Probability, Networks and Algorithms

PNA

State-dependent importance sampling for a slow-down tandem queue

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

REPORT PNA-R0811 AUGUST 2008

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 slow-down tandem queue

ABSTRACT

In this paper we investigate an advanced variant of the classical (Jackson) tandem queue, viz. a two-node system with server slow-down. The slow-down mechanism has the primary objective to protect the downstream queue from frequent overflows, and it does so by reducing the service speed of the upstream queue as soon as the number of jobs in the downstream queue reaches some pre-specified threshold. To assess the efficacy of such a policy, techniques are needed for evaluating overflow metrics of the second queue. We focus on the estimation of the probability of the following rare event: overflow in the downstream queue before exhausting the system, starting from any given state in the state space. Due to the rarity of the event under consideration, naïve, direct Monte Carlo simulation is often infeasible. We therefore rely on the application of importance sampling to obtain variance reduction. The principal contribution of this paper is that we construct an importance sampling scheme that is asymptotically efficient. In more detail, the paper addresses the following issues. (i) We rely on powerful heuristics to identify the exponential decay rate of the probability under consideration, and verify this result by applying sample-path large deviations techniques. (2) Immediately from these heuristics, we develop a proposal for a change of measure to be used in importance sampling. (3) We prove that the resulting algorithm is asymptotically efficient, which effectively means that the number of runs required to obtain an estimate with fixed precision grows subexponentially in the buffer size. We stress that our method to prove asymptotic efficiency is substantially shorter and more straightforward than those usually provided in the literature. Also our setting is more general than the situations analyzed so far, as we allow the process to start off at any state of the state space, and in addition we do not impose any conditions on the values of the arrival rate and service rates, as long as the underlying queueing system is stable.

2000 Mathematics Subject Classification: 60K25, 65C05, 60F10
Keywords and Phrases: Importance sampling, queues, asymptotic efficiency
Note: Part of this research has been funded by the Dutch Bsik/BRICKS project.

State-dependent Importance Sampling for a Slow-down Tandem Queue[*]

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

Abstract

In this paper we investigate an advanced variant of the classical (Jackson) tandem queue, viz. a two-node system with server slow-down. The slow-down mechanism has the primary objective to protect the downstream queue from frequent overflows, and it does so by reducing the service speed of the upstream queue as soon as the number of jobs in the downstream queue reaches some pre-specified threshold. To assess the efficacy of such a policy, techniques are needed for evaluating overflow metrics of the second queue. We focus on the estimation of the probability of the following rare event: overflow in the downstream queue before exhausting the system, starting from any given state in the state space.

Due to the rarity of the event under consideration, naïve, direct Monte Carlo simulation is often infeasible. We therefore rely on the application of importance sampling to obtain variance reduction. The principal contribution of this paper is that we construct an importance sampling scheme that is asymptotically efficient. In more detail, the paper addresses the following issues. (i) We rely on powerful heuristics to identify the exponential decay rate of the probability under consideration, and verify this result by applying sample-path large deviations techniques. (2) Immediately from these heuristics, we develop a proposal for a change of measure to be used in importance sampling. (3) We prove that the resulting algorithm is asymptotically efficient, which effectively means that the number of runs required to obtain an estimate with fixed precision grows subexponentially in the buffer size. We stress that our method to prove asymptotic efficiency is substantially shorter and more straightforward than those usually provided in the literature. Also our setting is more general than the situations analyzed so far, as we allow the process to start off at any state of the state space, and in addition we do not impose any conditions on the values of the arrival rate and service rates, as long as the underlying queueing system is stable.

[*] 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; part of this work was done while he was at Stanford University, Stanford, CA 94305, US.

1 Introduction

There is a vast body of literature on Jacksonian networks, and in particular tandem queues. Being applicable in a broad range of domains, such as communication networks, manufacturing, and logistics, they have been subject of intensive research for over fifty years. Owing to its special features, particularly the fact that its steady-state distribution is of product-form, various performance metrics could be analyzed explicitly.

Changing the model slightly, often means that the product-form is lost, and that the analysis becomes cumbersome. One such a variant of the classical Jacksonian tandem queue is the the two-node system with *server slow-down*, also known as a system with *backpressure.* This mechanism is designed to offer the second (or: *downstream*) queue some sort of protection against frequent overflows: as long as the number of jobs in the downstream queue is larger than some pre-specified threshold the server of the first (or: *upstream*) queue slows down, and it returns to its normal speed when the number of jobs in the second queue drops below the threshold again. For this model only partial results are available, see e.g. [20]. It is noted that the slow-down model is of significant practical interest, as a related mechanism has been proposed e.g. in the design of Metro Ethernet [12, 16].

Lacking explicit formulas for the queue's steady-state distribution, several alternative approaches can be pursued. In this paper we highlight two such approaches: we asymptotically characterize the probability of interest (when the buffer size of the downstream queue grows large, and the value of the threshold is scaled accordingly), but emphasis lies on the development of efficient simulation techniques, based on importance sampling (IS). It is noted that due to the rarity of the event under consideration, naïve, direct Monte Carlo simulation is often infeasible. The idea of IS is to simulate the system under a different probability distribution (often referred to as the 'new measure'), under which the event of interest occurs more frequently. After correcting the simulation output by means of likelihood ratios, an unbiased estimate is obtained. We refer to e.g. [10] for an introduction to IS and its background.

The asymptotics that we present in this work rely on powerful heuristics developed in [15], that identify the exponential decay rate associated to the probability under consideration as the solution of a, relatively easy, convex programming problem. The correctness of the heuristics is then proven by applying techniques known as *sample-path large-deviations* [6, 19]. Importantly, this procedure also reveals the so-called *typical path to overflow*: given that the rare event occurs, then with overwhelming probability this happens by a path 'close to' the typical path. Having this path at our disposal, the next step is to use this knowledge in designing IS algorithms.

When developing efficient IS schemes, various complications arise. The most important of these is that state-independent new measures, which often worked well in the case of a single queues [17, 18], usually fail for networks that are intrinsically more complex, see for instance [2, 9] for the case of the ordinary Jacksonian tandem model. It was concluded that the class of state-independent IS is not sufficiently rich to construct asymptotically efficient new measures, explaining the increasing interest in state-*dependent* IS schemes. An early reference on such state-dependent schemes, in the context of a certain class of queueing networks, is [3], but an analytic proof of efficiency properties was lacking there. Later such proofs for related new measures were given in, e.g., [8].

Let us now consider this paper's contributions in more detail. The primary goal is to analyze the probability of overflow in the downstream ('protected') queue, before the system idles, starting off from any given state. Special cases of this problem were already studied in [7, 13]. These papers exclusively consider the case of starting in the origin, which is substantially more straightforward to address. In [13] a 'pseudo-state-dependent' IS scheme was proposed for estimating the overflow in the second queue. Its asymptotic efficiency, for a limited set of initial parameters, was concluded, but just on the basis of empirical evidence. In [7] a provably asymptotically efficient new measure was proposed; as mentioned above, the analysis was restricted to the case that the system started empty, and in addition certain assumptions on the model parameters (i.e., arrival rate at the upstream queue, and service rates) were imposed. Our paper therefore generalizes [7, 13], in that all initial states are allowed, and that there are no assumptions on the values of the arrival rate and service rates, as long as the underlying queueing system is stable. An important additional contribution is that our proof of asymptotic efficiency is rather elementary and short, compared to that in [7].

It is mentioned that there are several technicalities related to the way the new measure should be constructed close to the axes, and close to the slow-down threshold. We considered a similar technique for the ordinary two-node Jacksonian tandem model in [15]. The approach followed there, however, could not be directly applied in the current model. The main complication lies in the discontinuity of the transition structure along the slow-down threshold. As a consequence, the typical paths to overflow can have a rather complex structure. In addition, the way the new measure should be constructed close to the threshold is non-trivial.

We like to stress that the focus on this paper lies on the analytic aspects of the problem, that is, the analysis of the decay rate and the proof of asymptotic efficiency. We decided to refrain from including numerical experiments in this paper, as these form a topic of research in their own right. This is due to the fact that it is still a rather nontrivial step from an asymptotically efficient procedure, as the one presented here, to an actual, efficient implementation of the algorithm. It is noted that several aspects that we did not mention above play a crucial role: it matters for instance very much whether a new measure requires computation 'on the fly' of new transition rates, or whether they can be precomputed. These issues we plan to address in forthcoming work.

The paper in structured as follows. The basics of IS are recapitulated in Section 2.1, whereas a model description is given in Section 2.2. Then, after having heuristically identified the shapes of the most probable path to overflow, we present in Section 3 IS schemes for estimating the probability of interest; as a by-product we find the corresponding exponential decay rate. The explanation of how sample-path large-deviations techniques are used to rigorize these findings, we leave to the Appendix. In Section 4 we slightly modify the IS schemes designed in Sections 3.4 – 3.6 and prove the asymptotic efficiency of the resulting scheme. We end the paper with some conclusions in Section 5.

Let us finish this introduction with a few words on the person in honor of whom this workshop has been organized, Reuven Rubinstein. Reuven has played a pivotal role at the interface of the simulation community and the applied probability. Relying on his unsurpassable intuition, he succeeded now and again to fuel the applied probabilists with new revolutionary ideas. We do recognize the crucial role 'intuition' plays in the design of 'good' simulation techniques – the present paper is very much in that spirit: it describes how heuristics can be transformed into provably efficient methods. We hope Reuven's beautiful contributions to the area will continue after his retirement.

2 Preliminaries

In this section we present a short overview on the main concepts in importance sampling, and we introduce our model.

