Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

State-dependent importance sampling for a Jackson tandem network

D.I. Miretskiy, W.R.W. Scheinhardt, M.R.H. Mandjes

REPORT PNA-R0807 APRIL 2008

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

Copyright © 2008, Stichting Centrum voor Wiskunde en Informatica
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

State-dependent importance sampling for a Jackson tandem network

ABSTRACT
This paper considers importance sampling as a tool for rare-event simulation. The focus is on estimating the probability of overflow in the downstream queue of a Jacksonian two-node tandem queue – it is known that in this setting 'traditional' stateindependent importance-sampling distributions perform poorly. We therefore concentrate on developing a state-dependent change of measure, that we prove to be asymptotically efficient. More specific contributions are the following. (i)We concentrate on the probability of the second queue exceeding a certain predefined threshold before the system empties. Importantly, we identify an asymptotically efficient importance-sampling distribution for any initial state of the system. (ii) The choice of the importance-sampling distribution is backed up by appealing heuristics that are rooted in large-deviations theory. (iii) Our method for proving asymptotic efficiency is substantially more straightforward than some that have been used earlier. The paper is concluded by simulation experiments that show a considerable speed up.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: importance sampling, tandem queue
Note: Part of this research has been funded by the Dutch Bsik/BRICKS project.

State-dependent Importance Sampling for a Jackson Tandem Network*

D.I. Miretskiy† W.R.W. Scheinhardt‡ M.R.H. Mandjes§

April 14, 2008

Abstract

This paper considers importance sampling as a tool for rare-event simulation. The focus is on estimating the probability of overflow in the downstream queue of a Jacksonian two-node tandem queue – it is known that in this setting 'traditional' state-independent importance-sampling distributions perform poorly. We therefore concentrate on developing a state-dependent change of measure, that we prove to be asymptotically efficient.

More specific contributions are the following. (i) We concentrate on the probability of the second queue exceeding a certain predefined threshold before the system empties. Importantly, we identify an asymptotically efficient importance-sampling distribution for *any* initial state of the system. (ii) The choice of the importance-sampling distribution is backed up by appealing heuristics that are rooted in large-deviations theory. (iii) Our method for proving asymptotic efficiency is substantially more straightforward than some that have been used earlier. The paper is concluded by simulation experiments that show a considerable speed up.

*Part of this research has been funded by the Dutch BSIK/BRICKS project.

†Corresponding author. Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands.

‡Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands. WS is also with CWI, Amsterdam, the Netherlands.

§Korteweg-de Vries Institute for Mathematics; The University of Amsterdam; Plantage Muidergracht 24; 1018 TV Amsterdam; The Netherlands. MM is also with CWI, Amsterdam, the Netherlands and EURANDOM, Eindhoven, the Netherlands.

1 Introduction

Rare event analysis of queueing networks has been attracting continuous and growing attention over the past decades. As explicit expressions are hardly available, one usually relies on asymptotic techniques to approximate small overflow probabilities. These asymptotics, however, often lack error bounds, and consequently it is not always clear whether their use is justified for given parameters. This explains why one often opts for simulation methods instead.

The use of simulation for estimating rare event probabilities has an inherent problem: the event under consideration occurs so rarely during the simulation, that it is extremely time consuming to obtain a reliable estimate; a rule of thumb is that the number of occurrences needed to obtain an estimate of a certain predefined accuracy is inversely proportional to the probability of interest. Perhaps the most prominent remedy to this problem is *importance sampling* (IS), i.e., simulating the system under a *new* probability measure, and correcting the simulation output by means of likelihood ratios (which essentially capture the likelihood of the realization under the old measure with respect to the new measure) to retain unbiasedness. Evidently, it makes sense to choose an IS distribution which guarantees frequent occurrence of the event of interest. The choice of a 'good' new measure is rather delicate though. It should be chosen such that the above-mentioned likelihood ratio tends to be small on the event of interest; choosing a 'wrong' new measure, one may even end up with an estimator with infinite variance. We refer to, e.g., Heidelberger [9] for more background on IS and its pitfalls.

'Classical' papers on the use of IS in queueing usually rely on a so-called 'state-independent' change of measure, i.e., for any state in the system the probabilistic law is changed in the same manner. Usually, large deviations techniques are used to motivate the choice of the new measure, and to prove that the resulting estimator has specific desirable properties (such as bounds on the likelihood on the event of interest). In this respect we mention the seminal paper by Parekh and Walrand [14], that focuses on the estimation of the probability of overflow in a single queue, but also on the probability of the *total* queue population in a network reaching some threshold. The new measure then corresponds to an unstable queueing system; for instance in the case of a single M/M/1 system the arrival and service rates should be swapped. A fundamental treatment of this change of measure, in fact even for the multi-server queue GI/GI/m (where it is tacitly assumed that the service times are light-tailed), was given by Sadowsky [15]. His main result is that the corresponding estimator is *asymptotically efficient* (or: asymptotically optimal), which effectively means that the variance of the estimator behaves roughly like the square of its first moment. In a setting in which the overflow probability decays exponentially in the buffer size B, asymptotic efficiency means that the number of replications needed to obtain an estimator with fixed relative error grows *subexponentially* fast with the 'rarity parameter' B.

Things complicate tremendously when looking at networks rather than one-node systems. For the Jacksonian two-node tandem queue (that is, Poisson arrivals, exponential service times at both queues), aiming at estimating the probability that the network population

exceeds a given threshold, [14] proposed to swap the arrival rate with the rate of the *slowest* server – this makes, heuristically, sense, as the slowest server corresponds to the bottleneck queue. In this case experimental results were not so encouraging as in the case of a single queue, and the quality of the simulation results was strongly affected by the specific values of the arrival and service rates. Later it was proved that this method is asymptotically efficient for some parameter values, but has unbounded variance for other values, see [8] and [2]. In fact, it was proven that *no* state-independent change of measure exists that is asymptotically efficient for all parameter values.
It was realized that the main problem of state-independent IS schemes is that the transition rates are changed in a 'uniform manner', i.e., irrespective of whether one of the queues is empty or not. As a result it cannot be guaranteed that the likelihood ratio is bounded on the event of interest, and therefore the IS scheme proposed in [14] performs poorly for some parameter values. Some of the first attempts to solve this problem can be found in [3] and [11], in which *state-dependent* IS schemes were proposed, i.e., IS distributions that are not uniform over the state space. Dupuis *et al.* [7] were the first to prove asymptotic efficiency for a state-dependent IS scheme for estimating overflow probabilities in a d-node Jackson network.

Several important questions are, however, still open; let us from now on concentrate on the two-node Jackson tandem network. In the first place, the majority of papers on this type of networks deals with the probability that, starting in a situation with both queues empty, the total network population exceeds a certain threshold. One may wonder, though, what the impact of the starting state is on the IS scheme. Also, it is not *a priori* clear how to change the simulation procedures if one is interested in the event of overflow in a specific queue (rather than the total queue).
The main topic of the present paper concerns the development of an asymptotically efficient IS algorithm for estimating the probability that the content *of the downstream queue* exceeds a certain threshold B before the system becomes empty, *starting in any initial state*, say $x \in \mathbb{N} \times \{0, \ldots, B-1\}$.
The search for an appropriate change of measure greatly benefits from powerful large-deviations based heuristics. We express the decay rate of the probability of our interest in terms of so-called 'cost functions', that assign cost to paths; the 'most likely path' is then defined as the 'cheapest' path from state x to the 'overflow set' $\mathbb{N} \times \{B, B+1, \ldots\}$ (that does not visit the origin). The intuition is that, conditional on the event that the second queue indeed reaches B before the system gets empty, the trajectory of the Markov process will be typically close to this most likely path. Then the idea is that knowledge of the most likely path helps in finding a good change of measure. The shape of the most likely path strongly depends on which of the two queues is the bottleneck (i.e., has the lowest service rate). When it comes to proving asymptotic efficiency, the two cases have to be dealt with differently. We remark that the most likely path can have a rather unexpected shape; there are situations that, starting in a state x in which the second queue is non-empty, this path is such that first the second queue becomes empty while the first queue fills (to end up in some state $(y, 0)$), and then the first queue drains while the second queue

builds up. Another interesting observation is that the most likely path is *not* continuous in the starting state x: two nearly identical initial states can reach the 'overflow set' in an entirely different manner. We also mention that a non-trivial technical issue we deal with is the *infinite* state space, in that the process can attain any value in $\mathbb{N} \times \{0, \ldots, B-1\}$, cf. [11]; this complication does not play a role when analyzing rare-event probabilities related to the *total* network population.
We expect that the above-mentioned large-deviations heuristic can be rather helpful when analyzing a broad class of networks; see also earlier results in [13] for the model that was introduced in [17], in which the service rate of the first queue depends on the content of the second queue.
The proof technique is essentially based on that of Dupuis *et al.* [7], but, as in De Boer and Scheinhardt [4], we have managed to simplify the proofs considerably. The change of measure is such that the most likely path is, roughly, followed (that is, with high probability), with corrections for the regions near the axes. The proof of asymptotic efficiency then relies on bounding the likelihood on the event of interest.

We end this section by detailing the structure of the paper. Model and preliminaries, as well as a short overview on the basics of IS, are presented in Section 2. In Section 3 we heuristically construct a state-dependent IS scheme for estimating the probability of our interest; interesting corollary results are (i) the most likely path, and (ii) the corresponding decay rate. In Section 4 we present a number of large-deviations properties of the system, where the main result is the correctness of the decay rate that was heuristically derived in the previous section. Section 5 shows that our IS scheme, after a minor adaptation that deals with visits to the axes, is indeed asymptotically efficient. We conclude this paper with supporting numerical results in Section 6, and conclusions in Section 7.

2 Model and preliminaries

We consider a two-node tandem Jackson network with Poisson arrivals at rate λ to the first station. Each job receives service at the first station, after which it is routed to the second station. After receiving service at the second station, the job leaves the system. Service times at station i have an exponential distribution with parameter μ_i, $i = 1, 2$. The waiting rooms at both stations are assumed to be infinitely large.
Let $Q(t) = \{(Q_1(t), Q_2(t)), t \geq 0\}$ be the joint queue-length process, as in [7] and [4], from which we will borrow some more notation. Then it is clear that this is a continuous-time Markov process, with possible jump directions $v_0 = (1, 0)$, $v_1 = (-1, 1)$ and $v_2 = (0, -1)$ with corresponding transition rates λ, μ_1 and μ_2 respectively. The process $Q(t)$ is regenerative if we impose the stability condition $\lambda < \min(\mu_1, \mu_2)$, which we will do from now on.
The queue-length process can also be described by the *embedded* discrete time Markov chain $Q_j = (Q_{1,j}, Q_{2,j})$, where $Q_{i,j}$ is the number of jobs in queue i after the j-th transition. Without loss of generality we will choose the parameters such that $\lambda + \mu_1 + \mu_2 = 1$, so that they also represent the *transition probabilities* of Q_j in the interior of the state space. To

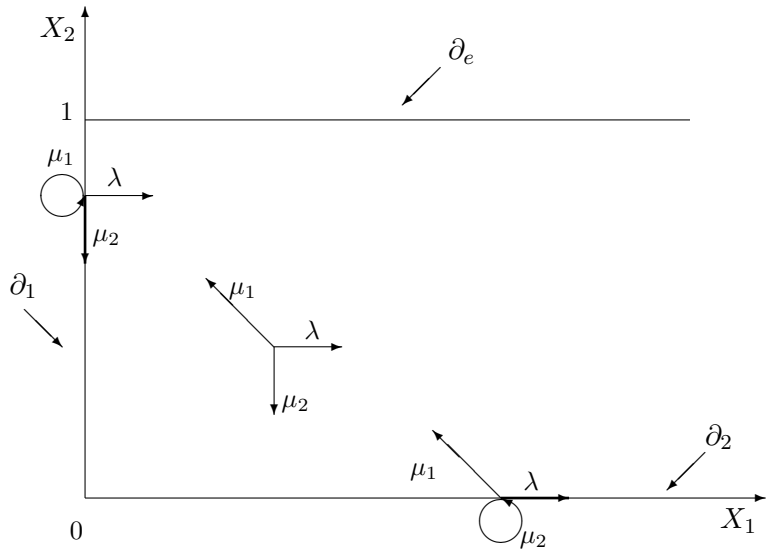

Figure 1: State space and transition structure for scaled process $X(t)$.

ensure that the same holds on the boundaries, we shall introduce socalled self-transitions shortly, see below.

