Rare event simulation for non-Markovian tandem queues

Anne Buijsrogge

RARE EVENT SIMULATION FOR NON-MARKOVIAN TANDEM QUEUES

Anne Buijsrogge

Graduation committee

Chairman & secretary:	Prof. dr. J.N. Kok
Promotors:	Prof. dr. R.J. Boucherie
	Prof. dr. ir. B.R.H.M. Haverkort
Co-promotors:	Dr. ir. P.T. de Boer
	Dr. ir. W.R.W. Scheinhardt

Members:

Prof. dr. J.L. van den Berg	University of Twente
Prof. dr. P. Dupuis	Brown University
Prof. dr. M.N.M. van Lieshout	University of Twente
Prof. dr. M.R.H. Mandjes	University of Amsterdam
Dr. A.A.N. Ridder	Vrije Universiteit Amsterdam

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

DSI Ph.D. Thesis Series No. 19-008
Digital Society Institute
P.O. Box 217
7500 AE Enschede, The Netherlands

This work is supported by the Netherlands Organization for Scientific Research (NWO), project number 613.001.105.

ISBN: 978-90-365-4788-8
ISSN: 2589-7721 (DSI Ph.D. Thesis Series No. 19-008)
DOI: 10.3990/1.9789036547888
https://doi.org/10.3990/1.9789036547888

Typeset in LaTeX. Printed by Ipskamp Printing, Enschede, The Netherlands

RARE EVENT SIMULATION FOR NON-MARKOVIAN TANDEM QUEUES

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof. dr. T.T.M. Palstra,
on account of the decision of the Doctorate Board,
to be publicly defended
on Friday the 21st of June 2019 at 16.45 hrs

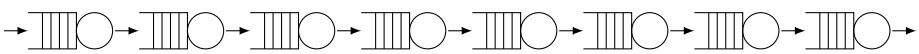
by

Anne Buijsrogge

born on the 12th of April 1991
in Smallingerland, the Netherlands

This dissertation has been approved by:

Prof.dr. R.J. Boucherie
Prof.dr.ir. B.R.H.M. Haverkort
Dr.ir. P.T. de Boer
Dr.ir. W.R.W. Scheinhardt



Acknowledgements

Dit proefschrift was nooit tot stand gekomen zonder de hulp en onvoorwaardelijke steun van verschillende personen. Een aantal hiervan wil ik graag in het bijzonder bedanken.

Ilze and Nelly, in my last year of my graduation I had the pleasure to work with you on my internship and graduation project. Your enthusiasm and passion for research was one of the reasons for me to pursue a PhD degree.

Pieter-Tjerk en Werner, met zijn drieën zijn we een team van perfectionisten, wat (soms tot mijn grote frustratie) leidt tot veel iteraties van papers en ook dit proefschrift. Bedankt voor alle waardevolle feedback die jullie mij hebben gegeven. Ik heb onwijs veel van jullie geleerd! Boudewijn en Richard, bedankt dat ook jullie deur altijd voor mij open stond.

Paul, thank you very much for the opportunity to visit Brown University and for all your suggestions that helped to improve this thesis. De rest van mijn commissie, Ad, Hans, Marie-Colette en Michel, wil ik ook bedanken voor de tijd die zij hebben geïnvesteerd in het lezen van het proefschrift en voor het bijwonen van mijn promotie.

Karol and Simone, thank you for your contributions that led to Chapters 2 and 3 respectively of this thesis. Michael and Paul, thank you for the collaboration that led to Chapter 8 of this thesis.

Tijdens mijn promotieonderzoek heb ik deel uit mogen maken van twee leuke vakgroepen: DACS en MOR. Ik wil iedereen bedanken voor de gezellige momenten bij de koffiemachine, de leuke lunchpauzes en alle andere leuke en gezellige momenten die hier niet onder vallen. Corine, het zal je niet ontgaan zijn dat je me wederom voor bent. Ik vind het nog steeds bijzonder dat we tien jaar lang min of meer dezelfde keuzes hebben gemaakt. Bedankt voor het vele theeleuten, ik mis het nu al!

Mijn vrienden binnen Arashi waarmee ik de afgelopen jaren op de mat heb gestaan wil ik bedanken voor het (soms letterlijk) opvangen van mijn frustraties rondom mijn onderzoek. Antoni en Daphne, ik vind het ontzettend leuk dat jullie mijn paranimfen willen zijn.

Papa, ik zal nooit vergeten hoe trots jij was toen ik begon met promoveren. Het doet ontzettend veel pijn dat je het resultaat nooit kan zien. Jelle en mama, bedankt voor jullie steun en onvoorwaardelijke liefde in de afgelopen jaren.

Tot slot, Tim. Bedankt voor je vertrouwen in mij. Ik kijk er naar uit om samen nog veel mooie plekken op deze wereld te ontdekken.

Anne
Enschede, mei 2019

Contents

CHAPTER 1

Introduction

Rare events hardly occur, as their name suggests, but *when* they occur, their impact on society can be huge and therefore it is very important to study such events. Plane crashes, tsunamis, large scale power outages and breakthrough of dikes are all well-known examples of disastrous rare events. Of course, many other examples exist.

Models of these events are often very complex so that they are hard, or nearly impossible, to study analytically. An alternative to an analytical study is computer simulation. While computer simulation is a solid way to estimate certain quantities of interest, especially for rare events it may take a very long time in order to obtain these.

In this thesis, we study the simulation of several rare events. In particular, we are interested in rare events in queueing systems, which are often used to model practical situations, including in logistics or telecommunication systems. For instance, the event in which some storage buffer becomes too full may lead to expensive loss of material, while an overflowing data buffer may lead to loss of important information. Even though such events may have a very low probability of occurring, their impact on the performance of the system as a whole can be profound, which explains why it may be important to obtain accurate estimates of such probabilities. Many other examples exist, but the ones we mentioned here can be modeled as overflow in a queueing system, which is the main topic of this thesis.

Besides queueing systems, we also consider a different class of stochastic models with a different rare event of interest. The motivation for considering these models is to show that, for these particular models, one method for rare event simulation works better than another method for rare event simulation.

1.1 Simulation of rare events

In this thesis, we are mainly interested in estimating the probability to reach a large number of customers in a queueing network during a busy cycle of the system using discrete event simulation, that is, we start with one customer in the system and estimate the probability to reach a large number of customers, say N, before the system is empty. This is also referred to as reaching the target

set, in our case having N customers in the system, before reaching the taboo set, which in our case is an empty system. We will address the drawbacks of computer simulation for rare events by means of the simplest queueing model that exists: the so-called $M|M|1$ queue.

Example 1.1. *In Figure 1.1, the so-called state transition diagram of an $M|M|1$ queue is shown. At such a queue, customers arrive according to a Poisson process with rate λ and service times are exponentially distributed with rate μ. Without loss of generality we can assume that $\lambda + \mu = 1$, and hence λ and μ are probabilities. As a result, the depicted state transition diagram can both represent a DTMC or a CTMC.*

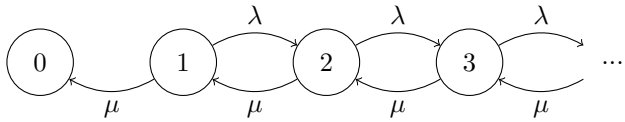


Figure 1.1 The state transition diagram of an $M|M|1$ queue.

The states represent the number of customers in the queue, the transitions represent arrivals, that happen with probability λ, and departures, that happen with probability μ. We start with one customer in the system, so we start in state 1. When there are zero customers in the system, we have reached the taboo set and so there is no arrow from state 0 to state 1. Another assumption that we make is $\rho = \lambda/\mu < 1$, otherwise the number of customers in the system can grow arbitrary large and the event of interest would not be rare. It is known that for the probability of interest, denoted by p_N, we have

$$p_N = \frac{(1-\rho)\rho^{N-1}}{1-\rho^N}, \tag{1.1}$$

see for example [11]. When N gets large, we see that p_N gets very small, for example, if $\lambda = 0.1$ and $\mu = 0.9$, then we find $p_{100} = 3.3465 \cdot 10^{-96}$.

Suppose that we want to estimate p_N by using simulation and suppose we perform S simulation trials to estimate this probability. The estimator of p_N will be denoted by $\hat{p}_N$. For each simulation i we let $I(\mathcal{P}_i) = 1$ if we have reached the target set before the taboo set and $I(\mathcal{P}_i) = 0$ otherwise. It will become clear in Section 1.2 why we use the notation $\mathcal{P}_i$ instead of i. Thus, an unbiased estimator for p_N is

$$\hat{p}_N = \frac{1}{S}\sum_{i=1}^{S} I(\mathcal{P}_i). \tag{1.2}$$

In Example 1.1, we see that for certain values of λ and μ the probability p_{100} is in the order of 10^{-96}. In order for this event to occur *once*, we have to perform roughly 10^{96} replications of the experiment. *Even* if we have a computer at

which we can perform 10^6 replications per second, it will take roughly 10^{85} years to perform 10^{96} replications. After waiting for years, we only expect our event to occur *once*. However, in order to get some meaningful confidence intervals we need more occurrences of our event of interest.

Luckily, for an $M|M|1$ queue the exact result is known and so there is no need to estimate this quantity by simulation. However, for a $G|G|1$ queue, which is a more general queueing system than an $M|M|1$ queue, no analytical result exists and thus computer simulation is necessary when estimating various types of quantities. In addition, also for networks of $G|G|1$ queues this is true. The generalization from $M|M|1$ queues to $G|G|1$ queues is important, because queueing networks in practice usually do not consist of $M|M|1$ queues.

In Section 1.2 and 1.3 we will explain how we can still use computer simulation in order to estimate quantities of interest, without waiting for around 10^{85} years, and we describe the challenges that come with these methods.

1.2 Importance sampling

The first method to perform rare event simulation that we discuss is importance sampling. In this method, the rare event is made less rare by changing the underlying probability distributions. The main challenge for importance sampling is to determine *how* to change the underlying probability distributions. While doing so, an important aspect is to consider the variance of the estimator and to make sure that the so-called *relative error*, formally defined as the standard deviation of the estimator divided by the mean of the estimator, does not increase too fast when N gets larger, see Section 1.2.1. The construction of such a change of measure is far from trivial and is one of the main challenges that is dealt with in this thesis. We explain importance sampling in more detail by considering Example 1.1.

Example 1.1. *(Continued) For the $M|M|1$ queue, it is known how to change the underlying probability distributions such that the relative error does not increase too fast. We get the new system with transition probabilities as in Figure 1.2. In fact, λ and μ are interchanged compared to the original system.*

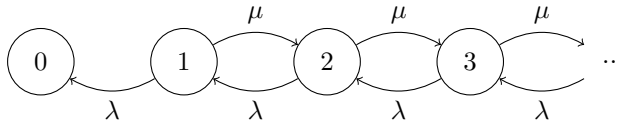


Figure 1.2 The state transition diagram of an $M|M|1$ queue using importance sampling.

Intuitively, it is easier to reach level N before reaching an empty system when simulating this new system, since $\mu > \lambda$. Furthermore, due to the same reasons, it is harder to end up in an empty system. To estimate our probability of interest

we cannot use (1.2) *since the probability of reaching level N before reaching an empty system in the new system is different from the similar probability in the original system. However, we know how likely it was to have an arrival or a departure in the old system compared to the new system, that is, when an arrival occurs in the new system we know that the event was actually $\frac{\lambda}{\mu}$ times less likely, whereas when a departure occurs in the new system we know that this event was actually $\frac{\mu}{\lambda}$ times more likely.*

Thus, we know the so-called likelihood ratio *of each event, an arriving customer or a departing customer, that is occurring. When we now multiply all likelihood ratios for each event, thus each transition until we reach the target set or the taboo set, then we keep track of all the likelihood ratios along a* path *during a busy cycle of the system. The likelihood ratio tells us how likely it is to take this particular path in the original system, compared to the same path in the changed system.*

In fact, we have now derived a different way to get an unbiased estimator $\hat{p}_N$ for the event of interest, by using importance sampling. We have

$$\hat{p}_N = \frac{1}{S}\sum_{i=1}^{S} L(\mathcal{P}_i)I(\mathcal{P}_i), \tag{1.3}$$

where $L(\mathcal{P}_i)$ is the likelihood ratio of a path $\mathcal{P}_i$ in simulation i, which can be formally defined as

$$L(\mathcal{P}_i) = \frac{d\mathbb{P}}{d\mathbb{Q}}(\mathcal{P}_i),$$

where $\frac{d\mathbb{P}}{d\mathbb{Q}}$ denotes the Radon-Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{Q}$, with $\mathbb{P}$ being absolutely continuous with respect to $\mathbb{Q}$. We can think of the Radon-Nikodym derivative as the ratio of probabilities (or probability densities) for a path $\mathcal{P}_i$ in simulation i under the original measure and the same, but under the new measure. The change from the original system in Figure 1.1 to the new system in Figure 1.2 is called a *change of measure*. Note that if $L(\mathcal{P}_i) = 1$ for all i, and thus we are simulating our original system, we find that (1.3) reduces to (1.2).

1.2.1 Performance measures

Though Example 1.1 intuitively suggests that we now have a faster simulation scheme, we need to show more formally that the simulation is efficient in some sense. Before we are able to do so, we make some remarks, define the relative error formally and show how ordinary simulation performs.

Let $\mathbb{E}^{\mathbb{Q}}[\cdot]$ denote the expected value under the new measure $\mathbb{Q}$. First of all,

we show that $\hat{p}_N$ is indeed unbiased, since by (1.3) we find

$$\begin{aligned}\mathbb{E}^{\mathbb{Q}}\left[\hat{p}_N\right] &= \frac{1}{S}\mathbb{E}^{\mathbb{Q}}\left[\sum_{i=1}^{S} L(\mathcal{P}_i)I(\mathcal{P}_i)\right] = \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})I(\mathcal{P})\right] \\ &= \int \frac{d\mathbb{P}}{d\mathbb{Q}}(\mathcal{P})I(P)d\mathbb{Q}(\mathcal{P}) = \int I(P)d\mathbb{P}(\mathcal{P}) = \mathbb{E}\left[I(\mathcal{P})\right] = p_N,\end{aligned}$$

by definition of $L(\mathcal{P})$. In the second equality we omitted the subscript i, since there is no explicit dependence on the simulation number anymore. The variance of the random variable $L(\mathcal{P})I(\mathcal{P})$ is

$$\sigma_N^2 = \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right] - \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})I(\mathcal{P})\right]^2,$$

and we remark that the unbiased estimator for the variance is

$$\hat{\sigma}_N^2 = \frac{1}{S-1}\sum_{i=1}^{S}(L(\mathcal{P}_i)I(\mathcal{P}_i) - \hat{p}_N)^2 = \frac{S}{S-1}\left(\frac{\sum_{i=1}^{S} L(\mathcal{P}_i)^2 I(\mathcal{P}_i)^2}{S} - \hat{p}_N^2\right).$$

Recall that the relative error is defined as the standard deviation of the estimator divided by the estimator itself, and note that the standard deviation of $L(\mathcal{P})I(\mathcal{P})$ equals σ_N. Thus, we have

$$RE = \frac{\sqrt{\mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right] - p_N^2}}{p_N}. \tag{1.4}$$

First, consider an estimator obtained by ordinary simulation, also called Monte Carlo simulation. In that latter case, the likelihood ratios equal 1, and so we have similar to (1.4) that the estimator of the relative error, denoted by $\hat{RE}$, equals

$$\hat{RE} = \frac{\sqrt{\frac{S}{S-1}\left(\frac{\sum_{i=1}^{S} L(\mathcal{P}_i)^2 I(\mathcal{P}_i)^2}{S} - \hat{p}_N^2\right)}}{\sqrt{S}\hat{p}_N} = \frac{\sqrt{\hat{p}_N - \hat{p}_N^2}}{\sqrt{S-1}\hat{p}_N} \approx \frac{1}{\sqrt{S-1}\sqrt{\hat{p}_N}}.$$

If we assume that p_N decays exponentially in N, hence so does the estimator $\hat{p}_N$, then we see that the estimate of the relative error grows exponentially fast with the overflow level N and so the number of simulations that we need in order to get a decent estimate of our probability of interest grows exponentially fast in N. When designing a method for rare event simulation, our goal is to ensure that the number of simulations that is required to get a decent estimate of the probability of interest grows *less* than exponentially fast in N.

In order to do so, we define what we call an *asymptotically efficient* estimator. We start with the mathematical definition, after which we go into some more detailed and intuitive explanation. The explanation on the efficiency of the importance sampling simulation that we provide below, is very similar to the one in [30].

Definition 1.2. *The estimator for p_N, here defined as $L(\mathcal{P})I(\mathcal{P})$, is asymptotically efficient – also called asymptotically optimal – if and only if*

$$\liminf_{N\to\infty} \frac{\log \mathbb{E}\left[L(\mathcal{P})I(\mathcal{P})\right]}{\log p_N} = \liminf_{N\to\infty} \frac{\log \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right]}{\log p_N} \geq 2.$$

Remark 1.3. *By Jensen's inequality, we immediately have*

$$\limsup_{N\to\infty} \frac{\log \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right]}{\log p_N} \leq 2,$$

and thus the condition in the definition of asymptotic efficiency is sometimes replaced by

$$\lim_{N\to\infty} \frac{\log \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right]}{\log p_N} = 2.$$

When using importance sampling, (1.4) is bounded as follows:

$$RE \leq \frac{\sqrt{\mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right]}}{p_N}.$$

Taking logarithms on both sides, scaling by N and taking the limsup, we find

$$\begin{aligned}\limsup_{N\to\infty} \frac{1}{N}\log(RE) &\leq \limsup_{N\to\infty} \frac{1}{N}\left(\frac{1}{2}\log \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right] - \log p_N\right) \\ &= \limsup_{N\to\infty} \frac{1}{N}\log p_N \left(\frac{\frac{1}{2}\log \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right]}{\log p_N} - 1\right),\end{aligned}$$

which is less than or equal to 0 if both the estimator is asymptotically efficient and $\frac{1}{N}\log p_N$ converges to some value strictly smaller than 0. This means that p_N decays exponentially fast in N, while the relative error decays sub-exponentially fast in N. Therefore, the number of simulations that we need in order to get a decent estimate of our probability of interest grows less than exponentially fast when the estimator for p_N is asymptotically efficient, hence, it will be easier to estimate quantities of interest using importance sampling rather than using ordinary Monte Carlo simulation.

Example 1.1. *(Continued) The proposed change of measure for the $M|M|1$ queue, that is, interchange λ and μ, gives in fact an asymptotically efficient estimator. To prove this, we note that for this queue we have*

$$\lim_{N\to\infty} \frac{1}{N}\log p_N = \log \rho = \log \frac{\lambda}{\mu},$$

see (1.1). *If we let n_a be the number of arriving customers on the path to reach level N and n_d be the number of departing customers on the path to reach level N,*

then we always have $n_a - n_d = N$ customers in the system when we have reached level N. Thus, for a path $\mathcal{P}$ where level N is reached before the system is empty, we have

$$L(\mathcal{P}) = \left(\frac{\lambda}{\mu}\right)^{n_a} \left(\frac{\mu}{\lambda}\right)^{n_d} = \left(\frac{\lambda}{\mu}\right)^{N},$$

and so we find

$$\begin{aligned}
&\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\mathbb{Q}} \left[L(\mathcal{P})^2 I(\mathcal{P})\right] \\
&= \limsup_{N\to\infty} \frac{1}{N} \log \left(\mathbb{E}^{\mathbb{Q}} \left[L(\mathcal{P})^2 \mid I(\mathcal{P}) = 1\right] \mathbb{P}^{\mathbb{Q}} \left(I(\mathcal{P}) = 1\right)\right) \\
&= \limsup_{N\to\infty} \frac{1}{N} \left(\log \left(\frac{\lambda}{\mu}\right)^{2N} + \log \mathbb{P}^{\mathbb{Q}} \left(I(\mathcal{P}) = 1\right)\right) \\
&\leq 2 \log \frac{\lambda}{\mu},
\end{aligned}$$

where the last step follows since $\mathbb{P}^{\mathbb{Q}} \left(I(\mathcal{P}) = 1\right) \leq 1$. This proves the statement. Note that this is a well-known result, and included in this chapter for illustration purposes.

We remark that for the example given in this section, importance sampling is a method which is easy to implement and has an easy proof for asymptotic efficiency. However, the main challenge for importance sampling is to find – or construct – the new measure $\mathbb{Q}$ such that it gives an asymptotically efficient estimator. This is in general a rather difficult problem. After constructing this new measure, still a proof of asymptotic efficiency has to be provided, which can also be a rather challenging task. In this thesis, we meet both of these challenges for more general queueing models.

1.2.2 Related literature

The work in this chapter is not new, in fact, there exists a vast amount of literature on related topics. In this section, we focus on related work regarding importance sampling for queueing networks concerning the same probability of interest as in this work. Needless to say, there also exists related literature on importance sampling where the probability of interest is slightly different or the queueing system is slightly different, see for example [18, 32, 33].

To construct an asymptotically efficient change of measure, it is important to know if the probability of interest decays exponentially fast in N, and if so how fast the probability decays. Thus, we need to determine the so-called *decay rate* of the probability of interest, sometimes also referred to as the *large deviations* of the process. For a single $GI|GI|m$ queue, the decay rate is determined by Sadowsky [36]. The decay rate of the probability of interest usually provides a

good suggestion for the change of measure, a method which has been used by Parekh and Walrand in [35]. Their suggestion for the change of measure is based on heuristics for the decay rate. We note that their change of measure is *state-independent*, that is, the change of measure does not depend on the number of customers in the system (or, the state of the system). Decay rates for related probabilities of interest in queueing systems can be found in for example [2, 3, 27].

Sadowsky [36] also shows that for the *single* $GI|GI|m$ queue the change of measure as proposed by Parekh and Walrand gives an asymptotically efficient estimator under some mild conditions. However, Glasserman and Kou [30] show that for the $M|M|1$ *tandem* queue the change of measure from Parekh and Walrand may or may not give an asymptotically efficient estimator. They provide necessary conditions and (other) sufficient conditions for asymptotic efficiency. De Boer [12] extends these results, but also shows that the change of measure from Parekh and Walrand is the *only* state-independent change of measure that can possibly yield an asymptotically efficient estimator for the $M|M|1$ tandem queue. Therefore, an extension to a *state-dependent* change of measure is required in order to get an estimator that is asymptotically efficient for all parameters. In a state-dependent change of measure, the change of measure *does* depend on the number of customers in the system.

For the 2-node Markovian tandem queue [19] and Jackson networks [22], Dupuis et al. propose such a state-dependent change of measure and prove that it results in an asymptotically efficient estimator for the probability of interest, by using the so-called *subsolution approach*. This method, as developed by Dupuis and Wang [21], can be used to construct a state-(in)dependent change of measure, and is not limited to queueing networks. Moreover, the use of subsolutions is not limited to importance sampling, but it can also be used for splitting, which is the topic of the next section.

1.3 Splitting

The second method for rare event simulation that we discuss in this thesis is splitting. In this method, we keep the probability measure as it is, but instead we 'reward' paths that seem promising, that is, when a path reaches some so-called (predetermined) *threshold*, we *split* a path into multiple (sub)paths that all continue independently, according to the same probability measure as before. This means that from reaching a threshold onwards, we simulate several paths. We also refer to the simulation of a path as the simulation of a *particle*.

The main challenge for splitting is to determine the number and locations of thresholds and the amount of particles that are simulated after reaching the threshold, the so-called *splitting rate* R, so that we get an asymptotically efficient estimator. The definition of asymptotically efficient, see Definition 1.3 below, is very similar to Definition 1.2, though it will not be in terms of a likelihood ratio as this is not used in splitting. In this section, we will explain both ordinary splitting and briefly discuss a special case of splitting called RESTART (Repetitive

Simulation Trials After Reaching Thresholds).

1.3.1 Ordinary splitting

In this section, we consider splitting in more detail. We will also explain splitting by means of an example.

Example 1.2. *An example of splitting can be found in Figure 1.3. We will think of this example as a possible realization of particles and splitting thresholds for an $M|M|1$ queue, although a similar graph may also apply to other processes. Note that we also included the levels c_j in this figure. For an explanation of these levels, see below this example.*

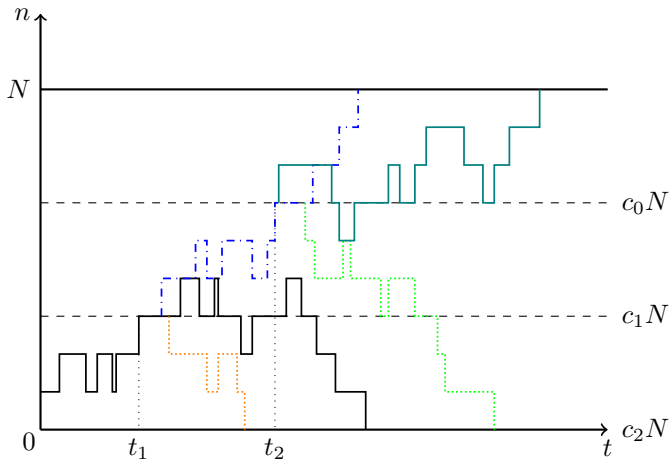


Figure 1.3 An example of splitting: a possible realization of particles and splitting thresholds.

On the horizontal axis, the time t is being displayed, while on the vertical axis we have the number of customers in the queue, denoted by n. The solid horizontal line is the overflow level N, which we want to reach, while the dashed horizontal lines are the splitting thresholds. After the arrival of customer 1 at $t = 0$, the process reaches the first threshold at time t_1. Using splitting rate $R = 3$, we see that the process splits into three particles out of which two, orange and black, reach the taboo set afterwards. The third, blue, particle hits the second threshold at time t_2, where it splits into $R = 3$ new particles. Out of these, the blue and the teal particle reach the overflow level N and the green particle reaches the taboo set.

This means that for this particular example, two out of the nine possible particles have reached the overflow level, and therefore we can estimate the probability to reach level N as $\hat{p}_N = \frac{2}{9}$.

Intuitively, the locations of splitting thresholds and splitting rates should be chosen such that on average one out of R particles reaches the next threshold.

This would also imply that the total number of particles that are simulated grows linear with the total number of thresholds. Note that it could happen that the total number of particles that are simulated grows exponentially with the total number of thresholds.

The thresholds can be defined in terms of some *importance function* $U(x)$, which takes lower values as x is closer to the target set. Using this function, we can determine the so-called *level sets* C_j of the importance function such that

$$G \subset C_0 \subset C_1 \subset \cdots \subset C_{J-1} \subset C_J = \mathcal{D},$$

where G is the target set and $\mathcal{D}$ is the domain of the process, scaled by the parameter N. To be more precise, the level sets C_j of the importance function $U(x)$ are defined as

$$C_j = \{x \in \mathcal{D} : U(x) \leq j \log(R)/N\}, \qquad 0 \leq j \leq J-1,$$

where J is the total number of level sets. By construction, we want the starting point $x_0 \in C_J \backslash C_{J-1}$ and hence we find $J = \lceil \frac{U(x_0)N}{\log(R)} \rceil$. Then, reaching a threshold is equivalent to reaching a certain level $c_j = \partial C_j$, where ∂C_j is the boundary of the level set C_j. (Note that in Figure 1.3 the levels are multiplied by N since in this figure the domain is unscaled.)

Suppose the splitting thresholds are known and suppose they are at levels $c_0, \ldots, c_{J-1}$. Then, when a particle reaches level c_{J-1}, the particle splits into R particles all of which continue independently according to the dynamics of the process. If some of these particles hit level c_{J-2}, these particles again split into R particles. This process continues until all particles either hit the taboo set, or the target set. When a particle hits a level that the particle has reached before, no splitting occurs. The estimator is defined as

$$\hat{p}_N = \frac{1}{S} \sum_{i=1}^{S} \frac{T(i)}{R^J}, \tag{1.5}$$

where $T(i)$ denotes the number of particles that reach the target set in simulation i. Note that R^J is the total number of *possible* particles that could be simulated. In practice, however, not all of these particles are simulated, since there will be (many) particles that reach the taboo set before reaching a next splitting threshold.

1.3.2 Performance measures

The estimator $\hat{p}_N$ in (1.5) is an unbiased estimator, since

$$\mathbb{E}\left[\hat{p}_N\right] = \frac{1}{S}\mathbb{E}\left[\sum_{i=1}^{S} \frac{T(i)}{R^J}\right] = R^{-J}\mathbb{E}\left[T\right] = R^{-J}\mathbb{E}\left[\sum_{i=1}^{R^J} I(i)\right] = p_N,$$

where $I(i)$ indicates whether particle i has reached the target set before the taboo set or not. For all particles that have not been simulated, see below (1.5), we naturally define $I(i) = 0$. Note that again, we omit the dependence on the simulation number whenever this is convenient.

For splitting, asymptotic efficiency of the estimator can be defined in terms of the second moment of the estimator T/R^J. However, the simulation of splitting requires the simulation of at most R^J particles, which can be computationally expensive. Therefore, we consider the work-normalized error, see [31], and so we define asymptotic efficiency as follows.

Definition 1.3. *The estimator for p_N, here defined as T/R^J, is asymptotically efficient – also called asymptotically optimal – if and only if*

$$\liminf_{N\to\infty} \frac{\log\left(w(N)R^{-2J}\mathbb{E}\left[T^2\right]\right)}{\log p_N} \geq 2,$$

where $w(N)$ is the expected computational effort per simulation run.

We will formally define $w(N)$ where needed in Chapter 7. The interpretation of asymptotic efficiency of the estimator for splitting is very similar to that for importance sampling, that is, if the estimator is asymptotically efficient, the relative error grows less than exponentially fast in N. In addition, for splitting it also means that the computational effort grows less than exponentially fast in N. In contrast to importance sampling, where it is known that the computational effort grows polynomially in N, for splitting it still needs to be shown whether the computational effort does not grow too fast (though it grows at least as fast as for importance sampling).

1.3.3 RESTART

RESTART is a special case of splitting, in the sense that the method is almost the same. Splitting thresholds that are used for ordinary splitting can also be used for RESTART. Thus, the main challenges for ordinary splitting and RESTART are the same. However, from an implementation point of view, RESTART is designed to give computational advantages. In this section, we briefly highlight the differences between RESTART and ordinary splitting.

The advantage of RESTART is that we spend less effort simulating particles that tend in the 'wrong' direction – from a queueing perspective this would mean towards an empty system – that is, we kill all particles that reach below the threshold they were born in. For each threshold we make sure there always exists at least one particle that is not killed, for example, the particle that is born at the starting point is never killed. This results in less computational effort compared to ordinary splitting. However, it might result in the estimator not being unbiased. To compensate for this, whenever a particle up-crosses a threshold, it splits (even if this particle has reached this threshold before). It turns out that this results in the estimator being unbiased, although with a much more complicated proof.

Example 1.2. *(Continued) An example of* RESTART *is shown in Figure 1.4. Again we will think of this example as a possible realization of particles and splitting thresholds for an $M|M|1$ queue, although a similar graph may also apply to other processes.*

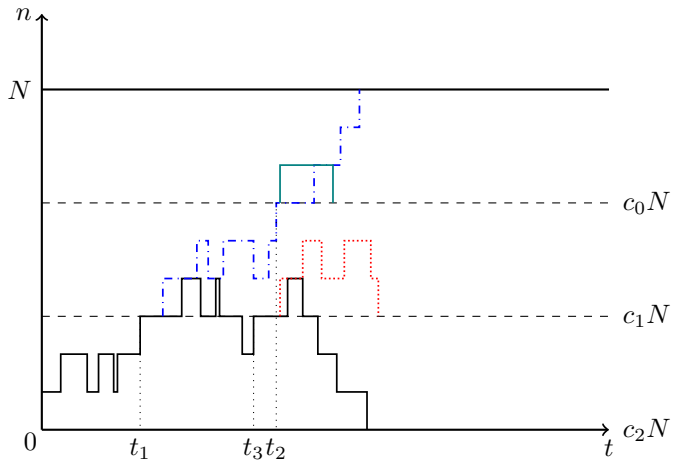


Figure 1.4 An example of RESTART: a possible realization of particles and splitting thresholds.

The figure is mostly the same as the figure for ordinary splitting, see Figure 1.3. This means that the axes are the same, as well as the splitting thresholds and the particles, if possible, in order to show the differences. Again, at time $t = 0$, we start with one customer in the system and at time t_1 the first particle reaches the first threshold, so it splits into $R = 3$ particles. Afterwards, we only see two particles, since the orange particle (see Figure 1.3 for the orange particle) immediately went below the threshold it was born in. The black particle is not killed when going below threshold c_1, since it was born at threshold c_2. Then, the next splitting occurs at time t_3, since the black particle again crossed threshold c_1, and also in this case only two particles stay above the threshold, the black particle and the red particle, and the third particle is immediately killed. At time t_2 the same threshold is crossed as for ordinary splitting, though again only two particles stay alive (the green particle is immediately killed, see Figure 1.3 for the green particle) and the teal particle stays alive for only a short period of time.

From this example, it is clear that the computational effort is much less for RESTART than for ordinary splitting. Therefore, from an implementation perspective, RESTART is preferred over ordinary splitting, especially when simulating over a long time interval. However, it is harder to analyze RESTART than ordinary splitting.

One final thing that has to be taken into account for RESTART in particular, depending on the process that is studied, is that it may be possible to jump over several thresholds at the same time. Then, one has to take care in defining the

thresholds each particle was born in. For splitting, the latter is not necessary, but for RESTART it is important to know when to kill a particle.

1.3.4 Related literature

In this section, we focus on related literature in the context of splitting and RESTART for queueing networks. We mainly focus on literature in which proofs for asymptotic efficiency are given. If we focus on the same probability of interest as in the current work, it turns out that there does not exist a lot of literature that explicitly concerns splitting in this setting. It remains unclear why there seems to be less interest in using splitting rather than importance sampling.

To construct a splitting scheme that results in an asymptotically efficient estimator, it is necessary to have knowledge about the decay rate of the probability of interest. Since this was necessary for importance sampling as well, the related literature concerning the decay rate has been discussed in Section 1.2.2.

In [28], splitting is also studied in the context of queueing systems. In this work, however, there are no proofs for efficiency of the proposed algorithms, that is, algorithms are only verified by using simulation. Splitting has also been studied in [34], where both examples of a 2-node $M|M|1$ tandem queue are considered as well as a system with server slowdown. The probability of interest in that paper is slightly different from ours, since the authors consider overflow in the second queue instead of overflow in the system as a whole.

From a more general analytical perspective, in [16] splitting is studied using subsolutions as importance functions. Although the main focus of the latter paper is not necessarily queueing systems, some examples of queueing systems are given considering for example the total buffer overflow.

RESTART was first introduced in [40], to speed up the simulation of splitting. Also in this work there are no proofs for efficiency, though the algorithm is verified by using simulation. Analytically, RESTART has been studied in [15], where also examples of queueing systems are included. Recently, in [39] a combination of RESTART and ordinary splitting is proposed, in the sense that one may consider not to immediately kill a particle when it goes below the threshold it was born in, but for example one threshold lower. Also for this latter algorithm, no proofs of efficiency are provided.

1.4 Contributions and outline of this thesis

In this section, we summarize the contributions and outline of this thesis. In Table 1.1 the contributions of this thesis can be found, in comparison with existing literature on total buffer overflow in queueing systems. In this table, we only mention existing literature that considers proofs for the decay rate or proofs for (conditions for) an asymptotically efficient estimator.
We remark that for a single queue, there is no need for a state-dependent change of measure, since there exists a state-independent change of measure that gives an

Table 1.1 Existing literature and contributions of this thesis. In this table, c.o.m. denotes change of measure and n.a. denotes not available.

Subject	Type of queue	Markovian	non-Markovian
Decay rate	Single	[11]	[36]
	Tandem	[30]	Chapter 2
State-independent c.o.m.	Single	Example 1.1	[36], Chapter 5
	Tandem	[12, 30]	Chapter 4
State-dependent c.o.m.	Single	n.a.	n.a.
	Tandem	[14, 19, 22], Chapter 6	Chapter 5
Splitting	Single	[16]	[16], Chapter 7
	Tandem	[16]	[16], Chapter 7

asymptotically efficient estimator. In addition, recall from Example 1.1 that for a single $M|M|1$ queue, the exact probability of interest is known and so simulation is not needed in order to estimate this probability.

In order to construct an asymptotically efficient change of measure or an asymptotically efficient splitting scheme, it is important to have some knowledge on the decay rate of the probability of interest. To this end, in **Chapter 2** we determine the decay rate of the probability to reach N customers during a busy cycle of a non-Markovian tandem queue. While doing so, we also obtain the decay rates of two related probabilities, namely, that the number of customers exceeds N in stationarity and upon arrival of a customer. The work in this chapter is based on the following publication:

> A. Buijsrogge, P.T. de Boer, K. Rosen, and W.R.W. Scheinhardt. Large deviations for the total queue size in non-Markovian tandem queues. *Queueing Systems*, 85 (3): 305–312, 2017

Although not directly related to the rest of this thesis, in **Chapter 3** we extend the work from Chapter 2 to unequal customer sizes. In particular, we obtain upper and lower bounds on the decay rate of the probability that the total size of all customers exceeds N during a busy cycle of the system. It turns out that upper and lower bounds only coincide for particular cases. This topic was first considered in [38].

In **Chapter 4** we discuss a method to find a state-independent change of measure for a non-Markovian tandem queue and we show that this is equivalent to a change of measure that was earlier, but implicitly, described by Parekh and Walrand [35] and later by Frater and Anderson [25]. We also show that this change of measure is the only exponential state-independent change of measure for which the corresponding estimator may be asymptotically efficient. Lastly, we provide some necessary conditions for this to be the case. Chapter 4 is based

on:

> A. Buijsrogge, P.T. de Boer, and W.R.W. Scheinhardt. On state-independent importance sampling for the $GI|GI|1$ tandem queue. *Accepted for publication in Probability in the Engineering and Informational Sciences*

As a result, when using importance sampling for non-Markovian tandem queues, a different approach is required in order to get an asymptotically efficient estimator for all input parameters. In **Chapter 5**, we extend the existing work on importance sampling for d-node Markovian tandem queues, as in [19] and [22], to d-node non-Markovian tandem queues. We present several state-dependent changes of measure for a d-node $GI|GI|1$ tandem queue based on the subsolution approach and we prove that, under a conjecture, these state-dependent changes of measure give an asymptotically efficient estimator for the probability of interest when all involved distributions have bounded support. This chapter is based on the following submission:

> A. Buijsrogge, P.T. de Boer, and W.R.W. Scheinhardt. Importance sampling for non-Markovian tandem queues using subsolutions. *Submitted*

Considering the results above, where we found several changes of measure to be asymptotically efficient, in **Chapter 6** we determine possibilities for the change of measure for the special case of a 2-node $M|M|1$ tandem queue to be asymptotically efficient using similar methods and analysis as in [14, 19, 22]. Even though there are already three changes of measure known that are asymptotically efficient, see [19, 22], it appears that there exists a whole family of changes of measure that result in an asymptotically efficient change of measure. We emphasize that the analysis of Chapter 5 and [14, 19, 22] is different, since the first one has a state description in continuous time and the latter one considers an embedded discrete time Markov chain. The results from this chapter are based on:

> A. Buijsrogge, P.T. de Boer, and W.R.W. Scheinhardt. Importance sampling for Markovian tandem queues using subsolutions: exploring the possibilities. *Submitted*

As mentioned earlier in Section 1.2.2, the use of subsolutions is not limited to the construction for an asymptotically efficient estimator in importance sampling, but they can also be useful to determine splitting thresholds. Therefore, in **Chapter 7** we consider splitting for non-Markovian tandem queues. In order to prove asymptotic efficiency for the estimator when the splitting thresholds are based on subsolutions, we need to study the decay rate of the probability of interest when starting with multiple customers in the system, instead of starting with one customer in the system. It turns out that *both* splitting thresholds based on subsolutions, as well as splitting thresholds based on this decay function yield

an asymptotically efficient estimator. This chapter is based on:

> A. Buijsrogge, P.T. de Boer, and W.R.W. Scheinhardt. Splitting for non-Markovian tandem queues using subsolutions. *Working paper*

Chapter 8 studies a somewhat different problem than all previous chapters. In this chapter, we study the probability that a stochastic process leaves a neighborhood of a metastable point during some long time interval $[0, T]$. We develop large deviations estimates when the time interval of interest depends on the large deviation parameter. Chapter 8 is based on:

> A. Buijsrogge, P. Dupuis, and M. Snarski. Splitting algorithms for rare event simulation over long time intervals. *Submitted*

CHAPTER 2

Large deviations for the total queue size in non-Markovian tandem queues

Large deviations for the total queue size in (networks of) queues are of interest since they provide insight into how the probability of overflow decays as the overflow level increases. Such results are well-known for Markovian tandem queues, see for example [30], but not for non-Markovian tandem queues. Thus, in this chapter, our interest is in the probability that the number of customers in a non-Markovian tandem queue reaches some high level N *during a busy cycle*, and the related probabilities that this number exceeds N *in stationarity* and *upon arrival of a customer*. In Sadowsky [36] the probability in a busy cycle has been considered for a single $GI|GI|m$ queue. In Bertsimas et al. [2] the Palm probability of a single queue in a network reaching some high level N upon arrival of a customer is considered; the associated decay rate is characterized using the sojourn time of a specific customer. Very related to this work is Ganesh [27], in which the large deviations behavior of the sojourn time for queues in series is considered. The exact asymptotics of the sojourn time for tandem queues have been determined by Foss [24].

In this chapter we will consider a d-node $GI|GI|1$ tandem queue with renewal input and independent, i.i.d. service processes. We characterize the decay rate for the probability of reaching a total of N customers during a busy cycle of the system. Also we show that the stationary probability of having N customers in the system, as well as the probability of having N customers in the system upon arrival, have the same decay rate.

In Section 2.1 we provide the model and introduce our notation. Section 2.2 presents the main result of this chapter, together with proofs.

2.1 Model and preliminaries

In Chapters 2–7 of this thesis, we consider d $GI|GI|1$ queues in tandem. Customers arrive at queue 1 according to a renewal process with inter-arrival times A_k (between customers k and $k+1$) distributed according to some positive random variable A. The service times at queue j, denoted as $B_k^{(j)}$ (for customer k), are independent and identically distributed according to some positive random vari-

able $B^{(j)}$. Furthermore, we assume that all processes are independent and that customers are served based on a first come first served (FCFS) principle. After service completion at queue $j < d$, each customer enters queue $j+1$ immediately, and customers leave the system after service completion at queue d. For stability, we assume $\mathbb{E}\left[B^{(j)}\right] < \mathbb{E}\left[A\right]$ $\forall j$. Note that when at least one of the queues would be unstable, then our event of interest would not be rare and hence the decay rate equals 0. See Figure 2.1 for a graphical illustration of the d-node tandem queue.

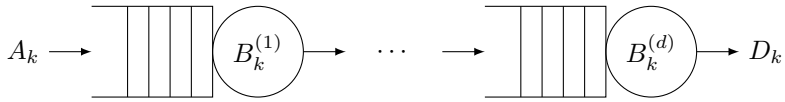


Figure 2.1 The d-node tandem queue.

Starting with customer 1 entering queue 1 and all other queues empty, we are interested in the probability of overflow during the busy cycle of the total queue. This can be written as $\mathbb{P}(K_N < K_0)$, where K_N is the index of the first customer who reaches the overflow level N and K_0 is the index of the first customer to see an empty system upon arrival. The indices K_N and K_0 can be expressed in terms of the inter-arrival times A_k (at queue 1) and the inter-departure times D_k (from queue d), as follows.

$$K_N = \min\left\{n \geq N : \sum_{k=1}^{n-1} A_k < \sum_{k=1}^{n-N+1} D_k\right\}, \tag{2.1}$$

$$K_0 = \min\left\{m : \sum_{k=1}^{m-1} A_k > \sum_{k=1}^{m-1} D_k\right\}. \tag{2.2}$$

For the inter-departure time D_k (between customers $k-1$ and k, for $k \geq 2$), we can write $D_k = B_k^{(d)} + I_k^{(d)}$, where $I_k^{(d)}$ is the, possibly zero, idle time of queue d after the departure of customer $k-1$, before customer k enters queue d. Consistently with this, D_1 is simply defined as the sojourn time of customer 1.

Other probabilities of interest that are related to $\mathbb{P}(K_N < K_0)$ are $\mathbb{P}(L \geq N)$ and $\mathbb{P}(L^{(a)} \geq N)$, where L denotes the total number of customers in the system in stationarity, and $L^{(a)}$ denotes the same number but immediately after an arbitrary arrival (including the customer that just arrived).

To characterize the decay rate, we need the following. For any random variable X, let $\Lambda_X(\theta) = \log \mathbb{E}\left[e^{\theta X}\right]$ denote its log moment generating function. For all $j = 1, \ldots, d$, we assume that $\Lambda_{B^{(j)}}(\theta)$ exists for some $\theta > 0$, and define $\theta^{(j)}$ as

$$\theta^{(j)} = \sup\left\{\theta : \Lambda_A(-\theta) + \Lambda_{B^{(j)}}(\theta) \leq 0\right\}. \tag{2.3}$$

Note that we only consider $\Lambda_A(-\theta)$ for $\theta \geq 0$ and so it always exists. Furthermore, we say $\theta^{(j)} = \infty$ when $\Lambda_A(-\theta) + \Lambda_{B^{(j)}}(\theta) < 0$ for all $\theta > 0$; note that this is equivalent to $\mathbb{P}(B^{(j)} > A) = 0$, see Lemma 2.9 in Section 2.3 for a formal proof.

Furthermore, we define $\theta^* = \min_j(\theta^{(j)})$, and assume that $\theta^* < \infty$, that is, we do not have $\mathbb{P}(B^{(j)} > A) = 0$ for all queues, so that the number of customers can grow arbitrarily large and the decay rates of the probabilities of interest will be in $(0, \infty)$. This gives rise to the following definition.

Definition 2.1. *The queue(s) j with $\theta^{(j)} = \theta^*$ will be called the θ-bottleneck queue(s).*

Remark 2.2. *The notion of θ-bottleneck queue as in Definition 2.1 can be different from the ρ-bottleneck queue, which is the queue with the largest server utilization $\rho_j = \mathbb{E}[B^{(j)}]/\mathbb{E}[A]$. However, both notions may yield the same bottleneck; for example, in case of an $M|M|1$ tandem queue, this is always the case.*

Finally, we assume that $\mathbb{P}(A > \sum_{j=1}^{d} B^{(j)}) > 0$, so that the system can become empty. For reference throughout this thesis, we summarize all assumptions that have been made so far. These assumptions are made throughout all chapters of this thesis concerning queueing systems, that is, Chapters 2-6.

Assumption 2.3.

1) *The system is stable, that is, for all j, $\mathbb{E}[B^{(j)}] < \mathbb{E}[A]$;*
2) *The probability of interest is non-trivial, that is, for at least one queue j we have $\mathbb{P}(B^{(j)} > A) > 0$ and $\mathbb{P}(A > \sum_{j=1}^{d} B^{(j)}) > 0$;*
3) *The log moment generating functions for all service time distributions exist, that is, for all j, $\Lambda_{B^{(j)}}(\theta) > -\infty$ for some $\theta > 0$.*

As a consequence of the stability and non-triviality assumption, $0 < \theta^* < \infty$. In Section 4.8 we discuss the interpretation of $\theta^* = \theta^{(j)}$, in particular when $\Lambda_A(-\theta^*) + \Lambda_{B^{(j)}}(\theta^*) < 0$.

2.2 Main result

In this section we present the main result of this chapter, namely the characterization of the decay rates of $\mathbb{P}(K_N < K_0)$, $\mathbb{P}(L \geq N)$ and $\mathbb{P}(L^{(a)} \geq N)$. In order to achieve this result, we will prove both a lower bound and an upper bound for the decay of $\mathbb{P}(K_N < K_0)$, which will also turn out to hold for the other decay rates. We will start with the lower bound, with a proof based on a coupling argument.

Lemma 2.4. *(Lower bound) For the decay of $\mathbb{P}(K_N < K_0)$ it holds that*

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(K_N < K_0\right) \geq \Lambda_A(-\theta^*).$$

Proof. We compare the tandem queue to a single queue with the same arrival process A_k and the service process of the j'th queue in the tandem, $B_k^{(j)}$. (This

is equivalent to comparing our tandem queue to a tandem queue with the same arrival process and all service times set to 0, except the service times of queue j.) Thus, by coupling the single queue to the tandem queue, we use the random variables $A_1, A_2, \ldots$ and $B_1^{(j)}, B_2^{(j)}, \ldots$ that are used for the tandem queue *also* for the single queue. The idea of the proof is to show that overflow is more likely in the tandem queue than in the single queue.

We define $\widehat{D}_i$, $\widehat{K}_0$ and $\widehat{K}_N$ analogous to D_i, K_0 and K_N but for the single queue and denote the inter-departure time of customer i at queue j in the tandem queue by $D_i^{(j)}$.

For $i < K_0$ it holds that $D_i^{(j)} = I_i^{(j)} + B_i^{(j)}$, and for $i < \widehat{K}_0$ it holds that $\widehat{D}_i = B_i^{(j)}$, as the single queue does not have idle times during its busy cycle. Since a customer cannot leave the last queue in the tandem before having left queue j, we find

$$\sum_{i=1}^{k} D_i \geq \sum_{i=1}^{k} D_i^{(j)} = \sum_{i=1}^{k} \widehat{D}_i + I_i^{(j)} \geq \sum_{i=1}^{k} \widehat{D}_i, \tag{2.4}$$

for all $k = 1, ..., \min(K_0 - 1, \widehat{K}_0 - 1)$, meaning that a customer leaves the tandem queue not earlier than that same customer leaves the coupled single queue.

Based on this we first show, by contradiction, that $\widehat{K}_0 \leq K_0$, that is, the single queue empties not later than the tandem queue. Suppose that $\widehat{K}_0 > K_0$, then (2.4) still holds for k up to $K_0 - 1$. By using (2.2) and (2.4) we have

$$\sum_{k=1}^{K_0-1} A_k > \sum_{k=1}^{K_0-1} D_k \geq \sum_{k=1}^{K_0-1} \widehat{D}_k,$$

which implies by definition of $\widehat{K}_0$ that $\widehat{K}_0 \leq K_0$. Therefore, our assumption $\widehat{K}_0 > K_0$ is wrong and so we have shown $\widehat{K}_0 \leq K_0$.

Next, we show that the tandem queue reaches the overflow level not later than the single queue. Suppose we have reached overflow in a busy cycle of the single queue, that is $\widehat{K}_N < \widehat{K}_0$. Then we have, by using (2.1) and (2.4),

$$\sum_{k=1}^{\widehat{K}_N-1} A_k < \sum_{k=1}^{\widehat{K}_N-N+1} \widehat{D}_k \leq \sum_{k=1}^{\widehat{K}_N-N+1} D_k,$$

and thus $K_N \leq \widehat{K}_N$.

Hence $\widehat{K}_N < \widehat{K}_0$ implies $K_N < K_0$, which means that overflow during a busy period in the single queue implies overflow during a busy period in the tandem queue. So we have for any j that

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(K_N < K_0\right) \geq \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(\widehat{K}_N < \widehat{K}_0\right) = \Lambda_A(-\theta^{(j)}),$$

where the second step follows by Theorem 1 in [36]. In particular, the above holds for j such that $\theta^{(j)} = \theta^*$, which completes the proof. □

The next step is to prove an upper bound. We will use a regenerative argument, for which we need that the expected total time spent at or above level N during a busy cycle in which level N is reached, is bounded from below, independently of N. Even though this sounds very plausible, we could not find a reference. Hence the next lemma, the proof of which is based on first principles, together with the technical assumption $\mathbb{P}(B^{(d)} > A) > 0$ (which will not be a limitation for the main result).

Let $L(t)$ be the total number of customers in the system at time t, and let T be the length of the first busy cycle; then we define the expected total time τ_N spent at or above level N during a busy cycle as $\tau_N = \int_0^T \mathbb{1}\{L(t) \geq N\}dt$.

Lemma 2.5. *Suppose that $\mathbb{P}(B^{(d)} > A) > 0$. Then some $c > 0$ exists such that for all $N = 1, 2, \ldots,$*

$$\mathbb{E}\left[\tau_N \mid K_N < K_0\right] \geq c.$$

Proof. Consider a busy cycle in which the overflow level N is reached and denote the moment that N is reached for the first time by t. Then the first arrival after t occurs at time $t_1 = t + A_{K_N}$, while the *second* departure after t occurs at some time $t_2 \geq t + B^{(d)}_{K_N - N + 2}$. (To see this, note that at time t, when customer K_N enters, customer $K_N - N + 1$ is the first to depart from the system, so the service of customer $K_N - N + 2$ at queue d cannot start earlier than at time t). It is not difficult to check that if $t_1 < t_2$, there will be at least N customers in the system between t_1 and t_2. Thus, for any N we have $\mathbb{E}\left[\tau_N \mid K_N < K_0\right] \geq \mathbb{E}\left[\max(0, t_2 - t_1) \mid K_N < K_0\right] \geq \mathbb{E}\left[\max(0, B^{(d)} - A)\right]$, which is nonzero due to $\mathbb{P}(B^{(d)} > A) > 0$. □

We are now ready to prove the upper bound, based on a regenerative argument and a Chernoff bound.

Lemma 2.6. *(Upper bound) For the decay of $\mathbb{P}(K_N < K_0)$, under the condition that $\mathbb{P}(B^{(d)} > A) > 0$, it holds that*

$$\begin{aligned}\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(K_N < K_0\right) &\leq \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(L \geq N\right) \\ &\leq \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(L^{(a)} \geq N\right) \\ &\leq \Lambda_A(-\theta^*),\end{aligned}$$

and a similar statement holds when we replace all lim sups *by* lim infs.

Proof. The proof for the lim infs and the lim sups is similar; we only give it explicitly for the lim sups. The same steps apply to prove the lim infs, in which the supremum has to be replaced by the infimum at the appropriate places.

The first inequality follows from a regenerative argument, as in [30], by which we have

$$\mathbb{P}(K_N < K_0) = \frac{\mathbb{E}\left[T\right]\ \mathbb{P}(L \geq N)}{\mathbb{E}\left[\tau_N \mid K_N < K_0\right]},$$

where T is the length of a busy cycle, which has a finite, constant, expectation due to stability of the system, and τ_N is the total time spent above level N during a busy cycle, which is bounded from below independently of N, see Lemma 2.5.

The remainder of the proof considers the system in stationarity, so time 0 and customer 0 are not necessarily related to the start of a busy cycle. For the second inequality then, fix some arbitrary time t in stationarity, and consider the last customer to arrive before time t, call this customer k. If the number of customers at time t is $\geq N$, then the queue length $L_k^{(a)}$ observed by – and including – customer k is also $\geq N$, because there can only be departures between the arrival of customer k and time t. So $\mathbb{P}(L \geq N) \leq \mathbb{P}(L_k^{(a)} \geq N)$. Furthermore, $L_k^{(a)} \geq N$ if and only if the sojourn time of customer $k-N+1$, denoted by S_{k-N+1}, exceeds the sum of $N-1$ inter-arrival times. So we have

$$\mathbb{P}(L_k^{(a)} \geq N) = \mathbb{P}\left(S_{k-N+1} \geq \sum_{i=k-N+1}^{k-1} A_i\right).$$

Note that this probability is independent of the age of A_k at time t, as the inter-arrival times are independent, so in fact $L_k^{(a)}$ has the same distribution as $L^{(a)}$, that is, customer k cannot be distinguished from an arbitrary customer in stationarity, which proves the second inequality.

For the last inequality, we analyze the right-hand side of the equation above (keeping customer index $k-N+1$ for convenience). We have for any $\theta > 0$, using the Chernoff bound, and the independence of S_{k-N+1} and $\sum_{i=k-N+1}^{k-1} A_i$,

$$\begin{aligned}\mathbb{P}\left(S_{k-N+1} \geq \sum_{i=k-N+1}^{k-1} A_i\right) &\leq \mathbb{E}\left[e^{\theta\left(S_{k-N+1} - \sum_{i=k-N+1}^{k-1} A_i\right)}\right] \\ &= \mathbb{E}\left[e^{\theta S_{k-N+1}}\right]\ \mathbb{E}\left[e^{-\theta \sum_{i=k-N+1}^{k-1} A_i}\right].\end{aligned}$$

In [27] it is shown that $\mathbb{E}[e^{\theta S_{k-N+1}}]$ is upper bounded by some constant C for all $\theta \in (0, \theta^*)$ (see just after equation (27) in the proof of Theorem 1). Note that the assumptions in [27] are more general than ours, so we can use this result. Hence, we have for any $\theta \in (0, \theta^*)$

$$\begin{aligned}&\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(S_{k-N+1} \geq \sum_{i=k-N+1}^{k-1} A_i\right) \\ &\leq \limsup_{N\to\infty} \frac{1}{N}\left(\log C + \log \mathbb{E}[e^{-\theta \sum_{i=k-N+1}^{k-1} A_i}]\right) = \Lambda_A(-\theta),\end{aligned}$$

where the last step follows by independence of the inter-arrival times. Taking $\theta \to \theta^*$ to achieve the best possible bound proves the statement. □

Theorem 2.7. *Consider a stable FCFS d-node GI\|GI\|1 tandem queue with arrival process and service processes at all queues i.i.d. and independent of each*

other, that satisfies Assumption 2.3. If $\theta^* < \infty$*, it holds that*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N < K_0) = \lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(L \geq N) = \lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(L^{(a)} \geq N) = \Lambda_A(-\theta^*). \tag{2.5}$$

Proof. When $\mathbb{P}(B^{(d)} > A) > 0$, statement (2.5) follows immediately from Lemmas 2.4 and 2.6 since all liminfs and limsups (with respect to each of the three probabilities) are equal to $\Lambda_A(-\theta^*)$.

