Simulation of a Jackson Tandem Network using State-dependent Importance Sampling

D.I. Miretskiy
Dept. of Applied Mathematics
University of Twente
The Netherlands
d.miretskiy@utwente.nl

W.R.W. Scheinhardt*
Dept. of Applied Mathematics
University of Twente
The Netherlands
werner@math.utwente.nl

M.R.H. Mandjes†
KdV Institute for Mathematics
The University of Amsterdam
The Netherlands
mmandjes@science.uva.nl

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 Jackson 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 is provably 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.

Keywords

Rare event simulation, importance sampling, state-dependent change of measure, asymptotic optimality, tandem queue

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 [8] 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 in many cases it was possible to prove that the resulting estimator is *asymptotically efficient* (or: asymptotically optimal), which effectively means that its variance 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 Jackson two-node tandem queue (that is, Poisson arrivals, exponential service times at both queues), aiming at estimating the probability that the *total* network population exceeds a given threshold, the seminal paper by Parekh and Walrand [13] 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

*Corresponding author; also affiliated to CWI, Amsterdam, the Netherlands.
†Also affiliated to CWI, Amsterdam, the Netherlands and EURANDOM, Eindhoven, the Netherlands.

SMCTools 2008, October 20, 2008, Athens, GREECE

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4370
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4370

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 [7] 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 [13] performs poorly for some parameter values. Some of the first attempts to solve this problem can be found in [3] and [9], in which *state-dependent* IS schemes were proposed, i.e., IS distributions that are not uniform over the state space. Dupuis *et al.* [6] 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. [9]; 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 [11] for the model that was introduced in [15], 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.* [6], 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 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. Section 4 shows that our IS scheme, after a minor adaptation that deals with visits to the axes, is indeed asymptotically efficient. Some details of the proofs are omitted but can be found in an extended version of this paper, see [12]. We conclude the paper with some discussion in Section 5, where we also spend some words on issues of implementation; supporting numerical results are presented in [12].

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 [6] 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 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 that we can use the same (continuous) state space $\mathcal{R}_+^{\in}$ for

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4370
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4370

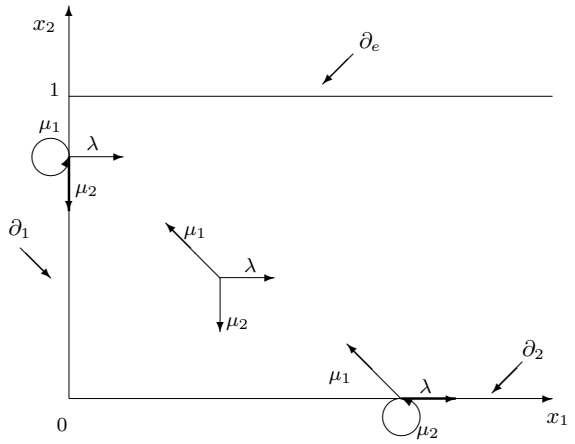

Figure 1: State space and transition structure for scaled process $X(t)$.

any B (although the true, discrete state space varies with the value of B as long as B is finite). 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 4 (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 [11], and in this section we present a method similar to the one in [11] 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

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4370
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4370

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 4. Another result of our method is the exponential decay rate of p_B^x, which will play a crucial role in the proofs of asymptotic efficiency of Section 4.

Before introducing our method we impose some restrictions on the path structures we consider. In [12] it is shown that it is sufficient to only consider paths that satisfy the following.

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 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 [14, 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 [11].

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. ♢

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. The following theorem shows the relevance of this function.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4370
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4370

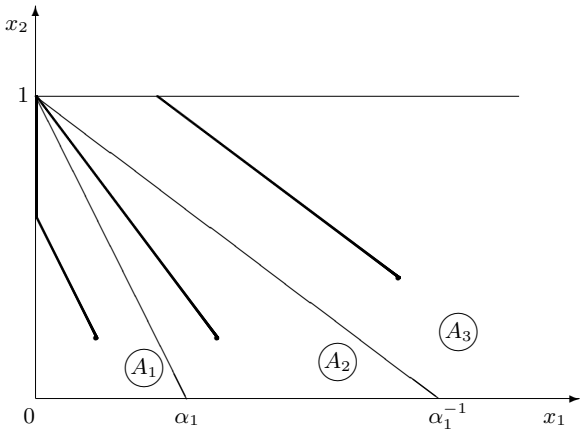

Figure 2: Partition of $\bar{D}$ and some optimal paths to overflow when $\mu_2 < \mu_1$.