2.1 Importance sampling

As we mentioned in the introduction, estimating small probabilities through direct, naïve, simulations is often infeasible, due to the rarity of the event involved; the simulation effort needed to obtain an estimate of given precision could be prohibitively large. We therefore have to use variance reduction techniques, and in this paper we focus on importance sampling (IS). IS performs simulations of the system under a new measure, which guarantees more frequent occurrence of the event of interest. After weighing the simulation output with the appropriate likelihood ratios (keeping track of the likelihood of the realization under the original measure with respect to the new measure), we obtain an unbiased estimator. The focus lies on *state-dependent* IS schemes, that is, schemes in which the new measure may depend on the current state of the system.

Let us now give a generic description of IS, as well as the concept of asymptotic efficiency. To this end, consider a family of rare events $\{A_B\}$, in the probability space $(\Omega, \mathfrak{F}, \mathbb{P})$; here B is the so-called 'rarity parameter'. To estimate $\mathbb{P}(A_B)$ via IS simulations one need to generate samples under a new probability measure $\mathbb{Q}$, with respect to which $\mathbb{P}$ is absolutely continuous. The probability $\mathbb{P}(A_B)$ can now alternatively be expressed as

$$\mathbb{P}(A_B) = \mathbb{E}^{\mathbb{Q}}[LI], \tag{1}$$

where I is an indicator function and L is the likelihood ratio (also known as Radon-Nikodým derivative) of a realization ('path') ω:

$$L = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\omega). \tag{2}$$

We see that we can sample under $\mathbb{Q}$, say n times, obtain observations $(L_1, I_1), \ldots, (L_n, I_n)$, and construct the unbiased estimator of $\mathbb{P}(A_B)$ by $n^{-1} \cdot \sum_{i=1}^{n} L_i I_i$.

Clearly some alternative measures $\mathbb{Q}$ perform better than others, in terms of the variance of the resulting estimator. Therefore the following concept has been introduced.

Definition 2.1. The IS scheme for $\mathbb{P}(A_B)$ is called *asymptotically efficient* if

$$\liminf_{B\to\infty} \frac{\log \mathbb{E}^{\mathbb{Q}}[L^2 I]}{\log \mathbb{P}(A_B)} \geq 2. \tag{3}$$

If, in addition, the probability of the $\{A_B\}$ decays exponentially in B, i.e.,

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(A_B) \in (-\infty, 0),$$

the definition of asymptotic efficiency reduces to

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathbb{E}^{\mathbb{Q}}[L^2 I] \leq 2 \lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(A_B).$$

Notice that $\mathbb{E}^{\mathbb{Q}}[L^2 I] = \mathbb{E}[LI]$, so the above criterion can alternatively be written as

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathbb{E}[LI] \leq 2 \lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(A_B). \tag{4}$$

2.2 Slow-down tandem queue

In this subsection we describe the two-node slow-down network. At the first node (or: station) jobs arrive according to a Poisson process of rate λ. Each job receives service at the first station, after which it is routed to the second one. After receiving service at the second node, the job leaves the system. Service times at the second station have an exponential distribution with parameter μ_2. At the first node, however, the service speed depends on the content of the second queue: normally, service times at the first station have an exponential distribution with parameter μ_1, but if the number of jobs in the second queue exceeds some pre-specified threshold (the 'slow-down threshold') than the parameter of the exponential distribution changes to μ_1^+, where $\mu_1^+ < \mu_1$. When the system 'stabilizes', that is, the number of jobs in the second queue drops again below the slow-down threshold, the service rate of the first station returns to its original value μ_1.

For convenience we choose the parameters such that $\lambda + \mu_1 + \mu_2 = 1$, without loss of generality (and thus $\lambda + \mu_1^+ + \mu_2 < 1$). 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, which is regenerative if we impose that it is stable, see [14] for more insights into this issue. Our main interest is to estimate the probability of reaching some high level B in the second queue before it returns to the origin, starting from any given state. Note that in our model the slow-down threshold scales with B, that is, the threshold has value θB in the remainder of this paper, for some $\theta \in (0, 1)$.

The queue-length process can also be recorded at jump epochs, and then it is 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. We define the possible jump directions of the process Q_j via vectors $v_0 = (1, 0)$, $v_1 = (-1, 1)$, and $v_2 = (0, -1)$, with corresponding jump rates λ, μ_1 (or μ_1^+) and μ_2 respectively.

For convenience we will also consider the so-called *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. More specifically, 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 visits the origin. We introduce the following subsets of the state space, with $x := (x_1, x_2)$:

$$\begin{aligned} D &:= \{x : x_1 > 0, 0 < x_2 < \theta\}, & \partial_1 &:= \{(0, x_2) : x_2 > 0\}, & \partial_2 &:= \{(x_1, 0) : x_1 > 0\}, \\ D^+ &:= \{x : x_1 > 0, \theta \le x_2 < 1\}, & \partial_1^+ &:= \{(0, x_2) : x_2 \in [\theta, 1)\}, & \partial_\theta &:= \{(x_1, \theta) : x_1 > 0\}, \\ & & & & \partial_e &:= \{(x_1, 1) : x_1 > 0\}. \end{aligned}$$

The full state space is $\bar{D} \cup \bar{D}^+$, where $\bar{D} := D \cup \partial_\theta \cup (\partial_1 \setminus \partial_1^+) \cup \partial_2$ and $\bar{D}^+ := D^+ \cup \partial_e \cup \partial_1^+ \cup \partial_\theta$. Note that the 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-loop transitions in the following way (see also Figure 1): for $k = 1, 2$,

$$\mathbb{P}(X_{j+1} = X_j | X_j \in \partial_k) = \mu_k, \quad \mathbb{P}(X_{j+1} = X_j | X_j \in \partial_1^+) = \mu_1^+/(\lambda + \mu_1^+ + \mu_2). \tag{5}$$

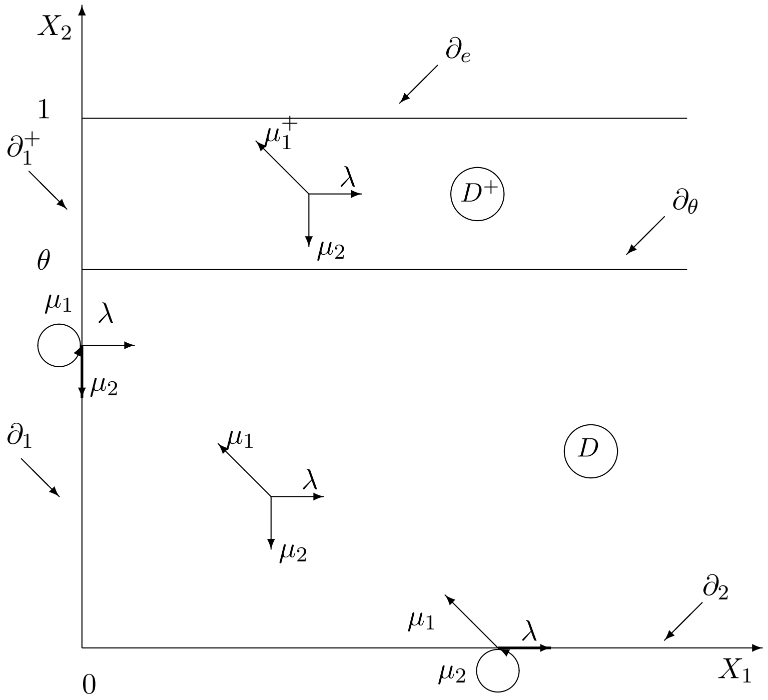

Figure 1: State space and transition structure for the scaled process X_j.

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 any visits to the origin:

$$\tau_B^x = \inf\{k > 0 : X_k \in \partial_e, X_j \neq 0 \text{ for } j = 1, \ldots, k-1\}, \tag{6}$$

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

$$p_B^x := \mathbb{E} I_B(A^x) = \mathbb{P}(\tau_B^x < \infty). \tag{7}$$

3 State-dependent importance sampling

In order to find a 'good' new 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 Subsections 3.1-3.3 we explain a method in which minimizing certain 'cost-functions' leads to the most probable path and a good corresponding new measure, given by new (state-dependent) transition rates $\tilde{\lambda}(x), \tilde{\mu}_1(x)$ and $\tilde{\mu}_2(x)$ below the slow-down threshold and $\tilde{\lambda}^+(x), \tilde{\mu}_1^+(x)$, $\tilde{\mu}_2^+(x)$ above it. Also, the minimal cost itself will be shown to be the decay rate of p_B^x as $B \to \infty$, which will play a pivotal role in the asymptotic efficiency proofs later.
The results of the minimization procedure are presented in three different subsections, since they are different, depending on the parameters settings. In Subsection 3.4 we treat the case $\mu_2 < \mu_1^+ < \mu_1$ in which the second server is always the bottleneck, Subsection 3.5 deals with the case $\mu_1^+ \leq \mu_2 < \mu_1$ in which either the first or the second server is the bottleneck, and in Subsection 3.6 we describe the case $\mu_1^+ < \mu_1 \leq \mu_2$, in which the first server is always the bottleneck. Beforehand we would like to point out that the new measure mentioned above and denoted by tildes, is not exactly the same as the asymptotically efficient new measure that will be introduced in Section 4 (denoted by bars), although it is closely related.

3.1 Path to overflow

The typical path to overflow in the very special case that the origin is the starting state and $\theta \in \{0, 1\}$, has already been identified in [1]. In that paper the time-reversed process is used to find the shape of the most probable path to overflow. In the more general setting that $\theta \in [0, 1]$, but again with the origin as the only starting state, the path to overflow was obtained in [13]. In this section we present a method similar to the one in [13] to find the optimal path starting from *any* state $x \in \bar{D} \cup \bar{D}^+$. The advantage of this method is that it provides us insight into the typical behavior conditional on observing the rare event under consideration; our choice for the new measure (which we prove to be asymptotically efficient) will be inspired on it.
Before introducing our method we state a property that says that, when searching the typical path to overflow, we can restrict ourselves to a (small) subset of all feasible paths. We leave the proof that this typical path satisfies these restrictions for the Appendix.