To show that (2.5) also holds in general, we consider a tandem queue where $\mathbb{P}(B^{(d)} > A) = 0$, and two corresponding systems, fed by the same arrival process. One is a queue in isolation as introduced in the proof of Lemma 2.4. More specifically, we consider a θ-bottleneck queue, that is, some queue j for which $\theta^{(j)} = \theta^*$. In this single queue we define $\widehat{K}_0$, $\widehat{K}_N$, $\widehat{L}$ and $\widehat{L}^{(a)}$ analogous to K_0, K_N, L and $L^{(a)}$ in the tandem queue. Note that $\mathbb{P}(B^{(j)} > A) > 0$ (otherwise we would have $\theta^* = \theta^{(j)} = \infty$), and hence (2.5) holds for this single queue system.

The other system we consider is the original tandem queue augmented with a suitably chosen additional queue $d+1$, for example, letting $B^{(d+1)} \sim B^{(j)}$ where queue j is a θ-bottleneck queue (another option is to choose $B^{(d+1)} \sim \exp(\mu)$ for some sufficiently large μ). In this system we analogously define $\widetilde{K}_0$, $\widetilde{K}_N$, $\widetilde{L}$ and $\widetilde{L}^{(a)}$. Clearly we then have $\mathbb{E}\left[B^{(d+1)}\right] < \mathbb{E}[A]$ and $\theta^{(d+1)} \geq \theta^*$, while we also have $\mathbb{P}(B^{(d+1)} > A) > 0$. As a result, also for this system (2.5) holds.

All three probabilities for the original tandem queue can now be bounded by the corresponding probabilities in the two other systems, as follows.

$$\begin{aligned}
&\mathbb{P}(\widehat{K}_N < \widehat{K}_0) \leq \mathbb{P}(K_N < K_0) \leq \mathbb{P}(\widetilde{K}_N < \widetilde{K}_0) \\
&\mathbb{P}(\widehat{L} \geq N) \quad\ \leq \mathbb{P}(L \geq N) \quad\ \leq \mathbb{P}(\widetilde{L} \geq N) \\
&\mathbb{P}(\widehat{L}^{(a)} \geq N) \leq \mathbb{P}(L^{(a)} \geq N) \leq \mathbb{P}(\widetilde{L}^{(a)} \geq N)
\end{aligned}$$

Each of these inequalities follows similarly as in the proof of Lemma 2.4 by coupling arguments; note that setting $B^{(d+1)} \equiv 0$ in the augmented tandem queue leads to the original tandem, and setting the service times of all but one queue in the original tandem queue leads to the single queue. Thus, the first inequality is straightforward from the proof of Lemma 2.4, and the second can be shown similarly. For the other two lines, we just need to consider the departure times in the three systems for the same customer to show that $\widehat{L}(t) \leq L(t) \leq \widetilde{L}(t)$ at any time t, and hence also in stationarity and upon arrivals.

Finally, we take logarithms above, then divide by N, and take limits. □

Note that when $\theta^* = \infty$, the total number of customers cannot grow arbitrarily large (see Section 2.1), and hence the decay rates in (2.5) are not properly defined (or are equal to $-\infty$).

Remark 2.8. *As mentioned in the introduction, Bertsimas et al. [2] and Ganesh [27] consider the decay of related overflow probabilities in a more general*

setting, where certain types of dependence for the arrival and service processes are allowed. We expect that the bounds in this chapter can be extended to this case as well, but this will take different techniques and additional effort, in particular to relate $\mathbb{P}(K_N < K_0)$, $\mathbb{P}(L \geq N)$ *and* $\mathbb{P}(L^{(a)} \geq N)$ *in the more general setting.*

2.3 Appendix

In this section, we present a formal proof of the equivalence between $\mathbb{P}(B^{(j)} > A) = 0$ and $\theta^* = \infty$, as shown in [8].

Lemma 2.9. $\mathbb{P}(B^{(j)} > A) = 0 \iff \theta^* = \infty$.

Proof. Suppose that $\mathbb{P}(B^{(j)} > A) = 0$, then

$$\Lambda_A(-\theta) + \Lambda_{B^{(j)}}(\theta) = \log \mathbb{E}\left[e^{\theta(B^{(j)} - A)}\right] < 0,$$

for all $\theta > 0$. For the reverse statement, suppose that $\mathbb{P}(B^{(j)} - A > 0) > 0$. Then we also have $\mathbb{P}(B^{(j)} - A > \epsilon) > 0$ for some $\epsilon > 0$. Hence,

$$\mathbb{E}[e^{\theta(B^{(j)} - A)}] > \int_{\epsilon}^{\infty} e^{\theta x} dF_{B^{(j)} - A}(x) > e^{\theta \epsilon} P(B^{(j)} - A > \epsilon),$$

which goes to ∞ as $\theta \to \infty$. Therefore, $\log \mathbb{E}\left[e^{\theta(B^{(j)} - A)}\right]$ can only be smaller than or equal to 0 if $\theta < \infty$. Hence, $\theta^* < \infty$ if $\mathbb{P}(B^{(j)} - A > 0) > 0$ and so the reverse statement holds. □

CHAPTER 3

Large deviations for the total queue size with unequal customer sizes

In Chapter 2, we considered the large deviations for the total queue size in non-Markovian tandem queues. The total queue size is independent of the size of the individual customers, and hence we could say that in that chapter all customers have size 1. However, in for example a production process, the size of packets – which we will refer to as customers – may change throughout the process and therefore customer sizes can have an influence on the probability of, for example, having a full warehouse. In particular, when the customer size changes throughout the production process, this significantly influences the probability of having a full warehouse.

In this chapter, we extend the results from Chapter 2 to unequal customer sizes, that is, customers can have a different size in a different queue. Note that in the remainder of this thesis we will assume equal customer sizes, all having size 1. We assume that all customers in a particular queue have the same size, which is a natural assumption in a production process. To the best of our knowledge, no such results exist in literature.

We consider the same three probabilities as in Chapter 2; the probability that the total size of all customers reaches some high level N *during a busy cycle*, and the probability that this number exceeds N *in stationarity* and *upon arrival of a customer*. We obtain lower and upper bounds for decay rate of the probability of interest, which coincide for a particular case. Based on this, we provide a conjecture for the decay rate of the probability of interest without any proof.

This chapter is organized as follows. In Section 3.1 we introduce new notation and derive some preliminary results. Then in Section 3.2 we consider lower and upper bounds on the decay rate of the probability of interest and we conclude in Section 3.3 by showing in which cases the lower and upper bounds coincide and we provide some intuition on what we expect to be the decay rate.

3.1 Preliminaries

In this chapter, we consider the same model as in Chapter 2, that is, we consider a d-node $GI|GI|1$ tandem queue. However, in this chapter we assume that a

customer has a fixed size $0 < g_j < \infty$ at queue j. This means that the size of a customer can change when a customer has received service at queue $j < d$ and hence moves from queue j to queue $j + 1$, which may happen in for example a production process.

We are interested in the event that the total size of all customers reaches some high level N during a busy cycle of the system, and the related probabilities that this number exceeds N in stationarity and upon arrival of a customer. As a result, we cannot use the mathematical definition of the stopping time K_N in Equation (2.1), as it can make a difference for the total size of all customers at which queue a customer is waiting or being served.

The exception is that we *can* use (2.1) when a customer has the same size $g_j = c$ in all queues j. Intuitively, when all customers have the same size $g_j = c$ in all queues j, we expect that the decay rate of our probabilities of interest, compared to Chapter 2, is scaled by this constant factor c. We formally show this result in Lemma 3.1. Remark that when the size of all customers is equal in all queues j, we know that the total size of all customers is dividable by c. Therefore, in that case, there are $M = N/c$ customers in the system, each having size c.

Recall that the decay rate obtained in Chapter 2 equals $\Lambda_A(-\theta^*)$, that is, the log-moment generating function of the inter-arrival times, see also (2.5), and that $\theta^* = \min_j \theta^{(j)}$, where $\theta^{(j)}$ can be found by solving (2.3). Also, recall that L denotes the total number of customers in the system in stationarity and $L^{(a)}$ denotes the same number but immediately after an arbitrary arrival.

Lemma 3.1. *For the d-node $GI|GI|1$ tandem queue with $g_j = c \in (0, \infty)$ in all queues j, we have, under Assumption 2.3,*

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_M < K_0) &= \lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(L \geq M) \\ &= \lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(L^{(a)} \geq M) = \frac{1}{c}\Lambda_A(-\theta^*).\end{aligned}$$

Proof. As we have $N = cM$ and $c < \infty$, we find

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_M < K_0) = \lim_{M\to\infty} \frac{1}{cM} \log \mathbb{P}(K_M < K_0).$$

A similar equation holds for the decay of $\mathbb{P}(L \geq M)$ and $\mathbb{P}(L^{(a)} \geq M)$. The result follows by using Theorem 2.7. □

Since it is not always possible to use the definition of the stopping time for K_N in (2.1), we need another characterization of our event of interest. To this end, we let t_N denote the first time at which the total size of all customers equals N, and we let t_0 denote the first time a customer sees an empty system upon arrival. Recall that we start with one customer in the system, hence $t_0 > 0$. In the following lemma, we prove the obvious result that when $g_j = c$ in all queues j, the probability of the events $\{K_M < K_0\}$ and $\{t_N < t_0\}$ is the same. This will allow us to use the result from Lemma 3.1 to obtain bounds on the decay rate of the probability of interest.

Lemma 3.2. *If $g_j = c \in (0, \infty)$ for all j, it holds that $\mathbb{P}(K_M < K_0) = \mathbb{P}(t_N < t_0)$, where $M = N/c$.*

Proof. By definition we have $t_N = \sum_{k=1}^{K_M - 1} A_k$, where A_k is defined as the inter-arrival time between customer k and $k+1$, see also Chapter 2, and $t_0 = \sum_{k=1}^{K_0 - 1} A_k$, since customer K_0 is the index of the first customer to see an empty system upon arrival.

Suppose the event $\{K_M < K_0\}$ occurs, then we know that

$$t_N = \sum_{k=1}^{K_M - 1} A_k < \sum_{k=1}^{K_0 - 1} A_k = t_0,$$

and so it follows that the event $\{t_N < t_0\}$ occurs. On the other hand, when the event $\{K_M > K_0\}$ happens, we find

$$t_N = \sum_{k=1}^{K_M - 1} A_k > \sum_{k=1}^{K_0 - 1} A_k = t_0.$$

Thus, $\{K_M < K_0\}$ implies $\{t_N < t_0\}$, and $\{K_M > K_0\}$ implies $\{t_N > t_0\}$. This concludes the proof. □

3.2 Bounds on the rate of decay

Using the introduced notation and the results of Lemmas 3.1 and 3.2, we can derive bounds on the decay rate for the d-node tandem queue with unequal customer sizes. To this end, we let T denote the total size of all customers in the system in stationarity, and $T^{(a)}$ denote the same number but immediately after an (arbitrary) arriving customer. We start with deriving a lower bound for the decay of $\mathbb{P}(t_N < t_0)$, $\mathbb{P}(T \geq N)$ and $\mathbb{P}(T^{(a)} \geq N)$.

Lemma 3.3. *(Lower bound) For a d-node $GI|GI|1$ tandem queue with unequal customer sizes it holds that,*

$$\liminf_{N \to \infty} \frac{1}{N} \log \mathbb{P}(t_N < t_0) \geq \max_j \frac{1}{g_{\min,j}} \Lambda_A(-\theta^{(j)}),$$

where $g_{\min,j} = \min_{i \leq j} g_i$, and a similar statement holds when we replace $\mathbb{P}(t_N < t_0)$ by $\mathbb{P}(T \geq N)$ or $\mathbb{P}(T^{(a)} \geq N)$.

Proof. As in the proof of Lemma 2.4, we apply a coupling argument. That is, we compare the d-node tandem queue to a j-node tandem queue fed by the same arrival process. We will only prove the lower bound of the decay rate of $\mathbb{P}(t_N < t_0)$, the lower bounds of the other decay rates follow by similar arguments.

Let $t_N^{(j)}$ denote the first time that the total size of all customers in the j-node tandem queue equals N, and $t_0^{(j)}$ be the first time that the j-node tandem queue

is empty. Then we have that $t_N \leq t_N^{(j)}$ for all j, since it is impossible that the total size of all customers in the j-node tandem queue equals N *before* it does so for the system as a whole. Moreover, we have $t_0 \geq t_0^{(j)}$ for all j, since the system as a whole can *only* be empty when the j-node tandem queue is empty. Thus, we have

$$\mathbb{P}(t_N < t_0) \geq \mathbb{P}(t_N^{(j)} < t_0^{(j)}), \tag{3.1}$$

for all j. Next, we compare the j-node tandem queue with unequal customer sizes to a j-node tandem queue with equal customer sizes. We note that the number of customers that is necessary in order to obtain a total customer size of N in the j-node tandem queue is at most $\overline{M} = \lceil \frac{N}{g_{\min,j}} \rceil$, each having size $g_{\min,j}$. Thus,

$$\mathbb{P}(t_N^{(j)} < t_0^{(j)}) \geq \mathbb{P}(K_{\overline{M}}^{(j)} < K_0^{(j)}), \tag{3.2}$$

where $K_{\overline{M}}^{(j)}$ is defined analogously to $K_{\overline{M}}$, but for the j-node tandem queue, and $K_0^{(j)}$ is defined analogously to K_0, but for the j-node tandem queue. Thus, we find from (3.1) and (3.2) that

$$\begin{aligned}
\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}(t_N < t_0) &\geq \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}(t_N^{(j)} < t_0^{(j)}) \\
&\geq \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_{\overline{M}}^{(j)} < K_0^{(j)}) \\
&= \frac{1}{g_{\min,j}} \Lambda_A(-\theta_{\min,j}),
\end{aligned}$$

where the final equality follows from Lemma 3.1, applied to $c = g_{\min,j}$, and by noting that for the j-node tandem queue $\theta^* = \min_j \theta^{(j)}$ should be replaced by $\theta_{\min,j} = \min_{i \leq j} \theta^{(i)}$.

In particular, the tightest lower bound for the decay rate is the maximum value taken over $j = 1, \ldots, d$ of the right-hand side of this equation. Since by definition $\theta_{\min,j} \leq \theta^{(j)}$ for all j, and thus we find $\Lambda_A(-\theta_{\min,j}) \geq \Lambda_A(-\theta^{(j)})$, which concludes the proof. □

By comparing the system to a j-node tandem queue with equal customer sizes, rather than to a d-node tandem queue with equal customer sizes, the obtained lower bound is tighter. In particular, when $g_1 \geq g_2 \geq \ldots \geq g_d$ we find that the lower bound equals $\max_j \frac{1}{g_j} \Lambda_A(-\theta^{(j)})$.

Next, we derive an upper bound for the decay of $\mathbb{P}(t_N < t_0)$, $\mathbb{P}(T \geq N)$ and $\mathbb{P}(T^{(a)} \geq N)$.

Lemma 3.4. *(Upper bound) For a d-node $GI|GI|1$ tandem queue with unequal customer sizes it holds that,*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(t_N < t_0) \leq \frac{1}{g_{\max}} \Lambda_A(-\theta^*),$$

where $g_{\max} = \max_j g_j$*, and a similar statement holds when we replace* $\mathbb{P}(t_N < t_0)$ *by* $\mathbb{P}(T \geq N)$ *or* $\mathbb{P}(T^{(a)} \geq N)$*.*

Proof. We will only prove the upper bound of the decay rate of $\mathbb{P}(t_N < t_0)$, the upper bounds of the other decay rates follow by similar arguments. The total size of all customers in the queue can only be at least N when there have been at least $\underline{M} = \lceil \frac{N}{g_{\max}} \rceil$ customers in the system, each having size $g_{\max}$. Thus we find, by using Lemma 3.2,

$$\begin{aligned}\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(t_N < t_0) &\leq \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_{\underline{M}} < K_0) \\ &= \frac{1}{g_{\max}} \Lambda_A(-\theta^*),\end{aligned}$$

where the equality follows from Lemma 3.1, applied to $c = g_{\max}$. □

3.3 Main result

In this section, we start by summarizing our findings from the previous section.

Theorem 3.5. *For a d-node* $GI|GI|1$ *tandem queue with unequal customer sizes, we have, under Assumption 2.3,*

$$\max_j \frac{1}{g_{\min,j}} \Lambda_A(-\theta_j) \leq \limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(t_N < t_0) \leq \frac{1}{g_{\max}} \Lambda_A(-\theta^*),$$

and similar statements holds when we replace $\mathbb{P}(t_N < t_0)$ *by* $\mathbb{P}(T \geq N)$ *or* $\mathbb{P}(T^{(a)} \geq N)$ *and/or when we replace the* $\limsup$ *by* $\liminf$*.*

We remark that, when we consider equal customer sizes, both bounds coincide and we obtain the same result as in Lemma 3.1 (which should be the case, since both bounds are obtained using this lemma). A more interesting case where our lower and upper bounds coincide is when the θ-bottleneck queue, see Definition 2.1, is queue 1 and when queue 1 is the queue with the largest customer sizes, that is, $g_{\max} = g_1$. This is the topic of the following corollary.

Corollary 3.6. *For a d-node* $GI|GI|1$ *tandem queue with* $\theta^* = \theta^{(1)}$ *and* $g_{\max} = g_1$*, we have, under Assumption 2.3,*

$$\begin{aligned}\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(t_N < t_0) &= \lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(T \geq N) \\ &= \lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(T^{(a)} \geq N) = \frac{1}{g_1} \Lambda_A(-\theta^{(1)}).\end{aligned}$$

Finally, we present a conjecture for the decay rate for the probability of interest, based on the results of this chapter. This decay rate is equal to the lower bound that has been derived in Lemma 3.3 when queue 1 is the θ-bottleneck queue or when $g_1 \geq g_2 \geq \ldots \geq g_d$. The intuition is provided below.

Conjecture 3.7. *For a d-node $GI|GI|1$ tandem queue with unequal customer sizes, we expect, under Assumption 2.3,*

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}(t_N < t_0) &= \lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}(T \geq N)\\ &= \lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}(T^{(a)} \geq N) = \max_j \frac{1}{g_j}\Lambda_A(-\theta^{(j)}).\end{aligned}$$

Intuition. Since customers have different sizes in different queues, it is not the θ-bottleneck that determines the decay rate but rather a combination of $\Lambda_A(-\theta^{(j)})$ and the size of the customers at that particular queue j. For example, consider a 2-node tandem queue and suppose $\theta^{(2)} = \theta^{(1)} + \varepsilon$, where $\varepsilon > 0$ is small. When $g_1 = g_2$, queue 1 will determine the rate of decay. However, when $g_2 > g_1$, the largest contribution towards the total number of customers in the system will be from queue 2 and hence the combination of $\Lambda_A(-\theta^{(2)})$ and g_2 will determine the decay rate. Thus, we could say that queue 2 is the g-θ-bottleneck queue. Intuitively, also in the more general case, we expect that the g-θ-bottleneck queue will determine the decay rate. □

CHAPTER 4

State-independent importance sampling for non-Markovian tandem queues

In this chapter, we consider importance sampling for $GI|GI|1$ tandem queues. Recall from Chapter 1 that we are interested in estimating the probability that in a busy cycle of the queueing system the total number of customers reaches some high level N.

As was mentioned in the introduction of this thesis, one of the first papers to consider importance sampling in queueing networks is by Parekh and Walrand [35]. Their interest is in the same probability as in the current chapter. To estimate this probability for the single queue, they propose a simple, explicit, change of measure, and for networks of queues they implicitly describe how to find a change of measure. Their proposed change of measure is state-independent, that is, the change of measure does not depend on the current state of the system. In the remainder of this thesis, the change of measure proposed by Parekh and Walrand will be referred to as the P&W change of measure. To determine this change of measure, an equation needs to be solved. Frater and Anderson [25] partially solved the equation proposed by P&W, resulting in a simpler but still implicit description of the P&W state-independent change of measure for a class of $GI|GI|1$ tandem queues.

In this chapter, another, simpler, method is considered to obtain a change of measure for $GI|GI|1$ tandem queues, based on knowledge of the decay rate of the probability of interest which has been determined in Chapter 2. This method is not implicit, and we show that it is equivalent to the earlier method from P&W, in the sense that it results in the same change of measure. Secondly, we show that the change of measure proposed by P&W is the only exponential state-independent change of measure that may give an asymptotically efficient estimator. Lastly, we provide necessary conditions for this exponential state-independent change of measure to give an asymptotically efficient estimator.

Based on results for the $M|M|1$ tandem queue in [12, 30], it is clear that for the $GI|GI|1$ tandem queue the P&W change of measure does not always give an asymptotically efficient estimator. In [19, 22], Dupuis et al. prove that a certain state-*dependent* change of measure is asymptotically efficient for Markovian networks. This change of measure roughly coincides with the P&W change of measure in most of the state space, but deviates from it near the edges. We

expect the same for the $GI|GI|1$ case, which motivates our interest in the (state-independent) P&W change of measure: even though it fails to be asymptotically efficient in some cases, it seems plausible that it will be an important ingredient for any asymptotically efficient state-dependent change of measure. In Chapter 5 we will see that this is indeed the case.

This chapter is structured as follows. In Section 4.1, we introduce the model and the change of measure as derived from the decay rate obtained in Chapter 2. In Section 4.2 we show that this is equivalent to the change of measure of Frater and Anderson [25] and thus to that of P&W. In Section 4.3 we show that this P&W change of measure is the only exponential state-independent change of measure that can give an asymptotically efficient estimator. Other necessary conditions for the state-independent change of measure to give an asymptotically efficient estimator are presented in Section 4.4. In Section 4.5 we give some numerical results and the conclusions are presented in Section 4.6.

4.1 Model and preliminaries

4.1.1 The model

In this chapter, we again consider d $GI|GI|1$ queues in tandem; in Section 4.4 and 4.5 we consider the special case $d = 2$. For a detailed introduction of the model, the notation and the assumptions, we refer to Section 2.1. We recall that K_N is defined in (2.1) as the index of the first customer who reaches the overflow level N. Likewise, K_0 is defined in (2.2) as the index of the first customer after customer 1 who sees an empty system upon arrival. Let $\mathcal{K} = \min(K_0, K_N)$. Then the indicator $\mathbb{1}\{\mathcal{K} = K_N\}$ defines if we have reached our rare event in the busy cycle or not, and the probability of this rare event, denoted by p_N, is equal to $\mathbb{E}[\mathbb{1}\{\mathcal{K} = K_N\}]$. We note that this notation is slightly different from (but not inconsistent with) that of Chapter 2. The introduction of $\mathcal{K}$ is new and necessary in the current chapter.

In addition to the notation introduced in Chapter 2, we denote the distribution functions of A_k and $B_k^{(j)}$ by F_A and $F_{B^{(j)}}$ respectively, and their moment generating functions by $M_A(t)$ and $M_{B^{(j)}}(t)$; note that in Chapter 2 we have defined $\Lambda_A(t) = \log M_A(t)$ and $\Lambda_{B^{(j)}}(t) = \log M_{B^{(j)}}(t)$, which we will also use in this chapter.

4.1.2 Importance sampling simulation

In importance sampling, the rare event is made less rare by changing the underlying probability distribution. For a single $GI|GI|1$ queue, so $d = 1$, it is suggested by Parekh and Walrand [35] to apply an exponential tilt $\theta = \theta^{(1)}$, with $\theta^{(1)}$ as in (2.3), for both the inter-arrival times and the service times in the following

way,

$$dF_A^\theta(a) = \frac{e^{-\theta a}}{M_A(-\theta)} dF_A(a), \tag{4.1}$$

$$dF_{B^{(1)}}^\theta(b) = \frac{e^{\theta b}}{M_{B^{(1)}}(\theta)} dF_{B^{(1)}}(b), \tag{4.2}$$

where $F_A^\theta(a)$ and $F_{B^{(1)}}^\theta(b)$ denote the distribution functions under the change of measure. It is shown by Sadowsky in [36] that this change of measure results in an asymptotically efficient estimator, assuming that $\mathbb{E}\left[B^{(1)}\right] < \mathbb{E}\left[A\right]$ (stability), that $\mathbb{P}(B^{(1)} > A) > 0$ (non-triviality), and that $\mathbb{P}(B^{(1)} < M) = 1$ for some finite constant M (bounded service times). The last assumption is the only real restriction, but it was claimed that this is a mere technicality, and not essential for the result to hold.

Let us now consider d $GI|GI|1$ queues in tandem and let $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_d)$ be a vector of exponential tilts. Then $\mathbb{E}^{\boldsymbol{\theta}}[\cdot]$ and $\mathbb{P}^{\boldsymbol{\theta}}(\cdot)$ denote expected values and probabilities under this change of measure $\boldsymbol{\theta}$, and we denote the distribution function and the moment generating function of a random variable X under this change of measure as $F_X^{\boldsymbol{\theta}}(x) = \mathbb{P}^{\boldsymbol{\theta}}(X \leq x)$ and $M_X^{\boldsymbol{\theta}}(t) = \mathbb{E}^{\boldsymbol{\theta}}\left[e^{tX}\right]$, respectively. For the distribution functions of A and $B^{(j)}$ we have

$$dF_A^{\boldsymbol{\theta}}(a) = \frac{e^{-\theta_0 a}}{M_A(-\theta_0)} dF_A(a), \tag{4.3}$$

$$dF_{B^{(j)}}^{\boldsymbol{\theta}}(b) = \frac{e^{\theta_j b}}{M_{B^{(j)}}(\theta_j)} dF_{B^{(j)}}(b), \qquad j = 1, \ldots, d. \tag{4.4}$$

Note that the difference compared with Equation (4.1)–(4.2) is that now the inter-arrival time distribution and service time distributions of all queues j may be tilted differently. As a result, the moment generating functions of A and $B^{(j)}$ under the change of measure $\boldsymbol{\theta}$ are

$$M_A^{\boldsymbol{\theta}}(t) = \frac{M_A(t-\theta_0)}{M_A(-\theta_0)}, \tag{4.5}$$

$$M_{B^{(j)}}^{\boldsymbol{\theta}}(t) = \frac{M_{B^{(j)}}(t+\theta_j)}{M_{B^{(j)}}(\theta_j)}, \qquad j = 1, \ldots, d,$$

and we find the expected values under the change of measure $\boldsymbol{\theta}$ to be

$$\mathbb{E}^{\boldsymbol{\theta}}[A] = \frac{M_A'(-\theta_0)}{M_A(-\theta_0)} = \frac{-d\Lambda_A}{d\theta}(-\theta_0), \tag{4.6}$$

$$\mathbb{E}^{\boldsymbol{\theta}}\left[B^{(j)}\right] = \frac{M_{B^{(j)}}'(\theta_j)}{M_{B^{(j)}}(\theta_j)} = \frac{d\Lambda_{B^{(j)}}}{d\theta}(\theta_j), \qquad j = 1, \ldots, d. \tag{4.7}$$

In the sequel it will become clear that the "best" tilt $\boldsymbol{\theta^*} = (\theta_0^*, \ldots, \theta_d^*)$ is such that $\theta_0^* = \theta_{j^*}^*$ and $\theta_j^* = 0$, $j \neq 0, j^*$, where j^* is the bottleneck queue in the sense of Definition 2.1. Next, we discuss how to find queue j^* and the tilt-parameter $\theta_{j^*}^*$. In Section 4.2 it turns out that the change of measure described above is in fact the P&W change of measure, although this is not immediately clear from [35].

4.1.3 Specific change of measure $\boldsymbol{\theta^*}$

Based on our knowledge of the decay rate determined in Chapter 2, see (2.5), we propose a specific change of measure. We start by solving the equations in (2.3), then we know by Definition 2.1 that queue j^* is such that $\theta^{(j^*)} = \min_j \theta^{(j)} = \theta^*$. In this chapter, we make the following (additional) assumptions.

Assumption 4.1. *In addition to Assumption 2.3 , we assume that*

- *the bottleneck queue is unique, that is, $\theta^* < \theta^{(j)}$ for all $j \neq j^*$; and*
- *the inequality in definition* (2.3) *for $j = j^*$ holds with equality:*

$$\Lambda_A(-\theta^*) + \Lambda_{B^{(j^*)}}(\theta^*) = 0,$$

or equivalently,

$$M_A(-\theta^*)M_{B^{(j^*)}}(\theta^*) = 1. \tag{4.8}$$

In Section 4.8 we remark on what happens when the inequality in definition (2.3) does not hold with equality.

Due to the uniqueness assumption, we are now ready to introduce the change of measure based on Chapter 2: it is simply a $\boldsymbol{\theta}$-tilt as given in (4.3)–(4.4), where we choose the exponential tilt to be $\boldsymbol{\theta} = \boldsymbol{\theta^*}$ with $\boldsymbol{\theta^*} = (\theta^*, 0, \ldots, 0, \theta^*, 0, \ldots, 0)$. This means that we only tilt the inter-arrival times and the service times of the bottleneck queue j^*, with the same tilting parameter θ^*.

We will refer to this change of measure as *the $\boldsymbol{\theta^*}$-tilt*. As mentioned earlier, in Section 4.2 we will show that this $\boldsymbol{\theta^*}$-tilt coincides with the P&W change of measure for cases in which this is properly defined (that is, when (4.8) holds), and in Section 4.3 we will show that it is the only reasonable exponential change of measure, since taking $\boldsymbol{\theta} \neq \boldsymbol{\theta^*}$ will result in an estimator that is not asymptotically efficient.

4.2 Comparison with Frater and Anderson

In this section we compare our method to obtain j^* and θ^* for the change of measure for the $GI|GI|1$ tandem queue to the earlier developed method by Frater and Anderson [25]. They presented one way to obtain a change of measure for the $GI|GI|1$ tandem queue in the early 90s. Their method is based on Parekh and Walrand [35] and is written in an implicit form. In Section 4.2.1 we present the method of Frater and Anderson; then in Section 4.2.2 we show that the two are equivalent in all cases where they are properly defined.

4.2.1 Method by Frater and Anderson [25]

In [25], the change of measure proposed by Parekh and Walrand is further explored. Based on large deviations theory, Parekh and Walrand defined a cost

function H that needs to be minimized in order to find the change of measure. Frater and Anderson simplify this function (see (37) in [25]) to

$$H(\lambda_1', \mu_1', \ldots, \mu_d', R) = \frac{1}{\lambda_1' - \mu_R'} \left[\lambda_1' h_A\left(\frac{1}{\lambda_1'}\right) + \sum_{j=1}^{d} \mu_j' h_{B^{(j)}}\left(\frac{1}{\mu_j'}\right) \right], \quad (4.9)$$

where λ_1' is the arrival rate at queue 1 and μ_j' the service rate of queue j, where each rate is just the inverse of the corresponding expectation, and where the primes denote that the values should be optimized to find the change of measure. Furthermore, $h_A(\cdot)$ and $h_{B^{(j)}}(\cdot)$ denote the Cramér transforms of the inter-arrival time distribution and the service time distribution at queue j respectively (where the Cramér transform of a random variable X is defined as $h_X(y) = \sup_s [sy - \log M_X(s)]$). Finally, R is the index of the *rightmost unstable queue* under the change of measure, that is, R is the largest index j for which $\mu_j' < \lambda_1'$ under the change of measure. (Note that Frater and Anderson write M instead of R.)

Then they explain how to find the minimum of (4.9). They show that for all queues $j \neq R$ the optimal value of μ_j' is μ_j and since $h_{B^{(j)}}(1/\mu_j) = 0$ (see [35]) this implies that H reduces to a function of λ_1', μ_R' and R in the following way,

$$H(\lambda_1', \mu_R', R) = \frac{1}{\lambda_1' - \mu_R'} \left[\lambda_1' h_A\left(\frac{1}{\lambda_1'}\right) + \mu_R' h_{B^{(R)}}\left(\frac{1}{\mu_R'}\right) \right], \quad (4.10)$$

see (43) in [25]. Next, they note the two problems that remain in order to find the change of measure:

1. to find the value of R that is optimal, that is, the value of R that minimizes $H(\lambda_1', \mu_R', R)$,

2. given R, to find the values of λ_1' and μ_R' that minimize $H(\lambda_1', \mu_R', R)$.

Assuming the first problem is solved, that is, given R, the solution of the second problem is not hard, using a similar method as for the single $GI|GI|1$ queue, and Frater and Anderson show how to obtain the optimal values of λ_1' and μ_R', referring to [35]. From these values, again using [35], the change of measure now follows, which prescribes exponential tilting of the distributions such that their rates become equal to the optimal rates. This change of measure turns out to be precisely as in (4.3)–(4.4) above, with the tilting vector given by $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}} \equiv (\widetilde{\theta}^{(R)}, 0, \ldots, 0, \widetilde{\theta}^{(R)}, 0, \ldots, 0)$, with $\widetilde{\theta}_j = 0$ for all $j \neq 0, R$, and $\widetilde{\theta}_0 = \widetilde{\theta}_R = \widetilde{\theta}^{(R)} > 0$, where the latter is such that it satisfies

$$\Lambda_A(-\widetilde{\theta}^{(R)}) + \Lambda_{B^{(R)}}(\widetilde{\theta}^{(R)}) = 0. \quad (4.11)$$

As a result, the expectations of A and $B^{(R)}$ under the change of measure become,

using (4.6) and (4.7),

$$1/\lambda_1' = \mathbb{E}^{\widetilde{\boldsymbol{\theta}}}[A] \quad = \frac{-d\Lambda_A}{d\theta}(-\widetilde{\theta}^{(R)}), \quad \text{and}$$
$$1/\mu_R' = \mathbb{E}^{\widetilde{\boldsymbol{\theta}}}\left[B^{(R)}\right] = \frac{d\Lambda_{B^{(R)}}}{d\theta}(\widetilde{\theta}^{(R)}),$$

as they should, so given the optimal value of R the problem is solved.

However, finding the optimal R is difficult since (the index of) the rightmost unstable queue under the change of measure depends on this change of measure itself. Only for a certain class of problems, Frater and Anderson show that R can be chosen simply as the 'rho-bottleneck' (see Remark 2.2 in Chapter 2). For the general case they need to calculate for each possible value of R the optimal λ_1' and μ_R', and then substitute these in $H(\lambda_1', \mu_R', R)$ to obtain a function $H(R)$ that only depends on R, after which the optimal value $\widetilde{R}$ needs to be picked such that it minimizes $H(R)$. When there are multiple candidates for R they seem to suggest that R should be chosen as large as possible, but this is not entirely clear to us.

Finally, it has to be checked whether under the resulting change of measure $\widetilde{\boldsymbol{\theta}}$, the corresponding $\widetilde{R}$ is indeed the rightmost unstable queue. If this is not the case it is not clear how to proceed, but we will show in Section 4.2.2 that $\widetilde{R}$ is indeed the rightmost unstable queue under the change of measure.

4.2.2 Comparison of the two methods

In this section we show that the method based on the decay rate (as described in Section 4.1.3) and the method by Frater and Anderson [25] (as described in Section 4.2.1) are equivalent. First of all it is clear that both methods consider the same type of exponential tilting based on (4.3)–(4.4), and that the tilting vectors $\boldsymbol{\theta}^*$ and $\widetilde{\boldsymbol{\theta}}$ have the same structure, so that only the inter-arrival times and the service times of one of the queues are tilted. Frater and Anderson find optimal values for λ_1' and μ_R', where R is the particular queue to be tilted, but as described above this optimization is equivalent to finding the corresponding $\widetilde{\theta}^{(R)}$. (In fact, their use of λ_1' and μ_j', $j = 1, \ldots, d$ in minimizing (4.9) and (4.10) can be seen as an alternative (one-to-one) parametrization to optimize the tilting parameters θ_0 and θ_j, $j = 1, \ldots, d$.) Given the optimal value of R, the value of the tilting parameter $\widetilde{\theta}^{(R)}$ is given in the same way as the $\theta^{(j)}$ in our method, compare (4.11) with (2.3), and note that (4.11) also shows that [25] and [35] only consider cases in which (2.3) holds with equality (as we assume for our j^*, see Assumption 4.1).

As a consequence, the change of measure is exactly the same for both methods *if* the same queue is tilted. Therefore we only need to show that the bottleneck queue j^* as described in Section 4.1.3 minimizes the function $H(R)$, and then do the 'Frater and Anderson check' to see if queue j^* is indeed the rightmost unstable queue under the change of measure, as described in Section 4.2.1. We show these statements in the following two lemmas.

The first lemma relates the $\boldsymbol{\theta^*}$-bottleneck queue to minimizing $H(R)$. We start by briefly motivating how to rewrite $H(R)$. As mentioned, for fixed R, Frater and Anderson choose the values for λ' and μ'_R which minimize the function $H(\lambda'_1, \mu'_R, R)$ in (4.10). The optimization is done in exactly the same manner as was done by Parekh and Walrand in [35] for the single $GI|GI|1$ queue. We will not copy the details but only mention they set the partial derivatives of $H(\lambda'_1, \mu'_R, R)$ with respect to λ' and μ'_R equal to zero, and combine this with properties of the Cramér transform and with the implicit assumption that (4.11) holds; for more details see Equations (37)-(44) in [35]. The result is simply that $H(R)$ can be written as

$$H(R) = -\Lambda_A(-\widetilde{\theta}^{(R)}).$$

Lemma 4.2. *$H(j^*) < H(j)$ for all $j \neq j^*$.*

Proof. As mentioned before, $\widetilde{\theta}^{(R)}$ coincides with our θ^* when $j^* = R$. Indeed, $H(j) = -\Lambda_A(-\theta^{(j)})$ is minimal for the choice $j = j^*$ since $\theta^* < \theta^{(j)}$ for all $j \neq j^*$, and $-\Lambda_A(-\theta)$ is a strictly increasing function of θ. □

In the second lemma we check that queue j^* is the rightmost unstable queue in the $\boldsymbol{\theta^*}$-tilted system.

Lemma 4.3. *Under Assumption 4.1, queue j^* is the rightmost unstable queue in the $\boldsymbol{\theta^*}$-tilted system and in particular $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$.*

Proof. We show that: (i) queue j^* is unstable under the $\boldsymbol{\theta^*}$-tilt and $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$; and (ii) all queues $k > j^*$ are stable under the $\boldsymbol{\theta^*}$-tilt, which proves the lemma.

(i) We say that a queue is unstable when the service rate of that queue is smaller than the local arrival rate to that queue. Under the $\boldsymbol{\theta^*}$-tilt the service rate at queue j^* is $1/\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]$, while the arrival rate at queue j^* is $\min\{1/\mathbb{E}^{\boldsymbol{\theta^*}}[A], 1/\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(1)}\right], \ldots, 1/\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*-1)}\right]\}$. We will show that both $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$ and $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] > \mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(k)}\right]$ for all $k = 1, \ldots, j^* - 1$, implying instability of queue j^*. To show that $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$ we let $f(\theta) = \Lambda_A(-\theta) + \Lambda_{B^{(j^*)}}(\theta)$. We know that $f(0) = f(\theta^*) = 0$ and $f'(0) < 0$. By convexity of the log moment generating functions it must hold that $f'(\theta^*) > 0$ and so it follows, using (4.6) and (4.7), that $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$. We conclude this part of the proof by showing that $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] > \mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(k)}\right] = \mathbb{E}\left[B^{(k)}\right]$ for all $k = 1, \ldots, j^* - 1$. For a graphical interpretation, see Figure 4.1.

Again using (4.6) and (4.7), we have for any $k \neq j^*$

$$\begin{aligned}\mathbb{E}\left[B^{(k)}\right] &= \frac{d\Lambda_{B^{(k)}}(0)}{d\theta} \leq \frac{\Lambda_{B^{(k)}}(\theta^*)}{\theta^*} \\ &< \frac{\Lambda_{B^{(j^*)}}(\theta^*)}{\theta^*} \leq \frac{d\Lambda_{B^{(j^*)}}(\theta^*)}{d\theta} = \mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right], \end{aligned} \tag{4.12}$$

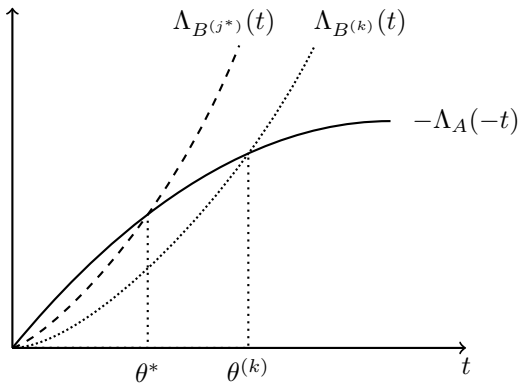


Figure 4.1 A graphical interpretation of the inequalities presented in (4.12).

where the first and the final inequality follow from convexity of $\Lambda_{B^{(k)}}(\theta)$ and $\Lambda_{B^{(j^*)}}(\theta)$, and the second inequality follows by definition and uniqueness of θ^* (that is, if $\Lambda_{B^{(j^*)}}(\theta^*) > \Lambda_{B^{(k)}}(\theta^*)$ queue k would be the bottleneck queue instead of queue j^*, and if $\Lambda_{B^{(j^*)}}(\theta^*) = \Lambda_{B^{(k)}}(\theta^*)$ the bottleneck queue would not be unique). Hence we have that queue j^* is unstable under the $\boldsymbol{\theta^*}$-tilt and in particular $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] > \mathbb{E}^{\boldsymbol{\theta^*}}\left[A\right]$.

(ii) Finally, we show that queue j^* is the *rightmost* unstable queue under the $\boldsymbol{\theta^*}$-tilt. If $j^* = d$ this statement is trivial, so suppose for the remainder of the proof that $j^* < d$. By (i), the arrival rate for queue $j^* + 1$ is equal to the service rate of the unstable queue j^*. Queue $j^* + 1$ is stable, as we have $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*+1)}\right] = \mathbb{E}\left[B^{(j^*+1)}\right] < \mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]$ by Equation (4.12). Since queue $j^* + 1$ is stable, the arrival rate for queue $j^* + 2$ also equals the service rate of queue j^*. Now look at any queue $k \in [j^* + 2, d]$ (if any). If all queues between j^* and k are stable, queue k is also stable because $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(k)}\right] = \mathbb{E}\left[B^{(k)}\right] < \mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]$, which follows immediately from Equation (4.12), and hence the arrival rate at queue k equals the service rate of queue j^*. Thus, by induction, the result follows. □

Theorem 4.4. *Under Assumption 4.1 and when $\widetilde{R}$ is unique, we have $j^* = \widetilde{R}$, and hence the $\boldsymbol{\theta^*}$-tilt described in Section 4.1.3 and the P&W method as described in Frater and Anderson give the same change of measure.*

Proof. Consider the $\boldsymbol{\theta^*}$-tilt with bottleneck queue j^*, then j^* is the rightmost unstable queue in the $\boldsymbol{\theta^*}$-tilted system by Lemma 4.3. We also know, in view of the uniqueness of j^*, that $\theta^* < \min_{j \neq j^*} \theta^{(j)}$ and by Lemma 4.2 that j^* minimizes $H(R)$. Hence $j^* = \widetilde{R}$. The equivalence of the two corresponding changes of measure is then immediate from the fact that both are based on (4.3)–(4.4) with $\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta^*}$. □

We note that $-H(j^*) = \Lambda_A(-\theta^*)$ is the rate of decay, see (2.5). This implies

that the large deviations approximation made in Parekh and Walrand is actually good.

4.3 The θ-tilt is not asymptotically efficient when $\theta \neq \theta^*$

Having determined that the $\boldsymbol{\theta}$-tilt is the same as the P&W change of measure by Parekh and Walrand [35], in this section we show that it is the only exponential state-independent change of measure that may give an asymptotically efficient estimator. In Section 4.3.1 we introduce the likelihood ratio $L^{\boldsymbol{\theta}}$ of a path that reaches level N in a busy cycle of the system and give the mathematical definition of asymptotic efficiency in terms of the second moment of this random variable. Then in Section 4.3.2 we show the main result of this section, Theorem 4.9.

4.3.1 Definitions

Suppose we use the exponential change of measure $\boldsymbol{\theta}$. Remembering that by definition we have $\mathcal{K} = \min(K_0, K_N)$, we let the likelihood ratio $L^{\boldsymbol{\theta}}$ of a path that consists of $\mathcal{K}$ arrivals be,

$$L^{\boldsymbol{\theta}} = \prod_{k=1}^{\mathcal{K}-1} \frac{dF_A}{dF_A^{\boldsymbol{\theta}}}(A_k) \prod_{j=1}^{d} \prod_{k=1}^{k_j} \frac{dF_{B^{(j)}}}{dF_{B^{(j)}}^{\boldsymbol{\theta}}}(B_k^{(j)}). \tag{4.13}$$

Here, k_j is the number of initiated services in queue j just before the $\mathcal{K}$-th arrival, formally defined as $k_j = \mathcal{K} - 1 - \sum_{k=1}^{j} n_k + \mathbb{1}\{n_j > 0\}$ for $j = 1, \ldots, d$, where n_j is the number of customers in queue j just before the $\mathcal{K}$-th arrival. When $\mathcal{K} = K_N$ it holds that $\sum_{k=1}^{d} n_k = N - 1$, so in that case we can also write $k_j = \mathcal{K} - N + \sum_{k=j+1}^{d} n_k + \mathbb{1}\{n_j > 0\}$.

Remark 4.5. *In principle, one could reduce the estimator variance a bit further by dividing the likelihood ratio in* (4.13) *by the likelihood ratio of the remaining service times upon reaching level N (but for a clearer presentation we decided not to do this).*

Under the tilt $\boldsymbol{\theta}$, $L^{\boldsymbol{\theta}}\mathbb{1}\{\mathcal{K} = K_N\}$ is an unbiased estimator for p_N, that is, $p_N = \mathbb{E}^{\boldsymbol{\theta}}\left[L^{\boldsymbol{\theta}}\mathbb{1}\{\mathcal{K} = K_N\}\right]$. The goal of importance sampling simulation is to get an asymptotically efficient estimator, which can be defined as follows (see also Definition 1.2 or [30]).

Definition 4.6. *An unbiased estimator is asymptotically efficient if*

$$\liminf_{N\to\infty} \frac{\log \mathbb{E}^{\boldsymbol{\theta}}\left[\left(L^{\boldsymbol{\theta}}\right)^2 \mathbb{1}\{\mathcal{K} = K_N\}\right]}{\log p_N} \geq 2.$$

Note that we always have $\limsup_{N\to\infty} \log \mathbb{E}^{\boldsymbol{\theta}}\left[\left(L^{\boldsymbol{\theta}}\right)^2 \mathbb{1}\{\mathcal{K}=K_N\}\right] / \log p_N \leq 2$ by Jensen's inequality. Hence, alternatively we could replace the inequality in Definition 4.6 by an equality sign (and the liminf by a limit).
The meaning of the definition is that for an asymptotically efficient estimator, the second moment vanishes at twice the rate of the estimator itself. As a consequence, the relative error increases sub-exponentially.

Using (2.5), we find that the estimator is asymptotically efficient if

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}}\left[\left(L^{\boldsymbol{\theta}}\right)^2 \mathbb{1}\{\mathcal{K}=K_N\}\right] \leq 2\Lambda_A(-\theta^*). \tag{4.14}$$

4.3.2 Main result

In this section we show that using an exponential tilt other than the P&W change of measure cannot give an asymptotically efficient estimator. By the above we need to show that an estimator based on the tilt $\boldsymbol{\theta} \neq \boldsymbol{\theta^*}$ satisfies

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}}\left[\left(L^{\boldsymbol{\theta}}\right)^2 \mathbb{1}\{\mathcal{K}=K_N\}\right] > 2\Lambda_A(-\theta^*). \tag{4.15}$$

Before we state the theorem, we need the following lemmas. Even though the statements seem obvious, they are not entirely trivial (especially the first one when $d > 1$); we present the proofs in Section 4.7.

Lemma 4.7. *Suppose we have d $GI|GI|1$ queues in tandem. Under the change of measure $\boldsymbol{\theta^*}$ for which $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$, we have for all N that $\mathbb{P}^{\boldsymbol{\theta^*}}(\mathcal{K}=K_N) \geq \mathbb{P}^{\boldsymbol{\theta^*}}(E) > 0$, where E is the event that the system never empties. Moreover, we have as $N\to\infty$ that $\frac{K_N}{N} \to \frac{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]}{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]-\mathbb{E}^{\boldsymbol{\theta^*}}[A]}$ with probability 1.*

Lemma 4.8. *Consider a sequence $\{X_N\}$ of random variables that converges to a constant c with probability 1 as $N\to\infty$ and let E be an event with $\mathbb{P}(E) > 0$. Then $\mathbb{E}[\lim_{N\to\infty} X_N \mid E] = \mathbb{E}[\lim_{N\to\infty} X_N] = c$.*

Now we are ready to prove our theorem.

Theorem 4.9. *Consider d $GI|GI|1$ queues in tandem. Under Assumption 4.1 the $\boldsymbol{\theta^*}$-tilt is the only exponential state-independent change of measure that can possibly give an asymptotically efficient estimator.*

Proof. Consider an exponential tilt $\boldsymbol{\theta} \neq \boldsymbol{\theta^*}$, then the goal is to show (4.15) when $\boldsymbol{\theta} \neq \boldsymbol{\theta^*}$. To rewrite the second moment of the likelihood ratio in terms of the expectation under $\boldsymbol{\theta^*}$, rather than $\boldsymbol{\theta}$, notice that

$$\mathbb{E}^{\boldsymbol{\theta}}\left[\left(L^{\boldsymbol{\theta}}\right)^2 \mathbb{1}\{\mathcal{K}=K_N\}\right] = \mathbb{E}^{\boldsymbol{\theta^*}}\left[L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta^*}} \mathbb{1}\{\mathcal{K}=K_N\}\right].$$

Let E denote the event that the system never empties, as in Lemma 4.7. We find

$$\begin{aligned}
&\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}^*} \left[L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*} \mathbb{1}\{\mathcal{K} = K_N\} \right] \\
&\geq \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta}^*} \left[L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*} \middle| E \right] + \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}^{\boldsymbol{\theta}^*}(E) \\
&\geq \liminf_{N\to\infty} \log \mathbb{E}^{\boldsymbol{\theta}^*} \left[\left(L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*} \right)^{\frac{1}{N}} \middle| E \right] \quad \text{(by Jensen's inequality and Lemma 4.7)} \\
&\geq \log \mathbb{E}^{\boldsymbol{\theta}^*} \left[\liminf_{N\to\infty} \left(L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*} \right)^{\frac{1}{N}} \middle| E \right] \quad \text{(by Fatou's Lemma)} \\
&\geq \mathbb{E}^{\boldsymbol{\theta}^*} \left[\liminf_{N\to\infty} \frac{1}{N} \log \left(L^{\boldsymbol{\theta}} L^{\boldsymbol{\theta}^*} \right) \middle| E \right] \quad \text{(by Jensen's inequality)} \\
&\geq \mathbb{E}^{\boldsymbol{\theta}^*} \left[\liminf_{N\to\infty} \frac{1}{N} \log L^{\boldsymbol{\theta}} \middle| E \right] + \mathbb{E}^{\boldsymbol{\theta}^*} \left[\liminf_{N\to\infty} \frac{1}{N} \log L^{\boldsymbol{\theta}^*} \middle| E \right]. \qquad (4.16)
\end{aligned}$$

From (4.13) and then (4.3)–(4.4) it follows that

$$\frac{1}{N} \log L^{\boldsymbol{\theta}} = \Lambda_A(-\theta_0) \frac{\mathcal{K}-1}{N} + \frac{\theta_0}{N} \sum_{k=1}^{\mathcal{K}-1} A_k + \sum_{j=1}^{d} \left(\Lambda_{B^{(j)}}(\theta_j) \frac{k_j}{N} - \frac{\theta_j}{N} \sum_{k=1}^{k_j} B_k^{(j)} \right),$$

and so the first term of the right-hand side of (4.16) is greater than or equal to

$$\begin{aligned}
f(\boldsymbol{\theta}) = \mathbb{E}^{\boldsymbol{\theta}^*} &\left[\liminf_{N\to\infty} \left(\Lambda_A(-\theta_0) \frac{\mathcal{K}-1}{N} + \frac{\theta_0}{N} \sum_{k=1}^{\mathcal{K}-1} A_k \right) \middle| E \right] \\
&+ \sum_{j=1}^{d} \mathbb{E}^{\boldsymbol{\theta}^*} \left[\liminf_{N\to\infty} \left(\Lambda_{B^{(j)}}(\theta_j) \frac{k_j}{N} - \frac{\theta_j}{N} \sum_{k=1}^{k_j} B_k^{(j)} \right) \middle| E \right].
\end{aligned}$$

Observe that with probability 1 we have $\lim_{N\to\infty} \frac{1}{K_N} \sum_{k=1}^{K_N-1} A_k = \mathbb{E}^{\boldsymbol{\theta}^*}[A]$ and $\lim_{N\to\infty} \frac{1}{k_j} \sum_{k=1}^{k_j} B_k^{(j)} = \mathbb{E}^{\boldsymbol{\theta}^*}\left[B^{(j)}\right]$ for all $j = 1, \ldots, d$. Conditional on the event E, for which we have $\mathbb{P}^{\boldsymbol{\theta}^*}(E) > 0$, we can replace $\mathcal{K}$ by K_N and note that with probability 1 the liminf is a constant as $K_N - N \leq k_j \leq K_N$. Then applying Lemmas 4.7 and 4.8 we can remove the conditioning from all terms of $f(\boldsymbol{\theta})$ and change the liminf to a limit in the first term. Thus, we have

$$\begin{aligned}
f(\boldsymbol{\theta}) = &\left(\Lambda_A(-\theta_0) + \theta_0 \mathbb{E}^{\boldsymbol{\theta}^*}[A] \right) \mathbb{E}^{\boldsymbol{\theta}^*} \left[\lim_{N\to\infty} \frac{K_N}{N} \right] \\
&+ \sum_{j=1}^{d} \left(\mathbb{E}^{\boldsymbol{\theta}^*} \left[\liminf_{N\to\infty} \left(\Lambda_{B^{(j)}}(\theta_j) - \theta_j \mathbb{E}^{\boldsymbol{\theta}^*} \left[B^{(j)} \right] \right) \frac{k_j}{N} \right] \right).
\end{aligned}$$

We now first show that a unique minimum of the above, and hence the tightest lower bound of (4.16), is achieved at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ and conclude the proof by showing

that $f(\boldsymbol{\theta^*}) = \Lambda_A(-\theta^*)$. To find the minimum of $f(\boldsymbol{\theta})$ we note that we only have to consider $\boldsymbol{\theta}$ such that $\Lambda_{B^{(j)}}(\theta_j) - \theta_j \mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j)}\right] \leq 0$ for all $j = 1, \ldots, d$, so that we can take this constant out of the liminf which then becomes limsup. It is not hard to see that such $\boldsymbol{\theta}$ exists (for example, by convexity of $\Lambda_{B^{(j)}}(\theta_j)$ and by (4.7), we have for all $\boldsymbol{\theta}$ with $0 \leq \theta_j \leq \theta_j^*$ that $\Lambda_{B^{(j)}}(\theta_j) \leq \theta_j \mathbb{E}^{\boldsymbol{\theta}}\left[B^{(j)}\right] \leq \theta_j \mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j)}\right]$). For all such $\boldsymbol{\theta}$ we can write

$$\begin{aligned} f(\boldsymbol{\theta}) = &\left(\Lambda_A(-\theta_0) + \theta_0 \mathbb{E}^{\boldsymbol{\theta^*}}[A]\right) \mathbb{E}^{\boldsymbol{\theta^*}}\left[\lim_{N\to\infty} \frac{K_N}{N}\right] \\ &+ \sum_{j=1}^{d} \left(\Lambda_{B^{(j)}}(\theta_j) - \theta_j \mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j)}\right]\right) \mathbb{E}^{\boldsymbol{\theta^*}}\left[\limsup_{N\to\infty} \frac{k_j}{N}\right]. \end{aligned} \tag{4.17}$$

We take partial derivatives of $f(\boldsymbol{\theta})$:

$$\begin{aligned} \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_0} &= \left(-\mathbb{E}^{\boldsymbol{\theta}}[A] + \mathbb{E}^{\boldsymbol{\theta^*}}[A]\right) \mathbb{E}^{\boldsymbol{\theta^*}}\left[\lim_{N\to\infty} \frac{K_N}{N}\right], \\ \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_j} &= \left(\mathbb{E}^{\boldsymbol{\theta}}\left[B^{(j)}\right] - \mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j)}\right]\right) \mathbb{E}^{\boldsymbol{\theta^*}}\left[\limsup_{N\to\infty} \frac{k_j}{N}\right], \qquad j = 1, \ldots, d. \end{aligned}$$

These partial derivatives are zero if and only if $\theta_j = \theta_j^*$, $j = 0, \ldots, d$, since all limsups exist and are strictly positive constants. Since the log-moment generating functions $\Lambda_A(-\theta_0)$ and $\Lambda_{B^{(j)}}(\theta_j)$, $j = 1, \ldots, d$, are strictly convex functions (unless their distributions are deterministic), the right hand side of (4.17) is a strictly convex function (unless all distributions are deterministic, but this is ruled out by the non-triviality and stability assumption). Therefore, and because $\boldsymbol{\theta^*}$ is one of the values of $\boldsymbol{\theta}$ for which (4.17) holds, we are justified in concluding that $\boldsymbol{\theta^*}$ is indeed a global minimum. Hence $f(\boldsymbol{\theta})$ is minimal only for $\boldsymbol{\theta} = \boldsymbol{\theta^*}$.

