QUEUEING NETWORKS: RARE EVENTS AND FAST SIMULATIONS

DENIS MIRETSKIY

Queueing networks: rare events and fast simulations

Denis Miretskiy

Graduation committee

Chairman

prof.dr.ir. A.J. Mouthaan	University of Twente

Promoters

prof.dr. M.R.H. Mandjes	University of Amsterdam
prof.dr. R.J. Boucherie	University of Twente

Co-promoter

dr.ir. W.R.W. Scheinhardt	University of Twente

Members

prof.dr. I.J.B.F Adan	Eindhoven University of Technology
dr. J. Blanchet	Columbia University
dr.ir. P.T. de Boer	University of Twente
dr.ir. E.A. van Doorn	University of Twente
dr. A.A.N. Ridder	VU University Amsterdam
prof.dr. W.H.M. Zijm	University of Twente

BRICKS

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

Beta
Research School for Operations Management and Logistics

Beta Dissertation Series **D 126**
Research School for Operations Management and Logistics

ISBN 978-90-365-2909-9

QUEUEING NETWORKS: RARE EVENTS AND FAST SIMULATIONS

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Thursday the 12$^{\text{th}}$ of November 2009 at 15.00

by

Denis Miretskiy

born on the 17$^{\text{th}}$ of March 1982
in Volzhsky, Russia

This dissertation has been approved by

prof.dr. R.J. Boucherie (promoter)
prof.dr. M.R.H. Mandjes (promoter)
dr.ir. W.R.W. Scheinhardt (co-promoter)

Acknowledgements

From the very beginning I would like to thank my supervisor Werner Scheinhardt. I am deeply grateful for the time we spent in the discussions, for his faithful suggestions, his great patience and certainly for the great work atmosphere we had for the last four years. Above all I would like to thank Werner for his help outside the university. I am sure that this thesis could not have its current form without the help and supervision of my promoter Michel Mandjes. I would like to express my gratitude for his irreplaceable ideas, motivation and for steering me in the right direction. Finally, I would like to thank my other promoter and chair of the Stochastic Operations Research group Richard Boucherie for giving me the opportunity to complete my Ph.D. research in The Netherlands. I am also grateful to him for introducing new horizons of precision and accuracy to me.

I would like to acknowledge the rest of my graduation committee: Ton Mouthaan, Jose Blanchet, Pieter-Tjerk de Boer, Henk Zijm, Erik van Doorn, Ivo Adan and Ad Ridder for their valuable inputs and efforts in reviewing this thesis.

I would like to thank all of my SOR colleagues for a pleasant and efficient job climate. My special gratitude goes to my constant officemate Tom Coenen, who was not only an excellent companion, but one of my main guides into Dutch culture. I would like to acknowledge all members of the SOR group: Ahmad Al Hanbali, Frank Warrink, Jan-Kees van Ommeren, Jasper Goseling, Judith Timmer, Maartje Zonderland, Maurits de Graaf, Nelly Litvak, Nikky Kortbeek, Roland de Haan, Thyra Kamphuis-Kuijpers and Yana Volkovich. My distinct recognition goes to Yana, who was not only a co-worker but a friend, for all great times we had in the past years.

I would like to express my gratitude to my former supervisors Victor Goryainov and Alexander Gnedin for arising and maintaining an interest for stochastics in me.

Next, I am happy to thank all my friends in The Netherlans, namely Jon, Barbara, Domi, Zsofi, Uros, Paul, Tanya, Dima, Markus, Georgia, Mitya, Peter, Connie, Des, Clare, Ove, Romen, Natcha, Natasha and Sabrina for being such great guys. My special appreciation goes to Jon my roommate, friend and a great person to watch ice hockey with. Thanks to Domi for lots of 'Goulash' events, and to Paul for keeping me updated about important changes in Enschede's life. I would like to mention my friends in Russia, especially Ilya, Katya, Lesha, Andrey, Sveta and Sveta, Nastya,

Anya, Dima and Anton. And of course I am very thankful to my mom, dad and all my family for their support and belief!

Last but not least I am infinitely thankful to my girlfriend Katya. It is difficult or even impossible to underestimate her input to my life and to this work in particular. Katya, thank you for your love, understanding, support and everything else you have been doing for me.

Contents

Chapter 1

Introduction

Almost everyone has heard about rare events, but not everyone realizes what they actually are. We will start this thesis with some examples of rare events.

Winning millions in a lottery is definitely rare. We do not know much about how to win, but we can estimate the probability of success and make a decision regarding this event, namely, to invest our money in lottery tickets or not.

Even if the class of rare events were restricted to the above, it would be worthwhile to investigate. However, it is much wider and consists not only of events that we hope for. Some rare events might have dramatic or even tragic consequences. The crash of the Concorde on July 25, 2000 is an example.

Until that day the Concorde was considered to be one of the safest planes. However, a sequence of unlucky and extremely rare occurrences turned the regular flight into a disaster and changed the history of civil supersonic aviation. During takeoff the Concorde ran over a strip of metal, which had fallen from another aircraft, and damaged one of the tyres. The debris hit the wing structure, leading to a rupture of one of the fuel tanks. A major fire broke out almost immediately. During the next several minutes all four engines were out of operation and the aircraft crashed onto a hotel. See [18] for the official report of the French authorities.

This tragic occurrence definitely might have been prevented. If a metal strip had not fallen from the previous aircraft; if the tyre had been produced from a different material; if the fuel tank had had a better protection... Then 113 people would not have been killed on July 25, 2000.

This example not only shows that analyzing rare events may be very important, even though they hardly ever occur, but also shows that a rare event usually consist of a concatenation of events that are less rare.

This work is dedicated to constructing efficient and reliable methods for estimating probabilities of a particular type of rare events, namely rare events in queueing networks. When compared to the example above, these rare events may seem rela-

tively harmless, but, as we argue, it is of substantial interest to be able to estimate the corresponding probabilities.

Let us consider the following example. Alice needs to send Bob a piece of important information. This data can be sent via a number of different routes which can be identified by a number of traversed nodes (routers). Even one incorrectly working (e.g. overflowed) router can cause loss or contortion of the information, which may have significant consequences on the data transmission.

The rare event in this example is the overflow in some router's buffer, which may occur, if during an extended period of time (which is a concatenation of many unit time periods) the number of arriving packets is unusually large and/or if the processing speed of the device is lower than usual. Even though this situation is very different from the one in the previous example, the rare event can again be seen as a combination of a number of more or less 'regular' events that occur consecutively.

Precise and efficient techniques to estimate loss (or corruption) probabilities can be considered to be important tools in communication network design. Having such a technique, one can find the optimum trade-off between cost and reliability of the network.

Queueing systems are probably the most natural modelling tool in communications networking. Incorrect operation of highly reliably routers can be considered as a rare event. This combination explains our interest in rare events in queueing networks.

In the remainder of this chapter we describe some queueing models, including the models of our interest, in Section 1.1. The concept of (fast) simulation is explained in Section 1.2. We outline two important simulation techniques, importance sampling and multilevel splitting, in Section 1.3 and Section 1.4 respectively. We end this chapter by outlining the contributions of the thesis, in Section 1.5.

1.1 Some queueing models

Here we describe the queueing models which are extensively studied in this thesis, namely, the tandem Jackson queue and the so-called slowdown network. The slowdown network is our main interest, while the tandem Jackson network is an interesting 'test case' to compare different estimation techniques. We start by describing the $M/M/1$ queue, which is the running example in this chapter to illustrate some of the concepts that are used. General background on queueing theory can be found in textbooks, such as [3, 9, 41].

1.1.1 $M/M/1$ system

Consider a queueing system consisting of an infinitely large waiting room (or buffer) and a single server (station). Jobs arrive one at a time, receive service in order of arrival and leave the system, see Figure 1.1.

Figure 1.1: $M/M/1$

Here we assume that jobs arrive to the system according to a Poisson process with rate λ. The time between two consecutive job arrivals is usually referred to as the interarrival time. In this context the inter-arrival times are i.i.d. The service requirements of the jobs, the so-called service times, are assumed to be independent and exponentially distributed with parameter μ. The service and interarrival times are assumed to be mutually independent.

In order to simplify and uniformize the description of queues (and queueing networks), the so-called Kendall notation, introduced in [40], is often used. The model above is then denoted as $M/M/1$ under the FIFO (*first-in-first-out*) service discipline.

1.1.2 Tandem Jackson network

Jackson networks were introduced by Jackson in [36], and are a frequently studied object. Here we describe a special case of Jackson networks, the two-node Jackson queue.

Thus, we consider a two-node Jackson network with Poisson arrivals at rate λ to the first (or: *upstream*) station. Each job receives service at the first station, after which it is routed to the second (or: *downstream*) station. After receiving service at the second station, the job leaves the system. Service times at station i have exponential distributions with parameter μ_i, $i = 1, 2$. The waiting rooms at both stations are assumed to be infinitely large. Both queues have FIFO as service discipline. Burke's theorem states that the departure process of the first queue, i.e., the arrival process of the second queue, is Poisson with parameter $\min\{\lambda, \mu_1\}$, see [8]. In other words, the tandem Jackson network can be described as two $M/M/1$ queues in tandem, see Figure 1.2. It will turn out that the system's qualitative behavior critically depends on whether the first queue is the bottleneck ($\mu_1 \leq \mu_2$) or if the second queue is the bottleneck ($\mu_2 < \mu_1$).

1.1.3 Slowdown network

Even though Jackson networks can be applied to model various applications, often they do not reflect the real process precisely enough. In order to better model some real-life processes, more complicated queueing models are needed.

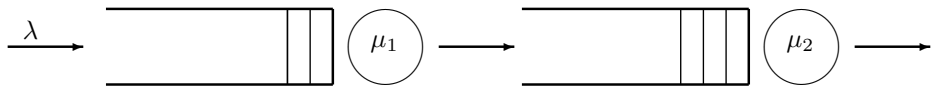

Figure 1.2: Tandem Jackson network

The system with *server slowdown*, also known as a system with *backpressure*, see [68] can be seen as one such important extension of the standard Jackson network. This mechanism is designed to offer the downstream queue some sort of protection against frequent overflows and works as follows: as long as the number of jobs in the downstream queue is smaller than some pre-specified threshold, the server of the upstream queue works in the normal regime, but when the number of jobs in the second queue grows above the threshold, the first server works in a 'slow' regime. It is noted that this property is of significant practical interest, as a related mechanism has been proposed e.g. in the design of Metro Ethernet [47, 58]. It also can be applied in manufacturing.

The slowdown network we consider has Poisson arrivals of rate λ and the same structure as the tandem Jackson network. The service times at the downstream station are exponential 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 the 'slowdown threshold', let us say m, then the parameter of the exponential distribution changes to μ_1^+, where $\mu_1^+ < \mu_1$. When the number of jobs in the second queue drops again below the slowdown threshold, the service rate of the first station returns to its original value μ_1. See also Figure 1.3.

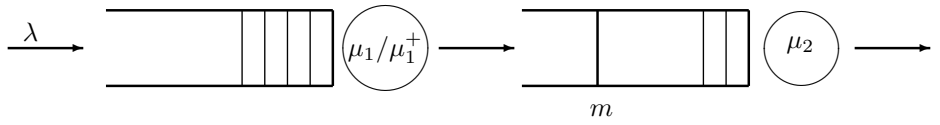

Figure 1.3: Slowdown network

Depending on the parameter values we distinguish three cases. Namely, the cases where the bottleneck is always the first ($\mu_1^+ < \mu_1 \leq \mu_2$) or the second queue ($\mu_2 < \mu_1^+ < \mu_1$), and the so-called 'shifting bottleneck' case ($\mu_1^+ \leq \mu_2 < \mu_1$).

1.1.4 Rare events considered in this thesis

Networks of queues can be applied for modelling in a broad range of situations, such as communication networks, manufacturing, production management, logistics, insurance, etc. See also [42] for survey.

In this thesis we concentrate on a particularly relevant problem – overflow in queueing networks. Let us consider a queueing network of given buffer capacity. When the number of jobs (either in a single node, or in the total system) reaches a certain level, the system cannot perform normally and jobs may be lost. When this threshold is large enough, overflow is rare. It is often important to know in advance how rare it is.

The overflow probabilities we consider can be seen as *first hitting probabilities* in a continuous time Markov chain (CTMC) $Q(t)$. These can be expressed as follows:

$$\mathbb{P}(\text{process } Q(t) \text{ hits } T \text{ before } A, \text{ given } \ Q(0) \notin T \cup A), \tag{1.1}$$

where T and A are target and tabu sets, respectively. In this thesis we will only consider probabilities of the form (1.1), in the context of the networks we described in Section 1.1.2 and in Section 1.1.3. Typically, the process $Q(t)$ will describe the system with infinite buffers, and T denotes the set in which the corresponding finite-buffer system would have overflow.

Example 1.1. Overflow in the single server queue with a finite buffer (let us say of capacity B) can be seen as an elementary rare event. In order to estimate its probability one can consider an $M/M/1$ queue with infinite buffer-size. More formally, let $Q(t)$ be the queue length process. Then define the rare set $T := \{B,\ B{+}1, B{+}2,\ \ldots\}$, where B is the buffer capacity, and the tabu set A only consists of the empty state $\{0\}$. Finally, we let $Q(0) = 1$ and estimate the first hitting probability (1.1). $\diamondsuit$

1.2 Monte Carlo and fast simulations

Now that the type of probability of our interest is defined we consider how to estimate it. Explicit expressions are hard to obtain, asymptotic approximations often lack error bounds. Numerical methods are inefficient on "large" state spaces, indeed, to obtain the probability of interest we need to numerically solve a linear system of equations, where the number of equations and unknowns is the cardinality of the state space, which can be very large or even infinite. Given this, one often relies on simulation methods to obtain performance measures of interest.

The Monte Carlo method is a well-known simulation technique. The idea of this method is to simulate the system multiple times and then estimate the unknown probability γ as the fraction of 'successful samples':

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} I_i, \tag{1.2}$$

where n is the number of replications, and I_i is 1 if the i-th sample was successful and 0 otherwise. Obviously, for estimating rare event probabilities standard Monte Carlo has an inherent problem: it is extremely time consuming to obtain reliable estimates. The relative error (RE) of the Monte Carlo estimator (1.2) is

$$\mathrm{RE}(\hat{\gamma}) = \frac{\sqrt{\mathrm{Var}(\hat{\gamma})}}{\gamma} = \sqrt{\frac{1-\gamma}{n\gamma}} \approx \frac{1}{\sqrt{n\gamma}}. \tag{1.3}$$

Indeed, from (1.3), it is clear that the number of samples needed to obtain an estimate of a certain predefined accuracy is inversely proportional to the probability of interest. This encourages us to use variance reduction techniques such as *Importance Sampling (IS)* or *Multilevel Splitting (MS)*.

In Importance Sampling one simulates the system under a *new* probability measure such that the event of interest occurs more frequently. In our case we change the probabilities such that the process $Q(t)$ will have a drift toward the target set T, away from the tabu set A. The simulation output is then corrected by means of likelihood ratios to retain unbiasedness. The likelihood ratios essentially capture the likelihood of the realization under the old measure with respect to the new measure. The choice of a 'good' new measure is a rather delicate and challenging task.

The other technique, Multilevel Splitting, is conceptually easier, in the sense that one can simulate under the normal probability measure. The main idea of MS is to 'decompose' a rare event in a sequence of 'nested' events that are not so rare. Recall examples in Section 1.1. When the sample path of $Q(t)$ approaches the target set T to a certain distance, it splits into a number of new paths, which are then all simulated independently of each other. This process may repeat itself several times, hence the term *multilevel* splitting.

1.3 Importance sampling

Importance sampling was described in e.g. [7, 34, 60]. In this section we provide a brief description of the main concepts in IS.

1.3.1 Basics of importance sampling

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 which is such that $\mathbb{P}(A_B) \to 0$ as $B \to \infty$. To estimate $\mathbb{P}(A_B)$ via IS simulations one needs to generate samples under a new probability measure $\mathbb{Q}$, with respect to which $\mathbb{P}$ is absolutely continuous. In the case of a CTMC the new measure $\mathbb{Q}$ is just given by replacing the original transition rates q_{ij} from state i to state j by new rates $\tilde{q}_{ij}$. The probability $\mathbb{P}(A_B)$ can now be expressed as

$$\mathbb{P}(A_B) = \mathbb{E}^{\mathbb{Q}}[L(\omega)I(\omega)], \tag{1.4}$$

where $I(\omega)$ is the indicator function which is 1 if $\omega \in A_B$ or 0 else. $L(\omega)$ is the likelihood ratio (also known as Radon-Nikodým derivative) of the realization (path) ω:

$$L(\omega) = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\omega). \tag{1.5}$$

We sample under $\mathbb{Q}$, say n times, to obtain observations $(L_1, I_1), \ldots, (L_n, I_n)$. We can construct an unbiased estimator of $\mathbb{P}(A_B)$ by

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} L_i I_i. \tag{1.6}$$

Example 1.2. We proceed with an example of a simple IS algorithm. Consider an $M/M/1$ queue with arrival rate λ and service rate μ. We assume $\lambda < \mu$, which implies stability of the system, and ensures that having some large number of jobs in the buffer during a busy cycle is a rare event, see also Example 1.1. In [59] it was shown that a good (in fact optimal, in a sense to be discussed later) new measure then corresponds to an unstable system: under the new measure $\mathbb{Q}$ the arrival rate equals μ and the service rate is λ. The likelihood ratio of a sample path ω is the product of likelihood ratios of all jumps, which are λ/μ for job arrivals and μ/λ for job departures. $\diamondsuit$

1.3.2 Design approach

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. A natural and powerful approach to determine this is *large deviations* (LD) theory. LD was described in many textbooks, see e.g., [17, 19, 66]. Here we give a brief review of the theory.

LD enables us to simplify the analysis of stochastic systems significantly. In fact, LD transforms a stochastic problem into a deterministic optimization problem. The outcome of this optimization is a deterministic function $\phi^*(t)$ which is called the optimal (or most probable) path towards the rare event. There are two important side results of this procedure. Firstly, it tells us how the probability of the rare event decays as B grows large: the so-called *logarithmic decay rate* is given by

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(A_B) = -\inf \int \ell(\phi(t), \dot{\phi}(t))dt,$$

where the infimum is taken over all possible paths $\phi(t)$, and $\ell(\phi(t), \dot{\phi}(t))$ is the so-called local rate function. It depends both on the position $\phi(t)$ at time t and on the time derivative (or speed vector) $\dot{\phi}(t)$ at time t. Secondly, LD provides a basis for an efficient IS algorithm, i.e., it tells how the system probabilistically behaves, conditional on the occurrence of the rare event. This knowledge can immediately be translated into a change of measure that we can use when we apply IS.

In case the considered stochastic process is a CTMC the local rate function $\ell(\phi(t), \dot{\phi}(t))$ can be represented using a family of specific cost functions I, which can be interpreted as Kullback-Leibler distances between exponential distributions. More precisely, for exponential random variables with parameters λ and $\tilde{\lambda}$ we define

$$I(\tilde{\lambda} \mid \lambda) := \lambda - \tilde{\lambda} + \tilde{\lambda} \log \frac{\tilde{\lambda}}{\lambda}, \tag{1.7}$$

see also [66, pages 14 and 20]. Note that the function (1.7) is convex and equals 0 at $\tilde{\lambda} = \lambda$. Intuitively, we can think of this value as the cost we need to pay to let a Poisson process $N(t)$ with parameter λ behave like a Poisson process with parameter $\tilde{\lambda}$, per time unit:

$$\mathbb{P}(N(t) \approx \tilde{\lambda} t) \approx e^{-I(\tilde{\lambda}|\lambda)t}. \tag{1.8}$$

As we mentioned before, LD provides us with knowledge of the optimal path and of the new measure. The stochastic process under this measure will follow the optimal trajectory with high probability.

1.3.3 Asymptotic efficiency

Clearly some alternative measures $\mathbb{Q}$ perform better than others, in terms of the variance of the resulting estimator (1.6).

The best solution would be a so-called *zero variance* change of measure, see e.g. [13, 76]. We can define a new measure $\mathbb{Q}^*$ such that $\mathrm{Var}^{\mathbb{Q}^*}[L(\omega)I(\omega)] = 0$ as follows

$$\mathbb{Q}^*(\omega) = \frac{\mathbb{P}(\omega)I(\omega)}{\mathbb{P}(A_B)},$$

for any ω. Obviously, $L(\omega)I(\omega) = \mathbb{P}(A_B)$ with probability one. This means that only one simulation run under the new measure $\mathbb{Q}^*$ is needed to obtain an unbiased estimator with zero variance. However, this idea cannot be used in practice since the value $\mathbb{P}(A_B)$ in the definition of the new measure $\mathbb{Q}^*$ is unknown.

IS schemes with *bounded relative error*, see e.g. [4, 35, 56], can be an achievable alternative. For such schemes, the relative error of the estimator remains bounded when the rarity parameter B grows infinitely large. Formally, a scheme $\mathbb{Q}^*$ has bounded relative error if some constant c exists such that

$$\mathrm{RE}(\hat{\gamma}) = \frac{\sqrt{\mathrm{Var}^{\mathbb{Q}^*}(\hat{\gamma})}}{\mathbb{P}(A_B)} < c,$$

for all B. In practice this property means that for *any* value of the rarity parameter B one needs to perform only a pre-specified number of simulation runs to achieve an estimate with given accuracy.

Unfortunately, it is difficult to construct an IS scheme with bounded relative error even for simple models, see e.g. [11], so we need to introduce another quality concept.

Asymptotic efficiency (or: *asymptotic optimality*) of IS scheme effectively says that the variance of the estimator behaves roughly like the square of its first moment. As a consequence, the computational effort is substantially smaller compared to that of the standard Monte Carlo scheme.

Definition 1.1. *The IS scheme for* $\mathbb{P}(A_B)$ *is called asymptotically efficient if*

$$\liminf_{B\to\infty} \frac{\log \mathbb{E}^{\mathbb{Q}}[L^2(\omega)I(\omega)]}{\log \mathbb{P}(A_B)} \geq 2. \tag{1.9}$$

If $\mathbb{P}(A_B)$ decays exponentially in B, i.e.,

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

asymptotic efficiency means that the number of replications needed to obtain an estimator with fixed relative error grows *subexponentially* fast with the rarity parameter B. In this case the definition of asymptotic efficiency reduces to

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

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

$$\limsup_{B\to\infty} \frac{1}{B} \log \mathbb{E}[L(\omega)I(\omega)] \leq 2 \lim_{B\to\infty} \frac{1}{B} \log \mathbb{P}(A_B). \tag{1.10}$$

Example 1.3. We continue our study of overflows in an $M/M/1$ system during a busy cycle, started in the previous examples. Here we prove that the new measure $\mathbb{Q}$, described in Example 1.2 is asymptotically efficient. The likelihood ratio of a path ω under the new measure $\mathbb{Q}$ with respect to the original measure is $L(\omega) = (\frac{\lambda}{\mu})^a (\frac{\mu}{\lambda})^d$, where a and d are the numbers of arrivals and departures in ω. It is clear that for any successful sample path, i.e. a path that hits level B before returning to the origin we have $a - d = B - 1$. Hence one can conclude that $L(\omega)I(\omega) = (\frac{\lambda}{\mu})^{B-1}$ if ω is a successful path, and 0 otherwise. Combining this with the well-known fact that the logarithmic decay rate of the described probability is $\log(\mu/\lambda)$, one obtains that inequality (1.10) holds and, consequently, the new measure $\mathbb{Q}$ is asymptotically efficient.

In [63] it was shown that a similar type of *change of measure* is optimal for the multi-server GI/GI/m system with light-tailed service times. $\diamondsuit$

Obviously asymptotic efficiency is a desirable characteristic of any IS scheme. For single queues the probabilistic law is usually changed in the same manner for any state in the system. This is a so-called *state-independent* change of measure. These schemes perform well for single queues, but may be inefficient for *networks*, including those considered in Section 1.1. To deal with this, *state-dependent* IS schemes may be needed. In such schemes, the new transition rates are not uniform over the state space. The construction of asymptotically efficient algorithms for simulation of even simple tandem Jackson networks was an open problem for several decades.

1.3.4 Literature on importance sampling for hitting probabilities in networks of queues

We focus here on literature on Jacksonian networks, which are of a Markovian nature (and hence all sojourn times involved are exponentially distributed); in the case of heavy-tailed distributions entirely different techniques are needed, see, e.g., [5, 6].

'Classical' papers on the use of IS in queueing networks usually rely on a state-independent change of measure, i.e., for any state in the system the probabilistic law is changed in the same manner. Usually, large deviations techniques are used to motivate the choice of the new measure, as indicated in Section 1.3.2, and to prove that the resulting estimator has specific desirable properties (such as bounded relative error or asymptotic efficiency). We first mention the seminal paper by Parekh and Walrand [59], that focuses on the estimation of the probability of overflow in a single queue (see Example 1.1–1.3), but also on the probability that the *total* queue population in a network reaches some high value before it returns to the empty state. Things complicate tremendously when looking at networks rather than one-node systems. For the tandem Jackson network, [59] proposed to swap the arrival rate with the rate of the *slowest* server. Heuristically this makes sense, as only the slowest server becomes unstable in this way. However, the experimental results were not so encouraging as in the case of a single queue, and the quality of the simulation results was strongly affected by the specific values of the arrival and service rates. Later it was proved that this method is asymptotically efficient for some parameter values, but has unbounded variance for other values, see [11, 31]. In fact, it was proven that *no* state-independent change of measure exists that is asymptotically efficient for all parameter values. This problem was also studied in [26, 27]; see [64] for the special case of both servers having the same rate.

It was realized that the main problem of state-independent IS schemes is that the transition rates are changed in a uniform manner, in particular irrespective of whether one of the queues is empty or not. As a result it cannot be guaranteed that the likelihood ratio is bounded on the event of interest, and therefore the IS scheme proposed in [59] performs poorly for some parameter values, namely those for which the process under the new measure often visits the boundaries where likelihoods are large. Some of the first attempts to solve this problem can be found in [43, 57, 76, 77] in which state-dependent IS schemes were proposed and in [10, 14], where the cross-entropy approach (see e.g. [61, 62]) was applied. For all these schemes, asymptotic efficiency was empirically concluded, but no analytic proofs were presented. The same holds true for [37], in which the overflow in one individual buffer of a Jackson network was considered.

Dupuis *et al.* [22] were the first to prove asymptotic efficiency for a state-dependent IS scheme for estimating overflow probabilities in a d-node Jackson tandem network. Later in [24], the authors generalized their method to estimate probabilities of first entry to a general rare set in any d-node open Jackson network starting from the empty state. In both works the large deviations decay rate is

found by solving a calculus of variations problem, in which the cost function to be optimized is the relative entropy between the old and the new probability measures. In [23] it was shown that even though this problem converges to the large deviations optimal control problem in some sense, the latter does not always correspond to an asymptotically efficient IS scheme. In other words, understanding the large deviations behavior (the most probable path) is in general not sufficient for an asymptotically efficient IS scheme. To overcome this problem, the large deviations solution is modified in [22, 24], using a suitable game representation, such that the result is a *classical subsolution* of a corresponding *Isaacs* equation. The resulting IS scheme is then proven to be asymptotically efficient by a standard (though cumbersome) verification argument.

For the slowdown model, a first 'pseudo-state-dependent' IS scheme for estimating the probability of overflow in the downstream queue was introduced in [48], see also Chapter 2 of this thesis. Its asymptotic efficiency was concluded for a limited set of initial parameters, but just on the basis of empirical evidence. Later in [21], a provably asymptotically efficient new measure was proposed for the slowdown model in the shifting bottleneck case. Both in [21] and [48] the analysis was restricted to the case that the system starts empty.

1.4 Multilevel splitting

In this section we describe how to estimate the probability of first entry to some rare set, see (1.1), using the MS method.

1.4.1 Basics of multilevel splitting

Consider some Markov process $\{Q_k\}$ with state space D and a finite number of possible jump directions in each state. Although this is not essential we will assume $\{Q_k\}$ to be a random walk for ease of exposition. We are interested in the probability that $\{Q_k\}$ hits the (rare) target set T before the 'tabu' set A, starting from some state $s \notin T \cup A$.

To apply MS, we define a family of nested sets $\{L_k\}$, $k = 0, \ldots, m$ on the state space D such that

$$T = L_m \subset L_{m-1} \subset \ldots \subset L_1 \subset L_0 \subset D.$$

This family $\{L_k\}$ is usually defined in terms of the level sets of a so-called *importance function* and should be chosen such that every state that belongs to the boundary of L_k has similar importance, i.e., the probability of reaching T before A should be approximately equal for every state $x \in \ell_k = \partial L_k$. Given this family, we start R_0 independent simulations of the process, starting at the initial state s. All paths that end up in A are to be terminated, but every path that reaches level $\ell_1 = \partial L_1$ is split

'into' R_1 independent copies. We continue to simulate all the (new) paths until they cross the next level $\ell_2 = \partial L_2$ or hit the tabu set A, and so on. When a sample path reaches the m-th level ℓ_m (i.e., the target set T) it is terminated as well. Now we construct the estimator of (1.1) by dividing the number of samples that eventually reaches the target set T by $R_0 \cdot \ldots \cdot R_{m-1}$, which is the total number of potential sample path. By averaging a number of independent replications of this estimator we obtain an estimate of the probability of interest $\mathbb{P}(A_B)$. The challenge is to choose a family of nested sets $\{L_k\}$, $k = 0, \ldots, m$ with corresponding splitting factors R_k, such that this estimate has bounded relative error or is asymptotically efficient.

Example 1.4. Again we consider the problem of estimating the probability of collecting a large number of jobs B during the busy cycle of a stable $M/M/1$ queue. We choose m such that $B/m \in \mathbb{N}$ and define a family of nested sets

$$L_k := \left\{ \frac{B}{m}k,\ \frac{B}{m}k + 1,\ \ldots \right\},\ k = 0, \ldots, m.$$

The levels are then given by $\ell_k = \{\frac{B}{m}k\}$, $k = 0, \ldots, m$. Since the probability of reaching ℓ_k starting from ℓ_{k-1} is roughly

$$\left(\frac{\lambda}{\mu}\right)^{B/m},$$

where λ and μ are arrival and service rates, respectively, and it is optimal to choose each of the splitting factors equal to the inverse of the probability of hitting the next level, see [29], we obtain in our case

$$R \approx \left(\frac{\mu}{\lambda}\right)^{B/m}.$$

$\diamondsuit$

1.4.2 Restart

Restart (Repetitive Simulation Trials After Reaching Thresholds) is a modification of the classical MS algorithm and was first presented in [73]. The method was enhanced and studied in more detail in [71, 72, 74, 75]. Although we do not study Restart in this thesis, restricting ourselves to the classical MS algorithm, we spend a few lines on this method for the sake of completeness.

The only differences between Restart and classical MS algorithms are in the 'splitting' and 'terminating' rules. Similarly to MS, Restart prescribes to make R_{k-1} independent copies of the path that have reached ℓ_{k-1}. The first $R_{k-1} - 1$ paths are to be split with the factor R_k if they reach the next level before returning to ℓ_{k-1} and terminated otherwise. The last path has a different termination rule. We split it with factor R_k if it hits ℓ_k before ℓ_{k-1}; we terminate it if it hits the tabu set A

before hitting ℓ_{k-1} for the second time; and we *re-split* it with factor R_{k-1} if it hits ℓ_{k-1} for the second time before hitting A or ℓ_k.

Clearly, Restart requires less machine time per cycle than the classical MS method due to the effective truncation. However, the variance of the Restart estimator is larger, see [28].

1.4.3 Asymptotic efficiency

The condition of asymptotic efficiency as introduced in (1.9) is appropriate when the simulation effort for the individual terms in (1.2) and (1.6) is comparable, as is the case for most IS schemes. However, this is not always the case for the MS method. Obviously, an inappropriately chosen splitting factor may lead to a large number of sample paths to be simulated and consequently, the computational time to perform one simulation run can become large.

Therefore, in the setting of MS it is useful to use the concept of *work-normalized* variance, which is the product of the variance and the expected computational effort per simulation run, see e.g. [32]. In other words, the work-normalized variance is equivalent to the variance resulting from a fixed computational budget. Keeping this concept in mind, we will call an estimator asymptotically efficient if

$$\liminf_{B\to\infty} \frac{\log\left(w(B)\mathbb{E}\hat{p}_B^2\right)}{\log \mathbb{E}\hat{p}_B} \geq 2, \tag{1.11}$$

where $w(B)$ represents the expected computational effort per replication of $\hat{p}_B$. We can make various choices for the specific form of $w(B)$, see [29]. Notice that when $w(B)$ is constant, (1.11) coincides with Definition 1.1.

1.4.4 Literature on multilevel splitting

To the best of our knowledge the MS method was first stated in [33], based on the ideas of [38]. The Restart variant was introduced in [73]. We will give a short overview of the most relevant literature here. More extended overviews of the MS method can be found in [28, 65].

There are not many examples of *asymptotically efficient* MS schemes for estimating general types of rare events in the present literature. Most articles deal either with restrictive models or with effective heuristics for particular (queueing) models, usually providing good estimates without rigorous analysis.

In [29], the authors provided conditions for asymptotic efficiency for a certain class of systems with strong assumptions on the state space and transition structure. The main achievement of that paper is that the splitting factor should be inversely proportional to the probability of hitting the next level. In case of a smaller splitting factor the variance of the estimator grows fast, in the opposite case the computational effort is too high. In both cases (1.11) does not hold. This research, with emphasis

on the large deviations properties was continued in [30]. The necessary condition of asymptotic efficiency of MS was formulated there in a general setting.

Sometimes it is proposed to split the procedure into two phases. During the first (or learning) phase the unknown probabilities of hitting a next level are estimated. They are used in the second (or main) phase in order to estimate the probability of interest. In [45] this method was carefully studied and an optimal division of the time budget between two phases was achieved.

Some heuristic methods for constructing importance functions for MS simulations were provided in [70, 75]. There the importance function was chosen to be linear in the content of every individual node of the network. Numerically, this shows a good performance, but no rigorous analysis was provided.

The recent work in [16] *does* enable one to construct an asymptotically efficient MS scheme for estimating the probability of first entrance to a rare set, when the decay rate of the probability is known for all starting states. The authors used control-theoretic techniques, similar to the ones used in [22], to derive and prove their results.

Finally, we like to mention [12] where the IS and MS (Restart) methods are compared, with the conclusion Restart is 'both less promising and less risky' than IS, by which the author means that in contrast to IS, Restart cannot perform with zero or infinite variance, i.e., extremely good and bad, respectively. We will provide our own comparison of IS and MS in the last chapter of this thesis.

1.5 Contribution and outline

This thesis concerns the problem of rare event simulation in two particular queueing networks: the tandem Jackson network and the slowdown network as described in Section 1.1.2 and Section 1.1.3, focusing on first hitting probabilities as introduced in Section 1.1.4. We present several IS schemes in this thesis, beginning with simple, but naïve state-independent algorithms and ending up with a family of simple and efficient state-dependent schemes. The last chapter is dedicated to a MS scheme, and the comparison with the IS schemes.

Being more specific, in Chapter 2 we focus on the overflow probabilities in the downstream queue of both the two-node Jackson network and the slowdown system during a busy cycle (i.e., we start at the empty state), where we apply *state-independent* IS. To find a good change of measure (i.e., to construct a good IS scheme), we first identify the most likely path to overflow. For the standard tandem queue (without slowdown), this path was known, but we develop an appealing novel heuristic, which can also be applied to the model with slowdown or the situation where the starting state is general. The knowledge of the most likely path is then directly used to devise importance sampling algorithms, both for the standard tandem system and for the system with slowdown. Our experiments indicate that the corre-

sponding new measure is sometimes asymptotically optimal, and sometimes not. We systematically analyze the cases that may occur. We also provide a non-trivial stability condition for the slowdown model, which was not known before. This chapter is based on [48] and (for the stability result) on [52].

In Chapter 3 we concentrate on *state-dependent* IS schemes for simulating overflow in the downstream queue of the tandem Jackson network. We identify the optimal overflow paths starting from *any* state. It turns out that the shape of these paths critically depend on the parameter values. When the second buffer is the bottleneck, the path simply grows in the second coordinate (content of the second queue). In the opposite case the behavior is less trivial: the optimal path can then even decrease in the second coordinate for some time. We design an asymptotically efficient IS scheme for any initial state of the system and prove its asymptotic efficiency. In the proof we rely on elementary arguments that are substantially easier than existing (and less general) analogues, see e.g. [22]. Numerical experiments show that the schemes indeed lead to useful results. This chapter is based on [55].

In Chapter 4 we apply the methods used in the previous chapter to construct a family of asymptotically efficient state-dependent IS schemes for the slowdown system for general starting state. The discontinuity of the measure around the slowdown threshold was an additional complication in the analysis of the schemes. However we managed to prove asymptotic efficiency of the scheme similarly as we did in Chapter 3. Chapter 4 is based on [51].

In Chapter 5 we continue our study of the tandem Jackson network. The scheme from Chapter 3 is asymptotically efficient, but it has the drawback that it is difficult to use in practice, as computation of the new measure is time consuming. In Chapter 5 we provide a good compromise: a simplified IS scheme which is asymptotically efficient for all parameter values, giving relative errors that are comparable to those from the 'fully state-dependent' counterpart in Chapter 3 (although slightly larger), and at the same time it is almost as simple to implement and is of a computational complexity that is comparable to the state-independent schemes in Chapter 2. Numerical studies back up our findings. Chapter 5 is based on [54].

In Chapter 6 we present a simple and easy-to-implement IS scheme for the slowdown system; in this sense it can be considered as the counterpart of Chapter 5. Due to the additional complications caused by the discontinuity in statistics around the slowdown threshold (as we had in Chapter 3), we were not able to prove asymptotic efficiency of the proposed scheme for general starting states. However, we did prove it for the major part of the state space. Numerical studies show a high accuracy for *any* parameter value. This chapter is based on [50, 52].

Finally, Chapter 7 is dedicated to the MS method. Here we present a family of MS schemes for rare event simulation for a class of models, which includes both networks of our interest. Namely, we design a simple and asymptotically efficient MS scheme for estimating first hitting probabilities as in (1.1). We assume that the logarithmic decay rate of the probability of interest is known, and use this as a basis for choosing the splitting levels. The proof of asymptotic efficiency relies on some elementary

combinatorics and a number of simple facts from the theory of branching processes. When we apply the proposed MS scheme to the tandem Jackson network or the slowdown network, our numerical findings show high accuracy and time efficiency. We also compare the efficiency of the developed IS and MS schemes. This chapter is based on [53].

Chapter 2

State-independent importance sampling

We begin this chapter with a more detailed description of the models of our interest (the tandem Jackson network and the slowdown network). We also specify the scaling concept, which allows us to use the same state space for any value of the rarity parameter B.

The first contribution of this chapter is to present the most probable path to overflow in the second queue for the various cases. In order to find them, we assign costs to scaled paths in terms of some large deviations based cost functions. The 'most likely path' is then the 'cheapest' path from the origin to the 'overflow set' while the origin is not visited. The intuition behind this is that, conditional on the event that the scaled process indeed reaches the rare set before the system becomes empty, the trajectory of the Markov process will be typically close to this most likely path. A rigorous proof of this is deferred to Chapters 3 and 4 (for tandem and slowdown network, respectively).

The cost-minimization procedure gives us also a change of measure that can be used for a state-independent IS scheme. We study this scheme with its benefits and drawbacks, using both analytic and numerical approaches. On the one hand this IS scheme is very easy to implement. On the other hand the quality of the estimator (in terms of the relative error) can be very poor, especially when the values of arrival and service rates are close to each other, which is often the case in practice. A better understanding of the performance of the scheme eventually helps us to construct asymptotically efficient IS schemes in the next chapters of this thesis.