Our main interest is to estimate the probability that $Q(t)$ (or equivalently, Q_j) reaches some high level B in the second buffer before it returns to the origin, starting from any state. Thereto, it will be convenient to also consider the scaled processes $X(t) = Q(Bt)/B$ (in continuous time) and $X_j = Q_j/B$ (in discrete time). The advantage of these scalings is the invariance of the state space for any B. In particular, our target probability is equivalent to the probability that the second component of either the scaled process X_j or the scaled process $X(t)$ reaches 1 before the process returns to the origin.

We introduce the following subsets of the state space

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

and denote the state space by $\bar{D} = D \cup \partial_e \cup \partial_1 \cup \partial_2$ (realize that we can exclude $x_2 > 1$ from the state space). Note that transition v_k is impossible when queue k is empty, i.e., when $X_j \in \partial_k$. We modify the process X_j to deal with this by allowing some self-transitions in the following way, see also Figure 1:

$$\mathbb{P}(X_{j+1} = X_j | X_j \in \partial_k) = \mu_k, \text{ for } k = 1, 2. \tag{1}$$

Next, we introduce the stopping time τ_B^x, which is the first time that the process X_j hits level 1, starting from state $x = (x_1, x_2)$, without visits to the origin:

$$\tau_B^x = \inf\{k > 0 : X_k \in \partial_e, X_j \neq 0 \text{ for } j = 1, \ldots, k-1\}, \tag{2}$$

and we define $\tau_B^x = \infty$ if X_j hits the origin before ∂_e. It will also be convenient to let $I_B(A^x)$ be the indicator of the event $\tau_B^x < \infty$ for the path $A^x = (X_j, j = 0, \ldots : X_0 = x)$.

Thus we can write the probability of our interest as

$$p_B^x = \mathbb{E} I_B(A^x) = \mathbb{P}(\tau_B^x < \infty). \tag{3}$$

It is clear that it is not efficient to estimate p_B^x via straightforward simulations when B is large, due to the rarity of the event of interest. In order to reduce the simulation time we will employ Importance Sampling (IS), i.e., we perform simulations under a new measure $\mathbb{Q}$, which replaces the transition rates corresponding to v_0, v_1, v_2 by other values. In particular, we will use a *state-dependent* IS scheme.
This means that the transition rates under the new measure $\mathbb{Q}$ may depend on the current state x of the process; they will be denoted by $\bar{\lambda}(x)$, $\bar{\mu}_1(x)$ and $\bar{\mu}_2(x)$ respectively.
The probability p_B^x can now also be expressed as

$$p_B^x = \mathbb{E}^{\mathbb{Q}}[L(A^x) I_B(A^x)], \tag{4}$$

where $L(A^x)$ is the likelihood ratio (also known as Radon-Nikodym derivative) of the path A^x. It is given by

$$L(A^x) = \prod_{j=0}^{\tau_B^x - 1} \frac{\mathbb{P}(Y_j)}{\mathbb{Q}(Y_j | X_j)}, \tag{5}$$

where $Y_j = B(X_{j+1} - X_j)$, unless $X_{j+1} = X_j$ in which case $Y_j = v_k$, if $X_j \in \partial_k$. Furthermore, $\mathbb{P}(Y_j)$ is the stochastic kernel of the scaled process X_j under the old measure, being equal to λ, μ_1 or μ_2 if $j = 0, 1, 2$, respectively, and $\mathbb{Q}(Y_j | X_j)$ is the kernel under the new measure, given by $\bar{\lambda}(x)$, $\bar{\mu}_1(x)$ or $\bar{\mu}_2(x)$ when the current state is $X_j = x$.

Definition 2.1. *The IS scheme for p_B^x is called asymptotically efficient if*

$$\liminf_{B \to \infty} \frac{\log \mathbb{E}^{\mathbb{Q}}[L^2(A^x) I_B(A^x)]}{\log \mathbb{E}^{\mathbb{Q}}[L(A^x) I_B(A^x)]} \geq 2. \tag{6}$$

In our case it is known that p_B^x decays exponentially in B, so that the *exponential decay rate* is well defined, i.e.,

$$\lim_{B \to \infty} \frac{1}{B} \log p_B^x \ \in \ (0, \infty).$$

As a result, (6) can be rewritten in the following form:

$$\limsup_{B \to \infty} \frac{1}{B} \log \mathbb{E}[L(A^x) I_B(A^x)] \leq 2 \lim_{B \to \infty} \frac{1}{B} \log p_B^x.$$

3 Optimal path and related change of measure

In order to find a good change of measure for IS simulations, the first step is usually to find the most probable path to overflow, i.e., the way in which overflow most probably occurs, conditional on its occurrence. In Section 3.1 we explain a method in which minimizing certain 'cost-functions' leads to the most probable path and a good corresponding change

of measure, given by new (state-dependent) transition rates $\tilde{\lambda}(x), \tilde{\mu}_1(x)$ and $\tilde{\mu}_2(x)$. Then, we split the problem, since the minimization procedure gives different results in different cases. In Section 3.2 we treat the case $\lambda < \mu_2 < \mu_1$, in which the second server is the bottleneck, while Section 3.3 deals with the case $\lambda < \mu_1 \leq \mu_2$, in which the first server is the bottleneck. Beforehand, we would like to point out that the change of measure mentioned above, denoted by tildes, is not the same as the asymptotically efficient change of measure that will be introduced in Section 5 (denoted by bars), although it is closely related.

3.1 Cost and structure of path to overflow

The typical path to overflow in the particular case that the origin is the starting point, has already been identified for the d-node Jackson tandem network in [1], and hence also for our tandem system. In that paper, the time-reversed process is used to find the shape of the most probable path to overflow. This path to overflow was also obtained as a corollary result in [13], and in this section we present a method similar to the one in [13] to find the optimal path starting from *any* state $x \in \bar{D}$. The advantage of this method is that it also provides a 'good' change of measure, which ensures that most simulation runs under this new measure will be close to the optimal path. This new measure will be the basis for another change of measure, which is used in our (state-dependent) IS scheme, as presented in Section 5. Another result of our method is the exponential decay rate of p_B^x, which will be determined in Section 4, and which will play a crucial role in the proofs of asymptotic efficiency of Section 5.
Before introducing our method we impose some restrictions on the path structures we consider, leaving the proof that the typical path to overflow indeed satsifies these restrictions to Section 4, see Lemma 4.4. We will only consider the following paths.

Property 3.1.

- Each path is a concatenation of subpaths, which are straight lines on any of the subsets D, δ_1 and δ_2, and the new measure stays constant along each subpath, i.e., $\tilde{\lambda}(x) = \tilde{\lambda}$, $\tilde{\mu}_1(x) = \tilde{\mu}_1$ and $\tilde{\mu}_2(x) = \tilde{\mu}_2$, for any state x on the same subpath;
- Each path does not have more than one subpath in each subset if $\mu_2 < \mu_1$;
- Each path does not have more than two subpaths in each subset if $\mu_2 \geq \mu_1$.

With every path that satisfies Property 3.1 we associate a 'cost', the main idea (to be proved in Section 4) being that the minimal cost of the path to overflow in the second buffer, starting from state x, can be interpreted as the decay rate of the probability of interest. Our method is based on the family of cost functions I, defined by

$$I(\tilde{\lambda} \mid \lambda) := \lambda - \tilde{\lambda} + \tilde{\lambda} \log \frac{\tilde{\lambda}}{\lambda}, \tag{7}$$

see also [16, pages 14 and 20]. Note that the function (7) is convex and equals 0 at $\tilde{\lambda} = \lambda$. Intuitively, we can think of the value $I(\tilde{\lambda} \mid \lambda)$ as the cost we need to pay to let a Poisson process with parameter λ behave like a Poisson process with parameter $\tilde{\lambda}$, per time unit.

We will now explain our cost method in more detail in the following two examples. More background can be found in the Appendix of [13].

Example 3.2. As an example, consider a straight path through the interior of the state space, staying away from the boundaries, from some state x to another state y, where $x_1 \geq y_1$ and $x_2 < y_2$. We then need to construct a new measure $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$, such that $\tilde{\mu}_1 > \tilde{\mu}_2$ and $\tilde{\lambda} \leq \tilde{\mu}_1$. This measure ensures that our path has constant north-west drift, or in other words, due to the scaling, our path has a constant slope

$$\alpha = \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\lambda} - \tilde{\mu}_1}. \tag{8}$$

The total cost of such a path, per unit time is

$$\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2) := I(\tilde{\lambda} \mid \lambda) + I(\tilde{\mu}_1 \mid \mu_1) + I(\tilde{\mu}_2 \mid \mu_2). \tag{9}$$

To find the cost per unit horizontal (vertical) distance, we need to divide this by the horizontal speed $\tilde{\lambda} - \tilde{\mu}_1$ (vertical speed $\tilde{\mu}_1 - \tilde{\mu}_2$). Thus, minimizing the cost of any straight path from x to y in this case boils down to minimizing

$$(y_2 - x_2)\frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\mu}_1 - \tilde{\mu}_2}, \tag{10}$$

over $\tilde{\mu}_1$ and $\tilde{\mu}_2$, such that $\tilde{\lambda} \leq \tilde{\mu}_1$ and $\tilde{\mu}_1 > \tilde{\mu}_2$ hold, as well as

$$\tilde{\lambda} = \tilde{\mu}_1 + \frac{y_1 - x_1}{y_2 - x_2}(\tilde{\mu}_1 - \tilde{\mu}_2);$$

in addition, we should have that

$$\frac{y_2 - x_2}{\tilde{\mu}_1 - \tilde{\mu}_2} = \frac{y_1 - x_1}{\tilde{\lambda} - \tilde{\mu}_1}$$

to guarantee that y is indeed the ending state of the path when it starts at x.
It is easily checked that the total cost (10) with ending state $y = (0, 1)$ attains its minimum when triplet $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ is a solution to

$$\begin{cases} \tilde{\lambda} = \tilde{\mu}_1 - \frac{x_1}{1-x_2}(\tilde{\mu}_1 - \tilde{\mu}_2) \\ \tilde{\lambda} + \tilde{\mu}_1 + \tilde{\mu}_2 = \lambda + \mu_1 + \mu_2 \\ \tilde{\lambda}\tilde{\mu}_1\tilde{\mu}_2 = \lambda\mu_1\mu_2 \\ \tilde{\lambda} \leq \tilde{\mu}_1 \text{ and } \tilde{\mu}_1 > \tilde{\mu}_2 \\ \tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2 > 0. \end{cases} \tag{11}$$

The reason we have chosen the specific ending state $(0, 1)$ is that it is the most frequent ending state for our network. Notice also that if $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ is the solution to (11) for some starting state x, it also minimizes this system if we replace x by any state that belongs to the straight line between x and $y = (0, 1)$. ♢

Example 3.3. Let us now give an example for another type of path with starting state $x \in D$ and ending state $(0,1)$, consisting of two (straight) subpaths. The first subpath belongs to the interior and has north-west drift. The second part belongs to the vertical boundary and has north drift. Thus, it may be denoted as $(x_1, x_2) \to (0, x_2 + \alpha^{-1}x_1) \to (0,1)$, for same slope α. Property 3.1 tells us that the new measure stays constant along each subpath, so the total cost of such a path is

$$\alpha^{-1}x_1 \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\mu}_1 - \tilde{\mu}_2} + (1 - x_2 - \alpha^{-1}x_1)\frac{\mathbb{I}(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2)}{\hat{\lambda} - \hat{\mu}_2},$$

where $\alpha = (\tilde{\mu}_1 - \tilde{\mu}_2)/(\tilde{\lambda} - \tilde{\mu}_1)$, see (8). The first term in the sum is the cost of the first subpath under some new measure $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and the second term is the cost of the second (vertical) subpath under some measure $(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2)$. Optimizing this expression such that $\tilde{\lambda} < \tilde{\mu}_1$, $\tilde{\mu}_2 < \tilde{\mu}_1$, $\hat{\lambda} \leq \hat{\mu}_1$ and $\hat{\mu}_2 < \hat{\mu}_1$, for the case $\mu_2 < \mu_1$, over all parameters marked with tildes and hats, it is readily verified that the minimal cost of this path type is obtained when the new measure is given by

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2) = (\mu_2, \mu_1, \lambda),$$

i.e., by simply interchanging the arrival rate λ and the service rate of the second station μ_2 for both subpaths. $\diamondsuit$

By considering all possible path types we obtain the overal minimum cost, corresponding to the most probable path, and the corresponding (state-dependent) change of measure $\tilde{\lambda}$, $\tilde{\mu}_1$ and $\tilde{\mu}_2$. Finally, we also have

$$\gamma_x := \text{ minimal cost over all paths } x \to \delta_e,$$

at our disposal. In Theorem 4.1 we will prove that this is in fact the exponential decay rate of the probability p_B^x as $B \to \infty$.
We now present the results of our minimum-cost-path method for both cases of the tandem network.