Property 3.1. We only consider paths that satisfy the following:

(i) Each path is a concatenation of subpaths, which are straight lines on any of the subsets $D, D^+, \partial_1 \setminus \partial_1^+, \partial_1^+$ and ∂_2, and the measure stays constant along each subpath, i.e., $\tilde{\lambda}(x) = \tilde{\lambda}$, $\tilde{\mu}_1(x) = \tilde{\mu}_1$, $\tilde{\mu}_2(x) = \tilde{\mu}_2$, $\tilde{\lambda}^+(x) = \tilde{\lambda}^+$, $\tilde{\mu}_1^+(x) = \tilde{\mu}_1^+$ and $\tilde{\mu}_2^+(x) = \tilde{\mu}_2^+$, for any state x on the same subpath;

(ii) each path does not have more than two subpaths in each subset if $\mu_1^+ < \mu_1 \leq \mu_2$, and more than one subpath per subset otherwise.

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 queue starting from state x is the decay rate of the probability of interest (see Section 3.3). Our method is based on the family of cost functions I, defined by

$$I(\tilde{\lambda} \mid \lambda) = \lambda - \tilde{\lambda} + \tilde{\lambda} \log \frac{\tilde{\lambda}}{\lambda}; \tag{8}$$

see also [19], p. 14, 20. Note that the function (8) is convex and equals 0 at $\tilde{\lambda} = \lambda$. Intuitively, the value $I(\tilde{\lambda} \mid \lambda)$ is the cost we need to 'pay' per time unit to let a Poisson process with parameter λ behave like a Poisson process with parameter $\tilde{\lambda}$.

In the following we will explain our cost method in some detail. More background can be found in Section 3.1 of [15] and in the Appendix of [13].

3.2 Example

As a leading example, we here consider a path consisting of two linear pieces, 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 $x_2 < \theta < y_2$ (the last condition meaning that the path crosses the slow-down threshold). We focus on computing the typical path that connects x with y (and in particular the point where it crosses the threshold), and the corresponding new measure.

To this end, we construct new measures $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$, such that $\tilde{\mu}_1 > \tilde{\mu}_2$, $\tilde{\lambda} \leq \tilde{\mu}_1$, $\tilde{\mu}_1^+ > \tilde{\mu}_2^+$ and $\tilde{\lambda}^+ \leq \tilde{\mu}_1^+$. Under these measures our path consists of two linear subpaths and each of them has a constant north-west drift. In other words: below the slow-down threshold our path has a constant slope $-\alpha$, with

$$\alpha = \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\mu}_1 - \tilde{\lambda}}, \tag{9}$$

while above the threshold it has a constant slope $-\alpha^+$, with

$$\alpha^+ = \frac{\tilde{\mu}_1^+ - \tilde{\mu}_2^+}{\tilde{\mu}_1^+ - \tilde{\lambda}^+}. \tag{10}$$

Below the slow-down threshold, the cost of this path is, per unit time,

$$\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = I(\tilde{\lambda} \mid \lambda) + I(\tilde{\mu}_1 \mid \mu_1) + I(\tilde{\mu}_2 \mid \mu_2). \tag{11}$$

To find the cost per unit horizontal (vertical) distance, we need to divide this cost by the horizontal speed $\tilde{\mu}_1 - \tilde{\lambda}$ (vertical speed $\tilde{\mu}_1 - \tilde{\mu}_2$). Similar expressions apply for the costs per unit time and unit distance, when the process is *above* the slow-down threshold. Thus, minimizing the cost of any path that consists of two straight subpaths (one strictly below the threshold and one above it) from x to y in this case boils down to minimizing

$$(\theta - x_2)\frac{\mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)}{\tilde{\mu}_1 - \tilde{\mu}_2} + (y_2 - \theta)\frac{\mathbb{I}(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)}{\tilde{\mu}_1^+ - \tilde{\mu}_2^+}, \tag{12}$$

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

$$\kappa(x) := x_1 - \frac{\tilde{\mu}_1 - \tilde{\lambda}}{\tilde{\mu}_1 - \tilde{\mu}_2}(\theta - x_2) = y_1 + \frac{\tilde{\mu}_1^+ - \tilde{\lambda}^+}{\tilde{\mu}_1^+ - \tilde{\mu}_2^+}(y_2 - \theta), \tag{13}$$

where $(\kappa(x), \theta)$ is the state in which the optimal path crosses the slow-down threshold.

One way to solve the minimization problem (12) is the following; from now on we focus on the ending state $y = (0, 1)$, as this will later turn out to be the most likely point of entering ∂_e in many situations. For each fixed crossing state $(\kappa(x), \theta)$, we can find the cost of the path through that state. Then, we minimize this cost over all possible values of κ. Note that the optimal value $\kappa(x)$ is a function of the starting state x. This property complicates the shape of the optimal paths significantly, as well as the analysis of the new measure.

The total cost of the bottom part of the optimal path, i.e., the subpath from x to $(\kappa(x), \theta)$ attains its minimum when the triplet $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ is a solution to

$$\begin{cases} \tilde{\lambda} = \tilde{\mu}_1 + \frac{\kappa(x) - x_1}{\theta - x_2}(\tilde{\mu}_1 - \tilde{\mu}_2) \\ \tilde{\lambda} + \tilde{\mu}_1 + \tilde{\mu}_2 = \lambda + \mu_1 + \mu_2 \\ \tilde{\lambda}\tilde{\mu}_1\tilde{\mu}_2 = \lambda\mu_1\mu_2 \\ \tilde{\lambda} \leq \tilde{\mu}_1 \text{ and } \tilde{\mu}_1 > \tilde{\mu}_2 \\ \tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2 > 0 \end{cases} \tag{14}$$

Similarly, the total cost of the subpath above the threshold from $(\kappa(x), \theta)$ to $(0, 1)$ is minimal when $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$ is a solution to

$$\begin{cases} \tilde{\lambda}^+ = \tilde{\mu}_1^+ - \frac{\kappa(x)}{1 - \theta}(\tilde{\mu}_1^+ - \tilde{\mu}_2^+) \\ \tilde{\lambda}^+ + \tilde{\mu}_1^+ + \tilde{\mu}_2^+ = \lambda + \mu_1^+ + \mu_2 \\ \tilde{\lambda}^+\tilde{\mu}_1^+\tilde{\mu}_2^+ = \lambda\mu_1^+\mu_2 \\ \tilde{\lambda}^+ \leq \tilde{\mu}_1^+ \text{ and } \tilde{\mu}_1^+ > \tilde{\mu}_2^+ \\ \tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+ > 0 \end{cases} \tag{15}$$

Notice also that if $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ is the solution to (14) 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 $(\kappa(x), \theta)$. Similarly, $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$ which is the solution of (15) stays unchanged for the whole top part of the optimal path.

It will be useful to define the functions γ_1 and γ_2 as the cost of the subpaths (of the optimal path to overflow) below and above the thresholds, i.e.,

$$\gamma_1(x_1, x_2) := -(x_1 - \kappa(x)) \log \frac{\tilde{\lambda}(x_1, x_2)}{\lambda} - (\theta - x_2) \log \frac{\tilde{\mu}_2(x_1, x_2)}{\mu_2}, \tag{16}$$

$$\gamma_2(\kappa(x), \theta) := -\kappa(x) \log \frac{\tilde{\lambda}^+(\kappa(x), \theta)}{\lambda} - (1 - \theta) \log \frac{\tilde{\mu}_2^+(\kappa(x), \theta)}{\mu_2}, \tag{17}$$

where $\tilde{\lambda}$ and $\tilde{\mu}_2$ are given by the solution to (14), $\tilde{\lambda}^+$ and $\tilde{\mu}_2^+$ by the solution to (15), and $\kappa(x)$ is given in (13). Then clearly the total cost of the path $(x_1, x_2) \to (\kappa(x), \theta) \to (0, 1)$ can be expressed as $\gamma_1(x_1, x_2) + \gamma_2(\kappa(x), \theta)$.

3.3 Decay rate as minimal cost

Once we have considered all possible path types with their minimal cost, we can obtain the overall minimum cost, corresponding to the most probable path, and the corresponding (state-dependent) new measures $(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x))$ and $(\tilde{\lambda}^+(x), \tilde{\mu}_1^+(x), \tilde{\mu}_2^+(x))$. Defining

$$\gamma(x) := \text{ overall minimal cost over all paths } x \to \partial_e,$$

the following theorem states that this is in fact the exponential decay rate of the probability p_B^x as $B \to \infty$. It is based on a large deviation principle for the process $X(t)$ (with a local rate function that is closely related to our cost function) that can be found in the Appendix.

Theorem 3.2. *The exponential decay rate of p_B^x is equal to the minimal cost of overflow $\gamma(x)$, i.e.,*

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

We now present the value of $\gamma(x)$ (as well as the corresponding new measures) for the three cases mentioned above (that is, second server is bottleneck, 'shifting bottleneck', and first server is bottleneck).