To show that $f(\boldsymbol{\theta^*}) = \Lambda_A(-\theta^*)$, we take $\boldsymbol{\theta} = \boldsymbol{\theta^*}$ in (4.17) above. With $\theta_0^* = \theta_{j^*}^* = \theta^*$, and $\theta_j^* = 0$ for all other j, only two terms remain: one for the inter-arrival time A, involving $\mathbb{E}^{\boldsymbol{\theta^*}}\left[\lim_{N\to\infty} \frac{K_N}{N}\right]$, which is given in Lemma 4.7, and one for the service times of the bottleneck queue $B^{(j^*)}$, involving $\mathbb{E}^{\boldsymbol{\theta^*}}\left[\limsup_{N\to\infty} \frac{k_{j^*}}{N}\right]$, for which we can use $k_{j^*} = \mathcal{K} - N + \sum_{j=j^*+1}^{d} n_j + \mathbb{1}\{n_{j^*} > 0\}$. This leads to

$$\begin{aligned} f(\boldsymbol{\theta^*}) = &\left(\Lambda_A(-\theta^*) + \theta^* \mathbb{E}^{\boldsymbol{\theta^*}}[A]\right) \frac{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]}{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] - \mathbb{E}^{\boldsymbol{\theta^*}}[A]} \\ &+ \left(\Lambda_{B^{(j^*)}}(\theta^*) - \theta^* \mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]\right) \left(\frac{\mathbb{E}^{\boldsymbol{\theta^*}}[A]}{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] - \mathbb{E}^{\boldsymbol{\theta^*}}[A]}\right. \\ &\left. + \mathbb{E}^{\boldsymbol{\theta^*}}\left[\limsup_{N\to\infty} \frac{\sum_{j=j^*+1}^{d} n_j}{N}\right]\right). \end{aligned}$$

Since queues $j^*+1, \ldots, d$ are stable queues under the $\boldsymbol{\theta^*}$-tilt and we have, by assumption, that $\Lambda_A(-\theta^*) + \Lambda_{B^{(j^*)}}(\theta^*) = 0$, we find $f(\boldsymbol{\theta^*}) = \Lambda_A(-\theta^*)$. Thus, (4.15) holds when $\boldsymbol{\theta} \neq \boldsymbol{\theta^*}$. □

Remark 4.10. *Note that the tilt $\boldsymbol{\theta} = \boldsymbol{\theta^*}$ can still give an asymptotically efficient estimator, but this is not guaranteed.*

Remark 4.11. *In case of a single queue, Sadowsky showed that the $\boldsymbol{\theta^*}$-tilt* is *the unique change of measure that is asymptotically efficient, see [36, Theorem 3]; however he assumes bounded support for the service time distribution, which we do not need.*

4.4 Necessary conditions for asymptotic efficiency when $d = 2$

Having found that the $\boldsymbol{\theta^*}$-tilt is the only exponential state-independent change of measure that can possibly give an asymptotically efficient estimator, we show that additional conditions are needed for this change of measure to actually give an asymptotically efficient estimator. In this section we assume that we have two queues in tandem ($d = 2$). First we derive the conditions in Section 4.4.1, then we zoom in to the Markovian case and compare with earlier work in Section 4.4.2.

4.4.1 Derivation of necessary conditions

To work with (4.14) we first rewrite the likelihood $L^{\boldsymbol{\theta}}$ as given in (4.13), using (4.3)–(4.4). Taking $d = 2$ and $\boldsymbol{\theta} = \boldsymbol{\theta^*}$ (with $\theta_0^* = \theta_{j^*}^* = \theta^*$, for which we have $M_A(-\theta^*)M_{B^{(j^*)}}(\theta^*) = 1$, and $\theta_{3-j^*}^* = 0$), we find

$$L^{\boldsymbol{\theta^*}} = \frac{M_A(-\theta^*)^{\mathcal{K}-1-k_{j^*}}}{e^{-\theta^*\left(\sum_{k=1}^{\mathcal{K}-1} A_k - \sum_{k=1}^{k_{j^*}} B_k^{(j^*)}\right)}}. \tag{4.18}$$

To rewrite the denominator of (4.18) we note the following relation for I_j, the idle time of queue j during the busy cycle, when $\mathcal{K} = K_N$,

$$I_j = \sum_{k=1}^{\mathcal{K}-1} A_k - \sum_{k=1}^{k_j} B_k^{(j)} + \bar{B}^{(j)},$$

where $\bar{B}^{(j)}$ is the residual service time of the customer in service (if any) in queue j just before the overflow level N is reached; in the event that queue j is empty just before N is reached (which is unlikely when $j = j^*$), we set $\bar{B}^{(j)} = 0$. Combining with (4.14) and (4.18) we have asymptotic efficiency when

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta^*}} \left[\frac{M_A(-\theta^*)^{2(\mathcal{K}-1-k_{j^*})}}{e^{-2\theta^*\left(I_{j^*} - \bar{B}^{(j^*)}\right)}} \mathbb{1}\{\mathcal{K} = K_N\} \right] \leq 2\Lambda_A(-\theta^*).$$

For the numerator we distinguish between two cases, depending on which queue is the bottleneck. When this is queue 1 ($j^* = 1$), we have $\mathcal{K}-1-k_{j^*} = n_1 - \mathbb{1}\{n_1 >$

$0\}$, so that we have asymptotic efficiency if and only if

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta^*}}\Big[M_A(-\theta^*)^{2(n_1-\mathbb{1}\{n_1>0\})} e^{2\theta^*\left(I_1-\bar{B}^{(1)}\right)} \mathbb{1}\{\mathcal{K}=K_N\}\Big] \leq 2\Lambda_A(-\theta^*). \tag{4.19}$$

When queue 2 is the bottleneck queue ($j^*=2$), we have $\mathcal{K}-1-k_{j^*} = n_1+n_2-\mathbb{1}\{n_2>0\} = N-1-\mathbb{1}\{n_2>0\}$, so that the condition is

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\boldsymbol{\theta^*}}\left[M_A(-\theta^*)^{-2\mathbb{1}\{n_2>0\}} e^{2\theta^*(I_2-\bar{B}^{(2)})} \mathbb{1}\{\mathcal{K}=K_N\}\right] \leq 0, \tag{4.20}$$

where we used that $\limsup_{N\to\infty} \frac{1}{N}\log M_A(-\theta^*)^{2(N-1)}$ equals the right-hand side in (4.14), $2\Lambda_A(-\theta^*)$.

In the sequel we will give necessary conditions for these inequalities to hold by considering specific sample paths which are very unlike the 'typical' paths that lead to overflow. The advantage of this approach, which is also used in Glasserman and Kou [30] for the Markovian case, is that the chosen unlikely paths are easy to analyze, and the process spends much time on the boundaries of the state space which we know is problematic for asymptotic efficiency, at least in the Markovian case.

The specific paths we will consider are illustrated in Figure 4.2, and will be used in the proofs of the following theorems. After stating the theorems we will consider the Markovian case, also comparing with De Boer [12] and Glasserman and Kou [30].

We start with the necessary condition for asymptotic efficiency when queue 2 is the bottleneck queue, since this is the easiest case, and show what it looks like for some special cases, including the $M|M|1$ tandem queue case.

Theorem 4.12. *Consider 2 $GI|GI|1$ queues in tandem and suppose queue 2 is the bottleneck queue ($j^*=2$). Under Assumption 4.1 a necessary condition for asymptotic efficiency of the $\boldsymbol{\theta^*}$-tilt is*

$$\limsup_{N\to\infty} \frac{1}{N} \log\left(\int_0^\infty e^{2\theta^* x}[1-F_{B^{(1)}}(x)]\; dF_{A,N-1}^{\boldsymbol{\theta^*}}(x)\right) \leq 0, \tag{4.21}$$

where $F_{A,N-1}^{\boldsymbol{\theta^*}}(x)$ is the $(N-1)$-fold convolution of $F_A^{\boldsymbol{\theta^*}}(x)$, the probability distribution function of A under tilt $\boldsymbol{\theta^*}$.
More specifically, when the service times of the first queue are exponentially distributed with rate μ_1, this condition becomes

$$M_A(\theta^*-\mu_1) \leq M_A(-\theta^*), \tag{4.22}$$

and for an $M|M|1$ tandem queue with arrival rate λ and service rates μ_1 and μ_2 (with $\mu_1>\mu_2$), the condition becomes

$$2\mu_2 \leq 2\lambda+\mu_1. \tag{4.23}$$

Proof. Consider the specific sample path with no service completions before level N is reached, that is, the path that moves $N-1$ steps to the right from $(1,0)$ to $(N,0)$, see Figure 4.2, left panel. Based on this path we find a lower bound on the left-hand side of (4.20). Since for this path $\mathbb{1}\{n_2 > 0\} = 0$, $\bar{B}^{(2)} = 0$ and $I_2 = \sum_{k=1}^{N-1} A_k$, it follows that

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\theta}^*}\left[M_A(-\theta^*)^{-2\mathbb{1}\{n_2>0\}} e^{2\theta^*(I_2-\bar{B}^{(2)})}\mathbb{1}\{\mathcal{K}=K_N\}\right] \\
&\geq \mathbb{E}^{\boldsymbol{\theta}^*}\left[e^{2\theta^*\sum_{k=1}^{N-1}A_k}\mathbb{1}\{\textstyle\sum_{k=1}^{N-1}A_k < B_1^{(1)}\}\right] \\
&= \int_0^\infty e^{2\theta^* x}[1-F_{B^{(1)}}(x)]\; dF^{\boldsymbol{\theta}^*}_{A,N-1}(x),
\end{aligned}$$

where the inequality follows because we only consider one possible path to reach the overflow level. Thus the general necessary condition for asymptotic efficiency in (4.21) follows.
When $B^{(1)} \sim \exp(\mu_1)$, the argument of the logarithm of the left hand side in (4.21) reduces to

$$\begin{aligned}
\int_0^\infty e^{2\theta^* x}[1-F_{B_1^{(1)}}(x)]dF^{\boldsymbol{\theta}^*}_{A,N-1}(x) &= \int_0^\infty e^{(2\theta^*-\mu_1)x}dF^{\boldsymbol{\theta}^*}_{A,N-1}(x) \\
&= \left[M_A^{\boldsymbol{\theta}^*}(2\theta^*-\mu_1)\right]^{N-1},
\end{aligned}$$

so the necessary condition for asymptotic efficiency becomes

$$M_A^{\boldsymbol{\theta}^*}(2\theta^*-\mu_1) \leq 1,$$

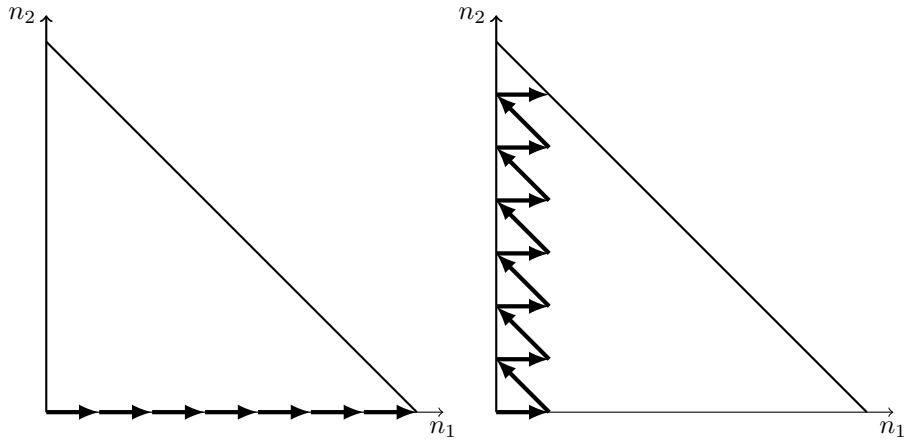


Figure 4.2 Sample paths considered in the proofs of Theorems 4.12 (left, $j^* = 2$) and 4.13 (right, $j^* = 1$).

which, using (4.5), leads to (4.22). Finally, (4.23) follows immediately from (4.22) by noting that $\theta^* = \mu_2 - \lambda$ for an $M|M|1$ tandem queue with $j^* = 2$. □

Next, we provide a necessary condition for an asymptotically efficient estimator when queue 1 is the bottleneck queue. Here we will assume that the service times of the second queue are exponentially distributed (with rate μ_2) in order to give a useful expression for the necessary conditions. We also consider some special cases, including the $M|M|1$ tandem queue.

Theorem 4.13. *Consider 2 $GI|GI|1$ queues in tandem and suppose queue 1 is the bottleneck queue ($j^* = 1$) and that the service times of queue 2 are exponentially distributed with rate μ_2. Under Assumption 4.1 a necessary condition for asymptotic efficiency of the $\boldsymbol{\theta^*}$-tilt is*

$$\int_0^\infty e^{(2\theta^* - \mu_2)x} \int_0^x e^{-2\theta^* y} \, dF_{B^{(1)}}^{\boldsymbol{\theta^*}}(y) dF_A^{\boldsymbol{\theta^*}}(x) \leq M_A(-\theta^*)^2, \tag{4.24}$$

where, as before, $F_A^{\boldsymbol{\theta^*}}(x)$ and $F_{B^{(1)}}^{\boldsymbol{\theta^*}}(y)$ denote the probability distribution functions of A and $B^{(1)}$ under the tilt $\boldsymbol{\theta^*}$.
More specifically, when both queues have exponential services with rates μ_1 and μ_2 respectively, this condition becomes

$$\frac{\mu_1 - \theta^*}{\theta^* + \mu_1} \left[M_A(\theta^* - \mu_2) - M_A(-(\mu_1 + \mu_2)) \right] \leq M_A(-\theta^*)^3, \tag{4.25}$$

and for an $M|M|1$ tandem queue with arrival rate λ and service rates μ_1 and μ_2 (with $\mu_1 < \mu_2$), the condition becomes

$$\frac{\mu_1}{2\mu_1 - \lambda} \left[\frac{1}{2\lambda + \mu_2 - \mu_1} - \frac{1}{\lambda + \mu_1 + \mu_2} \right] \leq \frac{\lambda}{\mu_1^2}. \tag{4.26}$$

Proof. We determine a lower bound on the expected value in (4.19) by considering the sample path that alternates between 0 and 1 customers in queue 1, with no departures from queue 2, until level N is reached; that is, the path moves from $(1,0)$ to $(0,1), (1,1), (0,2), (1,2), (0,3), ...,$ to $(1, N-1)$, see Figure 4.2, right panel. It is not hard to see that on this path we have $B_k^{(1)} < A_k, k = 1, ..., N-1$, and also $\sum_{k=1}^{N-1} A_k < B_1^{(1)} + B_1^{(2)}$. Obviously every $B_k^{(1)}$ should be smaller than $B_1^{(1)} + B_1^{(2)}$ as well, but this condition is implied by the above.

Also on this path we have $n_1 = 0$, $\bar{B}^{(1)} = 0$, $\mathcal{K} = K_N = N$ and $k_1 = N - 1$, so

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\theta^*}}\left[M_A(-\theta^*)^{2(n_1-\mathbb{1}\{n_1>0\})}e^{2\theta^*(I_1-\bar{B}^{(1)})}\mathbb{1}\{\mathcal{K}=K_N\}\right]\\
&\geq \mathbb{E}^{\boldsymbol{\theta^*}}\left[e^{2\theta^*\left(\sum_{k=1}^{N-1}A_k-\sum_{k=1}^{N-1}B_k^{(1)}\right)}\mathbb{1}\{B_k^{(1)}<A_k,\,\forall k=1,...,N-1\}\cdot\right.\\
&\quad\left.\mathbb{1}\{\textstyle\sum_{k=1}^{N-1}A_k<B_1^{(1)}+B_1^{(2)}\}\right]\\
&\geq \mathbb{E}^{\boldsymbol{\theta^*}}\left[e^{2\theta^*\left(\sum_{k=1}^{N-1}A_k-\sum_{k=1}^{N-1}B_k^{(1)}\right)}\left(\prod_{k=1}^{N-1}\mathbb{1}\{B_k^{(1)}<A_k\}\right)\mathbb{1}\{\textstyle\sum_{k=1}^{N-1}A_k<B_1^{(2)}\}\right],
\end{aligned} \tag{4.27}$$

where the first inequality follows because there are more paths that reach the overflow level than just the one we consider here. Next, since $B^{(2)}$ has the memoryless property, we can write

$$\mathbb{1}\{\textstyle\sum_{k=1}^{N-1}A_k<B_1^{(2)}\}\overset{d}{=}\displaystyle\prod_{k=1}^{N-1}\mathbb{1}\{A_k<B_{1,k}^{(2)}\},$$

where the $B_{1,k}^{(2)}$ are i.i.d. copies of $B_1^{(2)}$, independent of all else and $\overset{d}{=}$ denotes an equality in distribution. Note that, by assuming $j^* = 1$, the service times of queue 2 remain exponentially distributed with rate μ_2 under the $\boldsymbol{\theta^*}$-tilt, and hence still have the memoryless property. As a consequence, the right-hand side of (4.27) can be written as

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\theta^*}}\left[\prod_{k=1}^{N-1}e^{2\theta^*\left(A_k-B_k^{(1)}\right)}\mathbb{1}\{B_k^{(1)}<A_k\}\mathbb{1}\{A_k<B_{1,k}^{(2)}\}\right]\\
&=\left(\int_0^\infty\int_0^x e^{2\theta^*(x-y)}\left[1-F_{B^{(2)}}(x)\right]\,dF_{B^{(1)}}^{\boldsymbol{\theta^*}}(y)dF_A^{\boldsymbol{\theta^*}}(x)\right)^{N-1},
\end{aligned}$$

where in the last step the independence of A_k and $B_k^{(1)}$ is used. The general necessary condition for asymptotic efficiency in (4.24) now follows from applying (4.19) and $B^{(2)} \sim \exp(\mu_2)$.
When $B^{(1)} \sim \exp(\mu_1)$ (and $B^{(2)} \sim \exp(\mu_2)$ as before), the left hand side of (4.24) reduces to

$$\begin{aligned}
&(\mu_1-\theta^*)\int_0^\infty e^{(2\theta^*-\mu_2)x}\int_0^x e^{-(\theta^*+\mu_1)y}\,dydF_A^{\boldsymbol{\theta^*}}(x)\\
&=\frac{\mu_1-\theta^*}{\theta^*+\mu_1}\left[M_A^{\boldsymbol{\theta^*}}(2\theta^*-\mu_2)-M_A^{\boldsymbol{\theta^*}}(-(\mu_1+\mu_2-\theta^*))\right].
\end{aligned}$$

from which (4.25) follows by using (4.5). Finally (4.26) follows from (4.25) by noting that $\theta^* = \mu_1 - \lambda$ for an $M|M|1$ tandem queue with $j^* = 1$. □

4.4.2 Comparison of necessary conditions for the $M|M|1$ tandem queue

In this section we will make a comparison with earlier papers for the Markovian case. Since these papers always consider simulation in discrete time, we will first explain how this relates to our current work.

4.4.2.1 Continuous-time vs. discrete-time models

In this chapter, we represent the GI/GI/1 queueing systems in continuous time, and we *simulate in continuous time*, by which we mean that we tilt the (typically continuous) distributions of the A_k and $B_k^{(j)}$. Alternatively, if all distributions are exponential, the system state can also be represented by a discrete-time Markov chain, embedded at transition epochs, which can also be simulated. We will refer to this as *simulation in discrete time*.

Parekh and Walrand [35] consider both simulation in continuous and in discrete time. For the single Markovian queue, they show that their heuristic for both continuous and discrete-time leads to the same change of measure, namely an interchange of the arrival and service rates (or probabilities). In the same way, any exponential change of measure in the discrete-time Markov chain (changing transition probabilities) can easily be shown to be equivalent to an exponential change of measure in the corresponding continuous-time Markov chain (changing transition rates).

We will now compare all known conditions for asymptotic efficiency from De Boer [12] and Glasserman and Kou [30], who apply simulation in discrete time, and this chapter. For ease of comparison, we will normalize the (continuous time) rates such that $\lambda + \mu_1 + \mu_2 = 1$, so that on the interior of the state space they coincide with the (discrete time) transition probabilities .

4.4.2.2 Queue 2 is bottleneck

First we consider the case in which queue 2 is the bottleneck, which now means that $\mu_1 > \mu_2$. With the normalization, our necessary condition in Theorem 4.12 becomes $\mu_1 + 4\mu_2 \leq 2$, and since $\mu_1 > \mu_2$ it follows in particular that a necessary condition for asymptotic efficiency is $\mu_2 < \frac{2}{5}$. This is stricter than $\mu_2 \leq \sqrt{2} - 1$, which was obtained in Glasserman and Kou [30], simulating in discrete time. Thus, if $\mu_2 \in [\frac{2}{5}, \sqrt{2} - 1]$ our estimator *cannot* be asymptotically efficient, while the estimator in [30] *could* be asymptotically efficient. Although this situation seems very unlikely, we cannot rule out the possibility since the two estimators are different.

4.4.2.3 Queue 1 is bottleneck

Next we look at the case in which queue 1 is the bottleneck, which for the $M|M|1$ tandem queue means that $\mu_1 < \mu_2$. For this case, importance sampling has never been studied analytically before, because in the Markovian network

both servers are interchangeable without changing the probability of overflow, see [41]. Nevertheless in [12] it is shown by numerical computations that in terms of asymptotic efficiency both servers are not interchangeable. Before we continue to summarize all known necessary conditions for the $M|M|1$ tandem queue in a figure, we present the 'missing' result in Glasserman and Kou [30], namely a necessary condition for asymptotic efficiency of the estimator for simulation in discrete time, when queue 1 is the bottleneck queue. The proof is completely analogous to their approach for the other case, except that we consider the path in the right panel of Figure 4.2 (in discrete time), rather than the left panel.

Proposition 4.14. *For an $M|M|1$ tandem queue, simulated in discrete time, with arrival rate λ and service rates μ_1 and μ_2 such that $\lambda < \mu_1 < \mu_2$ (queue 1 is the bottleneck), a necessary condition for asymptotic efficiency of the corresponding estimator is*

$$\mu_1^3(\mu_1 + \mu_2) \leq \lambda(\lambda + \mu_2)^2(\lambda + \mu_1 + \mu_2).$$

Proof. In this case the change of measure (here denoted as $\mathbb{Q}$) prescribes to interchange λ and μ_1. The definition for asymptotic efficiency is (compare with the continuous-time analog in Definition 4.6),

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\mathbb{Q}} \left[L^2 \mathbb{1}\{\mathcal{K} = K_N\} \right] \leq \log \frac{\lambda^2}{\mu_1^2}. \tag{4.28}$$

Here L is the likelihood ratio of the path as simulated in discrete time, that is, $L = \Pi_i \, \mathbb{P}(t_i)/\mathbb{Q}(t_i)$ where the product is taken over all transitions t_i on the path, and $\mathbb{P}(t_i)$ and $\mathbb{Q}(t_i)$ are the probabilities of t_i under the original and changed measure respectively. In order to get a lower bound on $\mathbb{E}^{\mathbb{Q}} \left[L^2 \mathbb{1}\{\mathcal{K} = K_N\} \right]$, we consider the path in the right panel of Figure 4.2 (in discrete time) and note the following.
For each transition t_i which is an arrival to queue 1, the contribution $\mathbb{P}(t_i)/\mathbb{Q}(t_i)$ to the likelihood ratio is $\frac{\lambda}{\lambda+\mu_2} / \frac{\mu_1}{\mu_1+\mu_2} = \frac{\lambda}{\mu_1} \frac{\mu_1+\mu_2}{\lambda+\mu_2}$. In order to reach the overflow level, there are $N-1$ arrivals to queue 1 (as we start with one customer in queue 1).
For each departure from queue 1, except for the first one, the contribution to the likelihood ratio is $\frac{\mu_1}{\lambda+\mu_1+\mu_2} / \frac{\lambda}{\lambda+\mu_1+\mu_2} = \frac{\mu_1}{\lambda}$. The contribution to the likelihood ratio of the first departure from queue 1 is $\frac{\mu_1}{\mu_1+\lambda} / \frac{\lambda}{\lambda+\mu_1} = \frac{\mu_1}{\lambda}$. In total there are $N-1$ departures from queue 1. Therefore, the total likelihood ratio for this path is

$$\left(\frac{\lambda}{\mu_1} \frac{\mu_1 + \mu_2}{\lambda + \mu_2} \right)^{N-1} \left(\frac{\mu_1}{\lambda} \right)^{N-1} = \left(\frac{\mu_1 + \mu_2}{\lambda + \mu_2} \right)^{N-1}.$$

Similarly the probability of this path, under the change of measure $\mathbb{Q}$, is

$\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{N-1}\left(\frac{\lambda}{\lambda+\mu_1+\mu_2}\right)^{N-2}\frac{\lambda}{\mu_1+\lambda}$. Hence we have that

$$\mathbb{E}^{\mathbb{Q}}\left[L^2\mathbb{1}\{\mathcal{K}=K_N\}\right] \geq \left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{N-1}\left(\frac{\lambda}{\lambda+\mu_1+\mu_2}\right)^{N-2}\cdot \frac{\lambda}{\mu_1+\lambda}\left(\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right)^{2(N-1)},$$

so that the left-hand side of (4.28) is at least

$$\log\left(\frac{\mu_1}{\lambda+\mu_2}\frac{\lambda}{\lambda+\mu_1+\mu_2}\frac{\mu_1+\mu_2}{\lambda+\mu_2}\right).$$

Solving (4.28) concludes the proof. □

4.4.2.4 Comparison

We are now ready to summarize all necessary conditions from [12], [30] and this section for the $M|M|1$ tandem queue (with the convention that $\lambda+\mu_1+\mu_2=1$) in Figure 4.3. For each estimator this figure shows for which parameter settings the estimator is *certainly not* asymptotically efficient, and for which settings it *could* be:

- Between the dash-dotted (blue) lines the change of measure as discussed in this chapter (simulated in continuous time) does not give an asymptotically efficient estimator according to Theorems 4.12 and 4.13.
- Between the dashed (red) lines the change of measure as discussed in [30] (simulated in discrete time) does not give an asymptotically efficient estimator according to [30] and Proposition 4.14.
- Between the dotted (yellow) and solid (black) line the change of measure as discussed in [30] (simulated in discrete time) does not give an asymptotically efficient estimator according to [12].

When we compare the areas where asymptotic efficiency is certainly not attained, as derived from considering some unlikely path, that is, the area between the blue dash-dotted lines (simulated in continuous time), and the area between the red dashed lines (simulated in discrete time), we see that the first is largest. The method used by De Boer [12] gives an even bigger area for the discrete-time estimator, but this approach is different and only for the case where queue 2 is the bottleneck. Unfortunately this method cannot be used for simulation in continuous time.

4.5 Numerical results

In this section, we give an example of the conditions that have been shown in the previous section. In order to easily show both bottleneck cases in one figure,

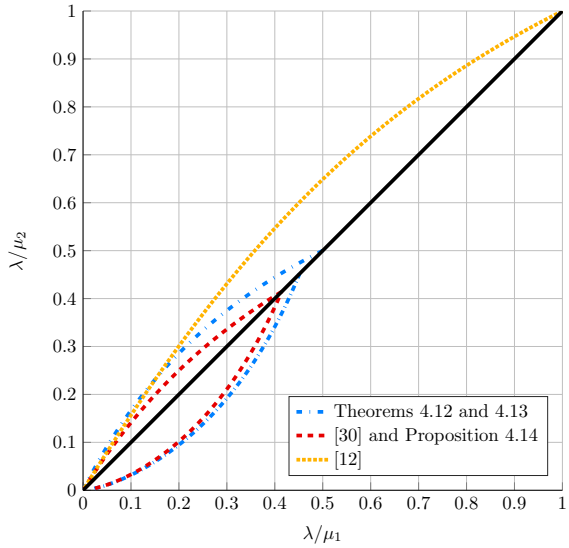


Figure 4.3 Summary of results for the $M|M|1$ tandem queue with arrival rate λ and service rates μ_1 and μ_2 for queues 1 and 2 respectively.

we consider a tandem queue with exponentially distributed service times. Also we compare our results for the $M|M|1$ tandem queue with the results obtained by De Boer [12].

In Figure 4.4 we give an example to show that the necessary conditions for asymptotic efficiency are not always satisfied.

In Tables 4.1 and 4.2 we show some simulation results for parameters as in Figures 4.3 and 4.4. In these tables RE denotes the relative error, that is, the standard deviation of the estimator divided by the mean of the estimator, which gives

$$\mathrm{RE} = \frac{\sqrt{\frac{1}{S-1}\sum_{i=1}^{S}(L(i)I(i)-\hat{p}_N)^2}}{\sqrt{S}\hat{p}_N}, \tag{4.29}$$

and AE is given by

$$\mathrm{AE} = \frac{\log \frac{1}{S}\sum_{i=1}^{S} L(i)^2 I(i)}{\log \frac{1}{S}\sum_{i=1}^{S} L(i)I(i)}, \tag{4.30}$$

where S is the total number of simulations, $L(i)$ is the likelihood ratio in simulation i, and $I(i)$ indicates whether level N has been reached in simulation i or not. This value should be 2 in case of asymptotic efficiency as N goes to infinity, see also Definition 4.6 and the text below it.

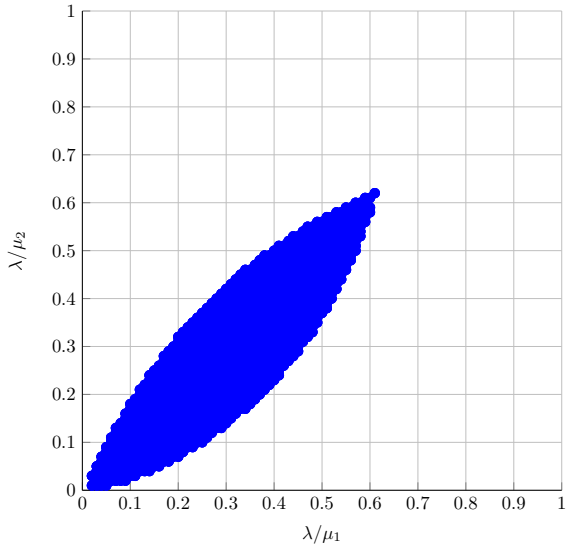


Figure 4.4 A tandem queue with $A \sim U[0,2]$. Here $\lambda = \frac{1}{\mathbb{E}[A]}$. The colored area shows for which parameter values the necessary conditions for asymptotic efficiency are *not* satisfied.

In Table 4.1 we present the results in case of a two node $M|M|1$ tandem queue, where the parameters are chosen such that the second queue is the bottleneck, and the necessary conditions for asymptotic efficiency given by De Boer [12] and Glasserman and Kou [30] *are* satisfied, while the conditions in Theorem 4.12 are *not* satisfied. In Table 4.2 we give results for a tandem queue with uniform arrivals and exponential service times at both queues, where the first queue is the bottleneck. Again, the parameters are such that the necessary conditions in Theorem 4.13 are not satisfied.

These tables suggest that the estimators are asymptotically efficient, as AE tends to 2 when N goes to infinity, although in fact they are not since they do not satisfy the conditions in Theorems 4.12 and 4.13. We can explain this in the following way. Firstly, in the proofs of Theorems 4.12 and 4.13 we considered very unlikely paths. So it is likely that these paths did not occur during these simulations and therefore it still seems that the estimator is asymptotically efficient.

Secondly, from Figures 4.5 and 4.6 we can indeed see that there are (very) unlikely paths with a large contribution to the likelihood ratio. These figures show AE for three fixed values of N against the number of simulation runs S. We see that, even though for increasing N the value of AE seems to increase to 2 (as in Tables 4.1 and 4.2), the value of AE is clearly decreasing as the number of simulations increases. Therefore the estimator cannot be asymptotically efficient. Moreover, the big jumps are caused by (rare) paths that have a large contribution to the likelihood ratio and they suggest that there exist paths that are even more

Table 4.1 Simulation results for a two node $M|M|1$ tandem queue with $\lambda = 0.04$, $\mu_1 = 0.6$ and $\mu_2 = 0.36$. The number of simulations is 10^6.

N	p_n	RE	AE
100	7.6124e-095	0.0128	1.9764
120	6.1567e-114	0.0052	1.9872
140	5.0555e-133	0.0051	1.9892
160	4.1991e-152	0.0085	1.9877
180	3.4939e-171	0.0078	1.9895
200	2.8198e-190	0.0082	1.9903
220	2.3300e-209	0.0049	1.9933
240	1.9092e-228	0.0078	1.9921
260	1.5865e-247	0.0068	1.9932
280	1.2925e-266	0.0049	1.9947
300	1.0714e-285	0.0066	1.9942

Table 4.2 Simulation results when $A \sim U[0, 2]$, $B^{(1)} \sim \exp(3)$ and $B^{(2)} \sim \exp(5.5)$. The number of simulations is 10^6.

N	p_n	RE	AE
100	7.5165e-068	0.0362	1.9536
120	1.8615e-081	0.0146	1.9711
140	4.5880e-095	0.0120	1.9771
160	1.1467e-108	0.0139	1.9788
180	2.9209e-122	0.0162	1.9801
200	7.7573e-136	0.0649	1.9732
220	1.9070e-149	0.0544	1.9767
240	4.5067e-163	0.0280	1.9822
260	1.0883e-176	0.0133	1.9872
280	2.7707e-190	0.0130	1.9882
300	6.7259e-204	0.0108	1.9898

unlikely to occur. Those paths probably have an even larger contribution to the likelihood ratio such that the estimator is not asymptotically efficient.

Next, we compare our results for the $M|M|1$ tandem queue with the results obtained numerically by De Boer [12]. Both our and [12]'s results concern P&W changes of measure, but ours in continuous time and [12]'s in discrete time; see Section 4.4.2.1. In order to compare all these results, we use the convention that $\lambda + \mu_1 + \mu_2 = 1$ and we transform Figure 3 from [12], see Figure 4.7, such that λ/μ_1 and λ/μ_2 are along the x-axis and y-axis respectively (which has been used more often throughout this chapter).
What we see in Figure 4.7 is that when queue 1 is the bottleneck queue our necessary condition is within the blue area. When queue 2 is the bottleneck

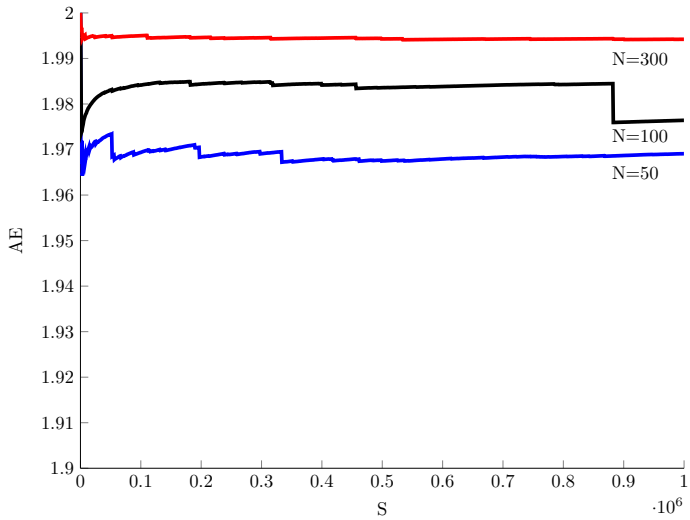


Figure 4.5 A possible explanation why it seems that the estimator in Table 4.1 is asymptotically efficient, while we proved that it is not.

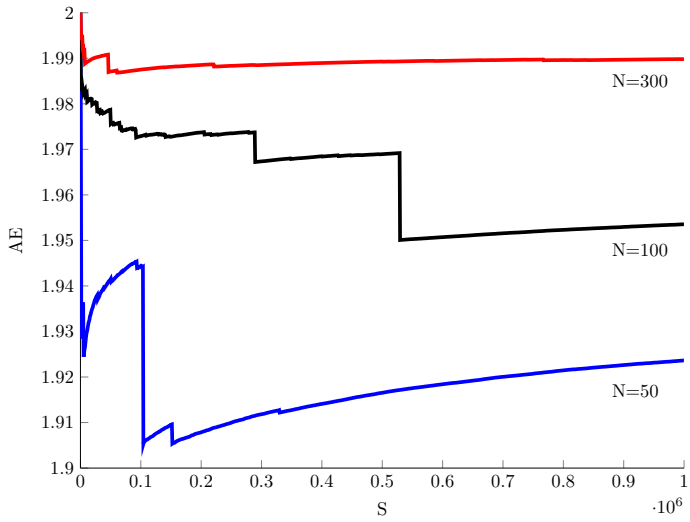


Figure 4.6 Similar possible explanation as in Figure 4.5, but corresponding to the situation as in Table 4.2.

queue it seems that our necessary condition does not coincide with the numerical results of [12]. This means that there are certain parameter choices where our estimator is not asymptotically efficient, while the numerical results of [12] tell that the estimator considered there has bounded relative error. This can be explained by the fact that in [12] the queueing system is simulated in discrete

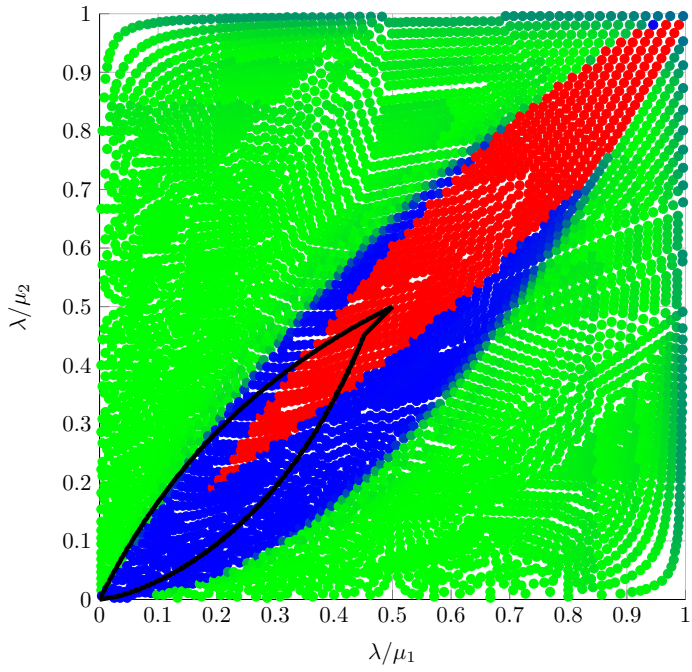


Figure 4.7 Figure 3 of [12], displayed with different x-axis and y-axis, together with our necessary conditions from Theorems 4.12 and 4.13. The green part has bounded relative error, the blue part does not give an asymptotically efficient estimator but has finite variance and the red part has infinite variance. Between the black lines our necessary conditions are not satisfied.

time, while we simulate it in continuous time.

4.6 Conclusions

Parekh and Walrand [35] introduced a method to estimate the probability that the total number of customers in a queueing network reaches some level N in a busy cycle using simulation, but unfortunately for a network of $GI|GI|1$ queues it is not clear how to do so. Frater and Anderson [25] found this change of measure for the $GI|GI|1$ tandem queue, but only specified it implicitly, in the form of a minimization over all possible "guesses" for which queue would become the rightmost unstable queue. Fortunately there is another way to *explicitly* find the change of measure for the $GI|GI|1$ tandem queue, based on the decay rate determined in Chapter 2. In this chapter, we have shown that these two methods result in the same change of measure for the $GI|GI|1$ tandem queue for all cases where they are properly defined.
Also, we proved that this change of measure is the only exponential change of

measure that can possibly result in an asymptotically efficient estimator. In other words, for a state-independent change of measure one should only consider using the P&W change of measure. However, using this change of measure does not guarantee asymptotic efficiency. We have identified some additional necessary conditions for this change of measure to be asymptotically efficient in case of a two node tandem queue.
For future research it seems useful to look for sufficient conditions for asymptotic efficiency, examine the tightness of the necessary conditions or maybe to improve the likelihood ratio with respect to Remark 4.5. However, it may be better to focus on state-dependent change of measures as these could be asymptotically efficient in the whole parameter space, which is the topic of the next chapter.

4.7 Appendix A

Here we present the proofs of Lemma 4.7 and Lemma 4.8, which we copy for convenience.

Lemma 4.7. *Suppose we have d $GI|GI|1$ queues in tandem. Under the change of measure $\boldsymbol{\theta^*}$ for which $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] > \mathbb{E}^{\boldsymbol{\theta^*}}[A]$, we have for all N that $\mathbb{P}^{\boldsymbol{\theta^*}}(\mathcal{K} = K_N) \geq \mathbb{P}^{\boldsymbol{\theta^*}}(E) > 0$, where E is the event that the system never empties. Moreover, we have as $N \to \infty$ that $\frac{K_N}{N} \to \frac{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]}{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] - \mathbb{E}^{\boldsymbol{\theta^*}}[A]}$ with probability* 1.

Proof of Lemma 4.7. Let $N_A(t)$ denote the number of arrivals to the system up to time t, $N_D(t)$ be the number of departures from the system up to time t and $N_{B^{(j^*)}}(t)$ be the number of departures from queue j^* at time t *if* its server would work continuously. Then by renewal theory, under the change of measure $\boldsymbol{\theta^*}$, with probability 1,

$$\lim_{t\to\infty} \frac{N_A(t) - N_D(t)}{t} \geq \lim_{t\to\infty} \frac{N_A(t) - N_{B^{(j^*)}}(t)}{t} = \frac{1}{\mathbb{E}^{\boldsymbol{\theta^*}}[A]} - \frac{1}{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]} > 0,$$

because $N_D(t) \leq N_{B^{(j^*)}}(t)$, since the number of departures of the system as a whole can never exceed the number of departures at the rightmost unstable queue *if* it would work continuously. Now if we would assume that $\mathbb{P}^{\boldsymbol{\theta^*}}(E) = 0$, that is, the system always empties, this would lead to a contradiction, since in the long run the expected number of arrivals to the system would then equal the number of departures from the system. Hence, $\mathbb{P}^{\boldsymbol{\theta^*}}(E) > 0$. Clearly we also have $\mathbb{P}^{\boldsymbol{\theta^*}}(\mathcal{K} = K_N) \geq \mathbb{P}^{\boldsymbol{\theta^*}}(E)$, because the event E implies that $\mathcal{K} = K_N$, so the first statement follows.

For the second statement we let X_t be the total number of customers in the system at time t. In Lemma 4.3 it has been shown that queue j^* is the rightmost unstable queue under the $\boldsymbol{\theta^*}$-tilt. Accordingly we write $X_t = X_t^{(1,\ldots,j^*)} + X_t^{(j^*+1,..,d)}$, where $X_t^{(j^*+1,\ldots,d)}$ is the total number of customers in

the (stable) queues $j^*+1,...,d$ at time t, and $X_t^{(1,...,j^*)}$ is the total number of customers in the (not necessarily stable) queues $1,...,j^*$ at time t. Hence, with probability 1 under the $\boldsymbol{\theta^*}$-tilt,

$$\begin{aligned}\lim_{t\to\infty}\frac{X_t}{t} &= \lim_{t\to\infty}\frac{X_t^{(1,...,j^*)}}{t}+0\\ &= \lim_{t\to\infty}\frac{N_A(t)-N_{D^{(j^*)}}(t)}{t},\end{aligned}$$

where $N_{D^{(j^*)}}(t)$ is the number of departures from queue j^* up to time t. Since with probability 1, from a certain time onwards queue j^* does not empty, since it is an unstable queue, it follows that

$$\lim_{t\to\infty}\frac{N_A(t)-N_{D^{(j^*)}}(t)}{t}=\frac{1}{\mathbb{E}^{\boldsymbol{\theta^*}}[A]}-\frac{1}{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]}.$$

Note that X_t is the number of customers at (continuous) time t, but we need a discrete time result, so let $\tilde{X}_k$ be the total number of customers in queue $1,...,d$ right after the k^{th} arrival and let t_k be the time of the arrival of customer k, then with probability 1

$$\lim_{k\to\infty}\frac{\tilde{X}_k}{k}=\lim_{k\to\infty}\frac{t_k}{k}\frac{X_{t_k}}{t_k}=\mathbb{E}^{\boldsymbol{\theta^*}}[A]\left(\frac{1}{\mathbb{E}^{\boldsymbol{\theta^*}}[A]}-\frac{1}{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]}\right).$$

But then also, when $N\to\infty$,

$$\frac{N}{K_N}=\mathbb{E}^{\boldsymbol{\theta^*}}[A]\left(\frac{1}{\mathbb{E}^{\boldsymbol{\theta^*}}[A]}-\frac{1}{\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right]}\right),\qquad \text{w.p. 1,}$$

as by definition $K_N=\min\{k:X_k=N\}$. This concludes the proof. □

Lemma 4.8. *Consider a sequence $\{X_N\}$ of random variables that converges to a constant c with probability 1 as $N\to\infty$ and let E be an event with $\mathbb{P}(E)>0$. Then $\mathbb{E}[\lim_{N\to\infty}X_N\mid E]=\mathbb{E}[\lim_{N\to\infty}X_N]=c$.*

Proof of Lemma 4.8. The lemma follows from elementary principles. Clearly, when $\mathbb{P}(\lim_{N\to\infty}X_N=c)=1$, also $\mathbb{E}[\lim_{N\to\infty}X_N]=c$. To show that conditioning on some event E with $\mathbb{P}(E)>0$ does not change this assertion, let event $F=\{\lim_{N\to\infty}X_N=c\}$ and note that $\mathbb{P}(F)=1$ implies $\mathbb{P}(F\mid E)=\frac{\mathbb{P}(F\cap E)}{\mathbb{P}(E)}=1$, and hence also $\mathbb{E}[\lim_{N\to\infty}X_N\mid E]=c$. □

4.8 Appendix B

In this section, we briefly show what may happen when

$$\Lambda_A(-\theta^{(j)})+\Lambda_{B^{(j)}}(\theta^{(j)})<0,$$

for some queue j and in particular when this queue is the bottleneck queue. First, we construct a specific probability density function for the service times of queue j such that $M_{B^{(j)}}(\theta) < \infty$ for some $\theta > 0$. Then, we choose a specific distribution for the inter-arrival times and we see that, under the P&W change of measure as discussed in this chapter, it can happen that we get a heavy-tailed distribution.

Consider the following density function for the service times of queue j:

$$dF_{B^{(j)}}(b) = cb^{-3}e^{-ab}db, \qquad b > 1,$$

for some constant $a > 0$ and the normalization constant $c = 1/\int_1^\infty \frac{e^{-ab}}{b^3}\mathrm{d}b$. Note that an explicit expression for this integral contains the error function, but by construction this constant is greater than $1/\int_1^\infty \frac{1}{b^3}\mathrm{d}b = 2$. This implies that

$$M_{B^{(j)}}(\theta) = \int_1^\infty e^{\theta b} dF_{B^{(j)}}(b)\mathrm{d}b = c\int_1^\infty b^{-3}e^{(\theta-a)b}db.$$

Remark that $M_{B^{(j)}}(\theta) < \infty$ for all $\theta \leq a$ and $M_{B^{(j)}}(\theta) = \infty$ for all $\theta > a$.

For simplicity, take the inter-arrival times deterministic, and hence we have $\Lambda_A(-\theta) = -\theta A$ for some constant A. If $A > \frac{\log M_{B^{(j)}}(a)}{a} = \frac{\log \frac{c}{2}}{a}$, it follows that

$$\theta^{(j)} = \sup\{\theta : -\theta A + \log(M_{B^{(j)}}(\theta)) < 0\} = a.$$

Important to note here is that, when queue j is the bottleneck queue, (4.4) reduces to

$$dF^{\boldsymbol{\theta^*}}_{B^{(j)}}(b) = \frac{e^{\theta^{(j)}b}}{M_{B^{(j)}}(\theta^{(j)})} dF_{B^{(j)}}(b) = 2b^{-3}db,$$

which is a heavy-tailed distribution. This means that, starting with a light-tailed distribution, then applying the change of measure $\boldsymbol{\theta^*}$, we end up with a heavy tailed distribution. In particular, we note that

$$M'_{B^{(j)}}(\theta) = \int_1^\infty be^{\theta b} dF_{B^{(j)}}(b)db = c\int_1^\infty b^{-2}e^{(\theta-a)b}db,$$

so that $M'_{B^{(j)}}(a) = c$. Note that $M_{B^{(j)}}(a) = \frac{c}{2}$ and thus we find, by using (4.7), $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] = 2$ and thus we have $\mathbb{E}^{\boldsymbol{\theta^*}}\left[B^{(j^*)}\right] < \mathbb{E}^{\boldsymbol{\theta^*}}\left[A\right]$ when $A > \max\{2, \frac{\log \frac{c}{2}}{a}\}$. This means that the heavy-tailed distribution results in a stable queueing system under the change of measure. Moreover, this shows why the second assumption in Assumption 4.1 is necessary in order to satisfy Lemma 4.3 and Theorem 4.9. We remark that this assumption will not be a necessary in the development of an asymptotically efficient change of measure in Chapter 5.

CHAPTER 5

State-dependent importance sampling for non-Markovian tandem queues

In Chapter 4, we have shown that the *state-independent* change of measure for a non-Markovian tandem queue, as suggested by Parekh and Walrand, is not always asymptotically efficient. Here, we continue to develop a change of measure that *does* give an asymptotically efficient estimator. In order to do so, a *state-dependent* change of measure is required.

Earlier, for the 2-node Markovian tandem queue in [19] and later for more general Jackson networks in [22], a state-dependent change of measure was developed that gives an asymptotically efficient estimator, by using the so-called *subsolution approach* developed by Dupuis and Wang [21]. In this chapter, we extend this approach to non-Markovian tandem queues.

To construct a state-dependent change of measure for non-Markovian tandem queues using the subsolution approach, we extend the state description for Markovian tandem queues – consisting of the number of customers in each queue – with the residual inter-arrival time and the residual service times of all queues. This gives complete knowledge of the system and we will see that this is sufficient information to construct a state-dependent change of measure.

At first we will analyze how the subsolution approach works for the single $GI|GI|1$ queue. Using this approach, we find the same change of measure as in [36], with a much shorter proof of asymptotic efficiency than the proof in [36]. Secondly, we consider the 2-node $GI|GI|1$ tandem queue. In that case the state description is small and therefore the proofs of asymptotic efficiency are still quite clean. We end with statements on the d-node $GI|GI|1$ tandem queue, where we present the results but omit the proofs (which are natural extensions of the proofs for the 2-node tandem queue).

This chapter is structured as follows. In Section 5.1 we introduce the model and notation, and we provide some background knowledge about the subsolution approach and importance sampling using subsolutions. Then in Section 5.2 we present a state-dependent change of measure and we prove that this change of measure gives an asymptotically efficient estimator for the probability of interest. Section 5.3 concludes this chapter with some numerical results.

5.1 Model and preliminaries

5.1.1 The model

In this chapter, we again consider the d-node $GI|GI|1$ tandem queue. We refer to Section 2.1 for a detailed introduction of the model, notation and assumptions. In addition, when considering the d-node tandem queue for $d > 1$, we make the following assumption.

Assumption 5.1. *The supports of all distributions are bounded, that is, there exist constants* $Q^{(j)} < \infty \ \forall j = 0, \ldots, d$ *such that* $\mathbb{P}(A_k < Q^{(0)}) = 1$, $\mathbb{P}(B_k^{(j)} < Q^{(j)}) = 1$.

When a state-dependent change of measure for a Markovian tandem queue was studied in [19], the state description consisted of the number of customers in each queue. To study a change of measure for a non-Markovian queue, we extend this state description by adding the residual inter-arrival time and the residual service times. Therefore, for the d-node tandem queue the state description is a vector with $2d+1$ components. As in [19], we define all processes embedded at a transition for the number of customers in a queue. For any vector $\mathbf{y}$ with $2d+1$ components $y_1, \ldots, y_{2d+1}$ we introduce shorthand notation $\bar{y}_j = y_{d+1+j}$, $j = 0, \ldots, d$. Using this, we let

$$\mathbf{Z}_i = (Z_{1,i}, \ldots, Z_{d,i}, \bar{Z}_{0,i}, \ldots, \bar{Z}_{d,i}),$$

denote the state of the system after i transitions. Here, $Z_{j,i}$, $j = 1, \ldots, d$, is the number of customers in queue j after i transitions, $\bar{Z}_{0,i}$ is the residual inter-arrival time after i transitions, and $\bar{Z}_{j,i}$, $j = 1, \ldots, d$, is the residual service time at queue j after i transitions. If $Z_{j,i} = 0$ for some j, then we set $\bar{Z}_{j,i} = 0$.
As we start with one customer in queue 1, we let $\mathbf{Z}_0 = (1, 0, \ldots, 0)$. We have $\mathbf{Z}_{i+1} = \mathbf{Z}_i + V_Z(\mathbf{Z}_i)$, where $V_Z(\mathbf{Z}_i)$ denotes the next transition when the state of the system after the i^{th} transition is $\mathbf{Z}_i$. For $i > 0$ we define $V_Z(\mathbf{Z}_i)$ in terms of the shortest residual time $\mathcal{Z}(\mathbf{Z}_i) = \min_{k \in \{0\} \cup \{j : Z_{j,i} > 0\}} \{\bar{Z}_{k,i}\}$, that is, the minimum of the residual inter-arrival time and the residual service times of all customers in service. In other words; $\mathcal{Z}(\mathbf{Z}_i)$ is the time until the next transition. In Remark 5.2 below we make a convention about what happens when $\mathcal{Z}(\mathbf{Z}_i)$ is not uniquely defined.

The possible transitions when the current state is $\mathbf{Z}_i$, corresponding to the cases in (5.1) below, are an arrival at queue 1, a customer going from queue j to queue $j+1$, $j \in \{1, \ldots, d-1\}$ and a departure from queue d, which is a departure from the system. As the process starts at $(1, 0, \ldots, 0)$, we need a different transition for $\mathbf{Z}_0$. It will become clear in Remarks 5.9 and 5.10 why this technicality is needed.

Let $\mathbf{e}_j$ denote the j^{th} unit vector, then for all $i > 0$ we have

$$V_Z(\mathbf{Z}_i) = -\mathcal{Z}(\mathbf{Z}_i)\left(\bar{\mathbf{e}}_0 + \sum_{k=1}^{d} \bar{\mathbf{e}}_k \mathbb{1}\{Z_{k,i} > 0\}\right) + \begin{cases} \mathbf{e}_1 + A_i\bar{\mathbf{e}}_0 + B_i^{(1)}\bar{\mathbf{e}}_1\mathbb{1}\{Z_{1,i}=0\} & \text{if } \mathcal{Z}(\mathbf{Z}_i) = \bar{Z}_{0,i}, \\ \mathbf{e}_{j+1} - \mathbf{e}_j + B_i^{(j)}\bar{\mathbf{e}}_j\mathbb{1}\{Z_{j,i}>1\} + B_i^{(j+1)}\bar{\mathbf{e}}_{j+1}\mathbb{1}\{Z_{j+1,i}=0\} & \\ & \text{if } \mathcal{Z}(\mathbf{Z}_i) = \bar{Z}_{j,i} \ j = 1, ..., d-1, \\ -\mathbf{e}_d + B_i^{(d)}\bar{\mathbf{e}}_d\mathbb{1}\{Z_{d,i}>1\} & \text{if } \mathcal{Z}(\mathbf{Z}_i) = \bar{Z}_{d,i}, \end{cases} \tag{5.1}$$

and

$$V_Z(\mathbf{Z}_0) = A_0\bar{\mathbf{e}}_0 + B_0^{(1)}\bar{\mathbf{e}}_1,$$

where A_i is the inter-arrival time *if* the i^{th} transition is an arrival. If the i^{th} transition is not an arrival, then $A_i = 0$. Similarly, we have that $B_i^{(j)}$ is the service time of a customer at queue j *if* a service is starting at queue j at the i^{th} transition. If the i^{th} transition is not a service at queue j, then $B_i^{(j)} = 0$. We note that this notation is slightly different compared to Chapters 2-4. The description of the next transition $V_Z(\mathbf{Z}_i)$ in (5.1) means that depending on the current state of the system it is known which type of transition to take and each of them can have infinitely many possibilities in terms of the residuals (depending on the distribution).
We note that in (5.1) we consider $\mathbb{1}\{Z_{j,i} > 1\}$ and not $\mathbb{1}\{Z_{j,i} > 0\}$, since we do not need a new service time for queue j when queue j is empty after the transition. We do need a new service time for queue j when a customer departs from queue $j-1$ and queue j was empty before, that is, $Z_{j,i} = 0$.

