the value function of an optimal control problem. When the quasipotential is smooth and in the setting of processes with smooth statistics, one can show that V satisfies $-H(x, DV(x)) = 0$. Under these circumstances, for *any* set Γ, a subsolution can be found by setting $a = \inf\{V(x) : x \in \Gamma\}$ and then choosing $\bar{U}(x) = -V(x) + a$. Difficulties associated with this approach are that the quasipotential is not usually smooth, nor is it usually available in any explicit form.

For Jackson networks, the invariant distribution, and hence the quasipotential, are known in closed form. In fact, one has $V(x) = \langle -r^*, x\rangle$, where r^* is defined as in Proposition 3.1. Although V is obviously smooth, because of the discontinuous statistics, it does not satisfy the subsolution property on all the boundaries and must be modified there (although it is a subsolution everywhere on the interior). Hence there are two questions: how do we identify the parts of the boundary where an alternative gradient is needed, and for those parts, what is a good gradient?

For points x such that an alternative gradient is required, all boundary behaviors must be accounted for by the subsolution. Thus if we consider a gradient r for use in a neighborhood of ∂_E, then we would like r to satisfy the equation $-H_F(-r) = 0$ for all $F \subset E$. By Lemma 7.1, this amounts to

$$-H_\emptyset(-r) = 0, \qquad -H_i(-r) = 0 \quad \text{for all } i \in E. \tag{9.2}$$

Moreover, if the optimal path goes from point x to point y on ∂_E, then the cost to travel from x to y should satisfy

$$\left\langle -r^*, x - y\right\rangle = \langle -r, x - y\rangle.$$

In other words, we would like r to be a *projection* of r^* such that $r_i = r_i^*$ for all $i \notin E$ so that the above display holds automatically. This leads to the equation

$$r = r^* - \sum_{i\in E} u_i e_i,$$

where $\{u_i : i \in E\}$ are unknown real numbers. There are in total $|E|$ unknowns and $|E| + 1$ equations (9.2). But Lemma 10.2 and Proposition 10.5 show that these $|E|$ unknowns are uniquely determined by these $|E| + 1$ equations, with $u_i = \log z_i^E$.

The remaining issue is to determine where alternative gradients are needed. This is related to how well the quasipotential approximates the invariant distribution at the *prelimit*. Consider a point $x \in \partial_E$, and consider the minimizing trajectory which connects the origin to x in the definition of the quasipotential. If the first time this trajectory reaches ∂_E after leaving the origin is when it hits x, then in some sense dynamics other than the interior dynamics do not contribute substantially to the behavior of the invariant distribution near x, and the subsolution property will hold without modification. On the other hand, if the trajectory runs along ∂_E to reach x, then we must distinguish between whether the trajectory in some sense "pushes into" ∂_E or not. This distinction is again quantified by the z_i^E introduced in Lemma 10.2, and in fact the trajectory pushes into ∂_E if and only if $z_i^E < 1$ for all $i \in E$. This will be shown in Sect. 13, where it is more convenient to first construct the time-reversed trajectories. It is only for the parts of the boundary where the condition $z_i^E < 1$ for all $i \in E$ holds that the prelimit invariant distribution reflects the different statistical behaviors of the boundary dynamics, and for which the quasipotential must be modified to produce a subsolution with respect to the boundary dynamics. This motivates the definition of the collection of important roots of the Hamiltonians in (10.8), which are essential to the construction of subsolutions.

10 The roots of Hamiltonians

As we will see, subsolutions for Jackson networks can be constructed from a small number of key roots to the various Hamiltonians. These roots only depend on the arrival/service rates and the matrix P. They are *independent* of the set Γ. A detailed discussion on the motivation and intuition behind these roots is presented in Sect. 9.

Later in this section we will need the notation

$$\mathbb{V}[e_i] \doteq \{v \in \mathbb{V} : v = e_i, \text{ or } v = -e_j + e_i \text{ for some } j = 1, \ldots, d\},$$

which denotes all the arrivals to node i. This should be distinguished from the previous notation $\mathbb{V}[-e_i]$ as defined by (5.2), which denotes all the departures from node i.

We will also need two well-known identities; see, e.g., [2, Chap. IV.2]. The first identity says that the arrival rate equals the output rate at node j in equilibrium, while the second identity amounts to saying that the total arrival rate to the network equals the total output rate of the network.

Lemma 10.1 *For each $j = 1, \ldots, d$, we have*

$$\lambda_j + \sum_{i=1}^{d} \mu_i \rho_i P_{ij} = \mu_j \rho_j. \tag{10.1}$$

Furthermore,

$$\sum_{j=1}^{d} \lambda_j = \sum_{j=1}^{d} \mu_j \rho_j P_{j0}. \tag{10.2}$$

Recall the definition of r^* as in (3.2). Fix an arbitrary nonempty subset $E \subset \{1, 2, \ldots, d\}$ and consider the system of linear equations of $\{z_i : i \in E\}$

$$\sum_{\substack{v \in \mathbb{V}[-e_i]: \\ v \neq -e_i + e_j \\ \text{any } j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} + \sum_{\substack{v \in \mathbb{V}[-e_i]: \\ v = -e_i + e_j \\ \text{some } j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} z_j = \mu_i z_i. \tag{10.3}$$

Later on we will define a vector r^E in terms of the solution to this system. The form of (10.3) is dictated by the requirement $H_i(-r^E) = 0$ for $i \in E$. See also the discussion in Sect. 9.

Lemma 10.2 *For every nonempty subset $E \subset \{1, \ldots, d\}$, the system of linear equations (10.3) has a unique solution $z^E = \{z_i^E : i \in E\}$ with strictly positive components.*

Proof For $i \in E$, let $u_i \doteq \mu_i z_i$. Then the system of equations (10.3) can be rewritten as

$$c + Bu = u, \tag{10.4}$$

where

$$u = (u_i)'_{i\in E}, \qquad c = (c_i)'_{i\in E}, \qquad B = [B_{ij}]_{i,j\in E},$$

$$c_i \doteq \sum_{\substack{v\in\mathbb{V}[-e_i]\\ v\neq -e_i+e_j\\ \text{any } j\in E}} \Theta[v]e^{-\langle r^*,v\rangle}, \qquad B_{ij} \doteq \Theta[v]e^{-\langle r^*,v\rangle}/\mu_j\big|_{v=-e_i+e_j}.$$

We first argue that B' is a substochastic matrix. Recalling that $r_i^* = \log\rho_i$ and the definition of $\Theta[v]$, we observe that identity (10.1) is equivalent to the identity

$$-\mu_j + \sum_{v\in\mathbb{V}[e_j]} \Theta[v]e^{-\langle r^*,v\rangle} = 0, \tag{10.5}$$

and therefore for every $j \in E$,

$$\sum_{i\in E} B_{ij} = \frac{1}{\mu_j} \sum_{\substack{v=-e_i+e_j\\ \text{some } i\in E}} \Theta[v]e^{-\langle r^*,v\rangle} \le \frac{1}{\mu_j} \sum_{v\in\mathbb{V}[e_j]} \Theta[v]e^{-\langle r^*,v\rangle} = 1. \tag{10.6}$$

Now we argue that B' has spectral radius less than one. Define

$$E_1 \doteq \left\{ j \in E : \sum_{i\in E} B_{ij} = 1 \right\}, \qquad E_2 \doteq E \setminus E_1.$$

For every $j \in E_1$, inequality (10.6) is indeed equality, and therefore $\Theta[e_j] = \lambda_j = 0$ and $\Theta[-e_i + e_j] = \mu_i P_{ij} = 0$ or $P_{ij} = 0$ for all $i \notin E$. Now fix an arbitrary $j \in E_1$. By Assumption [a] and since $\lambda_j = 0$, there must exist some $j_1, \ldots, j_k$ such that $\lambda_{j_1} P_{j_1 j_2} \cdots P_{j_k j} > 0$. Clearly $j_1 \notin E_1$, and therefore we can define a $k^* \in \{1, \ldots, k\}$ such that $\{j_{k^*+1}, \ldots, j_k, j\} \subset E_1$, while $j_{k^*} \notin E_1$. Since $P_{i j_{k^*+1}} = 0$ [with $j_{k^*+1} \doteq j$ if $k^* = k$] for all $i \notin E$, we have $j_{k^*} \in E$. It follows that $j_{k^*} \in E_2$. Recalling the definition of B_{ij},

$$[B']_{j j_k}[B']_{j_k j_{k-1}} \cdots [B']_{j_{k^*+1} j_{k^*}} = B_{j_{k^*} j_{k^*+1}} \cdots B_{j_{k-1} j_k} B_{j_k j} > 0.$$

Therefore the spectral radius of B' is less than one [2, Proposition I.6.3].

It follows that $(I - B)$ is invertible and $(I - B)^{-1} = I + B + B^2 + \cdots$ has nonnegative components. Therefore, $u^E \doteq (I - B)^{-1}c$ is the unique solution to the system of equations (10.4). Clearly $u^E \ge 0$. We now argue $u^E > 0$ by contradiction. Suppose this is not true, or the set $F \doteq \{i \in E : u_i^E = 0\}$ is nonempty. Then for every $i \in F$, (10.4) implies that $c_i = 0$ and $B_{ij} = 0$ for all those $j \in E \setminus F$. But $c_i = 0$ amounts to $P_{ij} = 0$ for all $j \notin E$ and $j = 0$, and $B_{ij} = 0$ amounts to $P_{ij} = 0$. Therefore, $P_{ij} = 0$ for all $i \in F$ and $j \notin F$ or $j = 0$. Thus, for $i \in F$, Assumption [b] is violated, a contradiction. Letting $z_i^E \doteq u_i^E/\mu_i$, we complete the proof. □

Corollary 10.3 *Suppose that a vector $\{z_i : i \in E\}$ satisfies (10.3) with "=" replaced by "$\ge$" [resp., "$\le$"]. Then $z_i^E \ge z_i$ [resp., "$\le$"] for every $i \in E$.*

Proof We only show for the case "$\geq$", the other direction being analogous. Let $u^E \doteq (\mu_i z_i^E)_{i\in E}$ and $u \doteq (\mu_i z_i)_{i\in E}$. Using the notation in the proof of Lemma 10.2, we have $c + Bu^E = u^E$ and $c + Bu \geq u$. Letting $w = u^E - u$, we have $w = Bw + (c + Bu - u)$. It follows readily that $w = (I - B)^{-1}(c + Bu - u) \geq 0$ since $(I - B)^{-1}$ has nonnegative components and $c + Bu - u \geq 0$. □

For every nonempty subset E, define the vector r^E by

$$r^E \doteq r^* - \sum_{i\in E} \log\left[z_i^E\right] \cdot e_i. \tag{10.7}$$

For completeness, we also define $r^{\emptyset} = r^*$.

Remark 10.4 Observe that for any i,

$$\sum_{v\in\mathbb{V}[-e_i]} \Theta[v] = \sum_{j=0}^{d} \mu_i P_{ij} = \mu_i.$$

This implies $z_i^{\{1,\ldots,d\}} = \rho_i$ for every i, which yields $r^{\{1,\ldots,d\}} = 0$.

Proposition 10.5 *For every subset $E \subset \{1, 2, \ldots, d\}$, the vector r^E satisfies*

$$H_F\left(-r^E\right) = 0$$

for all $F \subset E$.

This is in fact the key result for the construction of good subsolutions. When constructing a subsolution W, we will want the gradient to satisfy $H_E(DW(x)) \geq 0$ when $x \in \partial_E$, where DW denotes the gradient. Proposition 10.5 tells us that if $DW(x) = -r^E$ at a point $x \in \partial_E$, then at least up to a small error $H_F(DW(y)) \geq 0$ for all y in a neighborhood of x, where F is the set of indices such that $y \in \partial_F$ [note that $F \subset E$ for those y close to x].

Not every root r^E is needed in the construction of subsolutions. Define

$$\mathcal{F} \doteq \left\{F \subset \{1, \ldots, d\} : z_i^F < 1 \text{ for all } i \in F\right\}. \tag{10.8}$$

It will turn out that only the roots in $\{r^F : F \in \mathcal{F}\}$ are needed to construct subsolutions for importance sampling. This issue has been discussed in Sect. 9.

Proof of Proposition 10.5 We first consider the case of $E = \emptyset$, where $r^{\emptyset} = r^*$. By the definitions of $H_{\emptyset}$ and r^*, Lemma 10.1, and condition (5.1),

$$e^{H_{\emptyset}(-r^*)} = \sum_{j=1}^{d} \Theta[e_j] e^{-\langle r^*, e_j\rangle} + \sum_{j=1}^{d} \Theta[-e_j] e^{-\langle r^*, -e_j\rangle}$$

$$
\begin{aligned}
&+\sum_{i=1}^{d}\sum_{j=1}^{d}\Theta[-e_i+e_j]e^{-\langle r^*,-e_i+e_j\rangle}\\
&=\sum_{j=1}^{d}\frac{1}{\rho_j}\left(\lambda_j+\sum_{i=1}^{d}\mu_i P_{ij}\rho_i\right)+\sum_{j=1}^{d}\mu_j P_{j0}\rho_j\\
&=\sum_{j=1}^{d}[\mu_j+\lambda_j]\\
&=1,
\end{aligned}
$$

which yields $H_\emptyset(-r^*)=0$.