3.4 Importance sampling scheme for $\mu_2 < \mu_1^+ < \mu_1$

In this case, where the second queue is always the bottleneck, the new measure under which the path to overflow has minimal cost, in terms of the cost function (8), turns out to be given by

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_2, \mu_1, \lambda), & \text{if } x \in A_1, \\ \text{solution to (14)}, & \text{if } x \in A_2, \\ (\lambda, \mu_1, \mu_2), & \text{if } x \in A_3, \end{cases} \tag{18}$$

and

$$(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+) = \begin{cases} (\mu_2, \mu_1^+, \lambda), & \text{if } x \in A_1^+, \\ \text{solution to (15)}, & \text{if } x \in A_2^+, \\ (\lambda, \mu_1^+, \mu_2), & \text{if } x \in A_3^+. \end{cases} \tag{19}$$

Here the subsets A_i and A_i^+, $i = 1, 2, 3$ form a partition of the state space $\bar{D} \cup \bar{D}^+$ as depicted in Figure 2, where $\alpha_1 := (\mu_1 - \mu_2)/(\mu_1 - \lambda)$ and $\alpha_1^+ := (\mu_1^+ - \mu_2)/(\mu_1^+ - \lambda)$. We chose not give the precise definitions of the sets A_i and A_i^+ here, since they do not add much to the understanding. For some starting states x, Figure 2 also shows the shape of the most probable path to ∂_e.

Note that the new measure in the subsets A_1 and A_1^+, i.e., interchanging λ and μ_2, has been earlier found in [17] for the problem of reaching a large *total* network population. Measures similar to the ones in the other subsets were introduced in [15]. Also, we point out that the new measure is continuous in the state x, as can be verified by solving system (14) for $x = (\alpha_1\theta + \alpha_1^+(1-\theta), 0)$ and $x = (\theta\alpha_1 + (1-\theta)/\alpha_1^+, 0)$, yielding the solutions in the first and third lines of (18), respectively. A similar principle holds above the slow-down threshold as well.

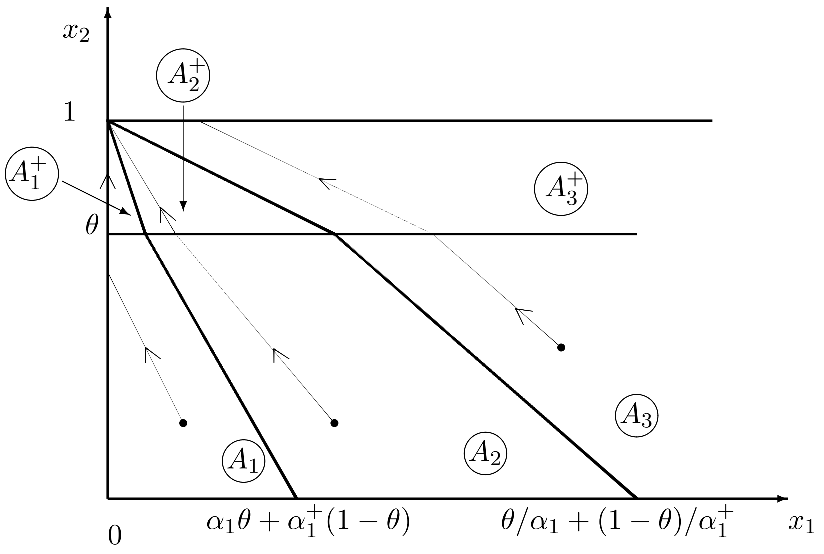
x_2
A_2^+
1
A_1^+
A_3^+
θ
A_3
A_1
A_2
0
$\alpha_1\theta + \alpha_1^+(1-\theta)$
$\theta/\alpha_1 + (1-\theta)/\alpha_1^+$
x_1

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

The cost $\gamma(x)$ of the optimal path, and hence the decay rate of p_B^x can be expressed as:

$$\gamma(x) = \begin{cases} (1 - x_1 - x_2)\gamma, & \text{if } x \in A_1, \\ \gamma_1(x_1, x_2) + \gamma_2(\kappa(x), \theta), & \text{if } x \in A_2, \\ 0, & \text{if } x \in A_3, \end{cases} \tag{20}$$

where

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

is the decay rate of the path $(0,0) \to (0,1)$, γ_1 and γ_2 are as in (16) and (17), and $\kappa(x)$ is given in (13). Importantly, we treat only paths with starting state below the slow-down threshold; starting states above the threshold are substantially easier to deal with.

3.5 Importance sampling scheme for $\mu_1^+ \leq \mu_2 < \mu_1$

In the case where the bottleneck may shift from the second to the first station due to the slowdown mechanism, the new measure under which the path to overflow has minimal cost, in terms of cost function (8), is given by

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_2, \mu_1, \lambda), & \text{if } x \in B_1, \\ \text{solution to (14)}, & \text{if } x \in B_2, \\ (\lambda, \mu_1, \mu_2), & \text{if } x \in B_3. \end{cases} \tag{21}$$

and

$$(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+) = \begin{cases} \text{solution to (15)}, & \text{if } x \in B_2^+, \\ (\lambda, \mu_2, \mu_1^+), & \text{if } x \in B_3^+. \end{cases} \tag{22}$$

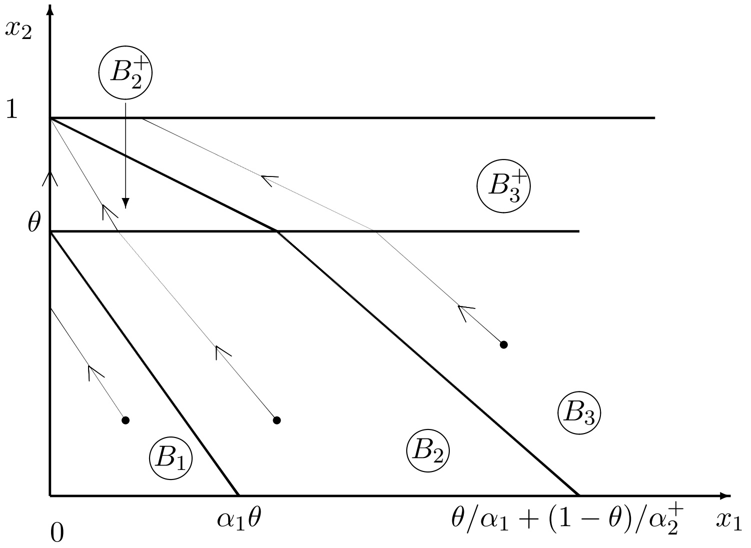

Figure 3: Partition of $\bar{D} \cup \bar{D}^+$ and some optimal paths to overflow when $\mu_1^+ \leq \mu_2 < \mu_1$.

Here, the five subsets B_1, B_2, B_3, B_2^+ and B_3^+ are shown in Figure 3, which is comparable to Figure 2. The main difference is that there is no set B_1^+, and the constant α_1^+ is replaced by $\alpha_2^+ := (\mu_2 - \mu_1^+)/(\mu_2 - \lambda)$, while α_1 is the same as introduced in the previous subsection.
The decay rate $\gamma(x)$ is now given by

$$\gamma(x) = \begin{cases} \theta(1 - x_1 - x_2)\gamma + (1-\theta)\log(1/z^+), & \text{if } x \in B_1, \\ \gamma_1(x_1, x_2) + \gamma_2(\kappa(x), \theta), & \text{if } x \in B_2, \\ (1-\theta)\log(\mu_2/\mu_1^+), & \text{if } x \in B_3. \end{cases} \tag{23}$$

Here z^+ is the unique solution in $(0, 1)$ of the equation

$$\lambda + \mu_1^+ + \mu_2(1 - z^+) = 2\sqrt{\frac{\lambda\mu_1^+}{z^+}}, \tag{24}$$

which follows from system (14) by taking $(x_1, x_2) = (0, 0)$. In fact, $(1-\theta)\log(1/z^+)$ is the cost of the vertical path $(0, \theta) \to (0, 1)$ *in the interior*, satisfying $\tilde{\lambda} = \tilde{\mu}_1^+$. See also Equations (30) and (33) in [11] and [13] respectively, for more details.

3.6 Importance sampling scheme for $\mu_1^+ < \mu_1 \leq \mu_2$

When the first queue is always the bottleneck *and* θ is not too small (to be made more precise at the end of this subsection), the new measure under which the path to overflow has minimal cost, in terms of cost function (8), is given by

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_1, \lambda, \mu_2), & \text{if } x \in C_1, \\ \text{solution to (14)}, & \text{if } x \in C_2, \\ (\lambda, \mu_2, \mu_1), & \text{if } x \in C_3. \end{cases} \tag{25}$$

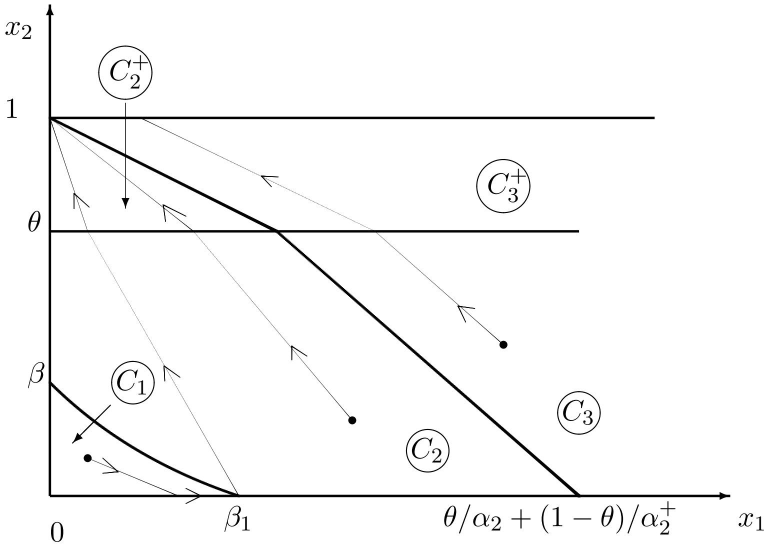

Figure 4: Partition of $\bar{D} \cup \bar{D}^+$ and some optimal paths to overflow when $\mu_1^+ < \mu_1 \leq \mu_2$.