2.1 Model descriptions

2.1.1 Tandem Jackson network

We consider a two-node tandem Jackson network with Poisson arrivals at rate λ to the first station. Each job receives service at the first station, after which it is routed to the second station. After receiving service at the second station, the job leaves the system. Service times at station i have an exponential distribution with parameter μ_i, $i = 1, 2$. The waiting rooms at both stations are assumed to be infinitely large, see also Figure 1.2.

Let $Q(t) = \{(Q_1(t), Q_2(t)), t \geq 0\}$ be the joint queue-length process. Then it is clear that this is a continuous-time Markov process, with possible jump directions $v_0 = (1, 0)$, $v_1 = (-1, 1)$ and $v_2 = (0, -1)$ with corresponding transition rates λ, μ_1 and μ_2, respectively. The process $Q(t)$ is regenerative if we impose the stability condition $\lambda < \min(\mu_1, \mu_2)$, which we will do from now on.

The queue-length process can also be described by the *embedded* discrete-time Markov chain $Q_j = (Q_{1,j}, Q_{2,j})$, where $Q_{i,j}$ is the number of jobs in queue i after the j-th transition. Without loss of generality we will choose the parameters such that $\lambda + \mu_1 + \mu_2 = 1$, so that they also represent the *transition probabilities* of Q_j in the interior of the state space. To ensure that the same holds on the boundaries, we shall introduce so-called self-transitions shortly, see below.

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

We introduce the following subsets of the state space

$$\begin{aligned} D &:= \{(x_1, x_2) : x_1 > 0, 0 < x_2 < 1\}, \\ \partial_1 &:= \{(0, x_2) : 0 < x_2 < 1\}, \\ \partial_2 &:= \{(x_1, 0) : x_1 > 0\}, \\ \partial_e &:= \{(x_1, 1) : x_1 > 0\}, \end{aligned} \tag{2.1}$$

and denote the state space by $\bar{D} = D \cup \partial_e \cup \partial_1 \cup \partial_2$ (realize that we can exclude $x_2 > 1$ from the state space).

Note that transition v_k is impossible when queue k is empty, i.e., when $X_j \in \partial_k$. In order to deal with this discontinuity we can modify process X_k in two different ways: to set $\mathbb{P}(v_k | X_j \in \partial_k) = 0$ and re-normalize jump transitions or to allow self-transitions.

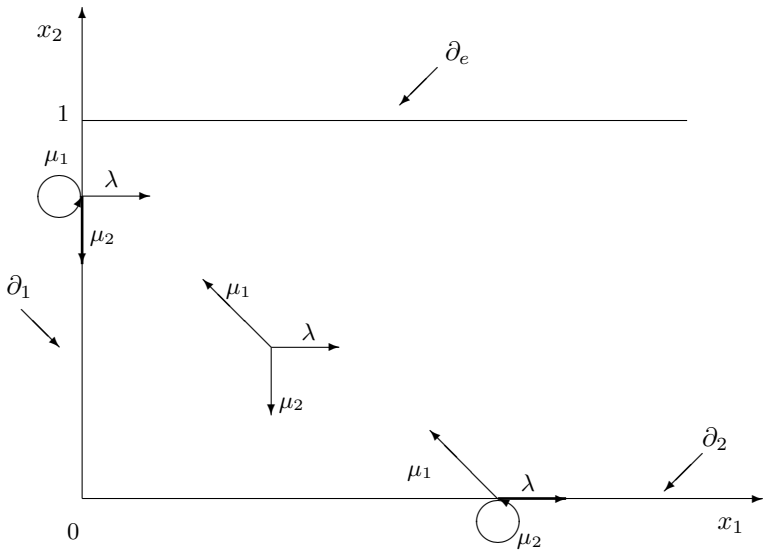

Figure 2.1: State space and transition rates for scaled process X_j.

Formally, the first rule can be stated as follows

$$\mathbb{P}(B\Delta X_j = v_k | X_{j-1} \in \partial_k) = 0, \quad k = 1, 2 \quad \text{and}$$

$$\mathbb{P}(B\Delta X_j = v_0 | X_{j-1} \in \partial_1) = \tfrac{\lambda}{\lambda+\mu_2}, \quad \mathbb{P}(B\Delta X_j = v_2 | X_j \in \partial_1) = \tfrac{\mu_2}{\lambda+\mu_2}, \tag{2.2}$$

$$\mathbb{P}(B\Delta X_j = v_0 | X_{j-1} \in \partial_2) = \tfrac{\lambda}{\lambda+\mu_1}, \quad \mathbb{P}(B\Delta X_j = v_1 | X_j \in \partial_2) = \tfrac{\mu_1}{\lambda+\mu_1},$$

where $\Delta X_j = X_j - X_{j-1}$. A second modification rule is formalized in the following way

$$\mathbb{P}(B\Delta X_j = v_k | X_{j-1} \in \partial_k) = 0 \text{ and } \mathbb{P}(X_j = X_{j-1} | X_{j-1} \in \partial_k) = \mu_k, \tag{2.3}$$

for $k = 1, 2$, while having everything else the same. On Figure 2.1 we modified queue-length process X_j according to the second rule, i.e. (2.3). Obviously, choice of the modification of X_k does not affect the probability of our interest. We will use the first type of the modification in the current chapter, since it is easer to illuminate some properties of the likelihood ratios in this way. We will use the second method in the rest of the thesis, since it simplifies the proofs. Next, we introduce the stopping time τ_B^s, which is the first time that the process X_j hits level 1, starting from state $s = (s_1, s_2)$, without visits to the origin:

$$\tau_B^s = \inf\{k > 0 : X_k \in \partial_e, X_j \neq 0 \text{ for } j = 1, \ldots, k-1\}, \tag{2.4}$$

and we define $\tau_B^s = \infty$ if X_j hits the origin before ∂_e. It will also be convenient to let $I_B(\omega^s)$ be the indicator of the event $\tau_B^s < \infty$ for the path

$$\omega^s = (X_j, j = 0, \dots : X_0 = s).$$

Thus we can write the probability of our interest as

$$p_B^s = \mathbb{E} I_B(\omega^s) = \mathbb{P}(\tau_B^s < \infty). \tag{2.5}$$

It is important to recall that in this chapter we will only consider the probability of 'overflow' in the second buffer during a busy cycle of the system. In other words we impose a restrictive condition on the starting state: $s = (1/B, 0)$ and will omit the index s from p_B^s and τ_B^s.

Due to the stability condition the overflow event becomes rare as B grows large, and hence p_B will become small. The following theorem specifies how this happens in the tandem Jackson network.

Theorem 2.1. *For the tandem Jackson network, the overflow probability p_B is asymptotically geometric in B with parameter ρ_2. More precisely,*

$$\lim_{B\to\infty} \frac{1}{B} \log p_B = \log \rho_2, \tag{2.6}$$

where $\rho_2 = \lambda/\mu_2$ is the utilization of the second queue.

Proof. The proof uses arguments that are somewhat similar to those used in the proof of Theorem 1 in [1]. At first let us recall that the limiting distribution of the process is well-known and given by

$$\pi\left(\frac{i}{B}, \frac{j}{B}\right) = \lim_{t\to\infty} \mathbb{P}\left(X(t) = \left(\frac{i}{B}, \frac{j}{B}\right)\right) = (1-\rho_1)(1-\rho_2)\rho_1^i \rho_2^j, \tag{2.7}$$

where $\rho_i = \lambda/\mu_i$. Denoting the indicator function of an event A by $\mathbf{1}(A)$, we can define the time spent by the process at level 1 during a busy cycle T, as

$$I_{(\cdot,1)} = \int_0^T \mathbf{1}(X_2(t) = 1) dt, \tag{2.8}$$

A simple renewal argument tells us that $\pi(0,0) = \lambda^{-1}/\mathbb{E}T$, and also that

$$\pi(\cdot, 1) = \frac{\mathbb{E} I_{(\cdot,1)}}{\mathbb{E}T} = p_B \mathbb{E}(I_{(\cdot,1)} \mid I_{(\cdot,1)} > 0) \lambda \pi(0,0).$$

Thus, using (2.7), we can write

$$p_B = \frac{1}{\lambda(1-\rho_1)} \frac{{\rho_2}^B}{\mathbb{E}(I_{(\cdot,1)} \mid I_{(\cdot,1)} > 0)}. \tag{2.9}$$

Also using (2.7), we can show that

$$\mathbb{E}(I_{(i/B,1)} \mid I_{(\cdot,1)} > 0) = \rho_1^i \, \mathbb{E}(I_{(0,1)} \mid I_{(\cdot,1)} > 0),$$

where $I_{(0,1)}$ is defined similarly as $I_{(\cdot,1)}$ in (2.8). This leads us to

$$\mathbb{E}(I_{(\cdot,1)} \mid I_{(\cdot,1)} > 0) = \frac{\mathbb{E}(I_{(0,1)} \mid I_{(\cdot,1)} > 0)}{1-\rho_1}. \tag{2.10}$$

Now we condition on the number of jobs in the first queue when the second queue reaches level B (i.e., $I_B(\omega^s) = 1$), and note that for all $i > 0$,

$$(I_{(0,1)} \mid X_1(\tau_B) = i) \stackrel{d}{=} \begin{cases} 0, \text{ if } (0,1) \text{ is not visited during the cycle,} \\ (I_{(0,1)} \mid X_1(\tau_B) = 0), \text{ if } (0,1) \text{ is visited during the cycle,} \end{cases}$$

due to the Markov property. Therefore we can write

$$\begin{aligned}
&\mathbb{E}(I_{(0,1)} \mid I_{(\cdot,1)} > 0) \\
&= \sum_{i=0}^{\infty} \mathbb{E}(I_{(0,1)} \mid X_1(\tau_B) = \frac{i}{B}, I_{(\cdot,1)} > 0) \; \mathbb{P}(X_1(\tau_B) = \frac{i}{B} \mid I_{(\cdot,1)} > 0) \\
&\leq \sum_{i=0}^{\infty} \mathbb{E}(I_{(0,1)} \mid X_1(\tau_B) = 0, I_{(\cdot,1)} > 0) \; \mathbb{P}(X_1(\tau_B) = \frac{i}{B} \mid I_{(\cdot,1)} > 0) \\
&= \mathbb{E}(I_{(0,1)} \mid X_1(\tau_B) = 0, I_{(\cdot,1)} > 0) \; = \; \frac{1}{(\lambda+\mu_2) q_B},
\end{aligned} \tag{2.11}$$

where $q_B = \mathbb{P}((0,0) \text{ reached before } (0,1) | X(0) = (0,1))$. It is clear that the sequence q_B is strictly decreasing in B and that $\lim_{B\to\infty} q_B = q_\infty \in (0,1)$ exists, which implies that for any positive B,

$$q_B > q_\infty. \tag{2.12}$$

Combining (2.9), (2.10), (2.11) and (2.12), we derive a lower bound on p_B and using the simple fact that $\mathbb{E}(I(\cdot,1) \mid I(\cdot,1) > 0) \geq \frac{1}{\mu_2}$ in (2.9), we also find an upper bound. This leads us to

$$\frac{\lambda+\mu_2}{\lambda} q_\infty {\rho_2}^B \leq p_B \leq \frac{\rho_2^{B-1}}{1-\rho_1},$$

from which we have

$$\log \rho_2 \leq \lim_{B\to\infty} \frac{1}{B} \log p_B \leq \log \rho_2.$$

□

Theorem 2.1 is important in itself, as it gives us already a rough estimate for the probability of interest (2.5) for large B. In fact it says that p_B is of the form $f(B)\rho_2^B$ where $\log f(B)/B \to 0$ as B grows large. To obtain p_B more precisely, we will use estimates based on simulations. Secondly, the theorem is important as it will help us to verify the asymptotic optimality of the estimators involved in these simulations. In the next chapters we extend Theorem 2.1 to general starting states (see Theorem 3.7) and for the slowdown system (see Theorem 4.1).

2.1.2 Slowdown network

In this network jobs enter the system at the upstream queue (as a Poisson process with rate λ), and after being served they are forwarded to the downstream queue; after service in this second queue, they leave the network. Service times at the second station are exponential with parameter μ_2 all the time, but the service speed at the first queue depends on the content of the second queue. Normally, service times at the first station are exponential with parameter μ_1, but if the number of jobs in the second queue is larger than some prespecified threshold m – the so-called *slowdown threshold* – then the service times are exponential with parameter μ_1^+, where $\mu_1^+ < \mu_1$. When the system 'stabilizes' and the number of jobs in the second queue is again strictly below the slowdown threshold, the rate of the first station returns to its original value μ_1, see also Figure 1.3.

For convenience we choose the parameters such that $\lambda + \mu_1 + \mu_2 = 1$, without loss of generality and hence $\lambda + \mu_1^+ + \mu_2 < 1$. Here we assume the waiting rooms at both stations to be infinitely large. We define the continuous and discrete-time joint queue-length process $Q(t) = \{(Q_1(t), Q_2(t)), t \geq 0\}$ and $Q_j = (Q_{1,j}, Q_{2,j})$ as in Section 2.1.1. We also define the possible jump directions of the process Q_j in the same way: $v_0 = (1, 0)$, $v_1 = (-1, 1)$ and $v_2 = (0, -1)$ with corresponding jump rates λ, μ_1 (or μ_1^+) and μ_2 (where it is noted that v_k is impossible if queue k is empty – then take $v_k = (0, 0)$ instead).

A first question is under what condition this process is stable – clearly for design purposes such a criterion is of crucial importance. Interestingly, the resulting criterion is substantially more involved than the usual '$\rho < 1$ conditions', see also [49]. Define $\psi := \mu_1/\mu_2$ and $\psi^+ := \mu_1^+/\mu_2$.

Theorem 2.2. *Case I: $\mu_1^+ < \mu_2$. The network is stable if*

$$\lambda < \frac{\mu_1(1-\psi^m)(1-\psi^+) + \mu_1^+\psi^m(1-\psi)}{(1-\psi^m)(1-\psi^+) + \psi^m(1-\psi)}.$$

Case II: $\mu_1^+ \geq \mu_2$. The network is stable if $\lambda < \mu_2$.

Proof. It is obvious that $\lambda < \mu_2$ is necessary, but not sufficient for stability. We deal with both cases separately.

Case I: $\mu_1^+ < \mu_2$. The proof relies on techniques from the theory of Quasi-Birth-and-Death (QBD) processes [46]; we aim to prove the positive recurrence of the discrete-time process Q_j. From now on we will treat Q_j as a discrete-time QBD with $Q_{1,j}$ and $Q_{2,j}$ being the level and the phase, respectively; note that this is not a 'standard QBD', as the number of phases is infinite.

We now introduce some QBD related notation. Let M_0, M_1 and M_2 be $(n+1) \times (n+1)$ dimensional matrices, with n being the number of phases (either finite or infinite). M_0 represents an increase in level (new job arrives to the system), M_1 no change in level (job leaves the system) and M_2 an decrease in level (job is forwarded

from the first queue to the second one); the precise definitions of these matrices were given in [68] for our model. [46, Thm. 7.2.3] now states that if the number of phases is *finite*, then the QBD is positive recurrent if $\boldsymbol{\alpha} M_0 \boldsymbol{e} < \boldsymbol{\alpha} M_2 \boldsymbol{e}$, where the vector $\boldsymbol{\alpha}$ is the solution to $\boldsymbol{\alpha} M = 0$, with $M := M_0 + M_1 + M_2$; $\boldsymbol{e}$ is an all-1 vector.

Application of this result (which is not legitimate in our case, due to the infinite number of phases!) would yield that the QBD is positive recurrent if

$$\sum_{i=0}^{m-1} \alpha_0 \psi^i (\lambda - \mu_1) + \sum_{i=m}^{\infty} \alpha_0 \psi^m (\psi^+)^{i-m} (\lambda - \mu_1^+) < 0, \tag{2.13}$$

where

$$\alpha_i = \begin{cases} \alpha_0 \psi^i, & i < m \\ \alpha_0 \psi^m (\psi^+)^{i-m}, & i \geq m \end{cases};$$

$$\alpha_0 = \left(\sum_{i=0}^{m-1} \psi^i + \psi^m \sum_{i=m}^{\infty} (\psi^+)^{i-m} \right)^{-1},$$

from which the first statement of the theorem would follow. There is a counterpart of [46, Thm. 7.2.3] that *does* deal with an infinite number of phases, though: [67, Thm. 5] states that the QBD with infinite number of phases is positive recurrent if $\boldsymbol{\alpha} M_0 \boldsymbol{e} < \boldsymbol{\alpha} M_2 \boldsymbol{e}$ provided that $\bar{M} = M_1 + M_2$. Here $\bar{M}$ is an infinite-dimensional matrix that describes the behavior of the phase-process of the QBD at level 0, see again [68] for its precise form. The condition $\bar{M} = M_1 + M_2$ effectively means that the phase process in level 0 is the same as for any other level. Obviously, this requirement fails in our case. In order to be able to apply [67, Thm. 5] we modify the QBD Q_j in order to satisfy the condition condition $\bar{M} = M_1 + M_2$:

$$v_1 = (0, 1), \text{ when } Q_{1,j} = 0,$$

where the other transition vectors remain unchanged. Now we can conclude that this modified process is stable if (2.13) holds. An elementary inspection yields that the cycle time (i.e., the number of transitions it takes the discrete-time process to reach the origin) of the modified QBD is *stochastically larger* than the cycle time of the original QBD, and hence stability of the original QBD is implied by the stability of the modified QBD.

Case II: $\mu_2 \leq \mu_1^+$. Here we cannot apply the reasoning mentioned above since the distribution α_i does not exist when $\mu_2 \leq \mu_1^+$. However, in this case, stability can be proven rather straightforwardly. Clearly, the expected cycle length in this case can be bounded from above by the expected cycle length for $m = 0$ (due to elementary coupling arguments). The latter value is finite, since it corresponds to the mean busy cycle length of the tandem Jackson network with parameters $(\lambda, \mu_1^+, \mu_2)$, which is stable under $\lambda < \mu_2 \leq \mu_1^+$. □

We remark that somewhat related results for the slowdown network with a *finite* second buffer were reported in [68, Thm. 15]. Also, interestingly, the slowdown

system can be stable *even when* $\lambda > \mu_1^+$*!* The intuition behind this is as follows. Consider the case when both $\lambda > \mu_1^+$ and the condition in Theorem 2.2 hold true. The content of the first queue typically increases when the number of jobs in the second queue is above m. However, it stays finite because the content of the second queue tends to decrease and the system returns to its normal state in which the number of jobs in the first queue tends to decrease. It is also worthwhile to mention that when the slow down threshold m is 0 or ∞, the criterion mentioned above reduces to the standard stability condition for the ordinary tandem Jackson network.

We focus on estimating the probability of reaching some high level B in the second queue before it returns to the origin, starting from any given state. Here we describe a number of notions that are useful with this goal in mind. From now on we let the threshold m scale linearly with B that is, $m \equiv \theta B$ for some $\theta \in (0,1)$. In terms of the scaled processes $X(t) = Q(Bt)/B$ (in continuous time) and $X_j = Q_j/B$ (in discrete time), we analyze the probability that its second coordinate attains the value 1 before reaching the origin. Note that an advantage of this scaling is again that we may use the state space $[0,\infty) \times [0,1]$ (for any value of B). We introduce the following subsets of the state space, with $x := (x_1, x_2)$:

$$\begin{array}{ll} D := \{x : x_1 > 0, 0 < x_2 < \theta\}, & \partial_2 := \{(x_1,0) : x_1 > 0\}, \\ D^+ := \{x : x_1 > 0, \theta \le x_2 < 1\}, & \partial_\theta := \{(x_1,\theta) : x_1 \ge 0\}, \\ \partial_1^+ := \{(0,x_2) : x_2 \in [\theta,1)\}, & \partial_e := \{(x_1,1) : x_1 \ge 0\}, \\ & \partial_1 := \{(0,x_2) : x_2 > 0\}. \end{array} \tag{2.14}$$

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 D in (2.1) and (2.14) refers to different sets.

Recall 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 as we did for tandem Jackson network, see (2.2) and (2.3). The probability of our interest is as follows

$$p_B^s = \mathbb{E} I_B(\omega^s) = \mathbb{P}(\tau_B^s < \infty),$$

see also (2.5); here τ_B^s is the first time that the process X_j hits level 1, starting from state $s = (s_1, s_2)$, without visits to the origin with $X_0 = s$, see also (2.4). Again, we will skip superindex s in this chapter.

2.2 Optimal path structure

In order to find a good change of measure for IS simulations, the first step is usually to find the 'optimal path to overflow', i.e., the way in which overflow most probably occurs, conditional on its occurrence. The optimal path is often used to state a 'law of large numbers for rare events', in the spirit of identifying a path $(x_1^*(t), x_2^*(t))$ such that

$$\mathbb{P}\left(\|(X_1(\cdot), X_2(\cdot)) - (x_1^*(\cdot), x_2^*(\cdot))\| > \varepsilon \mid X_2(\tau_B) = 1 \right) \to 0$$

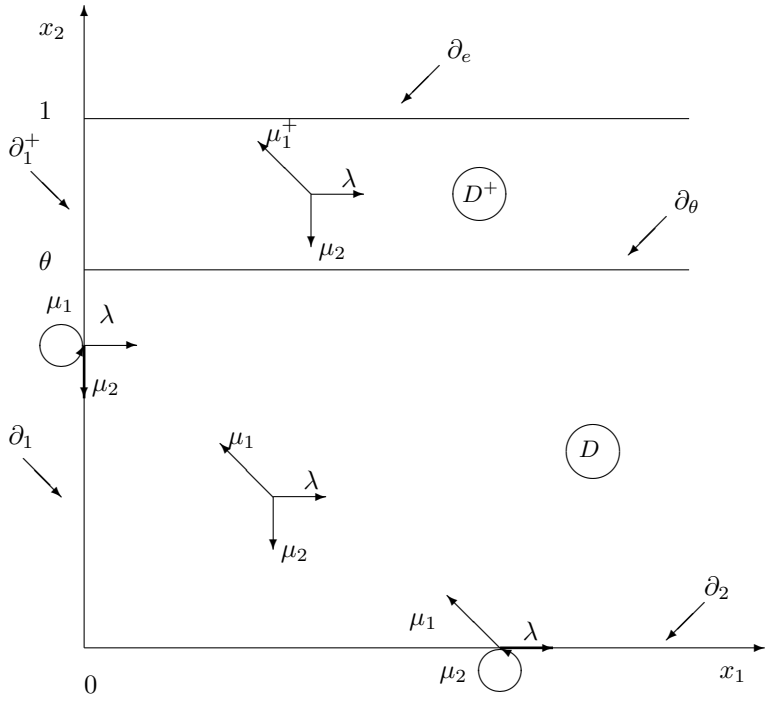

Figure 2.2: State space and transition rates for the scaled process X_j.

as $B \to \infty$, for all $\varepsilon > 0$, where $||\cdot||$ is some metric.

Such a path has already been identified for general Jackson networks in [2] (and hence also for our tandem system). In that paper, the time-reversed process is used to find the shape of the most probable path to overflow. In fact it is shown that this path can have two different forms, depending on the relation between μ_1 and μ_2. If the second server is the bottleneck ($\mu_2 < \mu_1$) the optimal path to overflow has a very simple shape: the second buffer fills up gradually, while the first queue remains virtually empty. On the other hand, when the first queue is the bottleneck we have a more complicated situation, in which the path consists of two parts. During the first part the second queue stays virtually empty while the number of jobs in the first buffer grows up to same value that is proportional to B. During the second part, the number of jobs in the first buffer decreases (virtually to 0) while the second buffer fills up to B.

In the remainder of this section we present another method (i.e., different from [2]), to find the optimal path. This method is heuristic by nature, but has important advantages. First, it not only yields the shape of the optimal path, but also gives a 'good' change of measure, which will ensure that most simulation runs under this new measure will be close to the optimal path. Secondly, we note that in the slowdown network, which is our ultimate interest in this thesis, we cannot use the method from

[2], since we do not know the explicit form of the stationary distribution in that case, and therefore we cannot use an analysis based on the time-reversed process. Our heuristic method, however, *can* be applied here.

The method is based on the use of a certain family of cost functions I

$$I(\tilde{\lambda} \mid \lambda) = \lambda - \tilde{\lambda} + \tilde{\lambda} \log\left(\frac{\tilde{\lambda}}{\lambda}\right), \tag{2.15}$$

see also (1.7). The idea behind this heuristic is the following. With $N(t)$ the number of arrivals generated by a Poisson process of rate λ, we may be interested in $\mathbb{P}(N(t) \approx \tilde{\lambda}t)$. Assume for ease that t is integer; then $N(t)$ is distributed as the sum of t i.i.d. Poisson random variables, each with mean λ. Cramér's theorem then says: for $\tilde{\lambda} > \lambda$

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(N(t) \geq \tilde{\lambda}t) = -\sup_{\theta} \Big(\theta\tilde{\lambda} - \log \mathbb{E}\exp(\theta N(1))\Big),$$

and for $\tilde{\lambda} < \lambda$

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(N(t) \leq \tilde{\lambda}t) = -\sup_{\theta} \Big(\theta\tilde{\lambda} - \log \mathbb{E}\exp(\theta N(1))\Big),$$

which can be interpreted as

$$\mathbb{P}(N(t) \approx \tilde{\lambda}t) \approx \exp\left(-t \sup_{\theta} \Big(\theta\tilde{\lambda} - \log \mathbb{E}\exp(\theta N(1))\Big)\right), \tag{2.16}$$

and it is easy to verify that the right hand side of the latter expression reduces to $e^{-I(\tilde{\lambda}|\lambda)t}$, see also (1.8). For instance, consider for any α a straight path from $(0,0)$ to $(\alpha,1)$ through the interior of the state space, staying away from the boundaries. We then need to replace the parameters by 'tilded' parameters $\tilde{\lambda}$, $\tilde{\mu}_1$ and $\tilde{\mu}_2$, such that $\tilde{\mu}_1 > \tilde{\mu}_2$ and $\tilde{\lambda} \geq \tilde{\mu}_1$, in order to have north-east drift. The total cost of such a path, per unit length in vertical direction is

$$\frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\mu}_1 - \tilde{\mu}_2}, \tag{2.17}$$

where

$$\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2) = I(\tilde{\lambda} \mid \lambda) + I(\tilde{\mu}_1 \mid \mu_1) + I(\tilde{\mu}_2 \mid \mu_2) \tag{2.18}$$

represents the total cost per unit time, and the denominator in (2.17) is the average speed by which the process moves up. If we would replace the denominator by $\tilde{\lambda} - \tilde{\mu}_1$, we would find the cost per unit length in horizontal direction. Finally we mention that the (negative) slope of this path is given by

$$\alpha = \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\lambda} - \tilde{\mu}_1}. \tag{2.19}$$

Minimizing (2.17) over the three tilded parameters, such that also $\tilde{\mu}_1 > \tilde{\mu}_2$ and $\tilde{\lambda} \geq \tilde{\mu}_1$ hold will then give the optimal values for the tilded parameters and the slope of the path for this particular shape. Equations (1.8) and (2.16) indicate that the exponent of the negative of (2.17) can be used as an approximation of the probability of interest, but conditional on a path of the given type (with $\tilde{\lambda} \geq \tilde{\mu}_1$). By considering all possible path types we will obtain an approximation of the probability of overflow in the second buffer during a busy cycle, as well as the most probable path (i.e. the minimizing values of $\tilde{\lambda}, \tilde{\mu}_1$ and $\tilde{\mu}_2$). The same ideas can be applied to the slowdown network.

Note that we can also associate cost to a non-straight path $\phi(t) = (\phi_1(t), \phi_2(t))$ with $t \in [0, T]$ for some T, namely as $\int_0^T \ell(\phi(t), \dot{\phi}(t))dt$, where ℓ is the so-called local rate function, see (5.5) in [66]. If $\dot{\phi}(t)$ is locally equal to $(\tilde{\lambda} - \tilde{\mu}_1, \tilde{\mu}_1 - \tilde{\mu}_2)$ with $\tilde{\lambda}\tilde{\mu}_1\tilde{\mu}_2 = \lambda\mu_1\mu_2$, the relation with our cost functions is given by $\ell(\phi(t), \dot{\phi}(t)) = \mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$, which can be seen by noting that the solution to (5.2) in [66] is given by $\theta_1 = \log(\tilde{\lambda}/\lambda)$ and $\theta_2 = -\log(\tilde{\mu}_2/\mu_2)$. We will come back to these issues in Chapter 3 and Chapter 4.

To proceed, let us formulate the following theorems, upon which our studies in this chapter will be based. Consider the slowdown system, with $\theta \in [0, 1]$ which includes the tandem case when $\theta = 0$ or 1. The state space consist of five subsets on which the transition parameters are constant, viz. the sets $\{(0, 0)\}$, ∂_1, ∂_2, D and D^+. See Figure 2.2.

Theorem 2.3. *The path with minimal cost in terms of cost-function (2.15), is the most probable path between any two points.*

We postpone the proof of Theorem 2.3 to the following chapters, since it is involved and lengthy. In Chapter 3 we prove this theorem for the tandem Jackson network (i.e., when $\theta = 0$ or 1), see Theorem 3.7. In Chapter 4 we present the proof for the slowdown network, see Theorem 4.1.

Theorem 2.4. *Consider the slowdown system, with $\theta \in [0, 1]$ then the typical path which starts in the empty state and leads to overflow in a single node or in the total queue consists of a concatenation of subpaths on the various subsets that are straight lines; each subset is traversed at most once.*

The great benefit of Theorem 2.4 is that the solution now boils down to optimizing over a finite number of possible path-types, i.e., we reduced the problem to a combinatorial problem. The proof of the theorem is based on the two following lemmas.

Lemma 2.5. *The optimal path between two states in the same subset is a straight line.*

Proof. This follows from Lemma 5.16 of [66], but to provide some more insight we present an alternative proof for the following special case. Let us consider two states

in the interior $x = (x_1, x_2)$ and $y = (y_1, y_2)$, both below or both above θ. We need to show that the optimal path from x to y is a straight line. To this end we consider a path from x to y via an additional point $z = (z_1, z_2)$ and find the values of z_1 and z_2 which minimize the total cost of such a path. First consider states x, y and z such that $x_1 \le z_1 \le y_1$ and $x_2 \le z_2 \le y_2$. The optimal cost per unit length in vertical direction of such a path is:

$$\inf \left\{ (z_2 - x_2) \frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\mu}_1 - \tilde{\mu}_2} + (y_2 - z_2) \frac{\mathbb{I}(\bar{\lambda},\, \bar{\mu}_1,\, \bar{\mu}_2)}{\bar{\mu}_1 - \bar{\mu}_2} \right\}. \tag{2.20}$$

The infimum is taken over variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$, $\bar{\mu}_2$, z_1 and z_2 that satisfy $\tilde{\lambda} \ge \tilde{\mu}_1, \tilde{\mu}_1 \ge \tilde{\mu}_2, \bar{\lambda} \ge \bar{\mu}_1, \bar{\mu}_1 \ge \bar{\mu}_2$. Note that the knowledge of starting and ending points for each subpath gives us a possibility to express $\tilde{\lambda}$ and $\bar{\lambda}$ from (2.20) in terms of the other variables, i.e.

$$\tilde{\lambda} = \frac{(z_1 - x_1)(\tilde{\mu}_1 - \tilde{\mu}_2) + (z_2 - x_2)\tilde{\mu}_1}{z_2 - x_2} \quad \text{and} \quad \bar{\lambda} = \frac{(y_1 - z_1)(\bar{\mu}_1 - \bar{\mu}_2) + (y_2 - z_2)\bar{\mu}_1}{y_2 - z_2}.$$

As a result we obtain the following:

$$z_1 = \frac{x_1 - y_1}{x_2 - y_2} z_2 + \frac{x_2 y_1 - x_1 y_2}{x_2 - y_2} \tag{2.21}$$

for any z_2. Equality (2.21) guarantees that z lies on the line which connects x and y. The same statement can be proved for arbitrary choices of x, y and z in an equal manner. This completes the proof. □

Lemma 2.6. *The optimal path does not have more than one subpath in each subset.*

Proof. At first we will focus on a path that has two subpaths in the interior. It is clear that the path could not be optimal if it includes two consecutive subpaths in the same subset, by Lemma 2.5, so we concentrate on paths that have two subpaths in the interior, with a connecting subpath on one of the boundaries in between, see Figure 2.3. We will show that it is not optimal to have two subpaths in the interior below θ, using the path from Figure 2.3 as a typical example. The minimal cost of the first four subpaths is:

$$\inf \left\{ \theta \frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_2} + \theta \frac{\mathbb{I}(\bar{\lambda},\, \bar{\mu}_1,\, \bar{\mu}_2)}{\bar{\mu}_2 - \bar{\mu}_1} + \beta \frac{\mathbb{I}(\hat{\lambda},\, \hat{\mu}_1,\, \hat{\mu}_2)}{\hat{\lambda} - \hat{\mu}_1} + \theta \frac{\mathbb{I}(\breve{\lambda},\, \breve{\mu}_1,\, \breve{\mu}_2)}{\breve{\mu}_1 - \breve{\mu}_2} \right\}, \tag{2.22}$$

where

$$\beta = \theta \left(\frac{\bar{\lambda} - \bar{\mu}_1}{\bar{\mu}_2 - \bar{\mu}_1} + \frac{\breve{\lambda} - \breve{\mu}_1}{\breve{\mu}_1 - \breve{\mu}_2} \right) > 0,$$

and the infimum is taken over all $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$, $\bar{\mu}_2$, $\hat{\lambda}$, $\hat{\mu}_1$, $\hat{\mu}_2$, $\breve{\lambda}$, $\breve{\mu}_1$ and $\breve{\mu}_2$, such that $\tilde{\lambda} > \tilde{\mu}_2$, $\tilde{\lambda} < \tilde{\mu}_1$, $\bar{\mu}_1 < \bar{\mu}_2$, $\hat{\lambda} > \hat{\mu}_1$, $\hat{\mu}_1 < \hat{\mu}_2$ and $\breve{\mu}_1 > \breve{\mu}_2$. We split

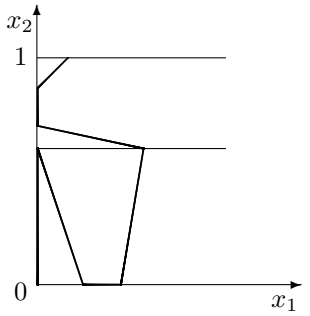

Figure 2.3: Example of a path with two vertical and two horizontal subpaths

the problem into two cases. When $\mu_1 > \mu_2$ we obtain the following result after optimization: $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_2, \mu_1, \lambda)$, $(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) = (\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2) = (\mu_1, \lambda, \mu_2)$ and $(\breve{\lambda}, \breve{\mu}_1, \breve{\mu}_2) = (\mu_1, \mu_2, \lambda)$. Since the second and third parts have the same cost of moving in horizontal direction, it is obvious that a path that does not contain the first part (following the vertical boundary) will always have lower cost than the path in Figure 2.3. When $\mu_1 < \mu_2$ it can be shown in a similar way that the cheapest path from $(0, 0)$ to $(x, 0)$ is again a straight line which follows the horizontal boundary, see also Section 2.3, Case 1. This means that the cost of the path in Figure 2.3 is always bounded from below by the cost of a path which satisfies the lemma.

Note that a path which consists of two subpath in the interior and a connecting subpath on the vertical axis also cannot be optimal. Using the same arguments one can prove that the optimal path does not not contain two (or more) subpaths in the interior above θ. This completes the proof. □

2.3 Tandem Jackson Network

2.3.1 Optimal path to overflow

For the tandem Jackson network, we consider the minimal costs of all possible path types that satisfy Theorem 2.4 with $\theta = 1$. As a result, we obtain the most probable path to overflow as the path with globally minimal cost, and the associated change of measure. The cost function itself will yield $-\log \rho_2$ as its optimum value, since we already know that ρ_2 is the geometric decay rate of the probability of interest for the tandem model, see Theorem 2.1.

We split the problem for the tandem Jackson network into two cases: **(1)** $\lambda < \mu_1 < \mu_2$, i.e. the first server is the bottleneck; and **(2)** $\lambda < \mu_2 < \mu_1$, i.e. the second server is the bottleneck. In the end of this subsection we will focus on the special case in which $\lambda < \mu_1 = \mu_2$ and argue that Case 1 can be extended to $\lambda < \mu_1 \leq \mu_2$.

Case 2, i.e. $\lambda < \mu_2 < \mu_1$

We prefer to start our analysis with Case 2, since this is the simplest problem. We consider a path that follows the vertical axis. To find the optimal tilded parameters for such a path we need to solve the minimization problem

$$I_2 = \inf \left\{ \frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_2} \right\}, \tag{2.23}$$

where the infimum is taken over all tilded variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, such that $\tilde{\mu}_1 > \tilde{\lambda}$ and $\tilde{\lambda} > \tilde{\mu}_2$, ensuring a north-west drift (i.e. to the left and up). Note that the denominator is again the average speed at which the process moves up; it is $\tilde{\lambda} - \tilde{\mu}_2$ instead of $\tilde{\mu}_1 - \tilde{\mu}_2$ since the first queue is stable ($\tilde{\lambda} < \tilde{\mu}_1$), so the second queue fills up at rate $\tilde{\lambda}$ rather than $\tilde{\mu}_1$. After taking partial derivatives with respect to all tilded variables and setting them equal to zero, some algebra leads us to the solutions $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\lambda, \mu_1, \mu_2)$ and $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_2, \mu_1, \lambda)$. However, only the second solution satisfies both boundary conditions $\tilde{\mu}_1 > \tilde{\lambda}$ and $\tilde{\lambda} > \tilde{\mu}_2$, so the minimal cost of this type of path is $I_2 = -\log(\rho_2)$ per unit vertical length.

We checked all other possible shapes of the path to overflow (for a detailed account, we refer to the Appendix) and conclude that for this case I_2 is in fact the minimal cost per unit movement in the vertical direction, and indeed ρ_2 is the decay rate.

Proposition 2.7. *If $\lambda < \mu_2 < \mu_1$ (Case 2) then the optimal path to overflow of the second buffer has the following shape: $(0,0) \to (0,1)$ and the decay rate is ρ_2. The corresponding change of measure is given by*

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_2, \mu_1, \lambda). \tag{2.24}$$

We note that the notation $(x_1, x_2) \to (y_1, y_2)$ stands for a straight line from state x to state y (for the scaled process); in this case the path follows the vertical boundary due to the north-west drift under the change of measure.

Now let us see what happens if the first queue is the bottleneck.

Case 1, i.e. $\lambda < \mu_1 < \mu_2$

We present the minimization problem for the path to overflow as described by [2]. Thus, assume we have tilded parameters that satisfy $\tilde{\mu}_1 < \tilde{\mu}_2$ and $\tilde{\mu}_1 < \tilde{\lambda}$, to ensure a path along the horizontal axis, south-east drift. For the second part of the path we have parameters $\bar{\lambda}$, $\bar{\mu}_1$ and $\bar{\mu}_2$ such that $\bar{\mu}_1 > \bar{\mu}_2$ and $\bar{\lambda} \leq \bar{\mu}_1$. The minimization problem is then given by

$$I_1 = \inf \left\{ -\alpha^{-1} \frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_1} + \frac{\mathbb{I}(\bar{\lambda},\, \bar{\mu}_1,\, \bar{\mu}_2)}{\bar{\mu}_1 - \bar{\mu}_2} \right\},$$

where the infimum is taken over variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$ and $\bar{\mu}_2$, that satisfy the given boundary conditions, and α is the (negative) slope of the second part of the path, i.e., $\alpha = (\bar{\mu}_1 - \bar{\mu}_2)/(\bar{\lambda} - \bar{\mu}_1)$, cf. (2.19). The solution for this problem can be found in two steps, first minimizing the first term over the tilded variables, which yields $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_1, \lambda, \mu_2)$, and then solving the remaining problem, yielding $(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) = (\mu_1, \mu_2, \lambda)$. The total path of this shape will cost us $I_1 = -\log(\rho_2)$ per vertical unit. Paths with other shapes have been checked as well, and indeed none of them has lower cost.

