Centrum Wiskunde & Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Rare-event simulation for tandem queues: a simple and efficient importance sampling scheme*

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

REPORT PNA-E0811 DECEMBER 2008

Centrum Wiskunde & 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)

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

Rare-event simulation for tandem queues: a simple and efficient importance sampling scheme*

ABSTRACT
This paper focuses on estimating the rare event of overflow in the downstream queue of a Jacksonian two-node tandem queue, relying on importance sampling. It is known that in this setting 'traditional' state-independent schemes perform poorly. More sophisticated state-dependent schemes yield asymptotic efficiency. Their drawback, however, is that they require a per-state computation of the new measure, so that it still consumes considerable machine time. The contribution of this paper is a scheme that combines asymptotic efficiency with low complexity. It retains the quality of the original state-dependent scheme, but its implementation is almost as simple as for state-independent analogues.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: queueing, importance sampling, rare events, large deviations, tandem network
Note: *Part of this research has been funded by the Dutch BSIK/BRICKS project. This article is also the result of joint research in the 3TU Centre of Competence NIRICT (Netherlands Institute for Research on ICT) within the Federation of Three Universities of Technology in The Netherlands.

Rare-event simulation for tandem queues: a simple and efficient importance sampling scheme*

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

December 15, 2008

Abstract

This paper focuses on estimating the rare event of overflow in the downstream queue of a Jacksonian two-node tandem queue, relying on importance sampling. It is known that in this setting 'traditional' state-independent schemes perform poorly. More sophisticated state-dependent schemes yield asymptotic efficiency. Their drawback, however, is that they require a per-state computation of the new measure, so that it still consumes considerable machine time.

The contribution of this paper is a scheme that combines asymptotic efficiency with low complexity. It retains the quality of the original state-dependent scheme, but its implementation is almost as simple as for state-independent analogues.

1 Introduction

Importance sampling (IS) is a powerful and flexible technique to speed up Monte Carlo simulations of rare events. The main idea behind IS is simulation of a system under a new probability measure which guarantees more frequent occurrence of the rare event of interest. To compensate for the influence of the new measure (i.e., to obtain an unbiased estimator), the simulation outputs need to be corrected by appropriate likelihood ratios. The challenge is to construct a 'good' new measure, by which we mean that the likelihood ratios of the rare event in which we are interested are 'small'. If this is not the case, the variance of the resulting estimator can blow up, or even become infinite. We refer to [1, 8], as well as the forthcoming book [13], for more background.

In the current paper, we consider a two-node Jacksonian tandem and our interest is in a rare event of collecting some large number of jobs in the second buffer before the system becomes empty. A first, naïve approach could be to consider *state-independent* IS schemes, that is, schemes in which the change of measure is static. In this context we mention the landmark paper by Parekh and Walrand [12], where the authors show the appealing result

*Part of this research has been funded by the Dutch BSIK/BRICKS project. This article is also the result of joint research in the 3TU Centre of Competence NIRICT (Netherlands Institute for Research on ICT) within the Federation of Three Universities of Technology in The Netherlands.

that for overflow in a single M/M/1 queue a swap of the arrival and service rate works excellently. Later Sadowsky [14] proved that this type of new measure was asymptotically efficient (or: asymptotically optimal) even for the (multi-server) GI/GI/m queue (with light-tailed service times), where asymptotic efficiency effectively means that the variance of the estimator behaves approximately as the square of its first moment. Application of a similar new measure to a two-node Jacksonian tandem (swapping the arrival rate with the slowest service rate) was not so encouraging – the method was asymptotically efficient for a specific set of parameter values, but led to unbounded variance for other values [7, 2].
It was clear that the class of state-independent new measures was not rich enough to obtain asymptotic efficiency, and therefore one started considering *state-dependent* IS, where the new measure is *not* uniform over the state space. In [3, 15] such measures were constructed, and asymptotic efficiency was empirically concluded, but without any analytic proofs. The first provably asymptotically efficient scheme (even for considerably more general networks) was proposed by Dupuis *et al.* [5], relying on a control-theoretical approach. In this approach, a first element is to find the exponential decay rate of the probability of interest, by relying on a large-deviations formulation (i.e., solving a variational problem). Understanding of this large-deviations behavior is a crucial step in the construction of an asymptotically efficient scheme, but, as argued in [6], in general not sufficient. To make the scheme work, the large-deviations-based new measure has to be subtly modified, as demonstrated in [5].