Remark 5.2. *If $\mathcal{Z}(\mathbf{Z}_i)$ is not uniquely defined, it is not clear in which order the transitions should happen. However, in most cases it is not important for our probability of interest which of the transitions happens first.* Only *if $\mathcal{Z}(\mathbf{Z}_i) = \bar{Z}_{0,i} = \bar{Z}_{d,i}$, that is, an arrival happens at the same time as a departure from the system, the order* does *matter, and we use the convention that the departure occurs first, that is, $\mathcal{Z}(\mathbf{Z}_i) = \bar{Z}_{d,i}$.*

As for the Markovian system in [19, 22], we will work with the scaled process. Therefore we define

$$\mathbf{X}_i = (X_{1,i}, \ldots, X_{d,i}, \bar{X}_{0,i}, \ldots, \bar{X}_{d,i}) = \frac{1}{N}\mathbf{Z}_i,$$

and we have

$$\mathbf{X}_{i+1} = \mathbf{X}_i + \frac{1}{N}V_X(\mathbf{X}_i),$$

where $V_X(\mathbf{X}_i) = V_Z(\mathbf{X}_i N)$ is the $(i+1)^{th}$ transition when the state of the scaled system is $\mathbf{X}_i$. The advantage of the scaled system is that the first d elements of

the state description are always in $[0,1]$, which does not increase as N increases. For the other $d+1$ elements, there is no advantage to have a scaled system, however, it is notationally easier to scale *all* elements of the state description rather than only the first d elements. Note that for the scaled system we will use a similar convention as in Remark 5.2. Similar to the unscaled system, we define $\mathcal{X}(\mathbf{X}_i) = \frac{\mathcal{Z}(\mathbf{X}_i)}{N}$. Hence, for $i > 0$, we have that

$$V_X(\mathbf{X}_i) = -\mathcal{X}(\mathbf{X}_i)N\left(\bar{\mathbf{e}}_0 + \sum_{k=1}^{d}\bar{\mathbf{e}}_k \mathbb{1}\{X_{k,i} > 0\}\right) + \begin{cases} \mathbf{e}_1 + A_i\bar{\mathbf{e}}_0 + B_i^{(1)}\bar{\mathbf{e}}_1\mathbb{1}\{X_{1,i}=0\} & \text{if } \mathcal{X}(\mathbf{X}_i) = \bar{X}_{0,i}, \\ \mathbf{e}_{j+1} - \mathbf{e}_j + B_i^{(j)}\bar{\mathbf{e}}_j\mathbb{1}\{X_{j,i} > \frac{1}{N}\} + B_i^{(j+1)}\bar{\mathbf{e}}_{j+1}\mathbb{1}\{X_{j+1,i}=0\} & \\ \qquad\qquad \text{if } \mathcal{X}(\mathbf{X}_i) = \bar{X}_{j,i} \ j = 1, ..., d-1, & \\ -\mathbf{e}_d + B_i^{(d)}\bar{\mathbf{e}}_d\mathbb{1}\{X_{d,i} > \frac{1}{N}\} & \text{if } \mathcal{X}(\mathbf{X}_i) = \bar{X}_{d,i}, \end{cases} \tag{5.2}$$

and

$$V_X(\mathbf{X}_0) = A_0\bar{\mathbf{e}}_0 + B_0^{(1)}\bar{\mathbf{e}}_1,$$

where $\mathbf{X}_0 = (\frac{1}{N}, 0, \ldots, 0)$.

Now that we have defined the full state description, we define the goal set δ_e and taboo set δ_0 in the following way

$$\begin{aligned} \delta_e &= \{\mathbf{x} : \sum_{k=1}^{d} x_k = 1 - \frac{1}{N}, \mathcal{X}(\mathbf{x}) = \bar{x}_0\}, \\ \delta_0 &= \{\mathbf{x} : x_k = \bar{x}_k = 0 \ \forall k \in \{1, \ldots, d\}\}, \end{aligned} \tag{5.3}$$

where the goal set is reached if there are $N-1$ customers in the system and the next event is an arrival to the system, and the taboo set is reached when there are no customers in the system.

Using the definitions above, we define the time to reach δ_e before δ_0 as

$$\tau_N = \inf\{i > 0 : \mathbf{X}_i \in \delta_e, \mathbf{X}_k \notin \delta_0 \ \forall k = 1, \ldots, i-1\},$$

and we set $\tau_N = \infty$ when δ_0 is reached before δ_e. Now we can write our probability of interest, p_N, in terms of τ_N, as

$$p_N = \mathbb{P}\left(\tau_N < \infty\right).$$

Remark 5.3. *Note that we could define δ_e and δ_0 differently, that is, there are more states for which we know that the total number of customers will reach N or that the system will be empty. However, the current definitions are easy to use and it turns out that these definitions are sufficient for our proofs, see Remarks 5.13 and 5.19.*

From Chapter 2 and [36], who considers the case $d = 1$, we know that the decay rate of p_N equals

$$\gamma = -\lim_{N\to\infty} \frac{1}{N} \log p_N = -\Lambda_A(-\theta^*), \tag{5.4}$$

where $\theta^* = \min_j\{\theta^{(j)}\}$ with $\theta^{(j)}$ as defined in (2.3). We note that which queue is the bottleneck, see Definition 2.1, also determines the form of the so-called *most likely path* to overflow: if the overflow level is reached, this is most likely done along a specific path (that dominates the probability to reach the overflow level and hence determines the decay rate). When queue j is the bottleneck queue, it is therefore expected that along the most likely path $x_j > 0$.

5.1.2 Preliminaries

In order to estimate the probability of interest using simulation, we use importance sampling to make our event of interest less rare by changing the underlying probability distribution. In this chapter, we make the change of measure state-dependent by using the subsolution approach.

5.1.2.1 Subsolution approach

The subsolution approach for importance sampling has been introduced in [21]. Later, in [19, 22], it has been used to find a state-dependent change of measure that results in an asymptotically efficient estimator for p_N in the context of a Markovian tandem queue [19] and Jackson networks [22]. The definition of a classical subsolution is as follows.

Definition 5.4. *A real valued function $W(\mathbf{x})$ is called a classical subsolution if*

1. *$W(\mathbf{x})$ is continuously differentiable,*
2. *$\mathbb{H}(\mathbf{x}, DW(\mathbf{x})) \geq 0$ for every $\mathbf{x}$,*
3. *$W(\mathbf{x}) \leq 0$ for $\mathbf{x} \in \delta_e$,*

where

$$\mathbb{H}(\mathbf{x}, DW(\mathbf{x})) = -\log\left(\mathbb{E}\left[e^{-\langle DW(\mathbf{x}), V_X(\mathbf{x})\rangle}\right]\right). \tag{5.5}$$

In addition to this definition, in order to use the subsolution as a basis for a change of measure we require that

4. $W((\frac{1}{N}, 0, \ldots, 0)) = \gamma$,

so that the change of measure can be useful (that is, asymptotically efficient, see Definition 5.7 below). This means that along the most likely path to reach the overflow level, subsolution properties 2 and 3 are satisfied with equality, see also [22].

In [19], there is an additional condition in the definition of classical subsolution for the boundaries, when at least one of the queues is empty, and in [22] there are dedicated boundary functions for $\mathbb{H}(\mathbf{x}, DW(\mathbf{x}))$, which are used when one of the queues is empty. Instead, we will include the boundaries in $\mathbb{H}(\mathbf{x}, DW(\mathbf{x}))$, as the boundaries are included in $V_X(\mathbf{x})$ by means of indicator functions.

Remark 5.5. *The function $W(\mathbf{x})$ that we will construct in the sequel, consists of (a combination of) affine functions so that for all these affine functions its derivative $DW(\mathbf{x})$ is a constant. Whenever we consider an affine function, we will denote the derivative by $\boldsymbol{\alpha}$.*

5.1.2.2 Importance sampling simulation

In this chapter, we will propose a particular change of measure and prove that it results in an asymptotically efficient estimator for p_N, see also Definition 1.2. In order to define the change of measure, we first introduce some notation for the probability measure. With some abuse of notation we let $dF_{V_X(\mathbf{x})}(\mathbf{v})$ be the probability measure for some random vector $V_X(\mathbf{x})$. Typically, the vector $V_X(\mathbf{x})$ contains one random variable, see (5.2), but it may result in zero or two random variables as well, for example, when a customer moves from queue j to $j+1$ and queue $j+1$ is empty there are two random variables in $V_X(\mathbf{x})$. For example, we can interpret $dF_{V_X(\mathbf{x})}(\mathbf{v})$ as $dF_A(a)$ if $\mathcal{X}(\mathbf{x}) = \bar{x}_0$ and $x_1 > 0$.

Remark 5.6. *The probability measure of $V_X(\mathbf{x})$ can be written in a more formal way, which we illustrate by considering the arrival transition in a single queue. Thus, consider a state $\mathbf{x} = (x_1, \bar{x}_0, \bar{x}_1)$ for which $x_1 > 0$ and $\mathcal{X}(\mathbf{x}) = \bar{x}_0$. Then we have*

$$F_{V_X(\mathbf{x})}(\mathbf{v}|\mathbf{x} : \bar{x}_0 < \bar{x}_1, x_1 > 0) = \mathbb{1}\{v_1 \geq 1\} \cdot F_A(\bar{v}_0 + \bar{x}_0 N) \cdot \mathbb{1}\{\bar{v}_1 \geq -\bar{x}_0 N\},$$

and thus the density

$$\frac{\partial^3 F_{V_X(\mathbf{x})}(\mathbf{v}|\mathbf{x} : \bar{x}_0 < \bar{x}_1, x_1 > 0)}{\partial v_1 \partial \bar{v}_0 \partial \bar{v}_1} = \delta(v_1 - 1) \cdot dF_A(\bar{v}_0 + \bar{x}_0 N) \cdot \delta(\bar{v}_1 + \bar{x}_0 N).$$

Remark that two of the three components of the transition $V_X(\mathbf{x})$ are deterministic, while the remaining component consists of a random sample from the inter-arrival time distribution F_A, added to the deterministic value $-\bar{x}_0 N$. Also in larger models, most components of the transition are deterministic, see (5.2); therefore, we prefer the shorter but less precise notation in which only the random component(s) are written down.

Let $W(\mathbf{x})$ be a classical subsolution , see Section 5.1.2.1. Then we define the change of measure as

$$d\bar{F}_{V_X(\mathbf{x})}(\mathbf{v}|\mathbf{x}) = e^{\langle -DW(\mathbf{x}), \mathbf{v} \rangle} e^{\mathbb{H}(\mathbf{x}, DW(\mathbf{x}))} dF_{V_X(\mathbf{x})}(\mathbf{v}), \tag{5.6}$$

where $\bar{F}_{V_X(\mathbf{x})}(\mathbf{v}|\mathbf{x})$ is the distribution function under the new measure and $\mathbb{H}(\mathbf{x}, DW(\mathbf{x}))$ is defined in (5.5). Note that the latter quantity can be interpreted as a normalization constant. In fact (5.6) means that we apply an exponential tilt to a random variable (that depends on the random vector $V_X(\mathbf{x})$), for example, when $\mathcal{X}(\mathbf{x}) = \bar{x}_0$ and $x_1 > 0$ we are tilting the inter-arrival times exponentially with some parameter.

While performing a simulation run and changing the underlying probability distributions at every step, we keep track of the likelihood ratio. The likelihood ratio for a successful path $\mathcal{P} = (\mathbf{X}_i, i = 0, ..., \tau_N)$ is

$$L(\mathcal{P}) = \prod_{i=0}^{\tau_N - 1} \frac{dF_{V_X(\mathbf{X}_i)}(V_X(\mathbf{X}_i))}{d\bar{F}_{V_X(\mathbf{X}_i)}(V_X(\mathbf{X}_i)|\mathbf{X}_i)} = \prod_{i=0}^{\tau_N - 1} e^{\langle DW(\mathbf{X}_i), V_X(\mathbf{X}_i)\rangle} e^{-\mathbb{H}(\mathbf{X}_i, DW(\mathbf{X}_i))}, \tag{5.7}$$

where the second equality follows by using (5.6).

We now define the estimator of p_N to be $L(\mathcal{P})I(\mathcal{P})$, where $I(\mathcal{P})$ indicates whether we have reached level N or not, that is, $I(\mathcal{P}) = \mathbb{1}\{\tau_N < \infty\}$. This estimator is unbiased under the new measure, denoted by $\mathbb{Q}$, since

$$p_N = \mathbb{E}\left[I(\mathcal{P})\right] = \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})I(\mathcal{P})\right].$$

Using the decay rate of p_N, see (2.5), we find that Definition 1.2 is equivalent to the following definition.

Definition 5.7. *The estimator for p_N is asymptotically efficient if and only if*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right] \leq -2\gamma.$$

5.2 Asymptotically efficient change of measure

In this section we present the main result of this chapter. For readability we present it for three cases separately. In Section 5.2.1, we start with $d = 1$, that is, the single queue, where the subsolution approach results in a state-independent change of measure. Although this change of measure has been proven to be asymptotically efficient in [36], we present an alternative proof using the method that will be extended to a state-dependent change of measure for the d-node tandem queue. Secondly, we consider $d = 2$ in Section 5.2.2 in detail, as in this case the state vector consists of 5 dimensions only. Lastly, in Section 5.2.3, we present the result for the d-node tandem queue, but we omit the proofs since they are very similar to the 2-node case.

The **approach** of the subsolution method, as developed in [21], is the same in all cases defined above and is as follows.

1. For all possible $\mathbf{x}$, we find solutions $\boldsymbol{\alpha}$ to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$.

2. We construct a function $W(\mathbf{x})$ that is both continuously differentiable and satisfies properties 3 and 4 in and below Definition 5.4, as $N \to \infty$. The function $W(\mathbf{x})$ will be as indicated in Remark 5.5; more precisely, for each $\mathbf{x}$ we will have $DW(\mathbf{x}) \approx \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is a solution corresponding to $\mathbf{x}$ as found in Step 1.

3. We then use the function $W(\mathbf{x})$ to define a change of measure as in (5.6).

After we have proposed the change of measure, we will prove asymptotic efficiency of this change of measure.

5.2.1 The single $GI|GI|1$ queue

For the single queue, the model, as presented in Section 5.1.1, simplifies significantly. We will highlight the most important simplifications. To start with, the state description reduces to $\mathbf{x} = (x_1, \bar{x}_0, \bar{x}_1)$. Furthermore, queue 1 is never empty during a busy cycle of the system and so $\mathcal{X}(\mathbf{x}) = \min\{\bar{x}_0, \bar{x}_1\}$ for $i > 0$. Due to the same reason we always have $\mathbb{1}\{x_1 = 0\} = 0$ and, when there is a departure from the system, $\mathbb{1}\{x_1 > \frac{1}{N}\} = 1$ (otherwise $\mathbf{x} \in \delta_0$, that is, the system will become empty after the transition). In view of Remark 5.9 below, we do not yet substitute $\mathbb{1}\{x_1 = 0\} = 0$ in the remainder. Thus, (5.2) in the case $d = 1$ becomes

$$\begin{aligned} V_X(\mathbf{x}) = &- \mathcal{X}(\mathbf{x})N\left(\bar{\mathbf{e}}_0 + \bar{\mathbf{e}}_1 \mathbb{1}\{x_1 > 0\}\right) + \\ &\begin{cases} \mathbf{e}_1 + A_i\bar{\mathbf{e}}_0 + B_i^{(1)}\bar{\mathbf{e}}_1\mathbb{1}\{x_1 = 0\} & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0, \\ -\mathbf{e}_1 + B_i^{(1)}\bar{\mathbf{e}}_1\mathbb{1}\{x_1 > \frac{1}{N}\} & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1. \end{cases} \end{aligned}$$

In the next sections we will follow the subsolution approach step by step, and conclude with a proof of asymptotic efficiency of the change of measure based on the developed subsolution.

5.2.1.1 Solution to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ for all $\mathbf{x}$

We have for all possible $\mathbf{x} \neq (\frac{1}{N}, 0, 0)$, from (5.5) and the above, that

$$\begin{aligned} \mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) &= -\log(\mathbb{E}\left[e^{-\langle \boldsymbol{\alpha}, V_X(\mathbf{x})\rangle}\right]) \\ &= -\mathcal{X}(\mathbf{x})N(\bar{\alpha}_0 + \bar{\alpha}_1) + \\ &\quad \begin{cases} \alpha_1 - \Lambda_A(-\bar{\alpha}_0) - \Lambda_{B^{(1)}}(-\bar{\alpha}_1\mathbb{1}\{x_1 = 0\}) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0, \\ -\alpha_1 - \Lambda_{B^{(1)}}(-\bar{\alpha}_1) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1, \end{cases} \end{aligned}$$

and

$$\mathbb{H}((\tfrac{1}{N}, 0, 0), \boldsymbol{\alpha}) = -\Lambda_A(-\bar{\alpha}_0) - \Lambda_{B^{(1)}}(-\bar{\alpha}_1).$$

Then, using that $\mathbb{1}\{x_1 = 0\} = 0$, a solution to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ for all $\mathbf{x}$ during a busy cycle of the system is

$$\boldsymbol{\alpha} = (\alpha_1, \bar{\alpha}_0, \bar{\alpha}_1) = (-\gamma, \theta^*, -\theta^*),$$

where θ^* equals $\theta^{(1)}$ in (2.3).

Remark 5.8. *Another solution to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ is $\boldsymbol{\alpha} = (0, 0, 0)$. As this is equivalent to no change of measure, we will not use this solution.*

Remark 5.9. *To justify the necessity of $\mathbf{X}_0 \neq (0, 0, 0)$, we note that when $\mathbf{X}_0$ would be equal to $(0, 0, 0)$ our proposed solution does not satisfy $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ for $\mathcal{X}(\mathbf{x}) = \bar{x}_0$. Using $\mathbf{X}_0$ solves this issue. We will justify our specific choice of $\mathbf{X}_0 = (\frac{1}{N}, 0, 0)$ in Remark 5.10.*

5.2.1.2 Construction of $W(\mathbf{x})$

To use the subsolution approach, we want to choose the function $W(\mathbf{x})$ such that $W(\mathbf{x}) \leq 0$ for $\mathbf{x} \in \delta_e$ and $W((\frac{1}{N}, 0, 0)) = \gamma$, see properties 3 and 4 in and below Definition 5.4. For the single queue we have $\delta_e = \{\mathbf{x} : x_1 = 1 - \frac{1}{N}, \mathcal{X}(\mathbf{x}) = \bar{x}_0\}$, see (5.3). Combining this, together with $DW(\mathbf{x}) = \boldsymbol{\alpha} = (-\gamma, \theta^*, -\theta^*)$, we find

$$W(\mathbf{x}) = -\gamma x_1 + \theta^*(\bar{x}_0 - \bar{x}_1) + \gamma, \tag{5.8}$$

satisfying all the requirements when $N \to \infty$: indeed, $W(\mathbf{x}) = -\gamma(1 - \frac{1}{N}) + \theta^*(\bar{x}_0 - \bar{x}_1) + \gamma \leq \frac{\gamma}{N}$ for all $\mathbf{x} \in \delta_e$ and $W((\frac{1}{N}, 0, 0)) = \gamma\left(1 - \frac{1}{N}\right)$ and so when $N \to \infty$ the boundary conditions for $W(\mathbf{x})$ are satisfied. In Figure 5.1 the function $W(\mathbf{x})$ is displayed as a function of x_1 and $\bar{x}_0 - \bar{x}_1$.

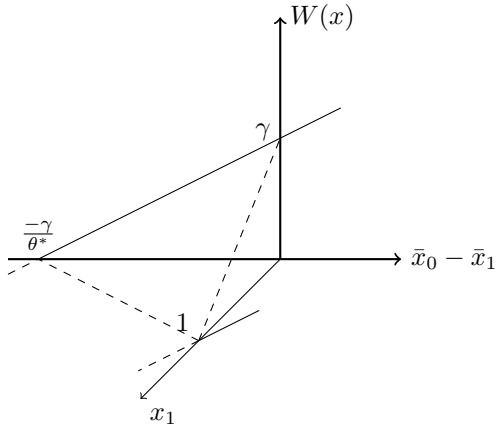


Figure 5.1 The classical subsolution $W(\mathbf{x})$ for the single queue as a function of x_1 and $\bar{x}_0 - \bar{x}_1$.

Remark 5.10. *To justify the specific choice of* $\bar{x}_0 = \bar{x}_1 = 0$ *in* $\mathbf{X}_0 = (\frac{1}{N}, 0, 0)$*, we note that any other choice does not satisfy* $W((\frac{1}{N}, \bar{x}_0, \bar{x}_1)) = \gamma$ *when* $N \to \infty$*. For the choice* $x_1 = \frac{1}{N}$ *we note that this is equivalent to starting with* 1 *customer in the system, and hence this is a natural choice.*

5.2.1.3 The change of measure

We now find that, using (5.6), the change of measure is for $\mathbf{x} \neq (\frac{1}{N}, 0, 0)$

$$d\bar{F}_{V_X(\mathbf{x})}(\mathbf{v}|\mathbf{x}) = \begin{cases} \frac{e^{-a\theta^*}}{M_A(-\theta^*)} dF_A(a) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0, \\ \frac{e^{b_1\theta^*}}{M_{B^{(1)}}(\theta^*)} dF_{B^{(1)}}(b_1) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1, \end{cases} \tag{5.9}$$

and

$$d\bar{F}_{V_X((\frac{1}{N},0,0))}(\mathbf{v}|(\tfrac{1}{N}, 0, 0)) = \frac{e^{-a\theta^* + b_1\theta^*}}{M_A(-\theta^*) M_{B^{(1)}}(\theta^*)} dF_A(a) dF_{B^{(1)}}(b_1). \tag{5.10}$$

Proposition 5.11. *The change of measure provided in* (5.9) *and* (5.10) *is the same change of measure as the state-independent change of measure from [36].*

Proof. The proof is straightforward by noting that the inter-arrival times are exponentially tilted with parameter $-\theta^*$ and the service times are exponentially tilted with parameter θ^*. □

5.2.1.4 Asymptotic Efficiency

In [36], it has been shown that the change of measure as mentioned above is asymptotically efficient under the condition that $\mathbb{P}(B_k^{(1)} < M) = 1$ for some finite constant M. This restriction to bounded service times is said to be a technicality, but the paper is not clear about how to remove it. Instead, we will give a different proof without the need for this condition.

Theorem 5.12. *Under Assumption 2.3, the state-independent change of measure based on* $DW(\mathbf{x}) = (-\gamma, \theta^*, -\theta^*)$*, see* (5.9) *and* (5.10)*, is asymptotically efficient for the single GI|GI|1 queue.*

Proof. We start with the log likelihood, $\log L(\mathcal{P})$, of any path $\mathcal{P} = (\mathbf{X}_i, i =$

$0, ..., \tau_N)$ such that $\tau_N < \infty$. By using (5.7), we find

$$\begin{aligned}
\log L(\mathcal{P}) &= N \sum_{i=0}^{\tau_N - 1} \langle DW(\mathbf{X}_i), \mathbf{X}_{i+1} - \mathbf{X}_i \rangle - \sum_{i=0}^{\tau_N - 1} \mathbb{H}(\mathbf{X}_i, DW(\mathbf{X}_i)) \\
&\leq N \sum_{i=0}^{\tau_N - 1} \langle DW(\mathbf{X}_i), \mathbf{X}_{i+1} - \mathbf{X}_i \rangle \\
&= N \langle DW(\mathbf{X}_0), \mathbf{X}_{\tau_N} - \mathbf{X}_0 \rangle \\
&= N \left(-\gamma(1 - \frac{1}{N} - \frac{1}{N}) + \theta^*(\bar{X}_{0,\tau_N} - \bar{X}_{0,0}) - \theta^*(\bar{X}_{1,\tau_N} - \bar{X}_{1,0}) \right) \\
&= N \left(-\gamma(1 - \frac{2}{N}) + \theta^*(\bar{X}_{0,\tau_N} - \bar{X}_{1,\tau_N}) \right) \\
&< -(N-2)\gamma, \qquad (5.11)
\end{aligned}$$

where we note that $\mathbf{X}_0 = (\frac{1}{N}, 0, 0)$, $\mathbf{X}_{\tau_N} = (1 - \frac{1}{N}, \bar{X}_{0,\tau_N}, \bar{X}_{1,\tau_N})$ and $\bar{X}_{0,\tau_N} < \bar{X}_{1,\tau_N}$ by definition of τ_N and δ_e. So we have

$$\begin{aligned}
\lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left[L(\mathcal{P}) I(\mathcal{P})\right] &= \lim_{N\to\infty} \frac{1}{N} \log \left(\mathbb{E}\left[L(\mathcal{P}) I(\mathcal{P}) | I(\mathcal{P}) = 1\right] \mathbb{P}(I(\mathcal{P}) = 1) \right) \\
&\leq \lim_{N\to\infty} \frac{1}{N} \log \left(\mathbb{E}\left[e^{-(N-2)\gamma} \right] p_N \right) \\
&= -2\gamma,
\end{aligned}$$

where the inequality follows from (5.11) and by definition of p_N, and the second equality follows by using (5.4). □

Remark 5.13. *In the proof of Theorem 5.12 we see why we defined δ_e as in* (5.3)*, and not simply $\delta_e = \{\mathbf{x} : x_1 = 1\}$; this allows for the second (strict) inequality to hold.*

5.2.2 The 2-node $GI|GI|1$ tandem queue

Although the subsolution approach for the single queue results in a state-independent and asymptotically efficient change of measure, it is known from both Chapter 4 (general distributions) and [12, 30] (Markovian distribution) that a state-independent change of measure for the $GI|GI|1$ tandem queue cannot be asymptotically efficient for all input parameters. We will see that the subsolution method, as it did for the Markovian case in [19, 22], results in a state-dependent change of measure.

In this section, we will use the same **approach** as in Section 5.2.1 and we conclude by proving asymptotic efficiency of the constructed change of measure.

5.2.2.1 Solutions to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq \mathbf{0}$ for all possible $\mathbf{x}$

We start the method by finding a solution to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ for all possible $\mathbf{x}$. As before, we let $DW(\mathbf{x}) = (\alpha_1, \alpha_2, \bar{\alpha}_0, \bar{\alpha}_1, \bar{\alpha}_2)$ and recall that $\mathcal{X}(\mathbf{x}) = \min_{k \in \{0\} \cup \{j : x_j > 0\}} \{\bar{x}_k\}$. Then we have from (5.5) for all possible $\mathbf{x} \neq (\frac{1}{N}, 0, 0, 0, 0)$

$$\begin{aligned}
\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) = &- \mathcal{X}(\mathbf{x}) N(\bar{\alpha}_0 + \bar{\alpha}_1 \mathbb{1}\{x_1 > 0\} + \bar{\alpha}_2 \mathbb{1}\{x_2 > 0\}) + \\
&\begin{cases}
- \log\left(e^{-\alpha_1} M_A(-\bar{\alpha}_0) M_{B^{(1)}}(-\bar{\alpha}_1 \mathbb{1}\{x_1 = 0\})\right) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0 \\
- \log\left(e^{\alpha_1 - \alpha_2} M_{B^{(1)}}(-\bar{\alpha}_1 \mathbb{1}\{x_1 > \frac{1}{N}\}) M_{B^{(2)}}(-\bar{\alpha}_2 \mathbb{1}\{x_2 = 0\})\right) & \\
& \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1 \\
- \log\left(e^{\alpha_2} M_{B^{(2)}}(-\bar{\alpha}_2 \mathbb{1}\{x_2 > \frac{1}{N}\})\right) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_2
\end{cases} \\
= &- \mathcal{X}(\mathbf{x}) N(\bar{\alpha}_0 + \bar{\alpha}_1 \mathbb{1}\{x_1 > 0\} + \bar{\alpha}_2 \mathbb{1}\{x_2 > 0\}) + \\
&\begin{cases}
\alpha_1 - \Lambda_A(-\bar{\alpha}_0) - \mathbb{1}\{x_1 = 0\} \Lambda_{B^{(1)}}(-\bar{\alpha}_1) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0, \\
-\alpha_1 + \alpha_2 - \mathbb{1}\{x_1 > \frac{1}{N}\} \Lambda_{B^{(1)}}(-\bar{\alpha}_1) - \mathbb{1}\{x_2 = 0\} \Lambda_{B^{(2)}}(-\bar{\alpha}_2) & \\
& \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1, \\
-\alpha_2 - \mathbb{1}\{x_2 > \frac{1}{N}\} \Lambda_{B^{(2)}}(-\bar{\alpha}_2) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_2,
\end{cases}
\end{aligned} \tag{5.12}$$

and

$$\mathbb{H}((\tfrac{1}{N}, 0, 0, 0, 0), \boldsymbol{\alpha}) = -\Lambda_A(-\bar{\alpha}_0) - \Lambda_{B^{(1)}}(-\bar{\alpha}_1).$$

The solutions to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ that we describe depend on the number of customers in queue 1 and the number of customers in queue 2. Note that there may be more solutions possible, but in order to find a state-dependent change of measure the current solutions turn out to be sufficient. The solutions are presented in Table 5.1. Clearly, $\boldsymbol{\alpha} = \mathbf{0}$ is always a solution to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ and is equivalent to no change of measure. However, using $DW(\mathbf{x}) = \boldsymbol{\alpha} = \mathbf{0}$ does not give a subsolution, since in that case it is impossible to satisfy properties 3 *and* 4 in Definition 5.4.

Table 5.1 Some solutions to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ for all possible $\mathbf{x}$. In this table, c.o.m. denotes change of measure.

	α_1	α_2	$\bar{\alpha}_0$	$\bar{\alpha}_1$	$\bar{\alpha}_2$
$x_1 > 0,\ x_2 > 0$	$\Lambda_A(-\theta^*)$	0	θ^*	$[-\widetilde{\theta}^{(1)}, -\theta^*]$	$[0, -\bar{\alpha}_1 - \theta^*]$
	$\Lambda_A(-\theta^*)$	$\Lambda_A(-\theta^*)$	θ^*	$[0, -\bar{\alpha}_2 - \theta^*]$	$[-\widetilde{\theta}^{(2)}, -\theta^*]$
$x_1 > 0,\ x_2 = 0$	$\Lambda_A(-\theta^*)$	0	θ^*	$[-\widetilde{\theta}^{(1)}, -\theta^*]$	$[0, -\bar{\alpha}_1 - \theta^*]$
$x_1 = 0,\ x_2 > 0$	$\Lambda_A(-\theta^*)$	$\Lambda_A(-\theta^*)$	θ^*	$[0, -\bar{\alpha}_2 - \theta^*]$	$[-\widetilde{\theta}^{(2)}, -\theta^*]$
$x_1 \geq 0,\ x_2 \geq 0$	0	0	0	0	0

In Table 5.1, $\widetilde{\theta}^{(1)} = \sup\{\theta : \Lambda_A(-\theta^*) + \Lambda_{B^{(1)}}(\theta) \leq 0\}$ and $\widetilde{\theta}^{(2)} = \sup\{\theta : \Lambda_A(-\theta^*) + \Lambda_{B^{(2)}}(\theta) \leq 0\}$; see Figures 5.2 and 5.3 for a graphical illustration of $\theta^{(1)}$, $\theta^{(2)}$, $\widetilde{\theta}^{(1)}$ and $\widetilde{\theta}^{(2)}$. Note that for all solutions proposed in Table 5.1 we

have

$$\mathbb{H}((\tfrac{1}{N},0,0,0,0),\boldsymbol{\alpha}) = -\Lambda_A(-\bar{\alpha}_0) - \Lambda_{B^{(1)}}(-\bar{\alpha}_1) \geq 0.$$

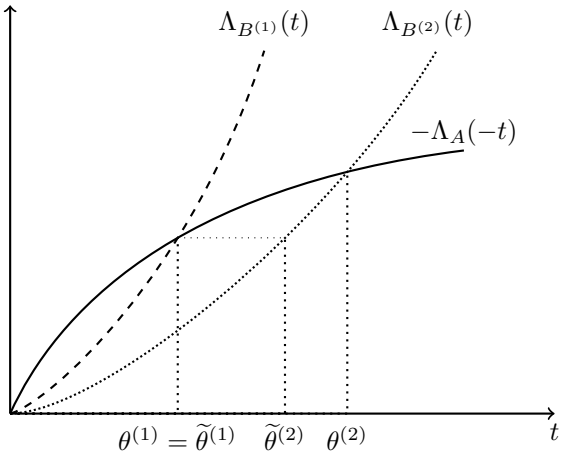


Figure 5.2 Illustration of $\theta^{(1)}$, $\theta^{(2)}$, $\widetilde{\theta}^{(1)}$ and $\widetilde{\theta}^{(2)}$ when queue 1 is the bottleneck queue, that is, $\theta^* = \theta^{(1)}$.

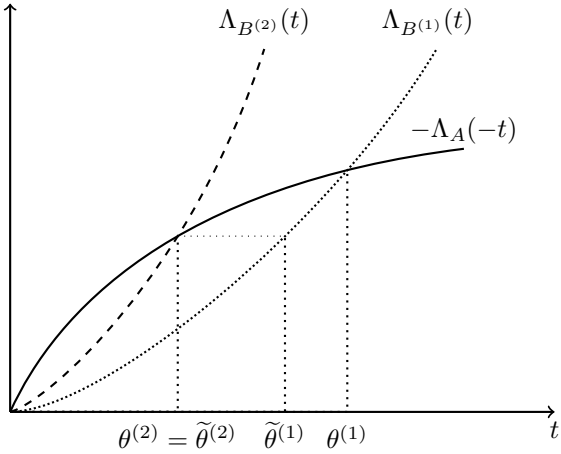


Figure 5.3 Similar to Figure 5.2, but when queue 2 is the bottleneck queue, that is, $\theta^* = \theta^{(2)}$.

Remark 5.14. *All solutions presented in Table 5.1 satisfy subsolution property 2 along the most likely path (as defined below Equation (5.4)) with equality. Recall that along the most likely path, $x_j > 0$ when queue j is the bottleneck queue. Note that the only solutions that do not necessarily satisfy $\mathbb{H}(\mathbf{x},\boldsymbol{\alpha}) = 0$, are*

those along the boundary $x_j = 0$ when queue j is the bottleneck queue (except for $\boldsymbol{\alpha} = \mathbf{0}$, which is always possible).

Remark 5.15. *We see from Table 5.1 that in some cases there is a possibility to choose $\bar{\alpha}_j > 0$. It is not obvious to do so, because then the service times of queue j are tilted in the 'wrong' way, since it would imply that on average the service times of queue j are even shorter than under the original measure. Nevertheless it is a solution that satisfies $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$.*

5.2.2.2 Construction of $W(\mathbf{x})$

Since there is no non-trivial solution for $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ that holds for all $\mathbf{x}$, that is, there is not a *single* affine function $W(\mathbf{x})$ such that for its derivative $\boldsymbol{\alpha}$ this condition holds for all $\mathbf{x}$, it turns out that a different change of measure needs to be applied in different 'regions' of the state space that are specified precisely later.

From Table 5.1, we can extract two non-trivial solutions that are dependent of the bottleneck queue; one which we can use for $x_1 > 0$ and one which we can use for $x_2 > 0$. These solutions, along with the trivial solution, can be found in Table 5.2 and are sufficient to obtain an asymptotically efficient estimator (as we will prove in Section 5.2.2.4, see Theorem 5.29). In the construction that is explained below, we will see that also cases in which both $x_j > 0$ can be handled using these subsolutions, see Equation (5.14) below. As it turns out, when both $x_j > 0$, either the solution for $x_1 > 0$ *or* the solution for $x_2 > 0$ will be used.

Table 5.2 Solutions to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ for all $\mathbf{x}$ used for the state-dependent change of measure.

	α_1	α_2	$\bar{\alpha}_0$	$\bar{\alpha}_1$	$\bar{\alpha}_2$
$x_2 > 0$	$\Lambda_A(-\theta^*)$	$\Lambda_A(-\theta^*)$	θ^*	0	$-\theta^*$
$x_1 > 0$	$\Lambda_A(-\theta^*)$	0	θ^*	$-\theta^*$	0
$x_1 \geq, x_2 \geq 0$	0	0	0	0	0

Remark 5.16. *The other solutions that are described in Table 5.1 can also be used to find a state-dependent change of measure that is proven to be asymptotically efficient using the same method that is described below, see Theorem 5.31.*

Remember that $\gamma = -\Lambda_A(-\theta^*)$ and so we define

$$\begin{aligned}\boldsymbol{\alpha}_1 &= (-\gamma, -\gamma, \theta^*, 0, -\theta^*),\\ \boldsymbol{\alpha}_2 &= (-\gamma, 0, \theta^*, -\theta^*, 0),\\ \boldsymbol{\alpha}_3 &= (0, 0, 0, 0, 0),\end{aligned}$$

which are used to determine the different changes of measure in each of the 'regions' of the state space. Thus, 'region' 1 – corresponding to $\boldsymbol{\alpha}_1$ – has to

cover $x_1 = 0$, 'region' 2 – corresponding to $\boldsymbol{\alpha}_2$ – has to cover $x_2 = 0$ and, finally, 'region' 3 – corresponding to $\boldsymbol{\alpha}_3$ has to cover $x_1 = x_2 = 0$. Note that $\boldsymbol{\alpha}_k$ is the notation for a solution to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$, whereas α_k is the notation for a component of some vector $\boldsymbol{\alpha}$. From these solutions $\boldsymbol{\alpha}_k$, we define three functions

$$W_k^\delta(\mathbf{x}) = \langle \boldsymbol{\alpha}_k, \mathbf{x} \rangle + \gamma - k\delta, \text{ for } k = 1, 2, 3, \tag{5.13}$$

for some $\delta > 0$ and their minimum

$$W^\delta(\mathbf{x}) = W_1^\delta(\mathbf{x}) \wedge W_2^\delta(\mathbf{x}) \wedge W_3^\delta(\mathbf{x}). \tag{5.14}$$

The idea of this construction, as in [19, 21, 22], and in particular subtracting $k\delta$, is that the minimum function $W^\delta(\mathbf{x})$ has gradient $\boldsymbol{\alpha}_1$ near $x_1 = 0$, gradient $\boldsymbol{\alpha}_2$ near $x_2 = 0$ and gradient $\boldsymbol{\alpha}_3$ near $x_1 = x_2 = 0$, as desired. On the interior, the gradient happens to be $\boldsymbol{\alpha}_1$.

In Figures 5.4 and 5.5 we give a rough illustration of the behavior of the function $W^\delta(\mathbf{x})$, considering the dependence on the scaled queue lengths x_1 and x_2 while neglecting the dependence on the scaled residuals $\bar{x}_0$, $\bar{x}_1$ and $\bar{x}_2$. For large N these can indeed be neglected due to scaling, in particular in the case of bounded supports of the inter-arrival time and service times. However, for unbounded supports there can be expections to this idea due to a very large residual inter-arrival time or a very large residual service time. For example, at $x_1 = 0$ we have $W_3^\delta(\mathbf{x}) < \min\{W_1^\delta(\mathbf{x}), W_2^\delta(\mathbf{x})\}$ if and only if $x_2 < \frac{2\delta + \theta^*(\bar{x}_0 - \bar{x}_2)}{\gamma}$. When $\bar{x}_0 - \bar{x}_2$ is large, $W_3^\delta(\mathbf{x})$ is the minimum function even when x_2 is not near 0. Indeed, in this case it is very likely that the system empties out before the next arriving customer and hence it makes sense that $W^\delta(\mathbf{x}) = W_3^\delta(\mathbf{x})$ (and no change of measure will be applied).

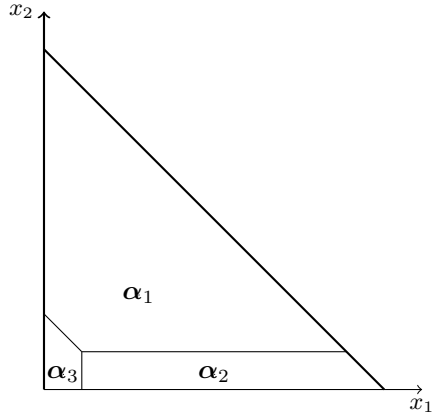


Figure 5.4 Main dimensions (x_1, x_2) (the scaled queue sizes) of the state space, neglecting $\bar{x}_0, \bar{x}_1, \bar{x}_2$ (the scaled residuals) by setting them equal to 0. The figure shows the three regions where the different $W_k^\delta(\mathbf{x})$, $k = 1, 2, 3$, are the minimum function.

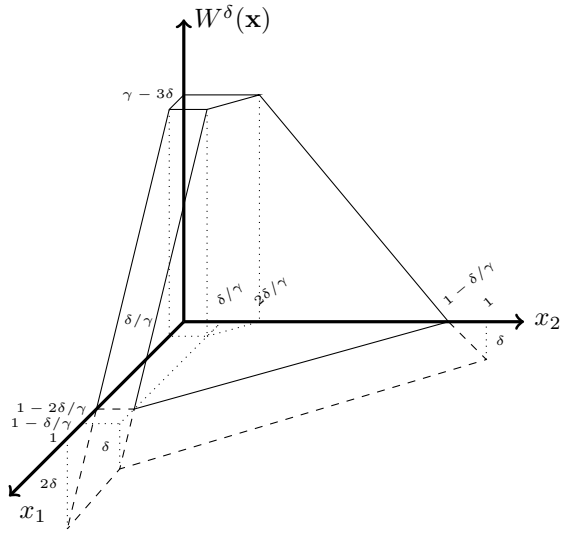


Figure 5.5 Illustration of the dependence of $W^\delta(\mathbf{x})$ on (x_1, x_2), neglecting the scaled residuals as in Figure 5.4. The figure here shows how for $\mathbf{x}$ in region k we have $W^\delta(\mathbf{x}) = W_k^\delta(\mathbf{x})$, and hence $DW^\delta(\mathbf{x}) = \boldsymbol{\alpha}_k$.

Note that $W^\delta(\mathbf{x})$ satisfies properties 2-4 in and below Definition 5.4 by construction. As we have different functions for different regions, $W^\delta(\mathbf{x})$ is not a continuously differentiable function, which is the first property of a classical subsolution. Therefore, we apply a similar mollification procedure as in previous work on Markovian systems, see [19, 22]. This mollification ensures that a gradient exists throughout the whole parameter space. We let

$$W^{\varepsilon,\delta}(\mathbf{x}) = -\varepsilon \log \sum_{k=1}^{3} e^{-W_k^\delta(\mathbf{x})/\varepsilon}. \tag{5.15}$$

When $\varepsilon \to 0$, $W^{\varepsilon,\delta}(\mathbf{x})$ converges to $W^\delta(\mathbf{x})$. Another result of this choice of $W^{\varepsilon,\delta}(\mathbf{x})$ is that

$$\begin{aligned} DW^{\varepsilon,\delta}(\mathbf{x}) &= \sum_{k=1}^{3} \rho_k(\mathbf{x}) \boldsymbol{\alpha}_k \\ &= (-(\rho_1(\mathbf{x}) + \rho_2(\mathbf{x}))\gamma, -\rho_1(\mathbf{x})\gamma, (\rho_1(\mathbf{x}) + \rho_2(\mathbf{x}))\theta^*, -\rho_2(\mathbf{x})\theta^*, -\rho_1(\mathbf{x})\theta^*), \end{aligned} \tag{5.16}$$

with

$$\rho_k(\mathbf{x}) = \frac{e^{-W_k^\delta(\mathbf{x})/\varepsilon}}{\sum_j e^{-W_j^\delta(\mathbf{x})/\varepsilon}}. \tag{5.17}$$

Throughout this chapter, we make the following assumptions on ε and δ, as in [19]. We remark that we let ε and δ depend on N, though for brevity we do not explicitly write this dependence.

Assumption 5.17. *We choose ε and δ dependent on N, such that*

- $\lim_{N\to\infty} \varepsilon = 0$,
- $\lim_{N\to\infty} \delta = 0$,
- $\lim_{N\to\infty} \varepsilon N = \infty$,
- $\lim_{N\to\infty} \frac{\varepsilon}{\delta} = 0$.

As a result, $W^{\varepsilon,\delta}(\mathbf{x})$ satisfies properties 3 and 4 in and below Definition 5.4 as $N \to \infty$, which follows immediately from the following lemma.

Lemma 5.18. *Under Assumption 2.3, we have for the 2-node $GI|GI|1$ tandem queue that $W^{\varepsilon,\delta}(\mathbf{x}) \leq \frac{\gamma}{N} - \delta$ for $\mathbf{x} \in \delta_e$ and $\gamma(1-\frac{1}{N}) - 3\delta - \varepsilon\log(3) \leq W^{\varepsilon,\delta}(\mathbf{X}_0) \leq \gamma - 3\delta$.*

Proof. We have for $\mathbf{x} \in \delta_e$ by (5.15),

$$W^{\varepsilon,\delta}(\mathbf{x}) \leq W_1^{\delta}(\mathbf{x}) = -\gamma(x_1 + x_2) + \theta^*(\bar{x}_0 - \bar{x}_2) + \gamma - \delta \leq \frac{\gamma}{N} - \delta,$$

where the final inequality follows as $x_1 + x_2 = 1 - \frac{1}{N}$ and $\bar{x}_0 - \bar{x}_2 \leq 0$ for $\mathbf{x} \in \delta_e$.

For $W^{\varepsilon,\delta}(\mathbf{X}_0)$, we note that $\mathbf{X}_0 = (\frac{1}{N}, 0, 0, 0, 0)$. By (5.15) we have

$$\begin{aligned}
W^{\varepsilon,\delta}(\mathbf{X}_0) &\leq W_3^{\delta}(\mathbf{X}_0) = \gamma - 3\delta, \\
W^{\varepsilon,\delta}(\mathbf{X}_0) &= -\varepsilon\log\left(e^{(-\gamma(1-\frac{1}{N})+\delta)/\varepsilon} + e^{(-\gamma(1-\frac{1}{N})+2\delta)/\varepsilon} + e^{(-\gamma+3\delta)/\varepsilon}\right) \\
&\geq -\varepsilon\log\left(3e^{(-\gamma+\frac{\gamma}{N}+3\delta)/\varepsilon}\right) \\
&= \gamma(1-\frac{1}{N}) - 3\delta - \varepsilon\log(3).
\end{aligned}$$

□

Remark 5.19. *The proof of Lemma 5.18 again explains our choice of δ_e: it is chosen such that $W^{\varepsilon,\delta}(\mathbf{x})$ can be upper bounded for $\mathbf{x} \in \delta_e$.*

Note that by construction we expect that also the second property of a classical subsolution is satisfied for $W^{\varepsilon,\delta}(\mathbf{x})$, that is, $\mathbb{H}(\mathbf{x}, DW^{\varepsilon,\delta}(\mathbf{x})) \geq 0$ when $N \to \infty$, which we will prove in Lemma 5.24.

5.2.2.3 The change of measure

In this section, we will give examples of the change of measure based on $W_k^\delta(\mathbf{x})$, see (5.13), for some parts of the state description. This gives some insight in the change of measure that is applied in the mollified function $W^{\varepsilon,\delta}(\mathbf{x})$ and how the change of measure that we use in the current chapter relates to previous work. When choosing $DW(\mathbf{x})$ in (5.6) as a constant vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \bar{\alpha}_0, \bar{\alpha}_1, \bar{\alpha}_2)$, the change of measure can be written in terms of $\bar{\alpha}_0$, $\bar{\alpha}_1$ and $\bar{\alpha}_2$ as

$$d\bar{F}_{V_X(\mathbf{x})}(\mathbf{v}|\mathbf{x}) = \begin{cases} \frac{e^{-a\bar{\alpha}_0 - b_1\bar{\alpha}_1 \mathbb{1}\{x_1=0\}}}{M_A(-\bar{\alpha}_0)M_{B^{(1)}}(-\bar{\alpha}_1\mathbb{1}\{x_1=0\})} dF_A(a)dF_{B^{(1)}}(b_1) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0, \\ \frac{e^{-b_1\bar{\alpha}_1\mathbb{1}\{x_1>\frac{1}{N}\} - b_2\bar{\alpha}_2\mathbb{1}\{x_2=0\}}}{M_{B^{(1)}}(-\bar{\alpha}_1\mathbb{1}\{x_1>\frac{1}{N}\})M_{B^{(2)}}(-\bar{\alpha}_2\mathbb{1}\{x_2=0\})} dF_{B^{(1)}}(b_1)dF_{B^{(2)}}(b_2) & \\ & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1, \\ \frac{e^{-b_2\bar{\alpha}_2\mathbb{1}\{x_2>\frac{1}{N}\}}}{M_{B^{(2)}}(-\bar{\alpha}_2\mathbb{1}\{x_2>\frac{1}{N}\})} dF_{B^{(2)}}(b_2) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_2, \end{cases} \tag{5.18}$$

for $\mathbf{x} \neq \mathbf{X}_0$ and

$$d\bar{F}_{V_X(\mathbf{X}_0)}(\mathbf{v}|\mathbf{X}_0) = \frac{e^{-a\bar{\alpha}_0 - b_1\bar{\alpha}_1}}{M_A(-\bar{\alpha}_0)M_{B^{(1)}}(-\bar{\alpha}_1)} dF_A(a)dF_{B^{(1)}}(b_1). \tag{5.19}$$

In each of the examples below, we consider the cases where our change of measure based on $W^{\varepsilon,\delta}(\mathbf{x})$ gets close, as $N \to \infty$, to an 'affine' change of measure as above, that is, based on some $W_k^\delta(\mathbf{x})$ with constant gradient $\boldsymbol{\alpha}$. It is two of these latter 'limiting' change of measures that we consider in the following.

Examples of the change of measure

Example 5.20. *Where $W^{\varepsilon,\delta}(\mathbf{x})$ gets close to $W_2^\delta(\mathbf{x})$, we let $\boldsymbol{\alpha} = DW_2^\delta(\mathbf{x}) = (-\gamma, 0, \theta^*, -\theta^*, 0)$. Then* (5.18) *and* (5.19) *reduce to*

$$d\bar{F}_{V_X(\mathbf{x})}(\mathbf{v}|\mathbf{x}) = \begin{cases} \frac{e^{-a\theta^* + b_1\theta^*}}{M_A(-\theta^*)M_{B^{(1)}}(\theta^*)} dF_A(a)dF_{B^{(1)}}(b_1) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0 \text{ and } x_1 = 0, \\ \frac{e^{-a\theta^*}}{M_A(-\theta^*)} dF_A(a) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0 \text{ and } x_1 > 0, \\ \frac{e^{b_1\theta^*\mathbb{1}\{x_1>\frac{1}{N}\}}}{M_{B^{(1)}}(\theta^*\mathbb{1}\{x_1>\frac{1}{N}\})} dF_{B^{(1)}}(b_1)dF_{B^{(2)}}(b_2) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1 \text{ and } x_2 = 0, \\ \frac{e^{b_1\theta^*\mathbb{1}\{x_1>\frac{1}{N}\}}}{M_{B^{(1)}}(\theta^*\mathbb{1}\{x_1>\frac{1}{N}\})} dF_{B^{(1)}}(b_1) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1 \text{ and } x_2 > 0, \\ dF_{B^{(2)}}(b_2) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_2, \end{cases}$$

for $\mathbf{x} \neq \mathbf{X}_0$ and

$$d\bar{F}_{V_X(\mathbf{X}_0)}(\mathbf{v}|\mathbf{X}_0) = \frac{e^{-a\theta^* + b_1\theta^*}}{M_A(-\theta^*)M_{B^{(1)}}(\theta^*)} dF_A(a)dF_{B^{(1)}}(b_1).$$

When queue 1 *is the bottleneck queue, this corresponds to the state-independent change of measure that has been studied in Chapter 4 and [25, 35]. On the other hand, when queue* 2 *is the bottleneck queue, the inter-arrival times are still exponentially tilted with parameter* $-\theta^* = -\theta^{(2)}$ *and so it also corresponds to the state-independent change of measure studied in those papers. However, it is not the service times of queue* 2 *that are exponentially tilted, but the service times of queue* 1 *that are exponentially tilted with parameter* $\theta^* = \theta^{(2)}$.

Remark 5.21. *We note that the change of measure used here along the horizontal boundary is different from the one in [19] for a* 2*-node Markovian tandem queue. Let* λ, μ_1 *and* μ_2 *be the exponential rates for the inter-arrival times, and service times of queue* 1 *and queue* 2 *under the original measure. If we would consider* $DW_2^\delta(\mathbf{x}) \equiv \boldsymbol{\alpha}_2 = (-\gamma, 0, \theta^*, -\widetilde{\theta}^{(1)}, 0)$, *which is a solution to* $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ *according to Table 5.1, then we find that under the change of measure the service times of queue* 1 *have an exponential distribution with rate* $\frac{\mu_1\lambda}{\mu_2}$. *This* does *correspond with the results from [19], although they consider the embedded discrete time Markov chain, while we consider the continuous time Markov chain.*

Example 5.22. *Where* $W^{\varepsilon,\delta}(\mathbf{x})$ *gets close to* $W_1^\delta(\mathbf{x})$, *we let* $\boldsymbol{\alpha} = DW_1^\delta(\mathbf{x}) = (-\gamma, -\gamma, \theta^*, 0, -\theta^*)$. *Then* (5.18) *and* (5.19) *reduce to*

$$d\bar{F}_{V_X(\mathbf{x})}(\mathbf{v}|\mathbf{x}) = \begin{cases} \frac{e^{-a\theta^*}}{M_A(-\theta^*)} dF_A(a) dF_{B^{(1)}}(b_1) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0 \text{ and } x_1 = 0, \\ \frac{e^{-a\theta^*}}{M_A(-\theta^*)} dF_A(a) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0 \text{ and } x_1 > 0, \\ dF_{B^{(1)}}(b_1) dF_{B^{(2)}}(b_2) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1 \text{ and } x_2 = 0, \\ dF_{B^{(1)}}(b_1) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1 \text{ and } x_2 > 0, \\ \frac{e^{b_2\theta^* \mathbb{1}\{x_2 > \frac{1}{N}\}}}{M_{B^{(2)}}(\theta^* \mathbb{1}\{x_2 > \frac{1}{N}\})} dF_{B^{(2)}}(b_2) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_2, \end{cases}$$

for $\mathbf{x} \neq \mathbf{X}_0$ *and*

$$d\bar{F}_{V_X(\mathbf{X}_0)}(\mathbf{v}|\mathbf{X}_0) = \frac{e^{-a\theta^*}}{M_A(-\theta^*)} dF_A(a) dF_{B^{(1)}}(b_1).$$

When queue 1 *is the bottleneck queue, this change of measure corresponds to the state-independent change of measure studied in Chapter 4 and [25, 35] in the sense that the inter-arrival times are exponentially tilted with parameter* $-\theta^* = -\theta^{(1)}$. *In contrast to an exponential tilt for the service times of queue* 1, *which happens for the state-independent change of measure, here we exponentially tilt the service times of queue* 2 *with parameter* $\theta^* = \theta^{(1)}$. *When queue* 2 *is the bottleneck queue, the change of measure described above corresponds to the state-independent change of measure that has been studied in Chapter 4 and [25, 35].*

5.2.2.4 Asymptotic Efficiency

Using (5.16), we can find a lower bound on $\mathbb{H}(\mathbf{x}, DW^{\varepsilon,\delta}(\mathbf{x}))$ which goes to 0 when $N \to \infty$. Recall that the constants $Q^{(j)} < \infty \ \forall j = 0, 1, 2$ are such that $\mathbb{P}(A_k < Q^{(0)}) = 1$ and $\mathbb{P}(B_k^{(j)} < Q^{(j)}) = 1$.