Proposition 2.8. *If $\lambda < \mu_1 < \mu_2$ (Case 1), then the optimal path to overflow of the second buffer has the following shape: $(0,0) \to (-\alpha^{-1}, 0) \to (0,1)$, where $\alpha = (\bar{\mu}_1 - \bar{\mu}_2)/(\bar{\lambda} - \bar{\mu}_1)$, and the decay rate is ρ_2. The corresponding change of measure is given by*

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_1, \lambda, \mu_2) \quad \textit{until } X_1(t) = -\alpha^{-1}, \tag{2.25}$$

$$(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) = (\mu_1, \mu_2, \lambda) \quad \textit{afterwards.} \tag{2.26}$$

Case with equal service rates, i.e. $\lambda < \mu_1 = \mu_2$

The special case where both service rates are equal can in principle be added to Case 1 or to Case 2. The constant $-\alpha^{-1}$ in formula (2.25) now equals zero and consecutive utilization of measures (2.25) and (2.26) is in fact equivalent to using the new measure in (2.24). Also both methods seem to provide the same path $(0,0) \to (0,1)$, so it might appear at first sight that it does not matter which case we extend. However, in the case of equal service rates, the optimal path is a vertical line upwards *in the interior, without horizontal drift*, which is different from a path on the vertical axis as in Case 2. This is why we prefer Case 1 over Case 2 to include the possibility that $\mu_1 = \mu_2$.

2.3.2 Importance sampling

We dedicate this section to the problems of IS simulation of the tandem Jackson network. We start our study from the case when the second buffer is the bottleneck, since this is the simplest case.

Case 2, i.e. $\lambda < \mu_2 < \mu_1$

Again we start our analysis with the simplest problem: the tandem Jackson network where the second node is the bottleneck. The optimal path to overflow is known, and under the new measure we will simply interchange λ and μ_2 as follows from Proposition 2.7.

To construct the probability estimator of a path, we need to know the likelihood of a sample path, which is just the product of the likelihoods of all individual transitions made until the path hits either the exit boundary ∂_e or state $(0,0)$, whichever

happens first. As an example let us introduce the likelihood ratio for a transition corresponding to an arrival in the first buffer (i.e., a jump to the right). It is important to note that the likelihood ratios in the interior and on the boundaries may be different. Let us first provide the likelihood ratio for a 'horizontal' jump from some state in the interior. This is given by the ratio of probabilities to make such a jump under the old and new measures, i.e. the ratio of $\lambda/(\lambda+\mu_1+\mu_2)$ and $\tilde{\lambda}/(\tilde{\lambda}+\tilde{\mu}_1+\tilde{\mu}_2)$, which gives $L = \lambda/\mu_2$. On the vertical boundary the ratio turns out to be the same, but on the horizontal boundary the likelihood ratio is different

$$L' = \frac{\frac{\lambda}{\lambda+\mu_1}}{\frac{\tilde{\lambda}}{\tilde{\lambda}+\tilde{\mu}_1}} = \frac{\lambda}{\mu_2}\frac{\mu_1+\mu_2}{\lambda+\mu_1}.$$

Similarly we can calculate the likelihood ratios for other types of jumps. Taking these into account we can find the likelihood ratio of an entire path to overflow as

$$L_2(\omega) = \left(\frac{\lambda}{\mu_2}\right)^{B-1+R}\left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{H}, \tag{2.27}$$

where R is the number of jobs in the first buffer when the second reaches level B for the first time, and H is the total number of visits to the horizontal axis under the new measure, both belonging to path ω.

Now let us see when the IS scheme (2.24) is asymptotically efficient. Definition 1.1, Theorem 2.1 and equality (2.27) together give that we need (note that R can be safely ignored, since $\lambda < \mu_2$),

$$\mathbb{E}^{\mathbb{Q}}\left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{2H} = \sum_{i=1}^{\infty}\left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{2i}\mathbb{Q}(H=i) < \infty.$$

If H is asymptotically geometric, i.e., if for some constants c and γ we have $\mathbb{Q}(H = i) \approx c\gamma^i$ as $i \to \infty$, then this holds when γ satisfies

$$\gamma\left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{2} < 1. \tag{2.28}$$

Although we have no formal proof, our simulation results confirm that the number of visits to the horizontal axis during a busy cycle indeed has an almost geometrical distribution. In Figure 2.4 we present a contour plot of the left-hand side of (2.28) as a function of μ_1 and μ_2; note that $\lambda = 1-\mu_1-\mu_2$ so that the domain is given by the triangular region $0 < 1-\mu_1-\mu_2 < \mu_2 < \mu_1$. The figure illustrates in which parameter region (2.28) holds, so that the estimator is asymptotically efficient. Note however that we cannot be sure that it is not asymptotically efficient in the remaining part of the domain.

Another way to assess asymptotic efficiency is to directly evaluate (1.9), which we also did empirically:

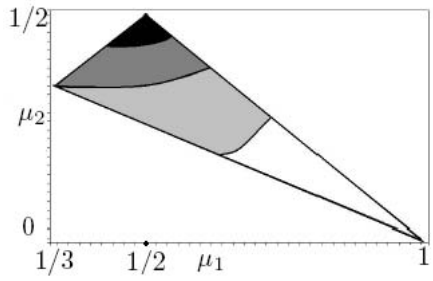

Figure 2.4: Contour plot of the left-hand side of (2.28) for the tandem Jackson network, Case 2, under the new measure (2.24). Values less than 0.5 are in white, less than 1 in light gray, less than 1.5 in dark gray and greater than 1.5 black.

$$\hat{\psi}_B = \frac{\log \frac{1}{N} \sum_{i=1}^{N} L^2(\omega_i^s) I_B(\omega_i^s)}{\log \frac{1}{N} \sum_{i=1}^{N} L(\omega_i^s) I_B(\omega_i^s)}. \tag{2.29}$$

In our case the denominator of (2.29) can be exchanged by $\log \rho_2$ thanks to Theorem 2.1. The curves in Figure 2.5 represent for two different parameter settings the estimate of (1.9) for $B = 50$ as N, the number of replications, increases until 10^6. The values of the (λ, μ_1, μ_2) are respectively $(0.1, 0.7, 0.2)$ (top curve) and $(0.3, 0.36, 0.34)$ (bottom curve). This is empirical evidence that for the first parameter setting we have an asymptotically efficient estimator, while for the second setting we do not.

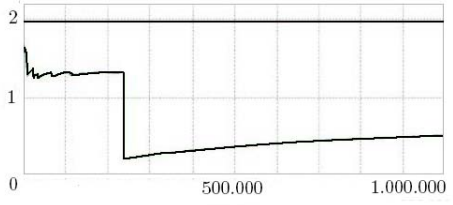

Figure 2.5: $\hat{\psi}_{50}$ against N. Case 2.

Finally, for the same two parameter settings but various values of B we present in Table 2.1 some estimates for the overflow probabilities with 95% confidence intervals, and estimates for the left hand side of (1.9). Simulations for this table (and upcoming tables) are based on $N = 10^6$ independent replications of the busy cycle.

Using the IS method we can decrease simulation time considerably. Under asymptotic efficiency the total time effort grows subexponentially in B. For $B = 20$ it takes 9 seconds to do the $N = 10^6$ replications to estimate the overflow probability with confidence interval of width $4.3 \cdot 10^{-9}$ for the first parameter setting in Table 2.1.

Compare this to straightforward simulations where, for a larger confidence interval of width $4 \cdot 10^{-8}$ we need $N \gg 10^6$, taking more than 2 hours. We do not have such a situation in the second column. For $B = 20$ it takes 37 seconds to obtain the estimate for the overflow probability and confidence intervals using IS, and 40 for similar result using direct simulations (again these values correspond to the first parameter settings). In this case IS simulations yield somewhat smaller simulation times compared to direct simulations, but the speedup is incomparably smaller than in the case of an asymptotically efficient change of measure.

$(\lambda, \mu_1, \mu_2) = (0.1, 0.7, 0.2)$			$(\lambda, \mu_1, \mu_2) = (0.3, 0.36, 0.34)$		
B	$\hat{\psi}_B$	p_B	B	$\hat{\psi}_B$	p_B
20	1.93	$1.11 \cdot 10^{-6} \pm 2.15 \cdot 10^{-9}$	20	0.67	$6.00 \cdot 10^{-2} \pm 6.25 \cdot 10^{-4}$
50	1.97	$1.03 \cdot 10^{-15} \pm 2.00 \cdot 10^{-18}$	50	1.30	$1.50 \cdot 10^{-3} \pm 6.35 \cdot 10^{-5}$
100	1.99	$9.21 \cdot 10^{-31} \pm 1.78 \cdot 10^{-33}$	100	1.60	$2.91 \cdot 10^{-6} \pm 6.95 \cdot 10^{-8}$

Table 2.1: Simulation results for the tandem Jackson network, Case 2

Remark 2.9. When we compare our region of asymptotic efficiency with that in [11], they seem to coincide, although [11] considers the probability that the *total* network population, i.e. $Q_1(t) + Q_2(t)$, reaches some high level B. However, since the optimal paths for both problems coincide for the current Case 2, the similarity need not surprise us.

Case 1, i.e. $\lambda < \mu_1 < \mu_2$

Now let us focus on the case where the first queue is the bottleneck of the system. In Proposition 2.8 we showed that a good change of measure for this problem is given by (2.25)–(2.26).

The likelihood ratio of an arbitrary path to overflow now has a more complicated structure than in Case 2:

$$L_1(\omega) = \left(\frac{\lambda}{\mu_2}\right)^{B-1-U} \left(\frac{\lambda}{\mu_1}\right)^{R-1} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right)^{V_1} \left(\frac{\mu_1+\lambda}{\mu_2+\lambda}\right)^{V_2} \left(\frac{\mu_1+\mu_2}{\mu_1+\lambda}\right)^{H_2}, \quad (2.30)$$

where V_1 is the number of visits to the vertical axis under measure (2.25); V_2 and H_2 are the numbers of visits to vertical and horizontal axes under measure (2.26) respectively; U is the value of the second component of the scaled process $X(t)$ when its first component reaches level α^{-1} for the first time; and R is the first component of $X(t)$ when the process hits the exit boundary ∂_e for the first time. We assume that V_1, V_2 and H_2 are (roughly) geometrical random variables with parameters γ_1, γ_2 and γ_3 respectively, which is indeed confirmed by simulation experiments. Also assuming independence as B grows large, the inequality that should hold for

asymptotic efficiency is now given by

$$\gamma_1 \left(\frac{\mu_1 + \mu_2}{\lambda + \mu_2} \right)^2 \gamma_2 \left(\frac{\mu_1 + \lambda}{\mu_2 + \lambda} \right)^2 \gamma_3 \left(\frac{\mu_1 + \mu_2}{\mu_1 + \lambda} \right)^2 < 1. \tag{2.31}$$

Unfortunately, simulations show that (2.31) never holds under the change of measure (2.25)–(2.26). On the other hand, Figure 2.6 suggests that in Case 1 we may have asymptotical efficiency for some parameters. The variance of the estimator strongly depends on the parameter settings. From top to bottom we have $(\lambda, \mu_1, \mu_2) = (0.13, 0.17, 0.7)$, $(0.25, 0.35, 0.4)$ and $(0.3, 0.33, 0.37)$.

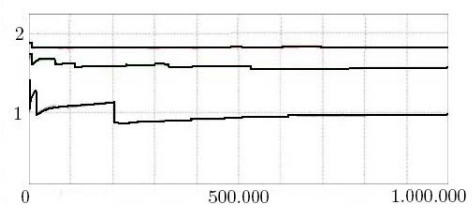

Figure 2.6: $\hat{\psi}_{50}$ against N. Case 1.

For two of these parameter settings and various values of B we also present in Table 2.2 some simulation results. It is clear that IS gives a considerable variance reduction and speedup compared to normal simulation, also when the estimator is (arguably) not asymptotically efficient.

$(\lambda, \mu_1, \mu_2) = (0.13, 0.17, 0.7)$			$(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$		
B	$\hat{\psi}_B$	p_B	B	$\hat{\psi}_B$	p_B
20	1.58	$7.50 \cdot 10^{-15} \pm 1.20 \cdot 10^{-15}$	20	0.30	$2.60 \cdot 10^{-2} \pm 2.39 \cdot 10^{-3}$
50	1.88	$5.64 \cdot 10^{-37} \pm 1.21 \cdot 10^{-37}$	50	1.09	$3.81 \cdot 10^{-5} \pm 2.80 \cdot 10^{-5}$
100	1.93	$1.73 \cdot 10^{-73} \pm 1.73 \cdot 10^{-74}$	100	1.34	$8.68 \cdot 10^{-10} \pm 4.05 \cdot 10^{-10}$

Table 2.2: Simulation results for the tandem Jackson network, Case 1

Remark 2.10. It is possible to consider various changes of measure that will result in the same optimal path. For instance, instead of switching from measure (2.25) to (2.26) once, we can also switch back and forth between these measures, depending on the current value of $X_2(t)$. Analysis of (2.30) shows that in particular visits to the horizontal axis during the second part of the cycle are harmful (i.e., they may result in a large value of the likelihood). We tried to exclude these by using the following more complicated change of measure. We start with measure (2.25); we always switch to measure (2.26) if $X_1(t) = -\alpha^{-1}$ and $X_2(t) > 0$; we always switch to the original measure ($\tilde{\lambda} = \lambda$, $\tilde{\mu}_1 = \mu_1$, $\tilde{\mu}_2 = \mu_2$) if $X_1(t) > -\alpha^{-1}$ and $X_2(t) = 0$;

we switch back to measure (2.25) if $X_1(t) \leq -\alpha^{-1}$ and $X_2(t) = 0$. Empirically it turns out that this change of measure (and other variants) is also not asymptotically efficient, although the variance of the estimator is a little less.

2.4 Slowdown network

2.4.1 Optimal path to overflow

In this section we will focus on the slowdown system in which the rate of the first server depends on the content of the second buffer. As in the tandem model we can identify different cases, depending on the values of the parameters, but now we distinguish three cases: **(3)** $\mu_2 < \mu_1^+ < \mu_1$, **(4)** $\mu_1^+ < \mu_2 < \mu_1$ and **(5)** $\mu_1^+ < \mu_1 < \mu_2$. The cases in which $\mu_1 = \mu_2$ or $\mu_1^+ = \mu_2$ can be dealt with in the same manner as for the tandem Jackson network. Cases 3 and 4 are comparable to Case 2 in the tandem model, where the second server is the bottleneck. The difference is in the situation when the second component of the scaled process $X(t)$ exceeds the slowdown threshold θ: in Case 3 the second server remains the bottleneck, i.e $\mu_1^+ > \mu_2$, while in Case 4 the first server becomes the bottleneck, i.e., $\mu_1^+ < \mu_2$. When the first server is the bottleneck there is only one possibility, in which the first server remains the bottleneck.

As mentioned earlier, we cannot use a reversibility argument as in [2] in the analysis of the slowdown system. However, we can employ our cost function approach, based on Theorem 2.4.

Case 3, i.e. $\mu_2 < \mu_1^+ < \mu_1$

Let us start from the situation in which the second server stays the bottleneck, analyzing the path that follows the vertical axis as in Case 2. This path now consists of two parts: below the slowdown threshold and above it. Using the same arguments as in (2.23) we need to find

$$I_3 = \inf \left\{ \theta \frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_2} + (1-\theta) \frac{\mathbb{I}(\bar{\lambda},\, \bar{\mu}_1^+,\, \bar{\mu}_2)}{\bar{\lambda} - \bar{\mu}_2} \right\}, \tag{2.32}$$

where the infimum is taken over variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$ and $\bar{\mu}_2$, such that we have north-west drift below the threshold (i.e., $\tilde{\mu}_1 > \tilde{\lambda}$ and $\tilde{\lambda} > \tilde{\mu}_2$) and above it (i.e., $\bar{\mu}_1^+ > \bar{\lambda}$ and $\bar{\lambda} > \bar{\mu}_2$). This can easily be solved by splitting it into two separate minimization problems that are completely analogous to (2.23), so the outcome will be to interchange the values of λ and μ_2. We have checked all other possible shapes of the path to overflow (see the second part of the Appendix) and conclude that indeed $I_3 = -\log \rho_2$ is the minimal cost for unit movement in the vertical direction.

Proposition 2.11. *If $\mu_2 < \mu_1^+ < \mu_1$ (Case 3) then the optimal path to overflow of the second buffer has the following shape: $(0,0) \to (0,\theta) \to (0,1)$. The corresponding change of measure is given by*

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_2, \mu_1, \lambda) \quad \textit{and} \quad (\bar{\lambda}, \bar{\mu}_1^+, \bar{\mu}_2) = (\mu_2, \mu_1^+, \lambda), \tag{2.33}$$

and the decay rate is ρ_2.

Case 4, i.e. $\mu_1^+ < \mu_2 < \mu_1$

Now let us concentrate on the network where the bottleneck shifts from the second server to the first server when the slowdown threshold is reached. We focus on a path that follows the vertical axis until the slowdown threshold, after which the process moves with north-east drift. The following minimization problem corresponds to this type of path:

$$I_4 = \inf \left\{ \theta \frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_2} + (1-\theta) \frac{\mathbb{I}(\bar{\lambda},\, \bar{\mu}_1^+,\, \bar{\mu}_2)}{\bar{\mu}_1^+ - \bar{\mu}_2} \right\}, \tag{2.34}$$

where we take the infimum over variables $\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2, \bar{\lambda}, \bar{\mu}_1$ and $\bar{\mu}_2$, such that $\tilde{\mu}_1 > \tilde{\lambda}$ and $\tilde{\lambda} > \tilde{\mu}_2$ and $\bar{\mu}_1^+ > \bar{\mu}_2$ and $\bar{\lambda} \geq \bar{\mu}_1^+$. Again we can decompose the optimization problem into two parts. The first part of (2.34) has the same solution as the first part of (2.32), and hence as (2.23). The second part of problem (2.34) has a more complicated solution, that in fact corresponds to the boundary case in which the path has no horizontal drift, i.e. $\bar{\lambda} = \bar{\mu}_1^+$. It is given by

$$\bar{\lambda} = \bar{\mu}_1^+ = \sqrt{\frac{\lambda \mu_1^+}{z^+}},\ \bar{\mu}_2 = \mu_2 z^+, \tag{2.35}$$

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

$$\lambda + \mu_1^+ + \mu_2(1 - z^+) = 2\sqrt{\frac{\lambda \mu_1^+}{z^+}}. \tag{2.36}$$

As an aside we note that this is the same equation as (30) in [44], and indeed the decay rate behavior as found in that paper can also be obtained using our heuristic. Since all other path types turn out to have higher cost (see the second part of the Appendix), this is the optimal path, and the corresponding cost is $I_4 = -\log\left(\rho_2^\theta (z^+)^{1-\theta}\right)$ per vertical unit.

Proposition 2.12. *If $\mu_1^+ < \mu_2 < \mu_1$ (Case 4), then the optimal path to overflow of the second buffer has the following shape: $(0,0) \to (0,\theta) \to (0^+,1)$. The corresponding change of measure is given by (2.24) and (2.35), and the decay rate is $\rho_2^\theta (z^+)^{1-\theta}$.*

The optimal path in this case looks very similar to the optimal path in Case 3. Indeed they coincide below θ, where the drift is to the north and east. Above θ the path is also vertical, but there is an essential difference, since there is *no* horizontal drift here. The notation 0^+ in Proposition 2.12 is meant to express this difference.

Case 5, i.e. $\mu_1^+ < \mu_1 < \mu_2$

This case has the least interest from a practical point of view, but we include it for the sake of completeness. In this section we will provide the shape of the most likely path to overflow in the second buffer.

Proposition 2.13. *If $\mu_1^+ < \mu_1 < \mu_2$ (Case 5), then the optimal path to overflow of the second buffer has the following shape: $(0,0) \to (\beta_1, 0) \to (\beta_2, \theta) \to (0,1)$, where $\beta_2 = (1-\theta)(\hat{\mu}_1^+ - \hat{\lambda})/(\hat{\mu}_1^+ - \hat{\mu}_2)$ and $\beta_1 = \beta_2 + \theta(\bar{\mu}_1 - \bar{\lambda})/(\bar{\mu}_1 - \bar{\mu}_2)$. The corresponding change of measure is given by*

$$\begin{aligned}
(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2) &= (\mu_1,\, \lambda,\, \mu_2) \text{ until } X_1 = \beta_1, \\
(\bar{\lambda},\, \bar{\mu}_1,\, \bar{\mu}_2) &= (\mu_1,\, \mu_2,\, \lambda) \text{ until } X_2 = \theta, \\
(\hat{\lambda},\, \hat{\mu}_1^+,\, \hat{\mu}_2) &= (\mu_1,\, \frac{\lambda\mu_1^+}{q\mu_1},\, q\mu_2) \text{ afterwards},
\end{aligned}$$

where q is the unique solution of

$$\mu_1\mu_2 q^2 + \mu_1(\mu_1 - \lambda - \mu_1^+ - \mu_2)q + \lambda\mu_1^+ = 0 \tag{2.37}$$

which guarantees $\hat{\lambda} < \hat{\mu}_1^+$ and $\hat{\mu}_1^+ > \hat{\mu}_2$, and the constants β_1 and β_2 are given by $\beta_2 = (1-\theta)(\lambda\mu_1^+ - \mu_1^2 q)/(\lambda\mu_1^+ - \mu_1\mu_2 q^2)$ and $\beta_1 = \beta_2 + \theta(\mu_2 - \mu_1)/(\mu_2 - \lambda)$. The decay rate is $\rho_2^{\theta} q^{1-\theta}$.

We omit all calculations due to similarity to Case 1.

2.4.2 Importance sampling

In this section we present our results for the IS simulations for the system with slowdown threshold. Let us first focus on the case where the second buffer remains the bottleneck, for which we have a much stronger result.

Case 3, i.e. $\mu_2 < \mu_1^+ < \mu_1$

In this case we use the change of measure given in (2.33). The rate of the first server is the same as under the original measure, i.e., it is equal to either μ_1 or μ_1^+ depending on the current state of the second buffer.

Proposition 2.14. *When $\lambda < \mu_2 < \mu_1^+ < \mu_1$ (Case 3), asymptotic efficiency does not depend on the value of μ_1^+. In other words, the overflow probability estimators under the measure (2.24) for the tandem Jackson network and (2.33) for the slowdown network are asymptotically efficient in the same parameter regions.*

Proof. The likelihood ratio of an arbitrary path ω that reaches the exit boundary ∂_e before the origin is very similar to (2.27), namely

$$L_3(\omega) = \left(\frac{\lambda}{\mu_2}\right)^{B-1+R'} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^{H'},$$

where R' is the number of jobs in the first buffer when the second one reaches level B for the first time and H' is the total number of visits to the horizontal axis under the new measure. It is enough to show that the second moments of L_2 and L_3 are asymptotically identical to prove the proposition. It is clear that the distribution of R' is not important since $\lambda/\mu_2 < 1$. The distribution of H' on the other hand does play a role, and in fact determines whether or not the estimator is asymptotically efficient for a certain parameter setting. Fortunately we have that H' converges in distribution to H as $B \to \infty$, so that comparison with (2.27) gives the statement of the proposition. □

As an illustration we simulate the system for two different parameter settings. In the first we take $(\lambda, \mu_1, \mu_2) = (0.1, 0.7, 0.2)$ and $\mu_1^+ = 0.3$, and for the second we take $(\lambda, \mu_1, \mu_2) = (0.3, 0.36, 0.34)$ and $\mu_1^+ = 0.35$. In these (and all further) simulations we will use $\theta = 0.8$ to define the slowdown threshold. Note the correspondence to the examples in Section 2.3, and that in both cases we have $\mu_1^+ > \mu_2$. Indeed in the first case the estimator is asymptotically efficient, and in the second case it is not, which can be illustrated by a similar picture as Figure 2.5, and also by the values of (2.29). The speedups obtained in this table are comparable to those in Table 2.1.

$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.1, 0.7, 0.3, 0.2)$			$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.35, 0.34)$		
B	$\hat{\psi}_B$	p_B	B	$\hat{\psi}_B$	p_B
20	1.95	$7.94 \cdot 10^{-7} \pm 1.89 \cdot 10^{-9}$	20	0.70	$5.8 \cdot 10^{-2} \pm 4.91 \cdot 10^{-4}$
50	1.98	$6.50 \cdot 10^{-16} \pm 1.68 \cdot 10^{-18}$	50	1.37	$1.46 \cdot 10^{-3} \pm 3.97 \cdot 10^{-5}$
100	1.99	$5.59 \cdot 10^{-31} \pm 1.48 \cdot 10^{-33}$	100	1.66	$2.64 \cdot 10^{-6} \pm 9.51 \cdot 10^{-8}$

Table 2.3: Simulation results for the slowdown system, Case 3

Case 4, i.e. $\mu_1^+ < \mu_2 < \mu_1$

In this case we use the change of measure given in Proposition 2.12. Specifically, we always use (2.24) when $X(t)$ is below θ and (2.35) else, until ∂_e is reached for the

first time. It is more difficult now to obtain an analogue of Proposition 2.14, since the process may have some cycles around level θ that influence the total likelihood of the path. Notice that this was not true in Case 3, due to the fact that the change of measure used below and above the threshold was essentially the same. Consider then a typical path to overflow of the form

$$(1,0) \to (X_1(t_1),\theta) \to (X_1(t_2),\theta) \to ... \to (X_1(t_n),\theta) \to (R,1),$$

where the $t_i, i = 1,2,\ldots,n$ are the $n \geq 1$ subsequent time epochs at which the process visits level θ before ∂_e. It can be shown that the total likelihood of such a path is given by

$$L_4(\omega) = \left(\frac{\lambda}{\mu_2}\right)^{\theta B} z^{+(1-\theta)B} \left(\frac{\mu_1+\mu_2}{\lambda+\mu_1}\right)^H \left(\frac{\sqrt{\frac{\lambda\mu_1^+}{z^+}}+\mu_2 z^+}{\lambda+\mu_1}\right)^V \left(\frac{\lambda}{\mu_2 z^+}\right)^C \times \left(\frac{\lambda}{\mu_2}\right)^D \left(\sqrt{\frac{\lambda z^+}{\mu_1^+}}\right)^U.$$

Here H is the number of visits to the horizontal axis and V is the number of visits to the vertical axis above the threshold. Furthermore C is the number of subpaths $(X_1(t_i),\theta) \to (X_1(t_{i+1}),\theta)$ below θ (starting with a downward jump), D is the 'total horizontal distance covered during these subpaths', i.e., $D = \sum(X_1(t_{i+1}) - X_1(t_i))$ where the sum is taken over subpaths that start with a jump downward, and similarly U is the total horizontal distance covered during subpaths above θ. Since it is difficult to see how the likelihood behaves (e.g., the random variables D and U may take positive or negative values), we content ourselves with some simulation results for the same scenarios as in Cases 2 and 3, but now taking $\mu_1^+ < \mu_2$. These can be found in Table 2.4. Again it seems clear that the change of measure (2.35) is asymptotically

$(\lambda,\mu_1,\mu_1^+,\mu_2) = (0.1,0.7,0.15,0.2)$			$(\lambda,\mu_1,\mu_1^+,\mu_2) = (0.3,0.36,0.32,0.34)$		
B	$\hat{\psi}_B$	p_B	B	$\hat{\psi}_B$	p_B
20	1.92	$3.79 \cdot 10^{-7} \pm 1.09 \cdot 10^{-9}$	20	0.45	$5.6 \cdot 10^{-2} \pm 1.01 \cdot 10^{-4}$
50	1.95	$1.26 \cdot 10^{-16} \pm 5.08 \cdot 10^{-19}$	50	1.34	$1.17 \cdot 10^{-4} \pm 2.85 \cdot 10^{-5}$
100	1.98	$3.54 \cdot 10^{-32} \pm 1.89 \cdot 10^{-34}$	100	1.45	$1.69 \cdot 10^{-6} \pm 1.23 \cdot 10^{-7}$

Table 2.4: Simulation results for the slowdown system, Case 4

efficient for the first parameter setting (since $\hat{\psi}_B$ tends to 2 as B grows), but not for the second, in which the loads of both queues are close to 1. Based on these and other simulation results we found that in the current Case 4, the region of asymptotical efficiency is somewhat smaller then we found in Case 2.

Case 5, i.e. $\mu_1^+ < \mu_1 < \mu_2$

In this section we briefly provide simulation results for Case 5, where the first server is always the bottleneck. The new measure is given by Proposition 2.13. See Table 2.5 for results, in which the left part corresponds to the left part of Table 2.2; note that the overflow probabilities are much smaller, due to the slowdown property of the system.

$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.13, 0.17, 0.14, 0.7)$			$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.25, 0.35, 0.28, 0.4)$		
B	$\hat{\psi}_B$	p_B	B	$\hat{\psi}_B$	p_B
20	1.69	$4.44 \cdot 10^{-15} \pm 1.43 \cdot 10^{-15}$	20	0.93	$1.28 \cdot 10^{-4} \pm 2.79 \cdot 10^{-5}$
50	1.86	$1.31 \cdot 10^{-37} \pm 8.23 \cdot 10^{-38}$	50	1.63	$3.51 \cdot 10^{-11} \pm 4.02 \cdot 10^{-12}$
100	1.93	$3.21 \cdot 10^{-75} \pm 5.02 \cdot 10^{-76}$	100	1.85	$5.61 \cdot 10^{-22} \pm 7.55 \cdot 10^{-23}$

Table 2.5: Simulation results for the slowdown system, Case 5

2.5 Conclusions

In this chapter we used a large deviations approach to identify the typical overflow trajectories and designed some (naive) state-independent IS schemes. These schemes can lead to estimators with high relative error, especially when the first queue is the bottleneck. We conclude that state-independent IS does in general not lead to schemes that are asymptotically efficient for *all* parameter values. This is why we will concentrate on state-dependent IS schemes in the next chapters, even though the current scheme performs well for some parameter settings (in fact much better than the state-dependent schemes to be introduced).

2.6 Appendix

In this Appendix we show how the optimal path to overflow in the second buffer can be found for the two-node Jackson network and the slowdown network. We start with the first model.

Cost and shape of the optimal path for tandem Jackson network

We need to consider only four possible types of paths, due to Theorem 2.4. These are illustrated in Figure 2.7; however note that for paths of type (4) that follow the horizontal axis and the interior, but not the vertical axis, the slope in the interior need not be positive. To obtain the optimal path we should calculate the minimal cost for each of these types of path, and then take the minimum over these four outcomes.

Note that the answer will depend on the case we consider. As an example, we will here consider the minimal cost of paths of type (4) for Case 2, i.e. the case in which $\lambda < \mu_2 < \mu_1$.

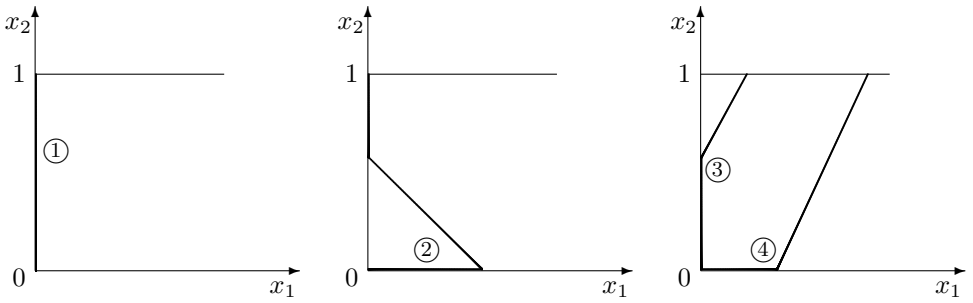

Figure 2.7: General forms of paths to overflow in second buffer

Paths of type (4)

It is clear that the cost of such paths consists of two parts: the cost of the subpath on the horizontal boundary and the cost of the subpath in the interior. Formally we have the following for the cost of the entire path,

$$\inf \left\{ \upsilon \frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\lambda} - \tilde{\mu}_1} + \frac{\mathbb{I}(\bar{\lambda},\, \bar{\mu}_1,\, \bar{\mu}_2)}{\bar{\mu}_1 - \bar{\mu}_2} \right\}, \tag{2.38}$$

where υ is the length of the horizontal part. The infimum is taken over variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$ and $\bar{\mu}_2$ that satisfy $\tilde{\lambda} > \tilde{\mu}_1$, $\tilde{\mu}_1 < \tilde{\mu}_2$ and $\bar{\mu}_1 > \bar{\mu}_2$. Below we will treat separately the cases in which the path in the interior has north-east drift, i.e. $\bar{\lambda} \geq \bar{\mu}_1$, or north-west drift, i.e. $\bar{\lambda} < \bar{\mu}_1$.

In the first case ($\bar{\lambda} \geq \bar{\mu}_1 > \bar{\mu}_2$) the horizontal subpath does not help us to reach high values in the second buffer, but it increases the cost of the overall path. So the first step to optimize (2.38) is to set $\omega = 0$. Now let us optimize the path in the interior under the condition $\bar{\lambda} \geq \bar{\mu}_1 > \bar{\mu}_2$; formally, we need to find a set of parameters that minimizes

$$\frac{\mathbb{I}(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2)}{\bar{\mu}_1 - \bar{\mu}_2}.$$

Calculating the partial derivatives with respect to $\bar{\lambda}, \bar{\mu}_1$ and $\bar{\mu}_2$, setting them equal to zero and solving the resulting system we find the following solutions,

$$(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) = (\lambda, \mu_1, \mu_2) \text{ and } (\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) = (\lambda, \mu_2, \mu_1).$$

Both of these parameter settings do not satisfy our condition $\bar{\lambda} \geq \bar{\mu}_1 > \bar{\mu}_2$. Hence the minimum is attained at the boundary where $\bar{\lambda} = \bar{\mu}_1$, which corresponds to a

vertical path in the interior. Minimizing over the two remaining variables, we find the minimizers of (2.38) to be the same as in Case 4 for the slowdown system, see Section 2.4, namely $\bar{\lambda} = \bar{\mu}_1 = \sqrt{\lambda\mu_1/z^+}$, $\bar{\mu}_2 = \mu_2 z^+$, where z^+ is the unique solution in $(0,1)$ of equation (2.36). The corresponding minimal cost is given by $-\log(z^+)$. We emphasize that the vertical path we found lies in the interior along the vertical boundary, which is different from the (optimal) path of type (1), in which there is a horizontal drift towards the vertical axis.

Let us now consider the other subtype of paths of type (4), for which we have $\bar{\lambda} < \bar{\mu}_1$ and $\bar{\mu}_2 < \bar{\mu}_1$. This time we should set $\omega = -\alpha^{-1}$, where α is the slope of the second subpath (see (2.19), with tildes replaced by bars), for if we choose $\omega < -\alpha^{-1}$ we will have a path of type (2), instead of (4). On the other hand, if we choose $\upsilon > -\alpha^{-1}$, say $\upsilon = -\alpha^{-1} + \delta$ with $\delta > 0$, then the cost of the path $(0,0) \to (-\alpha^{-1} + \delta, 0) \to (\delta, 1)$ is equal to the cost of the path $(0,0) \to (-\alpha^{-1}, 0) \to (0,1)$ plus the cost of the subpath $(0,0) \to (\delta, 0)$. In other words it is optimal to set $\delta = 0$. The minimization of (2.38) is now similar to that in Section 2.3 (Case 1), the only difference being that there $\mu_1 < \mu_2$. As a result, the infimum is not attained at $(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2) = (\mu_1, \mu_2, \lambda)$, since then $\bar{\lambda} < \bar{\mu}_1$ is not satisfied, but rather at the boundary where $\bar{\lambda} = \bar{\mu}_1$. In other words, we find the same vertical path as above, with $\omega = 0$. Therefore this is the optimum over all paths of type (4).

Paths of other types

In Section 2.3 (Case 2) it was found that the minimal cost for paths of type (1) is $-\log(\rho_2)$. Since $z^+ < \rho_2$, this means that it is cheaper to follow the vertical axis, than to follow a vertical path through the interior.

The cost of a path of type (2) consists of two parts: the cost of the subpath following the vertical axis and the cost of the remainder of the path. The optimal cost for the vertical boundary part is $-\log(\rho_2)$; for the remainder of the path the optimal shape is a vertical line in the interior (see the case for path (4) where $\bar{\lambda} < \bar{\mu}_1$ and $\bar{\mu}_2 < \bar{\mu}_1$). But since the cost of this is higher than following the vertical axis, we see that the optimal path of type (2) is actually the limiting case where the starting point and end point of the interior subpath are both equal to the origin. In other words, the optimal path of type (2) coincides with the path of type (1).

For paths of type (3) we can use similar arguments to show that also here the optimal path is the same as the path of type (1).

Cost and shape of the optimal path for slowdown network

In this second part of the Appendix we show how the most likely path to overflow in the second buffer for the slowdown network is found. Again Theorem 2.4 gives us all possible paths types, which are illustrated in Figure 2.8. Note that for paths of type (2) the slopes in the interior may have any sign, which also holds for the upper part of path type (3). Paths of type (2), (3) and (6) include paths in which the first part (following the horizontal or vertical boundary) is absent. In the following we

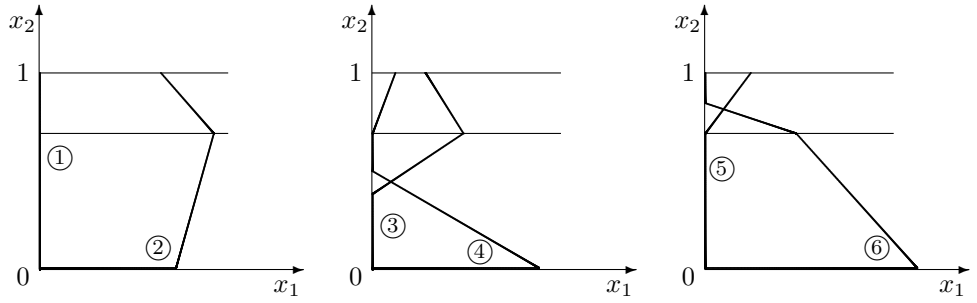

Figure 2.8: General forms of paths to overflow in second buffer

restrict ourselves to the most interesting Case 4, where the bottleneck shifts from the second to the first queue after hitting the slowdown threshold, i.e. $\mu_1^+ < \mu_2 < \mu_1$. For Cases 3 and 5 similar calculations hold. In Proposition 2.12 we claim that the optimal path is the special case of path type (5), in which the second part of the path is a vertical line in the interior. In the remainder we will mean this particular path when we mention the path of type (5), and we will compare other path types with it.

Paths of type (2)

Let us start with paths of type (2). We divide this type in three subtypes: paths without a first horizontal part, paths with horizontal part and north-west drift in the interior below the slowdown threshold, and paths with horizontal part and north-east drift in the interior below the slowdown threshold. In the latter two cases we need not consider paths with north-east drift above θ: for such paths of the third subtype it is not optimal to follow the horizontal boundary in eastern direction, so the optimal version of such a path belongs to the first subtype. It follows from Section 2.3 that any path of the second subtype with north-east drift above θ is also not optimal.

We first study paths without the first horizontal part. The optimal cost of such a path is

$$\inf\left\{\theta\frac{\mathbb{I}(\bar{\lambda},\,\bar{\mu}_1,\,\bar{\mu}_2)}{\bar{\mu}_1-\bar{\mu}_2}+(1-\theta)\frac{\mathbb{I}(\hat{\lambda},\,\hat{\mu}_1^+,\,\hat{\mu}_2)}{\hat{\mu}_1^+-\hat{\mu}_2}\right\}, \tag{2.39}$$

where the infimum is taken over variables $\bar{\lambda}$, $\bar{\mu}_1$, $\bar{\mu}_2$, $\hat{\lambda}$, $\hat{\mu}_1$ and $\hat{\mu}_2$ that satisfy $\bar{\lambda} \geq \bar{\mu}_1$, $\bar{\mu}_1 > \bar{\mu}_2$, $\hat{\mu}_1^+ > \hat{\mu}_2$, as well as the additional condition

$$\theta\frac{\bar{\lambda}-\bar{\mu}_1}{\bar{\mu}_1-\bar{\mu}_2} \geq (1-\theta)\frac{\hat{\mu}_1^+-\hat{\lambda}}{\hat{\mu}_1^+-\hat{\mu}_2},$$

which ensures that the path will not hit the vertical boundary below level 1 if $\hat{\mu}_1^+ > \hat{\lambda}$.