In the present paper, our aim is to analyze the probability of the population of the downstream queue in a two-node Jacksonian tandem exceeding some predefined threshold, before the system idles. It is noted that [5] focuses on overflow in either of the buffers during a busy cycle. A crucial difference (on the technical level) between these two settings, is that in our setting the states space (that is, the set of states the process can visit before one can determine whether the threshold has been reached before the system idle) is infinite, as in principle the first queue can grow beyond any value during such a run. The present paper is a follow-up of [9], where for this problem a state-independent scheme was proposed that worked well for just a limited set of parameter values, and [10], where an asymptotically efficient state-dependent scheme was presented. It is also stressed that [10] is more general than most of the previous papers, in that it considers the event of exceeding a threshold in the second buffer before emptying the system, starting from *any* given state, in contrast to all previous research where the origin was chosen as the starting state.
Importantly, the IS scheme of [10] has a substantial drawback as well: the state-dependence entails that one has to compute the new measure for any state in the state space, which may be time-consuming. Therefore one would like to device an IS algorithm that combines the attractive features of state-dependent and state-independent schemes. The present paper provides such a scheme: it is asymptotically efficient, but at the same time of low complexity (as determining the new measure requires minimal computational effort). Numerical experiments provide further insight in the performance of our method (including a comparison with existing methods). We remark that we could have cast our approach in a

control-theoretic framework, in line with, e.g., [5, 6], but we have refrained from doing so with the intention to make the paper accessible to an audience as broad as possible.

We finish this introduction with a description of the paper's structure. We describe the model of interest in detail in Section 2. Section 3 contains, in addition to a brief review of IS, the IS schemes themselves. The analytic proof of asymptotic efficiency is given in Section 4. Supporting numeric results are presented in Section 5. We end the paper with some final remarks in Section 6.

2 Model

We consider two M/M/1 queues in tandem, with jobs arriving at the first queue according to a Poisson process of rate λ, and the queues having service rates μ_1 and μ_2 respectively. Both stations have infinitely large waiting room. Moreover, we assume that the system is stable, i.e., $\lambda < \min\{\mu_1, \mu_2\}$. As we are interested in the probability that the number of jobs at the second station exceeds a given (typically high) level B before the system gets empty, we may rescale time; in the sequel we assume $\lambda + \mu_1 + \mu_2 = 1$ without loss of generality. It is also clear that all information required is captured 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. The possible transitions of the latter Markov chain are $v_0 = (1, 0)$, $v_1 = (-1, 1)$ and $v_2 = (0, -1)$ with corresponding probabilities λ, μ_1 and μ_2. However, note that the transition v_k is impossible while queue k is empty, and to resolve this issue we introduce self-transitions:

$$\mathbb{P}(Q_{j+1} = Q_j | Q_{k,j} = 0) = \mu_k, \text{ for } k = 1, 2.$$

For convenience we also introduce the scaled process $X_j = Q_j/B$. The main benefit of this is that X_j's state space does not depend on B: it is $\bar{D} = D \cup \partial_e \cup \partial_1 \cup \partial_2$, where

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

see Figure 1. Note that the probability of our interest is equal to the probability that process X_j reaches the exiting boundary ∂_e before reaching the origin. To define it more formally, we first introduce the stopping time τ_B^s, which denotes the first entrance of the process X_j to the exit boundary ∂_e starting from the state $s = (s_1, s_2)$ without visits to the origin:

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

where $\tau_B^s := \infty$ if X_j reaches the origin before ∂_e. Denote by $I_B(A^s)$ the indicator of the event $\{\tau_B^s < \infty\}$ for the path $A^s = (X_j, j = 0, \ldots : X_0 = s)$, as in [10]. Consequently the probability of interest reads

$$p_B^s := \mathbb{P}(\tau_B^s < \infty) = \mathbb{E} I_B(A^s). \tag{2}$$

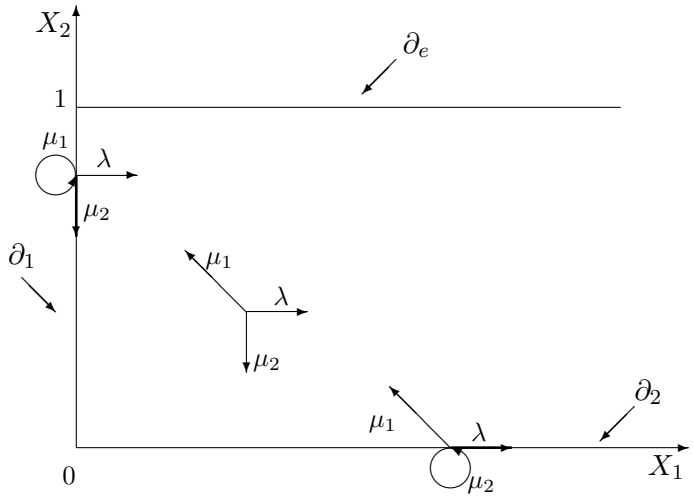

Figure 1: State space and transition structure for scaled process X_j.

3 Importance Sampling

This section has two goals: (i) a brief description of IS, (ii) the presentation of our IS scheme for estimating the probability p_B^s.