Assume from now on that E is nonempty. It suffices to show that $H_\emptyset(-r^E)=H_i(-r^E)=0$ for every $i\in E$, since Lemma 7.1 can then be inductively invoked to argue that $H_F(-r^E)=0$ for all $F\subset E$. Fix an arbitrary $i\in E$. By the definitions of $H_\emptyset$ and H_i,

$$
e^{H_\emptyset(-r^E)}-e^{H_i(-r^E)}=\sum_{v\in\mathbb{V}[-e_i]}\Theta[v]\big(e^{-\langle r^E,v\rangle}-1\big).
$$

As observed in Remark 10.4,

$$
\sum_{v\in\mathbb{V}[-e_i]}\Theta[v]=\mu_i, \tag{10.9}
$$

whereas it follows from the definition (10.7) that for any $v\in\mathbb{V}[-e_i]$,

$$
\begin{aligned}
\langle r^E,v\rangle&=\langle r^*,v\rangle-\sum_{k\in E}\log\big[z_k^E\big]\cdot\langle e_k,v\rangle\\
&=\langle r^*,v\rangle-\begin{cases}-\log z_i^E+\log z_j^E & \text{if } v=-e_i+e_j,\ j\in E,\\ -\log z_i^E & \text{otherwise.}\end{cases}
\end{aligned}
$$

Since $z^E=\{z_i^E:i\in E\}$ is a solution to the linear equations (10.3), it is now straightforward calculation to show that

$$
e^{H_\emptyset(-r^E)}-e^{H_i(-r^E)}=0,
$$

or $H_\emptyset(-r^E)=H_i(-r^E)$.

It remains to show that $H_\emptyset(-r^E)=0$. To this end, observe that the definition of $H_\emptyset$ and $H_\emptyset(-r^*)=0$ imply

$$
e^{H_\emptyset(-r^E)}-1=e^{H_\emptyset(-r^E)}-e^{H_\emptyset(-r^*)}=\sum_{v\in\mathbb{V}}\Theta[v]e^{-\langle r^*,v\rangle}\big(e^{\langle r^*-r^E,v\rangle}-1\big).
$$

For each $i\in E$, define subsets of $\mathbb{V}$ by

$$
\mathbb{V}_1^{[i]}\doteq\big\{v\in\mathbb{V}[-e_i]:v\neq-e_i+e_j,\ \text{any } j\in E\big\},
$$

$$\mathbb{V}_2^{[i]} \doteq \big\{v \in \mathbb{V}[-e_i] : v = -e_i + e_j,\ \text{some } j \in E\big\},$$
$$\mathbb{V}_3^{[i]} \doteq \{v \in \mathbb{V} : v = -e_j + e_i,\ j \notin E\},$$
$$\mathbb{V}_4^{[i]} \doteq \{v \in \mathbb{V} : v = e_i\}.$$

It is not difficult to see that $\{\mathbb{V}_k^{[i]} : i \in E, k = 1, \dots, 4\}$ are disjoint, and

$$e^{\langle r^* - r^E, v\rangle} - 1 = \begin{cases} 1/z_i^E - 1 & \text{if } v \in \mathbb{V}_1^{[i]}, \\ z_j^E/z_i^E - 1 & \text{if } v \in \mathbb{V}_2^{[i]} \text{ and } v = -e_i + e_j \text{ for some } j \in E, \\ z_i^E - 1 & \text{if } v \in \mathbb{V}_3^{[i]} \text{ or } v \in \mathbb{V}_4^{[i]}, \\ 0 & \text{otherwise.} \end{cases} \tag{10.10}$$

Therefore,

$$e^{H_\emptyset(-r^E)} - 1 = \sum_{i \in E} \big(d_1^{[i]} + d_2^{[i]} + d_3^{[i]} + d_4^{[i]}\big),$$

where for every $k = 1, \dots, 4$,

$$d_k^{[i]} \doteq \sum_{v \in \mathbb{V}_k^{[i]}} \Theta[v] e^{-\langle r^*, v\rangle} \big(e^{\langle r^* - r^E, v\rangle} - 1\big).$$

Furthermore, thanks to (10.10) again, for every $i \in E$,

$$d_1^{[i]} + d_2^{[i]} = \sum_{v \in \mathbb{V}_1^{[i]}} \Theta[v] e^{-\langle r^*, v\rangle} \big(1/z_i^E - 1\big) + \sum_{\substack{v = -e_i + e_j \\ \text{some } j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} \big(z_j^E/z_i^E - 1\big).$$

From (10.3) we obtain an alternative expression for the first sum, which gives

$$d_1^{[i]} + d_2^{[i]} = \mu_i \big(1 - z_i^E\big) + \sum_{\substack{v = -e_i + e_j \\ \text{some } j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} \big(z_j^E - 1\big).$$

Therefore,

$$\sum_{i \in E} \big(d_1^{[i]} + d_2^{[i]}\big) = \sum_{i \in E} \mu_i \big(1 - z_i^E\big) + \sum_{i \in E} \sum_{\substack{v = -e_i + e_j \\ j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} \big(z_j^E - 1\big).$$

For the double sum in the last display, one can use symmetry to interchange the roles of i and j and rewrite it as

$$\sum_{i \in E} \sum_{\substack{v = -e_j + e_i \\ j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} \big(z_i^E - 1\big).$$

Also note that

$$\sum_{i \in E} \big(d_3^{[i]} + d_4^{[i]}\big) = \sum_{i \in E} \sum_{\substack{v = -e_j + e_i \\ j \notin E}} \Theta[v] e^{-\langle r^*, v\rangle} \big(z_i^E - 1\big) + \sum_{\substack{i \in E \\ v = e_i}} \Theta[v] e^{-\langle r^*, v\rangle} \big(z_i^E - 1\big).$$

Therefore,

$$\sum_{i\in E}\left(d_1^{[i]}+d_2^{[i]}+d_3^{[i]}+d_4^{[i]}\right)=\sum_{i\in E}\left(z_i^E-1\right)\left[-\mu_i+\sum_{v\in\mathbb{V}[e_i]}\Theta[v]e^{-\langle r^*,v\rangle}\right].$$

Invoking (10.5), we complete the proof. □

Recall the definition of $\mathcal{F}$ in (10.8). The proof of the following result is deferred to the Appendix.

Lemma 10.6 *The following properties hold.*

1. $\emptyset\in\mathcal{F}$, $\{1,\ldots,d\}\in\mathcal{F}$.
2. *If* $E_1,E_2\in\mathcal{F}$, *then* $E_1\cup E_2\in\mathcal{F}$.
3. *If* $E,F\in\mathcal{F}$ *and* $E\subset F$, *then* $r^E\le r^F$.

Here we adopt the notation that $u\le v$ (or $\ge$, $<$, $>$) when u and v are vectors if the inequality holds component-wise. Taking $F=\{1,\ldots,d\}$ in Part 3 of this lemma, we have the following result.

Corollary 10.7 *If* $E\in\mathcal{F}$, *then* $z_i^E\in[\rho_i,1)$ *for all* $i\in E$.

11 Construction of classical subsolutions

Recall that a classical subsolution is a continuously differentiable function $W:\mathbb{R}_+^d\to\mathbb{R}$ such that $\mathbb{H}(x,DW(x))\ge 0$ for all $x\in\mathbb{R}_+^d$ and $W(x)\le 0$ for all $x\in\Gamma$. As was remarked at the beginning of Sect. 8, these functions can be used to bound the performance of related IS schemes. $W(0)$ is the key measure of performance, with large values indicating greater variance reduction. Hence methods for the systematic construction of subsolutions with optimal or nearly optimal $W(0)$ are very useful. Since the state dynamics are piecewise homogeneous, it is natural to start with the construction of piecewise affine subsolutions and then use an exponential weighting modification to obtain a classical subsolution [8, 10, 11].

11.1 Piecewise affine subsolution

A concave, piecewise affine function $\bar{W}:\mathbb{R}_+^d\to\mathbb{R}$ is said to be a *subsolution* to the Isaacs equation if $\mathbb{H}(x,D\bar{W}(x))\ge 0$ for all x where $D\bar{W}(x)$ is well defined and $\bar{W}(x)\le 0$ for all $x\in\Gamma$. It is often identified as the minimum of a finite collection of affine functions.

The idea is to start with $-\langle r^*,x\rangle$, which is a solution on all of the interior of $\mathbb{R}_+^d$, and then perturb this function slightly near the boundaries. These perturbations are characterized by a collection of nonnegative constants $\{C_F:F\in\mathcal{F}\}$ that satisfy

Condition 11.1 $C_\emptyset = 0$ and for any nonempty $F \in \mathcal{F}$,

$$C_F > \max\left\{\frac{\log(1/z_i^F)}{\log(1/z_i^G)} \cdot C_G : i \in G \subset F, G \neq F, G \in \mathcal{F}\right\}$$

with the convention that $\max\{\emptyset\} = 0$.

There are infinitely many such collections of constants. Indeed, for all those $F \in \mathcal{F}$ that are singletons, C_F can be defined as any positive number. Once those C_F are fixed, one can recursively define C_F for those F with $|F| = 2$, and so on. Moreover, it is not difficult to see that $C_F > 0$ for all nonempty $F \in \mathcal{F}$ and by Lemma 10.6 $C_G < C_F$ if G is a strict subset of F.

Fixing an arbitrary small positive number δ, we define the piecewise affine function $\bar{W}^\delta$ by

$$\bar{W}^\delta(x) \doteq \min_{F \in \mathcal{F}} \bar{W}_F^\delta(x), \qquad \bar{W}_F^\delta(x) \doteq 2\gamma + \langle 2r^F, x\rangle - C_F\delta. \tag{11.1}$$

Note that for all $x \in \Gamma$, by assumption (3.3)

$$\bar{W}^\delta(x) \leq \bar{W}_\emptyset^\delta(x) = 2\gamma + \langle 2r^\emptyset, x\rangle = 2\gamma + \langle 2r^*, x\rangle \leq 0. \tag{11.2}$$

Lemma 11.2 *Fix any collection of constants $\{C_F : F \in \mathcal{F}\}$ that satisfy Condition 11.1. Then there exists a constant $a > 0$ such that for any $E \subset \{1, \ldots, d\}$ and $x \in \partial_E$ and for any $\delta > 0$, if $G \in \mathcal{F}$ satisfies*

$$\bar{W}_G^\delta(x) - \bar{W}^\delta(x) < a\delta,$$

then $H_F(-r^G) \leq 0$ for all $F \subset E$. In particular, $\mathbb{H}(x, 2r^G) \geq 0$.

The proof of the lemma is deferred to the appendix. Applying this lemma when G is a minimizer in the definition of $\bar{W}^\delta(x)$, we see that the function $\bar{W}^\delta$ is indeed a piecewise affine subsolution. From now on, we will assume that $\{C_F : F \in \mathcal{F}\}$ is a collection of constants that satisfy Condition 11.1.

11.2 Modification

To obtain a smooth approximation of the piecewise affine subsolution $\bar{W}^\delta$, we use the *exponential weighting* modification as follows. Fix an arbitrary $\varepsilon > 0$. Define

$$W^{\varepsilon,\delta}(x) \doteq -\varepsilon \log \sum_{F \in \mathcal{F}} \exp\left\{-\frac{1}{\varepsilon}\bar{W}_F^\delta(x)\right\}. \tag{11.3}$$

The following lemma is standard, and we omit the proof [8].

Lemma 11.3 *For any $\varepsilon, \delta > 0$, the function $W^{\varepsilon,\delta}$ is continuously differentiable. Furthermore, for every $x \in \mathbb{R}_+^d$,*

$$-\log|\mathcal{F}| \cdot \varepsilon \leq W^{\varepsilon,\delta}(x) - \bar{W}^\delta(x) \leq 0,$$

$$DW^{\varepsilon,\delta}(x) = \sum_{F\in\mathcal{F}} \rho_F^{\varepsilon,\delta}(x)\cdot 2r^F, \qquad \rho_F^{\varepsilon,\delta}(x) \doteq \frac{\exp\{-\bar{W}_F^\delta(x)/\varepsilon\}}{\sum_{E\in\mathcal{F}}\exp\{-\bar{W}_E^\delta(x)/\varepsilon\}}.$$

In particular, for every x, $\{\rho_F^{\varepsilon,\delta}(x) : F\in\mathcal{F}\}$ defines a probability vector on $\mathcal{F}$. Moreover,

$$\rho_F^{\varepsilon,\delta}(x) = \frac{\exp\{[C_F\delta - \langle 2r^F, x\rangle]/\varepsilon\}}{\sum_{E\in\mathcal{F}}\exp\{[C_E\delta - \langle 2r^E, x\rangle]/\varepsilon\}}$$

is independent of γ and hence Γ.

Proposition 11.4 *The function $W^{\varepsilon,\delta}$ satisfies $W^{\varepsilon,\delta}(x)\le 0$ for $x\in\Gamma$. Furthermore, for all $x\in\mathbb{R}_+^d$,*

$$\mathbb{H}\big(x, DW^{\varepsilon,\delta}(x)\big) \ge \sum_{F\in\mathcal{F}} \rho_F^{\varepsilon,\delta}(x)\cdot \mathbb{H}\big(x, 2r^F\big) \ge -M\exp\{-a\delta/\varepsilon\},$$

where M and a (the constant given by Lemma 11.2) are two positive constants both independent of ε and δ.

This result establishes that $\bar{W}^{\varepsilon,\delta}$ is "approximately" a classical subsolution as long as ε/δ is small, which is sufficient for the subsequent analysis. See [11] for more discussion on exponential weighting modification.

Proof of Proposition 11.4 $W^{\varepsilon,\delta}(x)\le 0$ for $x\in\Gamma$ is trivial from Lemma 11.3 and (11.2). Consider any $x\in\mathbb{R}_+^d$ and assume $x\in\partial_E$. We will argue that, for all $F\subset E$,

$$2H_F\big(-DW^{\varepsilon,\delta}(x)/2\big) \le \sum_{G\in\mathcal{F}} \rho_G^{\varepsilon,\delta}(x)\cdot 2H_F\big(-r^G\big) \le M\exp\{-a\delta/\varepsilon\}. \tag{11.4}$$

The second part of the proposition is a special case of (11.4) with $F\doteq E$, thanks to Lemma 8.1.