Remark 5.23. *Due to these bounded supports of the service time distributions, we note that equality is achieved in (2.3) for all queues j. That is, we have $\Lambda_A(-\theta^{(j)}) + \Lambda_{B^{(j)}}(\theta^{(j)}) = 0$ for all queues j. In particular, equality holds for the bottleneck queue. Similarly, equality in the definition of $\widetilde{\theta}^{(1)}$ and $\widetilde{\theta}^{(2)}$ also holds.*

Lemma 5.24. *Under Assumptions 2.3 and 5.1, we have for the 2-node $GI|GI|1$ tandem queue for all possible* $\mathbf{x}$

$$\mathbb{H}(\mathbf{x}, DW^{\varepsilon,\delta}(\mathbf{x})) \geq -(\theta^* \max\{Q^{(0)}, Q^{(1)}, Q^{(2)}\} + \gamma)e^{-\delta/\varepsilon}.$$

Proof. We substitute (5.16) in (5.12) to find $\mathbb{H}(\mathbf{x}, DW^{\varepsilon,\delta}(\mathbf{x}))$ for $\mathbf{x} \neq \mathbf{X}_0$ equal to

$$\begin{aligned}
&- \mathcal{X}(\mathbf{x})N((\rho_1(\mathbf{x}) + \rho_2(\mathbf{x}))\theta^* - \rho_2(\mathbf{x})\theta^* \mathbb{1}\{x_1 > 0\} - \rho_1(\mathbf{x})\theta^* \mathbb{1}\{x_2 > 0\}) + \\
&\begin{cases}
-(\rho_1(\mathbf{x}) + \rho_2(\mathbf{x}))\gamma - \Lambda_A(-(\rho_1(\mathbf{x}) + \rho_2(\mathbf{x}))\theta^*) & \\
\quad -\mathbb{1}\{x_1 = 0\}\Lambda_{B^{(1)}}(\rho_2(\mathbf{x})\theta^*) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0 \\
(\rho_1(\mathbf{x}) + \rho_2(\mathbf{x}))\gamma - \rho_1(\mathbf{x})\gamma - \mathbb{1}\{x_1 > \frac{1}{N}\}\Lambda_{B^{(1)}}(\rho_2(\mathbf{x})\theta^*) & \\
\quad -\mathbb{1}\{x_2 = 0\}\Lambda_{B^{(2)}}(\rho_1(\mathbf{x})\theta^*) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1 \\
\rho_1(\mathbf{x})\gamma - \mathbb{1}\{x_2 > \frac{1}{N}\}\Lambda_{B^{(2)}}(\rho_1(\mathbf{x})\theta^*) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_2
\end{cases} \\
&\geq -\mathcal{X}(\mathbf{x})N(\rho_1(\mathbf{x})\theta^* \mathbb{1}\{x_2 = 0\} + \rho_2(\mathbf{x})\theta^* \mathbb{1}\{x_1 = 0\}) + \\
&\begin{cases}
-\rho_2(\mathbf{x})\mathbb{1}\{x_1 = 0\}\Lambda_{B^{(1)}}(\theta^*) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0, \\
\rho_2(\mathbf{x})\gamma - \rho_2(\mathbf{x})\mathbb{1}\{x_1 > \frac{1}{N}\}\Lambda_{B^{(1)}}(\theta^*) - \rho_1(\mathbf{x})\mathbb{1}\{x_2 = 0\}\Lambda_{B^{(2)}}(\theta^*) & \\
\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1, \\
\rho_1(\mathbf{x})\gamma - \rho_1(\mathbf{x})\mathbb{1}\{x_2 > \frac{1}{N}\}\Lambda_{B^{(2)}}(\theta^*) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_2,
\end{cases}
\end{aligned} \tag{5.20}$$

where the inequality follows from convexity of the log-moment generating functions, $\rho_k(\mathbf{x}) \in [0,1]$ and by definition of $\gamma = -\Lambda_A(-\theta^*) \geq 0$. By using $\mathbb{1}\{x_k > \frac{1}{N}\} \leq 1$ for $k = 1, 2$, $\Lambda_A(-\theta^*) + \Lambda_{B^{(j)}}(\theta^*) \leq 0$ for all queues j by definition of θ^*, γ and the bounded supports (and hence bounded $\mathcal{X}(\mathbf{x})$), we find that (5.20) is greater than or equal to

$$\begin{aligned}
&- \max\{Q^{(0)}, Q^{(1)}, Q^{(2)}\}(\rho_1(\mathbf{x})\theta^* \mathbb{1}\{x_2 = 0\} + \rho_2(\mathbf{x})\theta^* \mathbb{1}\{x_1 = 0\}) + \\
&\begin{cases}
-\rho_2(\mathbf{x})\mathbb{1}\{x_1 = 0\}\Lambda_{B^{(1)}}(\theta^*) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0, \\
-\rho_1(\mathbf{x})\mathbb{1}\{x_2 = 0\}\Lambda_{B^{(2)}}(\theta^*) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_1, \\
0 & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_2.
\end{cases}
\end{aligned} \tag{5.21}$$

For any $\mathbf{x}$ with $x_2 = \bar{x}_2 = 0$ we have from (5.17)

$$\rho_1(\mathbf{x}) = \frac{e^{-W_1^\delta(\mathbf{x})/\varepsilon}}{\sum_j e^{-W_j^\delta(\mathbf{x})/\varepsilon}} \leq \frac{e^{-(\langle \boldsymbol{\alpha}_1, \mathbf{x}\rangle + \gamma - \delta)/\varepsilon}}{e^{-(\langle \boldsymbol{\alpha}_2, \mathbf{x}\rangle + \gamma - 2\delta)/\varepsilon}} = e^{-(\bar{x}_1\theta^* + \delta)/\varepsilon} \leq e^{-\delta/\varepsilon}, \tag{5.22}$$

and for any $\mathbf{x}$ with $x_1 = \bar{x}_1 = 0$

$$\rho_2(\mathbf{x}) = \frac{e^{-W_2^\delta(\mathbf{x})/\varepsilon}}{\sum_j e^{-W_j^\delta(\mathbf{x})/\varepsilon}} \leq \frac{e^{-(\langle \boldsymbol{\alpha}_2, \mathbf{x}\rangle + \gamma - 2\delta)/\varepsilon}}{e^{-(\langle \boldsymbol{\alpha}_3, \mathbf{x}\rangle + \gamma - 3\delta)/\varepsilon}} = e^{-(\bar{x}_0\theta^* + \delta)/\varepsilon} \leq e^{-\delta/\varepsilon}. \tag{5.23}$$

Substituting (5.22) and (5.23) in (5.21) we find for $\mathbf{x} \neq \mathbf{X}_0$ that

$$\begin{aligned}
&\mathbb{H}(\mathbf{x}, DW^{\varepsilon,\delta}(\mathbf{x})) \\
&\geq -\max\{Q^{(0)}, Q^{(1)}, Q^{(2)}\}(\rho_1(\mathbf{x})\theta^* \mathbb{1}\{x_2 = 0\} + \rho_2(\mathbf{x})\theta^* \mathbb{1}\{x_1 = 0\}) \\
&\quad - \rho_2(\mathbf{x})\mathbb{1}\{x_1 = 0\}\Lambda_{B^{(1)}}(\theta^*)\mathbb{1}\{\mathcal{X}(\mathbf{x}) = \bar{x}_0\} \\
&\quad - \rho_1(\mathbf{x})\mathbb{1}\{x_2 = 0\}\Lambda_{B^{(2)}}(\theta^*)\mathbb{1}\{\mathcal{X}(\mathbf{x}) = \bar{x}_1\} \\
&\geq -e^{-\delta/\varepsilon}\Big(\theta^* \max\{Q^{(0)}, Q^{(1)}, Q^{(2)}\} \\
&\qquad\qquad + \Lambda_{B^{(1)}}(\theta^*)\mathbb{1}\{\mathcal{X}(\mathbf{x}) = \bar{x}_0\} + \Lambda_{B^{(2)}}(\theta^*)\mathbb{1}\{\mathcal{X}(\mathbf{x}) = \bar{x}_1\}\Big) \\
&\geq -e^{-\delta/\varepsilon}\Big(\theta^* \max\{Q^{(0)}, Q^{(1)}, Q^{(2)}\} + \max\left(\Lambda_{B^{(1)}}(\theta^*), \Lambda_{B^{(2)}}(\theta^*)\right)\Big) \\
&= -\Big(\theta^* \max\{Q^{(0)}, Q^{(1)}, Q^{(2)}\} + \gamma\Big)\, e^{-\delta/\varepsilon},
\end{aligned}$$

where the second and the third inequality follow as at most one of the indicators equals 1 at any time during a busy cycle of the system and the final equality follows by definition of θ^*. To conclude the proof we write for the initial state that

$$\begin{aligned}
\mathbb{H}(\mathbf{X}_0, DW^{\varepsilon,\delta}(\mathbf{X}_0)) &= -\Lambda_A(-(\rho_1(\mathbf{X}_0) + \rho_2(\mathbf{X}_0))\theta^*) - \Lambda_{B^{(1)}}(\rho_2(\mathbf{X}_0)\theta^*) \\
&\geq -(\rho_1(\mathbf{X}_0) + \rho_2(\mathbf{X}_0))\Lambda_A(-\theta^*) - \rho_2(\mathbf{X}_0)\Lambda_{B^{(1)}}(\theta^*) \geq 0.
\end{aligned}$$

□

Next, we show that $\sum_{i=0}^{\tau_N - 1}\langle DW^{\varepsilon,\delta}(\mathbf{X}_i), \mathbf{X}_{i+1} - \mathbf{X}_i\rangle$ approximates $W^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - W^{\varepsilon,\delta}(\mathbf{X}_0)$ and provide an upper bound for the error term, similar to Lemma 2 in [14].

Lemma 5.25. *Consider a 2-node $GI|GI|1$ tandem queue satisfying Assumptions 2.3 and 5.1. Then for any successful path $\mathbf{X}_i$, $i = 0, \ldots, \tau_N$, it holds that*

$$|N \sum_{i=0}^{\tau_N - 1}\langle DW^{\varepsilon,\delta}(\mathbf{X}_i), \mathbf{X}_{i+1} - \mathbf{X}_i\rangle - N(W^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - W^{\varepsilon,\delta}(\mathbf{X}_0))| \leq \frac{2C_1^2}{\varepsilon N}\tau_N,$$

where $C_1 = \sqrt{2}(\gamma + \theta^*)(\sqrt{2} + \sqrt{3}\max_j\{Q^{(j)}\}) < \infty$.

Proof. We can bound $|\langle DW^{\varepsilon,\delta}(\mathbf{X}_i), \mathbf{X}_{i+1} - \mathbf{X}_i\rangle - (W^{\varepsilon,\delta}(\mathbf{X}_{i+1}) - W^{\varepsilon,\delta}(\mathbf{X}_i))|$ for each step i by using the mean value theorem. Let $\mathbf{x} = \mathbf{X}_i$ and $\mathbf{y} = \mathbf{X}_{i+1} - \mathbf{X}_i$. By the mean value theorem we have $W^{\varepsilon,\delta}(\mathbf{x} + \mathbf{y}) - W^{\varepsilon,\delta}(\mathbf{x}) = \langle DW^{\varepsilon,\delta}(\mathbf{x} + \eta\mathbf{y}), \mathbf{y}\rangle$ for some $\eta \in [0, 1]$. For convenience, we denote $R_k(\mathbf{x}) = e^{-W_k^\delta(\mathbf{x})/\varepsilon}$, so that

$\rho_k(\mathbf{x}) = \frac{R_k(\mathbf{x})}{\sum_{j=1}^3 R_j(\mathbf{x})}$, and $R_k(\mathbf{x}+\eta\mathbf{y}) = R_k(\mathbf{x})e^{-\langle \boldsymbol{\alpha}_k,\mathbf{y}\rangle \eta/\varepsilon}$, which implies that $DW^{\varepsilon,\delta}(\mathbf{x}+\eta\mathbf{y}) = \sum_k \rho_k(\mathbf{x}+\eta\mathbf{y})\boldsymbol{\alpha}_k = \sum_k \frac{R_k(\mathbf{x}+\eta\mathbf{y})}{\sum_{j=1}^3 R_j(\mathbf{x}+\eta\mathbf{y})}\boldsymbol{\alpha}_k$. Thus,

$$
\begin{aligned}
&|\langle DW^{\varepsilon,\delta}(\mathbf{x}),\mathbf{y}\rangle - \big(W^{\varepsilon,\delta}(\mathbf{x}+\mathbf{y}) - W^{\varepsilon,\delta}(\mathbf{x})\big)| \qquad (5.24)\\
&= \left| \frac{\sum_{k=1}^3 R_k(\mathbf{x})\langle \boldsymbol{\alpha}_k,\mathbf{y}\rangle}{\sum_{j=1}^3 R_j(\mathbf{x})} - \frac{\sum_{k=1}^3 R_k(\mathbf{x}+\eta\mathbf{y})\langle \boldsymbol{\alpha}_k,\mathbf{r}\mathbf{y}\rangle}{\sum_{j=1}^3 R_j(\mathbf{x}+\eta\mathbf{y})} \right| \\
&= \left| \sum_{k=1}^3 \frac{R_k(\mathbf{x})\langle \boldsymbol{\alpha}_k,\mathbf{y}\rangle}{\sum_{j=1}^3 R_j(\mathbf{x})} \left(1 - e^{-\langle \boldsymbol{\alpha}_k,\mathbf{y}\rangle\eta/\varepsilon} \frac{\sum_{j=1}^3 R_j(\mathbf{x})}{\sum_{j=1}^3 R_j(\mathbf{x}+\eta\mathbf{y})}\right)\right| \\
&\le \big|\langle DW^{\varepsilon,\delta}(\mathbf{x}),\mathbf{y}\rangle\big| \cdot \\
&\quad \Big|\max\{1 - e^{(\min_k\langle \boldsymbol{\alpha}_k,\mathbf{y}\rangle - \max_k\langle \boldsymbol{\alpha}_k,\mathbf{y}\rangle)\eta/\varepsilon}, e^{(\max_k\langle \boldsymbol{\alpha}_k,\mathbf{y}\rangle - \min_k\langle \boldsymbol{\alpha}_k,\mathbf{y}\rangle)\eta/\varepsilon} - 1\}\Big|,
\end{aligned}
$$

where the inequality follows by definition of $DW^{\varepsilon,\delta}(\mathbf{x})$, see (5.16), $\rho(\mathbf{x})$ and $R_j(\mathbf{x}+\eta\mathbf{y})$, and by bounding $\langle \boldsymbol{\alpha}_k,\mathbf{y}\rangle$ by the minimum or maximum over k (whichever is appropriate). We use the crude bounds

$$
\begin{aligned}
|DW^{\varepsilon,\delta}(\mathbf{x})| &\le \max_k |\boldsymbol{\alpha}_k| \le \sqrt{2\gamma^2 + 2(\theta^*)^2} \le \sqrt{2}(\gamma+\theta^*),\\
|\mathbf{y}| &\le \frac{1}{N}\max\{\sqrt{1 + 2(Q^{(0)})^2 + (\max\{Q^{(0)},Q^{(1)}\})^2},\\
&\qquad \sqrt{2 + 2(Q^{(1)})^2 + (\max\{Q^{(1)},Q^{(2)}\})^2}, \sqrt{1 + 3(Q^{(2)})^2}\}\\
&\le \frac{1}{N}\sqrt{2 + 3\max_j\{(Q^{(j)})^2\}} \le \frac{\sqrt{2} + \sqrt{3}\max_j\{Q^{(j)}\}}{N},
\end{aligned}
$$

where the second inequality follows by considering all possible transitions. For example, in case of an arrival the number of customers in the system changes by 1, the residual inter-arrival time and the residual service time at queue 2 change by at most $Q^{(0)}$, and the residual service time at queue 1 changes by at most $\max\{Q^{(0)},Q^{(1)}\}$. Letting $C_1 = \sqrt{2}(\gamma+\theta^*)(\sqrt{2}+\sqrt{3}\max_j\{Q^{(j)}\})$, we find $|\langle \boldsymbol{\alpha}_k,\mathbf{y}\rangle| \le C_1$ for all $k = 1,2,3$, and so we can upper bound (5.24) by

$$
\begin{aligned}
\frac{C_1}{N}|\max\{1 - e^{-2\frac{C_1\eta}{\varepsilon N}}, e^{2\frac{C_1\eta}{\varepsilon N}} - 1\}| &= \frac{C_1}{N}\left(e^{2\frac{C_1\eta}{\varepsilon N}} - 1\right)\\
&= \frac{C_1}{N}\left(\frac{C_1\eta}{\varepsilon N} + \mathcal{O}\left(\frac{C_1^2\eta^2}{\varepsilon^2 N^2}\right)\right) \le \frac{2C_1^2}{\varepsilon N^2},
\end{aligned}
$$

where the inequality holds for sufficiently large εN and hence we have found

$$
|N\sum_{i=0}^{\tau_N-1}\langle DW^{\varepsilon,\delta}(\mathbf{X}_i), \mathbf{X}_{i+1} - \mathbf{X}_i\rangle - N(W^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - W^{\varepsilon,\delta}(\mathbf{X}_0))| \le \frac{2C_1^2}{\varepsilon N}\tau_N.
$$

□

Remark 5.26. *The bound in Lemma 5.25 is very crude and better bounds can be obtained by considering all possible transitions separately. However, in order to show asymptotic efficiency the current bound is sufficient.*

Remark 5.27. *Some solutions to* $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ *in Table 5.1 use different values rather than* θ^**, for example* $\widetilde{\theta}^{(1)}$*. It is easy to verify that for these different values, similar bounds as in Lemmas 5.24 and 5.25 can be obtained by using their definition. The only additional requirement is that the value replacing* θ^* *is finite. Note that* θ^* *is finite by Assumption 2.3.*

In order to show asymptotic efficiency, we need an asymptotic result involving τ_N, the total number of steps to reach level N (given that level N is reached before the system is empty). This is the subject of the following conjecture.

Conjecture 5.28. *If* σ_N *is a sequence of real numbers such that* $\sigma_N \to 0$ *when* $N \to \infty$*, then*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\mathbb{Q}} \left[e^{\tau_N \sigma_N} | \tau_N < \infty \right] = 0. \tag{5.25}$$

Intuition. We note that we can upper bound τ_N by $3K_N$, where K_N is the index of the first customer who reaches the overflow level N. Along the major part of the most likely path to reach the overflow level, the change of measure is very close to the state-independent change of measure as discussed in Chapter 4 (see also Examples 5.20 and 5.22). In Lemmas 4.3 and 4.7, it is shown that under this state-independent change of measure the system is instable and $\frac{K_N}{N} \to C < \infty$ with probability 1. This means that the left hand side of (5.25) would be upper bounded by $\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\mathbb{Q}} \left[e^{3K_N \sigma_N} | \tau_N < \infty \right]$, which behaves as $\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\mathbb{Q}} \left[e^{3CN\sigma_N} \right] = \lim_{N\to\infty} 3C\sigma_N = 0$. Also, suppose that (5.25) does not hold, that is, $\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\mathbb{Q}} \left[e^{\tau_N \sigma_N} | \tau_N < \infty \right] > 0$. This means that the random variable τ_N, and also the random variable K_N, would grow much faster than N as $N \to \infty$, which does not seem plausible.

For a mathematical proof of the conjecture, two problems remain. The first one is to show that the difference between the actual change of measure and the state-independent change of measure (which is small along the most likely path) does not influence the validity of (5.25). The second problem is to show that for the state-independent change of measure we indeed have $\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}^{\mathbb{Q}} \left[e^{3K_N \sigma_N} | \tau_N < \infty \right] = 0$. Even if the change of measure would be equal to the state-independent change of measure along the whole state space, we could not show this relation. Of course, there may be other possibilities to show that (5.25) holds. □

We can now prove the main theorem of this section.

Theorem 5.29. *Suppose we have a* 2*-node* $GI|GI|1$ *tandem queue satisfying Assumptions 2.3 and 5.1. Then, under Conjecture 5.28 and Assumption 5.17, the change of measure based on* $DW^{\varepsilon,\delta}(\mathbf{x})$ *in Equation* (5.16) *is asymptotically efficient.*

Proof. For a successful path $\mathcal{P}$ we have

$$\begin{aligned}\log L(\mathcal{P}) &= N\sum_{i=0}^{\tau_N-1}\langle DW^{\varepsilon,\delta}(\mathbf{X}_i),\mathbf{X}_{i+1}-\mathbf{X}_i\rangle - \sum_{i=0}^{\tau_N-1}\mathbb{H}(\mathbf{X}_i,DW^{\varepsilon,\delta}(\mathbf{X}_i))\\ &\leq \frac{2C_1^2}{\varepsilon N}\tau_N + NW^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - NW^{\varepsilon,\delta}(\mathbf{X}_0)\\ &\quad + (\theta^*\max\{Q^{(0)},Q^{(1)},Q^{(2)}\}+\gamma)e^{-\delta/\varepsilon}\tau_N,\end{aligned}$$

where the second step follows from Lemma 5.24 and 5.25. Define $\sigma_N = \frac{2C_1^2}{\varepsilon N} + (\theta^*\max\{Q^{(0)},Q^{(1)},Q^{(2)}\}+\gamma)e^{-\delta/\varepsilon}$, then it follows that

$$\log L(\mathcal{P}) \leq \tau_N\sigma_N + NW^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - NW^{\varepsilon,\delta}(\mathbf{X}_0).$$

Since $\mathbf{X}_{\tau_N}\in\delta_e$ for the successful path $\mathcal{P}$, we find, by using Lemma 5.18,

$$\begin{aligned}\log L(\mathcal{P}) &\leq \tau_N\sigma_N + N(\frac{\gamma}{N}-\delta) - N\left(\gamma(1-\frac{1}{N})-3\delta-\varepsilon\log(3)\right)\\ &= \tau_N\sigma_N + 2N\delta - \gamma(N-2) + \varepsilon N\log(3),\end{aligned}$$

and so we have

$$\begin{aligned}&\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right]\\ &\leq \limsup_{N\to\infty}\frac{1}{N}\log\mathbb{E}^{\mathbb{Q}}\left[e^{2\tau_N\sigma_N+4N\delta-2\gamma(N-2)+2\varepsilon N\log(3)}I(\mathcal{P})\right]\\ &= -2\gamma + \limsup_{N\to\infty}\frac{1}{N}\log\left(\mathbb{E}^{\mathbb{Q}}\left[e^{2\tau_N\sigma_N}|I(\mathcal{P})=1\right]\mathbb{P}^{\mathbb{Q}}(I(\mathcal{P})=1)\right)\\ &\leq -2\gamma + \limsup_{N\to\infty}\frac{1}{N}\log\mathbb{E}^{\mathbb{Q}}\left[e^{2\tau_N\sigma_N}|\tau_N<\infty\right],\end{aligned}$$

where the equality follows by Assumption 5.17 and the last step follows by noting that $\mathbb{P}^{\mathbb{Q}}(I(\mathcal{P})=1)\leq 1$. Using Conjecture 5.28 concludes the proof. □

Remark 5.30. *In the proof of Theorem 5.29 we bounded* $N\sum_{i=0}^{\tau_N-1}\langle DW^{\varepsilon,\delta}(\mathbf{X}_i),\mathbf{X}_{i+1}-\mathbf{X}_i\rangle$ *and* $\sum_{i=0}^{\tau_N-1}\mathbb{H}(\mathbf{X}_i,DW^{\varepsilon,\delta}(\mathbf{X}_i))$ *separately, resulting in bounds involving the maximum support of the various distributions. Looking at Equation* (5.18) *and* (5.19) *we see for example that the new probability density functions themselves do not depend on* $\mathcal{X}(\mathbf{X}_i)N$*, and thus the likelihood ratio in* (5.7) *does not depend on* $\mathcal{X}(\mathbf{X}_i)N$*, even though* $\mathcal{X}(\mathbf{X}_i)$ *is needed in order to determine which probability density function is used. However, to use the mean value theorem in Lemma 5.25 we do need to keep this quantity throughout the proofs.*

We can extend Theorem 5.29 to the following set of changes of measure, see also Remarks 5.16 and 5.27. We restrict ourselves to three regions only, since this is

the easiest from an implementation perspective. Let

$$DW^{\varepsilon,\delta}(\mathbf{x}) = (-(\rho_1(\mathbf{x}) + \rho_2(\mathbf{x}))\gamma, -\rho_1(\mathbf{x})\gamma, (\rho_1(\mathbf{x}) + \rho_2(\mathbf{x}))\theta^*, \\ - \rho_1(\mathbf{x})\bar{\alpha}_{1,1} - \rho_2(\mathbf{x})\bar{\alpha}_{1,2}, -\rho_1(\mathbf{x})\bar{\alpha}_{2,1} - \rho_2(\mathbf{x})\bar{\alpha}_{2,2}), \tag{5.26}$$

where

$$\begin{aligned}
\bar{\alpha}_{1,1} &\in [0, -\bar{\alpha}_{2,1} - \theta^*], \\
\bar{\alpha}_{2,1} &\in [-\widetilde{\theta}^{(2)}, -\theta^*], \\
\bar{\alpha}_{1,2} &\in [-\widetilde{\theta}^{(1)}, -\theta^*], \\
\bar{\alpha}_{2,2} &\in [0, -\bar{\alpha}_{1,2} - \theta^*].
\end{aligned}$$

Note that when queue j is the bottleneck queue, $\widetilde{\theta}^{(j)} = \theta^*$ and hence along the most likely path still the state-independent change of measure studied in Chapter 4 and [25, 35] is used.

We leave the proof of the following theorem to the reader, as this only requires some small modifications in the proofs of Lemmas 5.24, 5.25 and Theorem 5.29.

Theorem 5.31. *Suppose we have a 2-node $GI|GI|1$ tandem queue satisfying Assumptions 2.3 and 5.1. Then, under Conjecture 5.28 and Assumption 5.17, the change of measure based on $DW^{\varepsilon,\delta}(\mathbf{x})$ in Equation (5.26) is asymptotically efficient.*

5.2.3 The d-node $GI|GI|1$ tandem queue

In this section we present the steps to follow in order to find a state-dependent change of measure for the d-node $GI|GI|1$ tandem queue. As we needed a conjecture for the case $d = 2$, we also need a conjecture for the more general case. We will formulate similar lemmas as in Section 5.2.2 but omit the proofs, as these are just extensions of the proofs for the 2-node tandem queue and therefore do not require any additional techniques.

5.2.3.1 Solutions to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq \mathbf{0}$ for all possible $\mathbf{x}$

For the d-node tandem queue we have, for $\mathbf{x} \neq \mathbf{X}_0$,

$$\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) = -\mathcal{X}(\mathbf{x})N\left(\bar{\alpha}_0 + \sum_{k=1}^{d} \bar{\alpha}_k \mathbb{1}\{x_k > 0\}\right) + \\ \begin{cases} \alpha_1 - \Lambda_A(-\bar{\alpha}_0) - \Lambda_{B^{(1)}}(-\bar{\alpha}_1 \mathbb{1}\{x_1 = 0\}) & \\ \qquad\qquad \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0, & \\ \alpha_{j+1} - \alpha_j - \Lambda_{B^{(j)}}(-\bar{\alpha}_j \mathbb{1}\{x_j > \frac{1}{N}\}) - \Lambda_{B^{(j+1)}}(-\bar{\alpha}_{j+1}\mathbb{1}\{x_{j+1} = 0\}) & \\ \qquad\qquad \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_j,\ j = 1, ..., d-1, & \\ -\alpha_d - \Lambda_{B^{(d)}}(-\bar{\alpha}_d \mathbb{1}\{x_d > \frac{1}{N}\}) \qquad \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_d, & \end{cases}$$

and for $\mathbf{x} = \mathbf{X}_0$

$$H(\mathbf{X}_0, \boldsymbol{\alpha}) = -\Lambda_A(-\bar{\alpha}_0) - \Lambda_{B^{(1)}}(-\bar{\alpha}_1).$$

Some possible solutions to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$, similar to the solution presented in Table 5.2, can be found in Table 5.3 (including the trivial solution $\boldsymbol{\alpha} = \mathbf{0}$). Similarly to Table 5.2, the solutions presented consider cases where some $x_j > 0$. Due to the construction that is used, using the minimum of all subsolutions, these solutions also cover the interior of the state space.

Table 5.3 Some solutions to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ for all $\mathbf{x}$. All α_j's and $\bar{\alpha}_j$'s that are not explicitly mentioned in this table are equal to 0 in that particular case.

$x_1 > 0$	$\alpha_1 = \Lambda_A(-\theta^*)$	$\bar{\alpha}_0 = -\bar{\alpha}_1 = \theta^*$
$\vdots$	$\vdots$	$\vdots$
$x_j > 0$	$\alpha_1 = \ldots = \alpha_j = \Lambda_A(-\theta^*)$	$\bar{\alpha}_0 = -\bar{\alpha}_j = \theta^*$
$\vdots$	$\vdots$	$\vdots$
$x_d > 0$	$\alpha_1 = \ldots = \alpha_d = \Lambda_A(-\theta^*)$	$\bar{\alpha}_0 = -\bar{\alpha}_d = \theta^*$
All $\mathbf{x}$		

Clearly, solutions similar to the other solutions presented in Table 5.1 also exist. For example, define $\widetilde{\theta}^{(1)} = \sup\{\theta : \Lambda_A(-\theta^*) + \Lambda_{B^{(1)}}(\theta) \leq 0\}$. When $x_1 > 0$, another possibility would be $\bar{\alpha}_1 \in [-\widetilde{\theta}^{(1)}, -\theta^*]$ and $\bar{\alpha}_j \in [0, -\sum_{i \neq j} \bar{\alpha}_i]$, $j > 1$. Thus, when $x_j > 0$, we define $\widetilde{\theta}^{(j)} = \sup\{\theta : \Lambda_A(-\theta^*) + \Lambda_{B^{(j)}}(\theta) \leq 0\}$ and another possibility would be $\bar{\alpha}_j \in [-\widetilde{\theta}^{(j)}, -\theta^*]$ and for $k > 0$, $\bar{\alpha}_k \in [0, -\sum_{i \neq k} \bar{\alpha}_i]$, $k \neq j$.

5.2.3.2 Construction of $W(\mathbf{x})$

As for the 2-node tandem queue, there is no solution for $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$ that holds for all $\mathbf{x}$. Therefore, a different change of measure needs to be applied in different 'regions'. In this case, the following vectors are used to specify the different changes of measure in each of the 'regions':

$$\begin{aligned}
\boldsymbol{\alpha}_1 &= -\gamma(\mathbf{e}_1 + \ldots + \mathbf{e}_d) + \theta^*(\bar{\mathbf{e}}_0 - \bar{\mathbf{e}}_d), \\
&\vdots \\
\boldsymbol{\alpha}_j &= -\gamma(\mathbf{e}_1 + \ldots + \mathbf{e}_{d-(j-1)}) + \theta^*(\bar{\mathbf{e}}_0 - \bar{\mathbf{e}}_{d-(j-1)}), \\
&\vdots \\
\boldsymbol{\alpha}_d &= -\gamma\mathbf{e}_1 + \theta^*(\bar{\mathbf{e}}_0 - \bar{\mathbf{e}}_1), \\
\boldsymbol{\alpha}_{d+1} &= (0, \ldots, 0).
\end{aligned} \tag{5.27}$$

From this we find $d+1$ functions $W_k^\delta(\mathbf{x}) = \langle \boldsymbol{\alpha}_k, \mathbf{x} \rangle + \gamma - k\delta$, for $k = 1, \ldots, d+1$ for $\delta > 0$ and their minimum

$$W^\delta(\mathbf{x}) = W_1^\delta(\mathbf{x}) \wedge \ldots \wedge W_{d+1}^\delta(\mathbf{x}),$$

which again is a piecewise affine function that equals $W_k^\delta(\mathbf{x})$ in 'region' k. The constant γ in $W_k^\delta(\mathbf{x})$ is included to satisfy the properties of the subsolution, and the subtraction of $k\delta$ is needed such that the minimum function $W^\delta(\mathbf{x})$ is uniquely attained at each of the boundaries.
To make the minimum function a continuously differentiable function, we apply a similar mollification procedure as for the 2-node $GI|GI|1$ tandem queue. This mollification 'removes' the regions and ensures that a gradient exists throughout the whole parameter space. We let

$$W^{\varepsilon,\delta}(\mathbf{x}) = -\varepsilon \log \sum_{k=1}^{d+1} e^{-W_k^\delta(\mathbf{x})/\varepsilon}. \tag{5.28}$$

When $\varepsilon \to 0$, $W^{\varepsilon,\delta}(\mathbf{x}) \to W_1^\delta(\mathbf{x}) \wedge \ldots \wedge W_{d+1}^\delta(\mathbf{x})$. The assumptions on ε and δ can be found in Assumption 5.17. Similar to Lemma 5.18, we see that $W^{\varepsilon,\delta}(\mathbf{x})$ satisfies properties 3 and 4 in and below Definition 5.4.

Lemma 5.32. *Under Assumption 2.3, we have for the d-node $GI|GI|1$ tandem queue that $W^{\varepsilon,\delta}(\mathbf{x}) \leq \frac{\gamma}{N} - \delta$ for $\mathbf{x} \in \delta_e$ and $\gamma(1 - \frac{1}{N}) - (d+1)\delta - \varepsilon \log(d+1) \leq W^{\varepsilon,\delta}(\mathbf{X}_0) \leq \gamma - (d+1)\delta$.*

Proof. The proof is omitted, since it is similar to the proof of Lemma 5.18. □

The change of measure will be based on the gradient of $W^{\varepsilon,\delta}(\mathbf{x})$. From (5.28) it follows that

$$DW^{\varepsilon,\delta}(\mathbf{x}) = \sum_{k=1}^{d+1} \rho_k(\mathbf{x}) \boldsymbol{\alpha}_k, \tag{5.29}$$

where $\rho_k(\mathbf{x})$ is defined similar as in (5.17).

5.2.3.3 The change of measure

As for the 2-node tandem queue, we show that the change of measure based on $W_k^\delta(\mathbf{x})$ results in the state-independent change of measure studied in Chapter 4 and [25, 35] in some particular cases. This gives some insight in the change of measure that is applied in the mollified function $W^{\varepsilon,\delta}(\mathbf{x})$.

Choosing $DW(\mathbf{x})$ in (5.6) as a constant vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d, \bar{\alpha}_0, \ldots, \bar{\alpha}_d)$,

the change of measure can be written in terms of $\bar{\alpha}_0, \ldots, \bar{\alpha}_d$ as

$$d\bar{F}_{V_X(\mathbf{x})}(\mathbf{v}|\mathbf{x}) =$$
$$\begin{cases} \frac{e^{-a\bar{\alpha}_0 - b_1\bar{\alpha}_1 \mathbb{1}\{x_1=0\}}}{M_A(-\bar{\alpha}_0)M_{B^{(1)}}(-\bar{\alpha}_1 \mathbb{1}\{x_1=0\})} dF_A(a)dF_{B^{(1)}}(b_1) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_0, \\ \frac{e^{-b_j\bar{\alpha}_1 \mathbb{1}\{x_j > \frac{1}{N}\} - b_{j+1}\bar{\alpha}_{j+1}\mathbb{1}\{x_{j+1}=0\}}}{M_{B^{(j)}}(-\bar{\alpha}_j \mathbb{1}\{x_j > \frac{1}{N}\})M_{B^{(j+1)}}(-\bar{\alpha}_{j+1}\mathbb{1}\{x_{j+1}=0\})} dF_{B^{(j)}}(b_j)dF_{B^{(j+1)}}(b_{j+1}) & \\ & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_j, j = 1, \ldots d-1, \\ \frac{e^{-b_d\bar{\alpha}_d \mathbb{1}\{x_d > \frac{1}{N}\}}}{M_{B^{(d)}}(-\bar{\alpha}_d \mathbb{1}\{x_d > \frac{1}{N}\})} dF_{B^{(d)}}(b_d) & \text{if } \mathcal{X}(\mathbf{x}) = \bar{x}_d, \end{cases}$$

for $\mathbf{x} \neq \mathbf{X}_0$ and

$$d\bar{F}_{V_X(\mathbf{X}_0)}(\mathbf{v}|\mathbf{X}_0) = \frac{e^{-a\bar{\alpha}_0 - b_1\bar{\alpha}_1}}{M_A(-\bar{\alpha}_0)M_{B^{(1)}}(-\bar{\alpha}_1)} dF_A(a)dF_{B^{(1)}}(b_1).$$

We now consider the cases where our change of measure based on $W^{\varepsilon,\delta}(\mathbf{x})$ gets close, as $N \to \infty$, to an 'affine' change of measure as above, that is, based on some $W_k^\delta(\mathbf{x})$ with constant gradient. Then, if we let $\boldsymbol{\alpha} = DW^\delta_{d-(j-1)}(\mathbf{x}) = \boldsymbol{\alpha}_{d-(j-1)}$, see (5.27), it follows that indeed the resulting change of measure equals the state-independent change of measure studied in Chapter 4 and [25, 35] *if* queue j is the bottleneck queue.

5.2.3.4 Asymptotic Efficiency

To show that asymptotic efficiency also holds for the d-node $GI|GI|1$ tandem queue, assuming that Conjecture 5.28 is correct, we need the following two lemmas. Recall that the constants $Q^{(j)} < \infty \ \forall j = 0, ..., d$ are such that $\mathbb{P}(A_k < Q^{(0)}) = 1$, $\mathbb{P}(B_k^{(j)} < Q^{(j)}) = 1$.

Lemma 5.33. *Under Assumptions 2.3 and 5.1, we have for the d-node $GI|GI|1$ tandem queue for all possible* $\mathbf{x}$

$$\mathbb{H}(\mathbf{x}, DW^{\varepsilon,\delta}(\mathbf{x})) \geq -(\theta^* \max\{Q^{(0)}, \ldots, Q^{(d)}\} + \gamma)e^{-\delta/\varepsilon}.$$

Proof. The proof is omitted, since it is similar to the proof of Lemma 5.24. □

Lemma 5.34. *Suppose we have a d-node $GI|GI|1$ tandem queue satisfying Assumptions 2.3 and 5.1. Then for any successful path $\mathbf{X}_i$, $i = 0, \ldots, \tau_N$, it holds for all $\mathbf{X}_{\tau_N}$ that*

$$|N \sum_{i=0}^{\tau_N - 1} \langle DW^{\varepsilon,\delta}(\mathbf{X}_i), \mathbf{X}_{i+1} - \mathbf{X}_i \rangle - N(W^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - W^{\varepsilon,\delta}(\mathbf{X}_0))| \leq \frac{2C_2^2}{\varepsilon N}\tau_N,$$

where $C_2 = \sqrt{2}(\gamma + \theta^*)(\sqrt{2} + \sqrt{3}\max_j\{Q^{(j)}\}) < \infty$.

Proof. The proof is omitted, since it is similar to the proof of Lemma 5.25. □

Using Lemma 5.32–5.34, we can prove the following result.

Theorem 5.35. *Suppose we have a d-node $GI|GI|1$ tandem queue satisfying Assumptions 2.3 and 5.1. Then, under Conjecture 5.28 and Assumption 5.17, the change of measure based on $DW^{\varepsilon,\delta}(\mathbf{x})$ in Equation* (5.29) *is asymptotically efficient.*

Proof. The proof is omitted, since it is similar to the proof of Theorem 5.29. □

Remark 5.36. *We could also extend Theorem 5.35, as we did for the 2-node tandem queue in Theorem 5.31, and we claim that asymptotic efficiency also holds in that case under the same conditions as in Theorem 5.35.*

5.3 Numerical results for the 2-node tandem queue

In this section we present numerical results for the 2-node tandem queue to illustrate that the proposed method indeed works. In each of the examples below, we use (5.16) for the change of measure. We recall that the estimator for p_N is $L(\mathcal{P})I(\mathcal{P})$ and hence the estimator obtained via simulation is (1.3).

In all tables shown below, RE is the relative error, that is, the standard deviation of the estimator divided by its mean, see also (4.29), and AE is defined in (4.30). Thus, if the change of measure is asymptotically efficient, AE should converge to 2 when N tends to infinity, see Definition 5.7. Furthermore, in our tables we will include the number of times the overflow level has been reached (out of a total of $S = 10^6$ simulation runs) and the (rounded) simulation time in seconds. Note that the latter quantity is only there for reference and does not indicate if the estimator is asymptotically efficient or not.

To use the change of measure based on (5.16), we need to choose ε and δ so that they satisfy Assumption 5.17. It is not trivial to find suitable ε and δ that satisfy all requirements; we only explored several possibilities, while many more exist. In all the examples below, we set ε proportional to $\frac{1}{\sqrt{N}}$ and $\delta = -\varepsilon \log \varepsilon$, unless the condition Remark 5.37 below is not satisfied.

Remark 5.37. *For the choice $\delta = -\varepsilon \log \varepsilon$, it may happen for small values of N that $W^{\delta}(\mathbf{x})$ does not attain $W_k^{\delta}(\mathbf{x})$ for some k in the 'region' where it is supposed to be attained. For example, consider $W_2^{\delta}(\mathbf{x}) < W_1^{\delta}(\mathbf{x})$, which is equivalent to*

$$x_2 < \frac{\theta^*(\bar{x}_1 - \bar{x}_2)}{\gamma} + \frac{\delta}{\gamma}.$$

Taking into account that $W_2^{\delta}(\mathbf{x})$ is designed for the case $x_2 = 0$, the right-hand side of this equation should be positive. Thus, we need $\delta > \frac{Q^{(2)}}{N}\theta^*$ (recall that $Q^{(2)}$ is an upper bound on the support of the service times at queue 2). When we also

consider $W_1^\delta(\mathbf{x}) < W_3^\delta(\mathbf{x})$ *and* $W_2^\delta(\mathbf{x}) < W_3^\delta(\mathbf{x})$*, it turns out that we actually need*

$$\delta > \frac{\max\{Q^{(1)}, Q^{(2)}\}\theta^*}{N}. \tag{5.30}$$

Therefore, when $\delta = -\varepsilon \log \varepsilon$ *does not satisfy* (5.30)*, in our numerical experiments, we actually set* $\delta = \frac{\max\{Q^{(1)},Q^{(2)}\}\theta^*}{N^{0.95}}$*. We choose the denominator to be* $N^{0.95}$*, but of course other choices are possible to obtain the strict inequality in* (5.30)*. Unless mentioned otherwise, only for* $N = 20$ *this different value for* δ *is used. (Note that for a d-node tandem queue, a similar condition exists;* $\delta > \frac{\max\{Q^{(1)},\ldots,Q^{(d)}\}\theta^*}{N}$*.)*

We consider two types of tandem queues, a $D|U|1 - \cdot|U|1$ tandem queue and a $M|U|1 - \cdot|U|1$ tandem queue, and in both cases we vary the bottleneck queue. Starting with a $D|U|1 - \cdot|U|1$ tandem queue, we first consider the case where the θ-bottleneck queue is not unique, that is, the service times at both queues have the same distribution with the same parameters. In our case, both service times have a uniform distribution on the interval $[0, 2]$. Similar cases are known to fail when using a state-independent change of measure, see Chapter 4 and [12], and thus this is an interesting case to consider. The results can be found in Table 5.4. In this table, we see results that clearly support the theoretical results of the estimator being asymptotically efficient; at first the relative error is decreasing, after which it slowly increases. We also see that the number of times the overflow level N is reached, is decreasing with N. This behavior may seem strange, but it can be seen in all of the results in this section. Given our assumptions on ε and δ, we note that $\delta N \to \infty$ and hence 'region' 3 – where no change of measure is applied – is increasing in size. Therefore, we can expect that as N increases it becomes increasingly more difficult to reach the overflow level, even though we have asymptotic efficiency. In particular, we observed that when a system has small server utilizations, it may be hard to escape the 'region' of the state space where (almost) no change of measure is applied.

Next, we let one of the queues be the bottleneck queue. In Tables 5.5 and 5.6, queue 1 and queue 2 are the bottleneck queue respectively. Here we changed one of the service time distributions of the previous example from $U[0, 2]$ to $U[0.5, 1.5]$ so that there is a unique θ-bottleneck queue, but the server utilizations of both queues remain the same. When queue 1 is the bottleneck queue, we see that the relative error is still very large for $N = 20$, but for larger values of N the theoretical results are supported. Note that in this case $\delta = -\varepsilon \log \varepsilon$ for all values of N. We also see from this table that the simulation time is increasing, even though the number of times the overflow level is reached is decreasing. This is due to the fact that we have to reach a higher overflow level N. When queue 2 is the bottleneck queue, we have considered a much smaller value for ε, and thus δ differs from being $-\varepsilon \log \varepsilon$ for $N = 20, \ldots, 260$. This can also be seen from the table, by noting that both the number of times the overflow is reached and the

simulation time increase at $N = 300$. Besides this difference, the results in the table clearly support the theoretical results.

Lastly, we consider a $M|U|1 - \cdot|U|1$ tandem queue in Tables 5.7-5.9. We emphasize that the inter-arrival times are exponentially distributed, which means an unbounded support, and hence this is not covered by our theoretical results. Again, we start with the case where there is no unique bottleneck queue in Table 5.7. We see from this table that the relative error remains roughly constant, suggesting asymptotic efficiency. In Table 5.8, when queue 1 is the bottleneck queue, and in Table 5.9, when queue 2 is the bottleneck queue, we see that the relative error is slightly increasing, but still suggesting asymptotic efficiency. Remark that in Table 5.9, δ does not equal $-\varepsilon \log \varepsilon$ for $N = 20, \ldots, 100$, which can be seen by an increasing number of times level N is reached for $N = 140$.

Even though we do not have a proof of asymptotic efficiency for unbounded supports, which is the case for the inter-arrival times in Tables 5.7-5.9, all these tables show good results that suggest asymptotic efficiency of the estimator. We also did some experiments with exponentially distributed service times, where regardless of (5.30) in Remark 5.37 we set $\delta = -\varepsilon \log \varepsilon$, but we did not obtain good results there. This is most likely due to the fact that for Markovian service times (and unbounded service times in general), the condition on δ, see (5.30), cannot hold. As a result, it could happen that, for example, $W_2^\delta(\mathbf{x}) > W_1^\delta(\mathbf{x})$ when $x_2 = 0$, see also Remark 5.37 above, leading for $x_2 = 0$ to $\mathbb{H}(\mathbf{x}, DW^{\varepsilon,\delta}(\mathbf{x})) \approx \mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_1)$, which is negative for $x_2 = 0$ (and hence violates 2 in Definition 5.4).

Table 5.4 $D|U|1-\cdot|U|1$ tandem queue when both queues are the bottleneck queue. We choose $A = 1.1$, $B^{(1)} \sim U[0,2]$, $B^{(2)} \sim U[0,2]$ and $\varepsilon = \frac{0.06}{\sqrt{N}}$. We find $\theta^* = \theta^{(1)} = \theta^{(2)} = 0.6073$. The number of simulations is 10^6.

N	$\hat{p}_N$	RE	AE	#Overflow	Time (s)
20	1.3448e-03	0.0877	0.6465	680910	93
60	1.7785e-14	0.0536	1.7485	580520	307
100	8.7767e-26	0.0290	1.8833	455950	451
140	3.2087e-37	0.0112	1.9424	339723	520
180	1.0567e-48	0.0048	1.9711	242793	533
220	3.1997e-60	0.0037	1.9803	169426	512
260	9.4889e-72	0.0038	1.9833	118241	478
300	2.7188e-83	0.0044	1.9843	81788	440
340	7.6193e-95	0.0053	1.9844	57009	405
380	2.1168e-106	0.0067	1.9843	40519	376
420	5.9068e-118	0.0084	1.9842	29311	354
460	1.5638e-129	0.0110	1.9838	20811	333
500	4.2998e-141	0.0140	1.9836	15300	317

Table 5.5 $D|U|1-\cdot|U|1$ tandem queue when queue 1 is the bottleneck queue. We choose $A = 1.1$, $B^{(1)} \sim U[0,2]$, $B^{(2)} \sim U[0.5,1.5]$ and $\varepsilon = \frac{0.07}{\sqrt{N}}$. We find $\theta^* = \theta^{(1)} = 0.6073$, $\theta^{(2)} = 2.5215$. The number of simulations is 10^6.

N	$\hat{p}_N$	RE	AE	#Overflow	Time (s)
20	7.1456e-04	0.4771	0.2971	854566	838
60	2.2719e-15	0.0852	1.7363	802889	1761
100	1.0067e-26	0.0977	1.8469	764219	1781
140	3.4432e-38	0.0811	1.8981	708956	1952
180	9.0425e-50	0.0322	1.9385	632963	2235
220	2.2366e-61	0.0088	1.9688	540583	2566
260	5.6593e-73	0.0111	1.9711	441182	2838
300	1.3905e-84	0.0043	1.9847	346529	3059
340	3.4348e-96	0.0046	1.9859	262769	3219
380	8.6114e-108	0.0052	1.9865	195534	3316
420	2.1157e-119	0.0056	1.9873	142657	3382
460	5.3385e-131	0.0128	1.9830	103512	3408
500	1.3122e-142	0.0086	1.9868	75186	3374

Table 5.6 $D|U|1-\cdot|U|1$ tandem queue when queue 2 is the bottleneck queue. We choose $A = 1.1$, $B^{(1)} \sim U[0.5, 1.5]$, $B^{(2)} \sim U[0, 2]$ and $\varepsilon = \frac{0.01}{\sqrt{N}}$. We find $\theta^* = \theta^{(2)} = 0.6073$, $\theta^{(1)} = 2.5215$. The number of simulations is 10^6.

N	$\hat{p}_N$	RE	AE	#Overflow	Time (s)
20	7.8819e-05	0.0125	1.4655	478564	87
60	2.0245e-16	0.0138	1.8546	443297	245
100	5.0153e-28	0.0102	1.9250	426668	390
140	1.2576e-39	0.0117	1.9449	416708	550
180	3.1214e-51	0.0106	1.9593	408285	679
220	7.8784e-63	0.0129	1.9642	403760	808
260	1.9570e-74	0.0092	1.9738	398371	931
300	4.7882e-86	0.0237	1.9678	548146	1433
340	1.2159e-97	0.0227	1.9720	505585	1495
380	2.9018e-109	0.0074	1.9840	462204	1526
420	7.4330e-121	0.0107	1.9828	422362	1541
460	1.8204e-132	0.0061	1.9880	385678	1540
500	4.6143e-144	0.0107	1.9856	350976	1525

Table 5.7 $M|U|1-\cdot|U|1$ tandem queue when both queues are the bottleneck queue. We choose $A \sim \exp(0.5)$, $B^{(1)} \sim U[0, 3]$, $B^{(2)} \sim U[0, 3]$ and $\varepsilon = \frac{0.025}{\sqrt{N}}$. We find $\theta^* = \theta^{(1)} = \theta^{(2)} = 0.2690$. The number of simulations is 10^6.

N	$\hat{p}_N$	RE	AE	#Overflow	Time (s)
20	1.9166e-03	0.0065	1.3957	362952	16
60	1.8257e-10	0.0087	1.8067	322186	53
100	1.0024e-17	0.0088	1.8885	231546	68
140	4.5750e-25	0.0082	1.9247	175071	74
180	1.9576e-32	0.0096	1.9379	137692	76
220	7.9868e-40	0.0080	1.9537	109868	76
260	3.1289e-47	0.0080	1.9611	89605	74
300	1.2203e-54	0.0106	1.9620	74172	71
340	4.5110e-62	0.0080	1.9704	61372	68
380	1.6843e-69	0.0083	1.9731	51929	65
420	6.0889e-77	0.0081	1.9761	43639	62
460	2.2324e-84	0.0083	1.9779	37347	59
500	8.1358e-92	0.0087	1.9793	32318	56

Table 5.8 $M|U|1-\cdot|U|1$ tandem queue when queue 1 is the bottleneck queue. We choose $A \sim \exp(0.5)$, $B^{(1)} \sim U[0,3]$, $B^{(2)} \sim U[0,2]$ and $\varepsilon = \frac{0.025}{\sqrt{N}}$. We find $\theta^* = \theta^{(1)} = 0.2690$, $\theta^{(2)} = 0.8966$. The number of simulations is 10^6.

N	$\hat{p}_N$	RE	AE	#Overflow	Time (s)
20	3.3042e-04	0.0024	1.7592	249800	13
60	1.1016e-11	0.0027	1.9160	217878	39
100	3.6616e-19	0.0032	1.9435	150647	46
140	1.2177e-26	0.0037	1.9550	111215	48
180	4.0285e-34	0.0041	1.9628	84826	48
220	1.3518e-41	0.0046	1.9671	67284	47
260	4.5013e-49	0.0050	1.9706	54098	45
300	1.4843e-56	0.0055	1.9731	43870	43
340	4.9227e-64	0.0060	1.9751	36140	40
380	1.6372e-71	0.0069	1.9762	30178	38
420	5.4779e-79	0.0072	1.9780	25657	36
460	1.8101e-86	0.0077	1.9793	21731	34
500	6.0619e-94	0.0084	1.9802	18566	32

Table 5.9 $M|U|1-\cdot|U|1$ tandem queue when queue 2 is the bottleneck queue. We choose $A \sim \exp(0.5)$, $B^{(1)} \sim U[0,2]$, $B^{(2)} \sim U[0,3]$ and $\varepsilon = \frac{0.01}{\sqrt{N}}$. We find $\theta^* = \theta^{(2)} = 0.2690$, $\theta^{(1)} = 0.8966$. The number of simulations is 10^6.

N	$\hat{p}_N$	RE	AE	#Overflow	Time (s)
20	3.2750e-04	0.0038	1.6617	157082	9
60	1.0824e-11	0.0039	1.8891	144253	27
100	3.6318e-19	0.0039	1.9349	140723	43
140	1.2032e-26	0.0026	1.9656	199660	84
180	4.0069e-34	0.0032	1.9688	159602	88
220	1.3319e-41	0.0036	1.9723	129196	87
260	4.4494e-49	0.0041	1.9742	106298	86
300	1.4765e-56	0.0044	1.9767	88192	84
340	4.8887e-64	0.0047	1.9784	73833	82
380	1.6223e-71	0.0053	1.9792	62571	77
420	5.5025e-79	0.0059	1.9801	53472	73
460	1.8175e-86	0.0064	1.9811	45541	68
500	5.9896e-94	0.0066	1.9823	39072	64

CHAPTER 6

State-dependent importance sampling for Markovian tandem queues: exploring the possibilities

In this chapter, we consider importance sampling for the 2-node $M|M|1$ tandem queue with the same probability of interest as in Chapters 2, 4 and 5. Even though this problem has been studied before in for example [12, 14, 19, 22, 30, 35] and asymptotic efficiency of changes of measure has been shown in [14, 19, 22], still there are 'only' three changes of measure known that have been proven to be asymptotically efficient for the 2-node $M|M|1$ tandem queue: two when queue 2 is the bottleneck queue (that is, $\mu_2 \leq \mu_1$, where μ_i is the service rate of queue i), and one when queue 1 is the bottleneck queue ($\mu_1 \leq \mu_2$), see [19, 22].

In Chapter 5, we have shown that there are several possible changes of measure yielding an asymptotically efficient estimator for $GI|GI|1$ tandem queues with bounded support, see Theorem 5.31. Since the exponential distribution clearly does not have bounded support, this raises some thoughts on whether there exist more asymptotically efficient changes of measure for the 2-node $M|M|1$ tandem queue.

In this chapter, we give sufficient conditions for an asymptotically efficient change of measure for the $M|M|1$ tandem queue. These conditions follow naturally when using the same method as in [19, 22], but have never been stated as such. We show how to find changes of measure satisfying these conditions, without showing any numerical experiments (since we are only interested in the possibilities), and we believe the same could be done for other models.

In addition, in Chapter 5, we have considered both queue 1 and queue 2 being the bottleneck queue (for a 2-node $M|M|1$ tandem queue this is equivalent to $\mu_1 \leq \mu_2$ and $\mu_2 \leq \mu_1$ respectively). This is in slight contrast with most existing literature on importance sampling for 2-node *Markovian* tandem queues, see [12, 14, 19, 30], where only the second queue being the bottleneck queue has been discussed. Even though this does not affect the probability of interest, since both queues are interchangeable [41], it is particularly interesting to see the possibilities for a change of measure when $\mu_1 \leq \mu_2$. Therefore, we consider both $\mu_1 \leq \mu_2$ *and* $\mu_2 \leq \mu_1$.

A notable difference of this chapter compared to Chapter 5 is that in the latter

chapter we considered a system in continuous time, thereby adding residual times to the state description, while in this chapter we consider the embedded discrete time Markov chain of the $M|M|1$ tandem queue. Since the memoryless property holds for $M|M|1$ tandem queues, we do not need the residual times.

The contributions of this chapter are two-fold. After summarizing the subsolution method for importance sampling, partly referring to Chapter 5, and stating the results from [19, 22] in Section 6.1.3, our first contribution is in Section 6.2 where we state conditions for a change of measure for the d-node $M|M|1$ tandem queue based on subsolutions to give an asymptotically efficient estimator, and we prove that these conditions are sufficient for $d = 2$. The other contribution, in Section 6.3, is that we provide a whole family of changes of measure for the 2-node $M|M|1$ tandem queue that satisfy these conditions and hence result in an asymptotically efficient estimator. We conclude this chapter in Section 6.4.

6.1 Model and preliminaries

6.1.1 The model

In this chapter, we consider a d-node $M|M|1$ tandem queue, with arrival rate λ and service rates $\mu_1, \ldots, \mu_d$ for queues $1, \ldots, d$ respectively, and we are again interested in estimating the probability that the total number of customers in the system reaches some high level N during a busy cycle of the system. Since the case $d = 2$ has been studied mostly in literature, we mainly consider this case. However, it seems likely that all of the results can also be extended to $d > 2$ and so we will briefly comment on these extensions when necessary. Thus, we now let $d = 2$.

We consider the underlying embedded discrete time Markov chain and we assume without loss of generality $\lambda + \mu_1 + \mu_2 = 1$. Furthermore we assume that $\lambda < \min\{\mu_1, \mu_2\}$, so that the system is stable. Note that with this assumption, Assumption 2.3 is satisfied. As in [14, 19, 22], we let the state description be the number of customers in each queue, denoted by $\mathbf{Z}_i = (Z_{1,i}, Z_{2,i})$, where $Z_{j,i}$ is the number of customers in queue j after i transitions.

We let $\mathbf{v}_k$ denote the possible transitions and $\Theta(\mathbf{v}_k)$ their corresponding probabilities, that is, $\mathbf{v}_0 = (1, 0)$ corresponds to an arrival to the first queue, and has probability $\Theta(\mathbf{v}_0) = \lambda$. Similarly, we have $\mathbf{v}_1 = (-1, 1)$ and $\mathbf{v}_2 = (0, -1)$, $\Theta(\mathbf{v}_1) = \mu_1$ and $\Theta(\mathbf{v}_2) = \mu_2$. The transitions $\mathbf{v}_1$ and $\mathbf{v}_2$ can only occur when $Z_{1,i} > 0$ and $Z_{2,i} > 0$ respectively. If $Z_{j,i} = 0$ for some queue j, thus there are no customers in queue j, we add a self-loop transition with probability μ_j in order to make sure that the sum of all rates equals 1 (so that all rates are probabilities).

As in [19], we let $\mathbf{X}_i = \frac{1}{N}\mathbf{Z}_i$ be the scaled state description, which has the advantage that its elements are in $[0, 1]$ and therefore do not increase as N increases.

This allows us to make the following definitions:

$$\begin{aligned}
D &= \{(x_1, x_2) : x_j > 0,\ x_1 + x_2 < 1\},\\
\delta_1 &= \{(0, x_2) : 0 < x_2 < 1\},\\
\delta_2 &= \{(x_1, 0) : 0 < x_1 < 1\},\\
\delta_e &= \{(x_1, x_2) : x_j \geq 0,\ x_1 + x_2 = 1\},\\
\delta_0 &= \{(0, 0)\},
\end{aligned}$$

and this is sketched in Figure 6.1.