IS is one of the most popular and powerful tools to efficiently estimate rare-event probabilities. For example in our case, due to the rarity of the event under consideration, simulating the system under the original measure to estimate p_B^s is inefficient. In IS this problem is resolved by simulating the system under a *different* measure, under which the event of interest occurs frequently.

To estimate p_B^s, IS generates samples under a new probability measure $\mathbb{Q}$, with respect to which $\mathbb{P}$ is absolutely continuous. It is elementary that p_B^s can now alternatively be expressed as an expectation under $\mathbb{Q}$, viz. $p_B^s = \mathbb{E}^{\mathbb{Q}}[L(A^s)I_B(A^s)]$, where L is the likelihood ratio (also known as Radon-Nikodým derivative) of a realization ('path') ω, i.e., $L(\omega) = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\omega)$. Performing n independent runs, with observations $(L_i(A^s), I_{B,i}(A^s)), i = 1, \ldots, n$ an unbiased estimator is $n^{-1}\sum_{i=1}^n L_i(A^s)I_{B,i}(A^s)$. A notion developed to measure the efficiency of the new measure $\mathbb{Q}$ is *asymptotic efficiency*, which roughly requires that the second moment of the estimate behaves approximately as the square of its first moment (thus essentially minimizing the variance of the estimator). Since we know that $B^{-1}\cdot\log p_B^s$ converges to a positive constant as B grows large (see [10]; this constant, the exponential decay rate of p_B^s, will be denoted by $\gamma^s(s)$ in the sequel), we can write the definition of asymptotic efficiency in our case as follows.

Definition 3.1. *The IS scheme is asymptotically efficient if*

$$\limsup_{B\to\infty} \frac{1}{B}\log \mathbb{E}[L(A^s)I_B(A^s)] \leq 2\lim_{B\to\infty}\frac{1}{B}\log p_B^s. \tag{3}$$

If the probability of interest decays exponentially in the 'rarity parameter' (B in our case), which holds for our p_B^s, asymptotic efficiency effectively means that the number of replications needed to obtain an estimate of given accuracy grows subexponentially in the rarity

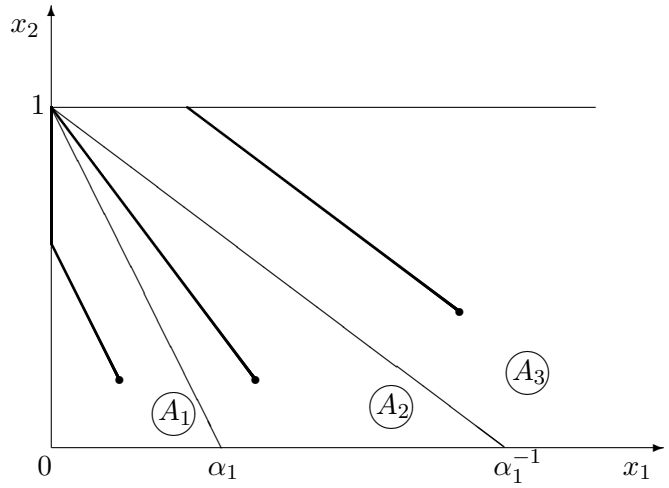

Figure 2: Partition of $\bar{D}$ and some optimal paths to overflow when $\mu_2 < \mu_1$.

parameter.

Before we propose the new measure to be used in this paper, we mention the main difference with the new measure introduced in [10]. In [10], the new measure $\mathbb{Q}$ is state-dependent, and needs to be computed after every transition (requiring a certain cubic equation to be solved numerically). In the present paper the new measure is still state-dependent, but its computation is substantially less demanding, as it requires just a single three-dimensional system to be solved.

3.1 IS scheme for the case $\mu_2 < \mu_1$

Recall that our goal is to modify the IS scheme described in [10], such that the scheme's complexity is reduced, but without compromising the asymptotic efficiency. Again, the scheme is based on the most probable path to overflow that we identified in [10], as well as the new measure that ensures that 'on average' the process follows this optimal trajectory. To ease the exposition of the new measures, we partitioned the state space as shown in Figure 2 into A_1, A_2 and A_3; here, $\alpha_1 := (\mu_1 - \mu_2)/(\mu_1 - \lambda)$. The same figure also provides some examples (solid lines) of the most probable path to the exit boundary for various starting states s.