The first inequality in (11.4) is trivial from the convexity of $H_F(\cdot)$. As for the second inequality, note that by the proof of Lemma 11.2, more precisely, inequality (A.6), $H_F(-r^G)\le 0$ for all those G such that $\bar{W}_G^\delta(x) - \bar{W}^\delta(x) < a\delta$. For all those G such that $\bar{W}_G^\delta(x) - \bar{W}^\delta(x)\ge a\delta$, the formula for $\rho_G^{\varepsilon,\delta}(x)$ in Lemma 11.3, together with the fact that $\bar{W}^\delta(x) = \bar{W}_E^\delta(x)$ for some $E\in\mathcal{F}$, yields that

$$\rho_G^{\varepsilon,\delta}(x) \le \exp\left\{\frac{-\bar{W}_G^\delta(x) + \bar{W}^\delta(x)}{\varepsilon}\right\} \le \exp\{-a\delta/\varepsilon\}.$$

Inequality (11.4) follows readily for some M independent of ε and δ. □

12 The importance sampling algorithm

Recall the formula for the saddle point in Lemma 8.1. One approach to the design of importance sampling schemes begins with the smooth function $W^{\varepsilon,\delta}(x)$ and then

uses the stochastic kernel defined by the saddle point $\bar{\Theta}^*(x, DW^{\varepsilon,\delta}(x))[v]$. While this would yield an algorithm with very good performance, it is more difficult to implement than a closely related mixture of saddle points.

For every $F \in \mathcal{F}$, define a stochastic kernel on $\mathbb{V}$ given $\mathbb{R}^d_+$ by the saddle point formula in Lemma 8.1

$$\bar{\Theta}_F[v|x] \doteq \bar{\Theta}^*\big(x, 2r^F\big)[v]. \tag{12.1}$$

Note that $\bar{\Theta}_F$ is piecewise constant in that $\bar{\Theta}^*(x, 2r^F) \equiv \Theta^*_E(2r^F)$ for all $x \in \partial_E$ and all $E \subset \{1, \ldots, d\}$. The change of measure we will associate with $W^{\varepsilon,\delta}$ is the mixture of $\{\bar{\Theta}_F : F \in \mathcal{F}\}$ given by

$$\bar{\Theta}^{\varepsilon,\delta}[\cdot|x] \doteq \sum_{F\in\mathcal{F}} \rho^{\varepsilon,\delta}_F(x)\bar{\Theta}_F[\cdot|x]. \tag{12.2}$$

In general, the parameters ε and δ are allowed to depend on n, denoted by ε_n and δ_n, respectively. To estimate p_n, the importance sampling algorithm uses a change of measure, say $\mathbb{Q}_n$, under which the transition kernel is

$$\bar{\Theta}^n \doteq \bar{\Theta}^{\varepsilon_n,\delta_n},$$

and the corresponding estimator $\hat{p}_n$ is defined as in (6.3). The next result establishes the asymptotic optimality of $\hat{p}_n$ in any order of moments, under suitable decay conditions on the parameters $\{\varepsilon_n, \delta_n\}$.

Theorem 12.1 *Suppose that $\delta_n \to 0$, $\varepsilon_n/\delta_n \to 0$, and $n\varepsilon_n \to \infty$. Then for any real number $b > 1$, the importance sampling estimator $\hat{p}_n$ satisfies*

$$\lim_n \frac{1}{n} \log E^{\mathbb{Q}_n}\big[\hat{p}_n^b\big] = -b\gamma.$$

In particular, $\hat{p}_n$ is asymptotically optimal.

The proof of this result, which uses ideas developed in [11], is deferred to the Appendix. It uses a verification argument, with the main difficulty due to the fact that the time interval of interest is potentially unbounded.

Remark 12.2 As we have remarked previously, a particular feature of the schemes constructed in this paper is that they are *independent* of γ and Γ, since Lemma 11.3 and Lemma 8.1 imply that $\rho^{\varepsilon,\delta}_F(x)$ and $\bar{\Theta}_F[\cdot|x]$ are independent of these quantities. In other words, for different target sets Γ, one can use the *same* state-dependent change of measure (12.2) for simulation, and indeed probabilities for distinct sets can be estimated simultaneously. This is very appealing for its simplicity and universality. However, we reiterate that with more knowledge of Γ it is sometimes possible to construct even simpler subsolutions, i.e., fewer affine pieces.

13 The optimal trajectory

A by-product of the subsolution analysis is the solution to the calculus of variation problem associated with the sample-path large deviations for Jackson networks. There are two versions, one for the discrete-time process Z, and the other for the continuous-time process Q. These two problems are closely related by a change of the time variable and are equivalent when the calculus of variations problem does not depend on time.

We wish to point out that the local rate functions $L(x,\cdot)$ and $\bar{L}(x,\cdot)$ appearing in the calculus of variation problems are identified as the inf-convolution of the rate functions for the dynamics in the neighborhood of x. This is due to the discontinuous nature of the underlying dynamics [3, 6].

Discrete time version The calculus of variation problem is

$$V(x^*) \doteq \inf\left\{\int_0^\tau L\big(\phi(t),\dot{\phi}(t)\big)\,dt : \phi(0)=0, \phi(\tau)=x^*, \phi(t)\in\mathbb{R}_+^d \text{ for all } t\right\}.$$

Here x^* is an arbitrary point in $\mathbb{R}_+^d$, and the local rate function L is defined, for all $x\in\partial_E$ and $\beta\in\mathbb{R}^d$, as

$$L(x,\beta) \doteq \inf\left\{\sum_{G\subset E}\rho_G L_G(\beta_G) : \rho_G\ge 0, \sum_{G\subset E}\rho_G=1, \sum_{G\subset E}\rho_G\beta_G=\beta\right\},$$

where L_G is the Legendre transform of the convex function H_G.

Continuous-time version For every $G\subset\{1,2,\ldots,d\}$, define the convex function $\bar{H}_G$ by

$$\bar{H}_G(\alpha) \doteq e^{H_G(\alpha)}-1 = \sum_{v\notin\cup_{i\in E}\mathbb{V}[-e_i]}\Theta[v]\cdot\big(e^{\langle\alpha,v\rangle}-1\big).$$

In other words, $\bar{H}_G$ is the Hamiltonian of the continuous-time process Q on the boundary ∂_G [9]. Furthermore, $\bar{H}_G(\alpha)=0$ (resp., "$\ge$", "$\le$") if and only if $H_G(\alpha)=0$ (resp., "$\ge$", "$\le$"). The calculus of variation problem associated with sample-path large deviations for the continuous-time process Q is

$$\bar{V}(x^*) \doteq \inf\left\{\int_0^\tau \bar{L}\big(\phi(t),\dot{\phi}(t)\big)\,dt : \phi(0)=0, \phi(\tau)=x^*, \phi(t)\in\mathbb{R}_+^d \text{ for all } t\right\}.$$

Here the local rate function $\bar{L}$ is defined, for all $x\in\partial_E$ and $\beta\in\mathbb{R}^d$, as

$$\bar{L}(x,\beta) \doteq \inf\left\{\sum_{G\subset E}\rho_G \bar{L}_G(\beta_G) : \rho_G\ge 0, \sum_{G\subset E}\rho_G=1, \sum_{G\subset E}\rho_G\beta_G=\beta\right\},$$

with $\bar{L}_G$ the Legendre transform of $\bar{H}_G$.

It will turn out that $V \equiv \bar{V}$ and the optimal trajectories take the same form for both problems. We introduce the notation

$$\bar{\mathcal{F}} \doteq \left\{F \subset \{1,\dots,d\} : z_i^F \le 1 \text{ for all } i \in F\right\}.$$

Recall that $\mathcal{F}$ has the analogous definition but with $\le$ replaced by $<$. Clearly $\mathcal{F} \subset \bar{\mathcal{F}}$. In particular, $\emptyset \in \bar{\mathcal{F}}$ and $\{1,\dots,d\} \in \bar{\mathcal{F}}$. Almost verbatim to the proof of Lemma 10.6 [we omit the details], one can argue that $F_1 \cup F_2 \in \bar{\mathcal{F}}$ if $F_1, F_2 \in \bar{\mathcal{F}}$. Therefore, for every $E \subset \{1,\dots,d\}$, we can define the *maximal* subset of E that belongs to $\bar{\mathcal{F}}$, that is,

$$\Pi[E] \doteq \bigcup_{F \subset E, F \in \bar{\mathcal{F}}} F.$$

Moreover, for any $F \in \bar{\mathcal{F}}$, define the vector

$$\beta_F^* \doteq \sum_{v \in \mathbb{V}} \Theta[v] e^{-\langle r^F, v\rangle} v - \sum_{i \in F} \left(1 - z_i^F\right) \sum_{v \in \mathbb{V}[-e_i]} \Theta[v] e^{-\langle r^F, v\rangle} v. \tag{13.1}$$

The structural properties and the cost of β_F^* are given by the following lemma, whose proof is deferred to the Appendix.

Lemma 13.1 *Fix any strict subset $E \subset \{1,\dots,d\}$ and let $F \doteq \Pi[E]$.*

1. $\langle \beta_F^*, e_i\rangle = 0$ *for all* $i \in F$, *and* $\langle -\beta_F^*, e_i\rangle > 0$ *for all* $i \in E \setminus F$.
2. *The set* $G \doteq \{j : \langle -\beta_F^*, e_j\rangle < 0\}$ *is nonempty. Furthermore, for all* $G_1 \subset G$, *we have* $F \cup G_1 \in \bar{\mathcal{F}}$.
3. *For any* $x \in \partial_F$, $L(x, \beta_F^*) = \bar{L}(x, \beta_F^*) = \langle -r^*, \beta_F^*\rangle$.

We are now ready to describe the optimal trajectory, which is done most conveniently in a time-reversed fashion. Assume that $x^* \in \partial_E$ for some strict subset $E \subset \{1,\dots,d\}$. Denote $F_0 = \Pi[E]$. The trajectory starts with constant velocity $-\beta_{F_0}^*$. Thanks to Lemma 13.1, Part 1, the trajectory immediately leaves ∂_E (unless $E = F_0$) and travels on ∂_{F_0} for a period of time until it leaves ∂_{F_0} and hits a new boundary, say ∂_{F_1}. By Lemma 13.1, Part 2, the trajectory will hit ∂_{F_1} in finite amount of time, and F_1 contains F_0 as a *strict* subset with $F_1 \in \bar{\mathcal{F}}$. The trajectory then switches to velocity $-\beta_{F_1}^*$ and will travel on ∂_{F_1} for a finite amount of time until it leaves ∂_{F_1} and hits, say ∂_{F_2}. Again, $F_2 \in \bar{\mathcal{F}}$ and contains F_1 as a strict subset. The trajectory then switches to velocity $-\beta_{F_2}^*$, and so on, until it hits the origin $\partial_{\{1,\dots,d\}} = \{0\}$. Denote this path by φ in the time-reversal setting with τ^* the hitting time to the origin. Switching to the forward-time setup, we define the trajectory

$$\phi^*(t) \doteq \varphi\left(\tau^* - t\right), \quad 0 \le t \le \tau^*.$$

Clearly $\phi^*(0) = 0$, $\phi^*(\tau^*) = x^*$, and $\phi^*(t) \in \mathbb{R}_+^d$. Moreover, ϕ^* is piecewise affine with at most d pieces.

Theorem 13.2 $V(x^*) = \bar{V}(x^*) = \langle -r^*, x^*\rangle$, *and ϕ^* is an optimal trajectory for both calculus of variation problems.*

Proof We will give the proof for the discrete-time version V. The continuous-time version $\bar{V}$ is almost verbatim and thus omitted.

Let $F^*(t)$ be such that $\phi^*(t) \in \partial_{F^*(t)}$. By the definition of ϕ^*, it is easy to see that $F^*(t) \in \bar{\mathcal{F}}$ and $\dot{\phi}^*(t) = \beta^*_{F^*(t)}$ for all $t \in [0, \tau^*)\backslash S$, where S is the finite set of times when the velocity changes. Therefore, by Part 3 of Lemma 13.1,

$$\begin{aligned}\int_0^{\tau^*} L\big(\phi^*(t), \dot{\phi}^*(t)\big)\,dt &= \int_0^{\tau^*} \big\langle -r^*, \dot{\phi}^*(t)\big\rangle dt \\ &= \big\langle -r^*, \phi^*\big(\tau^*\big) - \phi^*(0)\big\rangle \\ &= \big\langle -r^*, x^*\big\rangle.\end{aligned}$$

It remains to show that $V(x^*) \geq \langle -r^*, x^*\rangle$. This can be done using a verification argument.