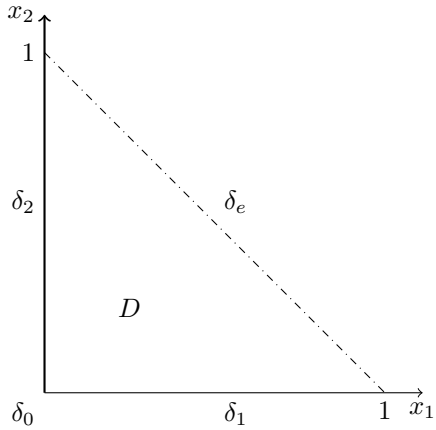


Figure 6.1 A sketch of the scaled state description of the event of interest.

Using these definitions we can define the first time that the process hits level N in a busy cycle as

$$\tau_N = \inf\{i > 0 : X_i \in \delta_e, \mathbf{X}_k \notin \delta_0\ \forall k = 1, \ldots, i-1\},$$

and we set $\tau_N = \infty$ if the process hits δ_0 before δ_e. Therefore our probability of interest can be written as

$$p_N = \mathbb{P}\left(\tau_N < \infty \mid \mathbf{X}_0 = (\tfrac{1}{N}, 0)\right).$$

It is known that the asymptotic decay rate of this probability is given by

$$\gamma \doteq -\lim_{N\to\infty} \frac{1}{N} \log p_N = -\Lambda_A(-\theta^*) = -\log\left(\frac{\lambda}{\min\{\mu_1, \mu_2\}}\right) = \min\{\gamma_1, \gamma_2\}, \quad (6.1)$$

where $\gamma_j = -\log(\lambda/\mu_j)$, see (2.5) or [30]. Using Definition 2.1, it follows that the queue for which we have $\gamma_j = \gamma$ is the bottleneck queue. For the $M|M|1$ tandem queue, the bottleneck queue is equivalent to the queue with the largest server utilization. In most papers on similar topics, see for example [12, 14, 19, 30], it is assumed that $\mu_2 \leq \mu_1$, as for the probability of interest the queues are interchangeable [41]. In this chapter we both consider $\mu_2 \leq \mu_1$ and $\mu_1 \leq \mu_2$.

6.1.2 Importance sampling simulation

To estimate our probability of interest using simulation, we use importance sampling. In importance sampling, we perform our simulation under some new measure $\mathbb{Q}$. While doing so, we keep track of the likelihood ratio $L(\mathcal{P})$ of a path $\mathcal{P} = (\mathbf{X}_i, i = 0, \ldots, \tau_N)$:

$$L(\mathcal{P}) = \prod_{i=0}^{\tau_N - 1} \frac{\Theta(\mathbf{Y}_i)}{\bar{\Theta}(\mathbf{Y}_i \mid \mathbf{X}_i)}, \tag{6.2}$$

where $\mathbf{Y}_i = N(\mathbf{X}_{i+1} - \mathbf{X}_i)$ if $\mathbf{X}_{i+1} \neq \mathbf{X}_i$ and $\mathbf{Y}_i = \mathbf{v}_k$ if $\mathbf{X}_{i+1} = \mathbf{X}_i$ and $\mathbf{X}_i \in \delta_k$, $k = 1, 2$, to include the self-loop transition when one of the queues is empty. Let $I(\mathcal{P}) = \mathbb{1}\{\tau_N < \infty\}$ indicate whether we have reached our event of interest during a busy cycle of the system or not. Then, under the new measure $\mathbb{Q}$, we have

$$p_N = \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})I(\mathcal{P})\right],$$

thus, $L(\mathcal{P})I(\mathcal{P})$ is the estimator for our probability of interest.

As in [19], we construct a subsolution $W(\mathbf{x})$ – see Definition 5.4 – and we use this function to specify a change of measure in the following way:

$$\bar{\Theta}(\mathbf{v}_i \mid \mathbf{x}) = \Theta(\mathbf{v}_i) e^{-\langle DW(\mathbf{x}), \mathbf{v}_i \rangle} e^{\mathbb{H}(\mathbf{x}, DW(\mathbf{x}))}, \tag{6.3}$$

where

$$\mathbb{H}(\mathbf{x}, DW(\mathbf{x})) = -\log\left(\sum_{i=0}^{2} \Theta(\mathbf{v}_i) e^{-\langle DW(\mathbf{x}), \mathbf{v}_i \rangle}\right), \tag{6.4}$$

which is in accordance with (5.5). If we compare (6.4) with the corresponding notation in [14, 19], a factor 2 is missing. However, we will also scale the function $W(\mathbf{x})$ accordingly, so that the change of measure remains the same.

As we are interested in finding a change of measure that gives an asymptotically efficient estimator for p_N, we need

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathbb{E}^{\mathbb{Q}}\left[L(\mathcal{P})^2 I(\mathcal{P})\right] \leq -2\gamma,$$

or equivalently,

$$\limsup_{N \to \infty} \frac{1}{N} \log \mathbb{E}\left[L(\mathcal{P})I(\mathcal{P})\right] \leq -2\gamma,$$

to hold, see Definition 1.2 combined with (6.1).

In [14, 19, 22], a subsolution is constructed (when $\mu_2 \leq \mu_1$ in [14, 19]) and asymptotic efficiency is proven for a specific choice of $W(\mathbf{x})$. In this chapter, we provide conditions on $W(\mathbf{x})$ and prove that under these conditions we get an asymptotically efficient estimator. Afterwards, we study the possibilities of the function $W(\mathbf{x})$, and the corresponding change of measure in (6.3).

6.1.3 Subsolution approach

In both [19] and [22] the change of measure for (2-node) tandem queues (and Jackson networks) has been studied. In those papers, the change of measure has been determined using subsolutions. We first briefly recap the ideas presented in those papers, mainly referring to Chapter 5. A formal definition of classical subsolution can be found in Definition 5.4, and we use the same approach as in that chapter, that is, we include the boundaries in $\mathbb{H}(\mathbf{x}, DW(\mathbf{x}))$, see the text below Definition 5.4.

It is known from [12, 30] that for an asymptotically efficient change of measure it is not possible that $DW(\mathbf{x})$ is constant throughout the whole state space as in [35], and in [19] this is 'confirmed' since the change of measure yielding an asymptotically efficient estimator differs from the change of measure in [35] along one of the boundaries of the state space and near the origin. Thus, in order to find an asymptotically efficient change of measure, we determine several (constant) changes of measure for various regions of the state space, in particular along the boundaries of the state space, and combine these so that we have a change of measure for the whole state space.

In order to determine such a change of measure that differs along various parts of the state space, in [19] there are multiple – say r – affine functions $W_k^\delta(\mathbf{x})$, $k = 1, \ldots, r$, considered such that for each of these functions $\mathbb{H}(\mathbf{x}, DW_k^\delta(\mathbf{x})) \geq 0$ for some part of the state space, so that all r functions cover the whole state space, and for at least one of these functions we have $W_k^\delta(\mathbf{x}) \leq 0$ for $\mathbf{x} \in \delta_e$. All these functions have the following form

$$W_k^\delta(\mathbf{x}) = \langle \boldsymbol{\alpha}_k, \mathbf{x} \rangle + c_k - d_k\delta, \tag{6.5}$$

where $\boldsymbol{\alpha}_k = (\alpha_{k,1}, \alpha_{k,2})$, which we assume to be finite, specifies the gradient of $W_k^\delta(\mathbf{x})$ and both c_k and $d_k > 0$ are constants. Combining these functions to $W^\delta(\mathbf{x})$ by taking the minimum of these functions for all $\mathbf{x}$ then results in a piecewise affine function, that unfortunately is not continuously differentiable. That is, we have, similar to (5.14),

$$W^\delta(\mathbf{x}) = W_1^\delta(\mathbf{x}) \wedge \ldots \wedge W_r^\delta(\mathbf{x}).$$

In order to satisfy the continuous differentiability, which is the first requirement for a function to be a classical subsolution, the functions $W_k^\delta(\mathbf{x})$ are combined into a (continuous) function $W^{\varepsilon,\delta}(\mathbf{x})$ by a similar mollification procedure as in (5.15), where in this case we obtain

$$W^{\varepsilon,\delta}(\mathbf{x}) = -\varepsilon \log \sum_{k=1}^{r} e^{-W_k^\delta(\mathbf{x})/\varepsilon}, \tag{6.6}$$

such that $W^{\varepsilon,\delta}(\mathbf{x})$ converges to $W^\delta(\mathbf{x})$ when $\varepsilon \to 0$. Throughout this chapter, we make the same assumptions on ε and δ as in Chapter 5, see Assumption 5.17.

The gradient of (6.6) is then used as change of measure in (6.3). It can be expressed as

$$DW^{\varepsilon,\delta}(\mathbf{x}) = \sum_{k=1}^{r} \rho_k(\mathbf{x})\boldsymbol{\alpha}_k, \tag{6.7}$$

where

$$\rho_k(\mathbf{x}) = \frac{e^{-W_k^\delta(\mathbf{x})/\varepsilon}}{\sum_{j=1}^{r} e^{-W_j^\delta(\mathbf{x})/\varepsilon}}, \tag{6.8}$$

see also (5.16) and (5.17).

The functions $\rho_k(\mathbf{x})$ are weight factors for the 'influence' of each function $W_k^\delta(\mathbf{x})$, and so of each different 'regional' change of measure, in the final change of measure. They can also be used to define a change of measure slightly different than in (6.3) as follows, see also [14, 19],

$$\bar{\Theta}(\mathbf{v}_i \mid \mathbf{x}) = \Theta(\mathbf{v}_i) \sum_{k=1}^{r} \rho_k(\mathbf{x}) e^{-\langle \boldsymbol{\alpha}_k, \mathbf{v}_i \rangle} e^{\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_k)}, \tag{6.9}$$

which we will refer to in this chapter later on. In fact, we will also show that asymptotic efficiency for the change of measure in (6.3) implies asymptotic efficiency for the change of measure in (6.9), similar as in [14]. We note that, from an implementation perspective, the change of measure in (6.9) is preferred over the change of measure in (6.3), see also Section 3.8.6 in [19].

6.1.4 Existing changes of measure

Now that the general ideas of [19, 22] have been presented, we will show the different functions $W^\delta(\mathbf{x})$ that have been used in [19, 22] to obtain an asymptotically efficient change of measure, as well as figures of the function $W^\delta(\mathbf{x})$ for both bottleneck queues j, in the following sections.

6.1.4.1 Change of measure from [19]

In [19], queue 2 is always considered to be the bottleneck queue because for the probability of interest the queues are interchangeable. In that paper, a subsolution $W^\delta(\mathbf{x})$ is determined by considering three functions $W_k^\delta(\mathbf{x})$, $k = 1, 2, 3$. The three functions from [19] are as follows

$$\begin{aligned} W_1^\delta(\mathbf{x}) &= -\gamma_2 x_1 - \gamma_2 x_2 + \gamma_2 - \delta, \\ W_2^\delta(\mathbf{x}) &= -\gamma_2 x_1 + \gamma_2 - 2\delta, \\ W_3^\delta(\mathbf{x}) &= \gamma_2 - 3\delta, \end{aligned} \tag{6.10}$$

and the function $W^\delta(\mathbf{x})$ is illustrated in Figure 6.2. We remark that also in the functions in (6.10) we scaled the results from [19] (except for the constant in front of δ) by a factor $1/2$, but this does not influence the resulting changes of measure in (6.3) and (6.9).

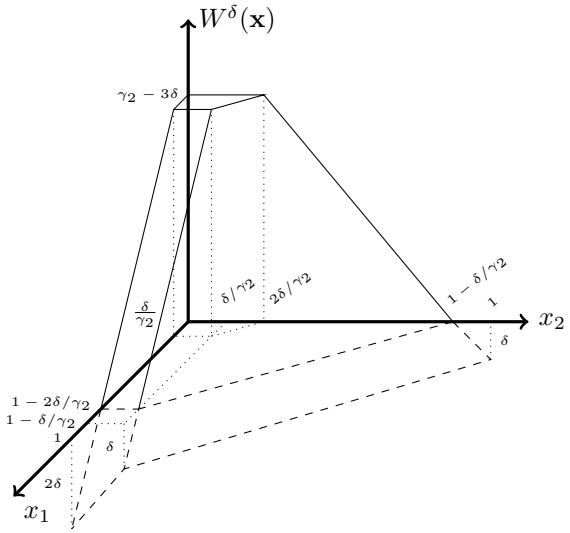


Figure 6.2 The function $W^\delta(\mathbf{x})$ from (6.10). Queue 2 is the bottleneck queue (so $\gamma_2 \leq \gamma_1$).

6.1.4.2 Changes of measure from [22]

In [22], the work in [19] is extended to Jackson networks and hence in [22] all queues being the bottleneck queue are considered, as in this chapter. Not only the probability that the total number of customers in the system reaches some high level N during a busy cycle of the system is considered in [22], the authors also consider buffer overflow in a single queue or in several queues at the same time. If we consider a 2-node tandem queue with queue 2 being the bottleneck queue, we find that the following four functions are used in [22]:

$$\begin{aligned} W_1^\delta(\mathbf{x}) &= -\gamma_1 x_1 - \gamma_2 x_2 + \gamma_2 - \ \delta, \\ W_2^\delta(\mathbf{x}) &= -\gamma_2 x_1 - \gamma_2 x_2 + \gamma_2 - 2\delta, \\ W_3^\delta(\mathbf{x}) &= -\gamma_1 x_1 \qquad\quad + \gamma_2 - 3\delta, \\ W_4^\delta(\mathbf{x}) &= \qquad\qquad\quad \gamma_2 - 4\delta, \end{aligned} \tag{6.11}$$

and $W^\delta(\mathbf{x})$ is illustrated in Figure 6.3.
In [22], the authors do not explicitly mention by which constant δ is multiplied, though implicit requirements with a proof for asymptotic efficiency are given. For simplicity we use $k\delta$ throughout this chapter. It turns out that these values are sufficient for asymptotic efficiency, as we show later in this chapter, but we will also see that they are by no means unique.

Since there is no limitation to queue 2 being the bottleneck queue in [22], we also present the result from [22] when queue 1 is the bottleneck queue. The result from [22], when queue 1 is the bottleneck queue, is different compared to when

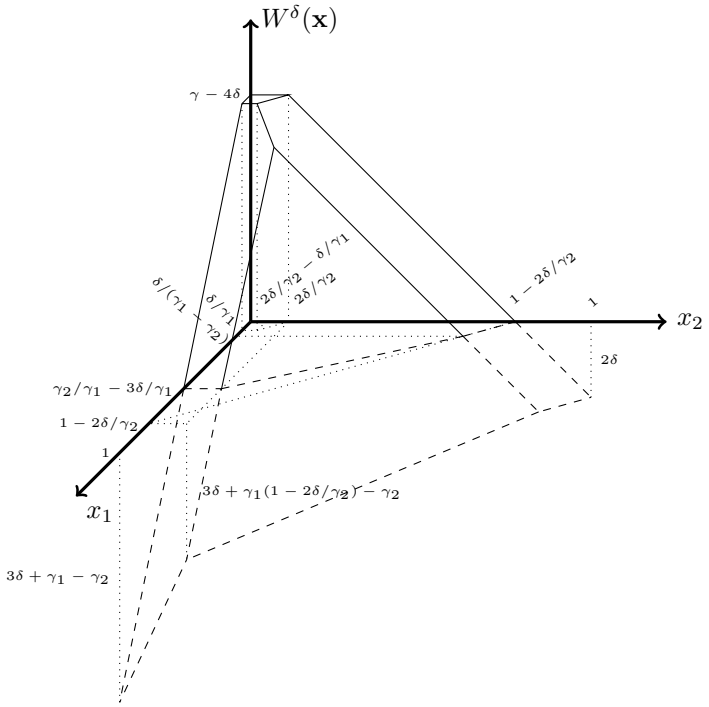


Figure 6.3 The function $W^\delta(\mathbf{x})$ from (6.11). Queue 2 is the bottleneck queue (so $\gamma_2 \leq \gamma_1$).

queue 2 is the bottleneck queue, even though for the probability of interest both queues are interchangeable. When queue 1 is the bottleneck queue, the following three functions are derived in [22]:

$$\begin{aligned} W_1^\delta(\mathbf{x}) &= -\gamma_1 x_1 - \gamma_2 x_2 + \gamma_1 - \quad \delta, \\ W_2^\delta(\mathbf{x}) &= -\gamma_1 x_1 \qquad\quad + \gamma_1 - 2\delta, \\ W_3^\delta(\mathbf{x}) &= \qquad\qquad\quad \gamma_1 - 3\delta, \end{aligned} \tag{6.12}$$

and the resulting function $W^\delta(\mathbf{x})$ is illustrated in Figure 6.4. These subsolutions are very similar to the ones in (6.11), but with one function less. Also here, the multiplication factor of δ is not explicitly mentioned, but we use the values above which are sufficient for asymptotic efficiency, as we will show later in this chapter.

6.1.4.3 Comparison of the existing changes of measure

In this section, we briefly comment on the similarities and differences of the changes of measure from [19] and [22]. We will do so by comparing the functions $W^\delta(\mathbf{x})$ for all different cases, see Figures 6.2–6.4.

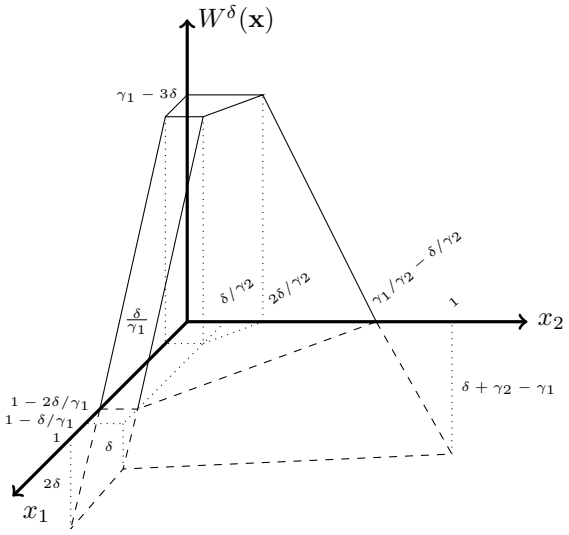


Figure 6.4 The function $W^\delta(\mathbf{x})$ from (6.12). Queue 1 is the bottleneck queue (so $\gamma_1 \leq \gamma_2$).

When queue 2 is the bottleneck queue, we see from Figures 6.2 and 6.3 that along the x_2-axis the function $W^\delta(\mathbf{x})$ is roughly the same. The only difference is that in Figure 6.3 the function, along the x_2-axis, is δ lower. In particular, this part of the state space covers the most likely path. In all other parts of the state space, the function $W^\delta(\mathbf{x})$ in Figure 6.3 is slightly steeper than in Figure 6.2 since $\gamma_2 < \gamma_1$. These observations suggest that any change of measure based on some function $W^\delta(\mathbf{x})$ that somehow lies 'in between' the functions in Figures 6.2 and 6.3 is also asymptotically efficient. In Section 6.3 we show that this is indeed the case.

When queue 1 is the bottleneck queue, there is not much to compare. However, since there are already two possibilities for the change of measure to be asymptotically efficient, and even more to expect, when queue 2 is the bottleneck queue, also the case when queue 1 is the bottleneck queue is studied in Section 6.3.

6.2 Sufficient conditions for asymptotic efficiency

Similar to [19, 22], the construction of the changes of measure in this chapter is based on finding appropriate subsolutions $W^{\varepsilon,\delta}(\mathbf{x})$. We start with a general proof for asymptotic efficiency for a change of measure based on the subsolution approach and the mollification procedure that is explained in Section 6.1.3. In our theorem, we provide sufficient conditions for the subsolution yielding an

asymptotic efficient change of measure, which we use later for the derivation of the possibilities for the change of measure. Afterwards, we discuss some general observations with respect to some of the conditions in our theorem.

6.2.1 Main result

Theorem 6.1. *Consider a 2-node $M|M|1$ tandem queue. Let $W_k^\delta(\mathbf{x})$ be affine functions for all $k = 1, \ldots, r$, as in (6.5), and let the classical subsolution $W^{\varepsilon,\delta}(\mathbf{x})$ be constructed by using (6.6). Then, using the gradient of the function $W^{\varepsilon,\delta}(\mathbf{x})$ as change of measure in (6.3) and (6.9) results in an asymptotically efficient estimator if there exist functions $f(N)$, $g(N)$ and $h(N)$ with $\lim_{N\to\infty} f(N) = \lim_{N\to\infty} g(N) = \lim_{N\to\infty} h(N) = 0$, such that*

1. $\sum_{k=1}^r \rho_k(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_k) \geq f(N)$;
2. $W^{\varepsilon,\delta}(\mathbf{x}) \leq g(N)$, *for all* $\mathbf{x} \in \delta_e$;
3. $W^{\varepsilon,\delta}(\mathbf{0}) \geq \gamma + h(N)$.

Proof. We start by showing that under the above conditions the change of measure in (6.3) is asymptotically efficient, after which it follows that also the change of measure in (6.9) is asymptotically efficient by using a similar argument as in Theorem 2 from [14].

From (6.2) and (6.3), it follows that the likelihood ratio of a path $L(\mathcal{P})$ is given by

$$\log L(\mathcal{P}) = N \sum_{j=0}^{\tau_N - 1} \langle DW^{\varepsilon,\delta}(\mathbf{X}_j), \mathbf{X}_{j+1} - \mathbf{X}_j \rangle - \sum_{j=0}^{\tau_N - 1} \mathbb{H}(\mathbf{X}_j, DW^{\varepsilon,\delta}(\mathbf{X}_j)).$$

We find, using (6.7) and (6.8), that for all states $\mathbf{x}$ we have

$$\mathbb{H}(\mathbf{x}, DW^{\varepsilon,\delta}(\mathbf{x})) \geq \sum_{k=1}^{r} \rho_k(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_k),$$

due to concavity of $\mathbb{H}(\mathbf{x}, DW^{\varepsilon,\delta}(\mathbf{x}))$ in its second argument, see Proposition 3.2 in [19]. Combining the two expressions above, we arrive at

$$\log L(\mathcal{P}) \leq N \sum_{j=0}^{\tau_N - 1} \langle DW^{\varepsilon,\delta}(\mathbf{X}_j), \mathbf{X}_{j+1} - \mathbf{X}_j \rangle - \sum_{j=0}^{\tau_N - 1} \sum_{k=1}^{r} \rho_k(\mathbf{X}_j)\mathbb{H}(\mathbf{X}_j, \boldsymbol{\alpha}_k) \tag{6.13}$$

$$\leq N \sum_{j=0}^{\tau_N - 1} \langle DW^{\varepsilon,\delta}(\mathbf{X}_j), \mathbf{X}_{j+1} - \mathbf{X}_j \rangle - f(N)\tau_N, \tag{6.14}$$

where the last inequality follows from Condition 1.

Similar to Lemma 2 in [14], also when using r regions, we can obtain the following bound. The idea of this Lemma in [14] is to replace the summation in (6.14) by $W^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - W^{\varepsilon,\delta}(\mathbf{0})$ and to give an upper bound on the error that is introduced. Thus, since by construction $|\boldsymbol{\alpha}_k| \leq c$ for some $0 \leq c < \infty$, see below (6.5), we find

$$|N \sum_{j=0}^{\tau_N - 1} \langle DW^{\varepsilon,\delta}(\mathbf{X}_j), \mathbf{X}_{j+1} - \mathbf{X}_j \rangle - N(W^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - W^{\varepsilon,\delta}(\mathbf{0}))| \leq \frac{5c^2}{\varepsilon N} \tau_N. \tag{6.15}$$

Next, we follow similar steps as in Theorem 1 of the same paper. By combining (6.14) and (6.15) we have

$$\begin{aligned} \log L(\mathcal{P}) &\leq N(W^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - W^{\varepsilon,\delta}(\mathbf{0})) + \frac{5c^2}{\varepsilon N}\tau_N - f(N)\tau_N \\ &\leq \left(g(N) - h(N) - \gamma\right) N + \left(\frac{5c^2}{\varepsilon N} - f(N)\right) \tau_N, \end{aligned}$$

where the second inequality follows from Conditions 2 and 3 when $\mathbf{X}_{\tau_N} \in \delta_e$, and thus we find, as in [14],

$$\begin{aligned} &\frac{1}{N} \log \mathbb{E}\left[L(\mathcal{P}) I(\mathcal{P})\right] \\ &= \frac{1}{N} \log \left(\mathbb{E}\left[L(\mathcal{P}) \mid I(\mathcal{P}) = 1\right] \cdot \mathbb{P}(I(\mathcal{P}) = 1)\right) \\ &\leq \frac{1}{N} \log \left(\mathbb{E}\left[e^{(g(N) - h(N) - \gamma)N + \left(\frac{5c^2}{\varepsilon N} - f(N)\right)\tau_N} \mid \tau_N < \infty \right] p_N \right) \\ &= g(N) - h(N) - \gamma + \frac{1}{N} \log \mathbb{E}\left[e^{\left(\frac{5c^2}{\varepsilon N} - f(N)\right)\tau_N} \mid \tau_N < \infty \right] + \frac{1}{N} \log p_N. \end{aligned} \tag{6.16}$$

To conclude the proof, we need Lemma 3 from [14], which states that for any sequence $\theta_N \geq 0$ such that $\lim_{N\to\infty} \theta_N = 0$, we have

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left[e^{\theta_N \tau_N} \mid \tau_N < \infty \right] = 0. \tag{6.17}$$

Thus, taking limits in (6.16) gives

$$\begin{aligned} &\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left[L(\mathcal{P}) I(\mathcal{P})\right] \\ &\leq \limsup_{N\to\infty} \left(g(N) - h(N) - \gamma + \frac{1}{N} \log \mathbb{E}\left[e^{\left(\frac{5c^2}{\varepsilon N} - f(N)\right)\tau_N} \mid \tau_N < \infty \right] + \frac{1}{N} \log p_N \right) \\ &= -2\gamma, \end{aligned} \tag{6.18}$$

where the last equation follows by using (6.1), $\lim_{N\to\infty} g(N) = \lim_{N\to\infty} h(N) = 0$ and since $\left(\frac{5c^2}{\varepsilon N} - f(N)\right) \to 0$ when $N \to \infty$ we can apply (6.17) to the fourth term of (6.18). This concludes the proof for the change of measure in (6.3).

For the change of measure in (6.9) we note that, similar to [14], we have

$$\begin{aligned}\log L(\mathcal{P}) &= \log \prod_{j=0}^{\tau_N-1} \frac{1}{\sum_{k=1}^r \rho_k(\mathbf{X}_j) e^{-\langle \boldsymbol{\alpha}_k, \mathbf{X}_{j+1}-\mathbf{X}_j\rangle} e^{\mathbb{H}(\mathbf{X}_j,\boldsymbol{\alpha}_k)}} \\ &= -\sum_{j=0}^{\tau_N-1} \log\left(\sum_{k=1}^r \rho_k(\mathbf{X}_j) e^{-\langle \boldsymbol{\alpha}_k, \mathbf{X}_{j+1}-\mathbf{X}_j\rangle} e^{\mathbb{H}(\mathbf{X}_j,\boldsymbol{\alpha}_k)}\right) \\ &\leq \sum_{j=0}^{\tau_N-1}\sum_{k=1}^r \rho_k(\mathbf{X}_j)\langle \boldsymbol{\alpha}_k, \mathbf{X}_{j+1}-\mathbf{X}_j\rangle - \sum_{k=1}^r \rho_k(\mathbf{X}_j)\mathbb{H}(\mathbf{X}_j,\boldsymbol{\alpha}_k) \\ &= \sum_{j=0}^{\tau_N-1} \langle DW^{\varepsilon,\delta}(\mathbf{X}_j), \mathbf{X}_{j+1}-\mathbf{X}_j\rangle - \sum_{k=1}^r \rho_k(\mathbf{X}_j)\mathbb{H}(\mathbf{X}_j,\boldsymbol{\alpha}_k),\end{aligned}$$

where the inequality follows by concavity of the logarithm and the last equality follows by definition of $DW^{\varepsilon,\delta}(\mathbf{x})$, see (6.7). Thus, we have the same bound as in (6.13) and so we also find that the change of measure in (6.9) is asymptotically efficient. □

Remark 6.2. *In* (6.18) *we see that we end up with some term* $g(N)-h(N)-\gamma$, *which goes to* $-\gamma$ *as* $N\to\infty$. *This term arises from bounding* $W^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - W^{\varepsilon,\delta}(\mathbf{0})$ *on a path that leads to the overflow level. However, we cannot have* $W^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - W^{\varepsilon,\delta}(\mathbf{0}) < -\gamma$ *when* $N\to\infty$, *since this would contradict with Jensen's inequality:*

$$\liminf_{N\to\infty} \frac{1}{N}\log \mathbb{E}\left[L(\mathcal{P})I(\mathcal{P})\right] \geq -2\gamma.$$

That is, it is impossible to obtain a tighter bound. As a result, we find that for an asymptotically efficient change of measure we need $W^{\varepsilon,\delta}(\mathbf{X}_{\tau_N}) - W^{\varepsilon,\delta}(\mathbf{0}) \to -\gamma$ *when* $N\to\infty$ *on at least one path that leads to reaching the overflow level, for example, the most likely path, see also [22]. Thus, on such a path, we need* $\lim_{N\to\infty} W^{\varepsilon,\delta}(\mathbf{x}) = 0$ *for* $\mathbf{x}\in\delta_e$ *and* $\lim_{N\to\infty} W^{\varepsilon,\delta}(\mathbf{0}) = \gamma$.

Remark 6.3. *It seems likely that Theorem 6.1 can also be extended to a d-node* $M|M|1$ *tandem queue (and Jackson networks), with the same sufficient conditions as in the current statement for* $d=2$. *In order to do so, observe that in the proof of Theorem 6.1, Lemma 3 from [14] is the only part that restricts to* $d=2$ *(and hence to tandem queues). Thus, one could either extend this result to* $d>2$ *(and Jackson networks), or use similar techniques as in [19, 22] in order to show that the theorem holds in a more general setting.*

6.2.2 General observations

Now that we have shown under which conditions we obtain an asymptotically efficient change of measure based on subsolutions, it remains to find $\boldsymbol{\alpha}_k$, c_k and d_k for all $k=1,\ldots,r$ such that Conditions 1, 2 and 3 of Theorem 6.1 are satisfied.

In this section, we make some general observations with respect to Condition 1 and 3 of Theorem 6.1 when considering a 2-node $M|M|1$ tandem queue, that are used later to construct the possibilities for the change of measure. These observations are independent of the bottleneck queue.

6.2.2.1 Observations with respect to Condition 1 of Theorem 6.1

We recall that the first condition is $\sum_{k=1}^{r} \rho_k(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_k) \geq f(N)$ for some $f(N)$ with $\lim_{N\to\infty} f(N) = 0$. In this section, we state some observations with respect to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha})$ for some general $\boldsymbol{\alpha}$, independent of k, after which we present some observations with respect to $\rho_k(\mathbf{x})$.

By considering all possibilities for $\mathbb{1}\{x_j > 0\}$, $j = 1, 2$, in a busy cycle of the system, we find from (6.4) that

$$\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) = \begin{cases} -\log(\lambda e^{-\alpha_1} + \mu_1 e^{\alpha_1-\alpha_2} + \mu_2 e^{\alpha_2}) & \text{if } x_1 > 0,\ x_2 > 0, \\ -\log(\lambda e^{-\alpha_1} + \mu_1 e^{\alpha_1-\alpha_2} + \mu_2) & \text{if } x_1 > 0,\ x_2 = 0, \\ -\log(\lambda e^{-\alpha_1} + \mu_1 + \mu_2 e^{\alpha_2}) & \text{if } x_1 = 0,\ x_2 > 0. \end{cases} \tag{6.19}$$

We start by finding solutions to $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) \geq 0$, for particular parts of the state space, or equivalently to

$$\lambda e^{-\alpha_1} + \mu_1 e^{(\alpha_1-\alpha_2)\mathbb{1}\{x_1>0\}} + \mu_2 e^{\alpha_2 \mathbb{1}\{x_2>0\}} \leq 1. \tag{6.20}$$

In Lemma 6.4 we consider $x_j > 0$ for $j = 1, 2$. We fix either α_1 or α_2, and we show for which values of α_2 or α_1 (6.20) holds. In Lemma 6.5 we state two relations considering (6.20), at one of the boundaries (either $x_1 = 0$ or $x_2 = 0$) that will be used several times later in this chapter. In Sections 6.3.1–6.3.4, we use the results of these Lemmas to lower bound $\sum_{k=1}^{r} \rho_k(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_k)$, such that Condition 1 of Theorem 6.1 is satisfied.

Lemma 6.4. *Suppose $x_j > 0$ for $j = 1, 2$. We prove the following statements:*

- *If $\alpha_2 = -\gamma_2$, then (6.20) holds iff $\alpha_1 \in [-\max_j \gamma_j, -\min_j \gamma_j]$;*
- *If $\alpha_1 = -\gamma_1$, then (6.20) holds iff $\alpha_2 \in [-\gamma_2, 0]$;*
- *If $\alpha_1 = -\gamma_2$, then (6.20) holds iff $\alpha_2 \in [-\gamma_2, \gamma_1 - \gamma_2]$;*
- *If $\alpha_1 = 0$, then (6.20) holds iff $\alpha_2 \in [0, \gamma_1 - \gamma_2]$.*

Proof. For the first statement, let $\alpha_2 = -\gamma_2$ and $x_j > 0$ for $j = 1, 2$. Then, (6.20) reduces to $\lambda e^{-\alpha_1} + (\mu_1\mu_2/\lambda)\, e^{\alpha_1} + \lambda \leq 1$ which holds if and only if

$$(1 - \mu_1 - \mu_2)^2 x^2 - (\mu_1 + \mu_2)(1 - \mu_1 - \mu_2)x + \mu_1\mu_2 \leq 0,$$

where we let $x = e^{-\alpha_1}$. Using elementary calculus we find that $\alpha_1 \in [-\max_j \gamma_j, -\min_j \gamma_j]$. The other statements follow similarly. □

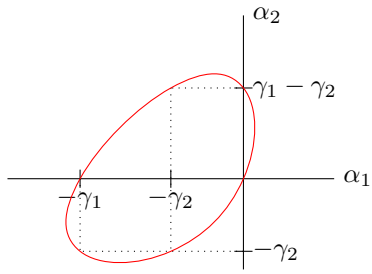


Figure 6.5 Sketch of the level set for which $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) = 0$ (that is, (6.20) holds with equality) for $\mathbf{x} > 0$, when queue 2 is the bottleneck queue (and so $\gamma_2 \leq \gamma_1$).

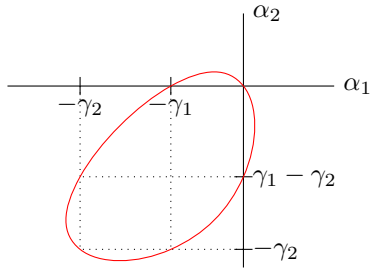


Figure 6.6 Similar to Figure 6.5, but when queue 1 is the bottleneck queue (and so $\gamma_1 \leq \gamma_2$).

As a result of Lemma 6.4, we can sketch the level set for all α_1, α_2 such that $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}) = 0$ for all $\mathbf{x} > 0$, by using the concavity of the function shown in [19], see Figures 6.5 and 6.6.

In the following lemma, we consider (6.20) at one of the boundaries given a certain choice for either α_1 or α_2. The first equation considers $x_1 > 0$, $x_2 = 0$ and $\alpha_1 = -\gamma_1$, the second equation considers $x_1 = 0$, $x_2 > 0$ and $\alpha_2 = -\gamma_2$.

Lemma 6.5. $\mu_1 + \lambda e^{-\alpha_2} + \mu_2 \leq 1$ *iff* $\alpha_2 \geq 0$ *and* $\lambda e^{-\alpha_1} + \mu_1 + \lambda \leq 1$ *iff* $\alpha_1 \geq -\gamma_2$.

Proof. The statements follow directly by elementary calculus, combined with $\lambda + \mu_1 + \mu_2 = 1$ and $\lambda > 0$. □

We conclude this section with a remark on $\rho_k(\mathbf{x})$. Using (6.8) we find that

$$\rho_k(\mathbf{x}) = \frac{e^{-W_k^\delta(\mathbf{x})/\varepsilon}}{\sum_{i=1}^r e^{-W_i^\delta(\mathbf{x})/\varepsilon}} \leq \frac{e^{-W_k^\delta(\mathbf{x})/\varepsilon}}{e^{-W_\ell^\delta(\mathbf{x})/\varepsilon}} = e^{\left(W_\ell^\delta(\mathbf{x}) - W_k^\delta(\mathbf{x})\right)/\varepsilon}, \tag{6.21}$$

for any ℓ, where the inequality follows trivially.

6.2.2.2 Observations with respect to Condition 3 of Theorem 6.1

In Remark 6.2 it is noted that we need $\bar{h}(N) \leq W^{\varepsilon,\delta}(\mathbf{0}) - \gamma \leq h(N)$ such that $\lim_{N\to\infty} h(N) = \lim_{N\to\infty} \bar{h}(N) = 0$. This is satisfied when $W_k^\delta(\mathbf{0}) - \gamma \geq \bar{h}_k(N)$ such that $\lim_{N\to\infty} \bar{h}_k(N) = 0$ for all k and $W_k^\delta(\mathbf{0}) - \gamma \leq h_k(N)$ such that $\lim_{N\to\infty} h_k(N) = 0$ for some k, see (6.6). By (6.5), we have $W_k^\delta(\mathbf{0}) = c_k - d_k\delta$. Therefore, we find

$$c_k = \gamma \quad \text{for all } k,$$

as a sufficient condition to satisfy Condition 3 since $\delta \to 0$ when $N \to \infty$ by Assumption 5.17.

As a result of the observations above, in order to construct the possibilities for the change of measure, it remains to find $\boldsymbol{\alpha}_k$ and d_k for all $k = 1, \ldots, r$ that satisfy Conditions 1, 2 and 3 of Theorem 6.1. This is the topic of the next section.

6.3 Construction of the subsolution

In this section, we construct possible subsolutions, based on the approach mentioned in Section 6.1.3, that satisfy the conditions in Theorem 6.1 and thus yield an asymptotically efficient estimator. It may not be clear at first sight that the method below results in subsolutions that satisfy all these conditions, since the construction is partly based on intuition. However, we conclude all sections by showing that the conditions are indeed satisfied.

For the 2-node $M|M|1$ tandem queue, we consider both possibilities for the bottleneck queue, that is, we consider both queue 1 and queue 2 as bottleneck queue. In addition, we focus on a maximum of 4 different regions. More regions may or may not be possible. However, this is undesirable from a practical point of view and does not contribute to an easier implementation of the change of measure.

We start by using three regions and queue 2 being the bottleneck queue, since this case has been studied most in literature. Afterwards, we consider three regions and queue 1 the bottleneck queue. We conclude this section with four regions, for which we again consider both queue 2 and queue 1 as the bottleneck respectively.

6.3.1 Three regions and queue 2 bottleneck

In this section, our starting point is to consider three functions $W_k^\delta(\mathbf{x})$, $k = 1, 2, 3$, and we let queue 2 be the bottleneck queue, thus $\gamma_2 \leq \gamma_1$. As the boundaries turned out to be crucial in designing an asymptotically efficient change of measure, we consider the following regions: (i) $x_2 > 0$, which also covers $x_1 = 0$; (ii) $x_1 > 0$, which also covers $x_2 = 0$; and (iii) $x_1 \geq 0$ and $x_2 \geq 0$, so that we have

covered the whole state space. All regions overlap in the sense that they all cover the case in which both $x_1 > 0$ and $x_2 > 0$. However, by construction of the continuously differentiable subsolution, in that part of the state space the function $W_k^\delta(\mathbf{x})$ of only one of the three regions will be used since we use the minimum of the functions $W_k^\delta(\mathbf{x})$. Clearly, the third region covers the whole state space, but it is important to note that there is no non-trivial solution that satisfies Condition 1 from Theorem 6.1 for the whole state space. The most important part of region three is that it covers $x_1 = x_2 = 0$.

The zero change of measure can be used when both $x_1 = x_2 = 0$, since this part of the state space is not covered by $x_1 > 0$ nor $x_2 > 0$. This in turn gives us $\boldsymbol{\alpha}_3 = (\alpha_{3,1}, \alpha_{3,2}) = (0, 0)$. The use of the zero change of measure is a choice, similar as in [19, 22], but we use this as a starting point. Remark though, that the zero change of measure satisfies $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_3) = 0$ for all $\mathbf{x}$ such that $x_1 = x_2 = 0$, and thus satisfies Condition 1 of Theorem 6.1. In particular, the condition in (6.20) for this part of the state description is equivalent to $\alpha_{3,1} \geq 0$. In addition, when *only* using the zero change of measure it is not possible to satisfy the conditions in Remark 6.2 and thus it is not possible to construct a subsolution that results in an asymptotically efficient estimator. The latter is also due to the fact that with the zero change of measure we are simulating the original system.

The ordering of the regions that we assign can be found in Table 6.1, the reasons for this ordering will become clear later in this section.

Table 6.1 Overview of proposed regions for the case $r = 3$.

Region	k
$x_2 > 0$	1
$x_1 > 0$	2
$x_1 \geq 0$, $x_2 \geq 0$	3

6.3.1.1 Finding $\boldsymbol{\alpha}_1$

To find $\boldsymbol{\alpha}_1$, we start with Condition 2, which is $W^{\varepsilon,\delta}(\mathbf{x}) \leq g(N)$ for all $\mathbf{x} \in \delta_e$, where $\lim_{N\to\infty} g(N) = 0$. In Remark 6.2 it is noted that for the most likely path equality should hold. Trivially we have $W^{\varepsilon,\delta}(\mathbf{x}) \leq W_1^\delta(\mathbf{x})$, with equality on the most likely path when $N \to \infty$, since $W_1^\delta(\mathbf{x})$ covers the most likely path. As a result, taking into account Section 6.2.2.2, we must have $\alpha_{1,2} = -\gamma_2$, where we recall that $\boldsymbol{\alpha}_k = (\alpha_{k,1}, \alpha_{k,2})$ for all k.

Using Condition 1, we can now determine $\alpha_{1,1}$. We only consider the term of the summation from this condition that involves $k = 1$, that is, $\rho_1(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_1)$ should be non-negative for large enough N. As $W_1^\delta(\mathbf{x})$ is *not* designed for $x_2 = 0$, the intuition is that the weight factor $\rho_1(\mathbf{x})$ tends to 0 for large enough N for all states $\mathbf{x}$ such that $x_2 = 0$, see also (6.29) below. Thus, for all $\mathbf{x}$ such that $x_2 > 0$

we see that, using (6.19) for $\boldsymbol{\alpha} = \boldsymbol{\alpha}_1$, we can only satisfy Condition 1, if we have

$$\lambda e^{-\alpha_{1,1}} + \mu_1\mu_2/\lambda e^{\alpha_{1,1}} + \lambda \leq 1, \tag{6.22}$$

$$\lambda e^{-\alpha_{1,1}} + \mu_1 + \lambda \leq 1, \tag{6.23}$$

where we used $\alpha_{1,2} = -\gamma_2$. As a result of Lemma 6.4, the first bullet, we find that (6.22) is satisfied when $\alpha_{1,1} \in [-\gamma_1, -\gamma_2]$. By using Lemma 6.5, we find that $\alpha_{1,1} \geq -\gamma_2$ is necessary in order to satisfy (6.23), and hence we must have $\alpha_{1,1} = -\gamma_2$. Therefore, we need

$$\boldsymbol{\alpha}_1 = (-\gamma_2, -\gamma_2),$$

in order to get an asymptotically efficient change of measure based on Theorem 6.1. For future reference, we remark that as a result of this condition on $\boldsymbol{\alpha}_1$ we find, using (6.19),

$$\rho_1(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_1) = \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0, \\ -\rho_1(\mathbf{x})\log(\mu_1 + 2\mu_2) & \text{if } x_1 > 0,\ x_2 = 0, \\ 0 & \text{if } x_1 = 0,\ x_2 > 0, \end{cases} \tag{6.24}$$

since $\lambda + \mu_1 + \mu_2 = 1$.

6.3.1.2 Finding $\boldsymbol{\alpha}_2$

Using the underlying idea of the construction of the subsolution – that is, the idea to construct several functions, each for different parts of the state space, that are combined through mollification to obtain a classical subsolution – we determine conditions on $\boldsymbol{\alpha}_2$ and d_k for $k = 1, 2, 3$. For some parts of the state space, it is known which function $W_k^\delta(\mathbf{x})$ has to be the minimum of the three functions or cannot be the minimum of the three. For example, when $x_1 = 0$, it follows that $W_2^\delta(\mathbf{x})$ cannot be the minimum function, since it is designed for $x_1 > 0$. As a result, we find that for some parts of the state space, some weight factors $\rho_k(\mathbf{x})$ must tend to 0 as $N \to \infty$.

To start with, we consider the origin of the state space, that is, $x_1 = x_2 = 0$. Here, we want $W_3^\delta(\mathbf{x})$ to be the minimum function, since this is the only function that is designed for this part of the state space. Thus, we need both $W_3^\delta(\mathbf{x}) < W_1^\delta(\mathbf{x})$ and $W_3^\delta(\mathbf{x}) < W_2^\delta(\mathbf{x})$ for $\mathbf{x} = \mathbf{0}$. Trivially, these inequalities result in

$$d_1 < d_3 \quad \text{and} \quad d_2 < d_3. \tag{6.25}$$

Secondly, we consider the boundary $x_2 = 0$ (and so $x_1 > 0$). At this part of the state space, we want $W_2^\delta(\mathbf{x})$ to be the minimum function. Thus, we want to have both

$$W_2^\delta(\mathbf{x}) < W_1^\delta(\mathbf{x}) \Leftrightarrow (\alpha_{2,1} - \alpha_{1,1})x_1 < (d_2 - d_1)\delta, \tag{6.26}$$

$$W_2^\delta(\mathbf{x}) < W_3^\delta(\mathbf{x}) \Leftrightarrow \ \alpha_{2,1} \qquad\qquad x_1 < (d_2 - d_3)\delta, \tag{6.27}$$

for all $\mathbf{x}$ such that $x_2 = 0$. Clearly, the first inequality holds for all x_1 whenever

$$\alpha_{2,1} \leq \alpha_{1,1} = -\gamma_2 \quad \text{and} \quad d_1 < d_2. \tag{6.28}$$

As a result, we immediately have, for all $\mathbf{x}$ such that $x_2 = 0$, by using (6.21) for $k = 1$ and $\ell = 2$,

$$\rho_1(\mathbf{x}) \leq e^{((\gamma_2+\alpha_{2,1})x_1+(d_1-d_2)\delta)/\varepsilon} \leq e^{(d_1-d_2)\delta/\varepsilon}, \tag{6.29}$$

for all $\alpha_{2,1} \leq -\gamma_2$ and since $d_1 < d_2$, the right-hand side tends to 0 as N tends to infinity. This intuitively implies that the weight factor $\rho_1(\mathbf{x})$ for $\mathbf{x}$ such that $x_2 = 0$ tends to 0 when $N \to \infty$, as suggested in Section 6.3.1.1, and thus the change of measure that is designed for $x_2 > 0$ hardly has any influence when $x_2 = 0$.

The second inequality, (6.27), is satisfied for all $x_1 > \frac{(d_3-d_2)\delta}{-\alpha_{2,1}}$, which is positive because $\alpha_{2,1} \leq -\gamma_2 < 0$, see (6.28), and $d_2 < d_3$. Thus, $W_2^\delta(\mathbf{x})$ is the minimum function for all $x_1 > \frac{(d_3-d_2)\delta}{-\alpha_{2,1}}$, and note that the right-hand side of this inequality tends to zero as $N \to \infty$. For all other, very small, x_1, the function $W_3^\delta(\mathbf{x})$ is the minimum function. It turns out in the sequel that this is not a problem for the resulting change of measure to be asymptotically efficient, since the function $W_3^\delta(\mathbf{x})$ can be used throughout the whole state space. More importantly, $W_1^\delta(\mathbf{x})$ is *not* the minimum function for all $\mathbf{x}$ such that $x_1 > 0$ and $x_2 = 0$ whenever (6.28) holds.

Combining all conditions on d_k, see (6.25) and (6.28), we find $d_1 < d_2 < d_3$. Here we see that the choice $d_k = k$, as in [19], satisfies all requirements that we have imposed until now in order to get an asymptotically efficient change of measure based on Theorem 6.1, but it is by no means unique.

Next, we consider the boundary $x_1 = 0$ (and so $x_2 > 0$). Here, we want $W_1^\delta(\mathbf{x})$ to be the minimum function, since all other functions are not designed for this part of the state space. Thus, for all $\mathbf{x}$ such that $x_1 = 0$, we want to have both

$$W_1^\delta(\mathbf{x}) < W_2^\delta(\mathbf{x}) \Leftrightarrow (\alpha_{1,2} - \alpha_{2,2})x_2 < (d_1 - d_2)\delta, \tag{6.30}$$

$$W_1^\delta(\mathbf{x}) < W_3^\delta(\mathbf{x}) \Leftrightarrow \alpha_{1,2} \qquad x_2 < (d_1 - d_3)\delta. \tag{6.31}$$

The first inequality is equivalent to

$$(\alpha_{2,2} + \gamma_2)x_2 > (d_2 - d_1)\delta,$$

which unfortunately holds only for $x_2 > \frac{(d_2-d_1)\delta}{\alpha_{2,2}+\gamma_2}$, provided that $\alpha_{2,2} > -\gamma_2$. Since we do *not* want $W_2^\delta(\mathbf{x})$ to be the minimum function for $\mathbf{x}$ such that $x_1 = 0$ and $x_2 \leq \frac{(d_2-d_1)\delta}{\alpha_{2,2}+\gamma_2}$, we need $W_3^\delta(\mathbf{x}) < W_2^\delta(\mathbf{x})$ for those states $\mathbf{x}$. This condition is equivalent to

$$\alpha_{2,2}x_2 > (d_2 - d_3)\delta, \tag{6.32}$$

and if $\alpha_{2,2} \geq 0$, this inequality holds for all x_2, so in particular for $x_2 \leq \frac{(d_2-d_1)\delta}{\alpha_{2,2}+\gamma_2}$. So we need

$$\alpha_{2,2} \geq 0. \tag{6.33}$$

As a result, by using (6.21) for $k = 2$ and $\ell = 3$, we find for all $\mathbf{x}$ such that $x_1 = 0$

$$\rho_2(\mathbf{x}) \leq e^{(-\alpha_{2,2}x_2+(d_2-d_3)\delta)/\varepsilon} \leq e^{(d_2-d_3)\delta/\varepsilon}. \tag{6.34}$$

Since $d_2 < d_3$, the right-hand side tends to 0 as $N \to \infty$ and this implies that the weight factor $\rho_2(\mathbf{x})$ for all $\mathbf{x}$ such that $x_1 = 0$ tends to 0 as $N \to \infty$. Therefore, the change of measure that is designed for $x_1 > 0$, has hardly any influence when $x_1 = 0$.

The second inequality, (6.31), is satisfied for all $x_2 > \frac{(d_3-d_1)\delta}{\gamma_2}$. Thus, $W_1^\delta(\mathbf{x})$ is smaller than $W_3^\delta(\mathbf{x})$ for all $x_2 > \frac{(d_3-d_1)\delta}{\gamma_2}$. As a result, for all $x_2 \leq \frac{(d_3-d_1)\delta}{\gamma_2}$ we have that $W_3^\delta(\mathbf{x})$ is the minimum function (which is not a limitation for the resulting change of measure to be asymptotically efficient, since this function can be used for the whole state space). Recall that it is more important is that $W_2^\delta(\mathbf{x})$ is *not* the minimum function for all $x_1 = 0$ and $x_2 > 0$, and this requirement is satisfied.

To derive a lower bound on $\alpha_{2,1}$ and an upper bound on $\alpha_{2,2}$, in contrast to the bounds in (6.28) and (6.33), we use Condition 1 of Theorem 6.1. In this case, we only consider the term of the summation from this condition that considers $k = 2$, since this involves $\boldsymbol{\alpha}_2$. That is, $\rho_2(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_2)$ should be non-negative for large enough N. Recall that we have derived an upper bound for $\rho_2(\mathbf{x})$ for all $\mathbf{x}$ such that $x_1 = 0$, see (6.34), that tends to 0 as $N \to \infty$. Thus, for all $\mathbf{x}$ such that $x_1 > 0$ we see that, using (6.19) for $\boldsymbol{\alpha} = \boldsymbol{\alpha}_2$, we can only satisfy Condition 1, if we have

$$\lambda e^{-\alpha_{2,1}} + \mu_1 e^{\alpha_{2,1}-\alpha_{2,2}} + \mu_2 e^{\alpha_{2,2}} \leq 1, \tag{6.35}$$

$$\lambda e^{-\alpha_{2,1}} + \mu_1 e^{\alpha_{2,1}-\alpha_{2,2}} + \mu_2 \leq 1. \tag{6.36}$$

It is clear that, as we have $\alpha_{2,2} \geq 0$ from (6.33), the second inequality is implied by the first inequality. Using Lemma 6.4, Figure 6.5, (6.28) and (6.33), we find that the first inequality is satisfied when

$$\alpha_{2,1} \in [-\gamma_1, -\gamma_2] \quad \text{and} \quad \alpha_{2,2} \in [0, \gamma_1 - \gamma_2].$$

Clearly, $\alpha_{2,1}$ and $\alpha_{2,2}$ have a dependence, for example $\alpha_{2,1} = -\gamma_1$ implies $\alpha_{2,2} = 0$ and $\alpha_{2,1} = -\gamma_2$ implies $\alpha_{2,2} \in [0, \gamma_1 - \gamma_2]$, see Lemma 6.4 and Figure 6.5. The dependence can be found in (6.35), however, this equation cannot be simplified. For future reference, we remark that as a result of these conditions on $\boldsymbol{\alpha}_2$, we find from (6.19), by using $\alpha_{2,1} \geq -\gamma_1$ and $\alpha_{2,2} \leq \gamma_1 - \gamma_2$, that

$$\rho_2(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_2) \geq \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0, \\ 0 & \text{if } x_1 > 0,\ x_2 = 0, \\ -\rho_2(\mathbf{x}) \log(3\mu_1) & \text{if } x_1 = 0,\ x_2 > 0, \end{cases} \tag{6.37}$$

since $\lambda + \mu_1 + \mu_2 = 1$.

6.3.1.3 Summary and proof that all conditions are satisfied

To summarize, we have found the following values for $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ that intuitively satisfy all conditions for an asymptotically efficient change of measure based on Theorem 6.1, see Table 6.2.

Table 6.2 Possibilities for $\boldsymbol{\alpha}_k$ when queue 2 is the bottleneck queue, provided that $d_1 < d_2 < d_3$.

	$\alpha_{k,1}$	$\alpha_{k,2}$	k	Condition
$x_2 > 0$	$-\gamma_2$	$-\gamma_2$	1	
$x_1 > 0$	$[-\gamma_1, -\gamma_2]$	$[0, \gamma_1 - \gamma_2]$	2	(6.35)
$x_1 \geq 0,\ x_2 \geq 0$	0	0	3	

We show that these possibilities for $\boldsymbol{\alpha}_k$, $k = 1, 2, 3$, indeed give an asymptotically efficient change of measure, by considering all conditions in Theorem 6.1. Recall that $\lambda + \mu_1 + \mu_2 = 1$ implies $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_3) = 0$. To start with Condition 1, using (6.24), (6.29), (6.34) and (6.37) we find the following lower bound

$$\begin{aligned}\sum_{k=1}^{3} \rho_k(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_k) &\geq \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0 \\ -e^{(d_1-d_2)\delta/\varepsilon} \log(\mu_1 + 2\mu_2) & \text{if } x_1 > 0,\ x_2 = 0 \\ -e^{(d_2-d_3)\delta/\varepsilon} \log(3\mu_1) & \text{if } x_1 = 0,\ x_2 > 0 \end{cases} \\ &\geq -e^{\max\{d_1-d_2, d_2-d_3\}\delta/\varepsilon} \log(3\mu_1),\end{aligned}$$

where the last step follows since queue 2 is the bottleneck queue, and thus $\mu_2 \leq \mu_1$. It follows that Condition 1 is satisfied. For Condition 2 we note that for all $\mathbf{x} \in \delta_e$ we have $W_1^\delta(\mathbf{x}) = -d_1\delta$, and thus Condition 2 is satisfied since $W^{\varepsilon,\delta}(\mathbf{x}) \leq W_1^\delta(\mathbf{x})$. To conclude with Condition 3, we find

$$\begin{aligned}W^{\varepsilon,\delta}(\mathbf{0}) &\geq -\varepsilon \log(3e^{-W_3^\delta(\mathbf{0})/\varepsilon}) = -\varepsilon \log 3 + W_3^\delta(\mathbf{0}) \\ &= -\varepsilon \log 3 + \gamma_2 - d_3\delta, \end{aligned} \tag{6.38}$$

which goes to γ_2 as $N \to \infty$ and hence all conditions are satisfied. Therefore, the change of measure for the given possible values of $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$ and d_k, $k = 1, 2, 3$, is asymptotically efficient. Note that also the conditions in Remark 6.2 are satisfied, based on the construction.