We now proceed by giving the new measure for starting points in A_1, A_2, and A_3. Let $(\lambda^{(\text{line})}, \mu_1^{(\text{line})}, \mu_2^{(\text{line})})$ solve

$$\begin{cases} \lambda^{(\text{line})} = \mu_1^{(\text{line})} - s_1(\mu_1^{(\text{line})} - \mu_2^{(\text{line})})/(1 - s_2) \\ \lambda^{(\text{line})} + \mu_1^{(\text{line})} + \mu_2^{(\text{line})} = \lambda + \mu_1 + \mu_2 \\ \lambda^{(\text{line})}\mu_1^{(\text{line})}\mu_2^{(\text{line})} = \lambda\mu_1\mu_2 \\ \lambda^{(\text{line})} \leq \mu_1^{(\text{line})} \text{ and } \mu_1^{(\text{line})} > \mu_2^{(\text{line})} \\ \lambda^{(\text{line})},\ \mu_1^{(\text{line})},\ \mu_2^{(\text{line})} > 0. \end{cases} \tag{4}$$

The superscript "(line)" indicates that the solution is in fact the optimal change of measure to reach the exit boundary following a straight line starting in s. Now we can define the

(overall) optimal new measure $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ through

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_2, \mu_1, \lambda), & \text{if } s \in A_1, \\ (\lambda^{(\text{line})}, \mu_1^{(\text{line})}, \mu_2^{(\text{line})}), & \text{if } s \in A_2, \\ (\lambda, \mu_1, \mu_2), & \text{if } s \in A_3. \end{cases} \tag{5}$$

Note that the dependence of $\tilde{\lambda}$ etc. on s is suppressed in the notation. Next we define

$$\gamma^s(x) := -x_1 \log \frac{\tilde{\lambda}}{\lambda} - (1 - x_2) \log \frac{\tilde{\mu}_2}{\mu_2}. \tag{6}$$

In the context of [10], $\gamma^s(x)$ can be interpreted as the residual 'cost' of moving from state x to ∂_e along the path to overflow that started in s. In particular $\gamma^s(s)$, the total cost of moving from s to ∂_e, is equal to the exponential decay rate of p_B^s, i.e., $B^{-1} \cdot \log p_B^s \to -\gamma^s(s)$, see Theorem 4.5 in [10].

Notice that the function $\gamma^s(x)$ is simply linear in x, since the new 'tilde-measure', i.e., $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$, depends only on the fixed initial state s, and *not* on the current state x. This is the main difference with the new measure studied in [10], where we used the optimal new measure for each current state x with its cost $\gamma^x(x)$. Therefore a cubic equation (corresponding to system (4) with s replaced by x) had to be solved for each state x in the sample path. In our current approach, computation of the tilde-measure requires the (numerical) solution of just a single cubic equation.

It is known, e.g. from our previous research [9], that the new measure $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$, which makes a sample path 'on average' follow the optimal trajectory to the rare set, is not necessarily asymptotically efficient; this is due to the possibility of several visits to the horizontal axis, which inflate the likelihood ratio, cf. [2, 12]. In order to resolve this, we first introduce the measure $(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2)$ as in [5], to be used when the current state is on or near the horizontal axis, through

$$(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2) := (\tilde{\lambda}, \mu_1 \lambda / \tilde{\lambda}, \mu_2). \tag{7}$$

The primary idea behind this 'hat-measure' is to make the likelihood ratios of the loops around the horizontal axis not greater than 1 (by ensuring $\hat{\mu}_2 = \mu_2$).

Having introduced the 'tilde-measure' and the 'hat-measure', we are now ready to define the (state-dependent) measure $(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x))$, of which we will prove asymptotic efficiency, and which is a combination of the two measures defined above and the original measure:

$$\begin{aligned} \bar{\lambda}(x) = &\quad \tilde{\lambda}^{\rho_1(x)} \hat{\lambda}^{\rho_2(x)} \lambda^{\rho_3(x)} M(x), \\ \bar{\mu}_1(x) = &\quad \tilde{\mu}_1^{\rho_1(x)} \hat{\mu}_1^{\rho_2(x)} \mu_1^{\rho_3(x)} M(x), \\ \bar{\mu}_2(x) = &\quad \tilde{\mu}_2^{\rho_1(x)} \hat{\mu}_2^{\rho_2(x)} \mu_2^{\rho_3(x)} M(x). \end{aligned} \tag{8}$$

Here $M(x)$ is a normalization function, and the $\rho_i(x)$ are positive weights, adding up to

unity, such that $\rho_1(x)$ is close to 1 on almost all of the state space (leaving the other weights to be close to zero), while $\rho_2(x)$ is close to 1 only near the horizontal axis; the reason that we include the normal measure with a weight $\rho_3(x)$ that should be close to 1 near the origin, is that applying the 'hat measure' there would also lead to high likelihood ratios. For the precise definition of the weights we have some freedom; here we follow the convenient choice in [5],

$$\rho_i(x) = \frac{e^{-W_i(x)/\epsilon}}{\sum_{j=1}^{3} e^{-W_j(x)/\epsilon}}, \quad i = 1, 2, 3, \tag{9}$$

where

$$W_1(x) := 2\gamma^s(x) - \delta, \quad W_2(x) := W_1(x_1, \delta/2\gamma^s(0)), \quad W_3(x) := 2\gamma^s(0) - 3\delta. \tag{10}$$

Not only does this choice ensure that the new measure (8) has the 'appropriate' form, in that $(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x)) \approx (\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ if $x \in D$, or $(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x)) \approx (\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2)$ if $x \in \partial_2$, etc.; it is also possible to express the new measures (5) and (7), and the natural measure in terms of the gradients of the functions $W_1(x), W_2(x)$ and $W_3(x)$ respectively, see [5, 10]. This will be useful in the proof of asymptotical efficiency in Section 4.