Fix arbitrarily $\varepsilon, \delta > 0$ and any absolutely continuous function ϕ such that $\phi(0) = 0$, $\phi(\tau) = x^*$, and $\phi(t) \in \mathbb{R}^d_+$ for all $t \in [0, \tau]$. Abusing the notation, define $\bar{W}^\delta$, $W^{\varepsilon,\delta}$ as in (11.1) and (11.3), except that γ is replaced by $\langle -r^*, x^*\rangle$. Note that

$$W^{\varepsilon,\delta}\big(x^*\big) - W^{\varepsilon,\delta}(0) = \int_0^\tau \big\langle DW^{\varepsilon,\delta}\big(\phi(t)\big), \dot{\phi}(t)\big\rangle dt. \tag{13.2}$$

However, for any $x \in \partial_E$, $L(x, \cdot)$ is the inf-convolution of $\{L_G : G \subset E\}$, and therefore it is the Legendre transform of the convex function $\max\{H_G : G \subset E\}$ [6, Appendix D]. In particular, for any $\alpha, \beta \in \mathbb{R}^d$,

$$L(x, \beta) \geq \langle \alpha, \beta\rangle - \max_{G\subset E} H_G(\alpha).$$

For any t, denote by $E(t)$ the subset of $\{1, \ldots, d\}$ such that $\phi(t) \in \partial_{E(t)}$. The previous inequality implies that for $t \notin S$,

$$L\big(\phi(t), \dot{\phi}(t)\big) \geq \big\langle -DW^{\varepsilon,\delta}\big(\phi(t)\big)/2, \dot{\phi}(t)\big\rangle - \max_{G\subset E(t)} H_G\big(-DW^{\varepsilon,\delta}\big(\phi(t)\big)/2\big).$$

It follows from the proof of Proposition 11.4, more precisely, inequality (11.4), that

$$\max_{G\subset E(t)} H_G\big(-DW^{\varepsilon,\delta}\big(\phi(t)\big)/2\big) \leq M\exp\{-a\delta/\varepsilon\}/2. \tag{13.3}$$

Therefore,

$$L\big(\phi(t), \dot{\phi}(t)\big) \geq \big\langle -DW^{\varepsilon,\delta}\big(\phi(t)\big)/2, \dot{\phi}(t)\big\rangle - M\exp\{-a\delta/\varepsilon\}/2.$$

Recalling (13.2), we arrive at

$$W^{\varepsilon,\delta}\big(x^*\big) - W^{\varepsilon,\delta}(0) \geq -2\int_0^\tau L\big(\phi(t), \dot{\phi}(t)\big)\,dt - \tau M\exp\{-a\delta/\varepsilon\}.$$

But by Lemma 11.3

$$\begin{aligned}W^{\varepsilon,\delta}(0) &\geq \bar{W}^\delta(0) - \varepsilon\cdot\log|\mathcal{F}| \\ &= \big\langle -2r^*, x^*\big\rangle - C_{\{1,\ldots,d\}}\delta - \varepsilon\cdot\log|\mathcal{F}|\end{aligned}$$

and

$$W^{\varepsilon,\delta}(x^*) \le \bar{W}^{\delta}(x^*) \le \bar{W}^{\delta}_{\emptyset}(x^*) = 0.$$

Therefore,

$$\langle 2r^*, x^*\rangle + C_{\{1,\dots,d\}}\delta + \varepsilon \cdot \log|\mathcal{F}| \ge -2\int_0^{\tau} L\big(\phi(t), \dot{\phi}(t)\big)\,dt - \tau M \exp\{-a\delta/\varepsilon\}.$$

Letting $\delta \to 0$, $\varepsilon/\delta \to 0$, we have

$$\int_0^{\tau} L\big(\phi(t), \dot{\phi}(t)\big)\,dt \ge \langle -r^*, x^*\rangle.$$

Since ϕ is arbitrary, we complete the proof. □

14 Numerical examples

In this section, we will give a number of numerical examples for illustration. For all the examples, three types of buffer overflow probabilities, that is,

1. total population overflow: $\Gamma = \{x \in \mathbb{R}^d_+ : x_1 + \cdots + x_d \ge 1\}$,
2. individual buffer overflow: $\Gamma = \{x \in \mathbb{R}^d_+ : x_1 \ge 1 \text{ or } \dots \text{ or } x_d \ge 1\}$,
3. overflow of the first buffer: $\Gamma = \{x \in \mathbb{R}^d_+ : x_1 \ge 1\}$,

will be estimated using the *same* change of measure; see Remark 12.2. The parameters in the simulation are set up as follows. Let $C_{\emptyset} = 0$ and for each $\{i\} \in \mathcal{F}$,

$$C_{\{i\}} \doteq \log\big(1/z_i^{\{i\}}\big).$$

For all $F \in \mathcal{F}$ such that $|F| \ge 2$, we recursively define

$$C_F \doteq (1+k)\max\left\{\frac{\log(1/z_i^F)}{\log(1/z_i^G)} \cdot C_G : i \in G \subset F, G \ne F, G \in \mathcal{F}\right\}$$

for some parameter $k > 0$. Also for any δ, we let $\varepsilon \doteq -\delta/\log\delta$. The choice of δ and k may vary for different numerical examples, and fine-tuning may be necessary for better performance [more discussion is given in Sect. 14.4]. Each estimate is based on 100,000 samples. We restrict ourselves to the examples for which the theoretical values can be obtained by recursively solving a linear equation by first step analysis. Of course the theoretical value cannot be found in general. We restrict ourselves to such examples in order to establish a benchmark.

14.1 A two-node feedback network

Two-node feedback networks are of great value for the illustration of the importance sampling schemes and optimal trajectories, since everything can be explicitly computed. Consider the following Jackson network with

$$\lambda = (\lambda_1, \lambda_2), \qquad \mu = (\mu_1, \mu_2), \qquad P = \begin{bmatrix} 0 & p \\ \bar{p} & 0 \end{bmatrix},$$

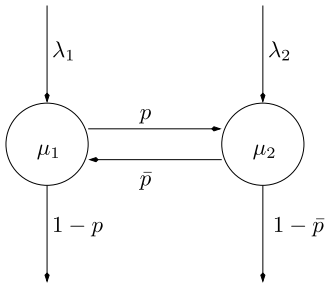

Fig. 1 Two-node feedback networks

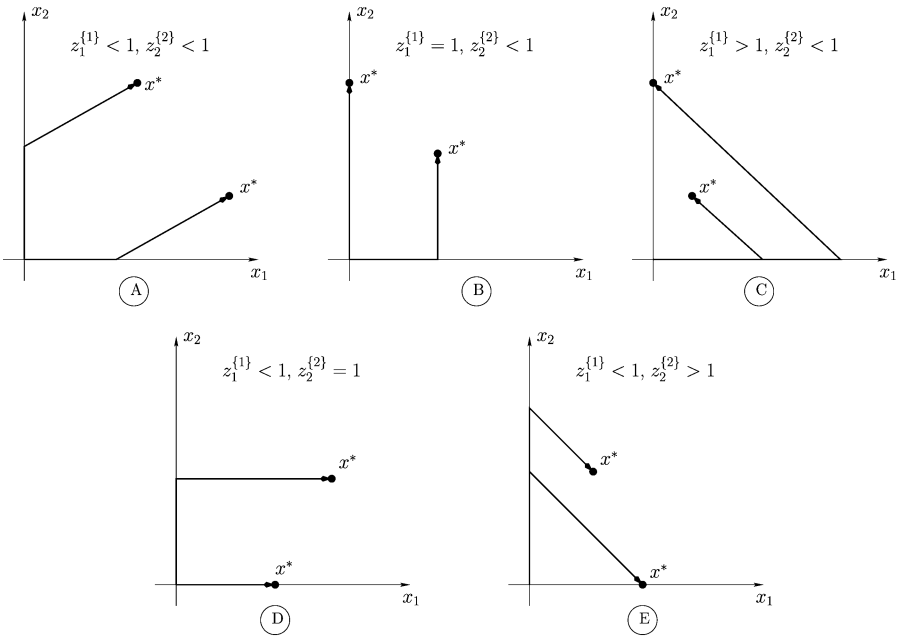

Fig. 2 Optimal trajectories

where $p\bar{p} < 1$. The utilization parameters are, by formula (2.2),

$$\rho_1 = \frac{1}{1-p\bar{p}} \cdot \frac{\lambda_1 + \bar{p}\lambda_2}{\mu_1}, \qquad \rho_2 = \frac{1}{1-p\bar{p}} \cdot \frac{\lambda_2 + p\lambda_1}{\mu_2}.$$

We assume that the system is stable, i.e., $\rho_1, \rho_2 \in (0, 1)$. See Fig. 1. Consider all possible subsets $E \subset \{1, 2\}$.

1. For $E = \emptyset$, $r^{\emptyset} = r^* = (\log \rho_1, \log \rho_2)$.
2. For $E = \{1, 2\}$, $r^{\{1,2\}} = (0, 0)$ by Remark 10.4.
3. For $E = \{1\}$, (10.3) yields $z_1^{\{1\}} = (1-p)\rho_1 + p\rho_1/\rho_2$ and $r^{\{1\}} = (\log \rho_1 - \log z_1^{\{1\}}, \log \rho_2)$.

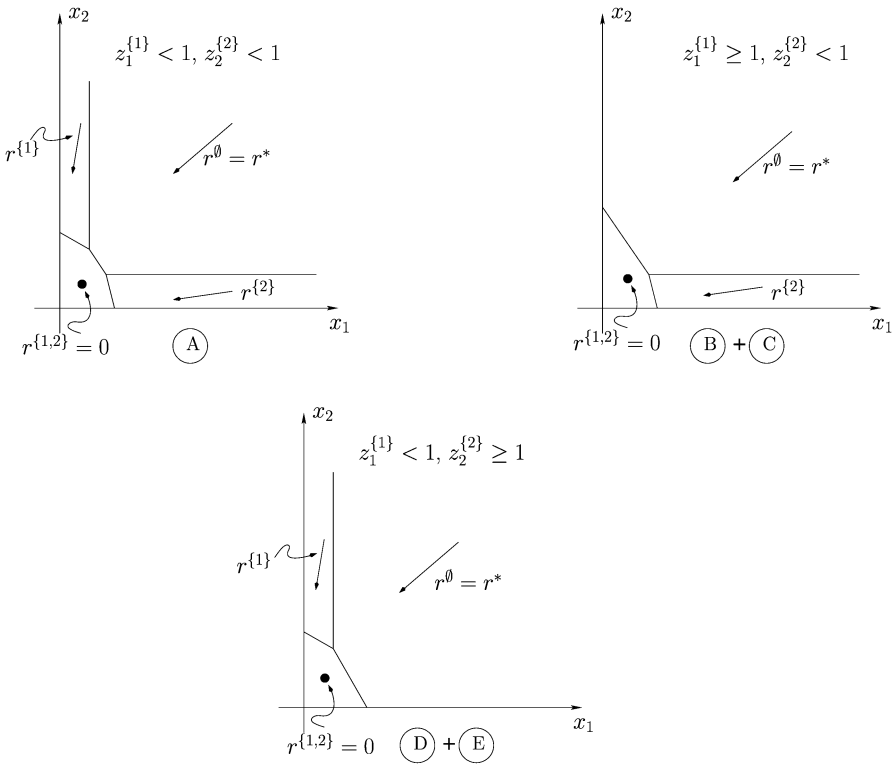

Fig. 3 Gradients of the piecewise affine subsolution

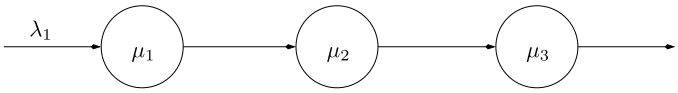

Fig. 4 A three node tandem Jackson network

4. For $E = \{2\}$, (10.3) yields $z_2^{\{2\}} = (1 - \bar{p})\rho_2 + \bar{p}\rho_2/\rho_1$ and $r^{\{2\}} = (\log \rho_1, \log \rho_2 - \log z_2^{\{2\}})$.

It is not difficult to show that $z_1^{\{1\}} \geq 1$ and $z_2^{\{2\}} \geq 1$ cannot hold simultaneously, and Theorem 13.2 yields the behavior of the optimal trajectories, as depicted in Fig. 2.

In these figures and Fig. 3 we see the relationship between the optimal trajectories, the values of z_i^E, and those parts of the boundary where the affine function $-\langle r^*, x\rangle$ should be perturbed to produce a subsolution (see also the discussion in Sect. 9). In A, $z_1^{\{1\}} < 1$ and $z_2^{\{2\}} < 1$, the time-reversed trajectory pushes into both boundaries, and the interior solution must be perturbed near both boundaries. In B, $z_1^{\{1\}} = 1$ and $z_2^{\{2\}} < 1$, the time-reversed trajectory is parallel to $\{x : x_1 = 0\}$ and pushes into the other boundary, and the interior solution is perturbed only near this second boundary.

Table 1 Two-node Jackson network, total population overflow

	$n = 20$	$n = 40$	$n = 60$
Theoretical value	9.60×10^{-11}	7.27×10^{-22}	5.44×10^{-33}
Estimate	9.64×10^{-11}	7.42×10^{-22}	5.54×10^{-33}
Std. Err.	0.09×10^{-11}	0.11×10^{-22}	0.17×10^{-33}
Rel. Err.	1.0%	1.5%	3.1%
95% C.I.	$[9.46, 9.82] \times 10^{-11}$	$[7.20, 7.64] \times 10^{-22}$	$[5.20, 5.88] \times 10^{-33}$

Table 2 Two-node Jackson network, individual buffer overflow

	$n = 20$	$n = 40$	$n = 60$
Theoretical value	2.66×10^{-11}	1.96×10^{-22}	1.46×10^{-33}
Estimate	2.60×10^{-11}	1.96×10^{-22}	1.39×10^{-33}
Std. Err.	0.06×10^{-11}	0.08×10^{-22}	0.08×10^{-33}
Rel. Err.	2.3%	4.1%	5.5%
95% C.I.	$[2.48, 2.72] \times 10^{-11}$	$[1.80, 2.12] \times 10^{-22}$	$[1.23, 1.55] \times 10^{-33}$

Table 3 Two-node Jackson network, overflow of the first buffer

	$n = 20$	$n = 40$	$n = 60$
Theoretical value	4.44×10^{-13}	3.83×10^{-26}	3.31×10^{-39}
Estimate	4.32×10^{-13}	3.89×10^{-26}	3.28×10^{-39}
Std. Err.	0.09×10^{-13}	0.07×10^{-26}	0.06×10^{-39}
Rel. Err.	2.0%	1.8%	1.8%
95% C.I.	$[4.14, 4.50] \times 10^{-13}$	$[3.75, 4.03] \times 10^{-26}$	$[3.16, 3.40] \times 10^{-39}$

Finally in C, $z_1^{\{1\}} > 1$ and $z_2^{\{2\}} < 1$, the time-reversed trajectory moves away from $\{x : x_1 = 0\}$ and pushes into the other boundary, and again the interior solution is perturbed only near the second boundary.