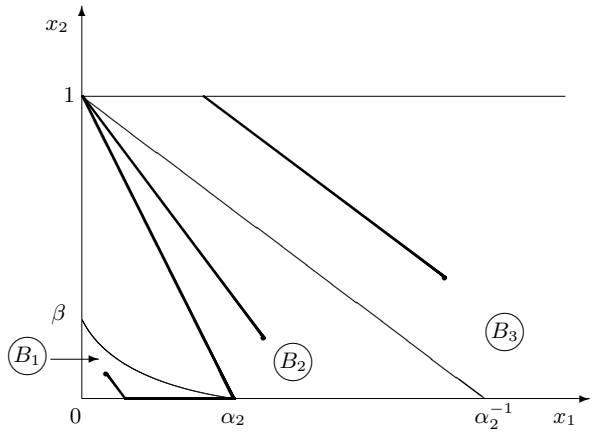

Figure 3: Partition of $\bar{D}$ and some optimal paths to overflow when $\mu_1 \leq \mu_2$.

Theorem 3.4. *The exponential decay rate of p_B^x equals the minimal cost, i.e.,*

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

The proof of this theorem is given in [12], and relies on the fact that the process $X(t)$ satisfies a large deviations principle with a local rate function that is closely related to (and on the interior essentially equal to) the cost function in (9).

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 [11] for more examples), yields the following new measure after some calculations:

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

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

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

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 [13] 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

$$\begin{array}{ll} (x_1, x_2) \to (0, x_2 + \alpha_1^{-1} x_1) \to (0, 1), & \text{if } x \in A_1, \\ (x_1, x_2) \to (0, 1), & \text{if } x \in A_2, \\ (x_1, x_2) \to (x_1 - \alpha_1^{-1} x_2, 1), & \text{if } x \in A_3. \end{array} \quad (13)$$

The resulting cost γ_x of the optimal path starting from $x = (x_1, x_2)$ is given by:

$$\gamma_x = \begin{cases} (1 - x_1 - x_2)\gamma, & \text{if } x \in A_1, \\ -x_1 \log \frac{\tilde{\lambda}(x)}{\lambda} - (1 - x_2) \log \frac{\tilde{\mu}_2(x)}{\mu_2}, & \text{if } x \in A_2, \\ 0, & \text{if } x \in A_3, \end{cases} \quad (14)$$

where

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

is the minimal cost of the path $(0, 0) \to (0, 1)$.

It may be useful to note that for any state x the new measure defined in (12) '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.

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

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

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_1, \lambda, \mu_2), & \text{if } x \in B_1, \\ \text{solution to (11)}, & \text{if } x \in B_2, \\ (\lambda, \mu_2, \mu_1), & \text{if } x \in B_3. \end{cases} \quad (15)$$

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

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

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

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

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

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4370
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4370

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 (15) 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:

$$\begin{array}{ll} (x_1,x_2) \to (x_1 + \alpha_3 x_2, 0) \to (\alpha_2, 0) \to (0,1), & \text{if } x \in B_1, \\ (x_1,x_2) \to (0,1), & \text{if } x \in B_2, \\ (x_1,x_2) \to (x_1 - \alpha_2^{-1} x_2, 1), & \text{if } x \in B_3, \end{array} \tag{16}$$

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 [11] 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):

$$\gamma_x = \begin{cases} \gamma - x_1 \log \frac{\mu_1}{\lambda}, & \text{if } x \in B_1, \\ -x_1 \log \frac{\tilde{\lambda}(x)}{\lambda} - (1 - x_2) \log \frac{\tilde{\mu}_2(x)}{\mu_2}, & \text{if } x \in B_2, \\ (1 - x_2) \log \frac{\mu_2}{\mu_1}, & \text{if } x \in B_3. \end{cases} \tag{17}$$