3.2 IS scheme for the case $\mu_1 \leq \mu_2$

In this subsection we present the IS scheme for the case when $\mu_1 \leq \mu_2$. Again, we start by partitioning the state space $\bar{D}$ as in [10], see Figure 3. To see whether a starting state s belongs to B_1 or B_2 we again need to solve system (4). Then s belongs to B_1 if and only if $f(s) \leq 0$, where

$$f(s) := \log \frac{\mu_2}{\lambda} + s_1 \log \frac{\lambda^{(\text{line})}}{\mu_1} + (1 - s_2) \log \frac{\mu_2^{(\text{line})}}{\mu_2}.$$

The constant β is the solution to $f(0, s_2) = 0$, while $\alpha_2 := (\mu_2 - \mu_1)/(\mu_2 - \lambda)$.
In the previous subsection, we arrived at a 'uniform' new measure $(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x))$ for all s, but in the case $\mu_1 \leq \mu_2$, we have to distinguish between two measures, depending on the starting state.

- At first, let us consider the case $s \in B_2 \cup B_3$. Then we define

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\lambda^{(\text{line})}, \mu_1^{(\text{line})}, \mu_2^{(\text{line})}), & \text{if } s \in B_2, \\ (\lambda, \mu_2, \mu_1), & \text{if } s \in B_3. \end{cases} \tag{11}$$

  The function $\gamma^s(x)$ is again defined by (6), but of course its shape is different from that in the previous subsection, since the 'tilde-measure' is now different. Similarly, for the IS simulations we propose to use the state-dependent new measure $(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x))$ again defined by (8) (with the 'tilde-measure' given by (11)), and the weights $\rho_i(x)$ given through (9) in conjunction with (10).

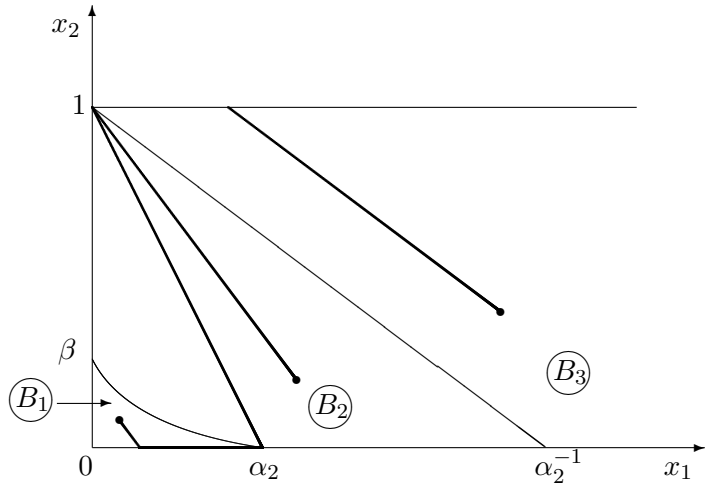

Figure 3: Partition of $\bar{D}$ and some optimal paths to overflow when $\mu_1 \leq \mu_2$.

- Now consider $s \in B_1$. We know from [10] that the optimal trajectory for this case consist of three straight subpaths, see also Figure 3. In our new measure, we need the stopping time $\tau^\star$, defined as the first time X_k visits $B_3 \cap \partial_2$, or, formally,

$$\tau^\star := \min\{k : X_{1,k} \geq \alpha_2^{-1} \text{ and } X_{2,k} = 0\}. \tag{12}$$

Now we define the new measure, being (μ_1, λ, μ_2) before time $\tau^\star$ and (μ_1, μ_2, λ) after it, by

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_1, \lambda I_{\{k<\tau^\star\}} + \mu_2 I_{\{k\geq\tau^\star\}}, \mu_2 I_{\{k<\tau^\star\}} + \lambda I_{\{k\geq\tau^\star\}}); \tag{13}$$

note that the new measure depends on time k, but (as with the starting state s) we omit this dependence in the notation. The residual cost $\gamma^s(x)$ is not given by (6) anymore, but rather by

$$\gamma^s(x) := \log\frac{\mu_2}{\lambda} - x_1 \log\frac{\tilde{\lambda}}{\lambda} + x_2 \log\frac{\tilde{\mu}_2}{\mu_2}, \quad \text{if } s \in B_1, \tag{14}$$

which is again a simple linear function in x; also $\gamma^s(s)$ is again the exponential decay rate of p_B^s, see Theorem 4.5 in [10].