As far as the importance sampling scheme is concerned, the piecewise affine subsolution $\bar{W}^\delta$ partitions the state space into several regions (this partition is independent of the value of γ), and is illustrated in Fig. 3. Note that the "boundary layer" in Fig. 3 on x_1 and/or x_2 axis has width δ under our choice of $\{C_F : F \in \mathcal{F}\}$.

For any target set Γ, the change of measure is the same and given by (12.1)–(12.2). Below is a collection of simulation results with

$$\lambda = (0.1, 0), \qquad \mu = (0.5, 0.4), \qquad p = 1, \qquad \bar{p} = 0.1.$$

The simulation parameters are set as $\delta = 0.1$, $k = 0.5$. See Tables 1, 2, 3.

Table 4 Three-node tandem Jackson network, total population overflow

	$n = 20$	$n = 40$	$n = 60$
Theoretical value	1.26×10^{-15}	4.29×10^{-33}	8.32×10^{-51}
Estimate	1.23×10^{-15}	4.26×10^{-33}	8.60×10^{-51}
Std. Err.	0.02×10^{-15}	0.07×10^{-33}	0.30×10^{-51}
Rel. Err.	1.6%	1.6%	3.6%
95% C.I.	$[1.19, 1.27] \times 10^{-15}$	$[4.12, 4.40] \times 10^{-33}$	$[8.00, 9.20] \times 10^{-51}$

Table 5 Three-node tandem Jackson network, individual buffer overflow

	$n = 20$	$n = 40$	$n = 60$
Theoretical value	2.38×10^{-11}	2.06×10^{-22}	1.79×10^{-33}
Estimate	2.68×10^{-11}	2.01×10^{-22}	1.40×10^{-33}
Std. Err.	0.12×10^{-11}	0.14×10^{-22}	0.13×10^{-33}
Rel. Err.	5.0%	6.8%	7.3%
95% C.I.	$[2.44, 2.92] \times 10^{-11}$	$[1.73, 2.29] \times 10^{-22}$	$[1.14, 1.66] \times 10^{-33}$

Table 6 Three-node tandem Jackson network, overflow of the first buffer

	$n = 20$	$n = 40$	$n = 60$
Theoretical value	7.93×10^{-18}	6.88×10^{-36}	5.97×10^{-54}
Estimate	7.73×10^{-18}	6.68×10^{-36}	6.02×10^{-54}
Std. Err.	0.74×10^{-18}	0.22×10^{-36}	0.20×10^{-54}
Rel. Err.	9.3%	3.2%	3.4%
95% C.I.	$[6.25, 9.21] \times 10^{-18}$	$[6.24, 7.12] \times 10^{-36}$	$[5.62, 6.42] \times 10^{-54}$

14.2 A three-node tandem Jackson networks

Consider the three-node tandem Jackson network with $\lambda_1 = 0.04$ and $\mu_1 = \mu_2 = \mu_3 = 0.32$. The simulation parameters are set again as $\delta = 0.1,\ k = 0.5$. See Fig. 4 and Tables 4, 5.

14.3 A four-node Jackson network

Consider the following four-node Jackson network with

$$\lambda = (0.05, 0.05, 0.05, 0), \qquad \mu = (0.15, 0.2, 0.15, 0.35),$$
$$P_{12} = 0.2, \qquad P_{32} = 0.3.$$

Since we would like to have the theoretical values available as a benchmark and the recursive algorithm for theoretical values becomes exceedingly slow for big n, especially in the case of the first buffer overflow, we will present the numerical results

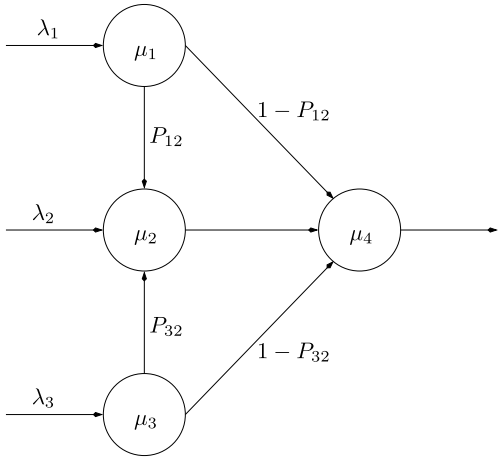

Fig. 5 A four-node Jackson network

Table 7 Four-node Jackson network, total population overflow

	$n = 20$	$n = 25$	$n = 30$
Theoretical value	7.08×10^{-6}	1.18×10^{-7}	1.82×10^{-9}
Estimate	7.10×10^{-6}	1.15×10^{-7}	1.81×10^{-9}
Std. Err.	0.10×10^{-6}	0.02×10^{-7}	0.05×10^{-9}
Rel. Err.	1.4%	1.7%	2.8%
95% C.I.	$[6.90, 7.30] \times 10^{-6}$	$[1.11, 1.19] \times 10^{-7}$	$[1.71, 1.91] \times 10^{-9}$

Table 8 Four-node Jackson network, individual buffer overflow

	$n = 20$	$n = 25$	$n = 30$
Theoretical value	2.21×10^{-7}	3.11×10^{-9}	4.44×10^{-11}
Estimate	2.34×10^{-7}	2.90×10^{-9}	4.56×10^{-11}
Std. Err.	0.17×10^{-7}	0.28×10^{-9}	0.64×10^{-11}
Rel. Err.	7.8%	9.0%	14.4%
95% C.I.	$[2.00, 2.68] \times 10^{-7}$	$[2.34, 2.46] \times 10^{-9}$	$[3.28, 5.84] \times 10^{-11}$

for $n = 20, 25, 30$. The simulation parameters are set as $\delta = 0.3,\ k = 0.3$. See Fig. 5 and Tables 7, 8, 9.

14.4 Remarks on numerical simulation

The choice of the parameters ε, δ and $\{C_F : F \in \mathcal{F}\}$ will effect the outcome of the numerical simulation, sometimes significantly, for a given n. For instance, in the example of the three-node tandem queue network, if we change $\delta = 0.1$ to $\delta = 0.15$ for $n = 20$, the numerical simulation results are better, especially for the problem of first buffer overflow. See Table 10.

Table 9 Four-node Jackson network, overflow of the first buffer

	$n = 20$	$n = 25$	$n = 30$
Theoretical value	8.03×10^{-10}	3.31×10^{-12}	1.36×10^{-14}
Estimate	7.54×10^{-10}	3.84×10^{-12}	1.49×10^{-14}
Std. Err.	0.40×10^{-10}	0.34×10^{-12}	0.17×10^{-14}
Rel. Err.	5.0%	10.2%	12.5%
95% C.I.	$[6.74, 8.8.34] \times 10^{-10}$	$[3.20, 4.48] \times 10^{-12}$	$[1.15, 1.83] \times 10^{-14}$

Table 10 Three-node tandem Jackson network, $n = 20$, $\delta_n = 0.15$

Overflow type	Total population	Individual buffer	1st buffer
Theoretical value	1.26×10^{-15}	2.38×10^{-17}	7.93×10^{-18}
Estimate	1.26×10^{-15}	2.40×10^{-17}	7.94×10^{-18}
Std. Err.	0.01×10^{-15}	0.08×10^{-17}	0.14×10^{-18}
Rel. Err.	0.8%	3.4%	1.8%
95% C.I.	$[1.24, 1.28] \times 10^{-15}$	$[2.24, 2.56] \times 10^{-17}$	$[7.66, 8.22] \times 10^{-18}$

Therefore, it is advisable to experiment with the parameters for fixed n to obtain better numerical performance.

A problem which becomes apparent in the numerical experimentation is that sometimes the constants $\{C_F : F \in \mathcal{F}\}$ may grow too fast. This phenomenon is of course more acute when the dimension is high and, for a given n, has some impact on the choice of δ. If δ is too small, the subsolution will fail to catch the important "boundary layer" information present in the prelimit. On the other hand, if δ is too big, the reduction of the value of the subsolution at the origin will impact the optimality. Sometimes one may be able to tweak the choice of $\{C_F\}$ to alleviate this situation, but not always. Therefore, it remains an interesting open question whether there are ways to automatically fine tune these parameters. Another possibility, as noted in the Introduction, is to forego the universal scheme in favor of a subsolution specifically designed for the problem at hand.

Acknowledgement We would like to thank anonymous referees for their careful reading of the paper and many helpful suggestions that have greatly improved the quality of the paper.

Appendix: Proofs of certain results

A.1 Proofs for Sect. 8

We will prove a slightly stronger version of Lemma 8.1, which will be useful in the proof of Theorem 12.1 to handle general moments of the importance sampling estimator. Fix an arbitrary $s \geq 1$. We define, for any $x \in \mathbb{R}_+^d$ and $\bar{\Theta}, \theta \in \mathcal{P}^+(\mathbb{V})$,

$$L_s(x, \alpha; \bar{\Theta}, \theta) \doteq \frac{s}{2}\big\langle \alpha, \mathbb{F}(x, \theta)\big\rangle + (s-1)\sum_{v \in \mathbb{V}} \theta[v] \log \frac{\bar{\Theta}[v]}{\Theta[v]} + R(\theta \| \Theta)$$

and

$$\mathbb{H}_s(x,\alpha) = \sup_{\bar{\Theta}\in\mathcal{P}^+(\mathbb{V})} \inf_{\theta\in\mathcal{P}^+(\mathbb{V})} L_s(x,\alpha;\bar{\Theta},\theta).$$

Note that $\mathbb{H}$ is the special case of $\mathbb{H}_s$ with $s=2$. The interpretation of $\mathbb{H}_s$ is that if one used the sth moment instead of the second moment of $\hat{p}_n$ as the criteria for optimality, $\mathbb{H}_s$ instead of $\mathbb{H}$ would be the associated Hamiltonian.

Lemma A.1 *Consider any $E\subset\{1,\dots,d\}$ and $x\in\partial_E$. For every $\alpha\in\mathbb{R}^d_+$,*

$$\mathbb{H}_s(x,\alpha) = -sH(x,-\alpha/2) = -sH_E(-\alpha/2).$$

The min/max condition holds in the definition of $\mathbb{H}_s$, and the saddle point is $(\Theta^*_E(\alpha),\Theta^*_E(\alpha))$ as defined in Lemma 8.1. In particular, for every x, $\mathbb{H}_s(\cdot)$ is concave.

Proof It follows from the definition of relative entropy that

$$\begin{aligned} L_s(x,\alpha;\bar{\Theta},\theta) &= \frac{s}{2}\big\langle\alpha,\mathbb{F}(x,\theta)\big\rangle - (s-1)\sum_{v\in\mathbb{V}}\theta[v]\log\frac{\theta[v]}{\bar{\Theta}[v]} \\ &\quad + (s-1)\sum_{v\in\mathbb{V}}\theta[v]\log\frac{\theta[v]}{\Theta[v]} + R(\theta\|\Theta) \\ &= \frac{s}{2}\big\langle\alpha,\mathbb{F}(x,\theta)\big\rangle - (s-1)R(\theta\|\bar{\Theta}) + sR(\theta\|\Theta). \end{aligned}$$

Assume that $x\in\partial_E$ for some $E\subset\{1,\dots,d\}$. It follows from the definition of H_E (7.2) that Θ^*_E as defined in Lemma 8.1 is a probability distribution on $\mathbb{V}$. We now show that $(\Theta^*_E(\alpha),\Theta^*_E(\alpha))$ is a saddle point, or for any $\bar{\Theta},\theta\in\mathcal{P}^+(\mathbb{V})$,

$$L_s\big(x,\alpha;\bar{\Theta},\Theta^*_E(\alpha)\big) \le L_s\big(x,\alpha;\Theta^*_E(\alpha),\Theta^*_E(\alpha)\big) \le L_s\big(x,\alpha;\Theta^*_E(\alpha),\theta\big).$$

The first inequality is trivial due to the fact that $s\ge 1$, the nonnegativity of relative entropy, and that $R(\theta\|\bar{\theta})=0$ if and only if $\theta=\bar{\theta}$. As for the second inequality, straightforward calculation yields that

$$L_s\big(x,\alpha;\Theta^*_E(\alpha),\theta\big) = -(s-1)H_E(-\alpha/2) + \frac{1}{2}\sum_{v\in\mathbb{V}}\theta[v]\big\langle\alpha,\pi[x,v]\big\rangle + R(\theta\|\Theta).$$

By elementary calculus, the right-hand side is minimized at $\theta=\Theta^*_E(\alpha)$. Therefore, the second inequality holds, and $(\Theta^*_E(\alpha),\Theta^*_E(\alpha))$ is a saddle point. In particular,

$$\mathbb{H}_s(x,\alpha) = L_s\big(x,\alpha;\Theta^*_E(\alpha),\Theta^*_E(\alpha)\big) = -sH_E(-\alpha/2).$$

This completes the proof. □

A.2 Proofs for Sect. 10

Proof of Lemma 10.6 Part 1 is trivial by the definition of $\mathcal{F}$ and Remark 10.4. As for Part 2, let $E_1, E_2 \in \mathcal{F}$ and $E \doteq E_1 \cup E_2$. Define $\bar{z} \doteq \{\bar{z}_i : i \in E\}$ by

$$\bar{z}_i \doteq \begin{cases} z_i^{E_1} & \text{if } i \in E_1 \setminus (E_1 \cap E_2), \\ z_i^{E_1} \wedge z_i^{E_2} & \text{if } i \in E_1 \cap E_2, \\ z_i^{E_2} & \text{if } i \in E_2 \setminus (E_1 \cap E_2). \end{cases}$$

It suffices to show that $z^E \le \bar{z}$. Thanks to Corollary 10.3, we only need to show that $\{\bar{z}_i : i \in E\}$ satisfies (10.3) with "=" replaced by "$\le$".