3.2 Optimal path results for $\lambda < \mu_2 < \mu_1$

When $\mu_2 < \mu_1$, the cost minimization starting in state x as outlined in the previous section (in particular Example 3.3; see also the Appendix in [13] for more examples), yields the following new measure after some calculations:

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_2, \mu_1, \lambda), & \text{if } x \in A_1, \\ \text{solution to (11)}, & \text{if } x \in A_2, \\ (\lambda, \mu_1, \mu_2), & \text{if } x \in A_3 \end{cases} \tag{12}$$

Here A_i, $i = 1, 2, 3$, is the following partition of the state space $\bar{D}$, see also Figure 2:

$$\begin{aligned} A_1 &:= \{x \in \bar{D} : x_2 \leq -x_1/\alpha_1 + 1\}, \\ A_2 &:= \{x \in \bar{D} : -x_1/\alpha_1 + 1 < x_2 < -\alpha_1 x_1 + 1\}, \\ A_3 &:= \{x \in \bar{D} : x_2 \geq -\alpha_1 x_1 + 1\}, \end{aligned} \tag{13}$$

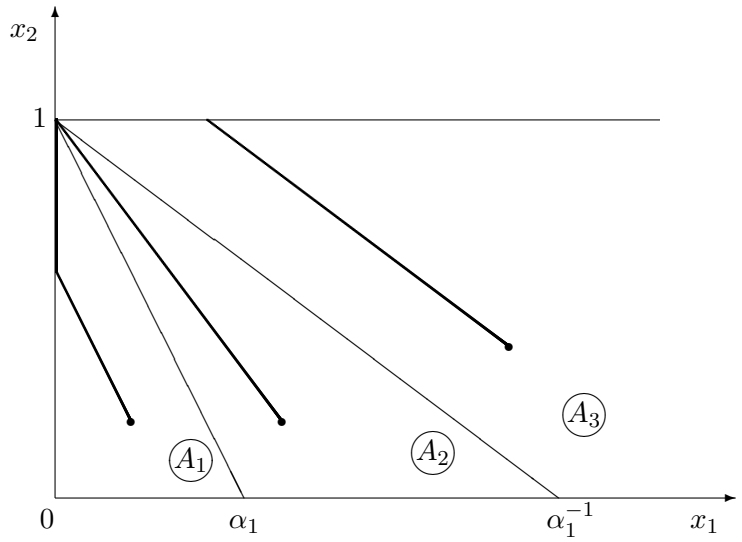

Figure 2: Partition of $\bar{D}$ and some optimal paths to overflow when $\mu_2 < \mu_1$.

with $\alpha_1 := (\mu_1 - \mu_2)/(\mu_1 - \lambda)$. Note that the path considered in Example 3.3 in the previous subsection is optimal for any starting state $x \in A_1$, and the corresponding new measure (exchanging λ and μ_2) was earlier found by Parekh and Walrand [14] for the problem of reaching a large total queue population. Also, we point out that the change of measure is continuous in the state x, as can be verified by solving system (11) for $x = (\alpha_1, 0)$ and $x = (\alpha_1^{-1}, 0)$, yielding the solutions in the first and third lines of (12), respectively.
The corresponding path from starting state $x = (x_1, x_2)$ to some state on ∂_e is given by

$$\begin{array}{ll} (x_1, x_2) \to (0, x_2 + \alpha_1^{-1} x_1) \to (0, 1), & \text{if } x \in A_1, \\ (x_1, x_2) \to (0, 1), & \text{if } x \in A_2, \\ (x_1, x_2) \to (x_1 - \alpha_1^{-1} x_2, 1), & \text{if } x \in A_3. \end{array} \tag{14}$$

The resulting cost γ_x of the optimal path is given by:

$$\gamma_x = \begin{cases} (1 - x_1 - x_2)\gamma, & \text{if } x \in A_1, \\ -x_1 \log \frac{\tilde{\lambda}(x)}{\lambda} - (1 - x_2) \log \frac{\tilde{\mu}_2(x)}{\mu_2}, & \text{if } x \in A_2, \\ 0, & \text{if } x \in A_3, \end{cases} \tag{15}$$

where

$$\gamma := -\log \frac{\lambda}{\mu_2},$$

is the minimal cost of the path $(0, 0) \to (0, 1)$.
We end this subsection with some interesting properties of the new measure defined in (12), to be used later. Intuitively, it says that for any state x, the new measure 'lies between' the Parekh and Walrand measure where λ and μ_2 are interchanged, and the 'normal' measure, where the parameters retain their original values. Moreover, the more jobs are present in the system at time zero, either in queue 1 or in queue 2, the 'less change of measure' we need. This perfectly coincides with the structure of the most probable path, see (14).

Lemma 3.4. *When $\mu_2 < \mu_1$, the functions $\tilde{\lambda}(x)$, $\tilde{\mu}_1(x)$ and $\tilde{\mu}_2(x)$ as defined in (12) are continuous and differentiable, satisfying the following for any $x \in \bar{D}$.*

(i) $\dfrac{\partial\tilde{\lambda}(x)}{\partial x_1} \leq 0,\ \dfrac{\partial\tilde{\mu}_2(x)}{\partial x_1} \geq 0,\ \dfrac{\partial\tilde{\lambda}(x)}{\partial x_2} \leq 0 \quad$ *and* $\quad \dfrac{\partial\tilde{\mu}_2(x)}{\partial x_2} \geq 0.$

(ii) $\tilde{\lambda}(x) \in [\lambda, \mu_2]$ *and* $\tilde{\mu}_2(x) \in [\lambda, \mu_2]$.

(iii) $\gamma = \max_{x\in\bar{D}} \gamma_x$.

Proof. (i) We only need to consider $x \in A_2$, since otherwise all partial derivatives are zero. Applying implicit differentiation to (11) one finds

$$\frac{\partial\tilde{\lambda}(x)}{\partial x_1} = -\frac{(1-x_2)(\tilde{\mu}_1(x)-\tilde{\mu}_2(x))\tilde{\lambda}(x)}{(1-x_2)^2\tilde{\lambda}(x) + (1-x_1-x_2)^2\tilde{\mu}_1(x) + x_1^2\tilde{\mu}_2(x)} \leq 0,$$

where the last inequality follows from the fact that $\tilde{\mu}_1(x) > \tilde{\mu}_2(x)$. The other statements follow similarly.
(ii) It follows from (12) that $\tilde{\lambda}(x) = \mu_2$ if $x \in \partial_1$ and $\tilde{\lambda}(x) = \lambda$ if $x \in A_3$, so applying the first statement of this lemma one can find that $\tilde{\lambda}(x) \in [\lambda, \mu_2]$. Using similar arguments one can obtain the same bounds for $\tilde{\mu}_2(x)$.
(iii) We show that the partial derivatives with respect to x_1 and x_2 of γ_x as given in (15) are not positive. For $x \in A_1 \cup A_3$ this is obvious, while for $x \in A_2$ it can be checked using implicit differentiation, similar to the proof of the first statement. □

Lemma 3.4 does not yield results on $\tilde{\mu}_1(x)$ since they are not needed in the sequel, but it may be interesting to note that $\tilde{\mu}_1(x)$ is *not* monotone. In fact, $\tilde{\mu}_1(x) = \mu$ when $x \in A_1$ and when $x \in A_3$, but also when $x_1 + x_2 = 1$, so it is also neither convex nor concave.

3.3 Optimal path results for $\lambda < \mu_1 \leq \mu_2$

The new measure under which the path to overflow has minimal cost in terms of (7) is as follows:

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_1, \lambda, \mu_2), & \text{if } x \in B_1, \\ \text{solution to (11)}, & \text{if } x \in B_2, \\ (\lambda, \mu_2, \mu_1), & \text{if } x \in B_3. \end{cases} \tag{16}$$

Again we partitioned the state space into three subspaces B_i, $i = 1, 2, 3$ as follows, see also Figure 3.

$$\begin{aligned} B_1 &:= \{x \in \bar{D} : f(x) \leq 0\}, \\ B_2 &:= \{x \in \bar{D} : f(x) > 0 \text{ and } x_2 < -\alpha_2 x_1 + 1\}, \\ B_3 &:= \{x \in \bar{D} : x_2 \geq -\alpha_2 x_1 + 1\}, \end{aligned} \tag{17}$$

where $\alpha_2 := (\mu_2 - \mu_1)/(\mu_2 - \lambda)$ and

$$f(x) := \gamma + x_1 \log\frac{\tilde{\lambda}(x)}{\mu_1} + (1-x_2)\log\frac{\tilde{\mu}_2(x)}{\mu_2},$$

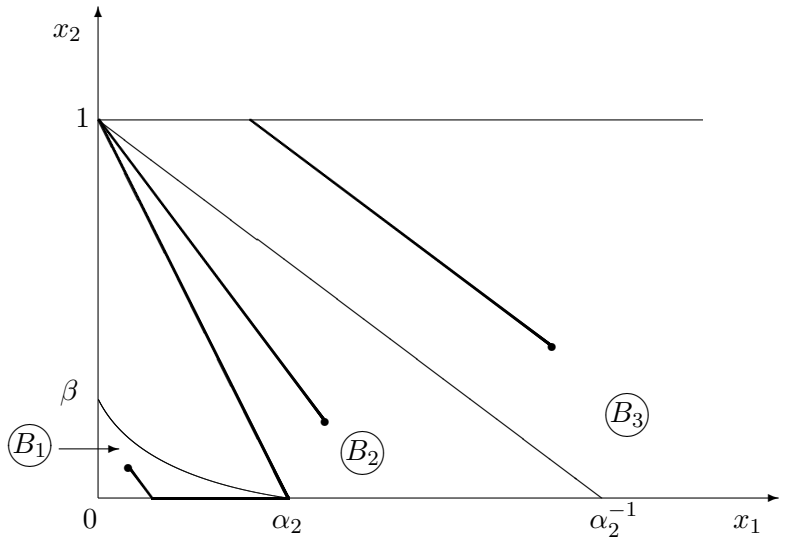

Figure 3: Partition of $\bar{D}$ and some optimal paths to overflow when $\mu_1 \leq \mu_2$.

with $\tilde{\lambda} \equiv \tilde{\lambda}(x)$ and $\tilde{\mu}_2 \equiv \tilde{\mu}_2(x)$ being the solution to (11). The zero level curve of the function $f(x)$ represents the boundary between subspaces B_1 and B_2, β is the unique solution to $f(0, x_2) = 0$. Interestingly, for the current case the change of measure is *not* continuous in states x that lie on this boundary (i.e., $f(x) = 0$), and the behavior on B_1 and B_2 is entirely different. In particular, the change of measure on B_2 has $\tilde{\lambda}(x) < \tilde{\mu}_1(x)$ and $\tilde{\mu}_2(x) < \tilde{\mu}_1(x)$, as opposed to the first line of (16) where both inequalities are reversed. This is also reflected in a different shape of the typical path from $x = (x_1, x_2)$ to ∂_e:

$$\begin{array}{ll} (x_1, x_2) \to (x_1 + \alpha_3 x_2, 0) \to (\alpha_2, 0) \to (0, 1), & \text{if } x \in B_1, \\ (x_1, x_2) \to (0, 1), & \text{if } x \in B_2, \\ (x_1, x_2) \to (x_1 - \alpha_2^{-1} x_2, 1), & \text{if } x \in B_3, \end{array} \tag{18}$$

where $\alpha_3 := (\mu_2 - \lambda)/(\mu_1 - \lambda)$. Note that the last part of any path with starting state $x \in B_1$ is just a special case of a path starting in B_2 (in this case starting in $(\alpha_2, 0)$), but the corresponding new measure on this line (i.e. the solution to system (11) for $x = (\alpha_2, 0)$) can be given explicitly as (μ_1, μ_2, λ). This was already known from [13] for the path starting in the origin.