With the function $\gamma^s(x)$ defined by (14), the proposed state-dependent new measure $(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x))$ is given by (8) (with the 'tilde-measure' given by (13)), and the weights $\rho_i(x)$, as before, through (9) and (10).

3.3 Overview of the IS scheme

For convenience we summarize the resulting IS scheme for the different cases.

- When $\mu_2 < \mu_1$ one needs to
  1. define the 'primary' new measure (5);
  2. define the 'hat'-measure (7);
  3. define weights $\rho_i(x)$ by (9), based on (10) and (6);
  4. apply (8).
- When $\mu_1 \leq \mu_2$ and $s \in B_2 \cup B_3$, the same procedure is followed, only replacing the 'primary' new measure (5) by (11) in step 1.
- When $\mu_1 \leq \mu_2$ and $s \in B_1$, again the same procedure is followed, this time replacing the 'primary' new measure by that in (13) *and* replacing (6) by (14) when determining the $W_i(x)$ and $\rho_i(x)$ in step 3.

Note that in the last case we always have $(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2) = (\mu_1, \lambda, \mu_2)$, both before and after time $\tau^\star$. In particular this means that when $k < \tau^\star$, the hat measure coincides with the tilde measure (and hence $\rho_1(x) = \rho_2(x)$).

4 Asymptotic Efficiency

We now prove the asymptotic efficiency of the IS scheme proposed in the previous section; the approach is due to [4].

Theorem 4.1. *If we choose the strictly positive parameters $\delta \equiv \delta_B$ and $\epsilon \equiv \epsilon_B$ such that as $B \to \infty$: (i) $\epsilon_B \to 0$, (ii) $\delta_B \to 0$, (iii) $B\epsilon_B \to \infty$, (iv) $\epsilon_B/\delta_B \to 0$, then the IS scheme defined by (8) is asymptotically efficient.*

Proof. Our first step is the decomposition of the likelihood $L(A)$ of any path $A = (X_j, j = 0, \ldots, \sigma)$ in three terms, cf. [10]. For this needs we define the following function

$$W(x) := -\epsilon \log \sum_{i=1}^{3} e^{-W_i(x)/\epsilon}, \tag{15}$$

which was firstly introduced in [5]. It is not difficult to see that

$$DW(x) = \sum_{i=1}^{3} \rho_i(x) DW_i(x), \tag{16}$$

where the weights $\rho_i(x)$ are defined by (9). Combining the definition of the likelihood ratio

with (15) and (16) one obtains

$$\begin{aligned}\log L(A) &= \frac{B}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1}-X_j\rangle \\ &+ \sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), v_k\rangle I\{X_j = X_{j+1}\in\partial_k\} \\ &- \sum_{j=0}^{\sigma-1}\log M(X_j).\end{aligned} \tag{17}$$

Now we bound all three summations in (17) and show that only the first sum has a significant impact on the log-likelihood.

(a) We start by analyzing the first term. For any path $A = (X_j, j = 0, ..., \sigma)$ and some positive constant C we can, in self-evident notation, construct the following bound:

$$\left|\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1}-X_j\rangle - (W(X_\sigma)-W(X_0))\right| \le \frac{C}{B^2\epsilon}\sigma + \frac{C}{B}I_{\{\mu_1\le\mu_2\}\cap\{s\in B_1\}}. \tag{18}$$

The proof of the above inequality is based on the approximation of the increment of the function $W(x)$ in terms of its gradient $DW(x)$, analogous to Lemma 5.5 in [10]. The accuracy of this representation can be shown by bounding the absolute value of any element of the corresponding Hessian matrix from above. The first term in the right-hand side of (18) corresponds to the sum of these contributions over all σ steps. The second term appears only if $\mu_1 \le \mu_2$ and $s \in B_1$, as a consequence of the non-smoothness of $\gamma^s(x)$ as a function of k, see (13), and therefore also of $W(x)$, after the $\tau^\star$-th transition; note that a similar problem was treated in Lemma 4.4 of [11]. Bearing in mind the definition of the function $W(x)$, see (15), we obtain

$$W(s) \ge 2\gamma^s(s) - \epsilon\log(3) - 3\delta \quad \text{and} \quad W(X_{\tau_B^s}) \le -\log\frac{\tilde\lambda}{\lambda}X_{1,\tau_B^s} - \delta \le -\delta.$$

Combining the two last inequalities with (18), we derive an upper bound for the first term in (17):

$$\sum_{j=0}^{\tau_B^s-1}\langle DW(X_j), X_{j+1}-X_j\rangle \le -2\gamma^s(s) + \eta(B) + \frac{C}{B^2\epsilon}\tau_B^s, \tag{19}$$

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

(b) We now proceed with the second term. For any path $A = (X_j, j = 0, ..., \sigma)$ and some positive constant $\gamma^\star$ we obtain by routine computations, as were done in in [10],

$$\sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), v_k\rangle I_{\{X_j = X_{j+1}\in\partial_k\}} \le \gamma^\star e^{-\delta/\epsilon}\sigma.$$