and

$$(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+) = \begin{cases} \text{solution to (15)}, & \text{if } x \in C_2^+, \\ (\lambda, \mu_2, \mu_1^+), & \text{if } x \in C_3^+. \end{cases} \tag{26}$$

The sets C_1, C_2, C_3, C_2^+ and C_3^+ are shown in Figure 4, where $\alpha_2 := (\mu_2 - \mu_1)/(\mu_2 - \lambda)$; the constants β and β_1 are given shortly. Interestingly, for the current case the behavior under the new measure is entirely different on C_1 and C_2, and the measure is *not* continuous in states that lie on the boundary between these two sets. For such states x, the cost of the path 'upwards', which is $\gamma_1(x) + \gamma_2(\kappa(x), \theta)$, is equal to the cost of the path 'to the right', which can be shown to be $\theta\gamma + (1-\theta)\log(1/q) - x_1 \log(\mu_1/\lambda)$ with q being the unique solution in $(0, \lambda\mu_1^+/\mu_1^2)$ of the equation

$$\mu_1\mu_2 q^2 + \mu_1(\mu_1 - \lambda - \mu_1^+ - \mu_2)q + \lambda\mu_1^+ = 0 \tag{27}$$

(see also Proposition 14 in [13] for more background). Thus, the boundary between C_1 and C_2 is the zero level curve of the function

$$f(x) = \theta\gamma + (1-\theta)\log(1/q) - x_1\log(\mu_1/\lambda) - \gamma_1(x_1, x_2) - \gamma_2(\kappa(x), \theta).$$

The intersection point of this curve with the horizontal axis lies at $(\beta_1, 0)$ with

$$\beta_1 := \theta(\mu_2 - \mu_1)/(\mu_2 - \lambda) + (1-\theta)(\lambda\mu_1^+ - \mu_1^2 q)/(\lambda\mu_1^+ - \mu_1\mu_2 q^2).$$

The intersection point $(0, \beta)$ with the vertical axis follows as the unique solution to

$$f(0, \beta) = \theta\gamma + (1-\theta)\log(1/q) - (\theta - \beta)\log(1/z) - (1-\theta)\log(1/z^+) = 0, \tag{28}$$

where z is the unique solution in $(0, 1)$ of

$$1 - \mu_2 z = 2\sqrt{\frac{\lambda\mu_1}{z}}, \tag{29}$$

which is the analogue of Equation (24), with μ_1^+ replaced by μ_1.
The main result of this subsection is the decay rate, which is in this case given by:

$$\gamma(x) = \begin{cases} \theta\gamma + (1-\theta)\log(1/q) - x_1 \log(\mu_1/\lambda), & \text{if } x \in C_1, \\ \gamma_1(x_1, x_2) + \gamma_2(\kappa(x), \theta), & \text{if } x \in C_2, \\ (\theta - x_2)\log(\mu_2/\mu_1) + (1-\theta)\log(\mu_2/\mu_1^+), & \text{if } x \in C_3. \end{cases} \tag{30}$$

We finally return to our assumption that θ should not be too small. In particular we used in the above that the threshold lies above the set C_1, i.e., $\theta > \beta$. When μ_2 is very large compared to λ and μ_1, the corresponding value of β may be rather large, because the cost of 'upward' paths will be much larger than the cost of paths 'to the right'. However, for most 'real-life' cases, β is quite small; in the most interesting (heavily loaded) cases, β is still below 0.1. Of course we could also consider cases where β is larger than θ. This will lead to minor changes in the structure of C_1, C_2 and C_2^+, and the minimal cost $\gamma(x)$ will then also change; we chose to leave this special case out.

3.7 Properties of the new measures

We like to summarize some important properties of the new measures $(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x))$ and $(\tilde{\lambda}^+(x), \tilde{\mu}_1^+(x), \tilde{\mu}_2^+(x))$ described in Sections 3.4–3.6 in the following proposition. The first two statements hold independently of the relation between the parameters λ, μ_1, μ_1^+ and μ_2, and show that the functions $\tilde{\lambda}(x), \tilde{\lambda}^+(x), \tilde{\mu}_2(x)$ and $\tilde{\mu}_2^+(x)$ depend monotonically on x. The last two statements give bounds which *do* depend on the relation between the parameters.

Proposition 3.3. *For any $x \in \bar{D} \cup \bar{D}^+$ the functions $\tilde{\lambda}(x)$, $\tilde{\mu}_1(x)$, $\tilde{\mu}_2(x)$ and $\tilde{\lambda}^+(x)$, $\tilde{\mu}_1^+(x)$, $\tilde{\mu}_2^+(x)$, as defined by either (18) and (19), or by (21) and (22), or by (25) and (26) satisfy the following:*

(i) $\dfrac{\partial\tilde{\lambda}(x)}{\partial x_1} \le 0,\ \dfrac{\partial\tilde{\mu}_2(x)}{\partial x_1} \ge 0,\ \dfrac{\partial\tilde{\lambda}(x)}{\partial x_2} \le 0$ *and* $\dfrac{\partial\tilde{\mu}_2(x)}{\partial x_2} \ge 0$
(if $\mu_1^+ < \mu_1 \le \mu_2$, then assume that $x \notin C_1$);

(ii) $\dfrac{\partial\tilde{\lambda}^+(x)}{\partial x_1} \le 0,\ \dfrac{\partial\tilde{\mu}_2^+(x)}{\partial x_1} \ge 0,\ \dfrac{\partial\tilde{\lambda}^+(x)}{\partial x_2} \le 0$ *and* $\dfrac{\partial\tilde{\mu}_2^+(x)}{\partial x_2} \ge 0;$

(iii) $\tilde{\lambda}(x) \in [\lambda, \mu_2]$ *and* $\tilde{\mu}_2(x) \in [\lambda, \mu_2]$ *if* $\mu_2 < \mu_1$, *and*
$\tilde{\lambda}(x) \in [\lambda, \sqrt{\lambda\mu_1/z}]$ *and* $\tilde{\mu}_2(x) \in [\mu_2 z, \mu_1] \cup \{\mu_2\}$ *if* $\mu_2 \ge \mu_1;$

(iv) $\tilde{\lambda}^+(x) \in [\lambda, \mu_2]$ *and* $\tilde{\mu}_2^+(x) \in [\lambda, \mu_2]$ *if* $\mu_2 < \mu_1^+$ *and*
$\tilde{\lambda}^+(x) \in [\lambda, \sqrt{\lambda\mu_1^+/z^+}]$ *and* $\tilde{\mu}_2^+(x) \in [\mu_2 z^+, \mu_1^+]$ *if* $\mu_2 \ge \mu_1^+$.

Here, as before, z^+ is defined by (24) and z by (29).

4 Asymptotic efficiency

For the special case in which the starting state is the origin, it is known from [13] that the new measures provided in Section 3 are not always asymptotically optimal. For example, in the simplest case, when $\mu_2 < \mu_1^+ < \mu_1$, multiple visits of the process $Q(t)$ to the horizontal axis ∂_2 may lead to large likelihood ratios of particular sample paths under the new measure (μ_2, μ_1, λ). This critically impacts the quality of the estimator. To avoid this behavior we use a technique similar to what was proposed in [8] and also used in [15]. It is based on using a specific measure around ∂_2, under which visits to ∂_2 are harmless to the likelihood ratio. Thus, in this section we will introduce new measures (indicated by bars) based on the measures from the previous section (indicated by tildes), and subsequently prove that these new measures are indeed asymptotically efficient. As in the previous section, we split the problem into three cases. In Section 4.1 we explain our method in detail for the situation in which the second server is always the bottleneck ($\mu_2 < \mu_1^+ < \mu_1$), and in Sections 4.2 and 4.3 we treat the other cases.

4.1 Asymptotic efficiency for $\mu_2 < \mu_1^+ < \mu_1$

In this subsection we present a modification of the scheme constructed in Subsection 3.4 and prove its asymptotic optimality. At first we introduce the function $W(x)$ for any point $x = (x_1, x_2)$ of the state space. This function will give us expressions for the new measures, denoted by bars, similar to how it was done in [8, 15]. In particular we will now find such a measure both below $(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x))$ and above $(\bar{\lambda}^+(x), \bar{\mu}_1^+(x), \bar{\mu}_2^+(x))$ the threshold.
For some small $\delta > 0$, let us first introduce three auxiliary functions $W_i(x), i = 1, 2, 3$:

$$\begin{aligned} W_1(x) &:= 2\gamma_2(x) I_{\{x \in \bar{D}^+\}} + 2\left(\gamma_1(x) + \gamma_2(\kappa(x), \theta)\right) I_{\{x \in \bar{D}\}} - \delta, \\ W_2(x) &:= 2\gamma(x_1, \delta/2\gamma) - \delta, \\ W_3(x) &:= 2\gamma(0) - 3\delta, \end{aligned} \tag{31}$$

where $\gamma(x)$, $\gamma_1(x)$ and $\gamma_2(\kappa(x), \theta)$ are as in (20), see also (16) and (17); we recall that γ equals $-\log(\lambda/\mu_2)$.
In the next step we introduce the minimum of these auxiliary functions:

$$\bar{W}(x) := W_1(x) \wedge W_2(x) \wedge W_3(x). \tag{32}$$

Since $W_2(x) = W_1(x_1, \delta/2\gamma)$, it follows that this minimum is only attained by W_2 in a narrow strip along the horizontal axis, namely when $x_2 \leq \delta/2\gamma$, unless x is close to the origin, in which case W_3 is the minimum. In all other states we simply have $\bar{W}(x) = W_1(x)$.
The last step is a mollification procedure, in which we define:

$$W(x) := -\epsilon \log \sum_{i=1}^{3} e^{-W_i(x)/\epsilon}. \tag{33}$$

The resulting function $W(x)$ is a 'smoothed' version of $\bar{W}(x)$, except on the threshold, where W_1 is not differentiable. The 'smoothness' of $W(x)$ depends on the choice of the parameter ϵ: the larger ϵ is chosen, the smoother the function $W(x)$ is. On the other hand, as $\epsilon \downarrow 0$ we see that $W(x)$ converges to the (non-smooth) function $\bar{W}(x)$.