6.3.1.4 Discussion

It is clear that the choice of $\boldsymbol{\alpha}_2 = (-\gamma_2, 0)$ results in the change of measure from [19], using $d_k = k$ for $k = 1, 2, 3$, and for this case $W^\delta(\mathbf{x})$ is illustrated in Figure 6.2. We show some other examples of the piecewise affine function $W^\delta(\mathbf{x})$

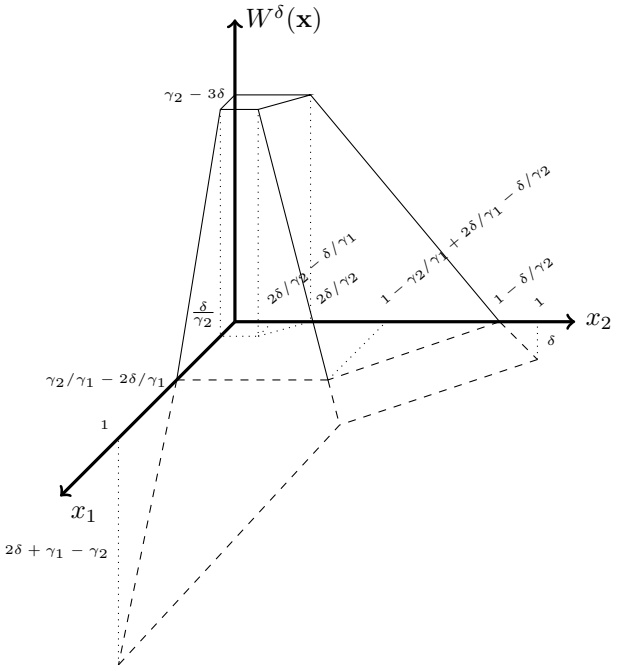


Figure 6.7 Display of $W^\delta(\mathbf{x})$ when queue 2 is the bottleneck queue, $\boldsymbol{\alpha}_2 = (-\gamma_1, 0)$ and $d_k = k$, $k = 1, 2, 3$.

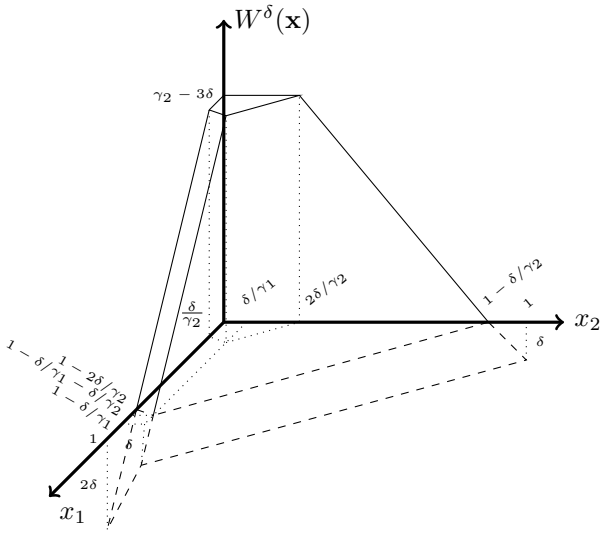


Figure 6.8 Display of $W^\delta(\mathbf{x})$ when queue 2 is the bottleneck queue, $\boldsymbol{\alpha}_2 = (-\gamma_2, \gamma_1 - \gamma_2)$ and $d_k = k$, $k = 1, 2, 3$.

below, where we let $\boldsymbol{\alpha}_2 = (-\gamma_1, 0)$ in Figure 6.7 and $\boldsymbol{\alpha}_2 = (-\gamma_2, \gamma_1 - \gamma_2)$ in Figure 6.8. In both cases we choose $d_k = k$, $k = 1, 2, 3$.
Choosing $\alpha_{2,2} = 0$, we find $\alpha_{2,1} \in [-\gamma_1, -\gamma_2]$ and thus $W^\delta(\mathbf{x})$ can be 'anything in between' Figure 6.2 and Figure 6.7. That is, the small area next to the x_1-axis in Figure 6.2 can be as steep as in Figure 6.7, and anything in between, while the resulting change of measure still gives an asymptotically efficient estimator. Choosing $\alpha_{2,1} = -\gamma_2$, we find $\alpha_{2,2} \in [0, \gamma_1 - \gamma_2]$ and thus $W^\delta(\mathbf{x})$ can be 'anything in between' Figure 6.2 and Figure 6.8. That is, the small area next to the x_1-axis in Figure 6.2 can be slightly twisted, while the resulting change of measure still gives an asymptotically efficient estimator.

We remark that the change of measure that is used for $x_2 > 0$ is the state-independent change of measure from [35]. The changes of measure that are found here, can be interpreted as 'protecting' the x_1-axis in the sense that we have to apply a different change of measure in that part of the state space. In Figure 6.7 we protect the x_1-axis quite a lot, rather than in Figure 6.8 where we only protect it slightly. Recall that the most likely path goes along the x_2-axis, where we have to apply the change of measure from [35], which also follows from the subsolution approach. It turns out that the change of measure along the most likely path is very important in the construction of an asymptotically efficient change of measure and that along the most likely path there is no variation possible for the change of measure. However, it turns out that for all other parts of the state space it is possible to apply a (slightly) different change of measure than the one from [19]. As we have seen in [22] and as we will see in Section 6.3.3, it is also possible to apply a different change of measure in the interior of the state space, rather than the same change of measure as along the most likely path or the change of measure that is applied along the x_1 axis.

6.3.2 Three regions and queue 1 bottleneck

In this section, we again consider three functions $W_k^\delta(\mathbf{x})$, $k = 1, 2, 3$, but in contrast to Section 6.3.1 we now let queue 1 be the bottleneck queue, so $\gamma_1 \leq \gamma_2$. The regions that we consider in this section are the same as in the previous section, see Table 6.1. By changing the bottleneck queue, the most likely path also changes and therefore we start the construction of an asymptotically efficient change of measure based on Theorem 6.1 by finding $\boldsymbol{\alpha}_2$.

6.3.2.1 Finding $\boldsymbol{\alpha}_2$

To find $\boldsymbol{\alpha}_2$, we start with Condition 2, as in Section 6.3.1.1, and we use Remark 6.2 in order to determine $\alpha_{2,1}$. When queue 1 is the bottleneck queue, the most likely path is covered by $W_2^\delta(\mathbf{x})$. Therefore, we find $\alpha_{2,1} = -\gamma_1$.

Next, we use Condition 1 to determine $\alpha_{2,2}$. Considering the term of the summation in this condition that involves $k = 2$, we need $\rho_2(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_2)$ to be non-negative for large enough N. For those parts of the state space where $x_1 = 0$,

we will find in the sequel that the weight factor $\rho_2(\mathbf{x})$ tends to 0 as $N \to \infty$. Since the function $W_2^\delta(\mathbf{x})$ is designed for $x_1 > 0$, it is expected that in that case the weight factor $\rho_2(\mathbf{x})$ does not tend to 0 as $N \to \infty$. Thus, for all $\mathbf{x}$ such that $x_1 > 0$ we see that, using (6.19) for $\boldsymbol{\alpha} = \boldsymbol{\alpha}_2$, we can only satisfy Condition 1 from Theorem 6.1, if we have

$$\begin{aligned} \mu_1 + \lambda e^{-\alpha_{2,2}} + \mu_2 e^{\alpha_{2,2}} &\leq 1, \\ \mu_1 + \lambda e^{-\alpha_{2,2}} + \mu_2 &\leq 1, \end{aligned}$$

where we used $\alpha_{2,1} = -\gamma_1$. By using the second bullet of Lemma 6.4 and Lemma 6.5 we find that both of the above conditions hold when $\alpha_{2,2} = 0$. Thus, in order to get an asymptotically efficient change of measure based on Theorem 6.1 we need

$$\boldsymbol{\alpha}_2 = (-\gamma_1, 0).$$

For future reference, we remark that as a result of this condition we find, using (6.19) and $\lambda + \mu_1 + \mu_2 = 1$,

$$\rho_2(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_2) = \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0, \\ 0 & \text{if } x_1 > 0,\ x_2 = 0, \\ -\rho_2(\mathbf{x}) \log(2\mu_1 + \mu_2) & \text{if } x_1 = 0,\ x_2 > 0. \end{cases} \tag{6.39}$$

6.3.2.2 Finding $\boldsymbol{\alpha}_1$

As in Section 6.3.1.2 we use the underlying idea of the construction of subsolutions in order to determine conditions on $\boldsymbol{\alpha}_2$ and d_k for $k = 1, 2, 3$. Since this approach has been fully explained in Section 6.3.1.2, we will skip most of the details and highlight the results.

By considering the origin of the state space, that is $x_1 = x_2 = 0$, we immediately find the same result as in (6.25), since that result is independent of $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$.

Next, we consider the boundary $x_2 = 0$ (and so $x_1 > 0$) and recall that in that case we want $W_2^\delta(\mathbf{x})$ to be the minimum function. The conditions that follow from this observation can be found in (6.26) and (6.27). Since we have determined that $\boldsymbol{\alpha}_2 = (-\gamma_1, 0)$, these conditions are equivalent to

$$\begin{aligned} (\alpha_{1,1} + \gamma_1)x_1 &> (d_1 - d_2)\delta, \\ \gamma_1 x_1 &> (d_3 - d_2)\delta. \end{aligned} \tag{6.40}$$

The first condition is satisfied when

$$\alpha_{1,1} \geq -\gamma_1 \quad \text{and} \quad d_1 < d_2, \tag{6.41}$$

since then the right-hand side of the inequality is negative. Similarly to (6.29) we find

$$\rho_1(\mathbf{x}) \leq e^{((-\gamma_1 - \alpha_{1,1})x_1 + (d_1 - d_2)\delta)/\varepsilon} \leq e^{(d_1 - d_2)\delta/\varepsilon}, \tag{6.42}$$

and so the weight factor $\rho_1(\mathbf{x})$ for $\mathbf{x}$ such that $x_2 = 0$ tends to zero when $N \to \infty$.

For the second condition, see (6.40), we find that this is satisfied for all $x_1 > \frac{(d_3-d_2)\delta}{\gamma_1}$, and so for small values of x_1 the function $W_3^\delta(\mathbf{x})$ is the minimum function. It turns out that this is not a problem for the resulting change of measure to be asymptotically efficient, since the function $W_3^\delta(\mathbf{x})$ can be used throughout the whole state space and since $W_2^\delta(\mathbf{x}) < W_1^\delta(\mathbf{x})$ for all $\mathbf{x}$ such that $x_2 = 0$. Therefore, it is ruled out that $W_1^\delta(\mathbf{x})$ is the minimum function for these $\mathbf{x}$, as desired.

At the boundary $x_1 = 0$ (and so $x_2 > 0$), we want $W_1^\delta(\mathbf{x})$ to be the minimum function. As a result, we want (6.30) and (6.31) to hold. These conditions are equivalent to

$$\alpha_{1,2}x_2 < (d_1 - d_2)\delta,$$
$$\alpha_{1,2}x_2 < (d_1 - d_3)\delta.$$

Clearly, both conditions are satisfied when the second condition holds, since $d_2 < d_3$. In particular, we need

$$\alpha_{1,2} < 0,$$

since $d_1 < d_3$, so that both conditions are satisfied for all $x_2 > \frac{(d_3-d_1)\delta}{-\alpha_{1,2}}$. To prevent $W_2^\delta(\mathbf{x})$ from being the minimum function for $\mathbf{x}$ such that $x_1 = 0$ and $x_2 \leq \frac{(d_3-d_1)\delta}{-\alpha_{1,2}}$, we need $W_3^\delta(\mathbf{x}) < W_2^\delta(\mathbf{x})$ for those states $\mathbf{x}$. This condition is equivalent to (6.32), and since $\alpha_{2,2} = 0$ and $d_2 < d_3$ this condition is always satisfied. As a result, we find for all $\mathbf{x}$ such that $x_1 = 0$ that (6.34) holds.

To get a tighter condition for $\alpha_{1,2}$ we use Condition 2 from Theorem 6.1. That is, for all $\mathbf{x}$ on the exit boundary we need to have $W^{\varepsilon,\delta}(\mathbf{x}) \leq g(N)$ such that $g(N) \to 0$ when $N \to \infty$. As we have $W^{\varepsilon,\delta}(\mathbf{x}) \to W^\delta(\mathbf{x})$ when $N \to \infty$, we in particular need $W_1^\delta(\mathbf{x})$ to be non-positive for large enough N when $x_1 = 0$ and $x_2 = 1$, since $W_1^\delta(\mathbf{x})$ is designed to be the minimum function at the boundary $x_1 = 0$ (and so $x_2 > 0$). Therefore we find

$$\alpha_{1,2} \leq -\gamma_1. \tag{6.43}$$

Using the same condition of Theorem 6.1, we also derive an upper bound on $\alpha_{1,1}$. Consider $W_2^\delta(\mathbf{x}) = -\gamma_1 x_1 + \gamma_1 - d_2\delta$, which is non-positive for all $x_1 \geq \frac{\gamma_1-d_2\delta}{\gamma_1}$. Thus, for all $x_1 < \frac{\gamma_1-d_2\delta}{\gamma_1}$ we need $W_1^\delta(\mathbf{x})$ to be non-positive for all $\mathbf{x} \in \delta_e$. On the exit boundary we have $x_1 + x_2 = 1$, so we find that for all $\mathbf{x} \in \delta_e$, using (6.43),

$$\begin{aligned} W_1^\delta(\mathbf{x}) &= \alpha_{1,1}x_1 + \alpha_{1,2}(1 - x_1) + \gamma_1 - d_1\delta \\ &\leq (\alpha_{1,1} + \gamma_1)x_1 - d_1\delta, \end{aligned}$$

which goes to zero as $N \to \infty$ if and only if $\alpha_{1,1} \leq -\gamma_1$. Combining with (6.41) we find

$$\alpha_{1,1} = -\gamma_1.$$

To conclude, we derive a lower bound on $\alpha_{1,2}$ using Condition 1. We only consider the term of the summation in Condition 1 that considers $k = 1$. That is, $\rho_1(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_1)$ should be non-negative for large enough N. We do *not* expect that $\rho_1(\mathbf{x}) \to 0$ when $N \to \infty$ for all $\mathbf{x}$ such that $x_2 > 0$, as $W_1^{\delta}(\mathbf{x})$ is designed for $x_2 > 0$. Thus, for all $\mathbf{x}$ such that $x_2 > 0$ we see that, using (6.19) for $\boldsymbol{\alpha} = \boldsymbol{\alpha}_1$, we can only satisfy Condition 1, if we have

$$\mu_1 + \lambda e^{-\alpha_{1,2}} + \mu_2 e^{\alpha_{1,2}} \leq 1, \tag{6.44}$$

$$\mu_1 + \mu_1 + \mu_2 e^{\alpha_{1,2}} \leq 1, \tag{6.45}$$

where we have used that $\alpha_{1,1} = -\gamma_1$. Using (6.43), it follows that the second inequality is implied by the first inequality. By using Lemma 6.4, the second bullet, we find that (6.44) is satisfied when $\alpha_{1,2} \in [-\gamma_2, -\gamma_1]$, and so is (6.45). For future reference, we remark that as a result of this condition we find, using (6.19) and $\alpha_{1,2} \geq -\gamma_2$,

$$\rho_1(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_1) \geq \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0, \\ -\rho_1(\mathbf{x}) \log(\mu_1 + 2\mu_2) & \text{if } x_1 > 0,\ x_2 = 0, \\ 0 & \text{if } x_1 = 0,\ x_2 > 0, \end{cases} \tag{6.46}$$

since $\lambda + \mu_1 + \mu_2 = 1$.

6.3.2.3 Summary and proof that all conditions are satisfied

Summarizing, we have found the following values for $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ that intuitively satisfy all conditions for an asymptotically efficient change of measure based on Theorem 6.1, see Table 6.3.

Table 6.3 Possibilities for $\boldsymbol{\alpha}_k$ when queue 1 is the bottleneck queue, provided that $d_1 < d_2 < d_3$.

	$\alpha_{k,1}$	$\alpha_{k,2}$	k
$x_2 > 0$	$-\gamma_1$	$[-\gamma_2, -\gamma_1]$	1
$x_1 > 0$	$-\gamma_1$	0	2
$x_1 \geq 0,\ x_2 \geq 0$	0	0	3

Again we show by considering all conditions in that theorem, indeed these possibilities for $\boldsymbol{\alpha}_k$, $k = 1, 2, 3$, give an asymptotically efficient change of measure. We start with Condition 1. First, recall that $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_3) = 0$, since $\boldsymbol{\alpha}_3 = \mathbf{0}$ and

$\lambda + \mu_1 + \mu_2 = 1$. Then, from (6.34), (6.39), (6.42) and (6.46), we find

$$\sum_{k=1}^{3} \rho_k(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_k) \geq \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0 \\ -e^{(d_1-d_2)\delta/\varepsilon} \log(\mu_1 + 2\mu_2) & \text{if } x_1 > 0,\ x_2 = 0 \\ -e^{(d_2-d_3)\delta/\varepsilon} \log(2\mu_1 + \mu_2) & \text{if } x_1 = 0,\ x_2 > 0 \end{cases}$$
$$\geq -e^{\max\{(d_1-d_2),(d_2-d_3)\}\delta/\varepsilon} \log(3\mu_2),$$

where the final inequality follows since $\mu_1 \leq \mu_2$ and so Condition 1 of Theorem 6.1 is satisfied. For Condition 2, we note for all $\mathbf{x} \in \delta_e$ we have $W_1^\delta(\mathbf{x}) \leq -d_1\delta$, and thus Condition 2 is satisfied since $W^{\varepsilon,\delta}(\mathbf{x}) \leq W_1^\delta(\mathbf{x})$. To conclude with Condition 3, we find a similar lower bound to $W^{\varepsilon,\delta}(\mathbf{0})$ as in (6.38), with γ_2 replaced by γ_1, since queue 1 is the bottleneck queue. Therefore, the change of measure for the given possible values of $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$ and d_k, $k = 1, 2, 3$ is asymptotically efficient.

6.3.2.4 Discussion

For $d_k = k$ for $k = 1, 2, 3$, it is clear that choosing $\boldsymbol{\alpha_1} = (-\gamma_1, -\gamma_2)$ results in the change of measure from [22] and for this case $W^\delta(\mathbf{x})$ is illustrated in Figure 6.4. Another example for $W^\delta(\mathbf{x})$, where we choose $\boldsymbol{\alpha}_1 = (-\gamma_1, -\gamma_1)$, is illustrated in Figure 6.9.

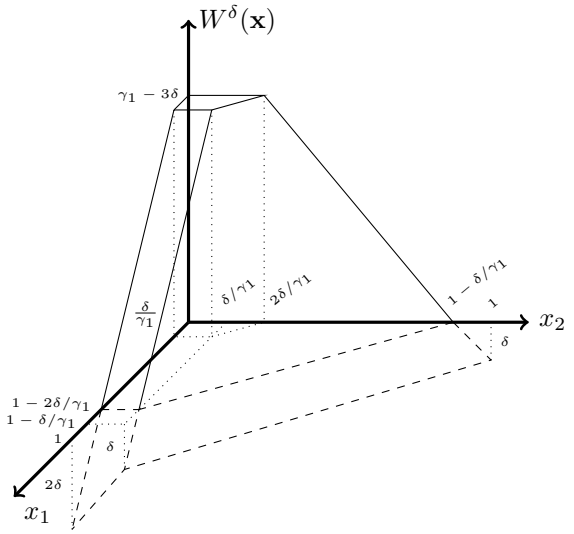


Figure 6.9 Display of $W^\delta(\mathbf{x})$ when queue 1 is the bottleneck queue, $\boldsymbol{\alpha}_1 = (-\gamma_1, -\gamma_1)$ and $d_k = k$, $k = 1, 2, 3$.

As can be seen in Table 6.3, $\alpha_{1,2}$ can range from $-\gamma_2$ to $-\gamma_1$ and thus $W^\delta(\mathbf{x})$ can be 'anything in between' Figure 6.4 and Figure 6.9. That is, the area next to the x_2-axis and in the interior in Figure 6.9 can be as steep as in Figure 6.4 and anything in between. Note that Figure 6.9 is very similar to Figure 6.2 in

the sense that γ_2 is replaced by γ_1, and so in particular when $\gamma_1 = \gamma_2$ the change of measure based on this function $W^\delta(\mathbf{x})$ can be used.

When queue 1 is the bottleneck queue, we find that the state-independent change of measure from [35] is only used around $x_2 = 0$. While when queue 2 is the bottleneck queue the change of measure based on $W^{\varepsilon,\delta}(\mathbf{x})$ can be interpreted as that we have to 'protect' the x_1-axis, see also Section 6.3.1, it seems natural that we have to 'protect' the x_2-axis when queue 1 is the bottleneck queue. Looking at Figures 6.4 and 6.9 we see that this is clearly not the case. One could argue that in this way we are 'overprotecting' the x_2-axis. However, it turns out that for an asymptotically efficient change of measure we only need to have the state-independent change of measure from [35] along the most likely path, which in this case is along the x_1-axis.

6.3.3 Four regions and queue 2 bottleneck

In this section, the starting point is to use four functions $W_k^\delta(\mathbf{x})$, $k = 1, 2, 3, 4$, and we let queue 2 be the bottleneck queue, thus $\gamma_2 \leq \gamma_1$. The regions that we consider can be found in Table 6.4, and the only difference compared to Section 6.3.1, where we considered three regions, is that we now have a dedicated region in case $x_j = 0$ for some queue j. The reasons for the ordering that we assigned in Table 6.4 will become clear later in this section.

Table 6.4 Overview of proposed regions for the case $r = 4$.

Region	k
$x_1 > 0,\ x_2 > 0$	1
$x_1 = 0,\ x_2 > 0$	2
$x_1 > 0,\ x_2 = 0$	3
$x_1 \geq 0,\ x_2 \geq 0$	4

The zero change of measure is used for $x_1 \geq 0$, $x_2 \geq 0$, for similar reasons as in Section 6.3.1. The only difference is that we now set $\boldsymbol{\alpha}_4 = \mathbf{0}$, instead of $\boldsymbol{\alpha}_3 = \mathbf{0}$ as in Section 6.3.1. The way the possibilities for the change of measure are constructed using 4 regions is very similar to the way this was done when using 3 regions in Sections 6.3.1 and 6.3.2, and hence we refer to these sections quite often. We start by considering the most likely path, and hence we start the construction of an asymptotically efficient change of measure based on Theorem 6.1 by finding $\boldsymbol{\alpha}_2$.

6.3.3.1 Finding $\boldsymbol{\alpha}_2$

As we consider conditions for that part of state space in which we have the most likely path, we can determine $\boldsymbol{\alpha}_2$ in exactly the same way as we determined $\boldsymbol{\alpha}_1$ in Section 6.3.1.1. The only difference is that the most likely path now lies in a

different region. As a result, we need

$$\boldsymbol{\alpha}_2 = (-\gamma_2, -\gamma_2),$$

in order to get an asymptotically efficient estimator based on Theorem 6.1.

6.3.3.2 Finding $\boldsymbol{\alpha}_3$

In order to find $\boldsymbol{\alpha}_3$, we proceed similarly to Section 6.3.1.2 and determine conditions based on the underlying idea of the construction of subsolutions. We start with the origin of the state space, that is, $x_1 = x_2 = 0$. In this case we find, similarly to the results in Section 6.3.1.2, that we need

$$d_1 < d_4 \quad \text{and} \quad d_2 < d_4 \quad \text{and} \quad d_3 < d_4. \tag{6.47}$$

Let us consider the boundary $x_2 = 0$ (and so $x_1 > 0$). In this part of the state space, we want $W_3^\delta(\mathbf{x})$ to be the minimum function. Thus, for all $\mathbf{x}$ such that $x_1 > 0$ and $x_2 = 0$ we need

$$W_3^\delta(\mathbf{x}) < W_1^\delta(\mathbf{x}) \Leftrightarrow (\alpha_{3,1} - \alpha_{1,1})x_1 < (d_3 - d_1)\delta, \tag{6.48}$$

$$W_3^\delta(\mathbf{x}) < W_2^\delta(\mathbf{x}) \Leftrightarrow (\alpha_{3,1} - \alpha_{2,1})x_1 < (d_3 - d_2)\delta, \tag{6.49}$$

$$W_3^\delta(\mathbf{x}) < W_4^\delta(\mathbf{x}) \Leftrightarrow \alpha_{3,1} x_1 < (d_3 - d_4)\delta, \tag{6.50}$$

and those conditions can only be satisfied when

$$\alpha_{3,1} \leq \alpha_{1,1} \quad \text{and} \quad d_1 < d_3 \quad \textit{and} \quad \alpha_{3,1} \leq \alpha_{2,1} = -\gamma_2 \quad \text{and} \quad d_2 < d_3. \tag{6.51}$$

Since $\alpha_{3,1} < 0$, the third condition is automatically satisfied for all $x_1 > \frac{(d_4 - d_3)\delta}{-\alpha_{3,1}}$. This is not a problem for the resulting change of measure to be asymptotically efficient, since it is implies that $W_4^\delta(\mathbf{x})$ is the minimum function for small values of x_1. More importantly, it is known from the second condition that $W_2^\delta(\mathbf{x})$ is *not* the minimum function for all $\mathbf{x}$ such that $x_1 > 0$ and $x_2 = 0$, as required.

As a result of (6.51), we find for all $\mathbf{x}$ such that $x_2 = 0$, using (6.21) for $k = 1$ and $\ell = 3$,

$$\rho_1(\mathbf{x}) \leq e^{((\alpha_{3,1} - \alpha_{1,1})x_1 + (d_1 - d_3)\delta)/\varepsilon} \leq e^{(d_1 - d_3)\delta/\varepsilon}, \tag{6.52}$$

and using (6.21) for $k = 2$ and $\ell = 3$,

$$\rho_2(\mathbf{x}) \leq e^{((\alpha_{3,1} - \alpha_{2,2})x_1 + (d_2 - d_3)\delta)/\varepsilon} \leq e^{(d_2 - d_3)\delta/\varepsilon}, \tag{6.53}$$

which both tend to zero as $N \to \infty$ since $d_1 < d_3$ and $d_2 < d_3$.

On the boundary $x_1 = 0$ (and so $x_2 > 0$), it is clear that we do *not* want $W_3^\delta(\mathbf{x})$ to be the minimum function. Instead, we want $W_2^\delta(\mathbf{x})$ to be the minimum function. That is, for all $\mathbf{x}$ such that $x_1 = 0$ and $x_2 > 0$ we need

$$W_2^\delta(\mathbf{x}) < W_1^\delta(\mathbf{x}) \Leftrightarrow (\alpha_{1,2} - \alpha_{2,2})x_2 > (d_1 - d_2)\delta, \tag{6.54}$$

$$W_2^\delta(\mathbf{x}) < W_3^\delta(\mathbf{x}) \Leftrightarrow (\alpha_{3,2} - \alpha_{2,2})x_2 > (d_3 - d_2)\delta, \tag{6.55}$$

$$W_2^\delta(\mathbf{x}) < W_4^\delta(\mathbf{x}) \Leftrightarrow -\alpha_{2,2}\, x_2 > (d_4 - d_2)\delta. \tag{6.56}$$

The first inequality is satisfied when

$$\alpha_{1,2} \geq \alpha_{2,2} = -\gamma_2 \quad \text{and} \quad d_1 < d_2. \tag{6.57}$$

For the second and third inequality we remark that they are only satisfied for $x_2 > \frac{(d_3-d_2)\delta}{\alpha_{3,2}+\gamma_2}$, provided that $\alpha_{3,2} > -\gamma_2$, and for $x_2 > \frac{(d_4-d_2)\delta}{\gamma_2}$ respectively, since we already imposed $d_2 < d_3$ and $d_2 < d_4$. Even though it is not an issue that $W_4^\delta(\mathbf{x})$ is the minimum function for small enough x_2, we need to prevent that $W_3^\delta(\mathbf{x})$ is the minimum function for states $\mathbf{x}$ such that $x_1 = 0$ and $x_2 \leq \frac{(d_3-d_2)\delta}{\alpha_{3,2}+\gamma_2}$. Thus, for those $\mathbf{x}$, we either need

$$W_1^\delta(\mathbf{x}) < W_3^\delta(\mathbf{x}) \Leftrightarrow (\alpha_{1,2} - \alpha_{3,2})x_2 < (d_1 - d_3)\delta, \quad \text{or} \tag{6.58}$$

$$W_4^\delta(\mathbf{x}) < W_3^\delta(\mathbf{x}) \Leftrightarrow \quad -\alpha_{3,2}\, x_2 < (d_4 - d_3)\delta. \tag{6.59}$$

For the first inequality we recall that $d_1 < d_3$ and so this inequality can only be satisfied when $\alpha_{1,2} < \alpha_{3,2}$. However, then the inequality is only satisfied for $x_2 > \frac{(d_3-d_1)\delta}{\alpha_{3,2}-\alpha_{1,2}}$. So, for $x_2 \leq \frac{(d_3-d_2)\delta}{\alpha_{3,2}+\gamma_2}$ we need the second inequality to be satisfied, which is equivalent to

$$\alpha_{3,2} \geq 0.$$

As a result of (6.57) and the previous display, we find for all $\mathbf{x}$ such that $x_1 = 0$, using (6.21) for $k = 1$ and $\ell = 2$,

$$\rho_1(\mathbf{x}) \leq e^{((\alpha_{2,2}-\alpha_{1,2})x_2+(d_1-d_2)\delta)/\varepsilon} \leq e^{(d_1-d_2)\delta/\varepsilon}, \tag{6.60}$$

and using (6.21) for $k = 3$ and $\ell = 4$,

$$\rho_3(\mathbf{x}) \leq e^{(-\alpha_{3,2}x_2+(d_3-d_4)\delta)/\varepsilon)} \leq e^{(d_3-d_4)\delta/\varepsilon}, \tag{6.61}$$

which both tend to zero as $N \to \infty$ since $d_1 < d_2$ and $d_3 < d_4$.

To determine a lower bound on $\alpha_{3,1}$ and an upper bound on $\alpha_{3,2}$, we use Condition 1 from Theorem 6.1. We only consider the term of the summation in this condition that involves $k = 3$. That is, $\rho_3(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_3)$ should be non-negative for large enough N. Recall that we have derived an upper bound for $\rho_3(\mathbf{x})$ for all $\mathbf{x}$ such that $x_1 = 0$, that tends to 0 as $N \to \infty$, see (6.61). Therefore, we see that, using (6.19) for $\boldsymbol{\alpha} = \boldsymbol{\alpha}_3$, we can only satisfy Condition 1 if we have

$$\begin{aligned} \lambda e^{-\alpha_{3,1}} + \mu_1 e^{\alpha_{3,1}-\alpha_{3,2}} + \mu_2 e^{\alpha_{3,2}} \leq 1, \\ \lambda e^{-\alpha_{3,1}} + \mu_1 e^{\alpha_{3,1}-\alpha_{3,2}} + \mu_2 \leq 1. \end{aligned} \tag{6.62}$$

Clearly, these are the same conditions as in (6.35) and (6.36), though the indices are different. This means that we get the same result for $\boldsymbol{\alpha}_3$ as for $\boldsymbol{\alpha}_2$ in Section 6.3.1, that is,

$$\alpha_{3,1} \in [-\gamma_1, -\gamma_2] \quad \text{and} \quad \alpha_{3,2} \in [0, \gamma_1 - \gamma_2],$$

and as additional requirement we have $\alpha_{3,1} \leq \alpha_{1,1}$, see (6.51). Clearly, there is a dependence of $\alpha_{3,1}$ and $\alpha_{3,2}$ that can be found in (6.62). As a result of these conditions on $\boldsymbol{\alpha}_3$ and the similarities with Section 6.3.1, a lower bound on $\rho_3(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_3)$ is the right-hand side of (6.37) with $\rho_2(\mathbf{x})$ replaced by $\rho_3(\mathbf{x})$.

To summarize the other conditions that were determined in this section, along with the conditions for $\boldsymbol{\alpha}_3$, we need

$$\alpha_{1,1} \geq \alpha_{3,1} \quad \text{and} \quad \alpha_{1,2} \geq -\gamma_2 \quad \text{and} \quad d_1 < d_2 < d_3 < d_4,$$

in order to get an asymptotically efficient estimator based on Theorem 6.1. In the next section we determine upper bounds on $\alpha_{1,1}$ and $\alpha_{1,2}$.

6.3.3.3 Finding $\boldsymbol{\alpha}_1$

To derive an upper bound for both $\alpha_{1,1}$ and $\alpha_{1,2}$, we remark that $W_1^\delta(\mathbf{x})$ is designed for all $\mathbf{x}$ such that $x_1 > 0$ and $x_2 > 0$. That is, for all such $\mathbf{x}$ we need

$$W_1^\delta(\mathbf{x}) < W_2^\delta(\mathbf{x}) \Leftrightarrow (\alpha_{2,1} - \alpha_{1,1})x_1 + (\alpha_{2,2} - \alpha_{1,2})x_2 > (d_2 - d_1)\delta, \tag{6.63}$$

$$W_1^\delta(\mathbf{x}) < W_3^\delta(\mathbf{x}) \Leftrightarrow (\alpha_{3,1} - \alpha_{1,1})x_1 + (\alpha_{3,2} - \alpha_{1,2})x_2 > (d_3 - d_1)\delta, \tag{6.64}$$

$$W_1^\delta(\mathbf{x}) < W_4^\delta(\mathbf{x}) \Leftrightarrow \qquad -\alpha_{1,1}\, x_1 \qquad\qquad -\alpha_{1,2}\, x_2 > (d_4 - d_1)\delta. \tag{6.65}$$

It is hard to derive conditions from these equations, though these equations in particular need to be satisfied for $x_1 = x_2$, in which case they are equivalent to

$$\begin{aligned} -2\gamma_2 - \alpha_{1,1} - \alpha_{1,2} &> \frac{(d_2 - d_1)\delta}{x_1}, \\ \alpha_{3,1} + \alpha_{3,2} - \alpha_{1,1} - \alpha_{1,2} &> \frac{(d_3 - d_1)\delta}{x_1}, \\ -\alpha_{1,1} - \alpha_{1,2} &> \frac{(d_4 - d_1)\delta}{x_1}. \end{aligned}$$

Since the right-hand side of these inequalities is negative, this gives $\alpha_{1,1} + \alpha_{1,2} \leq -2\gamma_2$, $\alpha_{1,1} + \alpha_{1,2} \leq 0$ and $\alpha_{1,1} + \alpha_{1,2} \leq \alpha_{3,1} + \alpha_{3,2}$. Thus, in particular we need

$$\alpha_{1,1} + \alpha_{1,2} \leq \min\{-2\gamma_2, \alpha_{3,1} + \alpha_{3,2}\}.$$

Provided that we first choose $\boldsymbol{\alpha}_1$ and afterwards choose $\boldsymbol{\alpha}_3$, we find using (6.57)

$$\alpha_{1,1} \leq -2\gamma_2 - \alpha_{1,2} \leq -\gamma_2. \tag{6.66}$$

Next, we take into account Condition 2 of Theorem 6.1, which states that for all $\mathbf{x} \in \delta_e$ we need $W^{\varepsilon,\delta}(\mathbf{x})$ to be non-positive for large enough N. Recall that $W^{\varepsilon,\delta}(\mathbf{x}) \to W^\delta(\mathbf{x})$ when $N \to \infty$, so in particular we need $W^\delta(\mathbf{x})$ to be non-positive for large enough N for all $\mathbf{x}$ such that $x_1 > 0$ and $x_2 > 0$ (and for those states we want $W^\delta(\mathbf{x}) = W_1^\delta(\mathbf{x})$). We have

$$\begin{aligned} W_1^\delta(\mathbf{x}) &= \alpha_{1,1}x_1 + \alpha_{1,2}x_2 + \gamma_2 - d_1\delta \\ &\leq (-\gamma_2 - \alpha_{1,2})x_1 + \alpha_{1,2} + \gamma_2 - d_1\delta, \end{aligned}$$

where the inequality follows from (6.66) and $x_1 + x_2 = 1$ for all $\mathbf{x} \in \delta_e$. As the right-hand side of this inequality needs to be non-positive for large enough N, we find

$$\alpha_{1,2} \leq -\gamma_2 + \frac{d_1\delta}{1-x_1}.$$

Since $x_1 > 0$ and $\delta \to 0$ for $N \to \infty$, it follows that $\alpha_{1,2} \leq -\gamma_2$. Considering (6.57), the result is that we need $\alpha_{1,2} = -\gamma_2$.

Next, we provide a lower bound on $\alpha_{1,1}$, using Condition 1 of Theorem 6.1. Similarly to all previous cases, we only consider the term of the summation in this condition for $k = 1$. Thus, we need $\rho_1(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_1)$ to be non-negative for large enough N. Since we have derived an upper bound for $\rho_1(\mathbf{x})$ when either $x_1 = 0$ or $x_2 = 0$, we find that, using (6.19) for $\boldsymbol{\alpha} = \boldsymbol{\alpha}_1$ and $\alpha_{1,2} = -\gamma_2$, we can only satisfy Condition 1 if

$$\lambda e^{-\alpha_{1,1}} + (\mu_1\mu_2/\lambda)e^{\alpha_{1,1}} + \lambda \leq 1.$$

Using Lemma 6.4, the first bullet, we find $\alpha_{1,1} \in [-\gamma_1, -\gamma_2]$. For future reference, we remark that as a result of the conditions on $\boldsymbol{\alpha}_1$, we find from (6.19), by using $-\gamma_1 \leq \alpha_{1,1} \leq -\gamma_2$,

$$\rho_1(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_1) \geq \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0, \\ -\rho_1(\mathbf{x}) \log(2\mu_1 + \mu_2) & \text{if } x_1 > 0,\ x_2 = 0, \\ -\rho_1(\mathbf{x}) \log(2\mu_1 + \lambda) & \text{if } x_1 = 0,\ x_2 > 0, \end{cases} \tag{6.67}$$

since $\lambda + \mu_1 + \mu_2 = 1$.

6.3.3.4 Summary and proof that all conditions are satisfied

To summarize, we have found the following values for $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$ and $\boldsymbol{\alpha}_4$ that intuitively satisfy all requirements for an asymptotically efficient change of measure based on Theorem 6.1, see Table 6.5.

Table 6.5 Possibilities for $\boldsymbol{\alpha}_k$ when queue 2 is the bottleneck queue, provided that $d_1 < d_2 < d_3 < d_4$.

	$\alpha_{k,1}$	$\alpha_{k,2}$	k	Conditions
$x_1 > 0,\ x_2 > 0$	$[-\gamma_1, -\gamma_2]$	$-\gamma_2$	1	
$x_1 = 0,\ x_2 > 0$	$-\gamma_2$	$-\gamma_2$	2	
$x_1 > 0,\ x_2 = 0$	$[-\gamma_1, \alpha_{1,1}]$	$[0, \gamma_1 - \gamma_2]$	3	(6.62)
				$\alpha_{3,1} + \alpha_{3,2} \geq \alpha_{1,1} - \gamma_2$
$x_1 \geq 0,\ x_2 \geq 0$	0	0	4	

We show that these possibilities for $\boldsymbol{\alpha}_k$, $k = 1, 2, 3, 4$, indeed give an asymptotically efficient change of measure, by considering all conditions in Theorem 6.1.

Recall that $\lambda + \mu_1 + \mu_2 = 1$ implies $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_4) = 0$. To start with Condition 1 of Theorem 6.1, we find that, using the result from (6.24) for $k = 2$, the result of (6.37) for $k = 3$, (6.52), (6.53), (6.60), (6.61) and (6.67),

$$\begin{aligned}
&\sum_{k=1}^{4} \rho_k(\mathbf{x}) \mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_k) \\
&\geq \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0 \\ -e^{(d_1-d_3)\delta/\varepsilon} \log(2\mu_1 + \mu_2) - e^{(d_2-d_3)\delta/\varepsilon} \log(\mu_1 + 2\mu_2) & \text{if } x_1 > 0,\ x_2 = 0 \\ -e^{(d_1-d_2)\delta/\varepsilon} \log(2\mu_1 + \lambda) - e^{(d_3-d_4)\delta/\varepsilon} \log(3\mu_1) & \text{if } x_1 = 0,\ x_2 > 0 \end{cases} \\
&\geq -2e^{\max\{(d_2-d_3),(d_1-d_2),(d_3-d_4)\}\delta/\varepsilon} \log(3\mu_1),
\end{aligned}$$

where the last step follows since queue 2 is the bottleneck queue, and hence $\lambda < \mu_2 \leq \mu_1$, and since $d_1 < d_2 < d_3 < d_4$. It follows that Condition 1 is satisfied. For Condition 2 we note that for all $\mathbf{x} \in \delta_e$ we have $W_2^\delta(\mathbf{x}) = -d_2\delta$, and thus Condition 2 is satisfied since $W^{\varepsilon,\delta}(\mathbf{x}) \leq W_2^\delta(\mathbf{x})$. Condition 3 follows similarly as in (6.38). Therefore, the change of measure for the given possible values of $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$, $\boldsymbol{\alpha}_4$ and d_k, $k = 1, 2, 3, 4$, is asymptotically efficient.

6.3.3.5 Discussion

From Table 6.5, we again see that the change of measure from [35] is used for $x_1 = 0$ and $x_2 > 0$, as expected. Clearly, the result in [22] satisfies the conditions in Table 6.5 but the possibilities for a change of measure are not limited to the results in [22]. Recall that in that paper also different overflow probabilities are considered, which limits the possibilities for an asymptotically efficient change of measure. The change of measure from [22], where in particular $\boldsymbol{\alpha}_1 = (-\gamma_1, -\gamma_2)$ and $\boldsymbol{\alpha}_3 = (-\gamma_1, 0)$, can be found in Figure 6.3.

From the results in Table 6.5 it is also clear that we could adapt Figure 6.3 such that along the x_1-axis we have a similar affine function as in Figure 6.7 or Figure 6.8. However, when choosing $\boldsymbol{\alpha}_3 = (-\gamma_2, \gamma_1 - \gamma_2)$, as we did for $\boldsymbol{\alpha}_2$ in Figure 6.8, we need $\alpha_{1,1} = -\gamma_2$ and so the functions $W_1^\delta(\mathbf{x})$ and $W_2^\delta(\mathbf{x})$ are almost the same. That is, the only difference then would be that their constants d_1 and d_2 differ.

Summarizing, at the x_1-axis we can make similar adaptations as when we considered three regions in Section 6.3.1. The difference is that now also in the interior we can 'push' the function a bit more towards the origin when we compare with Figure 6.2.

6.3.4 Four regions and queue 1 bottleneck

In this section, we again consider four functions $W_k^\delta(\mathbf{x})$, $k = 1, 2, 3, 4$, but in contrast to Section 6.3.3 we let queue 1 be the bottleneck queue, that is, $\gamma_1 \leq \gamma_2$. The regions that we consider in this section are the same as in the previous

section, see Table 6.4. The construction works very similar as when considering four regions when queue 2 is the bottleneck queue, hence we highlight the changes.

6.3.4.1 Finding $\boldsymbol{\alpha}_3$

As we consider conditions for that part of state space in which we have the most likely path, we can determine $\boldsymbol{\alpha}_3$ in exactly the same way as we determined $\boldsymbol{\alpha}_2$ in Section 6.3.2.1. The only difference is that the most likely path now lies in a different region. As a result, we need

$$\boldsymbol{\alpha}_3 = (-\gamma_1, 0),$$

in order to get an asymptotically efficient estimator based on Theorem 6.1.

6.3.4.2 Finding $\boldsymbol{\alpha}_1$

We again determine conditions for $\boldsymbol{\alpha}_1$ by using the underlying idea of the construction of subsolutions. Based on the conditions for the origin of the state space, we find as in Section 6.3.3, that (6.47) holds.

On the boundary $x_2 = 0$ (and so $x_1 > 0$), we want $W_3^\delta(\mathbf{x})$ to be the minimum function, and thus the conditions in (6.48)–(6.50) need to be satisfied. They can only hold true if

$$\alpha_{1,1} \geq -\gamma_1 \quad \text{and} \quad d_1 < d_3 \quad \textit{and} \quad \alpha_{2,1} \geq -\gamma_1 \quad \text{and} \quad d_2 < d_3. \tag{6.68}$$

The condition in (6.50) is automatically satisfied for $x_1 > \frac{(d_4-d_3)\delta}{\gamma_1}$. Recall from Section 6.3.3 that this is not an issue, since from (6.49) combined with (6.68) it follows that $W_2^\delta(\mathbf{x})$ is *not* the minimum function for all $\mathbf{x}$ such that $x_1 > 0$ and $x_2 = 0$. Using the requirements in (6.68), we again find the results in (6.52) and (6.53).

On the boundary $x_1 = 0$ (and so $x_2 > 0$), we want $W_2^\delta(\mathbf{x})$ to be the minimum function, and thus the conditions in (6.54)–(6.56) need to be satisfied. They can only hold true if

$$\alpha_{1,2} \geq \alpha_{2,2} \quad \text{and} \quad d_1 < d_2. \tag{6.69}$$

As in Section 6.3.3, the second and third inequality are only satisfied for $x_2 > \frac{(d_3-d_2)\delta}{\gamma_2}$ and $x_2 > \frac{(d_4-d_2)\delta}{\alpha_{2,2}}$, provided that $\alpha_{2,2} < 0$, respectively. Again, it is not a problem that $W_4^\delta(\mathbf{x})$ is the minimum function for small enough x_2, though we need to prevent that $W_3^\delta(\mathbf{x})$ is the minimum function for $x_2 \leq \frac{(d_3-d_2)\delta}{\gamma_2}$. In that case, we either need (6.58) or (6.59) to hold, and it is clear to see that (6.59) is always satisfied since $\alpha_{3,2} = 0$. Using the requirement in (6.69), we again find the results in (6.60) and (6.61).

For the interior, we want $W_1^\delta(\mathbf{x})$ to be the minimum function for all $\mathbf{x}$ such that $x_1 > 0$ and $x_2 > 0$. Therefore we need (6.63)–(6.65) to be satisfied. Similar

to the derivation just below these equations, we find that these equations are equivalent to

$$\alpha_{1,1} + \alpha_{1,2} \leq \min\{-\gamma_1, \alpha_{2,1} + \alpha_{2,2}\}.$$

We assume that we first choose $\boldsymbol{\alpha}_1$ and then choose $\boldsymbol{\alpha}_2$. As a result we find

$$\alpha_{1,2} \leq -\gamma_1 - \alpha_{1,1} \leq 0. \tag{6.70}$$

To determine conditions on $\alpha_{1,1}$ and $\alpha_{1,2}$, other than (6.68) and (6.69), we consider Condition 2 of Theorem 6.1. Recall from below (6.66) that $W^{\varepsilon,\delta}(\mathbf{x}) \to W^{\delta}(\mathbf{x})$. It is known that for all $\mathbf{x} \in \delta_e$, $W_3^{\delta}(\mathbf{x})$ is non-positive if and only if $x_1 \geq \frac{\gamma_1 - d_3\delta}{\gamma_1}$. So in particular for $x_1 = \frac{\gamma_1 - d_3\delta}{\gamma_1}$ and $x_2 = \frac{d_3\delta}{\gamma_1}$, we need either $W_1^{\delta}(\mathbf{x})$ or $W_2^{\delta}(\mathbf{x})$ to be non-positive for large enough N in order to satisfy Condition 2. Recall that $W_2^{\delta}(\mathbf{x})$ is designed for $x_1 = 0$ and $x_2 > 0$, so we need to consider $W_1^{\delta}(\mathbf{x})$. We find, using (6.70),

$$\begin{aligned} W_1^{\delta}(\mathbf{x}) &= \alpha_{1,1}x_1 + \alpha_{1,2}x_2 + \gamma_1 - d_1\delta \\ &\leq \alpha_{1,1}\frac{\gamma_1 - d_3\delta}{\gamma_1} + \gamma_1 - d_1\delta, \end{aligned}$$

which is less than or equal to zero if and only if $\alpha_{1,1} \leq \frac{-\gamma_1 + d_1\delta}{\gamma_1 - d_3\delta}\gamma_1$. Since $\delta \to 0$ as $N \to \infty$, we find $\alpha_{1,1} \leq -\gamma_1$, and as a result of (6.68) we find

$$\alpha_{1,1} = -\gamma_1.$$

Now that we have found an expression for $\alpha_{1,1}$, we again use that for each $\mathbf{x}$ such that $x_1 > 0$ and $x_2 > 0$ we must have on the exit boundary that $W_1^{\delta}(\mathbf{x})$ is non-positive for large enough N, since $W_1^{\delta}(\mathbf{x})$ is designed for this part of the state space. This results in a tighter upper bound for $\gamma_{1,2}$. That is,

$$W_1^{\delta}(\mathbf{x}) = -\gamma_1(1 - x_2) + \alpha_{1,2}x_2 + \gamma_1 - d_1\delta \leq 0,$$

which is equivalent to

$$\alpha_{1,2} \leq -\gamma_1 + \frac{d_1\delta}{x_2}.$$

Since $\delta \to 0$ as $N \to \infty$, we need

$$\alpha_{1,2} \leq \gamma_1.$$

We conclude with Condition 1 of Theorem 6.1. We again only consider the term of the summation in this condition that involves $k = 1$. That is, $\rho_1(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_1)$ should be non-negative for large enough N. Recall that we have derived an

upper bound for $\rho_1(\mathbf{x})$ when either $x_1 = 0$ or $x_2 = 0$. Thus, in order to satisfy Condition 1 we need

$$\mu_1 + \lambda e^{-\alpha_{1,2}} + \mu_2 e^{\alpha_{1,2}} \leq 1.$$

Using the second bullet from Lemma 6.4 we find that the inequality is satisfied whenever $\alpha_{1,2} \in [-\gamma_2, -\gamma_1]$. Thus we have found

$$\alpha_{1,1} = -\gamma_1 \quad \text{and} \quad \alpha_{1,2} \in [-\gamma_2, -\gamma_1].$$

For future reference, we remark that as a result of these conditions on $\boldsymbol{\alpha}_1$, we find from (6.19), by using $-\gamma_2 \leq \alpha_{1,2} \leq -\gamma_1$,

$$\rho_1(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_1) \geq \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0, \\ -\rho_1(\mathbf{x}) \log(\mu_1 + 2\mu_2) & \text{if } x_1 > 0,\ x_2 = 0, \\ -\rho_1(\mathbf{x}) \log(2\mu_1 + \mu_2 \lambda/\mu_1) & \text{if } x_1 = 0,\ x_2 > 0, \end{cases} \tag{6.71}$$

since $\lambda + \mu_1 + \mu_2 = 1$.

To summarize the other conditions that were determined in this section, along with the conditions for $\boldsymbol{\alpha}_1$, we need

$$\alpha_{2,1} \geq -\gamma_1, \quad \alpha_{2,2} \leq \alpha_{1,2}, \quad \alpha_{2,1} + \alpha_{2,2} \geq \alpha_{1,1} + \alpha_{1,2} \quad \text{and} \quad d_1 < d_2 < d_3 < d_4,$$

in order to get an asymptotically efficient estimator based on Theorem 6.1. In the next section we determine upper and lower bounds for $\alpha_{2,1}$ and $\alpha_{2,2}$, respectively.

6.3.4.3 Finding $\boldsymbol{\alpha}_2$

To derive bounds for both $\alpha_{2,1}$ and $\alpha_{2,2}$, we consider Condition 1. As before, we only consider the term of the summation in this condition that involves $k = 2$. That is, we need $\rho_2(\mathbf{x})\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_2)$ to be non-negative for large enough N. Since we have found an upper bound on $\rho_2(\mathbf{x})$ for all $\mathbf{x}$ such that $x_2 = 0$, we find, using (6.19) for $\boldsymbol{\alpha} = \boldsymbol{\alpha}_2$, that we can only satisfy Condition 1 if we have

$$\lambda e^{-\alpha_{2,1}} + \mu_1 e^{\alpha_{2,1} - \alpha_{2,2}} + \mu_2 e^{\alpha_{2,2}} \leq 1, \tag{6.72}$$

$$\lambda e^{-\alpha_{2,1}} + \mu_1 + \mu_2 e^{\alpha_{2,2}} \leq 1. \tag{6.73}$$

As the only requirements on $\boldsymbol{\alpha}_2$ we have until now are $\alpha_{2,1} \geq -\gamma_1$ and $\alpha_{2,2} \leq \alpha_{1,2}$, we use Figure 6.6, which is in turn based on Lemma 6.4, to find that the first inequality is satisfied when $\alpha_{2,1} \in [-\gamma_1, y]$ for some y and $\alpha_{2,2} \in [-\gamma_2, -\gamma_1]$. Recall that $\alpha_{2,2} \leq \alpha_{1,2}$. The dependence of $\alpha_{2,1}$ and $\alpha_{2,2}$ can be found in (6.72). Since we have $\alpha_{2,1} - \alpha_{2,2} \geq 0$, (6.73) is also satisfied. As a final remark, y can both be positive and negative, depending on whether $-\gamma_1$ is smaller than or greater than $\gamma_1 - \gamma_2$, see also Figure 6.6.

For future reference, we remark that as a result of these conditions on $\boldsymbol{\alpha}_2$, we find using (6.19), $-\gamma_1 \leq \alpha_{2,1} \leq y$ and $\alpha_{2,2} \geq -\gamma_2$

$$\rho_2(\mathbf{x})\mathbb{H}(\mathbf{x},\boldsymbol{\alpha}_2) \geq \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0, \\ -\rho_2(\mathbf{x})\log(\mu_1 + (\mu_1\mu_2/\lambda)e^y + \mu_2) & \text{if } x_1 > 0,\ x_2 = 0, \\ 0 & \text{if } x_1 = 0,\ x_2 > 0, \end{cases} \tag{6.74}$$

since $\lambda + \mu_1 + \mu_2 = 1$.

6.3.4.4 Summary and proof that all conditions are satisfied

To summarize, we have found the following values for $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$ and $\boldsymbol{\alpha}_4$ that intuitively satisfy all conditions for an asymptotically efficient change of measure based on Theorem 6.1, see Table 6.6.

Table 6.6 Possibilities for $\boldsymbol{\alpha}_k$ when queue 1 is the bottleneck queue, provided that $d_1 < d_2 < d_3 < d_4$.

	$\alpha_{k,1}$	$\alpha_{k,2}$	k	Conditions
$x_1 > 0,\ x_2 > 0$	$-\gamma_1$	$[-\gamma_2, -\gamma_1]$	1	
$x_1 = 0,\ x_2 > 0$	$[-\gamma_1, y]$	$[-\gamma_2, \alpha_{1,2}]$	2	(6.72)
				$\alpha_{2,1} + \alpha_{2,2} \geq -\gamma_1 + \alpha_{1,2}$
$x_1 > 0,\ x_2 = 0$	$-\gamma_1$	0	3	
$x_1 \geq 0,\ x_2 \geq 0$	0	0	4	

Again we show by considering all conditions in that theorem, indeed these possibilities for $\boldsymbol{\alpha}_k$, $k = 1, 2, 3, 4$, give an asymptotically efficient change of measure. We start with Condition 1, and recall that $\mathbb{H}(\mathbf{x}, \boldsymbol{\alpha}_4) = 0$. Then, we find that, using the result from (6.39) for $k = 3$, (6.52), (6.53), (6.71), (6.74), (6.60) and (6.61),

$$\begin{aligned} &\sum_{k=1}^{4} \rho_k(\mathbf{x})\mathbb{H}(\mathbf{x},\boldsymbol{\alpha}_k) \\ &\geq \begin{cases} 0 & \text{if } x_1 > 0,\ x_2 > 0 \\ -e^{(d_1-d_3)\delta/\varepsilon}\log(\mu_1 + 2\mu_2) & \\ \quad -e^{(d_2-d_3)\delta/\varepsilon}\log(\mu_1 + (\mu_1\mu_2/\lambda)e^y + \mu_2) & \text{if } x_1 > 0,\ x_2 = 0 \\ -e^{(d_1-d_2)\delta/\varepsilon}\log(2\mu_1 + \mu_2\lambda/\mu_1) & \\ \quad -e^{(d_3-d_4)\delta/\varepsilon}\log(2\mu_1 + \mu_2) & \text{if } x_1 = 0,\ x_2 > 0 \end{cases} \\ &\geq -2e^{\max\{(d_1-d_2),(d_2-d_3),(d_3-d_4)\}\delta/\varepsilon}\log(\max\{3\mu_2, 2\mu_2 + \mu_2^2 e^y/\lambda\}), \end{aligned}$$

where the final inequality follows since queue 1 is the bottleneck queue, and thus $\mu_1 \leq \mu_2$. It follows that Condition 1 is satisfied. For Condition 2 we note that

for all $\mathbf{x} \in \delta_e$ we have $W_1^\delta(\mathbf{x}) \leq -d_1\delta$, and thus Condition 2 is satisfied since $W^{\varepsilon,\delta}(\mathbf{x}) \leq W_1^\delta(\mathbf{x})$. Condition 3 follows similarly as in (6.38), with γ_2 replaced by γ_1, since queue 1 is the bottleneck queue. Therefore, the change of measure for the given possible values of $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, $\boldsymbol{\alpha}_3$, $\boldsymbol{\alpha}_4$ and d_k, $k = 1, 2, 3, 4$, is asymptotically efficient.

6.3.4.5 Discussion

From Table 6.6, we see that for $x_1 > 0$ and $x_2 = 0$ the change of measure from [35] is used, as expected. The difference compared to using three regions in Section 6.3.2 is that there is a possibility to apply a slightly different change of measure for $x_1 = 0$ and $x_2 > 0$. In Figure 6.10 we show an example when choosing $\boldsymbol{\alpha}_1 = (-\gamma_1, -\gamma_1)$, $\boldsymbol{\alpha}_2 = (0, -\gamma_1)$ and $\boldsymbol{\alpha}_3 = (-\gamma_1, 0)$. These values for $\boldsymbol{\alpha}_2$ can only occur when $-\gamma_1 \geq \gamma_1 - \gamma_2$.