(c) We finally consider the third term. For any $x \in D$ we have $\log M(x) \geq 0$. We skip the proof of the result as it consists of lengthy, but basic computations, that can be found in [10].
Upon combining *(a)*, *(b)* and *(c)*, we obtain the following upper bound on the likelihood ratio:

$$\log L(A^s) \leq -B\gamma^s(s) + B\eta(B) + \chi(B)\tau_B^s, \quad \text{where} \quad \chi(B) := \gamma^\star e^{-\delta/\epsilon} + \frac{C}{B\epsilon}.$$

After some elementary algebra this leads to

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

Using that $\lim_{B\to\infty}\chi(B) = 0$, due to assumptions *(iii)* and *(iv)*, in conjunction with

$$\lim_{B\to\infty}\frac{1}{B}\log \mathbb{E}(e^{\chi(B)\tau_B^s}|I_B(A^s)=1) = 0, \quad \text{when} \quad \lim_{B\to\infty}\chi(B) = 0,$$

see Lemma 5.6 in [10], we can neglect the penultimate item in the last expression. Now recalling that $B^{-1}\cdot\log p_B^s \to -\gamma^s(s)$, we conclude that

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

which completes the proof. □

5 Numerical Results

In this section we present two types of results. We start with some estimates of p_B^s obtained using our IS-scheme (8), see Table 1. In the rest of the section we compare the performance of the current scheme with that of the existing methods, in particular with [10]; see Table 2.
In Table 1 we present simulation results for three different parameter settings using the IS scheme defined in (8). We compute the weights $\rho_i(x)$ as in (9), choosing $\epsilon = 0.005$ and $\delta = -\epsilon\log\epsilon$ to enable comparison with [10]; see also [5] for the motivation of this choice. Each time we perform a fixed number of 10^6 simulation runs. In Table 1 we present the resulting estimates of p_B^s with 95%-confidence intervals. In the first two columns we have $\mu_2 < \mu_1$ while the third column has $\mu_1 < \mu_2$. In columns 1 and 3 we chose $s = (0,0)$ and the parameters λ, μ_1, μ_2 lie close together; the latter is challenging in the sense that such values are often problematic for IS. A comparison with Tables 1 and 2 in [9], where the same parameters were simulated using state-*in*dependent IS, indeed shows similar estimates, but with smaller confidence intervals. Column 2 shows a scenario in which the starting state is not the origin. The results may be compared with those in Table 1 in [10], showing similar results.
We now turn to comparing the performance of three different IS schemes (as well as

B	$(\lambda, \mu_1, \mu_2) = (0.3, 0.36, 0.34)$ $s = (0, 0)$	$(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$ $s = (0.6\,B, 0)$	$(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$ $s = (0, 0)$
20	$5.96 \cdot 10^{-2} \pm 3.65 \cdot 10^{-4}$	$2.00 \cdot 10^{-5} \pm 5.19 \cdot 10^{-8}$	$3.05 \cdot 10^{-2} \pm 2.27 \cdot 10^{-3}$
50	$1.52 \cdot 10^{-3} \pm 1.17 \cdot 10^{-5}$	$3.12 \cdot 10^{-12} \pm 9.70 \cdot 10^{-15}$	$6.15 \cdot 10^{-5} \pm 9.54 \cdot 10^{-6}$
100	$2.93 \cdot 10^{-6} \pm 2.37 \cdot 10^{-8}$	$1.82 \cdot 10^{-23} \pm 6.48 \cdot 10^{-26}$	$1.52 \cdot 10^{-9} \pm 4.01 \cdot 10^{-10}$

Table 1: Simulation results: 95%-confidence intervals for p_B^s

	st.-indep. [9]		st.-dep. old [10]			st.-dep. new		straightforward	
B	RE	time	RE	virtual time	time	RE	time	RE	time
20	$6.08 \cdot 10^{-3}$	12	$2.61 \cdot 10^{-3}$	$5 \cdot 10^6$	$55 + 16$	$3.12 \cdot 10^{-3}$	28	$3.95 \cdot 10^{-3}$	6
50	$2.21 \cdot 10^{-2}$	37	$3.12 \cdot 10^{-3}$	$9 \cdot 10^6$	$132 + 100$	$3.94 \cdot 10^{-3}$	80	$2.18 \cdot 10^{-2}$	9
100	$1.37 \cdot 10^{-2}$	77	N/A	N/A	N/A	$4.74 \cdot 10^{-3}$	168	$5.70 \cdot 10^{-1}$	9