Finally, we like to mention that for any $x \in B_2$, the new measure 'lies between' the normal measure and the measure that corresponds to the optimal path along the vertical axis. This latter measure follows from the value z as the unique solution in the interval $(0, 1)$ of the (essentially cubic) equation

$$\varphi(z) := \lambda + \mu_1 + \mu_2(1 - z) - 2\sqrt{\frac{\lambda\mu_1}{z}} = 0, \tag{18}$$

which follows from system (11) by taking $(x_1, x_2) = (0, 0)$. (The fact that there is a unique solution immediately follows from $\varphi(0) = -\infty$, $\varphi(1) = \lambda + \mu_1 - 2\sqrt{\lambda\mu_1} > 0$, and the fact that $\varphi'(z) = 0$ has just a single positive solution, viz. $\sqrt[3]{\lambda\mu_1/\mu_2^2}$.) 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 [10, Eqns. (30) and (33)] and [11] for more details.

4. ASYMPTOTIC EFFICIENCY

It is known from [11], where the starting state is the origin, that the new measures (12) and (15) 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 [6]. We will also introduce a protection strip along the lower part of the vertical boundary ∂_1 in the same manner, in the case when $\mu_1 \le \mu_2$.

We again split the problem into two cases: in Section 4.1 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 4.2 we treat the case in which the first server is the bottleneck ($\lambda < \mu_1 \le \mu_2$).

4.1 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 [6].

Let us first introduce three intermediate functions $W_i(x)$, $i = 1, 2, 3$, each with argument $x = (x_1, x_2)$:

$$\begin{aligned} W_1(x) &:= 2\gamma_x - \delta, \\ W_2(x) &:= W_1(x_1, \delta/2\gamma) = 2\gamma_{(x_1, \delta/2\gamma)} - \delta, \\ W_3(x) &:= 2\gamma - 3\delta, \end{aligned} \tag{19}$$

where δ is some small positive number, and γ_x is given by (14). 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 [6]. The last step in the construction is a mollification procedure which makes the resulting function $W(x)$ smooth. We do this by defining:

$$W(x) := -\epsilon \log \sum_{i=1}^{3} e^{-W_i(x)/\epsilon}, \tag{20}$$

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

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

The parameters δ and ϵ depend on B, and in the sequel we will need the following conditions for their asymptotic behavior as B grows large. Note that these are the same conditions as in [6] and [4].

Assumption 4.2. *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

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4370
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4370

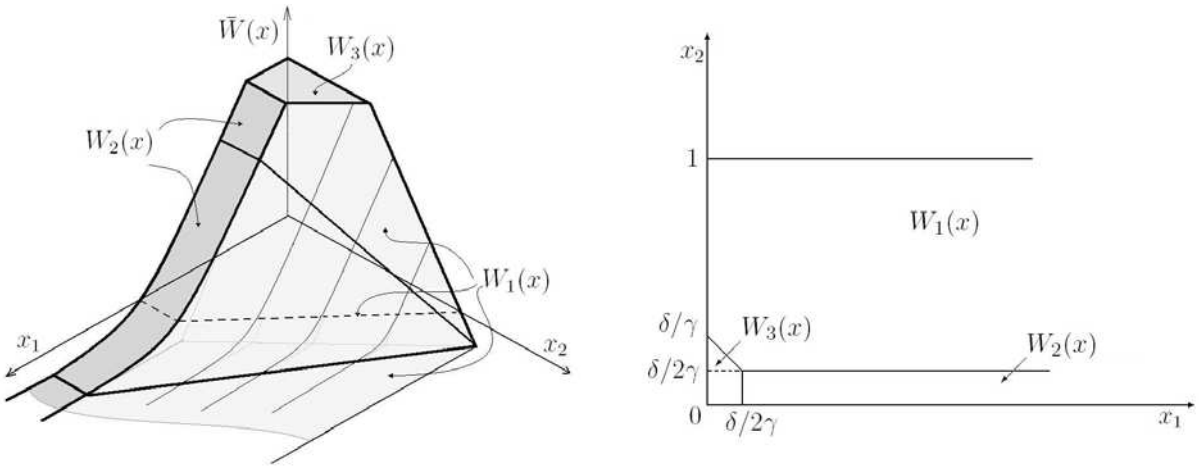

Figure 4: The function $\bar{W}(x)$ and the areas on which $\bar{W}(x) = W_i, i = 1, 2, 3$ (case $\mu_2 < \mu_1$).