Fix an arbitrary $i \in E = E_1 \cup E_2$. Assume for now that $i \in E_1$. Since $z^{E_2} < 1$, the left-hand side [LHS] of (10.3) with z replaced by $\bar{z}$ satisfies

$$\begin{aligned} \text{LHS} &= \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v \ne -e_i + e_j \\ \text{any } j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} + \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v = -e_i + e_j \\ \text{some } j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} \bar{z}_j \\ &\le \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v \ne -e_i + e_j \\ \text{any } j \in E_1}} \Theta[v] e^{-\langle r^*, v\rangle} + \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v = -e_i + e_j \\ \text{some } j \in E_1}} \Theta[v] e^{-\langle r^*, v\rangle} \bar{z}_j \\ &\le \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v \ne -e_i + e_j \\ \text{any } j \in E_1}} \Theta[v] e^{-\langle r^*, v\rangle} + \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v = -e_i + e_j \\ \text{some } j \in E_1}} \Theta[v] e^{-\langle r^*, v\rangle} z_j^{E_1} \\ &= \mu_i z_i^{E_1}. \end{aligned}$$

Similarly, for $i \in E_2$, LHS $\le \mu_i z_i^{E_2}$. Thus LHS $\le \mu_i \bar{z}_i$ for all $i \in E$, and therefore by Corollary 10.3 the proof of Part 1 is complete.

For Part 3, assume that $E, F \in \mathcal{F}$ and $E \subset F$. The proof of Part 2 implies, with $E_1 = E$ and $E_2 = F$, that $z^F \le \bar{z}$, where

$$\bar{z}_i \doteq \begin{cases} z_i^E \wedge z_i^F & \text{if } i \in E, \\ z_i^F & \text{if } i \in F \setminus E. \end{cases}$$

In particular, for $i \in E$,

$$z_i^F \le \bar{z}_i = z_i^E \wedge z_i^F,$$

or equivalently $z_i^F \le z_i^E$. It follows immediately that $r^E \le r^F$ by the definition (10.7) and because $z_i^F < 1$ for $i \in F \setminus E$. □

A.3 Proofs for Sect. 11

Lemma A.2 *Let $E_1, E_2 \subset \{1, \ldots, d\}$ be such that $E_1 \cap E_2 = \emptyset$. Let $E = E_1 \cup E_2$ and define the vector $\{\bar{z}_i : i \in E\}$ by*

$$\bar{z}_i \doteq \begin{cases} z_i^{E_1} & \textit{if } i \in E_1, \\ 1 & \textit{if } i \in E_2. \end{cases}$$

Suppose that $H_i(-r^{E_1}) \le 0$ [*resp.*, "$\ge$"] *for all* $i \in E_2$. *Then*

1. $\bar{z}_i \le z_i^E$ [*resp.*, "$\ge$"] *for all* $i \in E = E_1 \cup E_2$,
2. $1 < z_i^E$ [*resp.*, "$>$"] *for all those* $i \in E_2$ *such that* $H_i(-r^{E_1}) < 0$ [*resp.*, "$>$"].

Proof We will argue the "$\le$" case, the other case is analogous and thus omitted. For any $i \notin E_1$, thanks to Proposition 10.5 and the definitions of $H_\emptyset$ and H_i, it follows that

$$e^{H_i(-r^{E_1})} - 1 = e^{H_i(-r^{E_1})} - e^{H_\emptyset(-r^{E_1})} = \sum_{v \in \mathbb{V}[-e_i]} \Theta[v]\big(1 - e^{-\langle r^{E_1}, v\rangle}\big).$$

For $i \notin E_1$ and $v \in \mathbb{V}[-e_i]$,

$$e^{-\langle r^{E_1}, v\rangle} = \begin{cases} e^{-\langle r^*, v\rangle} & \text{if } v = -e_i \text{ or } v = -e_i + e_j,\ j \notin E_1, \\ e^{-\langle r^*, v\rangle} z_j^{E_1} & \text{if } v = -e_i + e_j,\ j \in E_1. \end{cases}$$

Thus by (10.9), for all $i \notin E_1$,

$$e^{H_i(-r^{E_1})} - 1 = \mu_i - \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v \ne -e_i + e_j \\ \text{any } j \in E_1}} \Theta[v] e^{-\langle r^*, v\rangle} - \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v = -e_i + e_j \\ \text{some } j \in E_1}} \Theta[v] e^{-\langle r^*, v\rangle} z_j^{E_1}. \tag{A.1}$$

Thus the assumption $H_i(-r^{E_1}) \le 0$ for $i \in E_2$ amounts to

$$\text{LHS} \doteq \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v \ne -e_i + e_j \\ \text{any } j \in E_1}} \Theta[v] e^{-\langle r^*, v\rangle} + \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v = -e_i + e_j \\ \text{some } j \in E_1}} \Theta[v] e^{-\langle r^*, v\rangle} z_j^{E_1} \ge \mu_i = \mu_i \bar{z}_i. \tag{A.2}$$

On the other hand, by the definition of $z_i^{E_1}$, for $i \in E_1$,

$$\text{LHS} = \mu_i z_i^{E_1} = \mu_i \bar{z}_i. \tag{A.3}$$

Combining inequalities (A.2) and (A.3) and by the definition of $\{\bar{z}_j : j \in E = E_1 \cup E_2\}$, we can reorganize the terms in LHS to conclude that for every $i \in E_1 \cup E_2$,

$$\text{LHS} = \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v \ne -e_i + e_j \\ \text{any } j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} + \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v = -e_i + e_j \\ \text{some } j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} \bar{z}_j \ge \mu_i \bar{z}_i.$$

Thanks to Corollary 10.3, we have

$$z_j^E \ge \bar{z}_j$$

for all $j \in E$. This ends the proof of Part 1.

As for Part 2, we assume from now on that $H_i(-r^{E_1}) < 0$ for some $i \in E_2$. Thanks to Part 1, $z_j^E \ge \bar{z}_j$ for all $j \in E$. That is, $z_j^E \ge 1$ for all $j \in E_2$ and $z_j^E \ge z_j^{E_1}$ for all

$j \in E_1$. Therefore,

$$\mu_i z_i^E = \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v \neq -e_i + e_j \\ \text{any } j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} + \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v = -e_i + e_j \\ \text{some } j \in E}} \Theta[v] e^{-\langle r^*, v\rangle} z_j^E$$

$$\geq \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v \neq -e_i + e_j \\ \text{any } j \in E_1}} \Theta[v] e^{-\langle r^*, v\rangle} + \sum_{\substack{v \in \mathbb{V}[-e_i] \\ v = -e_i + e_j \\ \text{some } j \in E_1}} \Theta[v] e^{-\langle r^*, v\rangle} z_j^{E_1}.$$

Since $H_i(-r^{E_1}) < 0$, inequality (A.2) is satisfied with $\geq$ replaced by $>$. It follows that $\mu_i z_i^E > \mu_i$ or $z_i^E > 1$. □

For any set $E \subset \{1, 2, \ldots, d\}$, Part 2 of Lemma 10.6 implies that

$$\eta[E] \doteq \bigcup_{F \subset E, F \in \mathcal{F}} F \tag{A.4}$$

is an element of $\mathcal{F}$ and is the *maximal* subset of E that belongs to $\mathcal{F}$.

Lemma A.3 *Given any subset $E \subset \{1, \ldots, d\}$ and $F \subset E$,*

$$H_F\left(-r^{\eta[E]}\right) \leq 0.$$

Proof Thanks to Lemma 7.1 and the fact that $H_\emptyset(-r^{\eta[E]}) = 0$ by Proposition 10.5, it suffices to argue for all $i \in E$ that

$$H_i\left(-r^{\eta[E]}\right) \leq 0. \tag{A.5}$$

By Proposition 10.5 again, (A.5) holds with equality for $i \in \eta[E]$. Suppose for now that $i \in E \setminus \eta[E]$ and that (A.5) does not hold. By Lemma A.2 [with $E_1 = \eta[E]$ and $E_2 = \{i\}$], we have $\eta[E] \cup \{i\} \in \mathcal{F}$, which contradicts the maximality of $\eta[E]$. Therefore (A.5) holds. □

Proof of Lemma 11.2 Assume $x = (x_1, \ldots, x_d)' \in \partial_E$. Since $\mathbb{H}(x, 2r^G) = -2H_E(-r^G)$ by Lemma 8.1, it suffices to show that for any $F \subset E$,

$$H_F\left(-r^G\right) \leq 0. \tag{A.6}$$

Define a to be a small positive constant such that

$$a \leq \min\{C_F - C_G : G, F \in \mathcal{F}, G \subset F, G \neq F\}$$

and for all nonempty $F \in \mathcal{F}$,

$$C_F \geq \max\left\{\frac{\log(1/z_i^F)}{\log(1/z_i^G)} \cdot (C_G + a) : i \in G \subset F, G \neq F, G \in \mathcal{F}\right\}.$$

Note that such a positive constant a always exists.

For any $F \subset \{1, \dots, d\}$, the definitions (10.7) and (11.1), and that $x_i = 0$ for $i \in E$ imply that

$$\bar{W}_G^\delta(x) - \bar{W}_F^\delta(x) = 2 \sum_{i \in F \setminus E} x_i \log z_i^F - 2 \sum_{i \in G \setminus E} x_i \log z_i^G + (C_F - C_G)\delta.$$

By definition, $\bar{W}_F^\delta(x) \geq \bar{W}^\delta(x)$, and therefore by the assumption of the lemma

$$\bar{W}_G^\delta(x) - \bar{W}_F^\delta(x) < a\delta.$$

Let $\eta[E]$ be the *maximal* subset of E that belongs to $\mathcal{F}$ as defined in (A.4). Plugging in $F = \eta[E]$ and $F = \eta[G \cup E] \supset G$, we arrive at the following two inequalities:

$$-2 \sum_{i \in G \setminus E} x_i \log z_i^G + \big(C_{\eta[E]} - C_G\big)\delta \ < \ a\delta, \tag{A.7}$$

$$2 \sum_{i \in \eta[G \cup E] \setminus E} x_i \log z_i^{\eta[G \cup E]} - 2 \sum_{i \in G \setminus E} x_i \log z_i^G + (C_{\eta[G \cup E]} - C_G)\delta \ < \ a\delta. \tag{A.8}$$

If $G \subset E$, then $G \subset \eta[E]$ by the maximality of $\eta[E]$. But now inequality (A.7) reduces to $C_{\eta[E]} - C_G < a$. However, since $C_F - C_G \geq a$ if G is a strict subset of F, $G = \eta[E]$, and by Lemma A.3 the proof is completed.

Assume from now on that $G \setminus E \neq \emptyset$. Note that the definition of η implies

$$G \cup E \supset \eta[G \cup E] \supset \eta[G] \cup \eta[E] = G \cup \eta[E].$$

It follows that

$$\eta[G \cup E] \setminus E = G \setminus E.$$

Therefore (A.8) yields that

$$C_{\eta[G \cup E]} < C_G + 2\delta^{-1} \sum_{i \in G \setminus E} x_i \big[\log\big(1/z_i^{\eta[G \cup E]}\big) - \log\big(1/z_i^G\big)\big] + a.$$

Let

$$c^* \doteq \max\left\{ \frac{\log(1/z_i^{\eta[G \cup E]})}{\log(1/z_i^G)} : i \in G \setminus E = \eta[G \cup E] \setminus E \right\}.$$

Note that $c^* \geq 1$. It follows from the definition of c^* and (A.7) that

$$\begin{aligned} C_{\eta[G \cup E]} &< C_G + \big(c^* - 1\big) \cdot 2\delta^{-1} \sum_{i \in G \setminus E} x_i \log\big(1/z_i^G\big) + a \\ &\leq C_G + \big(c^* - 1\big)(C_G - C_{\eta[E]} + a) + a \\ &\leq c^*(C_G + a). \end{aligned}$$

By the definition of a, if G is a strict subset of $\eta[G \cup E]$, then $C_{\eta[G\cup E]} \geq c^*(C_G + a)$, a contradiction. It follows that $\eta[G \cup E] = G$. Invoking Lemma A.3, we complete the proof. □

A.4 Proof of Theorem 12.1

The proof of this main result is analogous to that of [11], which is based on a verification type of argument. Even though [11] only treats the special case of $b = 2$ for tandem queueing networks, the differences in the proofs are largely notational. For this reason, we only give a sketch of the proof.

An important technicality of the proof is the validity of an exponential bound on the return time to the origin. We use a different but much simpler approach to establish a slightly stronger bound [22]; compare the current Lemma A.4 with [11, Proposition A.1].

Recall the definition

$$T_0 \doteq \inf\bigl\{k \geq 1 : X^n(k) = 0\bigr\} = \inf\bigl\{k \geq 1 : Z(k) = 0\bigr\}$$

and that

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

where Y is a sequence of iid random variables taking values in $\mathbb{V}$ with common distribution Θ. Denote by E_z the conditional expectation given that $Z(0) = z$ for each $z \in \mathbb{Z}_+^d$.