Paths of the second subtype consist of three parts: one part following the horizontal boundary and two parts traversing the interior with north-west drift (below and above θ). The optimal cost of such a path is:

$$\inf\left\{\upsilon_1\frac{\mathbb{I}(\tilde{\lambda},\,\tilde{\mu}_1,\,\tilde{\mu}_2)}{\tilde{\lambda}-\tilde{\mu}_1}+\theta\frac{\mathbb{I}(\bar{\lambda},\,\bar{\mu}_1,\,\bar{\mu}_2)}{\bar{\mu}_1-\bar{\mu}_2}+(1-\theta)\frac{\mathbb{I}(\hat{\lambda},\,\hat{\mu}_1^+,\,\hat{\mu}_2)}{\hat{\mu}_1^+-\hat{\mu}_2}\right\},\tag{2.40}$$

where

$$\upsilon_1=(1-\theta)\frac{\hat{\mu}_1^+-\hat{\lambda}}{\hat{\mu}_1^+-\hat{\mu}_2}+\theta\frac{\bar{\mu}_1-\bar{\lambda}}{\bar{\mu}_1-\bar{\mu}_2},$$

which guarantees that $(0,1)$ is the end point of the path (note that any path which hits ∂_e at some point $(x,1)$ with $x>0$ is not optimal, using the same arguments as for paths of type 4 in the two-node Jackson network). The infimum is taken over all variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$, $\bar{\mu}_2$, $\hat{\lambda}$, $\hat{\mu}_1^+$ and $\hat{\mu}_2$ that satisfy $\tilde{\lambda}>\tilde{\mu}_1$, $\tilde{\mu}_1<\tilde{\mu}_2$, $\bar{\lambda}\leq\bar{\mu}_1$, $\bar{\mu}_1>\bar{\mu}_2$, $\hat{\lambda}\leq\hat{\mu}_1^+$ and $\hat{\mu}_1^+>\hat{\mu}_2$.

It remains to consider paths of the third subtype: first following the horizontal boundary, then traversing the interior below θ with north-east drift and a last part in the interior above θ which has north-west drift. The minimal cost of such a path is

$$\inf\left\{\upsilon_2\frac{\mathbb{I}(\tilde{\lambda},\,\tilde{\mu}_1,\,\tilde{\mu}_2)}{\tilde{\lambda}-\tilde{\mu}_1}+\theta\frac{\mathbb{I}(\bar{\lambda},\,\bar{\mu}_1,\,\bar{\mu}_2)}{\bar{\lambda}-\bar{\mu}_1}+(1-\theta)\frac{\mathbb{I}(\hat{\lambda},\,\hat{\mu}_1^+,\,\hat{\mu}_2)}{\hat{\mu}_1^+-\hat{\mu}_2}\right\},\tag{2.41}$$

where the choice of

$$\upsilon_2=(1-\theta)\frac{\hat{\mu}_1^+-\hat{\lambda}}{\hat{\mu}_1^+-\hat{\mu}_2}-\theta\frac{\bar{\lambda}-\bar{\mu}_1}{\bar{\mu}_1-\bar{\mu}_2}$$

guarantees that $(0,1)$ is the end point of the path. The infimum is taken over all variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$, $\bar{\mu}_2$, $\hat{\lambda}$, $\hat{\mu}_1^+$ and $\hat{\mu}_2$ that satisfy $\tilde{\lambda}>\tilde{\mu}_1$, $\tilde{\mu}_1<\tilde{\mu}_2$, $\bar{\lambda}\geq\bar{\mu}_1$, $\bar{\mu}_1>\bar{\mu}_2$, $\hat{\lambda}\leq\hat{\mu}_1^+$ and $\hat{\mu}_1^+>\hat{\mu}_2$.

Optimization of the costs in (2.39) – (2.41) provides the optimal shape for paths of type (2), which turns out to be a vertical line in the interior. More precisely, the optimal values of the parameters with bars and hats are given by equations as in (2.35)–(2.36); as a result, the horizontal part always vanishes (i.e. $\upsilon_1=\upsilon_2=0$). Thus, it is clear that the cost of any path of type (2) is higher then the optimal cost of path (5).

Paths of type (3)

Let us continue with paths of type (3), these consist of three parts in general: a first part following the vertical boundary, a second part in the interior below θ with north-east drift and a third part in the interior above θ with north-west drift (north-east drift here is excluded since the optimal path can only end in $(0,1)$, as before).

The optimal cost of such a path is

$$\inf\left\{\upsilon_3\frac{\mathbb{I}(\tilde{\lambda},\,\tilde{\mu}_1,\,\tilde{\mu}_2)}{\tilde{\lambda}-\tilde{\mu}_2}+(\theta-\upsilon_3)\frac{\mathbb{I}(\bar{\lambda},\,\bar{\mu}_1,\,\bar{\mu}_2)}{\bar{\lambda}-\bar{\mu}_1}+(1-\theta)\frac{\mathbb{I}(\hat{\lambda},\,\hat{\mu}_1^+,\,\hat{\mu}_2)}{\hat{\mu}_1^+-\hat{\mu}_2}\right\},$$

where

$$\upsilon_3=\theta-(1-\theta)\frac{\hat{\mu}_1^+-\hat{\lambda}}{\hat{\mu}_1^+-\hat{\mu}_2}\frac{\bar{\mu}_1-\bar{\mu}_2}{\bar{\lambda}-\bar{\mu}_1}$$

is the length of the first part, and the infimum is taken over variables $\tilde{\lambda}$, $\tilde{\mu}_1$, $\tilde{\mu}_2$, $\bar{\lambda}$, $\bar{\mu}_1$, $\bar{\mu}_2$, $\hat{\lambda}$, $\hat{\mu}_1^+$ and $\hat{\mu}_2$ that satisfy $\tilde{\lambda}>\tilde{\mu}_2$, $\tilde{\lambda}<\tilde{\mu}_1$, $\bar{\lambda}\geq\bar{\mu}_1$, $\bar{\mu}_1>\bar{\mu}_2$ and $\hat{\mu}_1^+>\hat{\mu}_2$. Indeed a path which hits the vertical axis below level 1 and follows it afterwards is not optimal (Lemma 2.6); any path hitting level 1 at any point with positive first coordinate is not optimal either (see explanations for paths of type 2). After optimization we obtain that $(\tilde{\lambda},\,\tilde{\mu}_1,\,\tilde{\mu}_2)=(\mu_2,\,\mu_1,\,\lambda)$ and $\hat{\lambda}=\hat{\mu}_1^+$. As a result, $\upsilon_3=\theta$, so that we find in fact the optimal path of type (5) as optimal outcome for paths of type (3).

Paths of other types

First we mention that the calculations provided in Section 2.4 and our knowledge about the shape of the most probable path for the tandem Jackson network ensures that paths of type (1) have higher cost than the optimal path of type (5). These paths coincide below θ, but (1) is more expensive above it. After optimization, path (4), as well as path (3), converges to the path (5). Finally, we point out that the cost of paths of type (6) is higher then the cost of type (2), because following the vertical axis above θ is more expensive than traversing the interior along it, following a vertical path. This completes our discussion about the optimal shape for Case 3 for the slowdown network.

Chapter 3

State-dependent importance sampling for tandem Jackson network

In this chapter we focus on estimating the probability of overflow in the downstream queue of the tandem Jackson network, starting from *any* state, in contrast to Chapter 2 where the origin was the starting state. First of all, we provide a rigorous large deviations analysis of this network. Since, 'traditional' state-independent importance-sampling distributions may perform poorly, see also Chapter 2, we then introduce a very accurate state-dependent change of measure inspired by (and partially based on) [22]. Accuracy of this scheme is obtained by re-calculating the new measure after each transition. We analytically prove asymptotic efficiency of the scheme. The proof we provide is rather elementary and short, and relies on probabilistic arguments only.

Finally, we stress that the focus in this chapter lies on the analytic aspects of the problem, that is, the analysis of the decay rate and the proof of asymptotic efficiency, even though various numerical results are also presented.

3.1 Optimal path and related change of measure

As we stated before, the first step to construct a good change of measure for IS simulations, is usually to find the most probable path to overflow, i.e., the way in which overflow most probably occurs, conditional on its occurrence. In Section 3.1.1 we show how this can be found, together with the corresponding *state-dependent* change of measure $(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x))$. As in Chapter 2 we minimize cost functions of the type (2.15), the reasoning being based on large deviations analysis. However, this time we will provide a proof that the path with minimal cost is indeed the most

probable one (see Section 3.2). To actually find the most probable path and related change of measure, we split the problem in two cases, for which the minimization procedure gives different results. In Section 3.1.2 we treat the case $\lambda < \mu_2 < \mu_1$, in which the second server is the bottleneck, while Section 3.1.3 deals with the case $\lambda < \mu_1 \leq \mu_2$, in which the first server is the bottleneck. Beforehand, we would like to point out that the change of measure mentioned above, denoted by tildes, is not the same as the asymptotically efficient change of measure that will be introduced in Section 3.3 (denoted by bars), although it is closely related.

3.1.1 Cost and structure of path to overflow

The typical path to overflow in the particular case that the origin is the starting point, has already been identified for the d-node Jackson tandem network in [2], and hence also for our tandem system. In that paper, the time-reversed process is used to find the shape of the most probable path to overflow. This path to overflow was also obtained as a corollary result in Chapter 2, and in this section we present a method similar to the one in Chapter 2 to find the optimal path starting from *any* state $x \in \bar{D}$. The advantage of this method is that it also provides a 'good' change of measure, which ensures that most simulation runs under this new measure will be close to the optimal path. This new measure will be the basis for another change of measure, which is used in our (state-dependent) IS scheme, as presented in Section 3.3. Another result of the method is the exponential decay rate of p_B^s, which will be determined in Section 3.2, and which will play a crucial role in the proofs of asymptotic efficiency of Section 3.3.

Before introducing the method we impose some restrictions on the path structures we consider which may be seen as the generalization of Theorem 2.4. We leave the proof that the typical path to overflow indeed satisfies these restrictions to Section 3.2, see Lemma 3.6. The restrictions are as follows.

Property 3.1.

- Each path is a concatenation of subpaths, which are straight lines on any of the subsets D, ∂_1 and ∂_2, see also (2.1), and the new measure stays constant along each subpath, i.e., $\tilde{\lambda}(x) = \tilde{\lambda}$, $\tilde{\mu}_1(x) = \tilde{\mu}_1$ and $\tilde{\mu}_2(x) = \tilde{\mu}_2$, for any state x on the same subpath;

- Each path does not have more than one subpath in each subset if $\mu_2 < \mu_1$;

- Each path does not have more than two subpaths in each subset if $\mu_2 \geq \mu_1$.

With every path that satisfies Property 3.1 one associates a cost (coming from the theory of large deviations [17, 19, 25, 69]). The main idea (to be proved in Section 3.2) being that the minimal cost of the path to overflow in the second buffer, starting from state s, can be interpreted as the decay rate of the probability of interest. An

important role will be played by the family of cost functions (2.17):

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

We will now explain the cost method in more detail in the following two examples, one considering the path which is a straight line and the other considering the paths which consist of two straight subpaths. More background on the special case of the paths that start in the origin can be found in the Appendix of Chapter 2 of this thesis.

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

$$\alpha = \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\mu}_1 - \tilde{\lambda}}. \tag{3.1}$$

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

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

see also (2.18). To find the cost per unit vertical (horizontal) distance, we need to divide this by the vertical speed $\tilde{\mu}_1 - \tilde{\mu}_2$ (horizontal speed $\tilde{\mu}_1 - \tilde{\lambda}$). Thus, minimizing the cost of any straight path from x to y in this case boils down to minimizing

$$(y_2 - x_2) \frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\mu}_1 - \tilde{\mu}_2}, \tag{3.2}$$

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

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

in addition, we should have that

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

to guarantee that y is indeed the ending state of the path when it starts at x.

It is easily checked that the total cost (3.2) with ending state $y = (0, 1)$ attains its minimum when triplet $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ is a solution to

$$\begin{cases} \lambda^{(\mathrm{syst})} = \mu_1^{(\mathrm{syst})} - \frac{x_1}{1-x_2}(\mu_1^{(\mathrm{syst})} - \mu_2^{(\mathrm{syst})}) \\ \lambda^{(\mathrm{syst})} + \mu_1^{(\mathrm{syst})} + \mu_2^{(\mathrm{syst})} = \lambda + \mu_1 + \mu_2 \\ \lambda^{(\mathrm{syst})} \mu_1^{(\mathrm{syst})} \mu_2^{(\mathrm{syst})} = \lambda \mu_1 \mu_2 \\ \lambda^{(\mathrm{syst})} \leq \mu_1^{(\mathrm{syst})} \text{ and } \mu_1^{(\mathrm{syst})} > \mu_2^{(\mathrm{syst})} \\ \lambda^{(\mathrm{syst})}, \mu_1^{(\mathrm{syst})}, \mu_2^{(\mathrm{syst})} > 0, \end{cases} \tag{3.3}$$

where the superscript "(syst)" indicates that the optimal change of measure is the solution to system (3.3). The reason we have chosen the specific ending state $(0,1)$ is that it is the most frequent ending state for our network. Notice also that if $(\lambda^{(\text{syst})}(x), \mu_1^{(\text{syst})}(x), \mu_2^{(\text{syst})}(x))$ is the solution to (3.3) for some starting state x, it also minimizes this system if we replace x by any state that belongs to the straight line between x and $y = (0,1)$. ♢

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

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

where $\alpha = (\tilde{\mu}_1 - \tilde{\mu}_2)/(\tilde{\mu}_1 - \tilde{\lambda})$, see (3.1). The first term in the sum is the cost of the first subpath under some new measure $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and the second term is the cost of the second (vertical) subpath under some measure $(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2)$. We need to optimize this expression such that

$$\tilde{\lambda} < \tilde{\mu}_1, \quad \tilde{\mu}_2 < \tilde{\mu}_1, \quad \hat{\lambda} \le \hat{\mu}_1 \quad \text{and} \quad \hat{\mu}_2 < \hat{\mu}_1$$

over all parameters marked with tildes and hats. The result of this minimization may depend on the relation between the service rates. For instance, if $\mu_2 < \mu_1$, it is readily verified that the minimal cost of this path type is obtained when the new measure is given by

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

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

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

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

at our disposal. In Theorem 3.3 we will prove that this is in fact the exponential decay rate of the probability p_B^s as $B \to \infty$.

We now present the results of the minimum-cost-path method for both cases of the tandem network.

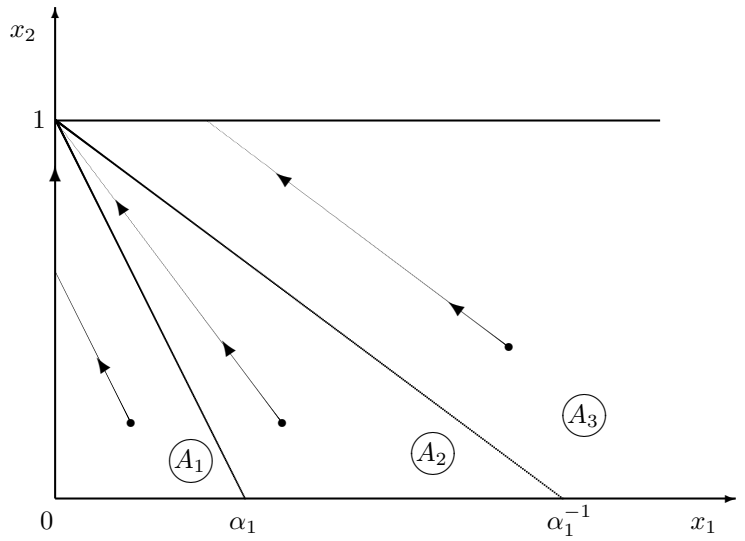

Figure 3.1: Partition of $\bar{D}$ and some optimal paths to overflow when $\mu_2 < \mu_1$.

3.1.2 Importance sampling for $\mu_2 < \mu_1$

When $\mu_2 < \mu_1$, the cost minimization for state x as outlined in the previous section (in particular Example 3.3; see also the Appendix in Chapter 2), yields the following new measure after some calculations:

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) = \begin{cases} (\mu_2, \mu_1, \lambda), & \text{if } x \in A_1, \\ (\lambda^{(\text{syst})}(x), \mu_1^{(\text{syst})}(x), \mu_2^{(\text{syst})}(x)), & \text{if } x \in A_2, \\ (\lambda, \mu_1, \mu_2), & \text{if } x \in A_3 \end{cases} \tag{3.4}$$

The partition A_i, $i = 1, 2, 3$ is also a result of the cost minimization procedure and is given by

$$\begin{aligned} A_1 &:= \{x \in \bar{D} : x_2 \leq -x_1/\alpha_1 + 1\}, \\ A_2 &:= \{x \in \bar{D} : -x_1/\alpha_1 + 1 < x_2 < -\alpha_1 x_1 + 1\}, \\ A_3 &:= \{x \in \bar{D} : x_2 \geq -\alpha_1 x_1 + 1\}, \end{aligned} \tag{3.5}$$

with $\alpha_1 := (\mu_1 - \mu_2)/(\mu_1 - \lambda)$. See also Figure 3.1. Note that the path considered in Example 3.3 in the previous subsection is optimal for any starting state $s \in A_1$, and the corresponding new measure (exchanging λ and μ_2) was earlier found by Parekh and Walrand [59] for the problem of reaching a large total queue population. Also, we point out that the change of measure is continuous in the state x, as can be verified by solving system (3.3) for $x = (\alpha_1, 0)$ and $x = (\alpha_1^{-1}, 0)$, yielding the solutions in the first and third lines of (3.4), respectively.

The corresponding path from starting state $s = (s_1, s_2)$ to some state on ∂_e is

given by

$$\begin{array}{ll} (s_1, s_2) \to (0, s_2 + \alpha_1^{-1} s_1) \to (0, 1), & \text{if } s \in A_1, \\ (s_1, s_2) \to (0, 1), & \text{if } s \in A_2, \\ (s_1, s_2) \to (s_1 - \alpha_1^{-1} s_2, 1), & \text{if } s \in A_3. \end{array} \tag{3.6}$$

The residual cost $\gamma(x)$ of the path at state x is given by:

$$\gamma(x) = \begin{cases} (1 - x_1 - x_2)\gamma, & \text{if } x \in A_1, \\ -x_1 \log \frac{\tilde{\lambda}(x)}{\lambda} - (1 - x_2) \log \frac{\tilde{\mu}_2(x)}{\mu_2}, & \text{if } x \in A_2, \\ 0, & \text{if } x \in A_3, \end{cases} \tag{3.7}$$

where

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

is the minimal cost of the path $(0, 0) \to (0, 1)$. The total cost of the optimal path starting in the state s is $\gamma(s)$.

We end this subsection with some interesting properties of the new measure defined in (3.4), to be used later. Intuitively, it says that for any state x, the new measure 'lies between' the Parekh and Walrand measure where λ and μ_2 are interchanged, and the original measure, where the parameters retain their original values. Moreover, the more jobs are present in the system at time zero, either in queue 1 or in queue 2, the 'less change of measure' we need. This perfectly coincides with the structure of the most probable path, see (3.6).

Lemma 3.1. *When $\mu_2 < \mu_1$, the functions $\tilde{\lambda}(x)$, $\tilde{\mu}_1(x)$ and $\tilde{\mu}_2(x)$ as defined in (3.4) are continuous and differentiable, satisfying the following for any $x \in \bar{D}$.*

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

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

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

Proof. (i) We only need to consider $x \in A_2$, since otherwise all partial derivatives are zero. Applying implicit differentiation to (3.3) one finds

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

where the last inequality follows from the fact that $\tilde{\mu}_1(x) > \tilde{\mu}_2(x)$. The other statements follow similarly.

(ii) It follows from (3.4) that $\tilde{\lambda}(x) = \mu_2$ if $x \in \partial_1$ and $\tilde{\lambda}(x) = \lambda$ if $x \in A_3$, so applying the first statement of this lemma one can find that $\tilde{\lambda}(x) \in [\lambda, \mu_2]$. Using similar arguments one can obtain the same bounds for $\tilde{\mu}_2(x)$.

(iii) We show that the partial derivatives with respect to x_1 and x_2 of $\gamma(x)$ as given in (3.7) are not positive. For $x \in A_1 \cup A_3$ this is obvious, while for $x \in A_2$ it can be checked using implicit differentiation, similar to the proof of the first statement. $\square$

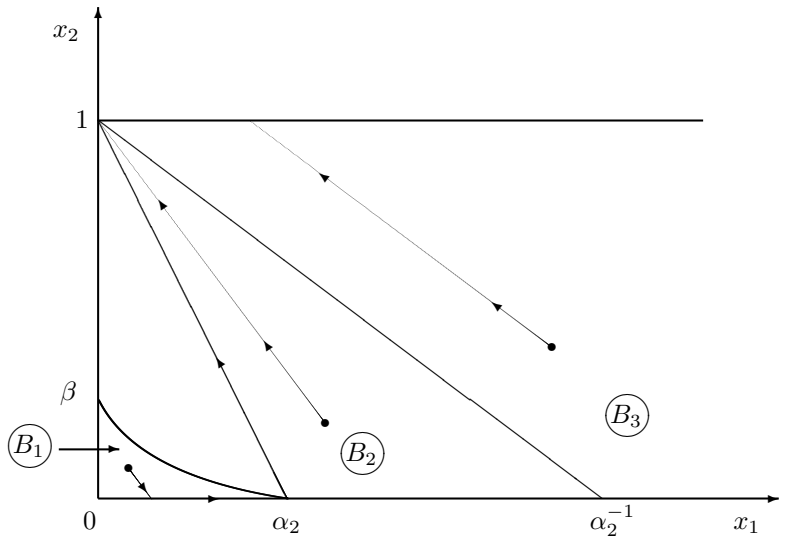

Figure 3.2: Partition of $\bar{D}$ and some optimal paths to overflow when $\mu_1 \leq \mu_2$.

Lemma 3.1 does not yield results on $\tilde{\mu}_1(x)$ since they are not needed in the sequel, but it may be interesting to note that $\tilde{\mu}_1(x)$ is *not* monotone. In fact, $\tilde{\mu}_1(x) = \mu_1$ when $x \in A_1$ and when $x \in A_3$, but also when $x_1 + x_2 = 1$, so it is also neither convex nor concave.

3.1.3 Importance sampling for $\mu_1 \leq \mu_2$

The new measure under which the path to overflow has minimal cost in terms of (2.15) is as follows:

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) = \begin{cases} (\mu_1, \lambda, \mu_2), & \text{if } x \in B_1, \\ (\lambda^{(\text{syst})}(x), \mu_1^{(\text{syst})}(x), \mu_2^{(\text{syst})}(x)), & \text{if } x \in B_2, \\ (\lambda, \mu_2, \mu_1), & \text{if } x \in B_3. \end{cases} \tag{3.8}$$

Again we partitioned the state space into three subspaces B_i, $i = 1, 2, 3$ as follows, see also Figure 3.2.

$$\begin{aligned} B_1 &:= \{x \in \bar{D} : f(x) \leq 0\}, \\ B_2 &:= \{x \in \bar{D} : f(x) > 0 \text{ and } x_2 < -\alpha_2 x_1 + 1\}, \\ B_3 &:= \{x \in \bar{D} : x_2 \geq -\alpha_2 x_1 + 1\}, \end{aligned} \tag{3.9}$$

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

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

with $\tilde{\lambda} \equiv \tilde{\lambda}(x)$ and $\tilde{\mu}_2 \equiv \tilde{\mu}_2(x)$ being the solution to (3.3). The zero level curve of the function $f(x)$ represents the boundary between subspaces B_1 and B_2, β is the unique solution to $f(0, x_2) = 0$. Interestingly, for the current case the change of measure is *not* continuous in states x that lie on this boundary (i.e., $f(x) = 0$), and the behavior on B_1 and B_2 is entirely different. In particular, the change of measure on B_2 has $\tilde{\lambda}(x) < \tilde{\mu}_1(x)$ and $\tilde{\mu}_2(x) < \tilde{\mu}_1(x)$, as opposed to the first line of (3.8) where both inequalities are reversed. This is also reflected in a different shape of the typical path from $s = (s_1, s_2)$ to ∂_e:

$$\begin{array}{ll} (s_1, s_2) \to (s_1 + \alpha_3 s_2, 0) \to (\alpha_2, 0) \to (0, 1), & \text{if } s \in B_1, \\ (s_1, s_2) \to (0, 1), & \text{if } s \in B_2, \\ (s_1, s_2) \to (s_1 - \alpha_2^{-1} s_2, 1), & \text{if } s \in B_3, \end{array} \tag{3.10}$$

where $\alpha_3 := (\mu_2 - \lambda)/(\mu_1 - \lambda)$. Note that the last part of any path with starting state $s \in B_1$ is just a special case of a path starting in B_2 (in this case starting in $(\alpha_2, 0)$), but the corresponding new measure on this line (i.e. the solution to system (3.3) for $x = (\alpha_2, 0)$) can be given explicitly as (μ_1, μ_2, λ), see (2.26).

The next result we give is $\gamma(x)$, the residual cost of the optimal path in state x in terms of (2.15):

$$\gamma(x) = \begin{cases} \gamma - x_1 \log \frac{\mu_1}{\lambda}, & \text{if } x \in B_1, \\ -x_1 \log \frac{\tilde{\lambda}(x)}{\lambda} - (1 - x_2) \log \frac{\tilde{\mu}_2(x)}{\mu_2}, & \text{if } x \in B_2, \\ (1 - x_2) \log \frac{\mu_2}{\mu_1}, & \text{if } x \in B_3. \end{cases} \tag{3.11}$$

We end this subsection with the analogue of Lemma 3.1 in the case when $\mu_1 \le \mu_2$. For this, we first introduce z as the unique solution in the interval $(0, 1)$ of the (essentially cubic) equation

$$\varphi(z) := 1 - \mu_2 z - 2\sqrt{\frac{\lambda \mu_1}{z}} = 0, \tag{3.12}$$

which follows from system (3.3) by taking $(x_1, x_2) = (0, 0)$. (The fact that there is a unique solution immediately follows from $\varphi(0) = -\infty$, $\varphi(1) = \lambda + \mu_1 - 2\sqrt{\lambda\mu_1} \ge 0$, and the fact that $\varphi'(z) = 0$ has just a single positive solution, viz. $\sqrt[3]{\lambda\mu_1/4\mu_2^2}$.) In fact, $-\log z$ is the cost of the vertical path $(0, 0) \to (0, 1)$ *in the interior* (i.e., in D), satisfying $\tilde{\lambda} = \tilde{\mu}_1$ (as opposed to the vertical path *following* ∂_1 in Example 3.3, where $\tilde{\lambda} < \tilde{\mu}_1$). See also (2.36) and [44, Eqns. (30) and (33)] for more details.

Lemma 3.2. *When $\mu_1 \le \mu_2$, the functions $\tilde{\lambda}(x)$, $\tilde{\mu}_1(x)$ and $\tilde{\mu}_2(x)$ as defined in (3.8) are continuous and differentiable, except in states x with $f(x) = 0$. For any such $x \in \bar{D}$, these functions satisfy the following.*

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

(ii) $\tilde{\lambda}(x) \in [\lambda, \sqrt{\lambda\mu_1/z}]$ *and* $\tilde{\mu}_2(x) \in [\mu_2 z, \mu_1] \cup \mu_2$, *where z is defined by (3.12).*

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

Proof. To prove these facts we can use similar arguments as in the proof of Lemma 3.1. An exception is the second statement, where the upper bound for $\tilde{\lambda}(x)$ and the lower bound for $\tilde{\mu}_2(x)$ is attained when $x \in \partial_1$ (see the first statement), i.e., $\tilde{\lambda}(x) \leq \sqrt{\lambda\mu_1/z}$ and $\tilde{\mu}_2(x) \geq \mu_2 z$, where z is the unique solution of (3.12) in the interval $(0, 1)$. □

Note that part *(i)* of Lemma 3.2 implies that the functions are monotone on $B_2 \cup B_3$, but not on all $\bar{D}$, due to the discontinuity on the boundary between B_1 and B_2.

3.2 Large deviations properties

The goal of this section is to formally prove that the cost of the optimal path to overflow is equal to the exponential decay rate of p_B^s, the probability of interest. We also illuminate some important and interesting large deviations properties of the process $X(t)$.

Consider any absolutely continuous function $\phi : [0, \infty) \to \bar{D}$, representing a path associated with the scaled process $X(t)$. The first aim is to define a so-called local rate function $\ell(\phi(t), \dot{\phi}(t))$, which depends both on the position at time t and on the time derivative (or speed vector) $\dot{\phi}(t)$ at time t. To do so, first define three auxiliary functions $L_i(y)$, where the argument y should be interpreted as a speed vector:

$$L_i(y) := \sup_{\theta} \left(\langle \theta, y \rangle - g_i(\theta) \right), \quad i = 0, 1, 2, \tag{3.13}$$

where

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

cf. [66, Eqn. (5.5)]. The second equality applies to ∂_1 and the third equality applies to ∂_2. The function $g_1(\theta)$ does not have a term with μ_1, because jumps of type v_1 from boundary ∂_1 are impossible, and likewise $g_2(\theta)$ does not have a term with μ_2. Finally, the local rate function ℓ can be defined as:

$$\ell(\phi(t), \dot{\phi}(t)) := \begin{cases} L_0(\dot{\phi}(t)), & \text{if } \phi(t) \in D \cup \partial_e, \\ [L_0 \oplus L_1\,](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_1, \\ [L_0 \oplus L_2](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_2, \end{cases} \tag{3.14}$$

where

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

for $i = 1, 2$, is the inf-convolution of the functions L_0 and L_i, see Chapter 7 of [19], the infimum being taken over all values ρ and vectors y_0 and y_i that satisfy the given conditions. Let us briefly explain the usefulness of this inf-convolution on the boundaries of the state space. Assume that the scaled process $X(t)$ follows a path $\phi(t) \in \partial_1$, such that $\partial\phi_2/\partial t > 0$ for $t \in [0, T]$. Hence, the first and second component of the vector y should be zero and strictly positive, respectively. It is clear that the original (unscaled) jump process $Q(t)$ can only increase its second component when it is not on ∂_1, since jumps of type v_1 are not allowed on ∂_1. Therefore, the inf-convolution provides a 'mixture' of the functions L_0 and L_1, supposing that the process $Q(t)$ spends a fraction of time ρ in the interior D and a fraction $1 - \rho$ on the vertical constraint. Note that ρ must be such that $\phi(t)$ has speed y with positive increment in the vertical direction and zero-increment in the horizontal direction, such that the scaled process $X(t)$ remains on ∂_1.

We are now ready to state the following theorem.

Theorem 3.3. *The process $X(t)$ satisfies a large deviations principle with local rate function (3.14), and therefore*

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

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

Sketch of the proof. The proof of this theorem is based on the results presented in [20]. Let us introduce a process $Z(t)$, which is the unconstrained version of $X(t)$, in other words $Z(t)$ is allowed to have negative values in both components. In addition we will assume that $Z(0) = X(0) = x \in \bar{D}$. One can use [20, Thms. 3.2 and 3.4] to show that the map $\Gamma : Z(t) \to X(t)$ exists and Theorem 2.2 from the same paper to show that it is Lipschitz continuous. Γ is known as the Skorokhod map and the question whether it exists is known as the Skorokhod problem; for more background we refer to [20].

Since the map Γ is Lipschitz continuous and the process $Z(t)$ satisfies a large deviation principle, see [66, Thm. 5.1], one can apply the contraction principle (see [66, Thm. 2.13]) and conclude that the process of our interest, $X(t)$, satisfies a large deviations principle with local rate function $\ell(\phi(t), \dot{\phi}(t))$ defined by (3.14). □

Using the local rate function ℓ, as defined in (3.14), the rate function of any path $\phi(t) = (\phi_1(t), \phi_2(t))$ with $t \in [0, T]$ for some T, can be defined as the integral of ℓ over time. The following lemma shows that for paths that stay in one of the subsets $D, \partial_1, \partial_2$, this rate function is minimal when the path is straight, with constant speed vector.

Lemma 3.4. *For any T, consider an absolutely continuous path $\phi(t)$ that remains in D (or in ∂_1, or in ∂_2) for all $t \in [0,T]$. Then,*

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

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

Proof. The proof of this lemma directly follows from convexity of the function ℓ and Jensen's inequality. We also refer to Lemma 2.5 and [66, p. 87]. □

Now assume that $\phi(t) \in D$, for $t \in (0,T)$ is a path between two states x and y. Lemma 3.4 tells us that the path $\phi(t)$ has minimal cost if the process $X(t)$ moves along a straight line at constant speed. A corresponding new measure can be defined as follows

$$\begin{aligned} \tilde{\lambda} &= \lambda e^{\theta_1}, \\ \tilde{\mu}_1 &= \mu_1 e^{\theta_2-\theta_1}, \\ \tilde{\mu}_2 &= \mu_2 e^{-\theta_2}, \end{aligned} \tag{3.15}$$

where $\theta = (\theta_1, \theta_2)$ is the maximizer of (3.13) with $i = 1$. In fact this is exactly the same change of measure we would find using the cost minimization procedure from Section 3.1, due to the immediate equality

$$\ell(\phi(t), \dot{\phi}(t)) = \mathbb{I}(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2). \tag{3.16}$$

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 (3.15). This is not difficult to see since, e.g. for paths on ∂_1, we find for some $\rho \in [0,1]$ that

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

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

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

for some x_1^* and x_2^*. Then we have the following relations between the rate function ℓ defined by (3.14) and the cost function $\mathbb{I}$ from the previous section, defined by (2.18):

Lemma 3.5. *(i) For paths in the interior D,*

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

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

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

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

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

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

Proof. Statement *(i)* is the same as (3.16). We continue to prove statement *(iii)*, the proof of *(ii)* being identical. The restriction $\phi(t) \subset \partial_2$ implies that $\dot{\phi}(t) = (\dot{\phi}_1(t), 0)$ for any $t \in [0, \sigma_2]$. The definition of the inf-convolution tell us that $\tilde{\lambda} v_0 + \tilde{\mu}_1 v_1 + \rho \tilde{\mu}_2 v_2 = \dot{\phi}(t)$. Hence we find that $\rho = \tilde{\mu}_1 / \tilde{\mu}_2$ and $\dot{\phi}_1(t) = \tilde{\lambda} - \tilde{\mu}_1$, from which we can conclude that

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

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

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

Lemma 3.6. *The optimal path from any starting state $s \in \bar{D}$ to ∂_e does not have more than*

(i) one subpath in each subset, if $\mu_2 < \mu_1$,

(ii) two subpaths in each subset, if $\mu_2 \geq \mu_1$.

Proof. Due to the one-to-one correspondence between the rate function of any path in terms of the local rate function ℓ and the cost function $\mathbb{I}$, see Lemma 3.5, the proof of this lemma is similar to the proof of Lemma 2.6.

The main idea of the proof is the following. We consider paths that do not satisfy the properties indicated in Lemma 3.6 (or in Lemma 2.6). Then we bound their costs by the cost of paths which do have that property. □

Theorem 3.7. *The exponential decay rate of p_B^s equals the minimal cost derived in Section 3.1, i.e.,*

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

Proof. By virtue of Lemma 3.6, the optimal path can be represented as a concatenation of k (at most 6) subpaths $\phi^{(1)}, \ldots, \phi^{(k)}$, which stay on different subsets of the state space (i.e., D, ∂_1 and ∂_2). If we denote the starting time and position of the i-th subpath by $t^{(i)}$ and $x^{(i)} = \phi^{(i)}(t^{(i)}) = \phi^{(i-1)}(t^{(i)})$ respectively, with the convention that $t^{(1)} = 0, x^{(1)} = s$, $t^{(k+1)} = \tau, x^{(k+1)} \in \partial_e$, then we can write the decay rate identified in Theorem 3.3 as

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

Using Lemma 3.5 we can rewrite the last expression as follows:

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

where in some of the terms we may need to replace the denominator $\tilde{\lambda} - \tilde{\mu}_1$ by $\tilde{\mu}_1 - \tilde{\mu}_2$, while also changing the prefactor to $x_2^{(i+1)} - x_2^{(i)}$, see Lemma 3.5. Using the fact that the new measure $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ determines the shape of each subpath $\phi^{(i)}$, and the shape of the whole path ϕ as a consequence, we may also take the first infimum over $x^{(1)}, \ldots, x^{(k)}$, rather than $t^{(1)}, \ldots, t^{(k)}$. Applying Lemma 3.5 to the last optimization problem we arrive at

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

which completes the proof. □

3.3 Asymptotic efficiency

It is known from Chapter 2, where the starting state is the origin, that the new measures (3.4) and (3.8) are not always asymptotically efficient. For example, when $\mu_2 < \mu_1$, multiple visits of the process $X(t)$ to the horizontal axis (∂_2) under the new measure (μ_2, μ_1, λ) may cause the likelihood ratio to become very large. To 'protect' the likelihood ratio we use a specific measure around ∂_2, under which these visits become harmless. This approach is similar to the one used in [22]. Similarly, a protection strip is needed along the lower part of the vertical boundary ∂_1 in the case when $\mu_1 \leq \mu_2$.

We again split the problem into two cases: in Section 3.3.1 we explain the method in detail for the situation in which the second server is the bottleneck ($\lambda < \mu_2 < \mu_1$), and in Section 3.3.2 we treat the case in which the first server is the bottleneck ($\lambda < \mu_1 \leq \mu_2$).

3.3.1 Asymptotic efficiency for $\mu_2 < \mu_1$

In order to construct an IS scheme that is provably asymptotically efficient, a function $W(x)$ is defined in the same manner as it was done in [22] for any point $x = (x_1, x_2)$ of the state space. This function will give us an expression for a new measure $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$. First consider the three intermediate functions $W_i(x), i = 1, 2, 3$:

$$\begin{aligned} W_1(x) &:= 2\gamma(x) - \delta, \\ W_2(x) &:= W_1(x_1, \delta/2\gamma) = 2\gamma(x_1, \delta/2\gamma) - \delta, \\ W_3(x) &:= 2\gamma - 3\delta, \end{aligned} \tag{3.17}$$

where δ is some small positive number, and $\gamma(x)$ is given by (3.7). In the next step the minimum of these three functions is defined, see also Figure 3.3:

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

Note that the particular choice of the functions W_i ensures that the shapes of the areas around the origin and ∂_2 on which $\bar{W}$ coincides with the functions W_i are the same as they were in [22]. Our function $W(x)$ is different from the one in [22] since it is based on the decay rate of the probability of interest, rather than on its linear approximation. The last step in the construction is a mollification procedure, which

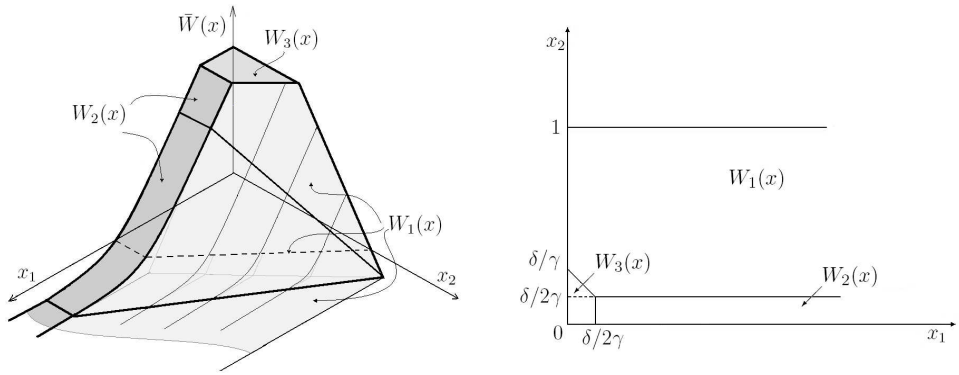

Figure 3.3: The function $\bar{W}(x)$ and the areas on which $\bar{W}(x) = W_i, i = 1, 2, 3$ (case $\mu_2 < \mu_1$).

makes the resulting function $W(x)$ smooth. Again following [22], this can be done by defining:

$$W(x) := -\epsilon \log \sum_{i=1}^{3} e^{-W_i(x)/\epsilon}, \tag{3.18}$$

where ϵ is a 'smoothness' parameter; the larger ϵ is chosen, the smoother the function $W(x)$ is. On the other hand, as $\epsilon \to 0$ we see that $W(x)$ converges to the (non-smooth) function $\bar{W}(x)$. The reason that we prefer to use (3.18) over other smoothing methods is its computational efficiency.