The next result we give is γ_x, the cost of the optimal path in terms of (7):

$$\gamma_x = \begin{cases} \gamma - x_1 \log \frac{\mu_1}{\lambda}, & \text{if } x \in B_1, \\ -x_1 \log \frac{\tilde{\lambda}(x)}{\lambda} - (1 - x_2) \log \frac{\tilde{\mu}_2(x)}{\mu_2}, & \text{if } x \in B_2, \\ (1 - x_2) \log \frac{\mu_2}{\mu_1}, & \text{if } x \in B_3. \end{cases} \tag{19}$$

We end this subsection with the analogue of Lemma 3.4 in the case when $\mu_1 \leq \mu_2$. For this, we first introduce z as the unique solution in the interval $(0, 1)$ of the (essentially cubic) equation

$$\varphi(z) := \lambda + \mu_1 + \mu_2(1 - z) - \sqrt{\frac{\lambda \mu_1}{z}} = 0, \tag{20}$$

which follows from system (11) by taking $(x_1, x_2) = (0, 0)$. (The fact that there is a unique solution immediately follows from $\varphi(0) = -\infty$, $\varphi(1) = \lambda + \mu_1 - \sqrt{\lambda\mu_1} > 0$, and the fact that $\varphi'(z) = 0$ has just a single positive solution, viz. $\sqrt[3]{\lambda\mu_1/4\mu_2^2}$.) In fact, $-\log z$ is the cost of the vertical path $(0, 0) \to (0, 1)$ *in the interior* (i.e., in D), satisfying $\tilde{\lambda} = \tilde{\mu}_1$ (as opposed to the vertical path *following* ∂_1 in Example 3.3, where $\tilde{\lambda} < \tilde{\mu}_1$). See also [12, Eqns. (30) and (33)] and [13] for more details.

Lemma 3.5. *When $\mu_1 \leq \mu_2$, the functions $\tilde{\lambda}(x)$, $\tilde{\mu}_1(x)$ and $\tilde{\mu}_2(x)$ as defined in (16) are continuous and differentiable, except in states x with $f(x) = 0$. For any such $x \in \bar{D}$, these functions satisfy the following.*

(i) $\dfrac{\partial\tilde{\lambda}(x)}{\partial x_1} \leq 0,\ \dfrac{\partial\tilde{\mu}_2(x)}{\partial x_1} \geq 0,\ \dfrac{\partial\tilde{\lambda}(x)}{\partial x_2} \leq 0, \quad$ *and* $\quad \dfrac{\partial\tilde{\mu}_2(x)}{\partial x_2} \geq 0, \quad .$

(ii) $\tilde{\lambda}(x) \in [\lambda, \sqrt{\lambda\mu_1/z}]$ *and* $\tilde{\mu}_2(x) \in [\mu_2 z, \mu_1] \cup \mu_2$, *where z is defined by (20).*

(iii) $\gamma = \max_{x\in\bar{D}} \gamma_x$.

Proof. To prove these facts we can use similar arguments as in the proof of Lemma 3.4. An exception is the second statement, where the upper bound for $\tilde{\lambda}(x)$ and the lower bound for $\tilde{\mu}_2(x)$ is attained when $x \in \partial_1$ (see the first statement), i.e., $\tilde{\lambda}(x) \leq \sqrt{\lambda\mu_1/z}$ and $\tilde{\mu}_2(x) \geq \mu_2 z$, where z is the unique solution of (20) in the interval $(0, 1)$. □

Note that part *(i)* of Lemma 3.5 implies that the functions are monotone on $B_2 \cup B_3$, but not on all $\bar{D}$, due to the discontinuity on the boundary between B_1 and B_2.

4 Large deviations properties

The goal of this section is to formally prove that the cost of the optimal path to overflow is equal to the exponential decay rate of p_B^x, the probability of interest. We also illuminate some important and interesting large deviations properties of the process $X(t)$.
Consider any absolutely continuous function $\phi : [0, \infty) \to \bar{D}$, representing a path associated with the scaled process $X(t)$. Our first aim is to define a so-called local rate function $\ell(\phi(t), \dot{\phi}(t))$, which depends both on the position at time t and on the time derivative (or speed vector) $\dot{\phi}(t)$ at time t. To do so, we first define three auxiliary functions $L_i(y)$, where the argument y should be interpreted as a speed vector:

$$L_i(y) := \sup_\theta \left(\langle\theta, y\rangle - g_i(\theta)\right), \quad i = 0, 1, 2, \tag{21}$$

where

$$\begin{aligned} g_0(\theta) &:= \lambda(e^{\theta_1} - 1) + \mu_1(e^{\theta_2-\theta_1} - 1) + \mu_2(e^{-\theta_2} - 1), \\ g_1(\theta) &:= \lambda(e^{\theta_1} - 1) + \mu_2(e^{-\theta_2} - 1), \\ g_2(\theta) &:= \lambda(e^{\theta_1} - 1) + \mu_1(e^{\theta_2-\theta_1} - 1), \end{aligned}$$

cf. [16, Eqn. (5.5)]. The second equality applies to ∂_1 and the third equality applies to ∂_2. The function $g_1(\theta)$ does not have a term with μ_1, because jumps of type v_1 from boundary

∂_1 are impossible, and likewise $g_2(\theta)$ does not have a term with μ_2. Finally let us define the local rate function ℓ as:

$$\ell(\phi(t),\dot{\phi}(t)) := \begin{cases} L_0(\dot{\phi}(t)), & \text{if } \phi(t) \in D \cup \partial_e, \\ [L_0 \oplus L_1\,](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_1, \\ [L_0 \oplus L_2](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_2, \end{cases} \tag{22}$$

where

$$[L_0 \oplus L_i](y) := \inf\{\rho L_0(y_0) + (1-\rho)L_i(y_i) : 0 \le \rho \le 1, \rho y_0 + (1-\rho)y_i = y\},$$

for $i = 1, 2$, is the inf-convolution of the functions L_0 and L_i, the infimum being taken over all values ρ and vectors y_0 and y_i that satisfy the given conditions. Let us briefly explain why we use this inf-convolution on the boundaries of the state space. Assume that the scaled process $X(t)$ follows a path $\phi(t) \in \partial_1$, such that $\partial\phi_2/\partial t > 0$ for $t \in [0, T]$. Hence, the first and second component of the vector y should be zero and strictly positive, respectively. It is clear that the original (unscaled) jump process $Q(t)$ can only increase its second component when it is not on ∂_1, since jumps of type v_1 are not allowed on ∂_1. Therefore, the inf-convolution provides a 'mixture' of the functions L_0 and L_1, supposing that the process $Q(t)$ spends a fraction of time ρ in the interior D and a fraction $1-\rho$ on the vertical constraint. Note that ρ must be such that $\phi(t)$ has speed y with positive increment in the vertical direction and zero-increment in the horizontal direction, such that the scaled process $X(t)$ remains on ∂_1.
We are now ready to state the following theorem.

Theorem 4.1. *The process $X(t)$ satisfies a large deviations principle with local rate function (22), i.e.,*

$$\lim_{B\to\infty} \frac{1}{B} \log p_B^x = -\inf \int_0^\tau \ell(\phi(t),\dot{\phi}(t))dt,$$

where $\tau := \inf\{t > 0 : \phi(t) \in \partial_e, \phi(s) \neq 0, s \in (0,t)\}$ and the infimum is taken over all absolutely continuous functions $\phi : [0;\infty) \to \bar{D}$ such that $\phi(0) = x$ and $\tau < \infty$.

Proof. The proof of this theorem is based on the results presented in [5]. Let us introduce a process $Z(t)$, which is the unconstrained version of $X(t)$, in other words $Z(t)$ is allowed to have negative values in both components. In addition we will assume that $Z(0) = X(0) = x \in \bar{D}$. One can use [5, Thms. 3.2 and 3.4] to show that the map $\Gamma : Z(t) \to X(t)$ exists and Theorem 2.2 from the same paper to show that it is Lipschitz continuous. Γ is known as the Skorokhod map and the question whether it exists is known as the Skorokhod problem; for more background we refer to [5].
Since the map Γ is Lipschitz continuous and the process $Z(t)$ satisfies a large deviation principle, see [16, Thm. 5.1], one can apply the contraction principle (see [16, Thm. 2.13]) and conclude that the process of our interest, $X(t)$, satisfies a large deviations principle with local rate function $\ell(\phi(t),\dot{\phi}(t))$ defined by (22). □

Using the local rate function ℓ, as defined in (22), we can define the rate function of any path $\phi(t) = (\phi_1(t), \phi_2(t))$ with $t \in [0, T]$ for some T, as the integral of ℓ over time. The following lemma shows that for paths that stay in one of the subsets $D, \partial_1, \partial_2$, this rate function is minimal when the path is straight, with constant speed vector.

Lemma 4.2. *For any T, consider an absolutely continuous path $\phi(t)$ that remains in D (or in ∂_1, or in ∂_2) for all $t \in [0, T]$. Then,*

$$\int_0^T \ell(\phi(t), \dot{\phi}(t))dt \geq T\ell\left(\phi(0), \frac{\phi(T) - \phi(0)}{T}\right).$$

Equality holds only if $\dot{\phi}(t)$ is a constant, i.e., $\phi(t)$ is a straight line.

Proof. The proof of this lemma can be found in [16, p. 87]; we mention that a related result was established in [13, Lemma 4]. □

Now assume that $\phi(t) \in D$, for $t \in (0, T)$ is a path between two states x and y. Lemma 4.2 tells us that the path $\phi(t)$ has minimal cost if the process $X(t)$ moves along a straight line at constant speed. We can define a corresponding new measure as follows

$$\begin{aligned} \tilde{\lambda} &= \lambda e^{\theta_1}, \\ \tilde{\mu}_1 &= \mu_1 e^{\theta_2 - \theta_1}, \\ \tilde{\mu}_2 &= \mu_2 e^{-\theta_2}, \end{aligned} \tag{23}$$

where $\theta = (\theta_1, \theta_2)$ is the maximizer of (21) with $i = 1$. In fact this is exactly the same change of measure we would find using the cost minimization procedure from Section 3, due to the immediate equality

$$\ell(\phi(t), \dot{\phi}(t)) = \mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2). \tag{24}$$

This equality however, does not hold on the boundaries. Instead, when $\phi(t)$ stays on ∂_1 or ∂_2 for $t \in [0, T]$, we have

$$\ell(\phi(t), \dot{\phi}(t)) \leq \mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2),$$

where the new measure $\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ is again defined as in (23). This is not difficult to see since, e.g. for paths on ∂_1, we find for some $\rho \in [0, 1]$ that

$$\ell(\phi(t), \dot{\phi}(t)) = I(\tilde{\lambda}|\lambda) + \rho I(\tilde{\mu}_1|\mu_1) + I(\tilde{\mu}_2|\mu_2) \leq \mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2).$$

However, we can still show equality between local rate functions and cost functions on the boundaries, but only for the optimal paths. To state this formally, let Φ_1 (Φ_2) be the set of paths that travels a distance $h > 0$ along ∂_1 (∂_2) at constant speed during a time σ_1 (σ_2), i.e.,

$$\begin{aligned} \Phi_1 &= \{\phi(t) \subset \partial_1 : \phi(0) = (0, x_2^*), \phi(\sigma_1) = (0, x_2^* + h)\}, \\ \Phi_2 &= \{\phi(t) \subset \partial_2 : \phi(0) = (x_1^*, 0), \phi(\sigma_2) = (x_1^* + h, 0)\}, \end{aligned}$$

for some x_1^* and x_2^*. Then we have the following relations between the rate function ℓ defined by (22) and the cost function $\mathbb{I}$ from the previous section, defined by (9):

Lemma 4.3. *(i) For paths in the interior D, we have*

$$\ell(\phi(t), \dot{\phi}(t)) = \mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2).$$

(ii) For paths on the vertical boundary ∂_1, we have

$$\inf_{\phi \in \Phi_1} \int_0^{\sigma_1} \ell(\phi(t), \dot{\phi}(t)) dt = h \inf \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\mu}_1 - \tilde{\mu}_2},$$

where the second infimum is taken over all $\tilde{\lambda}$, $\tilde{\mu}_1$ and $\tilde{\mu}_2$ such that $\tilde{\lambda} < \tilde{\mu}_1$ and $\tilde{\mu}_1 > \tilde{\mu}_2$.
(iii) For paths on the horizontal boundary ∂_2, we have

$$\inf_{\phi \in \Phi_2} \int_0^{\sigma_2} \ell(\phi(t), \dot{\phi}(t)) dt = h \inf \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_1},$$

where the second infimum is taken over all $\tilde{\lambda}$, $\tilde{\mu}_1$ and $\tilde{\mu}_2$ such that $\tilde{\lambda} > \tilde{\mu}_1$ and $\tilde{\mu}_2 > \tilde{\mu}_1$.