Lemma A.4 *There exists a constant $c > 0$ such that $E_z[\exp\{cT_0\}]$ is finite for all $z \in \mathbb{Z}_+^d$.*

Proof Let $\sigma_0 \doteq \inf\{k \geq 0 : Z(k) = 0\}$. Note that $\sigma_0 = T_0$ if $Z(0) \neq 0$ and $\sigma_0 = 0$ if $Z(0) = 0$. For each $z \in \mathbb{Z}_+^d$, define $h(z) \doteq E_z[\sigma_0]$. Since the network is positive recurrent, $h(z)$ is finite for every z. Define the process

$$S(k) \doteq \begin{cases} h(Z(k)) & \text{if } k \leq \sigma_0, \\ \sigma_0 - k & \text{if } k > \sigma_0. \end{cases}$$

There are two key properties regarding S, which can be established using an analogous argument to [11, Lemmas A.2, A.4, and A.5]. We omit the proof.

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

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

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

(ii) The increments of the process S are *uniformly* bounded.

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

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

Let B be the uniform bound of the increments of S. Define

$$\bar{\varepsilon} \doteq \min\left\{\frac{\varepsilon_0}{B}, \frac{1}{2B^2}\right\}.$$

Then for every $k \geq 0$, thanks to property (ii) and the fact that $\bar{\varepsilon} B \leq \varepsilon_0$,

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

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

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

This is equivalent to the process $\{(1-\bar{\varepsilon}/2)^{-k} e^{\bar{\varepsilon} S(k)}, \mathcal{F}_k\}$ being a supermartingale. In particular, given $Z(0) = z \neq 0$, the Optional Sampling Theorem and $S(T_0) = S(\sigma_0) = h(0) = 0$ imply that,

$$e^{\bar{\varepsilon} h(z)} = e^{\bar{\varepsilon} S(0)} \geq E_z\big[(1-\bar{\varepsilon}/2)^{-T_0}\big].$$

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

$$E_z\big[e^{cT_0}\big] < \infty$$

for all $z \in \mathbb{Z}_+^d$ and $z \neq 0$. For $Z(0) = 0$, we only need to note that

$$E_0\big[e^{cT_0}\big] = E_0\big[E_0\big[e^{cT_0}|Z(1)\big]\big] = e^c \sum_{i=1}^{d} \mathbb{P}_0\big\{Z(1) = e_i\big\} E_{e_i}\big[e^{cT_0}\big] < \infty.$$

This completes the proof. □

Proof of Theorem 12.1 To ease exposition, we write $W^n \doteq W^{\varepsilon_n,\delta_n}$, $\rho_F^n \doteq \rho_F^{\varepsilon_n,\delta_n}$, and set $\bar{\varepsilon}_n \doteq M \exp\{-a\delta_n/\varepsilon_n\}$, where M is given by Proposition 11.4. Fix an arbitrary real number $s \geq 1$. The value of s will be later determined, depending on n and b. Recall the definitions of L_s and $\mathbb{H}_s$ in Sect. A.1. By concavity of $\log x$, Lemma 11.3, and the definition of $\bar{\Theta}^n$ in (12.2), we have

$$L_s\big(x, DW^n(x); \bar{\Theta}^n[\cdot|x], \theta\big) \geq \sum_{F\in\mathcal{F}} \rho_F^n(x) L_s\big(x, 2r_F; \bar{\Theta}_F[\cdot|x], \theta\big)$$

for any $\theta \in \mathcal{P}^+(\mathbb{V})$. In particular,

$$\inf_{\theta\in\mathcal{P}^+(\mathbb{V})} L_s\big(x, DW^n(x); \bar{\Theta}^n[\cdot|x], \theta\big) \geq \sum_{F\in\mathcal{F}} \rho_F^n(x) \inf_{\theta\in\mathcal{P}^+(\mathbb{V})} L_s\big(x, 2r^F; \bar{\Theta}_F[\cdot|x], \theta\big).$$

However, by definition (12.1) of $\bar{\Theta}_F[\cdot|x]$, Lemma A.1, and Lemma 8.1,

$$\inf_{\theta\in\mathcal{P}^+(\mathbb{V})} L_s\big(x, 2r^F; \bar{\Theta}_F[\cdot|x], \theta\big) = \mathbb{H}_s\big(x, 2r^F\big) = \frac{s}{2}\mathbb{H}\big(x, 2r^F\big).$$

It follows from Proposition 11.4 that

$$\inf_{\theta\in\mathcal{P}^+(\mathbb{V})} L_s\big(x, DW^n(x); \bar{\Theta}^n[\cdot|x], \theta\big) \geq \sum_{F\in\mathcal{F}} \rho_F^n(x)\cdot\frac{s}{2}\mathbb{H}\big(x, 2r^F\big)$$
$$\geq -\frac{s}{2}\bar{\varepsilon}_n. \tag{A.9}$$

For any $\theta\in\mathbb{V}$ and $x\in\mathbb{R}_+^d$, Taylor's expansion yields that

$$\begin{aligned}
&n\sum_{v\in\mathbb{V}}\theta[v]\cdot\left[W^n\left(x+\frac{1}{n}\pi[x,v]\right)-W^n(x)\right]-\big\langle DW^n(x),\mathbb{F}(x,\theta)\big\rangle\\
&\quad= n\sum_{v\in\mathbb{V}}\theta[v]\cdot\left[W^n\left(x+\frac{1}{n}\pi[x,v]\right)-W^n(x)-\left\langle DW^n(x),\frac{1}{n}\pi[x,v]\right\rangle\right]\\
&\quad= \sum_{v\in\mathbb{V}}\theta[v]\cdot\frac{1}{2n}\pi[x,v]^T D^2W(\bar{x}_v)\pi[x,v],
\end{aligned}$$

where $\bar{x}_v$ is some point on the line connecting x and $x+\pi[x,v]/n$. By straightforward computation, it is not difficult to show that each component of the Hessian matrix $D^2W^n(x)$ is uniformly bounded by C/ε_n for some constant C that only depends on the system parameters. This, the boundedness of $\pi[x,v]$, the assumption $s\geq 1$, and inequality (A.9) imply that

$$\begin{aligned}
&\inf_{\theta\in\mathcal{P}^+(\mathbb{V})}\left\{\frac{s}{2}\sum_{v\in\mathbb{V}}\theta[v]\cdot n\left[W^n\left(x+\frac{1}{n}\pi[x,v]\right)-W^n(x)\right]\right.\\
&\qquad\left.+(s-1)\sum_{v\in\mathbb{V}}\theta[v]\log\frac{\bar{\Theta}^n[v|x]}{\Theta[v]}+R(\theta\|\Theta)\right\}\\
&\quad\geq -\frac{s}{2}\left[\bar{\varepsilon}_n+\frac{\bar{C}}{n\varepsilon_n}\right]
\end{aligned}\tag{A.10}$$

for some constant $\bar{C}$ that only depends on the system parameters.

Applying the relative entropy representation [11, Remark 3.1] to the left-hand side of (A.10) and using the notation $\beta_n=\bar{\varepsilon}_n+\bar{C}/(n\varepsilon_n)$, we arrive at

$$e^{-s\beta_n/2}\cdot\sum_{v\in\mathbb{V}}e^{-sn[W^n(x+\pi[x,v]/n)-W^n(x)]/2}\left(\frac{\Theta[v]}{\bar{\Theta}^n[v|x]}\right)^{s-1}\cdot\Theta[v]\leq 1.$$

This inequality amounts to saying that the process $U=\{U(k):k\geq 0\}$ where

$$U(k)\doteq e^{-ks\beta_n/2}e^{-snW^n(X^n(k))/2}\left(\prod_{j=0}^{k-1}\frac{\Theta[Y(j+1)]}{\bar{\Theta}^n[Y(j+1)|X^n(j)]}\right)^{s-1}$$

is a supermartingale under the original probability measure $\mathbb{P}$. In particular, by the Optional Sampling Theorem and the nonnegativity of U,

$$E^{\mathbb{P}}U(T_n \wedge T_0) \le E^{\mathbb{P}}U(0) = e^{-snW^n(0)/2}.$$

Since $W^n(x) \le 0$ for $x \in \Gamma$ and $\hat{p}_n = \hat{p}_n \cdot 1_{\{T_n < T_0\}}$, it follows that

$$e^{-snW^n(0)/2} \ge E^{\mathbb{P}}\big[U(T_n)1_{\{T_n<T_0\}}\big] \ge E^{\mathbb{P}}\big[e^{-s\beta_n T_n/2}\hat{p}_n^{s-1}\big].$$

Assume from now on that $s > b$. Then by Hölder's inequality,

$$\begin{aligned} E^{\mathbb{P}}\big[\hat{p}_n^{b-1}\big] &\le E^{\mathbb{P}}\big[e^{-\frac{s\beta_n T_n}{2}}\hat{p}_n^{s-1}\big]^{\frac{b-1}{s-1}} \cdot E^{\mathbb{P}}\big[e^{\frac{s\beta_n(b-1)T_n}{2(s-b)}} \cdot 1_{\{T_n<T_0\}}\big]^{\frac{s-b}{s-1}} \\ &\le e^{-\frac{sn(b-1)}{2(s-1)}W^n(0)} \cdot E^{\mathbb{P}}\big[e^{\frac{s\beta_n(b-1)T_n}{2(s-b)}} \cdot 1_{\{T_n<T_0\}}\big]^{\frac{s-b}{s-1}}. \end{aligned}$$

Let c be the constant determined by Lemma A.4 and choose s such that

$$\frac{s\beta_n(b-1)}{2(s-b)} = c \quad \text{or} \quad s = b \cdot \frac{2c}{2c - \beta_n(b-1)}.$$

It follows that

$$\begin{aligned} \frac{1}{n}\log E^{\mathbb{P}}\big[\hat{p}_n^{b-1}\big] &\le -\frac{s(b-1)}{2(s-1)}W^n(0) + \frac{s-b}{s-1}\frac{1}{n}\log E^{\mathbb{P}}\big[e^{cT_0}\big] \\ &= -\frac{bc}{2c+\beta_n}W^n(0) + \frac{b\beta_n}{2c+\beta_n}\frac{1}{n}\log E^{\mathbb{P}}\big[e^{cT_0}\big]. \end{aligned} \tag{A.11}$$

Note that by Lemma 11.3 and definition (11.1),

$$2\gamma - \log|\mathcal{F}| \cdot \varepsilon_n - C_{\{1,2,\ldots,d\}}\delta_n \le W^n(0) \le 2\gamma.$$

Letting $n \to \infty$ on both sides of (A.11), and observing that $\beta_n \to 0$, we have

$$\limsup \frac{1}{n}\log E^{\mathbb{P}}\big[\hat{p}_n^{b-1}\big] \le -b\gamma.$$

But

$$E^{\mathbb{Q}_n}\big[\hat{p}_n^{b}\big] = E^{\mathbb{P}}\big[\hat{p}_n^{b-1}\big],$$

and by Jensen's inequality

$$\liminf_n \frac{1}{n}\log E^{\mathbb{Q}_n}\big[\hat{p}_n^{b}\big] \ge \liminf_n \frac{b}{n}\log E^{\mathbb{Q}_n}[\hat{p}_n] = \liminf_n \frac{b}{n}\log p_n = -b\gamma.$$

This completes the proof. □

A.5 Proofs of Sect. 13

Proof of Lemma 13.1 To ease notation, we write, for each $v \in \mathbb{V}$,

$$\bar{\Theta}[v] \doteq \Theta[v] \cdot e^{-\langle r^*, v\rangle}.$$

Proof of Part 1 Fix an $i \in F$. Note that for $v \in \mathbb{V}$, $\langle v, e_i\rangle \neq 0$ only if $v \in \mathbb{V}[-e_i] \cup \mathbb{V}[e_i]$, and that $\mathbb{V}[e_i] \cap \mathbb{V}[-e_i] = \{-e_i + e_i\}$ with $\langle -e_i + e_i, e_i\rangle = 0$. It follows that

$$\sum_{v\in\mathbb{V}} \Theta[v] e^{-\langle r^F, v\rangle} \langle v, e_i\rangle = \sum_{v\in\mathbb{V}[-e_i]} \Theta[v] e^{-\langle r^F, v\rangle} \langle v, e_i\rangle + \sum_{v\in\mathbb{V}[e_i]} \Theta[v] e^{-\langle r^F, v\rangle} \langle v, e_i\rangle.$$

But by the definition (10.7) of r^F and $\bar{\Theta}[v]$,

$$\sum_{v\in\mathbb{V}[e_i]} \Theta[v] e^{-\langle r^F, v\rangle} \langle v, e_i\rangle = \sum_{\substack{v\in\mathbb{V}[e_i]\\ v\neq e_i - e_j\\ \text{any } j\in F}} \bar{\Theta}[v] z_i^F + \sum_{\substack{v\in\mathbb{V}[e_i]\\ v= e_i - e_j\\ \text{some } j\in F}} \bar{\Theta}[v] z_i^F / z_j^F,$$

and, thanks to (10.3),

$$\begin{aligned}\sum_{v\in\mathbb{V}[-e_i]} \Theta[v] e^{-\langle r^F, v\rangle} \langle v, e_i\rangle &= -\sum_{\substack{v\in\mathbb{V}[-e_i]\\ v\neq -e_i + e_j\\ \text{any } j\in F}} \bar{\Theta}[v]/z_i^F - \sum_{\substack{v\in\mathbb{V}[-e_i]\\ v= -e_i + e_j\\ \text{some } j\in F, j\neq i}} \bar{\Theta}[v] z_j^F / z_i^F \\ &= -\mu_i + \bar{\Theta}[-e_i + e_i]. \end{aligned} \tag{A.12}$$

Moreover, for $j \in F$ with $j \neq i$,

$$\sum_{v\in\mathbb{V}[-e_j]} \Theta[v] e^{-\langle r^F, v\rangle} \langle v, e_i\rangle = \bar{\Theta}[-e_j + e_i] z_i^F / z_j^F.$$

Therefore, by combining (A.12) with the second term in (13.1), we can write

$$\begin{aligned}\langle \beta_F^*, e_i\rangle &= \sum_{\substack{v\in\mathbb{V}[e_i]\\ v\neq e_i - e_j\\ \text{any } j\in F}} \bar{\Theta}[v] z_i^F + \sum_{\substack{v\in\mathbb{V}[e_i]\\ v= e_i - e_j\\ \text{some } j\in F, j\neq i}} \bar{\Theta}[v] z_i^F / z_j^F + z_i^F \big(-\mu_i + \bar{\Theta}[-e_i + e_i]\big) \\ &\quad - \sum_{j\in F, j\neq i} \big(1 - z_j^F\big) \bar{\Theta}[-e_j + e_i] z_i^F / z_j^F \\ &= \sum_{\substack{v\in\mathbb{V}[e_i]\\ v\neq e_i - e_j\\ \text{any } j\in F}} \bar{\Theta}[v] z_i^F + \sum_{\substack{v\in\mathbb{V}[e_i]\\ v= e_i - e_j\\ \text{some } j\in F, j\neq i}} \bar{\Theta}[v] z_i^F - \mu_i z_i^F + z_i^F \bar{\Theta}[-e_i + e_i] \\ &= z_i^F \left(-\mu_i + \sum_{v\in\mathbb{V}[e_i]} \bar{\Theta}[v]\right) \\ &= 0,\end{aligned}$$

where the last equality follows from (10.5).