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, see [8]. For conciseness, we often suppress the index B.

Assumption 4.1. *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) $\epsilon_B/\delta_B \to 0$.*

The following expression for the gradient of $W(x)$ is immediate from (33), and will play an important role in the representation of the state-dependent, asymptotically efficient new measure:

$$DW(x) = \sum_{k=1}^{3} \rho_k(x) DW_k(x), \text{ where } \rho_k(x) := \frac{e^{-W_k(x)/\epsilon}}{\sum_{i=1}^{3} e^{-W_i(x)/\epsilon}}. \tag{34}$$

Also, we have the following helpful property.

Proposition 4.2. *The gradients of the functions $W_i(x)$, $i = 1, 2, 3$ are given by:*

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

Proof. It is clear that $DW_1(x) = -2\gamma(1,1)$ if $x \in A_1 \cup A_1^+$ and $DW_1(x) = (0,0)$ if $x \in A_3 \cup A_3^+$. When $x \in A_2$, $DW_1(x)$ seems to be more complicated:

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

This gradient is more involved than its analog for the standard Jackson tandem (see Proposition 5.1 in [15]), not only because we now have two new measures (below and above the slowdown threshold), but also due to the strong dependence on x of the optimal path shape (and the optimal crossing state $(\kappa(x), \theta)$ in particular). Fortunately, the second term is zero because $\partial\gamma(x)/\partial\kappa(x) = 0$ due to the fact that $(\kappa(x), \theta)$ is the *optimal* crossing state. Also, applying implicit differentiation one can find the partial derivatives of all 'tilded' variables ($\tilde{\lambda}(x)$, etc.) and show that the vectors in the second and third lines sum up to zero.

The other statements (including the case when $x \in A_2^+$) follow easily from the definitions of $W_i(x)$, $i = 1, \ldots, 3$. □

Now we are ready to define the new measure, see also (27) and (29) in [15]:

$$\begin{aligned}
\bar{\lambda}(x) &:= \lambda e^{-\langle DW(x),v_0\rangle/2} e^{\mathbb{H}(DW(x))/2}, && \text{if } x \in \bar{D},\\
\bar{\mu}_i(x) &:= \mu_i e^{-\langle DW(x),v_i\rangle/2} e^{\mathbb{H}(DW(x))/2}, \quad i=1,2, && \text{if } x \in \bar{D},\\
\bar{\lambda}^+(x) &:= \frac{\lambda}{\lambda+\mu_1^+ + \mu_2} e^{-\langle DW(x),v_0\rangle/2} e^{\mathbb{H}^+(DW(x))/2}, && \text{if } x \in \bar{D}^+,\\
\bar{\mu}_1^+(x) &:= \frac{\mu_1^+}{\lambda+\mu_1^+ + \mu_2} e^{-\langle DW(x),v_1\rangle/2} e^{\mathbb{H}^+(DW(x))/2}, && \text{if } x \in \bar{D}^+,\\
\bar{\mu}_2^+(x) &:= \frac{\mu_2}{\lambda+\mu_1^+ + \mu_2} e^{-\langle DW(x),v_2\rangle/2} e^{\mathbb{H}^+(DW(x))/2}, && \text{if } x \in \bar{D}^+.
\end{aligned} \tag{35}$$

Note that the functions $\tilde{\lambda}(x)$, etc. from the previous section are transition *rates*, while the functions $\bar{\lambda}(x)$, etc. are transition *probabilities* under the new measure (just as λ and $\lambda/(\lambda + \mu_1^+ + \mu_2)$ are transition probabilities under the original measure when $x \in \bar{D}$ resp. $x \in \bar{D}^+$). The functions $\mathbb{H}(DW(x))$ and $\mathbb{H}^+(DW(x))$ in the new measure (35) are known as *Hamiltonians*, which we use to enable the comparison with [4, 8, 15]; in fact they provide the normalization such that the new transition probabilities sum up to 1. More precisely,

$$\mathbb{H}(DW(x)) := 2\log\left[\lambda e^{-\langle DW(x),v_0\rangle/2} + \mu_1 e^{-\langle DW(x),v_1\rangle/2} + \mu_2 e^{-\langle DW(x),v_2\rangle/2}\right]^{-1}$$

and

$$\mathbb{H}^+(DW(x)) := 2\log\left[\frac{\lambda e^{-\langle DW(x),v_0\rangle/2}}{\lambda+\mu_1^+ + \mu_2} + \frac{\mu_1^+ e^{-\langle DW(x),v_1\rangle/2}}{\lambda+\mu_1^+ + \mu_2} + \frac{\mu_2 e^{-\langle DW(x),v_2\rangle/2}}{\lambda+\mu_1^+ + \mu_2}\right]^{-1}.$$

Now that we defined the change of measure in (35), we are ready to prove that it is asymptotically efficient. We start with some lemmas that are similar to the ones in [4].

Lemma 4.3. *The likelihood $L(A)$ of a path $A = (X_j, j = 0, \ldots, \sigma)$ under the new measure (35) satisfies*

$$\begin{aligned}
\log L(A) &= \frac{B}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1} - X_j\rangle\\
&+ \sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), v_k\rangle I\{X_j = X_{j+1} \in \partial_k\}\\
&- \frac{1}{2}\sum_{j=0}^{\sigma-1}\left(\mathbb{H}(DW(X_j))I_{\{X_j\in D\}} + \mathbb{H}^+(DW(X_j))I_{\{X_j\in D^+\}}\right).
\end{aligned} \tag{36}$$

Proof. The proof is analogous to the proof of Lemma 1 in [4]. □

Lemma 4.4. *For any path $A = (X_j, j = 0, ..., \sigma)$ under the new measure (35), the first term in (36) satisfies*

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

for sufficiently large $B\epsilon$, where C and C^+ are some positive constants and σ^+ is the number of slow-down threshold crossings up to time σ.

Proof. Our argument is based on the representation

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

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 $r(x,y)$ is such that $\lim_{|y|\to 0} r(x,y) = 0$, except when x and $x+y$ are separated by the slow-down threshold. In the latter case we can bound $r(x,y)$ from above, uniform in x, as follows:

$$r(x,y) \leq 2BC^+,$$

where C^+ is some positive constant, based on a uniform upper bound on $|DW(x) - DW(x+y)|$. To end the proof, we refer to Lemma 5.5 in [15] for the following bound that holds when x and $x+y$ are not separated by the slow-down threshold,

$$\left|\frac{1}{2}y^T H(x)y + |y|^2 r(y)\right| \leq \frac{2C}{B^2\epsilon},$$

where C is some positive constant. □

Lemma 4.5. *For any $x \in D$ we have $\mathbb{H}(DW(x)) \geq 0$, and for any $x \in D^+$ we have $\mathbb{H}^+(DW(x)) \geq 0$.*

Proof. For any $x \in \bar{D}^+$ we have

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

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

where the last inequality is found by considering the convex function $f(x) := (x + \lambda\mu_1^+/x + \mu_2)/(\lambda + \mu_1^+ + \mu_2)$; since $f(\lambda) = f(\mu_1^+) = 1$ it follows that $f(x) < 1$ for any $x \in [\lambda, \mu_2] \subset [\lambda, \mu_1^+]$. Finally we also have

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

Combining these bounds with representation (34) and keeping in mind the concavity of $\mathbb{H}^+(x)$ (thanks to Proposition 3.2 in [8]) we obtain

$$\mathbb{H}^+(DW(x)) = \mathbb{H}^+\left(\sum_{i=0}^{3}\rho_i(x)DW(x)\right) \geq \sum_{i=0}^{3}\rho_i(x)\mathbb{H}^+\left(DW_i(x)\right) \geq 0,$$

for any $x \in \bar{D}^+$.
The proof of the other statement is analogous, or follows from Lemma 5.3 in [15]. □

Lemma 4.6. *Consider the slow-down network and recall the definition of τ_B^x in (6). For any sequence υ_B such that $\lim_{B\to\infty}\upsilon_B = 0$ the following limit holds:*

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

Proof. We define a new random variable τ which represents the same random period of time as τ_B^x, but for the case when $\theta = 0$, i.e., for the two-node tandem Jackson network with parameters $(\lambda, \mu_1^+, \mu_2)$. It is clear that $\tau_B^x \leq^{st} \tau$. From Lemma 5.6 in [15] we know that for the standard Jacksonian tandem network

$$\lim_{B\to\infty}\frac{1}{B}\log\mathbb{E}(e^{\upsilon_B\tau}|I_B(A^x)=1) = 0,$$

for any υ_B satisfying $\lim_{B\to\infty}\upsilon_B = 0$. This completes the proof. □

Theorem 4.7. *When $\mu_2 < \mu_1^+ < \mu_1$ and Assumption 4.1 holds, the new measure in (35), where the function W is based on (20), is asymptotically optimal.*