from [6, Prop. 3.2]:

$$\begin{aligned} \bar{\lambda}(p) &= N(p)\lambda e^{-\langle p,v_0\rangle/2}, \\ \bar{\mu}_1(p) &= N(p)\mu_1 e^{-\langle p,v_1\rangle/2}, \\ \bar{\mu}_2(p) &= N(p)\mu_2 e^{-\langle p,v_2\rangle/2}, \end{aligned} \tag{21}$$

where

$$\begin{aligned} N(p) &:= \left[\lambda e^{-\langle p,v_0\rangle/2} + \mu_1 e^{-\langle p,v_1\rangle/2} + \mu_2 e^{-\langle p,v_2\rangle/2}\right]^{-1} \\ &= e^{\mathbb{H}(p)/2}. \end{aligned} \tag{22}$$

Here $\mathbb{H}(p)$ is a function known as the *Hamiltonian*, which we use to simplify the notation and to enable the comparison with [6] 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 (21) as

$$\begin{aligned} \bar{\lambda}(x) &= \lambda e^{-\langle DW(x),v_0\rangle/2} e^{\mathbb{H}(DW(x))/2}, \\ \bar{\mu}_i(x) &= \mu_i e^{-\langle DW(x),v_i\rangle/2} e^{\mathbb{H}(DW(x))/2}, \quad i = 1, 2. \end{aligned} \tag{23}$$

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

$$DW(x) = \sum_{k=1}^{3} \rho_k(x) DW_k(x), \text{ with } \rho_k(x) = \frac{e^{-W_k(x)/\varepsilon}}{\sum_{i=1}^{3} e^{-W_i(x)/\varepsilon}} \tag{24}$$

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 (21) 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]; proofs can be found in [12].

Lemma 4.3. *The likelihood $L(A)$ of a path $A = (X_j,\ j = 0, \ldots, \sigma)$ under the new measure (23) satisfies*

$$\begin{aligned} \log L(A) &= \frac{B}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1} - X_j\rangle \\ &\quad + \sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), v_k\rangle I\{X_j = X_{j+1} \in \partial_k\} \\ &\quad - \frac{1}{2}\sum_{j=0}^{\sigma-1}\mathbb{H}(DW(X_j)). \end{aligned} \tag{25}$$

Lemma 4.4. *Consider the case $\mu_2 < \mu_1$. For any path $A = (X_j, j = 0, ..., \sigma)$ under the new measure (23), the first term in (25) satisfies*

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

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

Lemma 4.5. *For any x, $\mathbb{H}(DW(x)) \ge 0$.*

Lemma 4.6. *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:*

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

Theorem 4.7. *When $\mu_2 < \mu_1$ and Assumption 4.2 holds, the new measure in (23), with function W based on (14), is asymptotically efficient.*

Proof. We will sketch the proof, roughly following the proof of [4, Thm. 1]; some omitted details may be found in [12]. First we note that an upperbound on the first term of the log-likelihood expression in Lemma 4.3 can be found, using Lemma 4.4, as

$$\frac{B}{2}\sum_{j=0}^{\tau_B^x-1}\langle DW(X_j), X_{j+1} - X_j\rangle \le \frac{B}{2}(-2\gamma_x + \eta(B)) + \frac{C}{B\varepsilon}\tau_B^x, \tag{26}$$

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4370
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4370