Proof. Statement *(i)* is the same as (24). We continue to prove statement *(iii)*, the proof of *(ii)* being identical. The restriction $\phi(t) \subset \partial_2$ implies that $\dot{\phi}(t) = (\dot{\phi}_1(t), 0)$ for any $t \in [0, \sigma_2]$. The definition of the inf-convolution tell us that $\tilde{\lambda} v_0 + \tilde{\mu}_1 v_1 + \rho \tilde{\mu}_2 v_2 = \dot{\phi}(t)$. Hence we find that $\rho = \tilde{\mu}_1 / \tilde{\mu}_2$ and $\dot{\phi}_1(t) = \tilde{\lambda} - \tilde{\mu}_1$, from which we can conclude that

$$\inf \int_0^{\sigma_2} \ell(\phi(t), \dot{\phi}(t)) dt = h \inf \frac{I(\tilde{\lambda}|\lambda) + I(\tilde{\mu}_1|\mu_1) + (\tilde{\mu}_1/\tilde{\mu}_2) I(\tilde{\mu}_2|\mu_2)}{\tilde{\lambda} - \tilde{\mu}_1}.$$

Straightforward minimization shows that the latter equals $h \inf\{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)/(\tilde{\lambda} - \tilde{\mu}_1)\}$, because $\tilde{\mu}_2 = \mu_2$ and hence $I(\tilde{\mu}_2|\mu_2) = 0$. □

The next lemma validates our choice in the previous section to consider only paths that satisfy Property 3.1.

Lemma 4.4. *The optimal path from any starting state x to ∂_e does not have more than*

- *one subpath in each subset, if $\mu_2 < \mu_1$,*
- *two subpaths in each subset, if $\mu_2 \geq \mu_1$.*

Proof. Due to the one-to-one correspondence between the rate function of any path in terms of the local rate function ℓ and the cost function $\mathbb{I}$, see Lemma 4.3, the proof of this lemma is similar to the proof of Lemma 5 in [13]. □

Theorem 4.5. *The exponential decay rate of p_B^x equals the minimal cost derived in Section 3, i.e.,*

$$\lim_{B \to \infty} \frac{1}{B} \log p_B^x = -\gamma_x.$$

Proof. By virtue of Lemma 4.4, the optimal path can be represented as a concatenation of k (at most 6) subpaths $\phi^{(1)}, \ldots, \phi^{(k)}$, which stay on different subsets of the state space (i.e., D, ∂_1 and ∂_2). If we denote the starting time and position of the ith subpath by $t^{(i)}$

and $x^{(i)} = \phi^{(i)}(t^{(i)}) = \phi^{(i-1)}(t^{(i)})$ respectively, with the convention that $t^{(1)} = 0, x^{(1)} = x$, $t^{(k+1)} = \tau, x^{(k+1)} \in \partial_e$, then we can write the decay rate identified in Theorem 4.1 as

$$\inf_{\phi} \int_0^{\tau} \ell(\phi(t), \dot{\phi}(t))dt = \inf_{t^{(1)},\ldots,t^{(k)}} \sum_{i=1}^{k} \inf_{\phi^{(i)}} \int_{t^{(i)}}^{t^{(i+1)}} \ell(\phi^{(i)}(t), \dot{\phi}^{(i)}(t))dt.$$

Using Lemma 4.3 we can rewrite the last expression as follows:

$$\inf_{t^{(1)},\ldots,t^{(k)}} \sum_{i=1}^{k} \left(x_1^{(i+1)} - x_1^{(i)}\right) \inf_{\tilde{\lambda},\tilde{\mu}_1,\tilde{\mu}_2} \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_1},$$

where in some of the terms we may need to replace the denominator $\tilde{\lambda} - \tilde{\mu}_1$ by $\tilde{\mu}_1 - \tilde{\mu}_2$, while also changing the prefactor to $x_2^{(i+1)} - x_2^{(i)}$, see Lemma 4.3. Using the fact that the new measure $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ determines the shape of each subpath $\phi^{(i)}$, and the shape of the whole path ϕ as a consequence, we may also take the first infimum over $x^{(1)}, \ldots, x^{(k)}$, rather than $t^{(1)}, \ldots, t^{(k)}$. Applying Lemma 4.3 to the last optimization problem we arrive at

$$\inf_{x^{(1)},\ldots,x^{(k)}} \sum_{i=1}^{k} \left([\phi_{i+1}(t_{i+1})]_1 - [\phi_i(t_i)]_1\right) \inf_{\tilde{\lambda},\tilde{\mu}_1,\tilde{\mu}_2} \frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_1} = \gamma_x,$$

which completes the proof. □

5 Asymptotic efficiency

It is known from [13], where the starting state is the origin, that the new measures (12) and (16) are not always asymptotically efficient. For example, when $\mu_2 < \mu_1$, multiple visits of the process $Q(t)$ to the horizontal axis (∂_2) under the new measure (μ_2, μ_1, λ) may cause the likelihood ratio to become very large. We will 'protect' the likelihood ratio by using a specific measure around ∂_2, under which these visits become harmless. This approach is similar to the one used in [7]. We will also introduce a protection strip along the lower part of the vertical boundary ∂_1 in the same manner, in the case when $\mu_1 \leq \mu_2$.
We again split the problem into two cases: in Section 5.1 we explain our method in detail for the situation in which the second server is the bottleneck ($\lambda < \mu_2 < \mu_1$), and in Section 5.2 we treat the case in which the first server is the bottleneck ($\lambda < \mu_1 \leq \mu_2$).

5.1 Asymptotically efficient scheme for $\mu_2 < \mu_1$

In order to construct an IS scheme that is provably asymptotically efficient we introduce a function $W(x)$, defined for any point $x = (x_1, x_2)$ of the state space. This function will give us an expression for a new measure $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ in the same manner as it was done in [7]. Let us first introduce three intermediate functions $W_i(x), i = 1, 2, 3$:

$$\begin{aligned} W_1(x) &:= 2\gamma_x - \delta, \\ W_2(x) &:= W_1(x_1, \delta/2\gamma) = 2\gamma_{(x_1,\delta/2\gamma)} - \delta, \\ W_3(x) &:= 2\gamma - 3\delta, \end{aligned} \tag{25}$$

where δ is some small positive number, and γ_x is given by (15). In the next step we introduce the function which is the minimum of these three functions, see also Figure 4:

$$\bar{W}(x) := W_1(x) \wedge W_2(x) \wedge W_3(x).$$

Note that our particular choice of the functions W_i ensures that the shapes of the areas around the origin and ∂_2 on which $\bar{W}$ coincides with the functions W_i are the same as they were in [7]. The last step in the construction is a mollification procedure which makes the

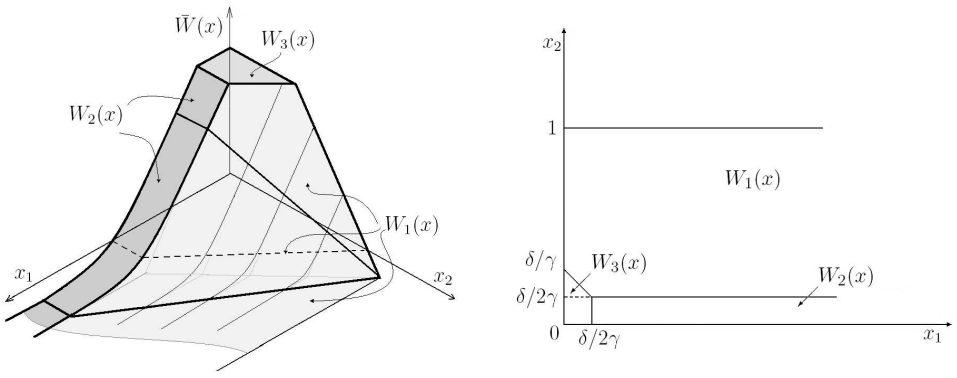

Figure 4: The function $\bar{W}(x)$ and the areas on which $\bar{W}(x) = W_i, i = 1, 2, 3$ (case $\mu_2 < \mu_1$).

resulting function $W(x)$ smooth. We do this by defining:

$$W(x) := -\epsilon \log \sum_{i=1}^{3} e^{-W_i(x)/\epsilon}, \tag{26}$$

where ϵ is a 'smoothness' parameter; the larger ϵ is chosen, the smoother the function $W(x)$ is. On the other hand, as $\epsilon \to 0$ we see that $W(x)$ converges to the (non-smooth) function $\bar{W}(x)$.
The function $W(x)$, and in particular its gradient, will play a main role in the representation of the state-dependent, asymptotically efficient new measure. However, before turning to this, we need some preliminaries, namely a relation between the gradients of the functions W_i and the measure from the previous sections, and some assumptions on the parameters δ and ϵ.

Proposition 5.1. *The gradients of the functions $W_i(x)$, $i = 1, 2, 3$ can be represented as follows:*

$$\begin{aligned} DW_1(x) &= 2\left(\log \frac{\lambda}{\tilde{\lambda}(x)}, \log \frac{\tilde{\mu}_2(x)}{\mu_2}\right), \\ DW_2(x) &= 2\left(\log \frac{\lambda}{\tilde{\lambda}(x_1, \delta/2\gamma)}, 0\right), \\ DW_3(x) &= (0, 0). \end{aligned}$$

Proof. It is clear that $DW_1(x) = -2\gamma(1,1)$ if $x \in A_1$. When $x \in A_2$, $DW_1(x)$ can be represented in the following form:

$$\begin{aligned} DW_1(x) &= 2\left(\log\frac{\lambda}{\tilde{\lambda}(x)}, \log\frac{\tilde{\mu}_2(x)}{\mu_2}\right) - 2x_1\left(\frac{\partial\tilde{\lambda}(x)/\partial x_1}{\tilde{\lambda}(x)}, \frac{\partial\tilde{\lambda}(x)/\partial x_2}{\tilde{\lambda}(x)}\right) \\ &\quad - 2(1-x_2)\left(\frac{\partial\tilde{\mu}_2(x)/\partial x_1}{\tilde{\mu}_2(x)}, \frac{\partial\tilde{\mu}_2(x)/\partial x_2}{\tilde{\mu}_2(x)}\right). \end{aligned}$$

Although we do not know $\tilde{\lambda}(x)$, $\tilde{\mu}_1(x)$ and $\tilde{\mu}_2(x)$ explicitly, we can find their partial derivatives with respect to x_1 and x_2 by implicit differentiation of (11), see Lemma 3.4 for more insight. After some elementary algebra we find that the sum of the last two vectors in the previous expression equals zero, which proves the first statement. The other two statements follow easily from the definitions of W_2 and W_3. □

The parameters δ and ϵ depend on B, and in the sequel we will need the following conditions for their asymptotical behavior as B grows large. Note that these are the same conditions as in [7] and [4].

Assumption 5.2. *The parameters $\delta \equiv \delta_B$ and $\epsilon \equiv \epsilon_B$ are strictly positive and satisfy the following limit conditions: as $B \to \infty$, (i) $\epsilon_B \to 0$, (ii) $\delta_B \to 0$, (iii) $B\epsilon_B \to \infty$, (iv) $\varepsilon_B/\delta_B \to 0$.*

We will now show how the new measure is constructed from the function W. We inherit the following expression from [7, Prop. 3.2]:

$$\begin{aligned} \bar{\lambda}(p) &= N(p)\lambda e^{-\langle p, v_0\rangle/2}, \\ \bar{\mu}_1(p) &= N(p)\mu_1 e^{-\langle p, v_1\rangle/2}, \\ \bar{\mu}_2(p) &= N(p)\mu_2 e^{-\langle p, v_2\rangle/2}, \end{aligned} \tag{27}$$

where

$$N(p) := \left[\lambda e^{-\langle p, v_0\rangle/2} + \mu_1 e^{-\langle p, v_1\rangle/2} + \mu_2 e^{-\langle p, v_2\rangle/2}\right]^{-1} = e^{\mathbb{H}(p)/2}. \tag{28}$$

Here $\mathbb{H}(p)$ is a function known as the *Hamiltonian*, which we use to simplify the notation and to enable the comparison with [7] and [4]. The vector p strongly depends on the current state of the process and is in fact taken to be the gradient $DW(x)$. We thus rewrite (27) as

$$\begin{aligned} &\bar{\lambda}(x) = \lambda e^{-\langle DW(x), v_0\rangle/2} e^{\mathbb{H}(DW(x))/2}, \\ &\bar{\mu}_i(x) = \mu_i e^{-\langle DW(x), v_i\rangle/2} e^{\mathbb{H}(DW(x))/2}, \quad i = 1, 2. \end{aligned} \tag{29}$$

We like to mention that we can express the gradient $DW(x)$ as a weighted sum of vectors $DW_k(x)$ at point x:

$$DW(x) = \sum_{k=1}^{3} \rho_k(x) DW_k(x), \text{ where } \rho_k(x) = \frac{e^{-W_k(x)/\varepsilon}}{\sum_{i=1}^{3} e^{-W_i(x)/\varepsilon}} \tag{30}$$

For the Hamiltonian we have the following results.