Proof. We will roughly follow the proof of Theorem 5.7 in [15], finding upper bounds on each of the three terms in Lemma 4.3.
To deal with the first term, we first bound $W(x)$ from below. Upon combining the fact that $W_2(x) \geq W_1(x) - \delta$ for any $x \in \bar{D} \cup \bar{D}^+$ (this is shown in the same manner as in Thm. 5.7 of [15]; use (20)), with the monotonicity of $\gamma(x)$ in both x_1 and x_2, and using definition (33), it is found that

$$\begin{aligned} W(x) &\geq -\epsilon\log(e^{-W_1(x)/\epsilon} + e^{(-W_1(x)+\delta)/\epsilon} + e^{-W_3(x)/\epsilon}) \\ &\geq -\epsilon\log(3e^{(-2\gamma(x)+3\delta)/\epsilon}) = 2\gamma(x) - \epsilon\log(3) - 3\delta. \end{aligned} \tag{37}$$

Using the same technique we obtain an upper bound for $W(X_{\tau_B^x})$:

$$W(X_{\tau_B^x}) \leq -\delta. \tag{38}$$

Combining the inequalities (37)-(38) with Lemma 4.4 (take $\sigma = \tau_B^x$), we now derive the following upper bound on the first term in Lemma 4.3:

$$\frac{B}{2}\sum_{j=0}^{\tau_B^x-1}\langle DW(X_j), X_{j+1}-X_j\rangle \leq \frac{B}{2}(-2\gamma(x)+\eta(B)) + \frac{C}{B\epsilon}\tau_B^x + C^+\tau_B^{x,+}, \tag{39}$$

where C and C^+ are some positive constants, $\eta(B)$ is such that $\lim_{B\to\infty}\eta(B) = 0$ (use Assumption 4.1), and $\tau_B^{x,+}$ is the number of slow-down threshold crossings up to time τ_B^x.

Now let us bound the second term in Lemma 4.3. For any $x \in \partial_2$ we have $\langle DW_2(x), -v_2\rangle = \langle DW_3(x), -v_2\rangle = 0$ and $\langle DW_1(x), -v_2\rangle = 2\log(\tilde{\mu}_2/\mu_2)$; applying (34) we arrive at

$$\langle DW(x), -v_2\rangle = 2\log\left(\frac{\tilde{\mu}_2}{\mu_2}\right)\rho_1(x) \geq -2\gamma\rho_1(x) \geq -2\gamma e^{-(W_1(x)-W_2(x))/\epsilon}, \tag{40}$$

where the first inequality comes from the fact that $\tilde{\mu}_2 \geq \lambda$ (see Proposition 3.3). It is also clear that $W_1(x) - W_2(x) = \delta$ for any $x \in A_1 \cap \partial_2$, where the functions $W_1(x)$ and $W_2(x)$ are defined by (31). The second statement of Proposition 3.3 guarantees that the difference $W_1(x) - W_2(x)$ decreases to 0 as x goes from $(\alpha_1, 0)$ to $(\alpha_1^{-1}, 0)$. From here we can immediately find $0 \leq W_1(x) - W_2(x) \leq \delta$, which implies:

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

Using the same technique and keeping Proposition 3.3 in mind, one can also show that

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

for any $x \in \partial_1$. Using these inequalities we can bound the second term in Lemma 4.3 from above:

$$\sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\tau_B^x-1}\langle DW(X_j), v_k\rangle I\{X_j = X_{j+1} \in \partial_k\} \leq \gamma e^{-\delta/\epsilon}\tau_B^x. \tag{41}$$

Finally note that Lemma 4.5 provides a straightforward bound on the last term of the log-likelihood expression in Lemma 4.3:

$$\mathbb{H}(DW(X_j))I_{\{X_j\in D\}} + \mathbb{H}^+(DW(X_j))I_{\{X_j\in D^+\}} \geq 0. \tag{42}$$

Upon combining (39), (41) and (42), we bound (36) in the following way:

$$\log(L(A)) \leq -B\gamma(x) + B\frac{\eta(B)}{2} + \chi(B)\tau_B^x + C^+\tau_B^{x,+},$$

where

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

Now for any path A^x we have

$$\begin{aligned}
\frac{1}{B}\log\mathbb{E}\left[L(A^x)I_B(A^x)\right] &= \frac{1}{B}\log(\mathbb{E}\left[L(A^x)|I_B(A^x)=1\right]\mathbb{P}\left[I_B(A^x)=1\right]) \\
&\leq \frac{1}{B}\log\left(\mathbb{E}\left[e^{-B\gamma(x)+B\eta(B)+\chi(B)\tau_B^x+C^+\tau_B^{x,+}}|I_B(A^x)=1\right]p_B^x\right) \\
&= -\gamma(x) + \frac{\eta(B)}{2} + \frac{1}{B}\log\mathbb{E}\left[e^{\chi(B)\tau_B^x}|I_B(A^x)=1\right] \\
&\quad +\frac{1}{B}\log\mathbb{E}\left[e^{C^+\tau_B^{x,+}}|I_B(A^x)=1\right] + \frac{1}{B}\log p_B^x.
\end{aligned}$$

Using that $\lim_{B\to\infty} \tau_B^{x,+}/B = 0$ a.s. when $I_B(A^x) = 1$, and that $\lim_{B\to\infty} \chi(B) = 0$ (see Assumption 4.1), and invoking Lemma 4.6 and Theorem 3.2, we conclude that

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

In view of criterion (4), this completes the proof. □

4.2 Asymptotic efficiency for $\mu_1^+ \le \mu_2 < \mu_1$

Remarkably, we can use the same function $W(x)$ for this case as in the previous subsection, see (33); also we define the new measures $\left(\bar\lambda(x), \bar\mu_1(x), \bar\mu_2(x)\right)$ and $\left(\bar\lambda^+(x), \bar\mu_1^+(x), \bar\mu_2^+(x)\right)$ in the same way, see (35).

One difference with the previous case is that Lemma 4.5 no longer holds. In fact the Hamiltonians can be negative now, but the next lemma shows that they vanish as B grows large; recall that due to Assumption 4.1 we have that $1/\epsilon \to \infty$ and $\delta/\epsilon \to \infty$ as $B \to \infty$.

Lemma 4.8. *For any $x \in D$ we have $\mathbb{H}(DW(x)) \ge 0$, and for any $x \in D^+$ we have*

$$\mathbb{H}^+(DW(x)) \ge -C^\star e^{-(\theta - \frac{\delta}{2\gamma})\gamma/\epsilon}$$

for some finite constant $C^\star > 0$.

Proof. Using the same technique as in the proof of Lemma 4.5, one can prove that the first statement holds, while for the second statement we find

$$\mathbb{H}^+(DW(x)) \ge \rho_2(0,\theta)\mathbb{H}^+(DW_2(0,x_2)),$$

if $x \in D^+$. The second factor on the right hand side is in fact a constant, since DW_2 does not depend upon its second argument. Furthermore it is negative, and

$$\rho_2(0,\theta) \le \frac{e^{-W_2(0,\theta)/\epsilon}}{e^{-W_1(0,\theta)/\epsilon}} \le e^{-(\theta - \frac{\delta}{2\gamma})\gamma/\epsilon},$$

so that the claim follows. □

Theorem 4.9. *When $\mu_1^+ \le \mu_2 < \mu_1$ and Assumption 4.1 holds, the new measure in (35), where the function W is based on (23) and (17), is asymptotically optimal.*

Proof. In order to prove this theorem we again bound all three terms of the log-likelihood ratio, see Lemma 4.3. For the first two terms we find exactly the same as in (39) and (41). As for the third term, this is now bounded in Lemma 4.8. Thus, we find again

$$\log(L(A)) \le -B\gamma(x) + B\frac{\eta(B)}{2} + \chi(B)\tau_B^x + C^+\tau_B^{x,+},$$

only now

$$\chi(B) = \gamma e^{-\delta/\epsilon} + \frac{C}{\epsilon B} + \frac{C^\star}{2} e^{-(\theta - \frac{\delta}{2\gamma})\gamma/\epsilon},$$

which also vanishes as B grows large. From now on we can follow the proof of Theorem 4.7 which leads us to the result. □

4.3 Asymptotic efficiency for $\mu_1^+ < \mu_1 \leq \mu_2$

For the case when $\mu_1^+ < \mu_1 \leq \mu_2$ we again use the function $W(x)$ defined by (33), to describe the IS scheme as in (35), which we can prove to be asymptotically efficient. It is important that in this case the function $W_2(x)$ plays a more important role than before. Because of the structure and the cost of the optimal path to overflow we now have

$$\bar{W}(x) = W_2(x), \text{ for any } x \in C_1 \cup \{x : x_2 \leq \delta/2\gamma\},$$

see also Figure 4. In the two previous subsections this was only valid on $\{x : x_2 \leq \delta/2\gamma\}$.

Theorem 4.10. *When $\mu_1^+ < \mu_1 \leq \mu_2$ and Assumption 4.1 holds, the new measure in (35), where the function W is based on (30) and (17), is asymptotically optimal.*

Proof. Not surprisingly, the proof of this theorem is almost the same as that of Theorem 4.9. Even Lemma 4.8 remains valid in this case. The only essential difference is the behavior of the function $W(x)$ on the constraints ∂_1 and ∂_2. This leads to a different bound on the second term of the log-likelihood in Lemma 4.3. As in Theorems 4.7 and Theorem 4.9 we have for any $x \in \partial_2$,

$$\langle DW(x), -v_2 \rangle \geq -2\log(1/z)\rho_1(x),$$

but for $\rho_1(x)$ we now have by Theorem 5.8 in [15] that

$$\rho_1(x) \geq \exp\left(-\frac{2\beta \log(1/z)}{\epsilon}\right),$$

where β is unique solution to the equation $f(0,\beta) = 0$, see also (28). We conclude, that for any $x \in \partial_2$,

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

For $x \in \partial_1$ we have, again due to Theorem 5.8 in [15]:

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

The rest of the proof can be done by mimicking the arguments used in Thm. 4.7 or Thm. 4.9. □

5 Conclusions

This paper focused on constructing IS schemes for estimating the probability of a specific rare event: overflow in the second queue of the slow-down network before the system idles, starting from any given state. We proved asymptotic efficiency of the proposed new measure. The analysis heavily relied on large-deviations argumentation.