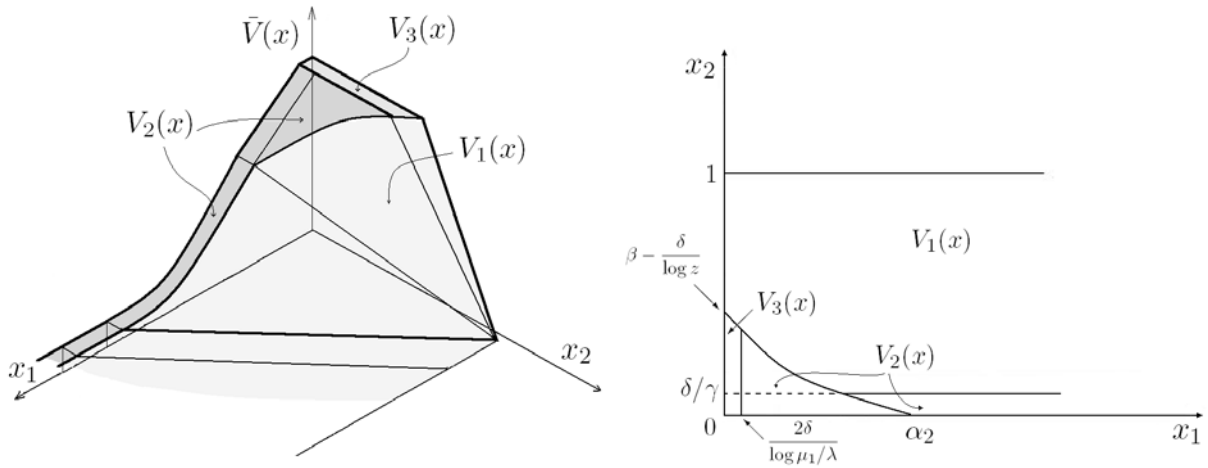

Figure 5: The function $\bar{V}(x)$ and the areas on which $\bar{V}(x) = V_i, i = 1, 2, 3$ (case $\mu_1 \leq \mu_2$).

where $\eta(B)$ is such that $\lim_{B\to\infty} \eta(B) = 0$.

For the second term in Lemma 4.3 we can find a result similar to the third statement of [6, Lemma B.1], namely the same as in [4]:

$$\sum_{k=1}^{2} \frac{1}{2} \sum_{j=0}^{\tau_B^x - 1} \langle DW(X_j), v_k \rangle I\{X_j = X_{j+1} \in \partial_k\} \leq \gamma e^{-\delta/\varepsilon} \tau_B^x. \tag{27}$$

The last term in Lemma 4.3 can also be bounded, using Lemma 4.5:

$$-\frac{1}{2} \sum_{j=0}^{\tau_B^x - 1} \mathbb{H}(DW(X_j)) \leq 0. \tag{28}$$

Combining (26), (27) and (28) we can rewrite (25) 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:

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

Using the fact that $\lim_{B\to\infty} \chi(B) = 0$ (see Assumption 4.2), Lemma 4.6 and Theorem 3.4 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.

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

We define a function based on the total cost function γ_x in (17), analogous to the function W in the previous section, see (20), as follows.

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

where

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

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

where γ_x is given in (17). See also Figure 5 for the function $\bar{V}(x) := V_1(x) \wedge V_2(x) \wedge V_3(x)$.

We note that V_1 is not defined entirely analogous to the way we defined W_1 in the previous section, the reason being that this ensures smoothness around the boundary between B_1 and B_2. We omit the details which can be found in [12]. The same holds for the proof of the following theorem, which is essentially the same as that of Theorem 4.7.

Theorem 4.8. *When $\mu_1 \leq \mu_2$ and Assumption 4.2 holds, the new measure in (23) with function V based on (17), is asymptotically efficient.*

5. DISCUSSION

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).

One of the reasons we did not include numerical results in this paper, is that it is still a nontrivial step to move from the asymptotically optimal algorithm presented here

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4370
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4370

to an actual, useful implementation. The main point here is that the simulation time is not only determined by the number of runs needed, but also by the simulation time per run (and hence it matters how much time the computation of the change of measure "on the fly" takes. As an alternative we could compute the new transition rates in all states of the state space beforehand, and store them. Numerical results based on this approach are presented in [12] and indeed show a considerable speedup. Clearly, the disadvantage is that we need to precompute more and more as B grows large, and in fact we developed and compared several approximate algorithms in [12] that reduce this burden. A perhaps more promising approach is to compute and use only the change(s) of measure that correspond(s) to the optimal path from the initial point, even when the sample path deviates from the optimal path. Clearly, such an approach will also rely heavily on the results in the current paper.

We strongly feel that the methods for constructing the change of measure and proving its efficiency as presented in the current paper are also 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 [15]. Such an analysis has recently been published in [5] 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).

6. ACKNOWLEDGMENTS

Part of this research has been funded by the Dutch BSIK–BRICKS project. The authors would like to thank P.T. de Boer for useful discussions.

7. 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, 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.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. State-dependent importance sampling for a Jackson tandem network. *Submitted*, 2008. See also Memorandum 1867, Dept. of Applied Mathematics, University of Twente.

[13] 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.

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

[15] 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.

Digital Object Identifier: 10.4108/ICST.VALUETOOLS2008.4370
http://dx.doi.org/10.4108/ICST.VALUETOOLS2008.4370