The function $W(x)$, and in particular its gradient, will play a main role in the representation of the state-dependent, asymptotically efficient new measure. However, before turning to this, we need some preliminaries, namely a relation between the gradients of the functions W_i and the measure from the previous sections, and some assumptions on the parameters δ and ϵ.

Proposition 3.4. *The gradients of the functions $W_i(x)$, $i = 1, 2, 3$ can be represented as follows:*

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

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

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

Although we do not know $\tilde{\lambda}(x)$, $\tilde{\mu}_1(x)$ and $\tilde{\mu}_2(x)$ explicitly, we can find their partial derivatives with respect to x_1 and x_2 by implicit differentiation of (3.3), see Lemma 3.1 for more insight. After some elementary algebra we find that the sum of the last two vectors in the previous expression equals zero, which proves the first statement. The other two statements follow easily from the definitions of W_2 and W_3. □

The parameters δ and ϵ depend on B, and in the sequel we will need the following conditions for their asymptotical behavior as B grows large. Note that these are the same conditions as in [15, 22].

Assumption 3.5. *The parameters $\delta \equiv \delta_B$ and $\epsilon \equiv \epsilon_B$ are strictly positive and satisfy the following limit conditions:*

$$(i)\ \epsilon_B \to 0, \quad (ii)\ \delta_B \to 0, \quad (iii)\ B\epsilon_B \to \infty, \quad (iv)\ \varepsilon_B/\delta_B \to 0,$$

as $B \to \infty$.

We will now show how the new measure is constructed from the function W. We inherit the following expression from [22, Prop. 3.2]:

$$\begin{aligned} \bar{\lambda}(p) &= N(p)\lambda e^{-\langle p, v_0\rangle/2}, \\ \bar{\mu}_1(p) &= N(p)\mu_1 e^{-\langle p, v_1\rangle/2}, \\ \bar{\mu}_2(p) &= N(p)\mu_2 e^{-\langle p, v_2\rangle/2}, \end{aligned} \tag{3.19}$$

where

$$N(p) := \left[\lambda e^{-\langle p, v_0\rangle/2} + \mu_1 e^{-\langle p, v_1\rangle/2} + \mu_2 e^{-\langle p, v_2\rangle/2}\right]^{-1} = e^{\mathbb{H}(p)/2}. \tag{3.20}$$

Here $\mathbb{H}(p)$ is a function known as the *Hamiltonian*, which we use here to simplify the notation and to enable the comparison with [15, 22]. The vector p strongly depends on the current state of the process and is in fact taken to be the gradient $DW(x)$. We thus rewrite (3.19) as

$$\bar{\lambda}(x) = \lambda e^{-\langle DW(x), v_0\rangle/2} e^{\mathbb{H}(DW(x))/2}, \tag{3.21}$$

$$\bar{\mu}_i(x) = \mu_i e^{-\langle DW(x), v_i\rangle/2} e^{\mathbb{H}(DW(x))/2}, \quad i = 1, 2.$$

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

$$DW(x) = \sum_{k=1}^{3} \rho_k(x) DW_k(x), \text{ where } \rho_k(x) = \frac{e^{-W_k(x)/\varepsilon}}{\sum_{i=1}^{3} e^{-W_i(x)/\varepsilon}}. \tag{3.22}$$

For the Hamiltonian we have the following results.

Lemma 3.8. *For any x, $\mathbb{H}(DW_1(x)) = \mathbb{H}(DW_3(x)) = 0$ and $\mathbb{H}(DW_2(x)) \geq 0$.*

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

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

and

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

Finally, for the third case we have

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

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

Clearly there is a difference between the new measures defined in Section 3.1 (indicated by tildes) and in this section (indicated by bars). In fact it is not difficult to see that the first one also follows from (3.19) if we replace W by W_1. However, this change of measure is not asymptotically efficient, while the other one is, due to the protection strips along the boundaries, as we will prove in the remainder of this subsection.

Lemma 3.9. *The likelihood $L(\omega)$ of a path $\omega = (X_j, j = 0, \ldots, \sigma)$ under the new measure (3.21) satisfies*

$$\begin{aligned}\log L(\omega) &= \frac{B}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1} - X_j\rangle \\ &+ \sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), v_k\rangle I\{X_j = X_{j+1} \in \partial_k\} \qquad (3.23) \\ &- \frac{1}{2}\sum_{j=0}^{\sigma-1}\mathbb{H}(DW(X_j)).\end{aligned}$$

Proof. This follows from the definitions of the likelihood ratio, see (1.5), and the new measure, see (3.21). We also refer to [15] for a detailed proof. □

Lemma 3.10. *Consider the case $\mu_2 < \mu_1$. For any path $\omega = (X_j, j = 0, ..., \sigma)$ under the new measure (3.21), the first term in (3.23) satisfies*

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

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

Proof. At first let us introduce the following representation

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

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

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

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

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

here

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

We now compute an upper bound for $|h_{11}(x)|$ as an example; the two other terms can be dealt with in the same manner. Using representation (3.22) one can write

$$\frac{\partial^2 W(x)}{\partial x_1^2} = \sum_{k=1}^{3} \left[\rho_k(x) \frac{\partial^2 W_k(x)}{\partial x_1^2} + \frac{\partial \rho_k(x)}{\partial x_1} \cdot \frac{\partial W_k(x)}{\partial x_1}\right], \tag{3.24}$$

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

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

Since the second fraction on the right-hand side turns out to be bounded as $\varepsilon \to 0$, and the same holds for the other terms in (3.24), we find that some positive constant C_1 exists, such that

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

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

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

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

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

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

Lemma 3.11. *Consider a path ω^s and recall that the event $\{I_B(\omega^s) = 1\}$ means that τ_B^s is finite. For any sequence θ_B such that $\theta_B \to 0$ $(B \to \infty)$, the following limit holds:*

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

Proof. Let us first give a sketch of the proof. This proof consist of three steps: *(i)* we bound the length of any path from ∂_e to the origin; *(ii)* using time reversibility arguments we show that the same bound applies to τ_B^0; *(iii)* we show that the path of our interest is shorter (in stochastic sense) than $(\tau_B^0 | I_B(\omega^0) = 1)$.

(i) Let σ_B be the length of the path from any state (B, α) to the origin, for any finite α; and let ω^B be the length of the path from any state (x_1, x_2), such that $x_1 + x_2 = B$. It is clear that

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E} e^{\theta_B \sigma_B} = \lim_{B\to\infty} \frac{1}{B+\alpha} \log \mathbb{E} e^{\theta_B \omega^{B+\alpha}} = \lim_{B\to\infty} \frac{1}{B} \log \mathbb{E} e^{\theta_B \omega^{B}} = 0, \quad (3.25)$$

where the last equality follows from the third statement of [15, Lemma 3].

(ii) Now consider the time-reversed network, see [39, Thm. 1.12]. It is not difficult to check that this is also a tandem queue, but with the first and second queue interchanged. The length of a path in the original system from the origin to level 1 in the second queue, without visits to the origin in the mean time, equals the length of a path from some state $(1, \alpha)$ to the origin in the reversed system, given that it does not visit any state $(1, \cdot)$ in between, hence

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

Combining the last statement with (3.25) we have

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}(e^{\theta_B \tau_B^0} | I_B(\omega^0) = 1) \leq \lim_{B\to\infty} \frac{1}{B} \log \mathbb{E} e^{\theta_B \sigma_B} = 0, \quad (3.26)$$

which is similar to the fourth statement of [15, Lemma 3], only there the exit boundary ∂_e was different.

(iii) We use stochastic coupling to show that

$$\left(\tau_B^s | I_B(\omega^0) = 1\right) \leq_{st} \left(\tau_B^0 | I_B(\omega^0) = 1\right). \quad (3.27)$$

To see this, we consider the ('original') process starting in the origin, and couple it to a similar process starting in state s. Then the above states that the time to overflow for the original process, given that overflow happens before reaching the

origin is stochastically larger than the time to overflow for the coupled process, given that *the original process* reaches overflow before the origin. Notice that the condition implies that also the coupled process reaches overflow before the origin (since the queue lengths cannot be negative). In other words, for any path with $I_B(\omega^0) = 1$, also $I_B(\omega^s) = 1$ must hold, but since the opposite does not hold in general, we have $\{I_B(\omega^0) = 1\} \subset \{I_B(\omega^s) = 1\}$. From this we can conclude

$$(\tau_B^s | I_B(\omega^s) = 1) \leq_{st} \left(\tau_B^s | I_B(\omega^0) = 1\right). \tag{3.28}$$

Using (3.26), (3.27) and (3.28) we can now write that for any state $x \in \bar{D}$,

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}(\mathrm{e}^{\theta_B \tau_B^s} | I_B(\omega^s) = 1) \leq \lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}(\mathrm{e}^{\theta_B \tau_B^0} | I_B(\omega^0) = 1) = 0, \tag{3.29}$$

which completes the proof of the lemma. □

Theorem 3.12. *When $\mu_2 < \mu_1$ and Assumption 3.5 holds, the new measure in (3.21), with function W based on (3.7) and (3.17), is asymptotically efficient.*

Proof. First note that Lemma 3.8 provides an upper bound on the last term of the log-likelihood expression in Lemma 3.9:

$$-\frac{1}{2} \sum_{j=0}^{\tau_B^s - 1} \mathbb{H}(DW(X_j)) \leq 0. \tag{3.30}$$

In order to bound the second term in Lemma 3.9 we will prove a result similar to the third statement of [22, Lemma B.1].

From Proposition 3.4 we know that $\langle DW_2(x), -v_2\rangle = \langle DW_3(x), -v_2\rangle = 0$ and also that $\langle DW_1(x), -v_2\rangle = 2\log(\tilde{\mu}_2(x)/\mu_2)$. Hence, applying (3.22), we have

$$\langle DW(x), -v_2\rangle = 2\log\left(\frac{\tilde{\mu}_2(x)}{\mu_2}\right)\rho_1(x) \geq 2\log\left(\frac{\tilde{\mu}_2(x)}{\mu_2}\right) e^{-(W_1(x)-W_2(x))/\varepsilon}. \tag{3.31}$$

It is clear that $W_1(x) - W_2(x) = \delta$ for any $x \in A_1 \cap \partial_2$, see (3.17). Also, Lemma 3.1 guarantees that both $\tilde{\lambda}(x_1, 0)$ and $\tilde{\lambda}(x_1, \delta/2\gamma)$ tend to λ as x_1 goes to α_1^{-1}. Similarly, $\tilde{\mu}_2(x_1, 0)$ and $\tilde{\mu}_2(x_1, \delta/2\gamma)$ go to μ_2 as x_1 increases up to α_1^{-1}. These relations implies that $W_1(x) - W_2(x)$ decreases to 0 as x moves along the horizontal axis from $(\alpha_1, 0)$ to $(\alpha_1^{-1}, 0)$. This immediately leads to $0 \leq W_1(x) - W_2(x) \leq \delta$ for any $x \in \partial_2$. Now keeping in mind that $\tilde{\mu}_2(x) \geq \lambda$ (see again Lemma 3.1) and hence $\log(\tilde{\mu}_2(x)/\mu_2) \geq -\gamma$, we can write

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

Using the same technique and keeping Lemma 3.1 in mind one can also show that

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

for any $x \in \partial_1$. Using these two inequalities we obtain the following bound for the second term in Lemma 3.9:

$$\sum_{k=1}^{2} \frac{1}{2} \sum_{j=0}^{\tau_B^s - 1} \langle DW(X_j), v_k \rangle I\{X_j = X_{j+1} \in \partial_k\} \leq \gamma e^{-\delta/\varepsilon} \tau_B^s. \tag{3.32}$$

To deal with the first term in Lemma 3.9, we first bound $W(x)$ using (3.18):

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

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

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

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

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

Using the three last inequalities and Lemma 3.10 we can derive an upper bound for the first term in Lemma 3.9,

$$\frac{B}{2} \sum_{j=0}^{\tau_B^s - 1} \langle DW(X_j), X_{j+1} - X_j \rangle \leq \frac{B}{2}(-2\gamma(s) + \eta(B)) + \frac{C}{B\varepsilon} \tau_B^s, \tag{3.33}$$

where $\eta(B)$ is such that $\lim_{B\to\infty} \eta(B) = 0$. Combining (3.30), (3.32) and (3.33) we can rewrite (3.23) in the following way

$$\log(L(\omega^s)) \leq -B\gamma(s) + B\eta(B) + \chi(B)\tau_B^s,$$

where

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

Now for any path ω^s we have:

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

Using the fact that $\lim_{B\to\infty} \chi(B) = 0$ (see Assumption 3.5), Lemma 3.11 and Theorem 3.7 we conclude that:

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

which completes the proof. □

3.3.2 Asymptotic efficiency for $\mu_1 \leq \mu_2$

We would like to define a function based on the total cost function $\gamma(x)$ in (3.11), analogous to the function W in the previous section, see (3.18). Suppose we define the functions $\hat{V}_i(x)$, $i = 1, 2, 3$, in the same way as (3.17). In particular we would find $\hat{V}_1$ on B_1 and B_2 from (3.11) as,

$$\hat{V}_1(x) = 2\gamma(x_1, 0) - \delta, \quad x \in B_1 \tag{3.34}$$

$$\hat{V}_1(x) = -2x_1 \log \frac{\tilde{\lambda}(x)}{\lambda} - 2(1 - x_2) \log \frac{\tilde{\mu}_2(x)}{\mu_2} - \delta, \quad x \in B_2. \tag{3.35}$$

where $(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x))$ is the solution to (3.3) if $x \in B_2$. As a result, $\hat{V}_1(x)$ would not be smooth on the boundary between the sets B_1 and B_2 (see also the discussion above (3.10)), and hence also the resulting mollified function would not be smooth. This would lead to problems when we try to prove the analogue of Lemma 3.10, where we used continuity and smoothness of $W(x)$. Fortunately, the functions $\hat{V}_1(x)$ and $\hat{V}_2(x)$ coincide on B_1 since they are equal on the boundary between B_1 and B_2 and both functions do not depend on their second argument (hence their gradients coincide). Hence, instead of using (3.34)-(3.35) we prefer to work with a function V_1 defined as (3.35) on *both* B_1 *and* B_2. Mollifying this function with functions V_2 and V_3 as in (3.18), will then provide a smooth function V.

To be more specific, we first define function $V_1(x)$, based on the second line of (3.11):

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

where

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) = \begin{cases} \text{solution to (3.3)} & \text{if } x \in B_1 \cup B_2, \\ (\lambda, \mu_2, \mu_1) & \text{if } x \in B_3. \end{cases}$$

The function $V_1(x)$ is an extension of the 'cost' proposed by the solution of (3.3) to the set B_1. In other words, for any $x \in B_1$ we replace the optimal cost (corresponding to the first type of path in (3.10)) by the cost that corresponds to the (suboptimal) path that leads straight from x to $(0, 1)$. We proceed with the definitions of $V_2(x)$ and $V_3(x)$:

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

where $\gamma(x)$ is given in (3.11). In this way, the minimum of V_1 and V_2 is attained by V_1 for $x \in B_2$ (as before), and by V_2 for $x \in B_1$ (rather than by $\hat{V}_1$ as before); see also Figure 3.4, where $\bar{V}(x) = V_1(x) \wedge V_2(x) \wedge V_3(x)$. The mollification procedure now ensures a smooth transition from B_1 to B_2 for the function $V(x)$ defined as in (3.18). Another minor problem is that the function $V_2(x)$ is not smooth around $(\alpha_2, 0)$. Specifically for $x_2 < \delta/\gamma_2$, the gradient of $V_2(x)$ is not continuous around

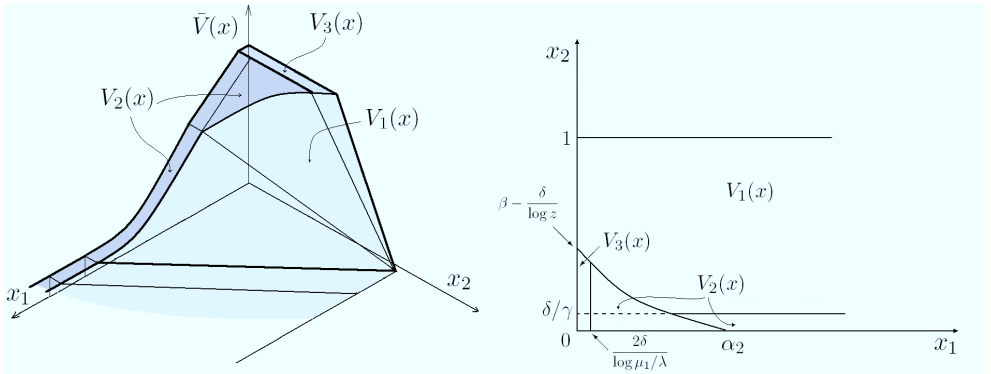

Figure 3.4: The function $\bar{V}(x)$ and the areas on which $\bar{V}(x) = V_i, i = 1, 2, 3$ (case $\mu_1 \leq \mu_2$).

the vertical line $(x_1, \cdot)$ where the first component satisfies $f(x_1, \delta/2\gamma) = 0$ with $f(x)$ as defined in (3.9). Without going into details, we propose to use any suitable mollification procedure to make $V_2(x)$ a smooth function, and from now on we will treat $V_2(x)$ as such. Thus, mollifying the functions $V_i(x)$, $i = 1, 2, 3$, in the same manner as we did in (3.18), we obtain a smooth and continuous function $V(x)$. Based on this function we define the new change of measure $(\bar{\lambda}, \bar{\mu}_1, \bar{\mu}_2)$ as in (3.21). As for the gradient in these equations, we like to notice that even though the functions $W_i(x)$ and $V_i(x)$ for $i = 1, 2, 3$ are different, they have similar gradients. In other words, we can replace $DW_i(x)$ by $DV_i(x)$ in Proposition 3.4 to obtain the shape of the gradients of the functions $V_i(x)$, $i = 1, 2, 3$. We will use these gradients in the following proofs.

Turning to the asymptotic efficiency proof of this new change of measure, we first mention that we can prove analogues of Lemmas 3.8, 3.9, 3.10 (using the smoothness of V) and 3.11 for our current case $\mu_1 \leq \mu_2$, as can be checked easily. Now we can proceed with the main result of this subsection.

Theorem 3.13. *When $\mu_1 \leq \mu_2$ and Assumption 3.5 holds, the new measure in (3.21) with function V based on (3.11), is asymptotically efficient.*

Proof. The proof is similar to that of Theorem 3.12, the main difference being the bound on the second term of the decomposition of $\log L(\omega)$ in Lemma 3.9. For any $x \in \partial_2$, we have $\langle DV_2, -v_2 \rangle = \langle DV_3, -v_2 \rangle = 0$ and $\langle DV_1, -v_2 \rangle = 2 \log(\tilde{\mu}_2(x)/\mu_2)$, so

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

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

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

Using the fact that the optimal cost $\gamma(x)$ for state $x = (0, \beta)$ is both equal to γ (corresonding to the path via state $(\alpha_2, 0)$) and to $(1-\beta)\log(1/z)$ (corresponding to the path along the vertical axis), this can be rewritten as $V_1(0,0) - V_2(0,0) = 2\beta\log(1/z)$. Also, the difference $V_1(x) - V_2(x)$ decreases in x_1 when $x \in \partial_2$, see Lemma 3.2. Combining the above with the fact that $\tilde{\mu}_2(x) \geq \mu_2 z$ (see Lemma 3.2) we obtain the following bound:

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

for any $x \in \partial_2$.

Now let us consider the situation when $x \in \partial_1$. Again $\langle DV_3(x), -v_1\rangle = 0$ holds, and in addition $\langle DV_2(x), -v_1\rangle = -2\log\frac{\mu_1}{\lambda}$ and $\langle DV_1(x), -v_1\rangle = 2\log(\lambda/\tilde{\lambda}(x)) - 2\log(\tilde{\mu}_2(x)/\mu_2) > 0$, where the last inequality is due to the simple observation that $\sqrt{\lambda z/\mu_1} > z$. Using (3.22) we have

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

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

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

Analogously to the previous proof we now conclude that

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

which completes the proof. □

3.4 Numerical results

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

$$\breve{\lambda}(x) = \lambda\sum_{k=1}^{3}\rho_k(x)e^{-\langle DW_k(x), v_0\rangle/2}e^{\mathbb{H}(DW_k(x))/2}, \tag{3.36}$$

$$\breve{\mu}_i(x) = \mu_i\sum_{k=1}^{3}\rho_k(x)e^{-\langle DW_k(x), v_i\rangle/2}e^{\mathbb{H}(DW_k(x))/2}, \quad i = 1, 2, \tag{3.37}$$

where $\rho_k(x)$ is defined by (3.22). This scheme is easier in implementation and its asymptotic efficiency immediately follows from the efficiency of the previous scheme, see Theorem 3.14.

Theorem 3.14. *Under Assumption 3.5,*

(i) when $\mu_2 < \mu_1$, the measure $\mathbb{Q}$ in (3.36)–(3.37), with W defined by (3.18), is asymptotically efficient;

(ii) when $\mu_1 \le \mu_2$, the measure $\mathbb{Q}$ in (3.36)–(3.37), with W replaced by the function V as defined in Section 3.3.2, is asymptotically efficient.

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

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

where the last inequality holds due to the fact that $\mathbb{H}(DW_k(x)) \ge 0$, thanks to Lemma 3.8, and concavity of the logarithm (note that $\sum_{k=1}^{3} \rho_k(x) = 1$). It is obvious that we have similar bounds for transitions v_1 and v_2. Summing these expressions over all steps of sample path $\omega = (X_j, j = 0 \dots \sigma)$ we will get the righthand side of expression (3.23), but without the last term. Since the function $W(x)$ stays the same, we may use the proof of Theorem 3.12 to verify the statement of the current problem.

(ii) The second claim is proved analogously to the first claim. □

All simulations were performed under new measure $\mathbb{Q}$ defined by (3.36)–(3.37) and the joint queue-length process around the boundaries was modified according to (2.3).

We present results of dynamic IS simulations for tandem Jackson networks with parameters $(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$ and $(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$ with the second and the first server being the bottleneck, respectively. The optimal path and new measure $\mathbb{Q}$ are very similar for both cases when the initial state x lies in $A_2 \cup A_3$ (respectively, $B_2 \cup B_3$); only when $x \in B_1$ (for the case $\mu_1 \le \mu_2$), there is an interesting difference with the other case ($\mu_1 > \mu_2$).

Let us start with the case when the second server is the bottleneck, i.e., $\mu_2 < \mu_1$. Here we simulate the network with parameters $(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$ with two different starting states: $(0.6\,B, 0)$ and $(B, 0)$. Both belong to the most interesting subspace A_2, see Figure 3.1, where the new measure is found as the solution to system (3.3).

We performed three different types of IS simulations. At first we simulate the system based on the asymptotic efficient state-dependent scheme obtained in Section 3.3, i.e., scheme (3.4) with protection along the horizontal boundary. In these (and all further) simulations we chose $\varepsilon = 0.005$ and $\delta = -\varepsilon \log \varepsilon$, as motivated in Remark 3.7 in [22]. Moreover, we always performed 10^6 simulation runs, leading to comparable computation times in the order of a few minutes; in fact these were approximately

$(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$, $s = (0.6\,B, 0)$			$(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$, $s = (B, 0)$		
B	p_B^s	RE	B	p_B^s	RE
20	$2.00 \cdot 10^{-5} \pm 2.74 \cdot 10^{-8}$	$7.01 \cdot 10^{-4}$	20	$1.29 \cdot 10^{-2} \pm 1.61 \cdot 10^{-5}$	$6.38 \cdot 10^{-4}$
50	$3.12 \cdot 10^{-12} \pm 4.69 \cdot 10^{-15}$	$7.67 \cdot 10^{-4}$	50	$3.95 \cdot 10^{-5} \pm 5.37 \cdot 10^{-8}$	$7.20 \cdot 10^{-4}$

Table 3.1: Simulation results; original state-dependent scheme

linear in the value of B, as could be expected. The results are presented in Table 3.1; the relative error is defined in (1.3). The obtained results are indeed good, but it is mentioned that the computation time needed is considerable when B grows large, due to the precalculation of the new measure $\mathbb{Q}$ for all states in A_2.

We tried to simplify the scheme to reduce the computation time, in the following way. We divide the set A_2 into a number of triangles D_i of equal area, each having state $(0, 1)$ as one of the corners, the other corners being given by points on the horizontal axis between $(\alpha_2, 0)$ and $(\alpha_2^{-1}, 0)$, at equal distances. In each of these subsets D_i we use a separate, fixed, new measure, based on the solution of system (3.3) where x lies in the middle of the two corners on the horizontal axis. In this way we only need to precalculate a few new measures, rather than dozens for $B = 50$ and hundreds for $B = 100$. In Table 3.2 the simulation results are given, when A_2 is divided into six subsets. Due to this precalculation reduction it became possible to add the extra line ($B = 100$) in Table 3.2 (and Table 3.3). As we can see, the

$(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$, $s = (0.6\,B, 0)$			$(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$, $s = (B, 0)$		
B	p_B^s	RE	B	p_B^s	RE
20	$2.00 \cdot 10^{-5} \pm 3.04 \cdot 10^{-8}$	$7.77 \cdot 10^{-4}$	20	$1.29 \cdot 10^{-2} \pm 2.28 \cdot 10^{-5}$	$8.94 \cdot 10^{-4}$
50	$3.12 \cdot 10^{-12} \pm 6.11 \cdot 10^{-15}$	$9.98 \cdot 10^{-4}$	50	$3.94 \cdot 10^{-5} \pm 9.36 \cdot 10^{-8}$	$1.20 \cdot 10^{-3}$
100	$1.82 \cdot 10^{-23} \pm 5.06 \cdot 10^{-26}$	$1.41 \cdot 10^{-3}$	100	$3.66 \cdot 10^{-9} \pm 1.13 \cdot 10^{-11}$	$1.57 \cdot 10^{-3}$

Table 3.2: Simulation results; simplified scheme with six domains

variance of the estimator increases as a result of the simplification of the original scheme (although the effect is relatively modest). This may be explained by the following. Under the simplified scheme a path that starts to, say, the left of the middle point of D_i will at some point hit the boundary between D_i and D_{i-1}, after which it follows this boundary. Hence, when B grows large and therefore the sizes of (the unscaled counterparts of) the D_i also increase, the sample path will move back and forth between two different changes of measure for a substantial period of time.

We also simulated the system using an even simpler scheme, getting rid of the set A_2 altogether, expanding A_1 and A_3 so that they meet at the line $x_1 + x_2 = 1$. That is, we simply used the measure $\mathbb{Q} = (\mu_2, \mu_1, \lambda)$ when the total population of the system is less than B and used no change of measure otherwise. Clearly, this method provides worse results, as can be concluded from Table 3.3. Looking at the relative error or at the confidence intervals, it is clear that this method is inferior to the ones presented above.

$(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$, $s = (0.6\,B, 0)$			$(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$, $s = (B, 0)$		
B	p_B^s	RE	B	p_B^s	RE
20	$2.00 \cdot 10^{-5} \pm 7.30 \cdot 10^{-8}$	$1.87 \cdot 10^{-3}$	20	$1.29 \cdot 10^{-2} \pm 1.78 \cdot 10^{-4}$	$7.02 \cdot 10^{-3}$
50	$3.12 \cdot 10^{-12} \pm 1.67 \cdot 10^{-14}$	$2.73 \cdot 10^{-3}$	50	$3.67 \cdot 10^{-5} \pm 4.34 \cdot 10^{-6}$	$6.00 \cdot 10^{-2}$
100	$1.81 \cdot 10^{-23} \pm 1.59 \cdot 10^{-25}$	$4.48 \cdot 10^{-3}$	100	$5.62 \cdot 10^{-9} \pm 7.09 \cdot 10^{-9}$	$6.40 \cdot 10^{-1}$

Table 3.3: Simulation results; simplified scheme with two domains

$(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$, $s = (0, 0)$			$(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$, $s = (0.6\,B, 0)$		
B	p_B^s	RE	B	p_B^s	RE
20	$2.48 \cdot 10^{-2} \pm 1.45 \cdot 10^{-3}$	$3.10 \cdot 10^{-2}$	20	$1.68 \cdot 10^{-1} \pm 1.95 \cdot 10^{-2}$	$5.90 \cdot 10^{-2}$
50	$3.61 \cdot 10^{-5} \pm 4.67 \cdot 10^{-6}$	$6.73 \cdot 10^{-2}$	50	$2.05 \cdot 10^{-3} \pm 4.12 \cdot 10^{-4}$	$1.02 \cdot 10^{-1}$

Table 3.4: Simulation results; original state-dependent scheme

Also, we tried a change of measure that replaces the original parameters on A_2 by a simple linear interpolation between the values on A_1 and A_3 (e.g., when $\mu_2 < \mu_1$ and $x \in A_2$ we let $\breve{\lambda}(x) = \alpha(x)\mu_2 + (1 - \alpha(x))\lambda$, where $\alpha(x) \in [0, 1]$ depends on the location of x relative to A_1 and A_3). Unfortunately, this approach gives only slightly better results than those in Table 3.3, but much worse than those in Table 3.2.

Finally, let us proceed to the case when the first server is the bottleneck, i.e., $\mu_1 \leq \mu_2$. Here we simulate the system with parameters $(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$ using the asymptotically efficient state-dependent IS scheme defined by (3.36)-(3.37), again with $\varepsilon = 0.005$ and $\delta = -\varepsilon \log \varepsilon$; see Table 3.4. We consider two possible starting states. Firstly, the origin, to enable the comparison with our earlier studies in Chapter 2, see Table 2.2. There we used a state-independent IS scheme to simulate this network, and we did not obtain asymptotic efficiency. Secondly, we let the state that belongs to the interesting subspace B_2, namely $(0.6B, 0)$, to be the starting state, see Figure 3.2. In the case of the origin being the starting state our current IS scheme provides a considerable variance reduction compared to the state-independent scheme designed in Chapter 2. Namely, we reduce the relative error by 40% for the case when $B = 20$ and by more than 80% when $B = 50$. The behavior of the scheme for the second starting state is also quite good.

3.5 Conclusions

In this chapter we focused on the event that, starting from an arbitrary state, the second queue in a tandem Jackson network reaches overflow before the system becomes empty. We proposed a particular change of measure, motivated by large-deviations arguments, and proved asymptotic efficiency of a subtly modified version (that differs close to the axes, and thus nicely controls the likelihood).

The next challenge is to transform the methods presented in this chapter into efficient simulation algorithm. 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. We will provide the final answer in Chapter 5.

Finally, we discuss our methodology in comparison with Dupuis *et al.* [22], who consider the problem from a control-theoretical point of view. First of all, in [22] the large deviations decay rate is found by solving a calculus of variation problem, in which the cost function to be optimized is the relative entropy between the old and the new probability measures. In our current work we essentially solve the same problem, but first heuristically from an intuitively appealing point of view in Section 3.1, and then more formally in Section 3.2. The function $\gamma(x)$ in (3.7) resp. (3.11) is simply the value function of the optimal control problem.

A reason why we put some emphasis on the heuristics, is that ideas of this type are useful for more general systems, such as d-node tandem Jackson or the slowdown network, see e.g. [24] and the next chapters of the thesis. The heuristics reveal how the system behaves conditional on the rare event under consideration, which is useful when generating a first guess for an efficient change of measure. Usually this first guess then needs to be adapted to tackle complications at the boundaries, as we describe after this.

When it comes to importance sampling, it is shown in [23] that this is also tightly connected to an optimal control problem. Even though this problem converges to the large deviations optimal control problem in some sense, the latter does not always correspond to an asymptotically efficient IS scheme. In other words, understanding the large deviations behavior (the most probable path) is in general not sufficient for an asymptotically efficient IS scheme. To overcome this problem, the large deviations solution is modified in [22], using a suitable game-theoretic representation, such that the result is a *classical subsolution* of a corresponding *Isaacs* equation. The resulting IS scheme is then proven to be asymptotically efficient by a standard (but cumbersome) verification argument.

In our work we take a different angle to the problem that the IS scheme as suggested by large deviations (as given in (3.4) and (3.8)) may not be asymptotically optimal. In Chapter 2 we identified the problem to be the visits of the process to the horizontal axis, which may lead to large likelihoods. We solve this by constructing a function W that leads to a modified change of measure, technically in the same way as in [22], and we credit its authors for this elegant mollification procedure, see (3.17) and (3.18). However, instead of detailing how the function W was found in [22] (namely as a subsolution to the corresponding *Isaacs* equation), we consider it as an appealing way to adapt the large deviations change of measure to one that is similar on most of the state space, but harmless around the horizontal boundary and the origin. In this way no knowledge of control theory, game theory or partial differential equations and Hamiltonians is required to understand both the way in which the IS change of measure is derived from the function W, see (3.19), and the subsequent

proof of asymptotic efficiency (which replaces the verification argument in [22]). Importantly, we stress that [22] deals with tandem networks of arbitrary dimension (where we just covered the two-queue case); in addition it provides structural insights, that are useful in other contexts too.

Chapter 4

State-dependent importance sampling for slowdown network

The primary goal of this chapter is to analyze the probability of overflow in the downstream ('protected') queue of the slowdown network, before the system idles, starting off from any given state, just as was done for the tandem network in Chapter 3. The special case of this problem in which the origin is the starting state was already studied in Chapter 2. There a state-independent IS scheme was proposed for estimating the overflow in the second queue. For a limited set of parameter values its asymptotic efficiency was concluded, but just on the basis of empirical evidence. In this chapter we generalize these findings, such that all initial states are allowed.

We design a *state-dependent* IS scheme, using similar techniques as for the tandem Jackson network in Chapter 3. 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 slowdown threshold. As a consequence, 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.

The most important results of this chapter are the state-dependent IS scheme itself (which requires recalculation of the new measure after every transition) and the proof of asymptotic efficiency of the designed IS scheme, which is similar to the one in Chapter 3 (but more involved). We postpone numerical studies to Chapter 6.

4.1 Optimal path and related change of measure

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 4.1.1-4.1.2 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 slowdown 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^s 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 4.1.3 we treat the case $\mu_2 < \mu_1^+ < \mu_1$ in which the second server is always the bottleneck, Subsection 4.1.4 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 4.1.5 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 4.2 (denoted by bars), although it is closely related.

4.1.1 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 [2]. 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 Chapter 2 of this thesis. Here we present a method similar to the one in Chapter 2 to find the optimal path starting from *any* state $s \in \bar{D} \cup \bar{D}^+$. 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 will restrict ourselves to a (small) subset of all feasible paths. This restriction is motivated by Theorem 4.1, which is a generalization of Theorem 2.4.

Property 4.1. 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) if $\mu_1^+ < \mu_1 \leq \mu_2$, then each path does not have more than two subpaths in each subset; otherwise each path does not have more than one subpath per subset.

See also Property 3.1. With every path that satisfies Property 4.1 we associate a 'cost', the main idea being that the minimal cost of the path to overflow in the

second queue starting from state s is the decay rate of the probability of interest (see Section 4.1.2). The method is similar to the one used in previous chapters and is based on the family of cost functions (2.15).

Example 4.2. As a leading example, we 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 slowdown 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 slowdown threshold our path has a constant slope $-\alpha$, with

$$\alpha = \frac{\tilde{\mu}_1 - \tilde{\mu}_2}{\tilde{\mu}_1 - \tilde{\lambda}}, \tag{4.1}$$

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

$$\alpha^+ = \frac{\tilde{\mu}_1^+ - \tilde{\mu}_2^+}{\tilde{\mu}_1^+ - \tilde{\lambda}^+}. \tag{4.2}$$

Below the slowdown threshold, the cost of this path is, per unit time,

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

see also (2.18). 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 slowdown 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

$$(\theta - x_2)\frac{\mathbb{I}(\tilde{\lambda},\, \tilde{\mu}_1,\, \tilde{\mu}_2)}{\tilde{\mu}_1 - \tilde{\mu}_2} + (y_2 - \theta)\frac{\mathbb{I}(\tilde{\lambda}^+,\, \tilde{\mu}_1^+,\, \tilde{\mu}_2^+)}{\tilde{\mu}_1^+ - \tilde{\mu}_2^+}, \tag{4.3}$$

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

$$\kappa(x) := x_1 - \frac{\tilde{\mu}_1 - \tilde{\lambda}}{\tilde{\mu}_1 - \tilde{\mu}_2}(\theta - x_2) = y_1 + \frac{\tilde{\mu}_1^+ - \tilde{\lambda}^+}{\tilde{\mu}_1^+ - \tilde{\mu}_2^+}(y_2 - \theta), \tag{4.4}$$

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

One way to solve the minimization problem (4.3) 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 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

$$\begin{cases} \lambda^{(\text{syst})} = \mu_1^{(\text{syst})} + \frac{\kappa(x)-x_1}{\theta-x_2}(\mu_1^{(\text{syst})} - \mu_2^{(\text{syst})}) \\ \lambda^{(\text{syst})} + \mu_1^{(\text{syst})} + \mu_2^{(\text{syst})} = \lambda + \mu_1 + \mu_2 \\ \lambda^{(\text{syst})}\mu_1^{(\text{syst})}\mu_2^{(\text{syst})} = \lambda\mu_1\mu_2 \\ \lambda^{(\text{syst})} \leq \mu_1^{(\text{syst})} \text{ and } \mu_1^{(\text{syst})} > \tilde{\mu}_2^{(\text{syst})} \\ \lambda^{(\text{syst})}, \mu_1^{(\text{syst})}, \mu_2^{(\text{syst})} > 0 \end{cases} \tag{4.5}$$

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

$$\begin{cases} \lambda^{+(\text{syst})} = \mu_1^{+(\text{syst})} - \frac{\kappa(x)}{1-\theta}(\mu_1^{+(\text{syst})} - \mu_2^{+(\text{syst})}) \\ \lambda^{+(\text{syst})} + \mu_1^{+(\text{syst})} + \mu_2^{+(\text{syst})} = \lambda + \mu_1^+ + \mu_2 \\ \lambda^{+(\text{syst})}\mu_1^{+(\text{syst})}\mu_2^{+(\text{syst})} = \lambda\mu_1^+\mu_2 \\ \lambda^{+(\text{syst})} \leq \mu_1^{+(\text{syst})} \text{ and } \mu_1^{+(\text{syst})} > \mu_2^{+(\text{syst})} \\ \lambda^{+(\text{syst})}, \mu_1^{+(\text{syst})}, \mu_2^{+(\text{syst})} > 0 \end{cases} \tag{4.6}$$

Notice also that if $(\lambda^{(\text{syst})}(x), \mu_1^{(\text{syst})}(x), \mu_2^{(\text{syst})}(x))$ is the solution to (4.5) for some 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, $(\lambda^{+(\text{syst})}(x), \mu_1^{+(\text{syst})}(x), \mu_2^{+(\text{syst})}(x))$ which is the solution of (4.6) 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.,

$$\gamma_1(x_1, x_2) := -(x_1 - \kappa(x)) \log \frac{\lambda^{(\text{syst})}(x_1, x_2)}{\lambda} - (\theta - x_2) \log \frac{\mu_2^{(\text{syst})}(x_1, x_2)}{\mu_2}, \tag{4.7}$$

$$\gamma_2(\kappa(x), \theta) := -\kappa(x) \log \frac{\lambda^{+(\text{syst})}(\kappa(x), \theta)}{\lambda} - (1 - \theta) \log \frac{\mu_2^{+(\text{syst})}(\kappa(x), \theta)}{\mu_2}, \tag{4.8}$$

where $\lambda^{(\text{syst})}$, $\mu_2^{(\text{syst})}$, $\lambda^{+(\text{syst})}$ and $\mu_2^{+(\text{syst})}$ are given by the joint solution to (4.5) and (4.6), $\kappa(x)$ is given in (4.4). Then clearly the total cost of the path $(s_1, s_2) \to (\kappa(s), \theta) \to (0, 1)$ can be expressed as $\gamma_1(s_1, s_2) + \gamma_2(\kappa(s), \theta)$. ♢

4.1.2 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 $(\tilde{\lambda}^+(x), \tilde{\mu}_1^+(x), \tilde{\mu}_2^+(x))$.