Table 2.1: $(\lambda, \mu_1, \mu_2) = (0.3, 0.36, 0.34)$ and $s = (0, 0)$

20	N/A	N/A	$7.01 \cdot 10^{-4}$	$1 \cdot 10^6$	$42 + 11$	$1.32 \cdot 10^{-3}$	7	$2.42 \cdot 10^{-1}$	3
50	N/A	N/A	$7.67 \cdot 10^{-4}$	$3 \cdot 10^6$	$104 + 68$	$1.58 \cdot 10^{-3}$	18	N/A	N/A
100	N/A	N/A	N/A	N/A	N/A	$1.81 \cdot 10^{-3}$	35	N/A	N/A

Table 2.2: $(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$ and $s = (0.6\,B, 0)$

20	$4.59 \cdot 10^{-2}$	15	$3.10 \cdot 10^{-2}$	$9 \cdot 10^6$	$46 + 11$	$3.79 \cdot 10^{-2}$	28	$5.40 \cdot 10^{-3}$	9
50	$3.67 \cdot 10^{-1}$	53	$6.73 \cdot 10^{-2}$	$22 \cdot 10^6$	$123 + 68$	$7.91 \cdot 10^{-2}$	84	$1.25 \cdot 10^{-1}$	11
100	$2.33 \cdot 10^{-1}$	116	N/A	N/A	N/A	$13.4 \cdot 10^{-2}$	189	N/A	N/A

Table 2.3: $(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$ and $s = (0, 0)$

Table 2: Comparison of different schemes

straightforward simulations). In Table 2 we present the results for the same three scenarios as in Table 1. For the same fixed number of replications (10^6) we compare the relative errors (RE) and machine running times (time; in seconds) of the different schemes. In the first column we always use the state-independent IS scheme designed in [9]; for the second column we use the state-dependent scheme described in [10], and the third column contains the outcomes of the current scheme. We also applied straightforward simulations to obtain the same estimates, see the fourth column.

The *virtual time* in the second column is an estimate of the time it would take to actually follow the IS scheme from [10], recalculating the path to overflow and the corresponding new measure after each transition. When the current state x is in subspace A_2 (or B_2) this means solving system (4) many times (with s replaced by x). To estimate the virtual time needed to do this, we multiplied the number of transitions in A_2 (or B_2) with the time needed to solve (4). However, when we did the simulations in [10] we actually used a method which is less time consuming, namely we precalculated the new measure for each state inside A_2 in advance. The real computation time therefore consists of two parts, which can be found under 'time' in the second column: the simulation time itself (first term) and the time needed to pre-compute the new measure (second term). Note that the pre-computation time grows as a square of the overflow level B.

From Tables 2.1 and 2.3 it becomes clear that both the scheme in [10] and the current scheme provide a relative error that is much smaller than with the state-independent scheme from [9]. (Note that the latter is not available in Table 2.2 since we only allowed the origin as starting state in [9]). This is due to our choice of the parameters: we chose the

values of the parameters λ, μ_1 and μ_2 very close to each other, since this is the most difficult case. Therefore, the IS scheme performs even better when arrival and service rates are clearly distinctive, as in Table 3, but this may also hold when we apply a state-independent scheme for these parameter values.
When we compare the current scheme with the old state-dependent scheme in [10], it becomes apparent that the relative error is slightly larger than in the old scheme, but of the same order. The big advantage is of course that running times are much lower, and the scheme is easier to implement.

6 Conclusions

In this work we designed an asymptotically efficient IS scheme for estimating the probability of overflow in the second buffer of a two-node tandem Jackson network. The IS scheme presented in this paper is the result of an investigation started in [9, 10]. The scheme constructed in [9] is easy to implement, but it is not always asymptotically efficient. The scheme from [10] *is* asymptotically efficient, but it has the drawback that it is difficult to use in practice, and simulation times are high. The IS scheme designed in this paper provides a good compromise: it is asymptotically efficient for all parameter values, giving relative errors that are comparable to those from the 'fully state-dependent' counterpart in [10] (although slightly larger), and at the same time it is almost as simple to implement and as fast as the state-independent schemes in [9].

References

[1] J.A. Bucklew. *Introduction to Rare Event Simulation*. Springer, 2004.

[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, V.F. Nicola, and R.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, A.D. Sezer, and H. Wang. Dynamic importance sampling for queueing networks. *Annals of Applied Probability*, 17(4):1306–1346, 2007.

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

[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.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.

[10] 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, *URL: http://eprints.eemcs.utwente.nl/12734/*.

[11] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. State-dependent importance sampling for a slow-down tandem queue. *Submitted*, 2008. See also Memorandum 1879, Dept. of Applied Mathematics, University of Twente, *URL: http://eprints.eemcs.utwente.nl/13251/*.

[12] 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.

[13] G. Rubino and B. Tuffin. *Rare Event Simulation using Monte Carlo Methods*. John Wiley & Sons, To appear in 2009.

[14] 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.

[15] T.S. Zaburnenko and V.F. Nicola. Efficient heuristics for simulating population overflow in tandem networks. *Proceedings of the Fifth St. Petersburg Workshop on Simulation*, pages 755–764, 2005.