Lemma 5.3. *For any x, $\mathbb{H}(DW_1(x)) = \mathbb{H}(DW_3(x)) = 0$ and $\mathbb{H}(DW_2(x)) \geq 0$.*

Proof. The first and second claims are due to a direct computation:

$$\begin{aligned}\mathbb{H}(DW_1(x)) &= 2\log N(DW_1(x)) \\ &= -2\log\left[\lambda e^{-\log(\lambda/\tilde{\lambda}(x))} + \mu_1 e^{-\log(\tilde{\mu}_2(x)/\mu_2)+\log(\lambda/\tilde{\lambda}(x))} + \mu_2 e^{\log(\tilde{\mu}_2(x)/\mu_2)}\right] \\ &= -2\log\left[\tilde{\lambda}(x) + \tilde{\mu}_1(x) + \tilde{\mu}_2(x)\right] = 0\end{aligned}$$

and

$$\mathbb{H}(DW_3(x)) = -2\log\left[\lambda + \mu_1 + \mu_2\right] = 0.$$

Finally, for the third case we have

$$\begin{aligned}\mathbb{H}(DW_2(x)) &= 2\log N(DW_2(x)) \\ &= -2\log\left[\lambda e^{-\log(\lambda/\tilde{\lambda}(x_1,\delta/2\gamma))} + \mu_1 e^{\log(\lambda/\tilde{\lambda}(x_1,\delta/2\gamma))} + \mu_2 e^{0}\right] \\ &= -2\log\left[\tilde{\lambda}(x_1,\delta/2\gamma) + \mu_1\frac{\lambda}{\tilde{\lambda}(x_1,\delta/2\gamma)} + \mu_2\right].\end{aligned}$$

To study the argument of the last logarithm, we consider the function $\psi(x) := x + \lambda\mu_1/x + \mu_2$, for which we have $\psi(\lambda) = \psi(\mu_1) = 1$. Also, $\psi(x)$ is convex, so that for all possible values of $\tilde{\lambda}(x_1, \delta/2\gamma)$ in $[\lambda, \mu_2] \subset [\lambda, \mu_1]$, one can conclude that $\psi(\tilde{\lambda}(x_1, \delta/2\gamma)) \leq 1$. This proves the last statement of this lemma. □

Clearly there is a difference between the new measures defined in Section 3 (indicated by tildes) and in this section (indicated by bars). In fact it is not difficult to see that the first one also follows from (27) if we replace W by W_1. However, this change of measure is not asymptotically efficient, while the other one is, due to the protection strips along the boundaries, as we will prove in the remainder of this subsection. We start with some lemmas that are similar to the ones in [4].

Lemma 5.4. *The likelihood $L(A)$ of a path $A = (X_j, j = 0, \ldots, \sigma)$ under the new measure (29) satisfies*

$$\begin{aligned}\log L(A) &= \frac{B}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1} - X_j\rangle \\ &+ \sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), v_k\rangle I\{X_j = X_{j+1} \in \partial_k\} \qquad (31) \\ &- \frac{1}{2}\sum_{j=0}^{\sigma-1}\mathbb{H}(DW(X_j)).\end{aligned}$$

Proof. The proof is the same as that of Lemma 1 in [4]. □

Lemma 5.5. *Consider the case $\mu_2 < \mu_1$. For any path $A = (X_j, j = 0, ..., \sigma)$ under the new measure (29), the first term in (31) satisfies*

$$\left| \frac{B}{2} \sum_{j=0}^{\sigma-1} \langle DW(X_j), X_{j+1} - X_j \rangle - \frac{B}{2}(W(X_\sigma) - W(X_0)) \right| \leq \frac{C}{B\varepsilon}\sigma,$$

for sufficiently large $B\varepsilon$, where C is some positive constant.

Proof. At first let us introduce the following representation

$$W(x+y) = W(x) + \langle DW(x), y \rangle + \frac{1}{2} y^T H(x) y + |y|^2 r(y),$$

where $y = X_{j+1} - X_j$ is a one-step increment of the scaled process X_j, the matrix $H(x)$ is the Hessian matrix of the function $W(x)$ and the function $r(y)$ satisfies $\lim_{|y|\to 0} r(y) = 0$. After transferring two terms to the left hand side and taking the absolute value we find

$$\begin{aligned} |\,(W(x+y) - W(x)) - \langle DW(x), y \rangle| &= \left| \frac{1}{2} y^T H(x) y + |y|^2 r(y) \right| \\ &\leq \frac{1}{2} |y^T| \cdot ||H(x)||_2 \cdot |y| + |y|^2 \, |r(y)| \\ &\leq |y|^2 ||H(x)||_{\max} + |y|^2 \, |r(y)|, \end{aligned}$$

where $||H(x)||_{\max}$ is the maximum norm of the Hessian matrix, given by

$$||H(x)||_{\max} = \max\{h_{11}(x), h_{12}(x), h_{22}(x)\};$$

here

$$h_{11}(x) := \left| \frac{\partial^2 W(x)}{\partial x_1^2} \right|, \;\; h_{12}(x) := \left| \frac{\partial^2 W(x)}{\partial x_1 \partial x_2} \right|, \;\; h_{22}(x) := \left| \frac{\partial^2 W(x)}{\partial x_2^2} \right|.$$

We now compute an upper bound for $|h_{11}(x)|$ as an example; the two other terms can be dealt with in the same manner. Using representation (30) one can write

$$\frac{\partial^2 W(x)}{\partial x_1^2} = \sum_{k=1}^{3} \left[\rho_k(x) \frac{\partial^2 W_k(x)}{\partial x_1^2} + \frac{\partial \rho_k(x)}{\partial x_1} \cdot \frac{\partial W_k(x)}{\partial x_1} \right], \tag{32}$$

where it follows from the definition of $\rho_k(x)$ that

$$\frac{\partial \rho_k(x)}{\partial x_1} = -\frac{1}{\varepsilon} \frac{\rho_k(x) \sum_{i \neq k} e^{-W_i(x)/\varepsilon} \left(\frac{\partial W_k(x)}{\partial x_1} - \frac{\partial W_i(x)}{\partial x_1} \right)}{\sum_i e^{-W_i(x)/\varepsilon}}.$$

Since the second fraction turns out to be bounded as $\varepsilon \to 0$, and the same holds for the other terms in (32), we find that some positive constant C_1 exists, such that

$$\left| \frac{\partial^2 W(x)}{\partial x_1^2} \right| < \frac{C_1}{\varepsilon}.$$

Due to similar bounds for the other second-order partial derivatives, and the simple fact that $|y| \leq \sqrt{2}/B$, we have for some positive constant C_2 that

$$|y|^2 ||H(x)||_{\max} \leq \frac{C_2}{B^{2\varepsilon}}.$$

Finally, if we choose B large enough (and hence $|y|$ small), we have for some positive constant C_3 that

$$|y|^2 |r(y)| \leq \frac{C_3}{B^2}.$$

The statement of the lemma is a direct consequence of these two bounds. □

Lemma 5.6. *Consider a two-node tandem Jackson network. For any sequence θ_B such that $\theta_B \to 0$ $(B \to \infty)$, and τ_B^x defined by (2), the following limit holds:*

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}(\mathrm{e}^{\theta_B \tau_B^x} | I_B(A^x) = 1) = 0.$$

Proof. Let us first give a sketch of the proof. This proof consist of three steps: *(i)* we bound the length of any path from ∂_e to the origin; *(ii)* using time reversibility arguments we show that the same bound applies to τ_B^0; *(iii)* we show that the path of our interest is shorter (in stochastic sense) than $(\tau_B^0 | I_B(A^0) = 1)$.
(i) Let σ_B be the length of the path from any state (B, α) to the origin, for any finite α; and let ω^B be the length of the path from any state (x_1, x_2), such that $x_1 + x_2 = B$. It is clear that

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E} e^{\theta_B \sigma_B} = \lim_{B\to\infty} \frac{1}{B+\alpha} \log \mathbb{E} e^{\theta_B \omega^{B+\alpha}} = \lim_{B\to\infty} \frac{1}{B} \log \mathbb{E} e^{\theta_B \omega^B} = 0, \tag{33}$$

where the last equality follows from the third statement of [4, Lemma 3].
(ii) Now consider the time-reversed network, see [10, Thm. 1.12]. It is not difficult to check that this is also a tandem queue, but with the first and second queue interchanged. The length of a path in the original system from the origin to level B in the second queue, without visits to the origin in the mean time, equals the length of a path from some state (B, α) to the origin in the reversed system, given that it does not visit any state $(B, \cdot)$ in between, hence

$$\left(\tau_B^0 | I_B(A^0) = 1\right) \leq_{\mathrm{st}} \sigma_B.$$

Combining the last statement with (33) we have

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}(e^{\theta_B \tau_B^0} | I_B(A^0) = 1) \leq \lim_{B\to\infty} \frac{1}{B} \log \mathbb{E} e^{\theta_B \sigma_B} = 0, \tag{34}$$

which is similar to the fourth statement of [4, Lemma 3], only there the exit boundary ∂_e was different.
(iii) We use stochastic coupling to show that

$$\left(\tau_B^x | I_B(A^0) = 1\right) \leq_{st} \left(\tau_B^0 | I_B(A^0) = 1\right). \tag{35}$$

To see this, we consider the ('original') process starting in the origin, and couple it to a similar process starting in state x. Then the above states that the time to overflow for the original process, given that overflow happens before reaching the origin is stochastically larger than the time to overflow for the coupled process, given that *the original process* reaches overflow before the origin. Notice that the condition implies that also the coupled process reaches overflow before the origin (since the queue lengths cannot be negative). In other words, for any path with $I_B(A^0) = 1$, also $I_B(A^x) = 1$ must hold, but since the opposite does not hold in general, we have $\{I_B(A^0) = 1\} \subset \{I_B(A^x) = 1\}$. From this we can conclude

$$(\tau_B^x | I_B(A^x) = 1) \leq_{st} \left(\tau_B^x | I_B(A^0) = 1\right). \tag{36}$$

Using (34), (35) and (36) we can now write that for any state $x \in \bar{D}$,

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}(\mathrm{e}^{\theta_B \tau_B^x} | I_B(A^x) = 1) \leq \lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}(\mathrm{e}^{\theta_B \tau_B^0} | I_B(A^0) = 1) = 0, \tag{37}$$

which completes the proof of the lemma. □

Theorem 5.7. *When $\mu_2 < \mu_1$ and Assumption 5.2 holds, the new measure in (29), with function W based on (15), is asymptotically efficient.*

Proof. We will roughly follow the proof of [4, Thm. 1], paying attention to some important issues. First note that Lemma 5.3 provides an upper bound on the last term of the log-likelihood expression in Lemma 5.4:

$$-\frac{1}{2} \sum_{j=0}^{\tau_B^x - 1} \mathbb{H}(DW(X_j)) \leq 0. \tag{38}$$

In order to bound the second term in Lemma 5.4 we will prove a result similar to the third statement of [7, Lemma B.1].
From Proposition 5.1 we know that $\langle DW_2(x), -v_2\rangle = \langle DW_3(x), -v_2\rangle = 0$ and also that $\langle DW_1(x), -v_2\rangle = 2\log(\tilde{\mu}_2(x)/\mu_2)$. Hence, applying (30), we have

$$\langle DW(x), -v_2\rangle = 2\log\left(\frac{\tilde{\mu}_2(x)}{\mu_2}\right)\rho_1(x) \geq 2\log\left(\frac{\tilde{\mu}_2(x)}{\mu_2}\right) e^{-(W_1(x)-W_2(x))/\varepsilon}. \tag{39}$$

It is clear that $W_1(x) - W_2(x) = \delta$ for any $x \in A_1 \cap \partial_2$, see (25). Also, Lemma 3.4 guarantees that $W_1(x) - W_2(x)$ decreases to 0 as x moves along the horizontal axis from $(\alpha_1, 0)$ to $(\alpha_1^{-1}, 0)$. This immediately leads to $0 \leq W_1(x) - W_2(x) \leq \delta$ for any $x \in \partial_2$. Now keeping in mind that $\tilde{\mu}_2(x) \geq \lambda$ (see again Lemma 3.4) and hence $\log(\tilde{\mu}_2(x)/\mu_2) \geq -\gamma$, we can write

$$\langle DW(x), -v_2\rangle \geq -2\gamma e^{-\frac{\delta}{\varepsilon}}.$$

Using the same technique and keeping Lemma 3.4 in mind one can also show that

$$\langle DW(x), -v_1\rangle \geq -2\gamma e^{-\frac{\delta}{\varepsilon}},$$

for any $x \in \partial_1$. Using these two inequalities we obtain the same bound for the second term in Lemma 5.4 as in [4]:

$$\sum_{k=1}^{2} \frac{1}{2} \sum_{j=0}^{\tau_B^x - 1} \langle DW(X_j), v_k \rangle I\{X_j = X_{j+1} \in \partial_k\} \leq \gamma e^{-\delta/\varepsilon} \tau_B^x. \tag{40}$$