Recall that $\gamma(s)$ is the overall minimal cost over all paths $s \to \partial_e$. The following theorem states that this is in fact the exponential decay rate of the probability p_B^s 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 4.1. *The exponential decay rate of p_B^s is equal to the minimal cost of overflow $\gamma(s)$, i.e.,*

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

We now present the value of $\gamma(x)$ for any state 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).

4.1.3 Importance sampling 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 (2.15), turns out to be given by

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) = \begin{cases} (\mu_2, \mu_1, \lambda), & \text{if } x \in A_1, \\ (\lambda^{(\text{syst})}(x), \mu_1^{(\text{syst})}(x), \mu_2^{(\text{syst})}(x)), & \text{if } x \in A_2, \\ (\lambda, \mu_1, \mu_2), & \text{if } x \in A_3, \end{cases} \tag{4.9}$$

and

$$(\tilde{\lambda}^+(x), \tilde{\mu}_1^+(x), \tilde{\mu}_2^+(x)) = \begin{cases} (\mu_2, \mu_1^+, \lambda), & \text{if } x \in A_1^+, \\ (\lambda^{+(\text{syst})}(x), \mu_1^{+(\text{syst})}(x), \mu_2^{+(\text{syst})}(x)), & \text{if } x \in A_2^+, \\ (\lambda, \mu_1^+, \mu_2), & \text{if } x \in A_3^+. \end{cases} \tag{4.10}$$

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 4.1, where $\alpha_1 := (\mu_1 - \mu_2)/(\mu_1 - \lambda)$ and $\alpha_1^+ := (\mu_1^+ - \mu_2)/(\mu_1^+ - \lambda)$. We chose not to give the precise definitions of the sets A_i and A_i^+ here, since they do not add much to the understanding, see also (3.5). For some starting states s, Figure 4.1 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 [59] for the problem of reaching a large *total* network population. Measures similar to the ones in the other subsets were introduced in Chapter 3. Also, we point out that the new measure is continuous in the state x, as can be verified by solving system (4.5) 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 (4.9), respectively. A similar principle holds above the slowdown threshold as well.

The residual cost $\gamma(x)$ of the optimal path at state x can be expressed as:

$$\gamma(x) = \begin{cases} (1 - x_1 - x_2)\gamma, & \text{if } x \in A_1, \\ \gamma_1(x_1, x_2) + \gamma_2(\kappa(x), \theta), & \text{if } x \in A_2, \\ 0, & \text{if } x \in A_3. \end{cases} \tag{4.11}$$

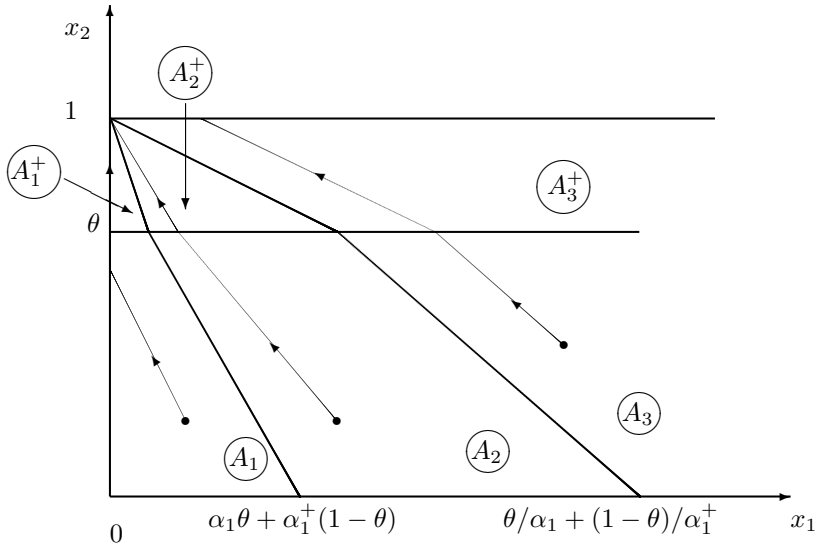

Figure 4.1: Partition of $\bar{D} \cup \bar{D}^+$ and some optimal paths to overflow when $\mu_2 < \mu_1^+ < \mu_1$.

We recall that $\gamma = -\log \frac{\lambda}{\mu_2}$ is the decay rate of the path $(0,0) \to (0,1)$, γ_1 and γ_2 are as in (4.7) and (4.8), and $\kappa(x)$ is given in (4.4). Importantly, we treat only paths with starting state below the slowdown threshold; starting states above the threshold are substantially easier to deal with.

4.1.4 Importance sampling 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 (2.15), is given by

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) = \begin{cases} (\mu_2, \mu_1, \lambda), & \text{if } x \in B_1, \\ (\lambda^{(\mathrm{syst})}(x), \mu_1^{(\mathrm{syst})}(x), \mu_2^{(\mathrm{syst})}(x)), & \text{if } x \in B_2, \\ (\lambda, \mu_1, \mu_2), & \text{if } x \in B_3. \end{cases} \tag{4.12}$$

and

$$(\tilde{\lambda}^+(x), \tilde{\mu}_1^+(x), \tilde{\mu}_2^+(x)) = \begin{cases} (\lambda^{+(\mathrm{syst})}(x), \mu_1^{+(\mathrm{syst})}(x), \mu_2^{+(\mathrm{syst})}(x)), & \text{if } x \in B_2^+, \\ (\lambda, \mu_2, \mu_1^+), & \text{if } x \in B_3^+. \end{cases} \tag{4.13}$$

Here, the five subsets B_1, B_2, B_3, B_2^+ and B_3^+ are shown in Figure 4.2, which is comparable to Figure 4.1. 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.

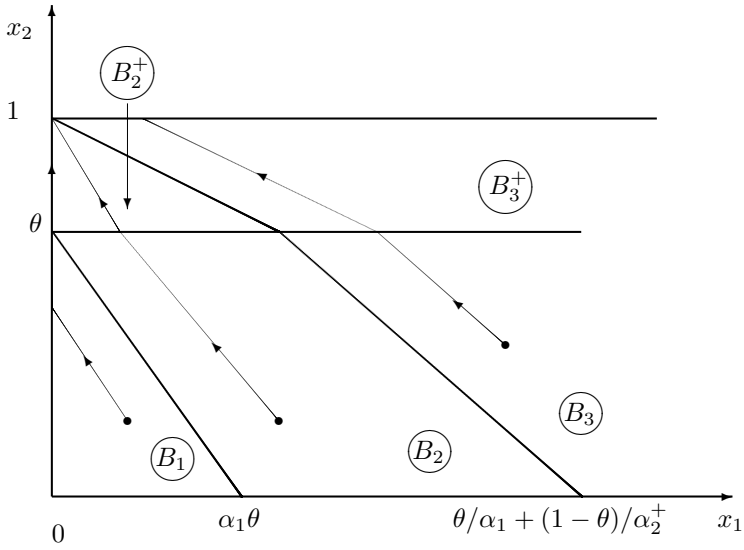

Figure 4.2: Partition of $\bar{D} \cup \bar{D}^+$ and some optimal paths to overflow when $\mu_1^+ \leq \mu_2 < \mu_1$

The residual cost $\gamma(x)$ is now given by

$$\gamma(x) = \begin{cases} \theta(1-x_1-x_2)\gamma + (1-\theta)\log(1/z^+), & \text{if } x \in B_1, \\ \gamma_1(x_1,x_2) + \gamma_2(\kappa(x),\theta), & \text{if } x \in B_2, \\ (1-\theta)\log(\mu_2/\mu_1^+), & \text{if } x \in B_3. \end{cases} \tag{4.14}$$

Here z^+ is the unique solution in $(0,1)$ of Equation (2.36). In fact, $(1-\theta)\log(1/z^+)$ is the cost of the vertical path $(0,\theta) \to (0,1)$ *in the interior*, satisfying $\tilde{\lambda} = \tilde{\mu}_1^+$.

4.1.5 Importance sampling 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 (2.15), is given by

$$(\tilde{\lambda}(x), \tilde{\mu}_1(x), \tilde{\mu}_2(x)) = \begin{cases} (\mu_1, \lambda, \mu_2), & \text{if } x \in C_1, \\ (\lambda^{(\text{syst})}(x), \mu_1^{(\text{syst})}(x), \mu_2^{(\text{syst})}(x)), & \text{if } x \in C_2, \\ (\lambda, \mu_2, \mu_1), & \text{if } x \in C_3. \end{cases} \tag{4.15}$$

and

$$(\tilde{\lambda}^+(x), \tilde{\mu}_1^+(x), \tilde{\mu}_2^+(x)) = \begin{cases} (\lambda^{+(\text{syst})}(x), \mu_1^{+(\text{syst})}(x), \mu_2^{+(\text{syst})}(x)), & \text{if } x \in C_2^+, \\ (\lambda, \mu_2, \mu_1^+), & \text{if } x \in C_3^+. \end{cases} \tag{4.16}$$

The sets C_1, C_2, C_3, C_2^+ and C_3^+ are shown in Figure 4.3, where $\alpha_2 := (\mu_2 - \mu_1)/(\mu_2 -$

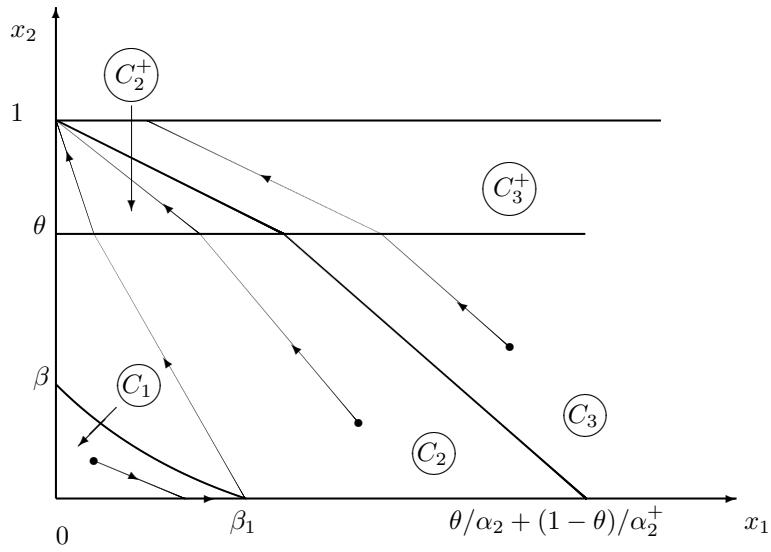

Figure 4.3: Partition of $\bar{D} \cup \bar{D}^+$ and some optimal paths to overflow when $\mu_1^+ < \mu_1 \leq \mu_2$.

λ); 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 Equation (2.37). 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

$$f(0, \beta) = \theta\gamma + (1-\theta)\log(1/q) - (\theta - \beta)\log(1/z) - (1-\theta)\log(1/z^+) = 0, \quad (4.17)$$

where z and z^+ follows from (3.12) and (2.36) respectively.

The main result of this subsection is the residual cost of the path to overflow at state x:

$$\gamma(x) = \begin{cases} \theta\gamma + (1-\theta)\log(1/q) - x_1 \log(\mu_1/\lambda), & \text{if } x \in C_1, \\ \gamma_1(x_1, x_2) + \gamma_2(\kappa(x), \theta), & \text{if } x \in C_2, \\ (\theta - x_2)\log(\mu_2/\mu_1) + (1-\theta)\log(\mu_2/\mu_1^+), & \text{if } x \in C_3. \end{cases} \quad (4.18)$$

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.

4.1.6 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 4.1.3–4.1.5 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 4.2. *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 (4.9) and (4.10), or by (4.12) and (4.13), or by (4.15) and (4.16) 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 (2.36) and z by (3.12).

4.2 Asymptotic efficiency

For the special case in which the starting state is the origin, it is known from Chapter 2 that the new measures provided in Section 4.1 are not always asymptotically efficient. 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 we used in Chapter 3. 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 4.2.1 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 4.2.2 and 4.2.3 we treat the other cases.

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

In this subsection we present a modification of the scheme constructed in Subsection 4.1.3 and prove its asymptotic efficiency. 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 Chapter 3. 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 $W_i(x), i = 1, 2, 3$:

$$\begin{aligned} W_1(x) &:= 2\gamma_2(x) I_{\{x \in \bar{D}^+\}} + 2\left(\gamma_1(x) + \gamma_2(\kappa(x), \theta)\right) I_{\{x \in \bar{D}\}} - \delta, \\ W_2(x) &:= 2\gamma(x_1, \delta/2\gamma) - \delta, \\ W_3(x) &:= 2\gamma(0) - 3\delta, \end{aligned} \tag{4.19}$$

where $\gamma(x)$, $\gamma_1(x)$ and $\gamma_2(\kappa(x), \theta)$ are as in (4.11), see also (4.7) and (4.8); we recall that γ equals $-\log(\lambda/\mu_2)$. In the next step we introduce the minimum of these auxiliary functions:

$$\bar{W}(x) := W_1(x) \wedge W_2(x) \wedge W_3(x). \tag{4.20}$$

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, see (3.18):

$$W(x) := -\epsilon \log \sum_{i=1}^{3} e^{-W_i(x)/\epsilon}. \tag{4.21}$$

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 some conditions for their asymptotic behavior, see Assumption 3.5.

The following expression for the gradient of $W(x)$ is immediate from (4.21), and will play an important role in the representation of the state-dependent, asymptotically efficient new measure:

$$DW(x) = \sum_{k=1}^{3} \rho_k(x) DW_k(x), \text{ where } \rho_k(x) := \frac{e^{-W_k(x)/\epsilon}}{\sum_{i=1}^{3} e^{-W_i(x)/\epsilon}}. \tag{4.22}$$

Also, we have the following helpful property.

Proposition 4.3. *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 $DW_1(x) = (-2\gamma, -2\gamma)$ 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 tandem Jackson network (see Proposition 3.4), 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 (3.19) and (3.21):

$$\begin{aligned}
\bar{\lambda}(x) &:= \lambda e^{-\langle DW(x),v_0\rangle/2} e^{\mathbb{H}(DW(x))/2}, && \text{if } x \in \bar{D},\\
\bar{\mu}_i(x) &:= \mu_i e^{-\langle DW(x),v_i\rangle/2} e^{\mathbb{H}(DW(x))/2}, \quad i=1,2, && \text{if } x \in \bar{D},\\
\bar{\lambda}^+(x) &:= \frac{\lambda}{\lambda+\mu_1^+ + \mu_2} e^{-\langle DW(x),v_0\rangle/2} e^{\mathbb{H}^+(DW(x))/2}, && \text{if } x \in \bar{D}^+,\\
\bar{\mu}_1^+(x) &:= \frac{\mu_1^+}{\lambda+\mu_1^+ + \mu_2} e^{-\langle DW(x),v_1\rangle/2} e^{\mathbb{H}^+(DW(x))/2}, && \text{if } x \in \bar{D}^+,\\
\bar{\mu}_2^+(x) &:= \frac{\mu_2}{\lambda+\mu_1^+ + \mu_2} e^{-\langle DW(x),v_2\rangle/2} e^{\mathbb{H}^+(DW(x))/2}, && \text{if } x \in \bar{D}^+.
\end{aligned} \tag{4.23}$$

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 (4.23) are known as *Hamiltonians*, which we use to enable the comparison with [15, 22]; 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 (4.23), we are ready to prove that it is asymptotically efficient. We start with some lemmas that are similar to the ones in Chapter 3.

Lemma 4.4. *The likelihood $L(\omega)$ of a path $\omega = (X_j, j = 0, \dots, \sigma)$ under the new measure (4.23) satisfies*

$$\begin{aligned}
\log L(\omega) &= \frac{B}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), X_{j+1}-X_j\rangle\\
&+ \sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\sigma-1}\langle DW(X_j), v_k\rangle I\{X_j = X_{j+1} \in \partial_k\}\\
&- \frac{1}{2}\sum_{j=0}^{\sigma-1}\left(\mathbb{H}(DW(X_j))I_{\{X_j\in D\}} + \mathbb{H}^+(DW(X_j))I_{\{X_j\in D^+\}}\right).
\end{aligned} \tag{4.24}$$

Proof. The proof is analogous to the proof of Lemma 3.9 or Lemma 1 in [15]. □

Lemma 4.5. *For any path $\omega = (X_j, j = 0, ..., \sigma)$ under the new measure (4.23), the first term in (4.24) satisfies*

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

for sufficiently large $B\epsilon$, where C and C^+ are some positive constants and σ^+ is the number of slowdown 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 slowdown 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 3.10 for the following bound that holds when x and $x+y$ are not separated by the slowdown 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 4.6. *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 (4.22) and keeping in mind the concavity of $\mathbb{H}^+(x)$ (thanks to Proposition 3.2 in [22]) 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 3.8. □

Lemma 4.7. *Consider the slowdown network and recall the definition of τ_B^s in (2.4). For any sequence υ_B such that $\lim_{B\to\infty}\upsilon_B = 0$ the following limit holds:*

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

Proof. We define a new random variable τ which represents the same random period of time as τ_B^s, but for the case when $\theta = 0$, i.e., for the tandem Jackson network with parameters $(\lambda, \mu_1^+, \mu_2)$. It is clear that $\tau_B^s \leq^{st} \tau$. From Lemma 3.11 we know that for the tandem Jackson network

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

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

Theorem 4.8. *When $\mu_2 < \mu_1^+ < \mu_1$ and Assumption 3.5 holds, the new measure in (4.23), where the function W is based on (4.11), is asymptotically efficient.*

Proof. We will roughly follow the proof of Theorem 3.12, finding upper bounds on each of the three terms in Lemma 4.4.

To deal with the first term, we first bound $W(s)$ from below. Upon combining the fact that $W_2(s) \geq W_1(s) - \delta$ for any $x \in \bar{D} \cup \bar{D}^+$ (this is shown in the same manner as in Theorem 3.12; use (4.11)), with the monotonicity of $\gamma(s)$ in both s_1 and s_2, and using definition (4.21), it is found that

$$\begin{aligned} W(s) &\geq -\epsilon\log(e^{-W_1(s)/\epsilon} + e^{(-W_1(s)+\delta)/\epsilon} + e^{-W_3(s)/\epsilon}) \\ &\geq -\epsilon\log(3e^{(-2\gamma(s)+3\delta)/\epsilon}) = 2\gamma(s) - \epsilon\log(3) - 3\delta. \end{aligned} \tag{4.25}$$

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

$$W(X_{\tau_B^s}) \leq -\delta. \tag{4.26}$$

Combining the inequalities (4.25)-(4.26) with Lemma 4.5 (take $\sigma = \tau_B^s$), we now derive the following upper bound on the first term in Lemma 4.4:

$$\frac{B}{2}\sum_{j=0}^{\tau_B^s-1}\langle DW(X_j), X_{j+1}-X_j\rangle \le \frac{B}{2}(-2\gamma(s)+\eta(B)) + \frac{C}{B\epsilon}\tau_B^s + C^+\tau_B^{s,+}, \quad (4.27)$$

where C and C^+ are some positive constants, $\eta(B)$ is such that $\lim_{B\to\infty}\eta(B)=0$ (use Assumption 3.5), and $\tau_B^{s,+}$ is the number of slowdown threshold crossings up to time τ_B^s.

Now let us bound the second term in Lemma 4.4. 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 (4.22) we arrive at

$$\langle DW(x), -v_2\rangle = 2\log\left(\frac{\tilde{\mu}_2}{\mu_2}\right)\rho_1(x) \ge -2\gamma\rho_1(x) \ge -2\gamma e^{-(W_1(x)-W_2(x))/\epsilon}, \quad (4.28)$$

where the first inequality comes from the fact that $\tilde{\mu}_2 \ge \lambda$ (see Proposition 4.2). 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 (4.19). The second statement of Proposition 4.2 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 $0 \le W_1(x) - W_2(x) \le \delta$, which implies:

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

Using the same technique and keeping Proposition 4.2 in mind, one can also show that

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

for any $x \in \partial_1$. Using these inequalities we can bound the second term in Lemma 4.4 from above:

$$\sum_{k=1}^{2}\frac{1}{2}\sum_{j=0}^{\tau_B^s-1}\langle DW(X_j), v_k\rangle I\{X_j = X_{j+1} \in \partial_k\} \le \gamma e^{-\delta/\epsilon}\tau_B^s. \quad (4.29)$$

Finally note that Lemma 4.6 provides a straightforward bound on the last term of the log-likelihood expression in Lemma 4.4:

$$\mathbb{H}(DW(X_j))I_{\{X_j\in D\}} + \mathbb{H}^+(DW(X_j))I_{\{X_j\in D^+\}} \ge 0. \quad (4.30)$$

Upon combining (4.27), (4.29) and (4.30), we bound (4.24) in the following way:

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

where

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

Now for any path ω^s we have

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

Using that $\lim_{B\to\infty}\tau_B^{s,+}/B=0$ a.s. when $I_B(\omega^s)=1$, and that $\lim_{B\to\infty}\chi(B)=0$ (see Assumption 3.5), and invoking Lemma 4.7 and Theorem 4.1, we conclude that

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

In view of criterion (1.10), this completes the proof. □

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

Remarkably, we can use the same function $W(x)$ for this case as in the previous subsection, see (4.21); 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 (4.23).

One difference with the previous case is that Lemma 4.6 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 3.5 we have that $1/\epsilon\to\infty$ and $\delta/\epsilon\to\infty$ as $B\to\infty$.

Lemma 4.9. *For any $x\in D$ we have $\mathbb{H}(DW(x))\geq 0$, and for any $x\in D^+$ we have*

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

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

Proof. Using the same technique as in the proof of Lemma 4.6, one can prove that the first statement holds, while for the second statement we find

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

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

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

so that the claim follows. □

Theorem 4.10. *When $\mu_1^+ \leq \mu_2 < \mu_1$ and Assumption 3.5 holds, the new measure in (4.23), where the function W is based on (4.14) and (4.8), is asymptotically efficient.*

Proof. In order to prove this theorem we again bound all three terms of the log-likelihood ratio, see Lemma 4.4. For the first two terms we find exactly the same as in (4.27) and (4.29). As for the third term, this is now bounded in Lemma 4.9. Thus, we find again

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

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 4.8 which leads us to the result. □

4.2.3 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 (4.21), to describe the IS scheme as in (4.23), 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 4.3. In the two previous subsections this was only valid on $\{x : x_2 \leq \delta/2\gamma\}$.

Theorem 4.11. *When $\mu_1^+ < \mu_1 \leq \mu_2$ and Assumption 3.5 holds, the new measure in (4.23), where the function W is based on (4.18) and (4.8), is asymptotically efficient.*

Proof. Not surprisingly, the proof of this theorem is almost the same as that of Theorem 4.10. Even Lemma 4.9 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 4.4. As in Theorems 4.8 and Theorem 4.10 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 3.13 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 (4.17). 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 3.13:

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

The rest of the proof can be done by mimicking the arguments used in Thm. 4.8 or Thm. 4.10. □

4.3 Conclusions

In this chapter we focused on constructing IS schemes for estimating the probability of a specific rare event: overflow in the second queue of the slowdown 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 tandem Jackson network started in Chapter 3. On the other hand, this chapter can be viewed as the generalization of research from Chapter 2 and [21]. We rigorized and further studied the empirical findings of Chapter 2. Also, in Chapter 2 as well as in [21] the only possible starting state is the origin. In our schemes one may pick an arbitrary starting state s. Here we developed IS schemes for all three possible cases (second queue bottleneck, 'shifting bottleneck', first queue bottleneck), unlike [21] 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 $\mu_1^+ \leq \mu_2 < \mu_1$). 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.

The IS scheme designed in this chapter has the same drawback as the scheme from Chapter 3. Namely, a direct implementation of this *asymptotically efficient* algorithm will have a high computational complexity, since we need to re-calculate a new measure (i.e., system of two cubic equations) after every transition. A simpler IS scheme as well as detailed simulation study and some practical guidelines, will be presented in Chapter 6. But first we will present such a simple IS scheme for the tandem Jackson network in Chapter 5.

4.4 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 4.1.

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':

$$L_i(y) := \sup_{\vartheta} \left(\langle \vartheta, y \rangle - g_i(\vartheta) \right), \ i = 1, \ldots, 4, \tag{4.31}$$

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 [66]. 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:

$$\ell(\phi(t), \dot{\phi}(t)) := \begin{cases}
L_1(\dot{\phi}(t)), & \text{if } \phi(t) \in D, \\
L_2(\dot{\phi}(t)), & \text{if } \phi(t) \in D^+ \cup \partial_e, \\
[L_1 \oplus L_3\ ](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_1 \setminus \partial_1^+, \\
[L_2 \oplus L_3\ ](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_1^+, \\
[L_1 \oplus L_2\ \oplus L_3\ ](\dot{\phi}(t)), & \text{if } \phi(t) = (0, \theta), \\
[L_1 \oplus L_4](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_2, \\
[L_1 \oplus L_2\ ](\dot{\phi}(t)), & \text{if } \phi(t) \in \partial_\theta,
\end{cases} \tag{4.32}$$

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 the simple structure of (4.32), i.e. the meaning of the inf-convolution on the boundaries of the state space and on the slowdown threshold. First 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_1 in the lower part of the interior D and a fraction $\rho_3 = 1 - \rho_1$ on the vertical constraint. Note that ρ_1 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.

On the slowdown threshold the inf-convolution has a similar meaning, i.e., the process spends a fraction of time ρ_1 below the slowdown threshold and a fraction $\rho_2 = 1 - \rho_1$ above it. It is important to note that the fraction of time the process spends on ∂_θ itself is virtually zero. This shows that (4.32) is a particular case of a more general rate function as in (7.9) in [19].

Now we are ready to state the following theorem.

Theorem 4.12. *The process $X(t)$ satisfies a large deviations principle with rate function (4.32), i.e.,*

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

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

Proof. We sketch the proof of this result, as it is reminiscent of results proven in [21]; see also the proof of Theorem 3.3.

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 [20] 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 [20].

Since the map Γ is Lipschitz continuous and the process $Z(t)$ satisfies a large deviation principle (see Theorem 7.2.3 of [19]), one can apply the contraction principle (see Theorem 2.13 of [66]) 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 (4.32). □

To prove Theorem. 4.1, we now recapitulate the main findings of Section 4 in Chapter 3. 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 (4.32) is minimal when the path is straight, with constant speed vector; see Lemma 3.4, and p. 87 of [66].

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:

$$\begin{aligned} \tilde{\lambda} &= \lambda e^{\vartheta_1}, \\ \tilde{\mu}_1 &= \mu_1 e^{\vartheta_2 - \vartheta_1}, \\ \tilde{\mu}_2 &= \mu_2 e^{-\vartheta_2}, \end{aligned} \tag{4.33}$$

where $\vartheta = (\vartheta_1, \vartheta_2)$ is the maximizer in (4.31) with $i = 1$. In fact this is exactly the same new measure we would find using the cost minimization procedure from Section 4.1, due to the immediate equality

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

see (3.16). 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 (4.33). However, for the optimal paths we still have equality between local rate functions and cost functions on the boundaries, see proof of the Lemma 3.4. 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 3.6, regulates the number of subpaths of the optimal path to overflow.

Lemma 4.13. *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, Theorem 4.1 follows; the proof is analogous to the proof of Theorem 3.7.

Chapter 5

Simpler scheme for tandem Jackson network

This chapter is a follow-up of Chapter 3, where an asymptotically efficient state-dependent IS scheme was presented. Importantly, the scheme of Chapter 3 has a substantial drawback: 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 devise an IS algorithm that combines the attractive features of state-dependent and state-independent schemes. The present chapter 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 earlier results from Chapters 2 and 3).

5.1 Importance sampling

This section is dedicated to the presentation of our IS scheme for estimating the probability p_B^s, see (2.5). Before we propose the new measure to be used in this chapter, we mention the main difference with the new measure introduced in Chapter 3. In Chapter 3, 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 this chapter 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.

5.1.1 Importance sampling for $\mu_2 < \mu_1$

Recall that our goal is to modify the IS scheme described in Chapter 3, 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 Chapter 3, 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 3.1 into A_1, A_2 and A_3. 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 $(\lambda^{(\text{syst})}, \mu_1^{(\text{syst})}, \mu_2^{(\text{syst})})$ solve

$$\begin{cases} \lambda^{(\text{syst})} = \mu_1^{(\text{syst})} - s_1(\mu_1^{(\text{syst})} - \mu_2^{(\text{syst})})/(1-s_2) \\ \lambda^{(\text{syst})} + \mu_1^{(\text{syst})} + \mu_2^{(\text{syst})} = \lambda + \mu_1 + \mu_2 \\ \lambda^{(\text{syst})}\mu_1^{(\text{syst})}\mu_2^{(\text{syst})} = \lambda\mu_1\mu_2 \\ \lambda^{(\text{syst})} \le \mu_1^{(\text{syst})} \text{ and } \mu_1^{(\text{syst})} > \mu_2^{(\text{syst})} \\ \lambda^{(\text{syst})},\ \mu_1^{(\text{syst})},\ \mu_2^{(\text{syst})} > 0. \end{cases} \tag{5.1}$$

The superscript "(syst)" indicates that this new measure is the solution to (5.1). Now we can define the (overall) optimal new measure $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ through

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_2, \mu_1, \lambda), & \text{if } s \in A_1, \\ (\lambda^{(\text{syst})}, \mu_1^{(\text{syst})}, \mu_2^{(\text{syst})}), & \text{if } s \in A_2, \\ (\lambda, \mu_1, \mu_2), & \text{if } s \in A_3. \end{cases} \tag{5.2}$$

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

$$\gamma^s(x) := -x_1 \log \frac{\tilde{\lambda}}{\lambda} - (1-x_2) \log \frac{\tilde{\mu}_2}{\mu_2}. \tag{5.3}$$

In the context of Chapter 3, $\gamma^s(x)$ 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 3.7.

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 Chapter 3, 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 (5.1) 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, see Chapter 2, 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. [11, 59]. In order to resolve this, we again use the measure $(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2)$ on or near the horizontal axis, through

$$(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2) := (\tilde{\lambda}, \mu_1\lambda/\tilde{\lambda}, \mu_2). \tag{5.4}$$

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 $(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x))$, similar to (3.21), of which we will prove asymptotic efficiency, and which is a combination of the two measures defined above and the original measure:

$$\begin{aligned} \bar{\lambda}(x) = &\quad \tilde{\lambda}^{\rho_1(x)} \hat{\lambda}^{\rho_2(x)} \lambda^{\rho_3(x)} M(x), \\ \bar{\mu}_1(x) = &\quad \tilde{\mu}_1^{\rho_1(x)} \hat{\mu}_1^{\rho_2(x)} \mu_1^{\rho_3(x)} M(x), \\ \bar{\mu}_2(x) = &\quad \tilde{\mu}_2^{\rho_1(x)} \hat{\mu}_2^{\rho_2(x)} \mu_2^{\rho_3(x)} M(x). \end{aligned} \tag{5.5}$$

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 use the procedure (3.22)

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

where

$$W_1(x) := 2\gamma^s(x) - \delta, \quad W_2(x) := W_1(x_1, \delta/2\gamma^s(0)), \quad W_3(x) := 2\gamma^s(0) - 3\delta. \tag{5.6}$$

Not only does this choice ensure that the new measure (5.5) 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 (5.2) and (5.4), and the natural measure in terms of the gradients of the functions $W_1(x), W_2(x)$ and $W_3(x)$ respectively, see (3.21). This will be useful in the proof of asymptotic efficiency in Section 5.2.

5.1.2 Importance sampling for $\mu_1 \leq \mu_2$

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 $\bar{D}$, see Figure 3.2.

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

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\lambda^{(\mathrm{syst})}, \mu_1^{(\mathrm{syst})}, \mu_2^{(\mathrm{syst})}), & \text{if } s \in B_2, \\ (\lambda, \mu_2, \mu_1), & \text{if } s \in B_3. \end{cases} \tag{5.7}$$

The function $\gamma^s(x)$ is again defined by (5.3), 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 (5.5) (with the 'tilde-measure' given by (5.7)), and the weights $\rho_i(x)$ given through (3.22) in conjunction with (5.6).

- Now consider $s \in B_1$. We know from Chapter 3 that the optimal trajectory for this case consist of three straight subpaths, see also Figure 3.2. 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,

$$\tau^\star := \min\{k : X_{1,k} \geq \alpha_2^{-1} \text{ and } X_{2,k} = 0\}. \tag{5.8}$$

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

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = (\mu_1, \lambda I_{\{k<\tau^\star\}} + \mu_2 I_{\{k\geq\tau^\star\}}, \mu_2 I_{\{k<\tau^\star\}} + \lambda I_{\{k\geq\tau^\star\}}); \tag{5.9}$$

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 (5.3) anymore, but rather by

$$\gamma^s(x) := \log\frac{\mu_2}{\lambda} - x_1 \log\frac{\tilde{\lambda}}{\lambda} + x_2 \log\frac{\tilde{\mu}_2}{\mu_2}, \quad \text{if } s \in B_1, \tag{5.10}$$

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 3.7.

With the function $\gamma^s(x)$ defined by (5.10), the proposed state-dependent new measure $(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x))$ is given by (5.5) (with the 'tilde-measure' given by (5.9)), and the weights $\rho_i(x)$, as before, through (3.22) and (5.6).

5.1.3 Overview of the importance sampling 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);
  2. define the 'hat'-measure (5.4);
  3. define weights $\rho_i(x)$ by (3.22), based on (5.6) and (5.3);
  4. apply (5.5).
- 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 (5.2) by (5.7) 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 (5.9) *and* replacing (5.3) by (5.10) 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 $\rho_1(x) = \rho_2(x)$).

5.2 Asymptotic efficiency

We now prove the asymptotic efficiency of the IS scheme proposed in the previous section.

Theorem 5.1. *Under Assumption 3.5 the IS scheme defined by (5.5) is asymptotically efficient.*

Proof. Our first step is the decomposition of the likelihood $L(\omega)$ of any path $\omega = (X_j, j = 0, \ldots, \sigma)$ in three terms, as we did in Lemma 3.9. For these needs we define smooth function $W(x)$, see (3.18). Combining the definition of the likelihood ratio (1.5) with simple representation of gradients of $W(x)$ (3.22) one obtains

$$\begin{aligned} \log L(\omega) &= \frac{B}{2} \sum_{j=0}^{\sigma-1} \langle DW(X_j), X_{j+1} - X_j \rangle \\ &+ \sum_{k=1}^{2} \frac{1}{2} \sum_{j=0}^{\sigma-1} \langle DW(X_j), v_k \rangle I\{X_j = X_{j+1} \in \partial_k\} \qquad (5.11) \\ &- \sum_{j=0}^{\sigma-1} \log M(X_j). \end{aligned}$$

Now we bound all three summations in (5.11) 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 $\omega = (X_j, j = 0, ..., \sigma)$ and some positive constant C we can, in self-evident notation, construct the following bound:

$$\left| \sum_{j=0}^{\sigma-1} \langle DW(X_j), X_{j+1} - X_j \rangle - (W(X_\sigma) - W(X_0)) \right| \leq \frac{C}{B^2 \epsilon} \sigma + \frac{C}{B} I_{\{\mu_1 \leq \mu_2\} \cap \{s \in B_1\}}. \tag{5.12}$$

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 3.10. 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 (5.12) corresponds to the sum of these contributions over all σ steps. The second term appears only if $\mu_1 \leq \mu_2$ and $s \in B_1$, as a consequence of the non-smoothness of $\gamma^s(x)$ as a function of k, see (5.9), and therefore also of $W(x)$, after the $\tau^\star$-th transition; note that a similar problem was treated in Lemma 4.5. Bearing in mind the definition of the function $W(x)$, see (3.18), we obtain

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

Combining the two last inequalities with (5.12), we derive an upper bound for the first term in (5.11):

$$\sum_{j=0}^{\tau_B^s - 1} \langle DW(X_j), X_{j+1} - X_j \rangle \leq -2\gamma^s(s) + \eta(B) + \frac{C}{B^2 \epsilon} \tau_B^s, \tag{5.13}$$

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

(b) We now proceed with the second term. For any path $\omega = (X_j, j = 0, ..., \sigma)$ and some positive constant $\gamma^\star$ we obtain by routine computations, as were done in Theorem 3.12 and Theorem 3.13.

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

(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 Lemma 3.8.

Upon combining *(a)*, *(b)* and *(c)*, we obtain the following upper bound on the likelihood ratio:

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

After some elementary algebra this leads to

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

Using that $\lim_{B\to\infty}\chi(B)=0$, due to Assumption 3.5, in conjunction with

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

see Lemma 3.11, 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

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

which completes the proof. □

5.3 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 (5.5), see Table 5.1. In the rest of the section we compare the performance of the current scheme with that of the existing methods, in particular with Chapter 3; see Table 5.2.

In Table 5.1 we present simulation results for three different parameter settings using the IS scheme defined in (5.5). We compute the weights $\rho_i(x)$ as in (3.22), choosing $\epsilon=0.005$ and $\delta=-\epsilon\log\epsilon$ to enable comparison with Chapter 3. Each time we perform a fixed number of 10^6 simulation runs. In Table 5.1 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 Table 2.1 and Table 2.2, 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 3.1, showing similar results.

We now turn to comparing the performance of three different IS schemes (as well as straightforward simulations). In Table 5.2 we present the results for the same three scenarios as in Table 5.1. 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 Chapter 2; for the second column we use the state-dependent scheme described in Chapter 3, and the third column contains the outcomes of the

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

	st.-indep., Chapter 2		st.-dep. old, Chapter 3			st.-dep. new	
B	RE	time	RE	virtual time	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
50	$2.21 \cdot 10^{-2}$	37	$3.12 \cdot 10^{-3}$	$9 \cdot 10^6$	$132 + 100$	$3.94 \cdot 10^{-3}$	80
100	$1.37 \cdot 10^{-2}$	77	N/A	N/A	N/A	$4.74 \cdot 10^{-3}$	168

Table 5.2.1: $(\lambda, \mu_1, \mu_2) = (0.3, 0.36, 0.34)$ 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
50	N/A	N/A	$7.67 \cdot 10^{-4}$	$3 \cdot 10^6$	$104 + 68$	$1.58 \cdot 10^{-3}$	18
100	N/A	N/A	N/A	N/A	N/A	$1.81 \cdot 10^{-3}$	35

Table 5.2.2: $(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$ 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
50	$3.67 \cdot 10^{-1}$	53	$6.73 \cdot 10^{-2}$	$22 \cdot 10^6$	$123 + 68$	$7.91 \cdot 10^{-2}$	84
100	$2.33 \cdot 10^{-1}$	116	N/A	N/A	N/A	$13.4 \cdot 10^{-2}$	189

Table 5.2.3: $(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$ and $s = (0, 0)$

Table 5.2: Comparison of different schemes

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 Chapter 3, 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 (5.1) 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 (5.1). However, when we did the simulations in Chapter 3 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 Table 5.2.1 and Table 5.2.3 it becomes clear that both the scheme in Chapter 3 and the current scheme provide a relative error that is much smaller than with the state-independent scheme from Chapter 2. (Note that the latter is not available in Table 5.2.2 since we only allowed the origin as starting state in Chapter 2). 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, see Table 5.2.2, but this may also hold when we apply a state-independent scheme for these parameter values, see Table 2.1 and Table 2.2.