One can look at this result from two different perspectives. On one hand, this is the continuation of our earlier work on rare-event simulation in a two-node Jacksonian tandem network, see [15].

On the other hand, this paper can be viewed as the generalization of already existed research on the slow-down network, see [7, 13]. We rigorized and further studied the empirical findings of [13]. Also, [7, 13] consider the restrictive case that the only possible starting state is the origin. In our paper we developed IS schemes for all three possible cases (second queue bottleneck, 'shifting bottleneck', first queue bottleneck), unlike [7] that specializes to a specific ordering of the parameters: $\lambda < \mu_1^+ < \mu_2 \leq \mu_1$ (which is covered by our 'shifting bottleneck' case $\mu_1^+ \leq \mu_2 < \mu_1$). In our schemes one may pick an arbitrary starting state x. An important by-product of our analysis is a precise description of the typical path to overflow (in the second queue), starting in an arbitrary state. Although our proofs use specific properties of the model at hand, we strongly feel that our methodology carries over to more general classes of queues.

As indicated in the introduction, a next challenge is to transform the methods presented in this paper into simulation programs. We stress that, even with an asymptotically efficient new measure at our disposal, new questions come up: should we compute the new measure 'on the fly' (that is, while running the program), or precompute it (and store it)? Also, it may pay off to partition the state space into a small number of sets, and to approximate the state-dependent change of measure by new measures that are constant on these sets. A detailed simulation study, as well as extensive practical guidelines, will be presented in a forthcoming paper.

A Appendix. Large deviations

The goal of this appendix is to establish the result that the cost of the optimal path to overflow is equal to the exponential decay rate of our probability of interest, see Theorem 3.2. We also highlight a number of important and interesting large-deviations properties of the process $X(t)$. Let us consider any absolutely continuous function $\phi : [0,\infty) \to \bar{D} \cup \bar{D}^+$, 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 it, first we define four auxiliary functions $L_i(y)$, where the argument y should be interpreted as a 'speed vector':

$$L_i(y) := \sup_{\vartheta} \left(\langle \vartheta, y \rangle - g_i(\vartheta) \right), \ i = 1, \ldots, 4, \tag{43}$$

and where

$$\begin{aligned}
g_1(\vartheta) &:= \lambda(e^{\vartheta_1} - 1) + \mu_1(e^{\vartheta_2 - \vartheta_1} - 1) + \mu_2(e^{-\vartheta_2} - 1),\\
g_2(\vartheta) &:= \lambda(e^{\vartheta_1} - 1) + \mu_1^+(e^{\vartheta_2 - \vartheta_1} - 1) + \mu_2(e^{-\vartheta_2} - 1),\\
g_3(\vartheta) &:= \lambda(e^{\vartheta_1} - 1) + \mu_2(e^{-\vartheta_2} - 1),\\
g_4(\vartheta) &:= \lambda(e^{\vartheta_1} - 1) + \mu_1(e^{\vartheta_2 - \vartheta_1} - 1);
\end{aligned}$$

cf. (5.5) in [19]. It is observed that $g_1(\cdot)$ corresponds to D, $g_2(\cdot)$ to D^+, $g_3(\cdot)$ to ∂_1, and $g_4(\cdot)$ to ∂_2.

Now we can define the local rate function ℓ as:

$$\ell(\phi(t),\dot{\phi}(t)) := \begin{cases} L_1(\dot{\phi}(t)), & \text{if } \phi(t) \in D, \\ L_2(\dot{\phi}(t)), & \text{if } \phi(t) \in D^+ \cup \partial_e, \\ [L_1 \oplus L_3\,](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_1 \setminus \partial_1^+, \\ [L_2 \oplus L_3\,](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_1^+, \\ [L_1 \oplus L_2\ \oplus L_3\,](\dot{\phi}(t)), & \text{if } \phi(t) = (0,\theta), \\ [L_1 \oplus L_4](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_2, \\ [L_1 \oplus L_2\,](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_\theta, \end{cases} \tag{44}$$

where, for $n \geq 2$, and y denoting a two-dimensional vector,

$$[L_1 \oplus \ldots \oplus L_n](y) := \inf \left\{ \sum_{i=1}^{n} \rho_i L_i(y_i) : \rho_i \geq 0, \sum_{i=1}^{n} \rho_i = 1, \sum_{i=1}^{n} \rho_i y_i = y \right\},$$

is the *inf-convolution* of the functions $L_1, \ldots, L_n$, the infimum being taken over all values ρ_i and vectors y_i, $i = 1, \ldots, n$, that satisfy the given conditions.
We now 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 $\phi(t) \in \partial_1 \setminus \partial_1^+$, such that $\partial\phi_2/\partial t > 0$ for $t \in [0,T]$. Hence, the first and the second components 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 the v_1 type are not allowed on ∂_1. Therefore the inf-convolution provides a 'mixture' of the functions L_1 and L_3, supposing that the process $Q(t)$ spends a fraction of time ρ in the lower part of 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 in ∂_1.

Now we are ready to state the following theorem.

Theorem A.1. *The process $X(t)$ satisfies a large deviations principle with rate function (44), 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 $\phi : [0,\infty) \to \bar{D} \cup \bar{D}^+$ such that $\phi(0) = x$ and $\tau < \infty$.

Proof. We sketch the proof of this result, as it is reminiscent of results proven in [7]; see also the proof of Thm. 4.1 in [15]. The main arguments are taken from [5, 6].
We first introduce the process $Z(t)$, which is an unconstrained version of $X(t)$, that is, $Z(t)$ is allowed to have negative values in both components. In addition we will assume that $Z(0) = X(0) = x \in \bar{D} \cup \bar{D}^+$. One can then use Theorems 3.2 and 3.4 of [6] 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 referred as the Skorokhod problem; for more details see [6].
Since the map Γ is Lipschitz continuous and the process $Z(t)$ satisfies a large deviation principle (see Theorem 7.2.3 of [5]), one can apply the contraction principle (see Theorem 2.13 of [19]) and conclude that the process of our interest, $X(t)$, satisfies a large deviations principle with rate function $\ell(\phi(t), \dot{\phi}(t))$ defined by (44). □

To prove Thm. 3.2, we now recapitulate the main findings of Section 4 in [15]. Using the local rate function ℓ 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. At first let us mention the following property: for the paths that stay in one of the subsets $D, D^+, \partial_1 \setminus \partial_1^+, \partial_1^+, \partial_2$, the rate function (44) is minimal when the path is straight, with constant speed vector; see Lemma 5 of [15], and p. 87 of [19].
Now we assume that $\phi(t) \in D$, for $t \in (0, T)$ is a path between two states x and y. We know 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:

$$\begin{aligned}\tilde{\lambda} &= \lambda e^{\vartheta_1}, \\ \tilde{\mu}_1 &= \mu_1 e^{\vartheta_2 - \vartheta_1}, \\ \tilde{\mu}_2 &= \mu_2 e^{-\vartheta_2},\end{aligned} \tag{45}$$

where $\vartheta = (\vartheta_1, \vartheta_2)$ is the maximizer in (43) with $i = 1$. In fact this is exactly the same new measure we would find using the cost minimization procedure from Section 3, due to the immediate equality

$$\ell(\phi(t), \dot{\phi}(t)) = \mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2); \tag{46}$$

see again [15]. 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 (45). However, for the optimal paths we still have equality between local rate functions and cost functions on the boundaries, see Lemma 6 in [15]. Let, e.g., Φ_1 be the set of paths that travels a distance $h > 0$ along $\partial_1 \setminus \partial_1^+$ at constant speed during a time σ, i.e.,

$$\Phi_1 \;\; = \;\; \{\phi(t) \subset \partial_1 \setminus \partial_1^+ : \phi(0) = (0, x_2^\star), \phi(\sigma) = (0, x_2^\star + h)\},$$

for some $x_2^\star$. Then we have the following relation

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

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 $\tilde{\mu}_1 > \tilde{\mu}_2$. We have a similar situation for paths that follow ∂_1^+ or the horizontal constraint ∂_2.
Our last result, which is an analogue of Lemma 7 in [15], regulates the number of subpaths of the optimal path to overflow.

Lemma A.2. *The optimal path from any starting state x to the exit boundary ∂_e does not have more than*

- *(i) two subpaths in each subset, if $\mu_1^+ < \mu_1 \leq \mu_2$, and*
- *(ii) one subpath in each subset otherwise.*

Finally, using all the results in this appendix, Thm. 3.2 follows; the proof is analogous to the proof of Thm. 8 in [15].

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 R.S. Ellis. *A weak convergence approach to the theory of Large deviations*. John Wiley & Sons, New York, 1997.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] R. Malhotra, M. Mandjes, W. Scheinhardt, and H. van den Berg. A feedback fluid queue with two congestion control thresholds. To appear in*: Mathematical Methods in Operations Research*, 2008.

[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] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. Tandem queue with server slow-down. *ACM Sigmetrics Performance Evaluation Review*, 35(3):51–52, 2007.

[15] 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/*.

[16] W. Noureddine and F. Tobagi. Selective back-pressure in switched Ethernet LANs. In *Global Telecommunications Conference, 2*, 1999.

[17] 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.

[18] 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.

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

[20] 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.