To deal with the first term in Lemma 5.4, we first bound $W(x)$ using (26):

$$W(x) \leq -\varepsilon \log\left(e^{W_1(x)/\varepsilon}\right) = -\varepsilon \log\left(e^{(-2\gamma_x+\delta)/(\varepsilon)}\right) = 2\gamma_x - \delta$$

and, using that $W_2(x) \geq W_1(x) - \delta$ and the monotonicity of γ_x,

$$\begin{aligned} W(x) &\geq -\varepsilon \log\left(e^{-W_1(x)/\varepsilon} + e^{(-W_1(x)+\delta)/\varepsilon} + e^{-W_3(x)/\varepsilon}\right) \\ &\geq -\varepsilon \log\left(3e^{(-2\gamma_x+3\delta)/\varepsilon}\right) = 2\gamma_x - \varepsilon \log(3) - 3\delta. \end{aligned}$$

Using the same technique we obtain similar bounds for $W(X_{\tau_B^x})$:

$$-\varepsilon \log(3) - 3\delta \leq W(X_{\tau_B^x}) \leq -\delta.$$

Using the three last inequalities and Lemma 5.5 we can derive an upper bound for the first term in Lemma 5.4,

$$\frac{B}{2} \sum_{j=0}^{\tau_B^x - 1} \langle DW(X_j), X_{j+1} - X_j \rangle \leq \frac{B}{2}(-2\gamma_x + \eta(B)) + \frac{C}{B\varepsilon}\tau_B^x, \tag{41}$$

where $\eta(B)$ is such that $\lim_{B\to\infty} \eta(B) = 0$.
Combining (38), (40) and (41) we can rewrite (31) in the following way

$$\log(L(A)) \leq -B\gamma_x + B\eta(B) + \chi(B)\tau_B^x,$$

where

$$\chi(B) := \gamma e^{-\delta/\varepsilon} + \frac{C}{B\varepsilon}.$$

Now for any path A^x we have:

$$\begin{aligned} &\frac{1}{B} \log \mathbb{E}\left[L(A^x) I_B(A^x)\right] = \frac{1}{B} \log(\mathbb{E}\left[L(A) | I_B(A^x) = 1\right] \mathbb{P}\left[I_B(A^x) = 1\right]) \\ &\leq \frac{1}{B} \log\left(\mathbb{E}\left[e^{-B\gamma_x + B\eta(B) + \chi(B)\tau_B^x} | I_B(A^x) = 1\right] p_B^x\right) \\ &= -\gamma_x + \eta(B) + \frac{1}{B} \log \mathbb{E}\left[e^{\chi(B)\tau_B^x} | I_B(A^x) = 1\right] + \frac{1}{B} \log p_B^x. \end{aligned}$$

Using the fact that $\lim_{B\to\infty} \chi(B) = 0$ (see Assumption 5.2), Lemma 5.6 and Theorem 4.5 we conclude that:

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathbb{E}\left[L(A^x) I_B(A^x)\right] \leq -2\gamma_x = 2 \lim_{B\to\infty} \frac{1}{B} \log p_B^x,$$

which completes the proof. □

5.2 Asymptotically efficient scheme for $\mu_1 \leq \mu_2$

We would like to define a function based on the total cost function γ_x in (19), analogous to the function W in the previous section, see (26). Suppose we define the functions $\hat{V}_i(x)$, $i = 1, 2, 3$, in the same way as (25). In particular we would find $\hat{V}_1$ on B_1 and B_2 from (19) as,

$$\hat{V}_1(x) = 2\gamma_{(x_1,0)} - \delta, \quad x \in B_1 \tag{42}$$

$$\hat{V}_1(x) = -2x_1 \log \frac{\tilde{\lambda}(x)}{\lambda} - 2(1-x_2) \log \frac{\tilde{\mu}_2(x)}{\mu_2} - \delta, \quad x \in B_2. \tag{43}$$

where $(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x))$ is the solution to (11) if $x \in B_2$. As a result, $\hat{V}_1(x)$ would not be smooth on the boundary between the sets B_1 and B_2 (see also the discussion above (18)), and hence also the resulting mollified function would not be smooth. This would lead to problems when we try to prove the analogue of Lemma 5.5, where we used continuity and smoothness of $W(x)$. Fortunately, the functions $\hat{V}_1(x)$ and $\hat{V}_2(x)$ coincide on B_1 since they are equal on the boundary between B_1 and B_2 and both functions do not depend on their second argument (hence their gradients coincide). Hence, instead of using (42)-(43) we prefer to work with a function V_1 defined as (43) on *both* B_1 *and* B_2. Mollifying this function with functions V_2 and V_3 as in (26), will then provide a smooth function V.
To be more specific, we first define function $V_1(x)$, based on the second line of (19):

$$V_1(x) = -2x_1 \log \frac{\tilde{\lambda}(x)}{\lambda} - 2(1-x_2) \log \frac{\tilde{\mu}_2(x)}{\mu_2} - \delta,$$

where

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) = \begin{cases} \text{solution to (11)} & \text{if } x \in B_1 \cup B_2, \\ (\lambda, \mu_2, \mu_1) & \text{if } x \in B_3. \end{cases}$$

The function $V_1(x)$ is an extension of the 'cost' proposed by the solution of (11) to the set B_1. In other words, for any $x \in B_1$ we replace the optimal cost (corresponding to the first type of path in (18)) by the cost that corresponds to the (suboptimal) path that leads straight from x to $(0, 1)$. We proceed with the definitions of $V_2(x)$ and $V_3(x)$:

$$V_2(x) = 2\gamma_{(x_1,\delta/2\gamma)} - \delta,$$
$$V_3(x) = 2\gamma - 3\delta,$$

where γ_x is given in (19). In this way, the minimum of V_1 and V_2 is attained by V_1 for $x \in B_2$ (as before), and by V_2 for $x \in B_1$ (rather than by $\hat{V}_1$ as before); see also Figure 5, where $\bar{V}(x) = V_1(x) \wedge V_2(x) \wedge V_3(x)$. The mollification procedure now ensures a smooth transition from B_1 to B_2 for the function $V(x)$ defined as in (26). Another minor problem is that the function $V_2(x)$ is not smooth around $(\alpha_2, 0)$. Specifically for $x_2 < \delta/\gamma_2$, the gradient of $V_2(x)$ is not continuous around the vertical line $(x_1, \cdot)$ where the first component satisfies $f(x_1, \delta/2\gamma) = 0$ with $f(x)$ as defined in (17). Without going into details, we propose to use any suitable mollification procedure to make $V_2(x)$ a smooth function, and from now on we will treat $V_2(x)$ as such. Thus, mollifying the functions $V_i(x)$, $i = 1, 2, 3$, in the

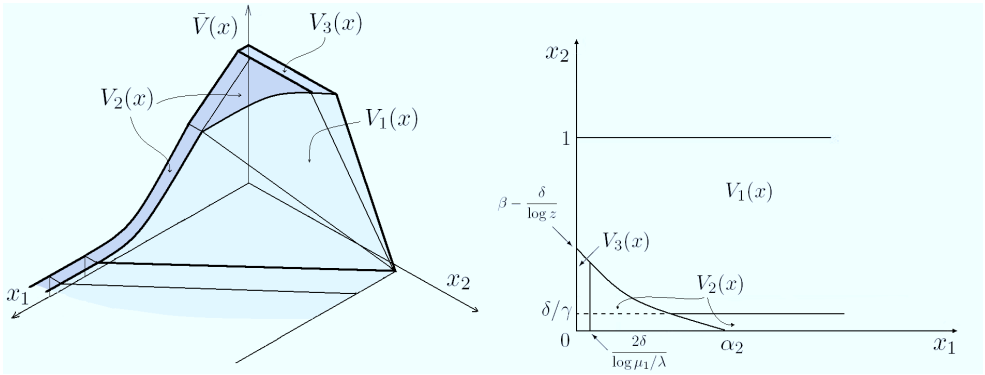

Figure 5: The function $\bar{V}(x)$ and the areas on which $\bar{V}(x) = V_i, i = 1, 2, 3$ (case $\mu_1 \leq \mu_2$).

same manner as we did in (26), we obtain a smooth and continuous function $V(x)$. Based on this function we define the new change of measure $(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2)$ as in (29). As for the gradient in these equations, we like to notice that even though the functions $W_i(x)$ and $V_i(x)$ for $i = 1, 2, 3$ are different, they have similar gradients. In other words, we can replace $DW_i(x)$ by $DV_i(x)$ in Proposition 5.1 to obtain the shape of the gradients of the functions $V_i(x)$, $i = 1, 2, 3$. We will use these gradients in the following proofs.
Turning to the asymptotic efficiency proof of this new change of measure, we first mention that we can prove analogues of Lemmas 5.3, 5.4, 5.5 (using the smoothness of V) and 5.6 for our current case $\mu_1 \leq \mu_2$, as can be checked easily. Now we can proceed with the main result of this subsection.

Theorem 5.8. *When $\mu_1 \leq \mu_2$ and Assumption 5.2 holds, the new measure in (29) with function V based on (19), is asymptotically efficient.*

Proof. The proof is similar to that of Theorem 5.7, the main difference being the bound on the second term of the decomposition of $\log L(A)$ in Lemma 5.4. For any $x \in \partial_2$, we have $\langle DV_2, -v_2 \rangle = \langle DV_3, -v_2 \rangle = 0$ and $\langle DV_1, -v_2 \rangle = 2 \log(\tilde{\mu}_2(x)/\mu_2)$, so

$$\langle DV(x), -v_2 \rangle \geq 2 \log \left(\frac{\tilde{\mu}_2(x)}{\mu_2} \right) e^{-(V_1(x) - V_2(x))/\varepsilon},$$

as in the previous case. For $x = (0, 0)$ we have that

$$V_1(0, 0) - V_2(0, 0) = 2 \log(1/z) - 2\gamma.$$

Using the fact that the optimal cost γ_x for state $x = (0, \beta)$ is both equal to γ (corresonding to the path via state $(\alpha_2, 0)$) and to $(1 - \beta) \log(1/z)$ (corresponding to the path along the vertical axis), this can be rewritten as $V_1(0, 0) - V_2(0, 0) = 2\beta \log(1/z)$. Also, the difference $V_1(x) - V_2(x)$ decreases in x_1 when $x \in \partial_2$, see Lemma 3.5. Combining the above with the fact that $\tilde{\mu}_2(x) \geq \mu_2 z$ (see Lemma 3.5) we obtain the following bound:

$$\langle DV(x), -v_2 \rangle \geq 2 \log(z) \exp \left(-\frac{2\beta \log(1/z)}{\varepsilon} \right),$$

for any $x \in \partial_2$.
Now let us consider the situation when $x \in \partial_1$. Again $\langle DV_3(x), -v_1\rangle = 0$ holds, and in addition $\langle DV_2(x), -v_1\rangle = -2\log\frac{\mu_1}{\lambda}$ and $\langle DV_1(x), -v_1\rangle = 2\log(\lambda/\tilde{\lambda}(x)) - 2\log(\tilde{\mu}_2(x)/\mu_2) > 0$, where the last inequality is due to the simple observation that $\sqrt{\lambda z/\mu_1} > z$. Using (30) we have

$$\langle DV(x), -v_1\rangle \geq \langle DV_2(x), -v_1\rangle = -2\log(\mu_1/\lambda)\rho_2(x) \geq -2\log(\mu_1/\lambda)e^{(V_3(x)-V_2(x))/\varepsilon}.$$

Since for any $x \in \partial_1$ we have $V_3(x) - V_2(x) = -2\delta$, we conclude that

$$\langle DV(x), -v_1\rangle \geq -2\log(\mu_1/\lambda)e^{-2\delta/\varepsilon}.$$

Analogously to the previous proof we now conclude that

$$\limsup_{B\to\infty} \frac{1}{B}\log \mathbb{E}\left[L(A^x)I_B(A^x)\right] \leq -2\gamma_x = 2\lim_{B\to\infty}\frac{1}{B}\log p_B^x,$$

which completes the proof. □

6 Numerical results

We provide some supporting simulation results in this section. To simplify the implementation of the IS scheme, we introduce a slightly different new measure $\mathbb{Q}$, defined as follows:

$$\breve{\lambda}(x) = \lambda\sum_{k=1}^{3}\rho_k(x)e^{-\langle DW_k(x),v_0\rangle/2}e^{\mathbb{H}(DW_k(x))/2}, \tag{44}$$

$$\breve{\mu}_i(x) = \mu_i\sum_{k=1}^{3}\rho_k(x)e^{-\langle DW_k(x),v_i\rangle/2}e^{\mathbb{H}(DW_k(x))/2}, \quad i = 1, 2, \tag{45}$$

where $\rho_k(x)$ is defined by (30).