When we compare the current scheme with the old state-dependent scheme in Chapter 3, 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.

5.4 Conclusions

In this chapter we analyzed the probability of overflow in the second queue of the tandem Jackson network. The focus was on efficient IS algorithm for estimating this probability, which outperforms methods developed earlier, see Chapter 3. The scheme constructed in this chapter provides a good compromise between the naive state-independent scheme from Chapter 2 and the very precise state-dependent scheme, designed in Chapter 3. It is almost as easy and fast as the first one, and approximately of the same accuracy as the latter one.

Chapter 6

Simpler scheme for slowdown network

This chapter is a follow-up of Chapter 4, in the same way as Chapter 5 is a follow-up of Chapter 3. An asymptotically efficient state-dependent scheme presented in Chapter 4 entails us to recompute the new measure after each transition, which is time-consuming.

The contribution of the current chapter is that we present a simple and efficient IS implementation for simulating the overflow probability in the slowdown model. On the one hand it is as easy to implement as the scheme in Chapter 2, while also performing comparably in terms of computational demand. On the other hand it allows any given starting state, while inheriting asymptotic efficiency (for the majority of cases) from Chapter 4. We provide a substantial number of numerical results, including a variety of parameter settings, and make a comparison with Chapter 2.

6.1 Importance sampling

In Chapter 4 we developed an asymptotically efficient IS-based method for estimating p_B^s, but this has, from a practical point of view, important drawbacks: the new measure is state-dependent, and needs to be recomputed at every transition (amounting to jointly solving two cubic equations), thus severely limiting the efficiency gain. In this section we present a new measure that is still state-dependent, but its computation is substantially less demanding, as it requires just a few cubic equations to be solved. As we will see, the speed-up of this new scheme is still substantial.

We give a detailed description of the IS scheme for the case $\mu_2 < \mu_1^+ < \mu_1$; for the other cases (i.e., $\mu_1^+ < \mu_2 \leq \mu_1$ and $\mu_1^+ \leq \mu_1 < \mu_2$), we just present the results. Throughout this section we fix the starting state s and assume it is situated below the slowdown threshold, i.e., $s \in \bar{D}$, which is evidently the most interesting case.

6.1.1 Importance sampling for $\mu_2 < \mu_1^+ < \mu_1$.

At first recall from Chapter 4 the most probable path to overflow and the pair of new measures $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$ that will ensure that any sample path will follow the optimal trajectory with high probability. To ease the exposition on the new measures we divide the state space as it is shown in Figure 4.1. This figure also provides some examples of the most probable overflow trajectories. We are particularly interested in the partition of the bottom part of the state space, i.e., in A_1, A_2 and A_3.

The new measures for $s \in A_1 \cup A_3$ are not difficult, as we will see in (6.9). However, to find the optimal new measure for $s \in A_2$ one first needs to jointly solve

$$\begin{cases} \lambda^{(\text{syst})} = \mu_1^{(\text{syst})} + \frac{\kappa^{(\text{syst})} - s_1}{\theta - s_2}(\mu_1^{(\text{syst})} - \mu_2^{(\text{syst})}) \\ \lambda^{(\text{syst})} + \mu_1^{(\text{syst})} + \mu_2^{(\text{syst})} = \lambda + \mu_1 + \mu_2 \\ \lambda^{(\text{syst})}\mu_1^{(\text{syst})}\mu_2^{(\text{syst})} = \lambda\mu_1\mu_2 \\ \lambda^{(\text{syst})} \le \mu_1^{(\text{syst})} \text{ and } \mu_1^{(\text{syst})} > \mu_2^{(\text{syst})} \\ \lambda^{(\text{syst})}, \mu_1^{(\text{syst})}, \mu_2^{(\text{syst})} > 0 \end{cases} \tag{6.1}$$

and

$$\begin{cases} \lambda^{+(\text{syst})} = \mu_1^{+(\text{syst})} - \frac{\kappa^{(\text{syst})}}{1-\theta}(\mu_1^{+(\text{syst})} - \mu_2^{+(\text{syst})}) \\ \lambda^{+(\text{syst})} + \mu_1^{+(\text{syst})} + \mu_2^{+(\text{syst})} = \lambda + \mu_1^+ + \mu_2 \\ \lambda^{+(\text{syst})}\mu_1^{+(\text{syst})}\mu_2^{+(\text{syst})} = \lambda\mu_1^+\mu_2 \\ \lambda^{+(\text{syst})} \le \mu_1^{+(\text{syst})} \text{ and } \mu_1^{+(\text{syst})} > \mu_2^{+(\text{syst})} \\ \lambda^{+(\text{syst})}, \mu_1^{+(\text{syst})}, \mu_2^{+(\text{syst})} > 0 \end{cases} \tag{6.2}$$

with the condition

$$\kappa^{(\text{syst})} := s_1 - \frac{\mu_1^{(\text{syst})} - \lambda^{(\text{syst})}}{\mu_1^{(\text{syst})} - \mu_2^{(\text{syst})}}(\theta - s_2) = \frac{\mu_1^{+(\text{syst})} - \lambda^{+(\text{syst})}}{\mu_1^{+(\text{syst})} - \mu_2^{+(\text{syst})}}(1 - \theta). \tag{6.3}$$

It can be verified that this amounts to solving two coupled cubic equations. The superscripts "(syst)" and "+(syst)" indicate that the solution is in fact the optimal change of measure to reach the state $(0, 1)$ following a concatenation of two straight line starting in s, with intersection in $(\kappa^{(\text{syst})}, \theta)$.

Now we can define the (overall) optimal new measures below and above the slowdown threshold, which depend only on the starting state s. The new measure below the slowdown threshold, as given through $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$, is

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_2, \mu_1, \lambda) & \text{if } s \in A_1, \\ (\lambda^{(\text{syst})}, \mu_1^{(\text{syst})}, \mu_2^{(\text{syst})}) & \text{if } s \in A_2, \\ (\lambda, \mu_1, \mu_2) & \text{if } s \in A_3. \end{cases} \tag{6.4}$$

Above the slowdown threshold the new measure, given through $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$, is

$$(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+) = \begin{cases} (\mu_2, \mu_1^+, \lambda) & \text{if } s \in A_1, \\ (\lambda^{+(\text{syst})}, \mu_1^{+(\text{syst})}, \mu_2^{+(\text{syst})}) & \text{if } s \in A_2, \\ (\lambda, \mu_1^+, \mu_2) & \text{if } s \in A_3. \end{cases} \tag{6.5}$$

Now, let us define $\gamma^s(x)$ to be the residual cost of moving from state x to ∂_e along the path to overflow that started in s:

$$\gamma^s(x) := \begin{cases} \gamma_1^s(x) + \gamma_2^s(\kappa^\star, \theta) & \text{if } x \in \bar{D}, \\ \gamma_2^s(x) & \text{if } x \in \bar{D}^+, \end{cases} \tag{6.6}$$

with

$$\gamma_1^s(x) := -(x_1 - \kappa^\star) \log \frac{\tilde{\lambda}}{\lambda} - (\theta - x_2) \log \frac{\tilde{\mu}_2}{\mu_2}, \quad \text{if } x \in \bar{D} \tag{6.7}$$

being the minimal cost of the bottom part of the path to overflow and

$$\gamma_2^s(x) := -x_1 \log \frac{\tilde{\lambda}^+}{\lambda} - (1 - x_2) \log \frac{\tilde{\mu}_2^+}{\mu_2}, \quad \text{if } x \in \bar{D}^+ \tag{6.8}$$

being the minimal cost of the top part of the optimal path to overflow; the optimal crossing state $(\kappa^\star, \theta)$ is as follows

$$\kappa^\star := \begin{cases} \max(0, s_1 - \alpha_1(\theta - s_2)) & \text{if } s \in A_1, \\ \kappa^{(\mathrm{syst})} & \text{if } s \in A_2, \\ s_1 - (\theta - s_2)/\alpha_1 & \text{if } s \in A_3. \end{cases} \tag{6.9}$$

Note that $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$, $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$ and $\kappa^\star$ (given by (6.4), (6.5) and (6.9) respectively) are fixed, i.e., they only depend on the fixed initial state s, and not on the current state x (as was the case in Chapter 4). It is also important to note that $\gamma^s(s)$, the total cost of moving from the starting state s to the exit boundary ∂_e, equals the decay rate of p_B^s, i.e., $B^{-1} \log p_B^s \to -\gamma^s(s)$, see Theorem 4.1.

Notice that the function $\gamma^s(x)$ is piecewise-linear in x, since the new tilde-measure, i.e., $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$, 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 Chapter 4, where we used the optimal new measure for each current state x with its cost $\gamma^x(x)$. Therefore a pair of cubic equations (corresponding to (6.1)–(6.3) with s replaced by x) had to be solved for *each* state x in the sample path. In our current approach, computation of the new measure requires the solution of (6.1)–(6.3) just once.

It is known, e.g. from our previous research in Chapter 2, that the new measures $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$, which make 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. [11, 59]. In order to resolve this, we first introduce the 'hat'-measure $(\hat{\lambda}, \hat{\mu}_1, \hat{\mu}_2)$ and $(\hat{\lambda}^+, \hat{\mu}_1^+, \hat{\mu}_2^+)$ as in Chapter 5, to be used when the current state is on or near the horizontal axis, through

$$(\tilde{\lambda}, \mu_1\lambda/\tilde{\lambda}, \mu_2) \text{ and } (\tilde{\lambda}^+, \mu_1^+\lambda/\tilde{\lambda}^+, \mu_2^+). \tag{6.10}$$

The main idea behind it 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 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:

$$\begin{array}{lll}
\bar{\lambda}(x) = & \tilde{\lambda}^{\rho_1}\hat{\lambda}^{\rho_2}\lambda^{\rho_3}M(x), & \text{if } x \in \bar{D}, \\
\bar{\mu}_1(x) = & \tilde{\mu}_1^{\rho_1}\hat{\mu}_1^{\rho_2}\mu_1^{\rho_3}M(x), & \text{if } x \in \bar{D}, \\
\bar{\mu}_2(x) = & \tilde{\mu}_2^{\rho_1}\hat{\mu}_2^{\rho_2}\mu_2^{\rho_3}M(x), & \text{if } x \in \bar{D}, \\
\bar{\lambda}^+(x) = & (\tilde{\lambda}^+)^{\rho_1}(\hat{\lambda}^+)^{\rho_2}(\lambda)^{\rho_3}M^+(x), & \text{if } x \in \bar{D}^+, \\
\bar{\mu}_1^+(x) = & (\tilde{\mu}_1^+)^{\rho_1}(\hat{\mu}_1^+)^{\rho_2}(\mu_1^+)^{\rho_3}M^+(x), & \text{if } x \in \bar{D}^+, \\
\bar{\mu}_2^+(x) = & (\tilde{\mu}_2^+)^{\rho_1}(\hat{\mu}_2^+)^{\rho_2}(\mu_2)^{\rho_3}M^+(x), & \text{if } x \in \bar{D}^+,
\end{array} \tag{6.11}$$

where $M(x)$ and $M^+(x)$ are normalization functions, and the $\rho_i(x)$ are *weights* and given by

$$\begin{aligned}
\rho_1(x) &= N(x) \cdot e^{-\frac{2\gamma^s(x_1,x_2)-\delta}{\epsilon}}, \\
\rho_2(x) &= N(x) \cdot e^{-\frac{2\gamma^s(x_1,\frac{\delta}{2}\log\frac{\mu_2}{\lambda})-\delta}{\epsilon}}, \\
\rho_3(x) &= N(x) \cdot e^{-\frac{2\gamma^s(0,0)-\delta}{\epsilon}}.
\end{aligned} \tag{6.12}$$

Here $N(x)$ is a normalization function and δ and ϵ are small positive numbers. We mention again that this measure is, albeit state-dependent, of low computational complexity, as it does not require to solve cubic systems for any point along the path (except s).

6.1.2 Importance sampling for $\mu_1^+ \leq \mu_2 < \mu_1$.

Here we present the IS scheme for the case when $\mu_1^+ \leq \mu_2 < \mu_1$. At first we provide the pair of new measures $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$ and $(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+)$ under which virtually any sample path follows the most probable trajectory with high probability. The bottom part of the state space $\bar{D}$ is divided in subspaces B_i as depicted in Figure 4.2. The new measure below θ, $(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2)$, is as follows:

$$(\tilde{\lambda}, \tilde{\mu}_1, \tilde{\mu}_2) = \begin{cases} (\mu_2, \mu_1, \lambda), & \text{if } s \in B_1, \\ (\lambda^{(\text{syst})}, \mu_1^{(\text{syst})}, \mu_2^{(\text{syst})}), & \text{if } s \in B_2, \\ (\lambda, \mu_1, \mu_2), & \text{if } s \in B_3, \end{cases} \tag{6.13}$$

whereas above θ we have

$$(\tilde{\lambda}^+, \tilde{\mu}_1^+, \tilde{\mu}_2^+) = \begin{cases} (\sqrt{\frac{\lambda\mu_1^+}{z^+}}, \sqrt{\frac{\lambda\mu_1^+}{z^+}}, \mu_2 z^+), & \text{if } s \in B_1 \\ (\lambda^{+(\text{syst})}, \mu_1^{+(\text{syst})}, \mu_2^{+(\text{syst})}), & \text{if } s \in B_2, \\ (\lambda, \mu_2, \mu_1^+), & \text{if } s \in B_3, \end{cases} \tag{6.14}$$

where z^+ is the unique in $(0,1)$ solution of (2.36). The optimal crossing state $(\kappa^\star, \theta)$ is now given by

$$\kappa^\star := \begin{cases} 0 & \text{if } s \in B_1, \\ \kappa^{(\text{syst})} & \text{if } s \in B_2, \\ s_1 - (\theta - s_2)/\alpha_1 & \text{if } s \in B_3. \end{cases} \tag{6.15}$$

The function $\gamma^s(x)$ is defined by (6.6) with (6.7), (6.8) and (6.15). The total cost of the path $\gamma^s(s)$ again is the decay rate of p_B^s, i.e., $B^{-1}\log p_B^s \to -\gamma^s(s)$, see Theorem 4.1. The new state-dependent measures $(\bar\lambda(x), \bar\mu_1(x), \bar\mu_2(x))$ and $(\bar\lambda^+(x), \bar\mu_1^+(x), \bar\mu_2^+(x))$ are given by (6.11), where the 'hat'-measures and the weights $\rho_i(x)$ are defined by (6.10) and (6.12), respectively.

6.1.3 Importance sampling for $\mu_1^+ < \mu_1 \le \mu_2$.

This case is the most difficult case. We partition the state space in subspaces C_i as in Figure 4.3.

- At first let us assume that $s \in C_2 \cup C_3$. Then we define the new measure below the slowdown threshold

$$(\tilde\lambda, \tilde\mu_1, \tilde\mu_2) = \begin{cases} (\lambda^{(\text{syst})}, \mu_1^{(\text{syst})}, \mu_2^{(\text{syst})}), & \text{if } s \in C_2, \\ (\lambda, \mu_2, \mu_1), & \text{if } s \in C_3, \end{cases} \tag{6.16}$$

  and above it

$$(\tilde\lambda^+, \tilde\mu_1^+, \tilde\mu_2^+) = \begin{cases} (\lambda^{+(\text{syst})}, \mu_1^{+(\text{syst})}, \mu_2^{+(\text{syst})}), & \text{if } s \in C_2, \\ (\lambda, \mu_2, \mu_1^+), & \text{if } s \in C_3, \end{cases} \tag{6.17}$$

  The optimal crossing state $(\kappa^\star, \theta)$ is now given by

$$\kappa^\star := \begin{cases} \kappa^{(\text{syst})} & \text{if } s \in C_2, \\ s_1 - (\theta - s_2)/\alpha_2 & \text{if } s \in C_3. \end{cases} \tag{6.18}$$

  The function $\gamma^s(x)$ is again defined by (6.6) with (6.7), (6.8) and (6.18). The total costs of the path, $\gamma^s(s)$, is the decay rate of p_B^s, i.e., $B^{-1}\log p_B^s \to -\gamma^s(s)$, see Theorem 4.1. The new measures $(\bar\lambda(x), \bar\mu_1(x), \bar\mu_2(x))$ and $(\bar\lambda^+(x), \bar\mu_1^+(x), \bar\mu_2^+(x))$ are given by (6.11), where the 'hat'-measures and the weights $\rho_i(x)$ are defined by (6.10) and (6.12), respectively.

- Let us now proceed to the case $s \in C_1$. In this case the optimal new measure below the slowdown threshold at the k-th transition is

$$(\tilde\lambda, \tilde\mu_1, \tilde\mu_2) = (\mu_1, \lambda I_{\{k<\tau^\star\}} + \mu_2 I_{\{k\ge\tau^\star\}}, \mu_2 I_{\{k<\tau^\star\}} + \lambda I_{\{k\ge\tau^\star\}}), \tag{6.19}$$

  where $\tau^\star := \min\{k : X_k \in C_3 \text{ and } X_{2,k} = 0\}$; see also the optimal trajectory that starts in C_1, Figure 4.3. Above it the new measure is defined by

$$(\tilde\lambda^+, \tilde\mu_1^+, \tilde\mu_2^+) = (\mu_1, \lambda\mu_1^+/q\mu_1, q\mu_2), \tag{6.20}$$

where q is defined by (2.37).

It is important, that the residual cost of the bottom part of the path to overflow, namely $\gamma_1(x)$, for this case is different from the one defined by (6.7). Here it is

$$\gamma_1^s(x) = \theta \log \frac{\mu_2}{\lambda} - (x_1 - \kappa^\star) \log \frac{\tilde{\lambda}}{\lambda} + x_2 \log \frac{\tilde{\mu}_2}{\mu_2}, \tag{6.21}$$

where

$$\kappa^\star = (1-\theta)(\lambda\mu_1^+ - q\mu_1^2)/(\lambda\mu_1^+ - q^2\mu_1\mu_2). \tag{6.22}$$

For this case $\gamma_2^s(x)$ is defined by (6.8), and $\gamma^s(x)$ again by (6.6) with (6.21), (6.8) and (6.22). As before, $\gamma^s(s)$ is the decay rate of p_B^s, i.e., $B^{-1}\log p_B^s \to -\gamma^s(s)$. The new state-dependent measures $(\bar{\lambda}(x), \bar{\mu}_1(x), \bar{\mu}_2(x))$ and $(\bar{\lambda}^+(x), \bar{\mu}_1^+(x), \bar{\mu}_2^+(x))$ are given by (6.11), where the 'hat'-measures and the weights $\rho_i(x)$ are defined by (6.10) and (6.12) respectively.

6.1.4 Overview of the importance sampling scheme

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

- When $\mu_2 < \mu_1^+ < \mu_1$ one needs to
  1. define the 'primary' new measures (6.4) and (6.5);
  2. define 'hat'-measures (6.10);
  3. define weights $\rho_i(x)$ by (6.12) based on (6.6)–(6.9);
  4. apply (6.11).
- When $\mu_1^+ \le \mu_2 < \mu_1$, the same procedure is followed, only replacing the 'primary' new measures (6.4) and (6.5) by (6.13) and (6.14) in step 1; and (6.9) by (6.15) in step 3.
- When $\mu_1^+ < \mu_1 \le \mu_2$ and $s \in C_2 \cup C_3$, we follow the same algorithm, only with the 'primary' new measure replaced by (6.16) and (6.17) in step 1; and using (6.18) instead of (6.9) in step 3.
- When $\mu_1^+ < \mu_1 \le \mu_2$ and $s \in C_1$, again the same procedure is followed, this time replacing the 'primary' new measure by that in (6.19) and (6.20); we also replace (6.7) by (6.21) and (6.9) by (6.22) in step 3.

In Chapter 4 we proved asymptotic efficiency of the fully state-dependent IS scheme, i.e., when the 'primary' new measure was dependent on the current state. Analyzing the simplified IS scheme (6.11) we have to deal with the additional complication that the discontinuity of $\gamma^s(x)$ around the slowdown threshold, see e.g., (6.6)–(6.8), (6.4) and (6.5).

6.2 Asymptotic efficiency

This section is dedicated to analysis of asymptotic efficiency of the IS scheme presented in the previous section.

Theorem 6.1. *If*

- $\mu_2 < \mu_1^+ < \mu_1$ *and* $s \in A_1 \cup A_3$, *or*
- $\mu_1^+ \leq \mu_2 < \mu_1$ *and* $s \in B_3$, *or*
- $\mu_1^+ < \mu_1 \leq \mu_2$ *and* $s \in C_1 \cup C_3$

and Assumption 3.5 holds then the new measure (6.11) is asymptotically efficient.

We skip proof of this theorem since it is very similar to proof of Theorem 5.1.

Situation is more involved when the starting state s does not belong to the subspaces indicated in Theorem 6.1. In this case we need to control the likelihood ratio of the paths with multiple crossings of the slowdown threshold. More precise, let us consider a path path $\omega = (X_j, j = 0, ..., \sigma)$ which makes σ^+ slowdown threshold crossings up to time σ, see also Lemma 4.5. Then the sum of likelihood ratios of all slowdown threshold crossings can be bounded as follows

$$R = \frac{B}{2} \log \left(\frac{\tilde{\lambda}}{\tilde{\lambda}^+} \right) \left| \sum_{i=1}^{\sigma^+} (-1)^i (\kappa - \eta_i) \right|, \tag{6.23}$$

where $B\eta_i$ is the number of jobs in the first buffer prior to the i-th crossing of the slowdown threshold. We impose the following regularity condition on (6.23).

Conjecture 6.2. *For any scaled sample path* ω^s *we believe the following holds true*

$$\lim_{B \to \infty} \frac{1}{B} \log \mathbb{E}(e^R | I_B(\omega^s) = 1) = 0,$$

where the bound R *is defined by (6.23).*

Intuitively, as B grows large the random variable σ^+ will essentially not grow with B, while due the scaled position(s) where the threshold is crossed by the sample path will become close to the point where the most probable path crosses the threshold, i.e., the η_i will converge to κ.

Proposition 6.3. *If Conjecture 6.2 is true then the new measure (6.11) is asymptotically efficient for any* $s \in \bar{D}$.

Proof of this fact is very similar to the proof of Theorem 5.1.

Finally, we like to mention that in the settings of Theorem 6.1, $\tilde{\lambda} = \tilde{\lambda}^+$ and therefore R defined by (6.23) is 0.

6.3 Numerical results

We start our numerical studies from the case with the shifting bottleneck, since it is the most interesting one. In Tables 6.1 and 6.2 we present simulation results for two different parameter settings using the new measure defined in (6.11). Instead of performing a fixed number of simulation runs such as in much of the IS literature, we simulated until the relative error of the estimator reached the value of 10^{-2}. In the tables we present 95% confidence intervals for p_B^s, the number of needed replications (# runs), the used machine time in seconds, and the number of 'succesful' replications (# succ.), i.e. the number of runs that resulted in buffer overflow.

We compare several starting states s, three values of the overflow level B, and two values of ϵ; the value of δ was taken to be $\delta = -\frac{1}{3}\epsilon \log \epsilon$. Note that the starting states in Table 6.1 belong to B_1, B_2 and B_3 respectively; we only include results for

$\epsilon = 0.01$					
s	B	p_B^s	# succ.	# runs	time
$(0,0)$	20	$3.79 \cdot 10^{-7} \pm 7.44 \cdot 10^{-9}$	18,565	41,985	< 1
	50	$1.28 \cdot 10^{-16} \pm 2.50 \cdot 10^{-18}$	36,999	193,128	2
	100	$3.48 \cdot 10^{-32} \pm 6.82 \cdot 10^{-34}$	66,473	1,097,097	8
$(0.7B,0)$	20	$6.11 \cdot 10^{-3} \pm 1.19 \cdot 10^{-4}$	8,946	8,946	1
	50	$3.79 \cdot 10^{-6} \pm 7.44 \cdot 10^{-8}$	24,969	24,969	10
	100	$3.25 \cdot 10^{-11} \pm 6.37 \cdot 10^{-13}$	49,528	49,528	37
$(1.5B,0)$	20	$5.18 \cdot 10^{-1} \pm 1.01 \cdot 10^{-2}$	11,287	12,888	< 1
	50	$1.35 \cdot 10^{-1} \pm 2.65 \cdot 10^{-3}$	66,997	77,942	2
	100	$1.05 \cdot 10^{-2} \pm 2.05 \cdot 10^{-4}$	316,351	367,327	21
$\epsilon = 0.001$					
$(0,0)$	20	$3.79 \cdot 10^{-7} \pm 7.44 \cdot 10^{-9}$	15,576	28,332	< 1
	50	$1.28 \cdot 10^{-16} \pm 2.52 \cdot 10^{-18}$	33,542	58,332	2
	100	$3.54 \cdot 10^{-32} \pm 6.95 \cdot 10^{-34}$	56,982	109,992	6
$(0.7B,0)$	20	$6.12 \cdot 10^{-3} \pm 1.20 \cdot 10^{-4}$	8,946	8,946	1
	50	$3.73 \cdot 10^{-6} \pm 7.32 \cdot 10^{-8}$	24,665	24,665	9
	100	$3.28 \cdot 10^{-11} \pm 6.43 \cdot 10^{-13}$	51,365	51,365	36

Table 6.1: Simulation results for $\theta = 0.8$ and $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.1, 0.7, 0.15, 0.2)$

starting states on the horizontal boundary, as these are more difficult to obtain than results for starting states in the interior. In Table 6.2 we only considered $s = (0,0)$ since for the other states the event of interest was not rare, and hence the results are not interesting. Also, for $s = (1.5B, 0)$ in Table 6.1 we omitted results for $\epsilon = 0.001$ as these were indistinguishable from those with $\epsilon = 0.01$.

Clearly, the IS scheme provides fast and reliable estimates. In some cases, especially when B grows large, the running times may be sensitive to the choice of ϵ and δ.

We also performed a few straightforward simulations (i.e., without IS) for comparison, using the same relative error of 10^{-2}. For the parameter settings of Table 6.1 with $B = 20$, this took 4521 seconds ($\pm 5 \cdot 10^9$ runs) for $s = (0,0)$, and 16 seconds ($\pm 2 \cdot 10^6$ runs) for $s = (0.7B, 0)$. In the settings of Table 6.2 with $B = 50$ it took

$\epsilon = 0.01$					
s	B	p_B^s	# succ.	# runs	time
$(0,0)$	20	$5.62 \cdot 10^{-2} \pm 1.11 \cdot 10^{-3}$	33,371	98,230	2
	50	$1.18 \cdot 10^{-3} \pm 2.31 \cdot 10^{-5}$	99,116	295,633	19
	100	$1.63 \cdot 10^{-6} \pm 3.19 \cdot 10^{-8}$	143,194	382,120	55
$\epsilon = 0.001$					
$(0,0)$	20	$5.63 \cdot 10^{-2} \pm 1.11 \cdot 10^{-3}$	39,496	91,596	2
	50	$1.19 \cdot 10^{-3} \pm 2.33 \cdot 10^{-5}$	99,567	241,332	18
	100	$1.63 \cdot 10^{-6} \pm 3.21 \cdot 10^{-8}$	128,864	320,120	49
$(0.35B,0)$	20	$2.02 \cdot 10^{-1} \pm 3.96 \cdot 10^{-3}$	39,937	45,557	2
	50	$1.34 \cdot 10^{-2} \pm 2.62 \cdot 10^{-4}$	51,327	52,628	9
	100	$1.31 \cdot 10^{-4} \pm 2.56 \cdot 10^{-6}$	62,174	62,356	25

Table 6.2: Simulation results for $\theta = 0.8$ and $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.32, 0.34)$.

$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.1, 0.7, 0.15, 0.2)$			
B	st.-ind., Chapter 2	st.-dep., [21]	current
20	$1.49 \cdot 10^{-3}$	$2.63 \cdot 10^{-3}$	$3.54 \cdot 10^{-3}$
50	$2.06 \cdot 10^{-3}$	$7.87 \cdot 10^{-3}$	$8.00 \cdot 10^{-3}$
100	$2.75 \cdot 10^{-3}$	$19.71 \cdot 10^{-3}$	$17.01 \cdot 10^{-3}$
$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.32, 0.34)$			
B	st.-ind., Chapter 2	st.-dep., [21]	current
20	$0.92 \cdot 10^{-3}$	$5.30 \cdot 10^{-3}$	$6.00 \cdot 10^{-3}$
50	$12.50 \cdot 10^{-3}$	$8.40 \cdot 10^{-3}$	$11.00 \cdot 10^{-3}$
100	$39.69 \cdot 10^{-3}$	$12.20 \cdot 10^{-3}$	$11.00 \cdot 10^{-3}$

Table 6.3: Comparison of relative errors for three IS schemes

118 seconds ($\pm 10^7$ runs).

To enable comparison with the state-independent scheme described in Chapter 2 and the state-dependent scheme in [21], we also fixed the number of runs to be 10^6 and compared the relative errors, see Table 6.3. Here, $s = (0,0)$, $\theta = 0.8$, and in the state-dependent schemes $\epsilon = 0.03/\sqrt{B}$ and $\delta = -\epsilon \log \epsilon$. As can be expected, both state-dependent schemes provide good estimates, but the performance of the state-independent scheme strongly depends on the parameters.

We also present some results for the cases in which either the first queue is always the bottleneck, see Table 6.4, or the second queue is always the bottleneck, see Table 6.5. These results suggest that the current scheme outperforms state-independent scheme from Chapter 2, see Tables 2.3 and 2.5. In both cases we fixed the relative error, but note that we took it to be 0.05 instead of 0.01 when the first queue is the bottleneck. The choice of $s = (0.35B, 0)$ in the case where $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.25, 0.35, 0.28, 0.4)$ corresponds to the point where the optimal path from $(0,0)$ to ∂_e leaves the horizontal axis. For the case $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.35, 0.34)$ we did not include results for $s = (3B, 0)$, since the 'new' measure here coincides with the old measure, i.e. it is optimal to use straightforward simulations here.

We now demonstrate techniques that enable selection of a proper value for the

$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.25, 0.35, 0.28, 0.4)$, RE $= 0.05$					
s	B	p_B^s	# succ.	# runs	time
$(0,0)$	20	$1.11 \cdot 10^{-4} \pm 1.09 \cdot 10^{-5}$	45,685	83,436	2
	50	$3.43 \cdot 10^{-11} \pm 3.36 \cdot 10^{-12}$	79,901	148,256	7
	100	$5.72 \cdot 10^{-22} \pm 5.60 \cdot 10^{-23}$	235,502	439,006	42
$(0.35B, 0)$	20	$6.18 \cdot 10^{-4} \pm 6.06 \cdot 10^{-5}$	38,190	40,333	1
	50	$2.56 \cdot 10^{-9} \pm 2.50 \cdot 10^{-10}$	92,005	92,234	5
	100	$4.64 \cdot 10^{-18} \pm 4.55 \cdot 10^{-19}$	206,100	206,182	25
$(3B, 0)$	20	$1.62 \cdot 10^{-1} \pm 1.58 \cdot 10^{-2}$	21,106	23,496	1
	50	$1.15 \cdot 10^{-3} \pm 1.13 \cdot 10^{-4}$	43,378	52,840	7
	100	$1.90 \cdot 10^{-7} \pm 1.86 \cdot 10^{-8}$	78,229	91,231	25

Table 6.4: Simulation results for $\theta = 0.8$ and first buffer being the bottleneck

$(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.35, 0.34)$, RE $= 0.01$					
$(0,0)$	20	$5.86 \cdot 10^{-2} \pm 1.44 \cdot 10^{-3}$	32,283	76,169	2
	50	$1.42 \cdot 10^{-3} \pm 2.79 \cdot 10^{-5}$	112,128	269,968	21
	100	$2.64 \cdot 10^{-6} \pm 5.18 \cdot 10^{-8}$	275,112	661,247	121
$(0.35B, 0)$	20	$2.11 \cdot 10^{-1} \pm 4.15 \cdot 10^{-3}$	37,178	42,163	2
	50	$1.50 \cdot 10^{-2} \pm 2.95 \cdot 10^{-4}$	82,133	92,301	15
	100	$2.15 \cdot 10^{-4} \pm 4.21 \cdot 10^{-6}$	114,694	124,994	35

Table 6.5: Simulation results for $\theta = 0.8$ and second buffer being the bottleneck

slowdown threshold $m = \theta B$. A first *caveat* is the following. It is natural to expect that smaller θ, will provide better protection of the second node and consequently smaller probability of overflow (with s being the origin), but this is not always the case. Indeed, numerical experiments show that, starting in $\theta = 1$, decreasing θ leads to a reduction of p_B^s. However, continuing to decrease θ, the probability of interest will start to *increase*. The same holds for the *stationary* probability of the process X_k, with the slowdown threshold θ, to be above level 1, denoted by $\pi^\theta(B)$, see Figs. 6.1 and 6.2. In these graphs we plotted $\pi^\theta(B)$ against θ for parameters $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.32, 0.34)$, with overflow levels $B = 20$ and $B = 50$.

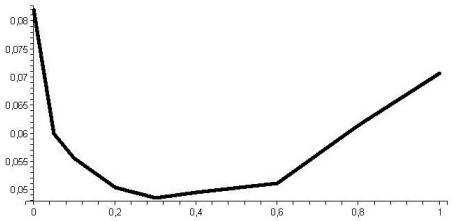

Figure 6.1: $\pi^\theta(20)$ against θ

For the case of 'shifting bottlenecks', i.e., $\mu_1^+ < \mu_2 < \mu_1$, we now provide an

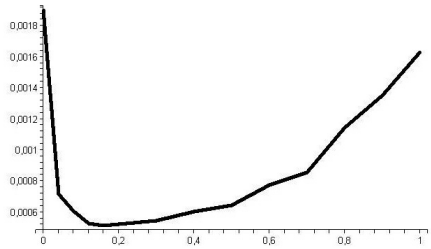

Figure 6.2: $\pi^{\theta}(50)$ against θ

explanation for the non-monotone behavior of $\pi^{\theta}(B)$. Clearly, $\pi^{\theta}(B)$ has decay rate $\theta \log \rho_2 + (1-\theta) \log z^+$, for $\theta \in (0,1]$ and z^+ from (2.36), see Section 6.1; ρ_2 is defined as λ/μ_2. However, when $\theta = 0$ the decay rate is just $\log \rho_2$, as the queue is then an ordinary tandem queue (without backpressure); in Chapter 2 it was shown that $z^+ < \rho_2$. This shows that the decay rate is discontinuous in $\theta = 0$, explaining that $\pi^0(B) > \pi^{0+}(B)$ (for B large).

The above type of justification for the observed non-monotonicity is valid only for the case that $\mu_1^+ \leq \mu_2 < \mu_1$. Another explanation for the decreasing nature of $\pi^{\theta}(B)$ for small θ is in the 'specific' behavior of the X_k around the origin. More precisely, consider the process X_k with starting state $s = (0,1)$, and compare threshold levels $\theta = 0$ and $\theta = 1/B$. In the latter case, the first server operates at full speed only when the second queue is empty. It is not difficult to see that the probability of transition $(1,0) \to (0,1)$ is higher when $\theta = 1/B$; and the probability of the 'terminating' transition $(0,1) \to (0,0)$ is $\mu_2/(\lambda + \mu_2)$, which does not depend on θ. This means that the probability of overflow starting from the origin, p_B^0, is higher when $\theta = 0$ than when $\theta = 1/B$, even though we have 'more slowdown' in the first case. One can generalize this type of arguments for the other states around the origin and the other values of θ. These arguments, unlike the ones based on decay rates, hold for all parameter values.

We now demonstrate how to develop procedures for optimally choosing the value of the slowdown threshold. The primary role of the backpressure mechanism is to control the probability of some undesirable event, viz. overflow in the second buffer (expressed in terms of p_B^s). However, introducing server slowdown has a negative side effect: the expected sojourn time of a job increases. In order to find an optimal value of θ, one could, for given coefficient α and β, minimize the following (dis-)utility function

$$u(\theta, B) = -\alpha \log^{-1} p_B^0 + \beta S(\theta, B),$$

where $S(\theta, B)$ is the mean sojourn time of a job, α is the penalty for overflow and β is the cost for each job being in the system per unit time; we assume s is the origin. The α and β should be chosen by the service provider, and should reflect the Service

Level Agreement (SLA) as agreed upon.

We present plots of the utility functions $u(\theta, 20)$ and $u(\theta, 50)$, with $\alpha = 10$ and $\beta = 3$, for a system with parameters $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.32, 0.34)$ in Figures 6.3 and 6.4 respectively.

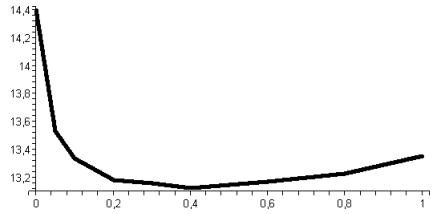

Figure 6.3: Utility function against θ for $B = 20$

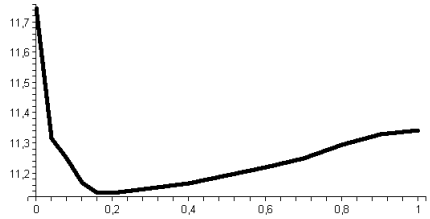

Figure 6.4: Utility function against θ for $B = 50$

We observe that the optimal points are located close to the minimum of $\pi^\theta(B)$, see Figures 6.1 and 6.2, as one may expect.

6.4 Conclusion

We analyzed the probability of overflow in the downstream node of the slowdown system in this chapter. Here we focused on the development of efficient IS-based simulation techniques for estimating this probability. The methods developed earlier are either of a poor quality (see Chapter 2), or very time consuming (see Chapter 4). We need to recalculate a new measure (by solving a system of two cubic equations) after each transition in the latter scheme. The IS scheme designed in this chapter requires to solve this system at most once. We then demonstrated how the techniques developed in this chapter help in tuning the design parameters involved, specifically focusing on selecting an appropriate value for the slowdown threshold.

Chapter 7

Multilevel splitting

7.1 Introduction

We dedicate this chapter to the study of multilevel splitting, the other main tool of rare event simulation (beside IS). At first sight, IS and MS are quite different. However, here we show that the design of both IS and MS schemes can be based on the same LD decay rate function.

In this chapter we provide a simple and asymptotically efficient MS scheme for estimating the probability that a discrete-state Markov process enters some rare set before a tabu set. The scheme can be seen as part of the class of asymptotically efficient MS schemes developed in the recent work [16]. In that paper it was shown how to construct an asymptotically efficient MS scheme for estimating the probability of first entrance to a rare set, when the decay rate of the probability is known for all starting states. Control-theoretic techniques were used to prove the results. The scheme uses splitting factors that may vary between levels, and are usually noninteger (which is then implemented by using a randomization procedure). In contrast, since we are interested in easy-to-implement (but still efficient) schemes, we use a fixed, pre-specified splitting factor R, to be used for all levels. Moreover, we accompany the scheme with a proof of asymptotic efficiency which is relatively easy, in the sense that it only uses probabilistic arguments and some simple bounds, thereby giving insight into why the scheme works so well.