Now consider an arbitrary $j \notin F$. A similar argument leads to

$$\begin{aligned}\langle -\beta_F^*, e_j\rangle &= -\sum_{v\in\mathbb{V}[e_j]} \Theta[v]e^{-\langle r^F, v\rangle}\langle v, e_j\rangle - \sum_{v\in\mathbb{V}[-e_j]} \Theta[v]e^{-\langle r^F, v\rangle}\langle v, e_j\rangle \\ &\quad + \sum_{i\in F}\big(1 - z_i^F\big)\Theta[-e_i + e_j]e^{-\langle r^F, -e_i+e_j\rangle} \\ &= -\sum_{v\in\mathbb{V}[e_j]} \bar{\Theta}[v] + \sum_{\substack{v\in\mathbb{V}[-e_j]\\ v=-e_j+e_i\\ \text{some } i\in F}} \bar{\Theta}[v]z_i^F + \sum_{\substack{v\in\mathbb{V}[-e_j]\\ v\neq -e_j+e_i\\ \text{any } i\in F}} \bar{\Theta}[v] \\ &= -\mu_j + \sum_{\substack{v\in\mathbb{V}[-e_j]\\ v=-e_j+e_i\\ \text{some } i\in F}} \bar{\Theta}[v]z_i^F + \sum_{\substack{v\in\mathbb{V}[-e_j]\\ v\neq -e_j+e_i\\ \text{any } i\in F}} \bar{\Theta}[v].\end{aligned}$$

Thanks to the proof of Lemma A.2, more precisely, (A.1),

$$\langle -\beta_F^*, e_j\rangle = 1 - e^{H_j(-r^F)}$$

for $j \notin F$. We claim that $H_j(-r^F) < 0$ for all $j \in E \setminus F$. To this end we will argue by contradiction, invoking Lemma A.2 with $E_1 \doteq F$ and $E_2 \doteq \{j\}$. Assume that $H_j(-r^F) \geq 0$. Then part 1 of Lemma A.2 implies that $\bar{z}_i \geq z_i^{F\cup\{j\}}$ for all $i \in F \cup \{j\}$. By definition of $\bar{z}_i$ and that $F \in \bar{\mathcal{F}}$, $\bar{z}_i \leq 1$ for every $i \in F \cup \{j\}$. It follows that $F \cup \{j\} \in \bar{\mathcal{F}}$, which contradicts the maximality of $F = \Pi[E]$.

Proof of Part 2 We now argue by contradiction that G is nonempty. Suppose that $G = \emptyset$. Then $H_j(-r^F) \leq 0$ for all $j \notin F$. By Lemma A.2 with $E_1 = F$ and $E_2 = \{1, \ldots, d\} \setminus F$, we arrive at

$$z_j^{\{1,\ldots,d\}} \geq 1$$

for all $j \notin F$. But this is impossible since $z_j^{\{1,\ldots,d\}} = \rho_j < 1$ by Remark 10.4. Therefore G is nonempty. The claim that $F \cup G_1 \in \bar{\mathcal{F}}$ for all $G_1 \subset G$ is immediate from Lemma A.2 by setting $E_1 \doteq F$ and $E_2 \doteq G_1$.

Proof of Part 3 Since by definition (10.7),

$$\langle r^* - r^F, \beta_F^*\rangle = \sum_{i\in F} \log\big[z_i^F\big] \cdot \langle e_i, \beta_F^*\rangle = 0,$$

it suffices to show that $L(x, \beta_F^*) = \bar{L}(x, \beta_F^*) = \langle -r^F, \beta_F^*\rangle$. We first argue the easy direction that $L(x, \beta_F^*) \geq \langle -r^F, \beta_F^*\rangle$ and $\bar{L}(x, \beta_F^*) \geq \langle -r^F, \beta_F^*\rangle$. Consider any $\{(\rho_G, \beta_G) : G \subset F\}$ such that

$$\rho_G \geq 0, \qquad \sum_{G\subset F} \rho_G = 1, \qquad \sum_{G\subset F} \rho_G\beta_G = \beta_F^*. \tag{A.13}$$

Note that for any $G \subset F$, $H_G(-r^F) = 0$ by Proposition 10.5. It now follows from the duality of L_G and H_G that

$$\begin{aligned}\sum_{G\subset F} \rho_G L_G(\beta_G) &\geq \sum_{G\subset F} \rho_G\big[\big\langle -r^F, \beta_G\big\rangle - H_G\big(-r^F\big)\big] \\ &= \sum_{G\subset F} \rho_G\big\langle -r^F, \beta_G\big\rangle \\ &= \big\langle -r^F, \beta_F^*\big\rangle. \end{aligned} \tag{A.14}$$

Taking infimum on the left-hand side, we arrive at $L(x, \beta_F^*) \geq \langle -r^F, \beta_F^*\rangle$. The other inequality $\bar{L}(x, \beta_F^*) \geq \langle -r^F, \beta_F^*\rangle$ is shown in the same fashion, observing that $\bar{H}_G(-r^F) = 0$.

In order to show the other direction "$\leq$", for every $G \subset F$, we denote by $\hat{\beta}_G$ the conjugate to $-r^F$ through the conjugacy of L_G and H_G, that is,

$$\hat{\beta}_G \doteq DH_G\big(-r^F\big). \tag{A.15}$$

Note that $\hat{\beta}_G$ is also the conjugate of $-r^F$ through the conjugacy of $\bar{L}_G$ and $\bar{H}_G$, namely,

$$\hat{\beta}_G = D\bar{H}_G\big(-r^F\big).$$

It suffices to find a collection of weights $\{\hat{\rho}_G\}$ such that $\{(\hat{\rho}_G, \hat{\beta}_G) : G \subset F\}$ satisfies the constraints (A.13). Indeed, if this is the case, the inequality in (A.14) becomes equality with $(\hat{\rho}_G, \hat{\beta}_G)$ in place of (ρ_G, β_G), and therefore

$$L\big(x, \beta_F^*\big) \leq \sum_{G\subset F} \hat{\rho}_G L_G(\hat{\beta}_G) = \big\langle -r^F, \beta_F^*\big\rangle.$$

An analogous argument works for $\bar{L}(x, \beta_F^*)$.

For any $G \subset F$, we define

$$\hat{\rho}_G \doteq \prod_{i\in G}\big(1 - z_i^F\big) \cdot \prod_{i\in F\setminus G} z_i^F.$$

Then

$$\sum_{G\subset F} \hat{\rho}_G = \prod_{i\in F}\big[\big(1 - z_i^F\big) + z_i^F\big] = 1.$$

By the definitions in (A.15), (7.2), and Proposition 10.5, it is not difficult to check that

$$\hat{\beta}_G = \sum_{v\in\mathbb{V}} \Theta[v]e^{-\langle r^F, v\rangle} v - \sum_{i\in G}\sum_{v\in\mathbb{V}[-e_i]} \Theta[v]e^{-\langle r^F, v\rangle} v.$$

On the other hand,

$$\sum_{G\subset F} \hat{\rho}_G\hat{\beta}_G = \sum_{G\subset F} \hat{\rho}_G \sum_{v\in\mathbb{V}} \Theta[v]e^{-\langle r^F, v\rangle} v - \sum_{G\subset F} \hat{\rho}_G \sum_{i\in G}\sum_{v\in\mathbb{V}[-e_i]} \Theta[v]e^{-\langle r^F, v\rangle} v$$

$$= \sum_{v \in \mathbb{V}} \Theta[v] e^{-\langle r^F, v \rangle} v - \sum_{i \in F} \sum_{\{G \subset F : i \in G\}} \hat{\rho}_G \sum_{v \in \mathbb{V}[-e_i]} \Theta[v] e^{-\langle r^F, v \rangle} v.$$

However, for every $i \in F$,

$$\sum_{\{G \subset F : i \in G\}} \hat{\rho}_G = \left(1 - z_i^F\right) \cdot \prod_{j \in F \setminus \{i\}} \left[\left(1 - z_j^F\right) + z_j^F\right] = 1 - z_i^F.$$

Therefore

$$\sum_{G \subset F} \hat{\rho}_G \hat{\beta}_G = \beta_F^*,$$

and the proof is completed. □

References

1. Asmussen, S.: Conditioned limit theorems relating a random walk to its associates, with applications to risk reverse process and GI/G/1 queue. Adv. Appl. Probab. **14**, 143–170 (1982)
2. Asmussen, S.: Applied Probability and Queues. Springer, New York (2003)
3. Atar, R., Dupuis, P.: Large deviations and queueing networks: methods for rate function identification. Stoch. Process. Appl. **84**, 255–296 (1999)
4. Bardi, M., Capuzzo-Dolcetta, I.: Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations. Birkhäuser, Basel (1997)
5. De Boer, P.T.: Analysis of state-independent importance sampling measures for the two-node tandem queue. ACM Trans. Model. Comput. Simul. **16**, 225–250 (2006)
6. Dupuis, P., Ellis, R.S.: A Weak Convergence Approach to the Theory of Large Deviations. Wiley, New York (1997)
7. Dupuis, P., Wang, H.: Importance sampling, large deviations, and differential games. Stoch. Stoch. Rep. **76**, 481–508 (2004)
8. Dupuis, P., Wang, H.: Subsolutions of an Isaacs equation and efficient schemes for importance sampling. Math. Oper. Res. **32**, 1–35 (2007)
9. Dupuis, P., Ishii, H., Soner, H.M.: A viscosity solution approach to the asymptotic analysis of queueing systems. Ann. Probab. **18**, 226–255 (1990)
10. Dupuis, P., Leder, K., Wang, H.: Large deviations and importance sampling for a tandem network with slow-down. QUESTA **57**, 71–83 (2007)
11. Dupuis, P., Sezer, A., Wang, H.: Dynamic importance sampling for queueing networks. Ann. Appl. Probab. **17**, 1306–1346 (2007)
12. Fleming, W.H., Soner, H.M.: Controlled Markov Processes and Viscosity Solutions. Springer, New York (1992)
13. Freidlin, M.I., Wentzell, A.D.: Random Perturbations of Dynamical Systems. Springer, New York (1984)
14. Glasserman, P., Juneja, S.: Uniformly efficient importance sampling for the tail distribution of sums of random variables. Math. Oper. Res. **33**, 36–50 (2008)
15. Glasserman, P., Kou, S.: Analysis of an importance sampling estimator for tandem queues. ACM Trans. Model. Comput. Simul. **4**, 22–42 (1995)
16. Glasserman, P., Wang, Y.: Counter examples in importance sampling for large deviations probabilities. Ann. Appl. Probab. **7**, 731–746 (1997)
17. Glynn, P.W., Iglehart, D.L.: Simulation methods for queues: an overview. Queueing Syst. Theory Appl. **3**, 221–256 (1988)
18. Heidelberger, P.: Fast simulation of rare events in queueing and reliability models. ACM Trans. Model. Comput. Simul. **4**, 43–85 (1995)
19. Juneja, S., Nicola, V.: Efficient simulation of buffer overflow probabilities in Jackson networks with feedback. ACM Trans. Model. Comput. Simul. **15**, 281–315 (2005)

20. Juneja, S., Shahabuddin, P.: Rare-event simulation techniques: an introduction and recent advances. In: Henderson, S., Nelson, B. (eds.) Handbook on Simulation, pp. 291–350. Elsevier, Amsterdam (2006)
21. Majewski, K., Ramanan, K.: How large queues build up in Jackson networks. Math. Oper. Res. (2009, to appear)
22. Meyn, S.P., Down, D.: Stability of generalized Jackson networks. Ann. Appl. Probab. **4**, 124–148 (1994)
23. Parekh, S., Walrand, J.: A quick simulation method for excessive backlogs in networks of queues. IEEE. Trans. Autom. Control **34**, 54–66 (1989)
24. Sadowsky, J.S.: Large deviations and efficient simulation of excessive backlogs in a $GI/G/m$ queue. IEEE. Trans. Autom. Control **36**, 1383–1394 (1991)
25. Sezer, A.D.: Asymptotically optimal importance sampling for a Jackson network with a tree topology. Preprint (2008)
26. Sezer, A.D.: Importance sampling for a Markov modulated queuing network. Stoch. Process. Appl. **119**, 491–517 (2009)