Theorem 6.1. *Under Assumption 5.2,*

(i) when $\mu_2 < \mu_1$, the measure $\mathbb{Q}$ in (44)–(45), with W defined by (26), is asymptotically efficient.

(ii) when $\mu_1 \leq \mu_2$, the measure $\mathbb{Q}$ in (44)–(45), with W replaced by the function V as defined in Section 5.2, is asymptotically efficient.

Proof. (i) It is clear that the log-likelihood ratio for a transition of type v_0 from any state x under $\mathbb{Q}$ satisfies

$$\begin{aligned}\log\frac{\lambda}{\breve{\lambda}(x)} &= -\log\sum_{k=1}^{3}\rho_k(x)e^{-\langle DW_k(x),v_0\rangle/2}e^{\mathbb{H}(DW_k(x))/2} \\ &\leq -\sum\rho_k(x)\log e^{-\langle DW_k(x),v_0\rangle/2} = \langle DW(x), v_0\rangle/2,\end{aligned}$$

$(0.6\,B, 0)$				$(B, 0)$			
B	ψ_B	p_B	RE	B	ψ_B	p_B	RE
20	1.96	$2.00 \cdot 10^{-5} \pm 2.74 \cdot 10^{-8}$	$7.01 \cdot 10^{-4}$	20	1.92	$1.29 \cdot 10^{-2} \pm 1.61 \cdot 10^{-5}$	$6.38 \cdot 10^{-4}$
50	1.98	$3.12 \cdot 10^{-12} \pm 4.69 \cdot 10^{-15}$	$7.67 \cdot 10^{-4}$	50	1.96	$3.95 \cdot 10^{-5} \pm 5.37 \cdot 10^{-8}$	$7.20 \cdot 10^{-4}$

Table 1: Simulation results; original state-dependent scheme

where the last inequality holds due to the fact that $\mathbb{H}(DW_k(x)) \geq 0$, thanks to Lemma 5.3, and concavity of the logarithm (note that $\sum_{k=1}^{3} \rho_k(x) = 1$). It is obvious that we have similar bounds for transitions v_1 and v_2. Summing these expressions over all steps of sample path $A = (X_j, j = 0 \ldots \sigma)$ we will get the righthand side of expression (31), but without the last term. Since the function $W(x)$ stays the same, we may use the proof of Theorem 5.7 to verify the statement of the current problem.
(ii) The second claim is proved analogously to the first claim. □

All simulations were performed under new measure $\mathbb{Q}$ defined by (44)–(45) and the joint queue-length process around the boundaries was modified according to (1).
Here we present results of dynamic IS simulations for the two-node Jackson tandem network with initial parameters $(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$. We restrict ourselves to the case when the second buffer is the bottleneck (i.e., $\mu_2 < \mu_1$), since the optimal path and new measure $\mathbb{Q}$ are very similar for both cases when the initial state x lies in $A_2 \cup A_3$ (respectively, $B_2 \cup B_3$); only when $x \in B_1$ (for the case $\mu_1 \leq \mu_2$), there is an interesting difference with the other case ($\mu_1 \geq \mu_2$), but the shape of the optimal path and corresponding new measure are then very similar to the ones studied before in [13].
We have two different starting states: $(0.6\,B, 0)$ and $(B, 0)$. Both belong to the most interesting subspace A_2, see Figure 2, where the new measure is found as the solution to system (11).
We performed three different types of IS simulations. At first we simulate the system based on the asymptotic efficient state-dependent scheme obtained in Section 5, i.e., scheme (12) with protection along the horizontal boundary. In these (and all further) simulations we chose $\varepsilon = 0.05$ and $\delta = -\varepsilon \log \varepsilon$, as motivated in Remark 3.7 in [7]. Moreover, we always performed 10^6 simulation runs, leading to comparable computation times in the order of a few minutes; in fact these were approximately linear in the value of B, as could be expected.
The results are presented in Table 1. The value ψ_B in the second column of each table (for both panels) is the estimator of the right hand side of (6), without the limit, and is used as an indication for the efficiency of the scheme. The obtained results are indeed good (that is, close to 2), but it is mentioned that the computation time needed is considerable when B grows large, due to the precalculation of the new measure $\mathbb{Q}$ for all states in A_2.
We tried to simplify the scheme to reduce the computation time, in the following way. We divide the set A_2 into a number of triangles D_i of equal area, each having state $(0, 1)$ as one of the corners, the other corners being given by points on the horizontal axis between $(\alpha_2, 0)$ and $(\alpha_2^{-1}, 0)$, at equal distances. In each of these subsets D_i we use a separate,

fixed, new measure, based on the solution of system (11) where x lies in the middle of the two corners on the horizontal axis. In this way we only need to precalculate a few new measures, rather than dozens for $B = 50$ and hundreds for $B = 100$. In Table 2 the simulation results are given, when A_2 is divided into six subsets. Due to this precalculation reduction it became possible to add the extra line ($B = 100$) in Table 2 (and Table 3).

$(0.6\,B, 0)$				$(B, 0)$			
B	ψ_B	p_B	RE	B	ψ_B	p_B	RE
20	1.95	$2.00 \cdot 10^{-5} \pm 3.04 \cdot 10^{-8}$	$7.77 \cdot 10^{-4}$	20	1.86	$1.29 \cdot 10^{-2} \pm 2.28 \cdot 10^{-5}$	$8.94 \cdot 10^{-4}$
50	1.97	$3.12 \cdot 10^{-12} \pm 6.11 \cdot 10^{-15}$	$9.98 \cdot 10^{-4}$	50	1.91	$3.94 \cdot 10^{-5} \pm 9.36 \cdot 10^{-8}$	$1.20 \cdot 10^{-3}$
100	1.98	$1.82 \cdot 10^{-23} \pm 5.06 \cdot 10^{-26}$	$1.41 \cdot 10^{-3}$	100	1.93	$3.66 \cdot 10^{-9} \pm 1.13 \cdot 10^{-11}$	$1.57 \cdot 10^{-3}$

Table 2: Simulation results; simplified scheme with six domains

As we can see, the variance of the estimator increases as a result of the simplification of the original scheme (although the effect is relatively modest). This may be explained by the following. Under the simplified scheme a path that starts to, say, the left of the middle point of D_i will at some point hit the boundary between D_i and D_{i-1}, after which it follows this boundary. Hence, when B grows large and therefore the sizes of (the unscaled counterparts of) the D_i also increase, the sample path will move back and forth between two different changes of measure for a substantial period of time.
We also simulated the system using an even simpler scheme, getting rid of the set A_2 altogether, expanding A_1 and A_3 so that they meet at the line $x_1 + x_2 = 1$. That is, we simply used the measure $\mathbb{Q} = (\mu_2, \mu_1, \lambda)$ when the total population of the system is less than B and used no change of measure otherwise. Clearly, this method provides worse

$(0.6\,B, 0)$				$(B, 0)$			
B	ψ_B	p_B	RE	B	ψ_B	p_B	RE
20	1.86	$2.00 \cdot 10^{-5} \pm 7.30 \cdot 10^{-8}$	$1.87 \cdot 10^{-3}$	20	1.10	$1.29 \cdot 10^{-2} \pm 1.78 \cdot 10^{-4}$	$7.02 \cdot 10^{-3}$
50	1.91	$3.12 \cdot 10^{-12} \pm 1.67 \cdot 10^{-14}$	$2.73 \cdot 10^{-3}$	50	1.18	$3.67 \cdot 10^{-5} \pm 4.34 \cdot 10^{-6}$	$6.00 \cdot 10^{-2}$
100	1.94	$1.81 \cdot 10^{-23} \pm 1.59 \cdot 10^{-25}$	$4.48 \cdot 10^{-3}$	100	1.30	$5.62 \cdot 10^{-9} \pm 7.09 \cdot 10^{-9}$	$6.40 \cdot 10^{-1}$

Table 3: Simulation results; simplified scheme with two domains

results, as can be concluded from Table 3. Looking at the relative error or at the confidence intervals, it is clear that this method is inferior to the ones presented above.
Finally, we tried a change of measure that replaces the original parameters on A_2 by a simple linear interpolation between the values on A_1 and A_3 (e.g., when $\mu_2 < \mu_1$ and $x \in A_2$ we let $\breve{\lambda}(x) = \alpha(x)\mu_2 + (1 - \alpha(x))\lambda$, where $\alpha(x) \in [0, 1]$ depends on the location of x relative to A_1 and A_3). Unfortunately, this approach gives only slightly better results than those in Table 3, but much worse than those in Table 2.

7 Conclusions

In this paper we focused on the event that, starting from an arbitrary state, the second queue in a two-node Jackson tandem network reaches overflow before the system becomes

empty. The main focus is on the development of efficient simulation techniques for estimating this probability. We have proposed a particular change of measure, motivated by large-deviations arguments, and we have proved asymptotic efficiency of a subtly modified version (that differs close to the axes, and thus nicely controls the likelihood).
We strongly feel that the methods used in the current paper are applicable to other, more complex queueing networks. For example, we expect that it can be applied to a so-called 'slow-down network', i.e., a tandem network with Poisson arrivals and exponential service times, in which the first server decreases its speed as soon as the second buffer reaches some prescribed utilization, see [17]. Such an analysis has recently been published in [6] for a specific parameter setting, with the origin as starting state, but several issues remain open (general parameter settings, general initial point, simplification of the asymptotic efficiency proof).

Acknowledgement
The authors would like to thank P.T. de Boer for useful discussions.

References

[1] V. Anantharam, P. Heidelberger, and P. Tsoucas. Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation. IBM Research Report REC 16280, 1990.

[2] P.T. de Boer. Analysis of state-independent importance-sampling measures for the two-node tandem queue. *ACM Transactions on Modeling and Computer Simulation*, 16(3):225–250, 2006.

[3] P.T. de Boer, Victor F. Nicola, and Reuven Y. Rubinstein. Adaptive importance sampling simulation of queueing networks. In *Proceedings of the 2000 Winter Simulation Conference (WSC'00)*, pages 646–655, Orlando, Florida, 2000.

[4] P.T. de Boer and W.R.W. Scheinhardt. Alternative proof with interpretations for a recent state-dependent importance sampling scheme. *Queueing Systems: Theory and Applications*, 57(2-3):61–69, 2007.

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

[6] P. Dupuis, K. Leder, and H. Wang. Large deviations and importance sampling for a tandem network with slow-down. *Queueing Systems: Theory and Applications*, 57(2-3):71 – 83, 2007.

[7] P. Dupuis, A.D. Sezer, and H. Wang. Dynamic importance sampling for queueing networks. *Annals of Applied Probability*, 17(4):1306–1346, 2007.

[8] P. Glasserman and S.-G. Kou. Analysis of an importance sampling estimator for tandem queues. *ACM Transactions on Modeling and Computer Simulation*, 1(5):22–42, 1995.

[9] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation*, 5(1):43–85, 1995.

[10] F.P. Kelly. *Reversibility and Stochastic Networks*. Wiley, New York, 1979.

[11] D.P. Kroese and V.F. Nicola. Efficient simulation of a tandem Jackson network. *ACM Transactions on Modeling and Computer Simulation*, 12(2):119–141, 2002.

[12] D.P. Kroese, W.R.W. Scheinhardt, and P.G. Taylor. Spectral properties of the tandem Jackson network, seen as quasy-birth-and-death process. *Annals of Applied Probability*, 14(4):2057–2089, 2004.

[13] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. Efficient simulation of a tandem queue with server slow-down. *Simulation*, 83(11):751–767, 2007.

[14] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control*, 34:54–66, 1989.

[15] J. S. Sadowsky. Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *IEEE Transactions on Automatic Control*, 36(12):1383–1394, 1991.

[16] A. Shwartz and A. Weiss. *Large deviations for performance analysis. Queues, communications and computing*. Chapman & Hall, London, UK, 1995.

[17] N. D. van Foreest, M.R.H. Mandjes, J.C.W. van Ommeren, and W.R.W. Scheinhardt. A tandem queue with server slow-down and blocking. *Stochastic Models*, 21(2-3):695–724, 2005.