We end the chapter with some supporting numerical results for the models of our interest, the tandem Jackson network and the slowdown network. We also compare these results with those for IS. In fact it turns out that MS can be a good alternative to IS for some parameter settings, but it is outperformed by IS for other settings.

7.2 Preliminaries

7.2.1 Model

We consider some Markov process $\{Q_k\}$ with state space D^B that has a finite number of possible jump directions v_i that are the same for each state x, with corresponding transition probabilities $\nu_i(x)$. We are interested in the probability that $\{Q_k\}$ hits the (rare) target set T^B before the 'tabu' set A^B, starting from some state $s \notin T^B \cup A^B$.

To clarify the situation we provide a simple queueing example, in which $\{Q_k\}$ is the joint-queue length after the k-th transition of the Markov chain that describes a tandem Jackson network. Then we may be interested in the event where, starting from some state, the queue of the second node reaches a level B before the entire system becomes empty. Then obviously, B is the 'rarity parameter' (in the sense that the event becomes more rare as we choose larger values for B), and we have $D^B = \mathbb{N}^2$; $T^B = \{x \in D^B : x_2 \geq B\}$ and $A^B = (0,0)$.

It is convenient to scale the process $\{Q_k\}$ with the parameter B, just as we did for the tandem Jackson network and the slowdown network. The scaled process $X_k = Q_k/B$ then makes jumps of size v_i/B, and has state space D, which is the scaled version[1] of D^B. The target and tabu sets T^B and A^B are scaled in the same manner, their scaled versions being given by T and A.

For such (disjoint) sets A and T and some state $s \notin A \cup T$, we define the stopping time

$$\tau_B^s = \inf\{k > 0 : X_k \in T,\ X_j \notin A\ \forall j = 1, \ldots, k-1,\ X_0 = s\},$$

where $\tau_B^s = \infty$ if $\{X_k\}$ hits the set A before T, see also (2.4). The probability of interest is now as follows:

$$p_B^s = \mathbb{P}\left(\tau_B^s < \infty\right),$$

see also (2.5). Importantly, we will assume that this probability decays exponentially in B, with decay rate

$$-\lim_{B\to\infty} B^{-1} \log p_B^s = \gamma(s). \tag{7.1}$$

In fact we will even assume that this convergence is uniform in s:

Assumption 7.1. *For any $\epsilon > 0$, some $B^* > 0$ exists such that for all $s \notin A \cup T$ we have $|B^{-1} \log p_B^s + \gamma(s)| < \epsilon$ for $B > B^*$.*

It is conceivable that under mild regularity conditions, a Laplace principle (and consequently Assumption 7.1) for random walks holds uniformly on compacts (where uniformity is with respect to the starting state s of the scaled process). We refer to Theorems 6.3.3 and 7.2.3 in [19] for cases with continuous and discontinuous statistics respectively.

[1] Formally the state space is a subset of D (namely a grid) for any finite B. As B grows large, the grid becomes denser, leading to D itself in the limit $B \to \infty$.

7.2.2 Scheme description

As we already stated in Section 1.4, to apply MS one first needs to define a family of nested sets $\{L_k\}$, $k = 0, \ldots, m$ such that $s \in \ell_0 = \partial L_0$ and

$$T = L_m \subset L_{m-1} \subset \ldots \subset L_1 \subset L_0 \subset D.$$

This family $\{L_k\}$ should be chosen such that every state that belongs to the boundary of L_k has similar importance, i.e., the probability of reaching T before A should be approximately equal for every state $x \in \ell_k = \partial L_k$. The sets L_i are typically chosen as the level sets of some function $f(x)$, which is called the *importance function*. Given this family, we start at the initial state s (which belongs to ℓ_0) with exactly R_0 particles. We continue to simulate each of them until they either cross level ℓ_1 or hit the tabu set A. All particles that end up in A are to be terminated without any replacement. Every particle that reaches level ℓ_1 is to be replaced by R_1 independent replicas. We continue to simulate all the (new) particles until they cross the next level ℓ_2 or hit the tabu set A, and so on. At stage k we start with some number of particles in level ℓ_{k-1} and simulate them until they reach ℓ_k or A. Then each particle that crossed ℓ_k is replaced by R_k independent copies, while all particles in A are terminated. We stop the procedure when the m-th level (i.e., the target set T) is reached. Now we construct the estimator as follows:

$$\hat{p}_B = \frac{X}{R_0 \cdot R_1 \cdot \ldots \cdot R_{m-1}}, \tag{7.2}$$

where X is the number of particles that eventually reaches the target set T before the tabu set A, see also Section 1.4.1. The estimate of p_B^s is constructed by averaging a number of independent replications of $\hat{p}_B$.

We now choose the importance function to be the logarithmic decay rate in (7.1), i.e., we chose $f(x) = \gamma(x)$ and describe the Multilevel Splitting scheme as follows:

1. Choose some integer R to be the splitting factor for all levels.
2. Compute the number of levels $n_B := \lfloor B\gamma(s)/\log R \rfloor$.
3. Define levels $\ell_k := \left(x \in D : \gamma(s) - \gamma(x) = \frac{k}{B} \log R \right)$, $k = 0, \ldots, n_B$.
4. Define $R' := \lfloor e^{B\gamma(s) - n_B \log R} \rfloor$, to be used as splitting factor at level n_B only.

(7.3)

The idea of the scheme is as follows: different states x in the same level have the same decay rate for their corresponding probabilities p_B^x, and the different levels are defined such that the total decay rate $\gamma(s)$ is 'evenly spread'; in other words, the distances between consecutive levels are equal in terms of decay rate. The corresponding probability of reaching the next level is roughly equal to $1/R$ due to the choice of n_B in step 2, so that on average only one particle out of R will reach the next level. Finally, since level n_B is in general not the boundary of the target set

T (due to the rounding in step 2), and the probability to reach T from this level is larger than $1/R$, we can do with the lower splitting factor R' at level n_B.

7.3 Asymptotic Efficiency

In this section we provide the proof of asymptotic efficiency of the MS scheme in (7.3). Having (7.1) in mind, we can rewrite the definition of asymptotic efficiency (1.11) as follows:

$$\limsup_{B\to\infty} B^{-1} \log\left(w(B)\mathbb{E}\hat{p}_B^2\right) \leq -2\gamma(s), \tag{7.4}$$

where $w(B)$ represents the expected computational effort per replication of $\hat{p}_B$. For the specific form of $w(B)$ we can make various choices. Here we assume that the required time effort increases linearly in the starting level. That is, we assume it takes $k+1$ time units to simulate a sample path of a particle starting from level ℓ_k, since with high probability it will reach A before ℓ_{k+1}, which takes more time when k is large; see also [29] for the motivation of this choice. The result is that we have

$$w(B) = \mathbb{E}\left[\sum_{k=0}^{n_B-1} R(k+1)\alpha(k) + R'(n_B+1)\alpha(n_B)\right], \tag{7.5}$$

where the random variable $\alpha(k)$ is the number of paths that have reached level ℓ_k, but have not reached ℓ_{k+1}.

From now on, in order to simplify the notation we omit the dependence on B in the notation n_B for the number of levels. Also we rewrite the estimator in (7.2) as follows:

$$\hat{p}_B = \frac{1}{R^n R'} \sum_{i=1}^{R^n R'} I_i. \tag{7.6}$$

Here we used that we have the same splitting factor R at each level, except the last one which is R'; furthermore the I_i are indicator random variables for each of the $R^n R'$ *possible* particles that may be simulated: $I_i = 1$ if the i-th particle hits the target set T before the tabu set A, and $I_i = 0$ otherwise. At first sight, it may seem that the number of particles needed to obtain this estimator grows exponentially in n, and consequently in B. However this is not the case, since we only need to simulate a few of all possible $R^n R'$ particles till the end. Suppose for instance that from the initial R particles only one reaches ℓ_1 before A, then the maximum number of possible particles to be simulated further is already reduced from $R^n R'$ to $R^{n-1} R'$.

In order to prove that (7.4) holds for our scheme, we first analyze the second moment of the estimator, for which we have the following result.

Lemma 7.1. *Under Assumption 7.1 the logarithm of the second moment of the estimator in (7.6) satisfies:*

$$\lim_{B\to\infty} \frac{1}{B} \log \mathbb{E}\hat{p}_B^2 \leq -2\gamma(s).$$

Proof. We first write

$$\frac{1}{B} \log \mathbb{E}\hat{p}_B^2 = \frac{1}{B} \log \frac{1}{R^{2n} R'^2} + \frac{1}{B} \log \mathbb{E} \left[\sum_{i=1}^{R^n R'} I_i \right]^2 \tag{7.7}$$

It is not difficult to see that the first term in the right-hand side of (7.7) has the following behavior

$$\lim_{B\to\infty} \frac{1}{B} \log \frac{1}{R^{2n} R'^2} = -2\gamma(s), \tag{7.8}$$

thanks to step 2 in (7.3). The second term is somewhat difficult to analyze.

$$\begin{aligned}
\frac{1}{B} \log \mathbb{E} \left[\sum_{i=1}^{R^n R'} I_i \right]^2 &= \frac{1}{B} \log \mathbb{E} \left(\sum_{i=1}^{R^n R'} I_i^2 + \sum_{i=1}^{R^n R'} \sum_{j=1, j\neq i}^{R^n R'} I_i I_j \right) \\
&= \frac{1}{B} \log \left(\sum_{i=1}^{R^n R'} \mathbb{E} I_i + \sum_{i=1}^{R^n R'} \sum_{j=1, j\neq i}^{R^n R'} \mathbb{E} I_i I_j \right) \\
&= \frac{1}{B} \log \left(R^n R' p_B^s + \sum_{i=1}^{R^n R'} \sum_{j=1, j\neq i}^{R^n R'} \mathbb{E}\left[I_i | I_j = 1 \right] p_B^s \right) \qquad (7.9) \\
&= \frac{1}{B} \log \left(R^n R' p_B^s \left[1 + \sum_{i=2}^{R^n R'} \mathbb{E}\left[I_i | I_1 = 1 \right] \right] \right) \\
&= \frac{1}{B} \log \left(R^n R' p_B^s \right) + \frac{1}{B} \log \left(1 + \sum_{i=2}^{R^n R'} \mathbb{E}\left[I_i | I_1 = 1 \right] \right).
\end{aligned}$$

In view of (7.1) and step 2 of (7.3) it is clear that the first term in the last line of (7.9) tends to 0 as B grows to infinity. For the second term, we condition on the level where particles 1 and k had their last common ancestor. Thus for some random

$S_i \in \ell_i$ and $S_0 \equiv s$ we have

$$\begin{aligned}
&\frac{1}{B}\log\left(1+\sum_{i=2}^{R^n R^{'}}\mathbb{E}\left[I_i|I_1=1\right]\right)\\
&= \frac{1}{B}\log\left(1+\sum_{i=0}^{n-1}(R-1)R^{n-i-1}R^{'}\mathbb{E}p_B^{S_i}+(R^{'}-1)\mathbb{E}p_B^{S_n}\right)\\
&\leq \frac{1}{B}\log\left(1+\sum_{i=0}^{n}(R-1)R^{n-i-1}R^{'}\mathbb{E}p_B^{S_i}\right) \qquad (7.10)\\
&\leq \frac{1}{B}\log\left(1+(n+1)e^{\epsilon B}\right)\leq\frac{1}{B}\log\left((2+n)e^{\epsilon B}\right)\\
&= \frac{1}{B}\log(2+n)+\epsilon.
\end{aligned}$$

The first inequality in (7.10) holds due to the fact that $R^{'}\leq R$; the second inequality in (7.10) follows from the combination of assumption (7.1) and the definition of the MS scheme (7.3): for all ϵ there exists B^* such that for all $s_i \in \ell_i$ we have that for all $B > B^*$ it holds that $R^{n_B-i}R^{'}p_B^{s_i} < e^{\epsilon B}$; finally the last inequality follows from the simple fact that ϵB is positive. Now it is not difficult to see that (7.10) and consequently (7.9) converges to zero when B grows to infinity. Taking into account expressions (7.7) and (7.8) one can obtain the following bound

$$\lim_{B\to\infty}\frac{1}{B}\log\mathbb{E}\hat{p}_B^2\leq -2\gamma(s).$$

□

Now let us analyze the expected computational effort.

Lemma 7.2. *When Assumption 7.1 is satisfied the logarithm of the expected workload (7.5) grows subexponentially in B, i.e.*

$$\lim_{B\to\infty}\frac{1}{B}\log w(B)=0.$$

Proof. We use similar arguments as in the proof of Lemma 7.1: for all ϵ there exists B^* such that for any i we have that for all $B > B^*$ it holds that

$$\mathbb{E}\alpha(i)=R^i p_B^{s,\ell_i}<e^{\epsilon B},$$

where p_B^{s,ℓ_i} is the probability that a sample path hits level ℓ_i, but does not hit level ℓ_{i+1}, starting from the initial state $s \in \ell_0$. Now for B large enough we obtain the following

$$w(B)\leq R\sum_{i=0}^{n-1}(i+1)R^i p_B^{s,\ell_i}+R^{'}(n+1)R^n p_B^{s,\ell_n}\leq R(n+1)^2 e^{\epsilon B}.$$

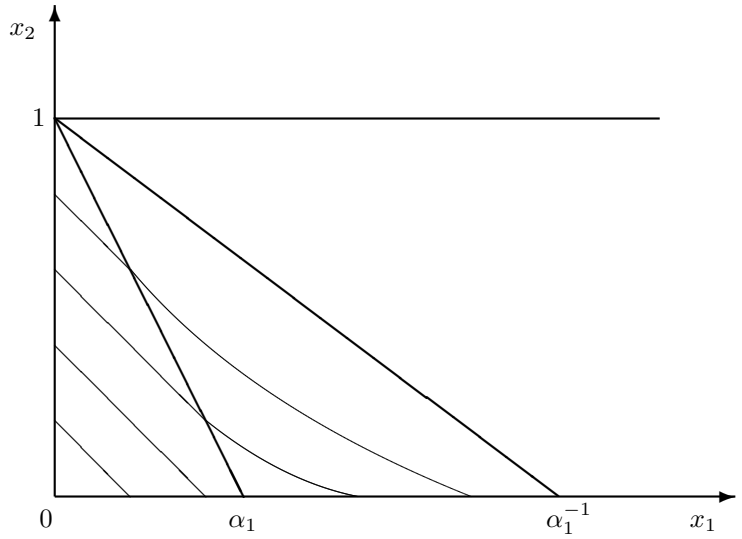

Figure 7.1: Splitting levels when $\mu_2 < \mu_1$

Taking the logarithm and sending B to infinity we obtain the following

$$\lim_{B\to\infty} \frac{1}{B} \log w(B) \leq \epsilon, \quad \forall \epsilon > 0,$$

which completes the proof. □

Combining the statements of Lemma 7.1 and Lemma 7.2 now immediately leads to the main result.

Theorem 7.3. *The Multilevel Splitting algorithm (7.3) is asymptotically efficient.*

7.4 Numerical Results

In this section we illustrate the efficiency of the MS scheme by applying it to the tandem Jackson and slowdown networks. As in the previous chapters we consider the rare event in which the second queue collects some large number of jobs B before the entire system empties.

First let us show that Assumption 7.1 holds for the tandem Jackson network. Consider the case where the second buffer is the bottleneck. We refer to Figure 7.1 for the shape of the splitting levels, see also Figure 3.1 and (3.5). In this case, we do not need Assumption 7.1 in full generality. We only require uniform convergence for states that belong to the splitting thresholds ℓ_i. This obviously holds since the union of all splitting levels is a subset of the compact $A_1 \cup A_2$.

We have a slightly more difficult situation when the first buffer is the bottleneck, since the union of all splitting levels is an unbounded set, see Figure 7.2. In order

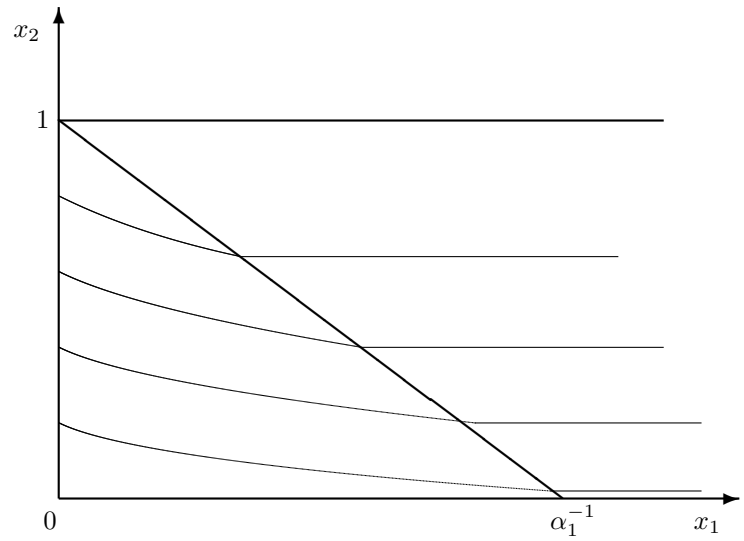

Figure 7.2: Splitting levels when $\mu_1 \leq \mu_2$

to deal with this problem, we truncate the state space in the following way: $[0,1) \times [0,K]$, where K is some constant. Keeping in mind the optimal path structure, see Chapter 3, one may conclude that the probability of our interest p_B^s stays almost the same for original and truncated state spaces, when K is large enough. The slowdown network can be treated in the same way.

Now let us provide numerical results. We estimate the probability p_B^s using MS scheme (7.3) and compare its performance with that of the (also asymptotically efficient) IS schemes developed in Chapters 5 and Chapter 6. In order to make a fair comparison we use the computer time from those IS simulations as a time budget for the MS scheme.

Let us start with the tandem Jackson network. In Table 7.1, Table 7.2 and Table 7.3 we present estimates of p_B^s for different starting states s and parameter settings, accompanied by their 95% confidence intervals and relative errors, as well as the relative errors obtained using the IS scheme from Chapter 5. (See also Table 5.1 and Table 5.2).

(λ, μ_1, μ_2)	B	p_B^s	RE(MS)	RE(IS)	time
$(0,0)$	20	$5.98 \cdot 10^{-2} \pm 2.57 \cdot 10^{-4}$	$2.19 \cdot 10^{-3}$	$3.12 \cdot 10^{-3}$	28
	50	$1.52 \cdot 10^{-3} \pm 1.72 \cdot 10^{-5}$	$5.77 \cdot 10^{-3}$	$3.94 \cdot 10^{-3}$	80
	100	$2.91 \cdot 10^{-6} \pm 5.80 \cdot 10^{-8}$	$10.1 \cdot 10^{-3}$	$4.74 \cdot 10^{-3}$	168

Table 7.1: $(\lambda, \mu_1, \mu_2) = (0.3, 0.36, 0.34)$, $R = 4$

Let us proceed with the numerical studies of the slowdown system, see Tables 7.4–7.7. We also refer to Tables 6.1– 6.5 for comparison.

(λ, μ_1, μ_2)	B	p_B^s	RE(MS)	RE(IS)	time
$(0.6B, 0)$	20	$1.99 \cdot 10^{-5} \pm 4.09 \cdot 10^{-7}$	$1.04 \cdot 10^{-2}$	$1.32 \cdot 10^{-3}$	7
	50	$3.19 \cdot 10^{-12} \pm 2.09 \cdot 10^{-13}$	$5.10 \cdot 10^{-2}$	$1.58 \cdot 10^{-3}$	18
	100	$1.87 \cdot 10^{-23} \pm 3.54 \cdot 10^{-24}$	$9.65 \cdot 10^{-2}$	$1.81 \cdot 10^{-3}$	35

Table 7.2: $(\lambda, \mu_1, \mu_2) = (0.1, 0.55, 0.35)$, $R = 4$

(λ, μ_1, μ_2)	B	p_B^s	RE(MS)	RE(IS)	time
$(0, 0)$	20	$3.29 \cdot 10^{-2} \pm 2.59 \cdot 10^{-4}$	$4.01 \cdot 10^{-3}$	$3.79 \cdot 10^{-2}$	28
	50	$7.00 \cdot 10^{-5} \pm 2.07 \cdot 10^{-6}$	$1.50 \cdot 10^{-2}$	$7.90 \cdot 10^{-2}$	84
	100	$1.92 \cdot 10^{-9} \pm 1.29 \cdot 10^{-10}$	$3.42 \cdot 10^{-2}$	$13.4 \cdot 10^{-2}$	189

Table 7.3: $(\lambda, \mu_1, \mu_2) = (0.3, 0.33, 0.37)$, $R = 8$

s	B	p_B^s	RE(MS)	RE(IS)	time
$(0, 0)$	20	$3.73 \cdot 10^{-7} \pm 4.91 \cdot 10^{-8}$	$6.58 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	1
	50	$1.45 \cdot 10^{-16} \pm 5.71 \cdot 10^{-17}$	$19.6 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	2
	100	$4.99 \cdot 10^{-32} \pm 2.58 \cdot 10^{-32}$	$25.8 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	6
$(0.7B, 0)$	20	$6.09 \cdot 10^{-3} \pm 2.03 \cdot 10^{-4}$	$1.70 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	1
	50	$3.40 \cdot 10^{-6} \pm 3.97 \cdot 10^{-7}$	$5.95 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	10
	100	$2.89 \cdot 10^{-11} \pm 9.72 \cdot 10^{-12}$	$17.1 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	37
$(1.5B, 0)$	20	$5.25 \cdot 10^{-1} \pm 4.37 \cdot 10^{-3}$	$0.42 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	1
	50	$1.35 \cdot 10^{-1} \pm 2.37 \cdot 10^{-3}$	$0.89 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	2
	100	$1.05 \cdot 10^{-2} \pm 2.31 \cdot 10^{-4}$	$1.12 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	21

Table 7.4: $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.1, 0.7, 0.15, 0.2)$, $R = 10$

s	B	p_B^s	RE(MS)	RE(IS)	time
$(0, 0)$	20	$5.68 \cdot 10^{-2} \pm 9.07 \cdot 10^{-4}$	$0.81 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	2
	50	$1.25 \cdot 10^{-3} \pm 4.45 \cdot 10^{-5}$	$1.40 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	18
	100	$1.81 \cdot 10^{-6} \pm 1.91 \cdot 10^{-7}$	$5.27 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	49
$(0.35B, 0)$	20	$2.02 \cdot 10^{-1} \pm 1.42 \cdot 10^{-3}$	$0.35 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	2
	50	$1.36 \cdot 10^{-2} \pm 3.71 \cdot 10^{-4}$	$1.39 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	9
	100	$1.29 \cdot 10^{-4} \pm 7.01 \cdot 10^{-6}$	$2.77 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	25

Table 7.5: $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.32, 0.34)$, $R = 4$

s	B	p_B^s	RE(MS)	RE(IS)	time
$(0, 0)$	20	$5.91 \cdot 10^{-2} \pm 9.35 \cdot 10^{-4}$	$0.80 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	2
	50	$1.46 \cdot 10^{-3} \pm 3.88 \cdot 10^{-5}$	$1.33 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	21
	100	$2.71 \cdot 10^{-6} \pm 7.17 \cdot 10^{-8}$	$1.42 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	121
$(0.35B, 0)$	20	$2.10 \cdot 10^{-1} \pm 2.52 \cdot 10^{-3}$	$0.54 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	2
	50	$1.54 \cdot 10^{-2} \pm 4.05 \cdot 10^{-4}$	$1.34 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	11
	100	$2.21 \cdot 10^{-4} \pm 8.89 \cdot 10^{-6}$	$2.05 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	35

Table 7.6: $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.3, 0.36, 0.35, 0.34)$, $R = 4$

s	B	p_B^s	RE(MS)	RE(IS)	time
$(0,0)$	20	$1.14 \cdot 10^{-4} \pm 6.20 \cdot 10^{-6}$	$2.71 \cdot 10^{-2}$	$5.00 \cdot 10^{-2}$	2
	50	$4.11 \cdot 10^{-11} \pm 7.15 \cdot 10^{-12}$	$8.69 \cdot 10^{-2}$	$5.00 \cdot 10^{-2}$	7
	100	$7.80 \cdot 10^{-22} \pm 2.81 \cdot 10^{-22}$	$18.0 \cdot 10^{-2}$	$5.00 \cdot 10^{-2}$	42
$(0.35B,0)$	20	$6.35 \cdot 10^{-4} \pm 5.24 \cdot 10^{-5}$	$4.21 \cdot 10^{-2}$	$5.00 \cdot 10^{-2}$	1
	50	$2.61 \cdot 10^{-9} \pm 3.63 \cdot 10^{-10}$	$7.09 \cdot 10^{-2}$	$5.00 \cdot 10^{-2}$	5
	100	$4.61 \cdot 10^{-18} \pm 9.22 \cdot 10^{-19}$	$10.2 \cdot 10^{-2}$	$5.00 \cdot 10^{-2}$	25

Table 7.7: $(\lambda, \mu_1, \mu_1^+, \mu_2) = (0.25, 0.35, 0.28, 0.4)$, $R = 10$

Clearly, the MS scheme (7.3) provides good results. Sometimes the relative error may be even lower than the one obtained via IS, see e.g. Table 7.3. However, in some cases the relative error of MS is quite high, see e.g. Table 7.4. This undesirable performance can be explained as follows. When the parameters of the network (λ, μ_1 (μ_1^+), μ_2) are clearly distinctive, as is the case in Table 7.4, then the IS scheme performs well, and consequently requires a relatively small amount of time to obtain a good estimate. On the other hand, this is the toughest case for MS since the queue-length process has a strong drift towards the origin. The relative error may also be decreased by finetuning the splitting factor R.

7.5 Conclusions

In this chapter we designed an asymptotically efficient MS scheme for estimating the probability of first entry to some rare set. The probabilities of overflow in the downstream buffer of a tandem Jackson or slowdown network are particular cases.

As we see from numerical results for these networks, MS can be a good alternative to IS, especially when the system has high loads (in both nodes), or when the rarity parameter B is not extremely large, i.e., when the event of our interest is not so rare. Even when the server loads are low and MS is obviously outperformed by IS, MS may be preferred, since it is conceptually easier than IS. The main reason for this is that we do not need to consider the boundaries and define mollifications around them, as we need to do in IS. This is a big advantage over IS, especially when we want to simulate large networks (which we did not consider in this thesis).

A disadvantage of MS is that it is less straightforward to implement than the IS schemes from Chapters 5 and 6, in the sense that for MS we need to calculate the decay rate $\gamma(x)$ for every state in $\bar{D}$. IS requires knowledge of the decay rate for a single state, namely the starting, state.

Bibliography

[1] V. Anantharam. The optimal buffer allocation problem. *IEEE Transactions on Information Theory*, 35:721–725, 1989.

[2] 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.

[3] S. Asmussen. *Applied probability and queues*. Springer, New York, 2003.

[4] S. Asmussen and P.W. Glynn. *Stochastic simulation: algorithms and analysis*. Springer-Verlag, New York, 2007.

[5] J.H. Blanchet and P.W. Glynn. Efficient rare-event simulation for the maximum of heavy-tailed random walks. *Annals of Applied Probability*, 18(4):1351–1378, 2008.

[6] J.H. Blanchet and J.C. Liu. State-dependent importance sampling for regularly varying random walks. *Advances in Applied Probability*, 40:1104–1128, 2008.

[7] J.A. Bucklew. *Introduction to Rare Event Simulation*. Springer, New York, 2004.

[8] P.J. Burke. The output of a queuing system. *Operations Research*, 4(6):699–704, 1956.

[9] J.W. Cohen. *The single server queue*. North Holland, Amsterdam, 1969.

[10] P.T. de Boer. *Analysis and efficient simulation of queueing models of telecommunication systems*. PhD thesis, University of Twente, 2000.

[11] 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.

[12] P.T. de Boer. Some observations on importance sampling and Restart. In *Proceedings of RESIM 2006*, pages 217–224, Bamberg, Germany, 2006.

[13] P.T. de Boer, P. L'Ecuyer, G. Rubino, and B. Tuffin. Estimating the probability of a rare event over a finite time horizon. In *Proceedings of the 2007 Winter Simulation Conference*, pages 403–411, Washington, USA, 2007.

[14] 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*, pages 646–655, Orlando, USA, 2000.

[15] 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.

[16] T. Dean and P. Dupuis. Splitting for rare event simulation: a large deviations approach to design and analysis. *Preprint*, 2008.

[17] A. Dembo and O. Zeitouni. *Large deviation techniques and applications*. Springer, New York, 1998.

[18] Bureau d'enquêtes et d'analyses pour la sécurité de l'aviation civile. Accident on 25 July 2000 at La Patte d'Oie in Gonesse (95) to the Concorde registered F-BTSC operated by Air France. *URL: http://www.bea-fr.org/docspa/2000/f-sc000725a/pdf/f-sc000725a.pdf*.

[19] P. Dupuis and R.S. Ellis. *A weak convergence approach to the theory of Large deviations*. Wiley, New York, 1997.

[20] 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.

[21] 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.

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

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

[24] P. Dupuis and H. Wang. Importance sampling for Jackson network. *Preprint*, 2008.

[25] R. Ellis. *Entropy, large deviations, and statistical mechanics*. Springer, New York, 1985.

[26] M.R. Frater and B.D.O. Anderson. Fast estimation of the statistics of excessive backlogs. *Australian Telecommunications Research*, 23(1):49–55, 1989.

[27] M.R. Frater, T.M. Lenon, and B.D.O. Anderson. Optimally efficient estimation of the statistics of rare events in queueing networks. *IEEE Transactions on Automatic Control*, 36:1395–1405, 1991.

[28] M.J.J. Garvels. *The splitting method in rare event simulation.* PhD thesis, University of Twente, 2000.

[29] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. Multilevel splitting for estimating rare event probabilities. *IBM Research Report*, RC 20478, 1996.

[30] P. Glasserman, P. Heidelberger, P. Shahabuddin, and T. Zajic. A large deviations perspective on the efficiency of multilevel splitting. *IEE Transactions on Automatic Control*, 43(12):1666–1679, 1998.

[31] 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.

[32] P.W. Glynn and W. Whitt. An asymptotic efficiency of simulation estimators. *Operations Research*, 40(3):505–520, 1992.

[33] J. Hammersley and D. Handscomb. *Monte Carlo Methods.* Methuen, London, 1965.

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

[35] P. Heidelberger, P. Shahabuddin, and V.F. Nicola. Bounded relative error in estimating transient measures of highly dependable non-Markovian systems. *ACM Transactions on Modeling and Computer Simulation*, 4(2):137–164, 1994.

[36] J.R. Jackson. Networks of waiting lines. *Operations Research*, 4(4):518–521, 1957.

[37] S. Juneja and V.F. Nicola. Efficient simulation of buffer overflow probabilities in Jackson networks with feedback. *ACM Transactions on Modeling and Computer Simulation*, 15(4):281–315, 2005.

[38] H. Kahn and T.E. Harris. Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Mathematics Series*, 12:27–30, 1951.

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

[40] D.G. Kendall. Stochastic processing occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Annals of Mathematical Statistics*, 24:338–354, 1953.

[41] L. Kleinrock. *Queueing Systems*. Wiley, New York, 1976.

[42] I.N. Kovalenko. Rare events in queueing systems – A survey. *Queueing Systems: Theory and Applications*, 16(1-2):1–49, 1994.

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

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

[45] A. Lagnoux-Renaudie. A two-step branching splitting model under cost constraint for rare event analysis. *Journal of Applied Probability*, 46(2):429–452, 2009.

[46] G. Latouche and V. Ramaswami. *Introduction to matrix analytic methods in stochastic modelling*. SIAM, Philadelphia, USA, 1999.

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

[48] 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.

[49] 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.

[50] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. Simple and efficient importance sampling scheme for a tandem queue with server slow-down. In *Proceedings of RESIM 2008*, pages 38–50, Rennes, France, 2008.

[51] 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/*.

[52] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. Backpressure-based control protocols: design and computational aspects. In *Proceedings of ITC 21*, Paris, France, 2009.

[53] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. An efficient multi-level splitting scheme. In *Proceedings of the Sixth St. Petersburg Workshop on Simulation*, St. Petersburg, Russia, 2009.

[54] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. Rare-event simulation for tandem queues: a simple and efficient importance sampling scheme. In *Proceedings of NET-COOP*, pages 107–120, Eindhoven, The Netherlands, 2009.

[55] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. State-dependent importance sampling for a Jackson tandem network. *Accepted for publication in ACM Transactions on Modeling and Computer Simulation*, 2009. See also Memorandum 1867, Dept. of Applied Mathematics, University of Twente, *URL: http://eprints.eemcs.utwente.nl/12734/*.

[56] M.K. Nakayama. General conditions for bounded relative error in simulations of highly reliable Markovian systems. *Advances in Applied Probability*, pages 687–727, 1996.

[57] V.F. Nicola and T.S. Zaburnenko. Importance sampling simulation of population overflow in two-node tandem network. In *Proceedings of the 2nd International Conference on the Quantative Evaluation of Systems*, pages 347–354, Torino, Italy, 2005.

[58] W. Noureddine and F. Tobagi. Selective back-pressure in switched Ethernet LANs. In *Proceedings of Global TelecommunicationsConference*, volume 2, pages 1256–1263, Rio de Janeiro, Brazil, 1999.

[59] 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.

[60] G. Rubino and B. Tuffin. *Rare Event Simulation using Monte Carlo Methods*. Wiley, New York, 2009.

[61] R.Y. Rubinstein. Rare event simulation via cross-entropy and importance sampling. In *Proceedings of RESIM 1999*, pages 1–17, Pisa, Italy, 1999.

[62] R.Y. Rubinstein. The cross-entropy method and rare events for maximal cut and bipartition problems. *ACM Transactions on Modeling and Computer Simulation*, 12(1):27–53, 2002.

[63] 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.

[64] W. Sandmann. Fast simulation of excessive population size in tandem Jackson networks. In *Proceedings of the 12th Annual IEEE International Symposium MASCOTS'04*, pages 347–354, 2004.

[65] P. Shahabuddin. Rare event simulation in stochastic models. In *Proceedings of the 1995 Winter Simulation Conference*, pages 178–185, Arlington, USA, 1995.

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

[67] R.L. Tweedie. Operator-geometric stationary distributions for Markov chains, with applications to queueing models. *Advances in Applied Probability*, pages 368–391, 1982.

[68] 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.

[69] S. Varadhan. *Large Deviations and Applications.* SIAM, Philadelphia, USA, 1984.

[70] J. Villén-Altamirano. Importance functions for Restart simulation of general Jackson networks. In *Proceedings of RESIM 2006*, pages 184–196, Bamberg, Germany, 2006.

[71] J. Villén-Altamirano. Rare event Restart simulation of two-stage networks. *European Journal of Operational Research*, 179(1):148–159, 2007.

[72] M. Villén-Altamirano, J. Gamo A. Martinez-Marron, and F. Fernandez-Cuesta. Enhancement of the accelerated simulation method Restart by concidering multiple thresholds. In *Proceedings of ITC 14*, pages 797–810, Antibes - Juan-les-Pins, France, 1994.

[73] M. Villén-Altamirano and J. Villén-Altamirano. Restart: a method for accelerating rare event simulations. In *Proceedings of ITC 13*, pages 71–76, Copenhagen, Denmark, 1991.

[74] M. Villén-Altamirano and J. Villén-Altamirano. Analysis of Restart simulation: theoretical basis and sensitivity study. *European Transaction on Telecommunication*, 13(4):373–386, 2002.

[75] M. Villén-Altamirano and J. Villén-Altamirano. On the efficiency of Restart for multidimensional state systems. *ACM Transactions on Modeling and Computer Simulation*, 16(3):251–279, 2006.

[76] T.S. Zaburnenko. *Efficient heuristics for simulating rare events in queueing networks.* PhD thesis, University of Twente, 2008.

[77] T.S. Zaburnenko and V.F. Nicola. Efficient heuristics for simulating population overflow in tandem networks. In *Proceedings of the Fifth St. Petersburg Workshop on Simulation*, pages 755–764, St. Petersburg, Russia, 2007.

Summary

This monograph focuses on rare events. Even though they are extremely unlikely, they can still occur and then could have significant consequences.

We mainly consider rare events in queueing networks. More precisely, we are interested in the probability of collecting some large number of jobs in the downstream queue of a two-node tandem network. We consider the Jackson network case, as well as a generalization, the so-called slowdown network. In practice these models can be used to model overflows in telecommunication networks. We chose these networks as a first step in developing a methodology that can be extended to more general networks.

We investigate rare events from two different sides. On the one hand we are interested in the *nature* of the event, i.e., how the event 'builds up'. At first we identify the structure of a specific path to overflow, which plays the role of our candidate for the *most probable* trajectory to overflow. We use some simple, but powerful *large deviations* based heuristics to this end. The shape of the trajectory crucially depends on both the starting state and the system parameters. We then provide a rigorous proof that this trajectory is indeed the most probable path to overflow. Thus our method combines simplicity (as it is easy to identify) and precision (as it is backed up by rigorous mathematical support).

On the other hand our ultimate goal is to design accurate and efficient techniques to estimate the probability of our interest; in particular we aim for techniques that are *asymptotically efficient*, which effectively means that the number of replications needed to obtain an estimator with predetermined relative error grows sub-exponentially when the probability of interest decays exponentially. We present several *importance sampling* schemes based on the large deviations results. We begin with naïve, state-independent algorithms and end up with a family of simple and efficient state-dependent schemes. We also develop a *multilevel splitting* scheme, which turns out to be efficient for a wider class of processes. Strengths and weaknesses of the importance sampling schemes and multilevel splitting schemes are also discussed in this work.

We start in Chapter 2 by identifying the most likely path to overflow. Here we develop a family of state-independent importance sampling schemes that are mimicking

the most probable path 'on average', i.e., without any adaptation. These schemes are very simple to implement. However, in general they are not asymptotically efficient.

In Chapter 3 and Chapter 4 we focus on the development of state-dependent importance sampling schemes for these two networks. We generalize the procedure described in Chapter 2 to identify the most likely path to overflow starting from *any* state. Based on this we design a family of importance sampling schemes for any initial state and prove its asymptotic efficiency. These schemes, in contrast to the ones from Chapter 2, are state-dependent, i.e., they require the recalculation of a new measure after each transition. We stress that the discontinuity of the measure around the slowdown threshold was an additional complication in the analysis of the schemes in Chapter 4. These asymptotically efficient schemes have a drawback, namely that the computational efforts required to determine the importance sampling algorithm are excessively large. In the following two chapters we reduce the complexity of the schemes while retaining their efficiency.

Thus, in Chapter 5 and Chapter 6 we design a family of simpler state-dependent important sampling schemes based on the results of the previous two chapters. The main idea is to reduce dependence of the new measure on the current state of the process without losing asymptotic efficiency of the schemes. As a result we have a family of simplified schemes which are asymptotically efficient for all parameter values for the two-node Jackson network. For the slowdown model, the same holds for the major part of starting states, while for the rest of the starting states we proved efficiency under some mild condition. The accuracy of these schemes is comparable to those from Chapter 3 and Chapter 4. At the same time they are almost as easy to implement as the state-independent schemes in Chapter 2.

Chapter 7 is dedicated to the multilevel splitting method. Here we design a family of asymptotically efficient multilevel splitting schemes for a rather broad class of models, which includes both networks of our interest. The proof of asymptotic efficiency relies on some elementary combinatorics and a number of simple facts from the theory of branching processes. Applying the proposed scheme to both networks of our interest, our numerical findings show high accuracy and time efficiency, comparable to what we obtained via importance sampling.

About the author

Denis Miretskiy was born on the 17$^{\text{th}}$ of March 1982 in Volzhsky, Russia. He studied Applied Mathematics at Volgograd State University, Russia and received his Specialist degree in June 2004. Later that year he continued his study in Utrecht University, The Netherlands. He received his M.Sc. degree in Applied Mathematics in June 2005. Subsequently, he became a Ph.D. student at the Stochastic Operations Research group at the University of Twente, The Netherlands. Denis defends his thesis on the 12$^{\text{th}}$ of November 2009.