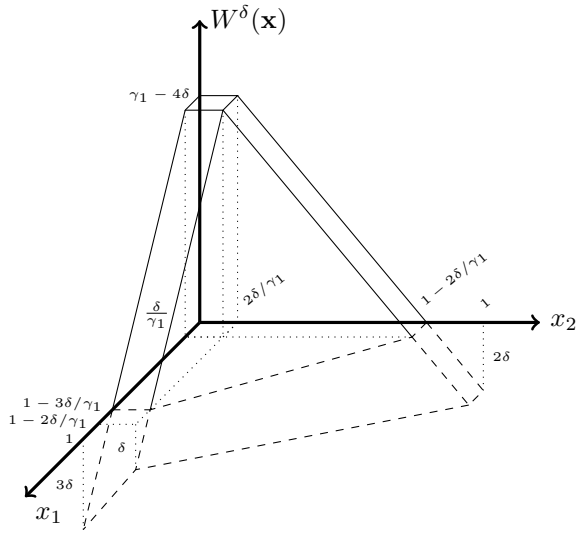


Figure 6.10 Display of $W^\delta(\mathbf{x})$ when queue 1 is the bottleneck queue, $\boldsymbol{\alpha}_1 = (-\gamma_1, -\gamma_1)$, $\boldsymbol{\alpha}_2 = (0, -\gamma_1)$ and $d_k = k$, $k = 1, 2, 3, 4$.

Comparing with Figure 6.9, we see that in the area close to the x_2-axis, we can apply a slightly different change of measure. Of course, there are several other possibilities.

6.4 Conclusions

In this chapter, we determined sufficient conditions for subsolution-based changes of measure to give asymptotically efficient estimators. As a result, for the 2-node $M|M|1$ tandem queue we explicitly gave a whole family (continuum) of changes

of measure that all lead to asymptotically efficient estimators, and the previously known changes of measure are just three members of this family. For d-node tandem queues, it seems likely that we can use a similar analysis to find a family of changes of measure that are asymptotically efficient.

For the case $d = 2$ we like to highlight one particular change of measure based on the subsolution $W^{\varepsilon,\delta}(\mathbf{x})$ in (6.6) (via either (6.3) or (6.9)), that uses the following three functions:

$$\begin{aligned} W_1^\delta(\mathbf{x}) &= -\gamma x_1 - \gamma x_2 + \gamma - \delta, \\ W_2^\delta(\mathbf{x}) &= -\gamma x_1 \qquad\quad + \gamma - 2\delta, \\ W_3^\delta(\mathbf{x}) &= \qquad\qquad\quad \gamma - 3\delta, \end{aligned}$$

where we recall that $\gamma = \min\{\gamma_1, \gamma_2\}$. We note that this is the same subsolution as in [19] if queue 2 is the bottleneck queue, while it also works when queue 1 is the bottleneck queue (with γ_2 replaced by γ_1). This matches nicely with the known fact that interchanging the queues leaves our probability of interest unchanged.

From an implementation point of view, in general it makes sense to use as few regions as possible, that is, 3 regions (or $d+1$ in the d-node case). However, when the event of interest is not total buffer overflow but for example individual buffer overflow or simultaneous buffer overflows, it may be more useful to implement the change of measure from [22] that is based on four regions.

Finally, we mention that future work could aim at investigating whether or not the method generalizes to more general models, which we expect to be the case. In Chapter 5 something similar has already been done for non-Markovian tandem queues, but one could also think about general Jackson networks as in [22], for which we expect similar results to hold (though more cumbersome to obtain), or even about more general (non-Markovian) networks.

CHAPTER 7

Splitting for non-Markovian tandem queues

In Chapter 5, we have used subsolutions in order to determine a state-dependent change of measure for non-Markovian tandem queues that gives an asymptotically efficient estimator. As was mentioned in the introduction of this thesis, subsolutions can not only be used to determine a change of measure that results in an asymptotically efficient estimator, they can also be used to determine splitting thresholds resulting in an asymptotically efficient estimator. This is the topic of the current chapter.

For the 2-node Markovian tandem queue this has explicitly been done in [16] for ordinary splitting and in [15] for RESTART. In both papers, splitting thresholds based on subsolutions have been proven to result in an asymptotically efficient estimator. In addition, in [16], it is remarked that the analysis is not limited to Markovian models, but it is assumed that the decay rate for the probability of interest, starting at some general point in the state space, is known. For d-node non-Markovian tandem queues this decay rate is not known.

Therefore, in this chapter, we first determine the decay rate for the probability of interest, starting at some general point in the state space in Section 7.2. This result is necessary in order to prove that we get an asymptotically efficient estimator by using the splitting scheme based on subsolutions, see Section 7.3. In addition, we also present a splitting scheme based on the decay rate and we prove that we indeed get an asymptotically efficient estimator for both splitting schemes. We conclude this chapter with some numerical results in Section 7.4.

7.1 Model and preliminaries

In this chapter, we again consider the d-node $GI|GI|1$ tandem queue, see Section 2.1 for a detailed introduction of the model, notation and assumptions, and we are again interested in the probability of having more than N customers in the system during a busy cycle. In addition, we refer to Chapter 5 for an introduction to the scaled state description, see Section 5.1, an introduction to subsolutions, see Definition 5.4, and the construction of subsolutions, see Section 5.2. In order to use a subsolution to determine the splitting thresholds, we do not need to

have a continuously differentiable subsolution and so we do not need to apply the mollification procedure from that section.

We again let $\mathbf{x} = (x_1, \ldots, x_d, \bar{x}_0, \ldots, \bar{x}_d)$ denote a state in the scaled system, where we recall that $x_1, \ldots, x_d$ are the scaled number of customers in the system and $\bar{x}_0, \ldots \bar{x}_d$ are the residual inter-arrival times and service times in queue $1, \ldots, d$ respectively. We assume that at time 0, we start in some particular state $\mathbf{x}$. We let A_k denote the inter-arrival time between the k^{th} and $(k+1)^{st}$ *arriving customer* after time 0. That is, the first arriving customer enters the system at time $\bar{x}_0 N$, which is already fixed, and hence A_1 denotes the inter-arrival time between the first and second arriving customer. We define $B_k^{(j)}$ as the service time at queue j of the k^{th} *departing customer from the system* and D_k as the inter-departure time between the $(k-1)^{th}$ and k^{th} *departing* customer for all $k \geq 2$. We conveniently define $D_1 = \sum_{j>i} B_1^{(j)} + \bar{x}_i N$, where i is the index of the last non-empty queue at time 0. That is, $x_i > 0$ and $x_j = 0$ for all queues $j > i$. By definition, it can either happen that D_1 is stochastic, when $x_d = 0$, or deterministic, when $x_d > 0$.

The most notable difference between the notation in Chapter 2 and the current chapter, is that we now consider the k^{th} arriving and k^{th} departing customer, which do not necessarily consider the same customer, while in Chapter 2 the k^{th} arriving and k^{th} departing customer *are* the same customer. As in Chapter 5, we mostly assume that the distributions of the inter-arrival times and service times at all queues j have a bounded support, see Assumption 5.1.

We again use stopping times K_N and K_0. The mathematical definition of the stopping times K_N and K_0 as in (2.1) and (2.2) slightly changes when we start from some state $\mathbf{x} \neq \mathbf{0}$. To this end, we define $K_N(\mathbf{x})$ and $K_0(\mathbf{x})$ as

$$K_N(\mathbf{x}) = \min\Bigg\{ n \geq (1 - x_1 - \ldots - x_d)N : \sum_{k=1}^{n-1} A_k + \bar{x}_0 N < \sum_{k=2}^{n-(1-x_1-\ldots-x_d)N+1} D_k + D_1 \Bigg\}, \tag{7.1}$$

$$K_0(\mathbf{x}) = \min\Bigg\{ m \geq 1 : \sum_{k=1}^{m-1} A_k + \bar{x}_0 N > \sum_{k=2}^{m-1+(x_1+\ldots+x_d)N} D_k + D_1 \Bigg\}, \tag{7.2}$$

and we define $\mathcal{K}(\mathbf{x}) = \min\{K_0(\mathbf{x}), K_N(\mathbf{x})\}$ analogously to $\mathcal{K}$ in Chapter 4. Note that the definitions in (7.1) and (7.2) reduce to (2.1) and (2.2) when $\mathbf{x} = \mathbf{0}$.

Lastly, recall from Chapter 2 that $\Lambda_X(\theta)$ denotes the log moment generating function of some random variable X and that in and below Equation (2.3) we have defined $\theta^{(j)} = \sup\Big\{\theta : \Lambda_A(-\theta) + \Lambda_{B^{(j)}}(\theta) \leq 0\Big\}$ and $\theta^* = \min_j(\theta^{(j)})$ respectively.

7.2 Decay rates from general starting point

In order to determine the decay rate when starting at some general point in the (scaled) state space, we use a similar approach as in Chapter 2. We start with the decay rate for the single queue, where we derive both an upper and a lower bound on the decay rate and show that they coincide. We provide a conjecture on the decay rate when supports are unbounded. Afterwards, we use the decay rate for the single queue in order to obtain a lower bound on the decay rate for the tandem queue, similar as in Chapter 2. For the upper bound on the decay rate we provide a conjecture, along with some intuition.

7.2.1 The single $GI|GI|1$ queue

For the single queue, (7.1) and (7.2) reduce to

$$K_N(\mathbf{x}) = \min\left\{ n \geq (1-x_1)N : \sum_{k=1}^{n-1} A_k + \bar{x}_0 N < \sum_{k=2}^{n-(1-x_1)N+1} B_k^{(1)} + \bar{x}_1 N \right\}, \tag{7.3}$$

$$K_0(\mathbf{x}) = \min\left\{ m \geq 1 : \sum_{k=1}^{m-1} A_k + \bar{x}_0 N > \sum_{k=2}^{m-1+x_1 N} B_k^{(1)} + \bar{x}_1 N \right\}, \tag{7.4}$$

since in that case $D_k = B_k^{(1)}$ and $D_1 = \bar{x}_1 N$. Recall that the latter is only true when $x_1 > 0$, though we can assume without loss of generality that we do not start from an empty system (in which case the decay rate can be found in [36]). In the remainder of this section, when considering a single queue, we drop the superscript (1) for convenience. We start with an upper bound on the (negative) decay rate, where we follow the same steps as in Theorem 1 of [36]. In that paper, the decay rate for the single queue, starting from $\mathbf{x} = \mathbf{0}$ has been determined. It turns out that method in that paper can be extended to a more general starting point.

Lemma 7.1 (Upper bound for bounded support)**.** *Consider a single $GI|GI|1$ queue satisfying Assumptions 2.3 and 5.1. For the decay of $\mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x}))$ it holds that,*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) \leq (1-x_1)\Lambda_A(-\theta^*),$$

uniformly in $\mathbf{x}$.

Proof. We follow the same steps as in Theorem 1 of [36] and adjust where necessary. For readability, we show most intermediate steps. Let $\mathcal{F}_k = \sigma(A_1, \ldots, A_k, B_2, \ldots, B_{k+1})$ be a sigma algebra, and note that $K_N(\mathbf{x})$ and $K_0(\mathbf{x})$ are $\mathcal{F}_{k-1}$ and $\mathcal{F}_{k+x_1N-2}$ stopping times respectively. As a result, $\mathcal{K}(\mathbf{x})$ is an $\mathcal{F}_{k+x_1N-2}$ stopping time.

Similar to [36], we need to show that

$$\{\mathcal{K}(\mathbf{x}) = K_N(\mathbf{x})\} \in \sigma(A_1, \ldots, A_{K_N(\mathbf{x})-1}, B_2, \ldots, B_{K_N(\mathbf{x})-(1-x_1)N+1}), \quad (7.5)$$

which says that *both* $\mathcal{K}(\mathbf{x})$ and $K_N(\mathbf{x})$ have the same sigma algebra. To show that (7.5) indeed holds, we remark the following. On $\{K_N(\mathbf{x}) = j\} \in \sigma(A_1, \ldots, A_{j-1}, B_2, \ldots B_{j-(1-x_1)N+1}) \subset \mathcal{F}_{j-1}$ it holds that

$$\sum_{k=1}^{j-1} A_k + \bar{x}_0 N < \sum_{k=2}^{j-(1-x_1)N+1} B_k + \bar{x}_1 N,$$

and thus also that

$$\sum_{k=1}^{\bar{j}-1} A_k + \bar{x}_0 N < \sum_{k=2}^{\bar{j}+x_1 N} B_k + \bar{x}_1 N,$$

for $\bar{j} \in [j - N + 1, j]$ and hence $K_0(\mathbf{x}) \in [j - N + 1, j]$ is not possible. Therefore, given the event $\{K_N(\mathbf{x}) = j\}$, to determine the event $\{\mathcal{K}(\mathbf{x}) = K_N(\mathbf{x}) = j\}$, we only have to check whether $K_0(\mathbf{x}) < j - N + 1$. That is, given $\{K_N(\mathbf{x}) = j\}$, to determine whether $\{\mathcal{K}(\mathbf{x}) = K_N(\mathbf{x}) = j\}$, we only need $A_1, \ldots, A_{j-N}, B_2, \ldots, B_{j-(1-x_1)N+1}$ to check if $K_0(\mathbf{x}) < j - N + 1$. Thus, (7.5) holds and as a result the event $\{\mathcal{K}(\mathbf{x}) = K_N(\mathbf{x})\}$ is independent of the inter-arrival times $\{A_k : K_N(\mathbf{x}) \leq k \leq K_N(\mathbf{x}) + x_1 N - 2\}$ and the service times $\{B_k : K_N(\mathbf{x}) - (1 - x_1)N + 1 < k \leq K_N(\mathbf{x}) + x_1 N - 1\}$.

Next, we let

$$S_k = \sum_{j=1}^{k} (B_{j+1} - A_j). \quad (7.6)$$

As noted before, $\mathcal{K}(\mathbf{x})$ is a $\mathcal{F}_{k+x_1 N-2}$-stopping time, and thus we have, for all θ such that $\Lambda(\theta) = \Lambda_A(-\theta) + \Lambda_B(\theta) < \infty$, Wald's identity, see also [1],

$$\mathbb{E}\left[e^{\theta S_{\mathcal{K}(\mathbf{x})+x_1 N-2} - (\mathcal{K}(\mathbf{x})+x_1 N-2)\Lambda(\theta)}\right] = \mathbb{P}^\theta(\mathcal{K}(\mathbf{x}) + x_1 N - 2 < \infty) = 1. \quad (7.7)$$

The second equality is due to the strong law of large numbers: we either have $\mathbb{P}^\theta(K_N(\mathbf{x}) + x_1 N - 2 < \infty) = 1$ or $\mathbb{P}^\theta(K_0(\mathbf{x}) + x_1 N - 2 < \infty) = 1$, depending on the sign of $\frac{d\Lambda}{d\theta}(\theta) = \mathbb{E}^\theta[B] - \mathbb{E}^\theta[A]$, and hence we also have $\mathbb{P}^\theta(\mathcal{K}(\mathbf{x}) + x_1 N - 2 < \infty) = 1$ by definition.

By Assumption 2.3 we have $\Lambda'(0) = \mathbb{E}[B] - \mathbb{E}[A] < 0$, so we can fix any $\theta \in (0, \theta^*)$ such that $\Lambda(\theta) < 0$. For all such θ, we have from (7.7)

$$\begin{aligned} 1 &= \mathbb{E}\left[e^{\theta S_{\mathcal{K}(\mathbf{x})+x_1 N-2} - (\mathcal{K}(\mathbf{x})+x_1 N-2)\Lambda(\theta)}\right] \\ &\geq \mathbb{E}\left[e^{\theta S_{K_N(\mathbf{x})+x_1 N-2}} \mid K_N(\mathbf{x}) < K_0(\mathbf{x})\right] e^{-(N-2)\Lambda(\theta)} \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})), \quad (7.8) \end{aligned}$$

where the inequality follows since $K_N(\mathbf{x}) + x_1 N - 2 \geq N - 2$. Furthermore, using (7.6) and the definition of $K_N(\mathbf{x})$, see (7.3), we find

$$
\begin{aligned}
S_{K_N(\mathbf{x})+x_1N-2} &= \sum_{j=1}^{K_N(\mathbf{x})+x_1N-2} (B_{j+1} - A_j) \\
&= \sum_{j=K_N(\mathbf{x})-(1-x_1)N+1}^{K_N(\mathbf{x})+x_1N-2} B_{j+1} - \sum_{j=K_N(\mathbf{x})}^{K_N(\mathbf{x})+x_1N-2} A_j + \sum_{j=1}^{K_N(\mathbf{x})-(1-x_1)N} B_{j+1} - \sum_{j=1}^{K_N(\mathbf{x})-1} A_j \\
&> \sum_{j=K_N(\mathbf{x})-(1-x_1)N+1}^{K_N(\mathbf{x})+x_1N-2} B_{j+1} - \sum_{j=K_N(\mathbf{x})}^{K_N(\mathbf{x})+x_1N-2} A_j - (\bar{x}_1 - \bar{x}_0)N, \qquad (7.9)
\end{aligned}
$$

and so we find by substituting (7.9) in (7.8) and the fact that

$$
\begin{aligned}
&\mathbb{E}\left[e^{\theta\left(\sum_{j=K_N(\mathbf{x})-(1-x_1)N+1}^{K_N(\mathbf{x})+x_1N-2} B_{j+1} - \sum_{j=K_N(\mathbf{x})}^{K_N(\mathbf{x})+x_1N-2} A_j - (\bar{x}_1-\bar{x}_0)N\right)} \middle| K_N(\mathbf{x}) < K_0(\mathbf{x}) \right] \\
&= e^{(N-2)\Lambda_B(\theta) + x_1N\Lambda_A(-\theta) - (\bar{x}_1-\bar{x}_0)N\theta},
\end{aligned}
$$

due to independence of $\{B_k : K_N(\mathbf{x}) - (1-x_1)N + 1 < k \leq K_N(\mathbf{x}) + x_1N - 1\}$ and $\{A_k : K_N(\mathbf{x}) \leq k \leq K_N(\mathbf{x}) + x_1N - 2\}$ of the event $\{\mathcal{K} = K_N(\mathbf{x})\}$, and $\Lambda(\theta) = \Lambda_A(-\theta) + \Lambda_B(\theta)$, that

$$
1 \geq e^{-\Lambda_A(-\theta)((1-x_1)N-2) - (\bar{x}_1-\bar{x}_0)N\theta} \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})). \qquad (7.10)
$$

From (7.10) it follows that

$$
\begin{aligned}
&\frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) && (7.11) \\
&\leq (1-x_1)\Lambda_A(-\theta) - \frac{2\Lambda_A(-\theta)}{N} + (\bar{x}_1 - \bar{x}_0)\theta && (7.12) \\
&\leq (1-x_1)\Lambda_A(-\theta) + \frac{Q^{(1)}\theta - 2\Lambda_A(-\theta)}{N}, && (7.13)
\end{aligned}
$$

where the final inequality follows by Assumption 5.1. Uniform convergence in $\mathbf{x}$ of (7.11) follows from (7.13), since given any $\varepsilon > 0$ there exists N_1 that depends *only* on ε, so that we can choose any $N \geq N_1$ such that for all $\mathbf{x}$

$$
\frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) \leq \Lambda_A(-\theta)(1-x_1) + \varepsilon.
$$

Thus, we find

$$
\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) \leq (1-x_1)\Lambda_A(-\theta),
$$

and taking $\theta \to \theta^*$ we obtain the tightest upper bound. This concludes the proof. □

The next lemma presents an upper bound when the supports are unbounded, which follows by extending the analysis in Lemma 7.1. Similarly to (4.6), we let $\mathbb{E}^{\theta}[A] = \frac{-d\Lambda_A}{d\theta}(-\theta)$. Remark that the only difference compared to (4.6) is that we do not use the vector notation here, since it is not necessary in this chapter.

Lemma 7.2 (Upper bound for unbounded support)**.** *Consider a single GI|GI|1 queue satisfying Assumption 2.3. For the decay of* $\mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x}))$ *it holds that,*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) \leq \begin{cases} 0 & \text{if } \bar{x}_1 - \bar{x}_0 \geq (1-x_1)\mathbb{E}[A], \\ \min_{\theta\in[0,\theta^*]}(1-x_1)\Lambda_A(-\theta) + (\bar{x}_1 - \bar{x}_0)\theta & \text{if } \frac{\bar{x}_1-\bar{x}_0}{1-x_1} \in (\mathbb{E}^{\theta^*}[A], \mathbb{E}[A]), \\ (1-x_1)\Lambda_A(-\theta^*) + (\bar{x}_1 - \bar{x}_0)\theta^* & \text{if } \bar{x}_1 - \bar{x}_0 \leq (1-x_1)\mathbb{E}^{\theta^*}[A], \end{cases}$$

uniformly in $\mathbf{x}$.

Proof. We follow the proof of Lemma 7.1 until (7.12). To determine the tightest upper bound of (7.12), when we let $N \to \infty$ and when supports are unbounded, note that $\Lambda_A(-\theta)$ is a decreasing function and whether $(\bar{x}_1 - \bar{x}_0)\theta$ is increasing or decreasing depends on the sign of $\bar{x}_1 - \bar{x}_0$. If $\bar{x}_1 \leq \bar{x}_0$, then it is non-increasing, and hence the tightest upper bound of (7.12) is attaind at $\theta \to \theta^*$. Otherwise, when $\bar{x}_1 > \bar{x}_0$, we take derivatives to determine the tightest upper bound. Let

$$f(\theta) = (1-x_1)\Lambda_A(-\theta) + (\bar{x}_1 - \bar{x}_0)\theta,$$

and hence

$$\frac{df(\theta)}{d\theta} = -(1-x_1)\mathbb{E}^{\theta}[A] + \bar{x}_1 - \bar{x}_0,$$

which equals 0 when $\mathbb{E}^{\theta}[A] = \frac{\bar{x}_1-\bar{x}_0}{1-x_1}$. Recall that $\theta \in (0,\theta^*)$ and thus $\mathbb{E}^{\theta}[A] \in (\mathbb{E}^{\theta^*}[A], \mathbb{E}[A])$. Taking this into account, we arrive at

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) \leq \begin{cases} (1-x_1)\Lambda_A(-\theta^*) + (\bar{x}_1 - \bar{x}_0)\theta^* & \text{if } \bar{x}_1 \leq \bar{x}_0, \\ \inf_{\theta\in(0,\theta^*)} f(\theta) & \text{if } \bar{x}_1 > \bar{x}_0, \end{cases}$$

where

$$\inf_{\theta\in(0,\theta^*)} f(\theta) = \begin{cases} f(0) = 0 & \text{if } \bar{x}_1 - \bar{x}_0 \geq (1-x_1)\mathbb{E}[A], \\ \min_{\theta\in(0,\theta^*)}(1-x_1)\Lambda_A(-\theta) + (\bar{x}_1 - \bar{x}_0)\theta & \text{if } \frac{\bar{x}_1-\bar{x}_0}{1-x_1} \in (\mathbb{E}^{\theta^*}[A], \mathbb{E}[A]), \\ f(\theta^*) = (1-x_1)\Lambda_A(-\theta^*) + (\bar{x}_1 - \bar{x}_0)\theta^* & \text{if } \bar{x}_1 - \bar{x}_0 \leq (1-x_1)\mathbb{E}^{\theta^*}[A]. \end{cases}$$

Again, the convergence is uniform in the starting state $\mathbf{x}$. This concludes the proof. □

Remark 7.3. *The result in Lemma 7.2 gives the same result for bounded supports as in Lemma 7.1, since* $\bar{x}_0$ *and* $\bar{x}_1$ *tend to* 0 *as* $N \to \infty$.

Next, we prove a lower bound for the decay with bounded support. Also the analysis in this proof is an extension of the proof of Theorem 1 from [36].

Lemma 7.4 (Lower bound for bounded support)**.** *Consider a single GI|GI|1 queue satisfying Assumptions 2.3 and 5.1. For the decay of* $\mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x}))$ *it holds that,*

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) \geq (1-x_1)\Lambda_A(-\theta^*),$$

uniformly in $\mathbf{x}$.

Proof. By Assumption 2.3 we have $\theta^* < \infty$, otherwise $\mathbb{P}(B > A) = 0$, and by Assumption 5.1 we have $B \leq Q^{(1)}$ with probability 1 for some constant $Q^{(1)}$. Also in this proof we use Wald's identity as in (7.7), though now our goal is to find a lower bound on the probability of interest. Hence, we need to upper bound $S_{K_N(\mathbf{x})+x_1N-2}$ and we find, by using (7.6) and the definition of $K_N(\mathbf{x})$ in (7.3),

$$\begin{aligned}
S_{K_N(\mathbf{x})+x_1N-2} &= \sum_{j=1}^{K_N(\mathbf{x})+x_1N-2} (B_{j+1} - A_j) \\
&= \sum_{j=K_N(\mathbf{x})-(1-x_1)N+1}^{K_N(\mathbf{x})+x_1N-2} B_{j+1} - \sum_{j=K_N(\mathbf{x})}^{K_N(\mathbf{x})+x_1N-2} A_j \\
&\quad + \sum_{j=1}^{K_N(\mathbf{x})-(1-x_1)N-1} B_{j+1} - \sum_{j=1}^{K_N(\mathbf{x})-2} A_j + B_{K_N(\mathbf{x})-(1-x_1)N+1} - A_{K_N(\mathbf{x})-1} \\
&\leq \sum_{j=K_N(\mathbf{x})-(1-x_1)N+1}^{K_N(\mathbf{x})+x_1N-2} B_{j+1} - \sum_{j=K_N(\mathbf{x})}^{K_N(\mathbf{x})+x_1N-2} A_j + (\bar{x}_0 - \bar{x}_1)N + Q^{(1)}, \qquad (7.14)
\end{aligned}$$

where the inequality follows since the inequality in (7.3) is reversed for $n = K_N(\mathbf{x}) - 1$ and since $B_{K_N(\mathbf{x})-(1-x_1)N+1} - A_{K_N(\mathbf{x})-1} \leq Q^{(1)}$ by assumption. Remark that also for the boundary case $K_N(\mathbf{x}) = (1-x_1)N$ we can use the bound in (7.14) since clearly $-\sum_{j=1}^{K_N(\mathbf{x})-1} A_j \leq 0$ and $Q^{(1)} \geq (\bar{x}_1 - \bar{x}_0)N$.

Due to Assumption 5.1, that considers bounded supports, it follows that

$\Lambda(\theta^*) = 0$. Thus, choosing $\theta = \theta^*$ we have from (7.7) and (7.14) that

$$\begin{aligned}
1 &= \mathbb{E}\left[e^{\theta^* S_{\mathcal{K}(\mathbf{x})+x_1N-2} - (\mathcal{K}(\mathbf{x})+x_1N-2)\Lambda(\theta^*)}\right] \\
&= \mathbb{E}\left[e^{\theta^* S_{K_N(\mathbf{x})+x_1N-2}} \mid K_N(\mathbf{x}) < K_0(\mathbf{x})\right] \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) \\
&\quad + \mathbb{E}\left[e^{\theta^* S_{K_0(\mathbf{x})+x_1N-2}} \mathbb{1}\{K_0(\mathbf{x}) < K_N(\mathbf{x})\}\right] \\
&< e^{\theta^*(Q^{(1)}+(\bar{x}_0-\bar{x}_1)N)} \\
&\quad \cdot \mathbb{E}\left[e^{\theta^* \sum_{j=K_N(\mathbf{x})-(1-x_1)N+1}^{K_N(\mathbf{x})+x_1N-2} B_{j+1} - \theta^* \sum_{j=K_N(\mathbf{x})}^{K_N(\mathbf{x})+x_1N-2} A_j} \mid K_N(\mathbf{x}) < K_0(\mathbf{x})\right] \\
&\quad \cdot \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) + \mathbb{E}\left[e^{\theta^* S_{K_0(\mathbf{x})+x_1N-2}}\right].
\end{aligned}$$

Using the independence of $\{A_k : K_N(\mathbf{x}) \le k \le K_N(\mathbf{x}) + x_1N - 2\}$ and $\{B_k : K_N(\mathbf{x}) - (1-x_1)N + 1 < k \le K_N(\mathbf{x}) + x_1N - 1\}$ of the event $\{\mathcal{K}(\mathbf{x}) = K_N(\mathbf{x})\}$, see proof of Lemma 7.1, we find that the last display is equivalent to

$$\begin{aligned}
&\mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) \\
&> \left(1 - \mathbb{E}\left[e^{\theta^* S_{K_0(\mathbf{x})+x_1N-2}}\right]\right) e^{-\theta^*(Q^{(1)}+(\bar{x}_0-\bar{x}_1)N) - (N-2)\Lambda_B(\theta^*) - (x_1N-1)\Lambda_A(-\theta^*)}.
\end{aligned} \tag{7.15}$$

It remains to show that $\mathbb{E}\left[e^{\theta^* S_{K_0(\mathbf{x})+x_1N-2}}\right] < 1$. By (7.4) and (7.6) we have,

$$S_{K_0(\mathbf{x})+x_1N-2} = \sum_{j=1}^{K_0(\mathbf{x})+x_1N-2} (B_{j+1} - A_j) \le - \sum_{j=K_0(\mathbf{x})}^{K_0(\mathbf{x})+x_1N-2} A_j + (\bar{x}_0 - \bar{x}_1)N.$$

Since $\{A_k : K_0(\mathbf{x}) \le k \le K_0(\mathbf{x}) + x_1N - 2\}$ does not depend on the event $\{K_0(\mathbf{x}) = k\}$, we find

$$\begin{aligned}
\mathbb{E}\left[e^{\theta^* S_{K_0(\mathbf{x})+x_1N-2}}\right] &\le \mathbb{E}\left[e^{-\theta^* \sum_{j=K_0(\mathbf{x})}^{K_0(\mathbf{x})+x_1N-2} A_j + \theta^*(\bar{x}_0-\bar{x}_1)N}\right] \\
&\le \mathbb{E}\left[e^{\Lambda_A(-\theta^*)(x_1N-1) + \theta^*(\bar{x}_0-\bar{x}_1)N}\right],
\end{aligned}$$

and by convexity of $\Lambda_A(-\theta^*)$ we find that $\Lambda_A(-\theta^*) \le -\theta^* \mathbb{E}^{\theta^*}[A]$. Thus, we find, provided that $\bar{x}_0 - \bar{x}_1 < \mathbb{E}^{\theta^*}[A]\,(x_1 - \frac{1}{N})$

$$\mathbb{E}\left[e^{\theta^* S_{K_0(\mathbf{x})+x_1N-2}}\right] < 1.$$

Clearly, for large enough N the bound on $\bar{x}_0 - \bar{x}_1$ holds under Assumption 5.1 for fixed x_1. Therefore, we find from (7.15) that

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) \ge (1 - x_1)\Lambda_A(-\theta^*),$$

since $\bar{x}_0 N \le Q^{(0)}$ and $\Lambda_A(-\theta^*) + \Lambda_B(\theta^*) = 0$. As in the proof of Lemma 7.1, we find that the convergence is uniform in $\mathbf{x}$. This concludes the proof. □

As a result of Lemmas 7.1 and 7.4, we immediately find the following.

Theorem 7.5 (Decay rate for bounded support)**.** *Consider a single GI|GI|1 queue satisfying Assumptions 2.3 and 5.1. For the decay of* $\mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x}))$ *it holds that,*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) = (1-x_1)\Lambda_A(-\theta^*),$$

uniformly in $\mathbf{x}$*.*

We expect that the upper bound of the decay rate in Lemma 7.2 is tight for several starting points $\mathbf{x}$, this statement can be found in Conjecture 7.6. An intuition is provided below.

Conjecture 7.6 (Decay rate for unbounded support)**.** *Consider a single GI|GI|1 queue satisfying Assumption 2.3. For the decay of* $\mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x}))$*, with* $\mathbf{x}$ *such that* $\bar{x}_1 - \bar{x}_0 > -x_1\mathbb{E}^{\theta^*}\left[B^{(1)}\right]$*, it holds that,*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) = \begin{cases} 0 & \text{if } \bar{x}_1 - \bar{x}_0 \geq (1-x_1)\mathbb{E}\left[A\right], \\ \min_{\theta\in(0,\theta^*)}(1-x_1)\Lambda_A(-\theta) + (\bar{x}_1-\bar{x}_0)\theta & \text{if } \frac{\bar{x}_1-\bar{x}_0}{1-x_1} \in (\mathbb{E}^{\theta^*}\left[A\right], \mathbb{E}\left[A\right]), \\ (1-x_1)\Lambda_A(-\theta^*) + (\bar{x}_1-\bar{x}_0)\theta^* & \\ & \text{if } -x_1\mathbb{E}^{\theta^*}\left[B^{(1)}\right] < \bar{x}_1 - \bar{x}_0 \leq (1-x_1)\mathbb{E}^{\theta^*}\left[A\right], \end{cases}$$

uniformly in $\mathbf{x}$*.*

Intuition. The most important observation is that we have already shown that the decay rate in the statement is an upper bound on the decay rate, see Lemma 7.2. We give some intuition case by case, why we expect this upper bound to be tight for all considered starting points $\mathbf{x}$. In Figure 7.1, we give a graphical illustration of all starting points $\mathbf{x}$ and their corresponding decay rate. Starting with all starting states in area I, $(\bar{x}_1-\bar{x}_0)N$ is so large that on average at least $(1-x_1)N$ arrivals happen before the first customer departs and thus we reach the overflow level. Therefore, the decay of the probability to reach the overflow level equals 0. Area III considers $\bar{x}_1 - \bar{x}_0 \leq (1-x_1)\mathbb{E}^{\theta^*}[A]$, which in particular covers all starting states that satisfy Assumption 5.1, that is, $\bar{x}_0 = \bar{x}_1 = 0$ (since for bounded supports $\bar{x}_0$ and $\bar{x}_1$ tend to zero as $N \to \infty$). We can interpret $\mathbb{E}^{\theta^*}[A]$ as the expected inter-arrival times under the change of measure θ^*, see also Chapter 4. Even though we do not consider importance sampling here, it turns out that this term is important in the sense that when $\bar{x}_1 - \bar{x}_0 \leq (1-x_1)\mathbb{E}^{\theta^*}[A]$, the decay rate equals $(1-x_1)\Lambda_A(-\theta^*) + (\bar{x}_1-\bar{x}_0)\theta^*$. In addition, area III covers states where a lot of customers depart, without reaching an empty system, prior to the first arriving customer. In that case, the decay rate should be very negative (which it is). Area II, where $\bar{x}_1 - \bar{x}_0 \in ((1-x_1)\mathbb{E}^{\theta^*}\left[A\right], (1-x_1)\mathbb{E}\left[A\right])$, considers a decay rate that is somewhere in between the previous two cases.

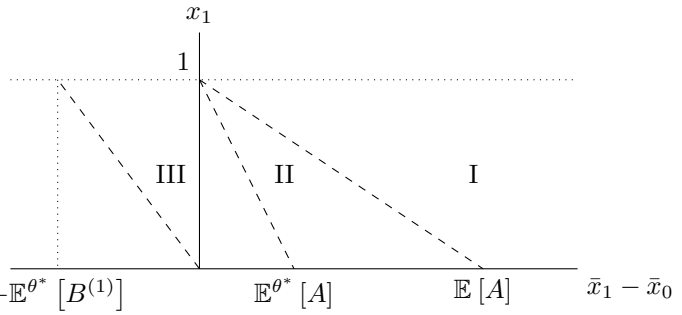


Figure 7.1 A graphical illustration of all possible starting states $\mathbf{x}$ and their corresponding decay rates. The areas are separated by the dashed lines. Area I corresponds to a decay rate of 0, area II corresponds to a decay rate of $\min_{\theta\in(0,\theta^*)}(1-x_1)\Lambda_A(-\theta)+(\bar{x}_1-\bar{x}_0)\theta$ and area III corresponds to a decay rate of $(1-x_1)\Lambda_A(-\theta^*)+(\bar{x}_1-\bar{x}_0)\,\theta^*$.

Remark that the decay rate equals $-W(\mathbf{x})$ in (5.8) in area III (and in area I and II the decay rate is upper bounded by $-W(\mathbf{x})$). This can be expected based on the interpretation of the subsolution $W(\mathbf{x})$. □

7.2.2 The d-node $GI|GI|1$ tandem queue

In this section, we derive a lower bound for the decay rate for the d-node tandem queue, based on the decay rate for the single queue, as we did in Lemma 2.4 for $\mathbf{x}=\mathbf{0}$. We use the same arguments, but we mention all steps for readability.

Lemma 7.7 (Lower bound for bounded support)**.** *Consider a d-node $GI|GI|1$ tandem queue satisfying Assumptions 2.3 and 5.1. For the decay of $\mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x}))$ it holds that,*

$$\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}(K_N(\mathbf{x})<K_0(\mathbf{x}))\geq\max_j\left((1-x_1-\ldots-x_j)\Lambda_A(-\theta^{(j)})\right),$$

uniformly in $\mathbf{x}$.

Proof. We follow the proof in Lemma 2.4 using (7.1) and (7.2) instead of (2.1) and (2.2). That is, we compare the tandem queue with starting state $\mathbf{x}=(x_1,\ldots,x_d,\bar{x}_0,\ldots,\bar{x}_d)$ to a single queue with starting state $\widehat{\mathbf{x}}=(x_1+\ldots+x_j,\bar{x}_0,\bar{x}_j)$ for some $j=1,\ldots,d$, with the same arrival process A_k and the service process of the j^{th} queue in the tandem queue.

For convenience, we add a hat to a random variable whenever it relates to the *single* queue. In particular, we let $\hat{D}_k$ denote the inter-departure time at the single queue, between the $(k-1)^{th}$ and k^{th} *departing customer from the d-node tandem queue*. Furthermore, we let $D_k^{(j)}$ be the inter-departure time at the j^{th} queue in tandem, between the $(k-1)^{th}$ and k^{th} *departing customer from the system*. In particular, $D^{(j)}_{(x_{j+1}+\ldots+x_d)N+1}=\sum_{i<k\leq j}B^{(k)}_{(x_{j+1}+\ldots+x_d)N+1}+\bar{x}_iN$ when

queue i is the last non-empty queue with index smaller than j. Recall that the first departing customer from queue j is the $((x_{j+1}+\ldots+x_d)N+1)^{th}$ departing customer from the system. This means that, using similar arguments as just before (2.4) gives us for all $k=1,\ldots,\min\{K_0(\mathbf{x})-1+(x_1+\ldots+x_d)N, \widehat{K}_0(\widehat{\mathbf{x}})-1+(x_1+\ldots+x_d)N\}$,

$$\begin{aligned}\sum_{i=2}^{k} D_i + D_1 &\geq \sum_{i=(x_{j+1}+\ldots+x_d)N+2}^{k} D_i^{(j)} + D^{(j)}_{(x_{j+1}+\ldots+x_d)N+1} \\ &\geq \sum_{i=(x_{j+1}+\ldots+x_d)N+2}^{k} \widehat{D}_i + \bar{x}_j N, \end{aligned} \tag{7.16}$$

since a customer cannot leave the d-node tandem queue before having left queue j and the first $(x_{j+1}+\ldots+x_d)N$ customers leaving the system did not receive service at queue j. In particular, the first departing customer in the *single* queue departs after $\bar{x}_j N$ time.

As in Lemma 2.4, we show by contradiction that $\widehat{K}_0(\widehat{\mathbf{x}}) \leq K_0(\mathbf{x})$, so that the single queue is empty not later than the tandem queue. Suppose now that $\widehat{K}_0(\widehat{\mathbf{x}}) > K_0(\mathbf{x})$, then (7.16) holds for k up to $K_0(\mathbf{x})-1+(x_1+\ldots+x_d)N$. Using (7.2) together with (7.16) we find

$$\begin{aligned}\sum_{k=1}^{K_0(\mathbf{x})-1} A_k + \bar{x}_0 N &> \sum_{k=2}^{K_0(\mathbf{x})-1+(x_1+\ldots+x_d)N} D_k + D_1 \\ &\geq \sum_{k=(x_{j+1}+\ldots+x_d)N+2}^{K_0(\mathbf{x})-1+(x_1+\ldots+x_d)N} \widehat{D}_k + \bar{x}_j N,\end{aligned}$$

which implies $\widehat{K}_0(\widehat{\mathbf{x}}) \leq K_0(\mathbf{x})$, so our assumption is wrong and we have shown that $\widehat{K}_0(\widehat{\mathbf{x}}) \leq K_0(\mathbf{x})$.

Also similar to Lemma 2.4, we show that the tandem queue reaches the overflow level not later than the single queue. Suppose $\widehat{K}_N(\widehat{\mathbf{x}}) < \widehat{K}_0(\widehat{\mathbf{x}})$, so that we have reached overflow in a busy cycle of the single queue. Then we find by using (7.2), where we recall that customer $(x_{j+1}+\ldots+x_d)N+1$ is the first customer to leave queue j, and (7.16)

$$\begin{aligned}\sum_{k=1}^{\widehat{K}_N(\widehat{\mathbf{x}})-1} A_k + \bar{x}_0 N &< \sum_{k=(x_{j+1}+\ldots+x_d)N+2}^{\widehat{K}_N(\widehat{\mathbf{x}})-(1-x_1-\ldots-x_d)N+1} \widehat{D}_k + \bar{x}_j N \\ &\leq \sum_{k=2}^{\widehat{K}_N(\widehat{\mathbf{x}})-(1-x_1-\ldots-x_d)N+1} D_k + D_1,\end{aligned}$$

and so we have found $K_N(\mathbf{x}) \leq \widehat{K}_N(\widehat{\mathbf{x}})$. Note that (7.16) holds, since $\widehat{K}_N(\widehat{\mathbf{x}}) \leq \widehat{K}_0(\widehat{\mathbf{x}}) \leq K_0(\mathbf{x})$.

Thus, $\widehat{K}_N(\widehat{\mathbf{x}}) < \widehat{K}_0(\widehat{\mathbf{x}})$ implies $K_N(\mathbf{x}) < K_0(\mathbf{x})$ and so we have for any j by using Theorem 7.5

$$\begin{aligned}\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) &\geq \liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}(\widehat{K}_N(\widehat{\mathbf{x}}) < \widehat{K}_0(\widehat{\mathbf{x}})) \\ &= (1 - x_1 - \ldots - x_j)\Lambda_A(-\theta^{(j)}),\end{aligned}$$

where we recall that the starting state for the single queue is $\widehat{\mathbf{x}} = (x_1 + \ldots + x_j, \bar{x}_0, \bar{x}_j)$. This inequality holds for all queues $j = 1, \ldots, d$, therefore it holds in particular for the maximum over all j. Since the decay rate for the single $GI|GI|1$ queue holds uniformly in $\mathbf{x}$, this also holds for the derived lower bound. □

In the next conjecture, we state that we expect the lower bound in Lemma 7.7 to be tight when we have bounded supports.

Conjecture 7.8 (Decay rate for bounded supports)**.** *Consider a d-node $GI|GI|1$ tandem queue with $d > 1$, satisfying Assumptions 2.3 and 5.1. For the decay of $\mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x}))$ it holds that,*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}(K_N(\mathbf{x}) < K_0(\mathbf{x})) = -\gamma(\mathbf{x}), \tag{7.17}$$

uniformly in $\mathbf{x}$*, where* $\gamma(\mathbf{x}) = -\max_j \left((1 - x_1 - \ldots - x_j)\Lambda_A(-\theta^{(j)})\right)$.

Intuition. Although we have proved the equality (7.17) as lower bound in Lemma 7.7, it remains to show that the upper bound is the same. If the supports are bounded, we expect that the residual inter-arrival time and residual service time do not influence the decay rate, as we have shown for the single queue in Theorem 7.5.

So suppose that we neglect the residual inter-arrival times and service times, then the probability to reach the overflow level clearly depends on the number of customers currently in the system. Moreover, it depends on which queue the customers are in. When all customers are in some queue with an index smaller than the bottleneck queue j^*, then we expect the decay rate to be $(1 - x_1 - \ldots - x_{j^*})\Lambda_A(-\theta^*)$, since we only need $(1 - x_1 - \ldots - x_{j^*})N$ additional customers to arrive in order to reach the overflow level.

On the other hand, when x_j is large for some queue j with index larger than the bottleneck queue j^*, then it is expected that the decay rate depends on $\theta^{(j^*)}$ and $\theta^{(j)}$ and the number of customers in each of the queues. That is, it depends on how difficult it is to fill queue j^* compared to how difficult it is to fill queue j.

Taking all of the above into account, we expect that the lower bound in Lemma 7.7 is tight. □

7.3 Asymptotically efficient splitting schemes

In this section, we propose a splitting scheme and we show that it results in an asymptotically efficient estimator, see Definition 1.3. The proofs that we provide

are similar to the proofs in [34], though in that paper the decay rate from any starting point $\mathbf{x}$, see also Section 7.2, is used as importance function (similar to Theorem 7.12 below). Recall that we are interested in estimating the probability to reach N customers during a busy cycle of the system, that is, $\mathbf{0}$ is the starting state (as in all previous chapters).

Let $U(\mathbf{x})$ denote the importance function, which we define more precisely later, and let the **splitting scheme** be as follows:

1. Choose the splitting rate R being some integer.
2. Let J_N be the smallest integer such that $J_N \frac{\log R}{N} \geq U(\mathbf{x})$ for all $\mathbf{x}$ in the domain. Obviously, J_N depends on N.
3. Let, for $j = 0, \ldots, J_N$, $C_j^N = \{\mathbf{x} \in \mathcal{D} : U(\mathbf{x}) \geq (J_N - j)\log(R)/N\}$ be the level sets. Therefore, the levels are given by $c_j^N = \{\mathbf{x} \in \mathcal{D} : U(\mathbf{x}) = (J_N - j)\log(R)/N\}$.

For convenience, we define $C_{-1}^N = \emptyset$. In [34], the splitting factor at one of the thresholds is slightly different than R, though this is not necessary for the proofs to hold and hence we choose to use the same splitting factor at all thresholds.

We denote, for any starting state $\mathbf{x}$

$$p_N(\mathbf{x}) = \mathbb{P}\left(K_N(\mathbf{x}) < K_0(\mathbf{x})\right),$$
$$\gamma(\mathbf{x}) = -\lim_{N\to\infty} \frac{1}{N} \log p_N(\mathbf{x}).$$

For $U(\mathbf{x})$ we consider a function based on a subsolution $W(\mathbf{x})$ as derived in Chapter 5. For the single queue, this function can be found in (5.8). For the d-node tandem queue we recall that for splitting we do not need a mollification procedure as in (5.28) and in fact, splitting requires a slightly less complicated subsolution (so that we also do not need to take the minimum of several functions). That is, we can use the following generalization of (5.8)

$$W(\mathbf{x}) = -(x_1 + \ldots + x_d)\gamma(\mathbf{0}) + (\bar{x}_0 - \bar{x}_d)\theta^* + \gamma(\mathbf{0}). \tag{7.18}$$

We choose

$$U(\mathbf{x}) = W(\mathbf{0}) - W(\mathbf{x}) = (x_1 + \ldots + x_d)\gamma(\mathbf{0}) - (\bar{x}_0 - \bar{x}_d)\theta^*,$$

independent of the bottleneck queue. As a result of the importance function, we find that $J_N = \lceil \frac{\gamma(\mathbf{0})N + \theta^* Q^{(d)}}{\log(R)} \rceil$.

From Section 1.3.2 it is known that $p_N = R^{-J_N}\mathbb{E}\left[T\right]$, where T is the number of particles that reach overflow level. To prove that the estimator for p_N is asymptotically efficient, see Definition 1.3, we need the following two lemmas.

Lemma 7.9. *Under Assumptions 2.3 and 5.1, provided that Conjecture 7.8 holds (for $d > 1$), we find*

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{E}\left[T^2\right] \leq 0.$$

Proof. Recall that $T = \sum_{i=1}^{R^{J_N}} I(i)$, where $I(i)$ indicates whether particle i has reached the overflow level or not. It follows that

$$\begin{aligned}
\frac{1}{N}\log\mathbb{E}\left[T^2\right] &= \frac{1}{N}\log\mathbb{E}\left[\left(\sum_{i=1}^{R^{J_N}} I(i)\right)^2\right] \\
&= \frac{1}{N}\log\mathbb{E}\left[\sum_{i=1}^{R^{J_N}} I(i)^2 + \sum_{i=1}^{R^{J_N}}\sum_{j=1,j\neq i}^{R^{J_N}} I(i)I(j)\right] \\
&= \frac{1}{N}\log\left(\sum_{i=1}^{R^{J_N}}\mathbb{E}\left[I(i)\right] + \sum_{i=1}^{R^{J_N}}\sum_{j=1,j\neq i}^{R^{J_N}}\mathbb{E}\left[I(i)I(j)\right]\right) \\
&= \frac{1}{N}\log\left(R^{J_N}p_N(\mathbf{0}) + \sum_{i=1}^{R^{J_N}}\sum_{j=1,j\neq i}^{R^{J_N}}\mathbb{E}\left[I(i)\mid I(j)=1\right]\mathbb{P}\left(I(j)=1\right)\right) \\
&= \frac{1}{N}\log\left(R^{J_N}p_N(\mathbf{0})\left[1 + \sum_{i=2}^{R^{J_N}}\mathbb{E}\left[I(i)\mid I(j)=1\right]\right]\right) \\
&= \frac{1}{N}\log\left(R^{J_N}p_N(\mathbf{0})\right) + \frac{1}{N}\log\left(1 + \sum_{i=2}^{R^{J_N}}\mathbb{E}\left[I(i)\mid I(1)=1\right]\right). \qquad (7.19)
\end{aligned}$$

Clearly, the first term in (7.19) goes to $\gamma(\mathbf{0}) - \gamma(\mathbf{0}) = 0$ as $N \to \infty$, since $J_N = \lceil \frac{\gamma(\mathbf{0})N + Q^{(d)}}{\log(R)} \rceil$. To analyze the second term of (7.19), we consider the level κ in which particle 1 and i had their last common ancestor. Thus we find,

$$\begin{aligned}
&\frac{1}{N}\log\left(1 + \sum_{i=2}^{R^{J_N}}\mathbb{E}\left[I(i)\mid I(1)=1\right]\right) \\
&\leq \frac{1}{N}\log\left(1 + \sum_{\kappa=0}^{J_N-1}(R-1)R^{\kappa}\sup_{\mathbf{z}\in D_\kappa^N} p_N(\mathbf{z})\right), \qquad (7.20)
\end{aligned}$$

where $D_\kappa^N = C_\kappa^N \setminus C_{\kappa-1}^N$. An important difference compared to [34] is that in our case we most likely jump over the level c_κ^N so that we have to consider the supremum of all $\mathbf{z} \in D_\kappa^N$.

To determine a bound on (7.20), fix any $\varepsilon > 0$. By Conjecture 7.8 there exists an $N_1(\varepsilon) < \infty$ such that we have for all $N \geq N_1(\varepsilon)$

$$\frac{1}{N}\log p_N(\mathbf{z}) \leq -\gamma(\mathbf{z}) + \varepsilon.$$

Fix any such N, then we know that for all $\kappa = 0, \ldots, J_N - 1$ we have

$$\sup_{\mathbf{z}\in D_\kappa^N} p_N(\mathbf{z}) \leq \sup_{\mathbf{z}\in D_\kappa^N} e^{-\gamma(\mathbf{z})N + \varepsilon N}. \qquad (7.21)$$

When Assumption 5.1 is satisfied, and so we have bounded inter-arrival and service times, we find from (7.17) and (7.18) that

$$\begin{aligned}-\gamma(\mathbf{z}) + W(\mathbf{z}) &= \max_j \left((1 - z_1 - \ldots - z_j)\Lambda_A(-\theta^{(j)})\right) \\ &\quad + (z_1 + \ldots + z_d)\Lambda_A(-\theta^*) - \Lambda_A(-\theta^*) + (\bar{z}_0 - \bar{z}_d)\theta^* \\ &\leq (\bar{z}_0 - \bar{z}_d)\theta^* \leq \theta^* \frac{Q^{(0)}}{N},\end{aligned}$$

where the inequality follows since $\Lambda_A(-\theta^{(j)}) \leq \Lambda_A(-\theta^*) \leq 0$ for all queues j. Therefore, we can upper bound (7.21) by

$$\begin{aligned}\sup_{\mathbf{z}\in D_\kappa^N} e^{-W(\mathbf{z})N + \theta^* Q^{(0)} + \varepsilon N} &< e^{-W(\mathbf{0})N + (J_N - \kappa + 1)\log R + \theta^* Q^{(0)} + \varepsilon N} \\ &\leq R^{-\kappa+1} e^{-W(\mathbf{0})N + \gamma(\mathbf{0})N + \theta^* Q^{(d)} + \log R + \theta^* Q^{(0)} + \varepsilon N},\end{aligned}$$

where the first inequality follows since for all $\mathbf{z} \in D_\kappa^N$ we have $U(\mathbf{z}) = W(\mathbf{0}) - W(\mathbf{z}) < (J_N - \kappa + 1)\log(R)/N$ and the second inequality follows since $J_N \leq \frac{\gamma(\mathbf{0})N + \theta^* Q^{(d)}}{\log R} + 1$. In particular we can now upper bound (7.20) for the given N by

$$\begin{aligned}&\frac{1}{N}\log\left(1 + \sum_{\kappa=0}^{J_N-1}(R-1)R^\kappa R^{-\kappa+1} e^{-W(\mathbf{0})N + \gamma(\mathbf{0})N + \theta^* Q^{(d)} + \log R + \theta^* Q^{(0)} + \varepsilon N}\right) \\ &= \frac{1}{N}\log\left(1 + \sum_{\kappa=0}^{J_N-1}(R-1)R^2 e^{\theta^*(Q^{(d)} + Q^{(0)}) + \varepsilon N}\right) \\ &= \frac{1}{N}\log\left(1 + J_N(R-1)R^2 e^{\theta^*(Q^{(d)} + Q^{(0)}) + \varepsilon N}\right) \\ &\leq \frac{1}{N}\log\left((J_N(R-1)R^2 + 1) e^{\theta^*(Q^{(d)} + Q^{(0)}) + \varepsilon N}\right),\end{aligned}$$

where the first equality follows since $W(\mathbf{0}) = \gamma(\mathbf{0})$. Since J_N increases (roughly) linearly in N, we find

$$\limsup_{N\to\infty} \frac{1}{N}\log \mathbb{E}\left[T^2\right] \leq \varepsilon,$$

for all $\varepsilon > 0$. This concludes the proof. □

Let $\mathcal{N}_j^N$ denote the number of particles in generation j, that is, the number of particles that have reached threshold $J_N - j$ but not threshold $J_N - j - 1$. Then we define the expected computational effort $w(N)$ as follows

$$w(N) = \sum_{\kappa=0}^{J_N} \mathbb{E}\left[R(\kappa+1)\mathcal{N}_\kappa^N\right], \tag{7.22}$$

see also [29, 34] for a motivation of this choice. Roughly, the idea is that it is assumed that it takes $j + 1$ time units to simulate a particle starting from level $c_{J_N - j}^N$. We have the following result.

Lemma 7.10. *Under Assumption 2.3 and 5.1 it holds that*

$$\limsup_{N\to\infty} \frac{1}{N} \log w(N) \leq 0.$$

Proof. Fix any $\varepsilon > 0$. By Theorem 2.7 there exists $N_1(\varepsilon) < \infty$ such that we have for all $N \geq N_1(\varepsilon)$

$$\frac{1}{N} \log \tilde{p}_{\Sigma(\mathbf{z})N}(\mathbf{0}) \leq -\Sigma(\mathbf{z})\gamma(\mathbf{0}) + \varepsilon,$$

where $\Sigma(\mathbf{z}) = z_1 + \ldots + z_d$ and $\tilde{p}_{\Sigma(\mathbf{z})N}$ denotes the probability to reach $\Sigma(\mathbf{z})N$ customers in the system during a busy cycle of the system. Fix any such N, then we have for all $\kappa = 0, \ldots J_N$

$$\mathbb{E}\left[\mathcal{N}_\kappa^N\right] \leq \sup_{\mathbf{z}\in D^N_{J_N-\kappa}} \tilde{p}_{\Sigma(\mathbf{z})N}(\mathbf{0}) R^\kappa \leq \sup_{\mathbf{z}\in D^N_{J_N-\kappa}} R^\kappa e^{-\Sigma(\mathbf{z})\gamma(\mathbf{0})N+\varepsilon N}.$$

Recall that by definition we have $W(\mathbf{z}) = -(z_1+\ldots+z_d)\gamma(\mathbf{0}) + (\bar{z}_0 - \bar{z}_d)\theta^* + \gamma(\mathbf{0})$ and so the right-hand side of the previous display equals

$$\begin{aligned}
&\sup_{\mathbf{z}\in D^N_{J_N-\kappa}} R^\kappa e^{W(\mathbf{z})N - \gamma(\mathbf{0})N - (\bar{z}_0-\bar{z}_d)\theta^* N + \varepsilon N} \\
&\leq R^\kappa e^{-\kappa \log R + W(\mathbf{0})N - \gamma(\mathbf{0})N - (\bar{z}_0-\bar{z}_d)\theta^* N + \varepsilon N} \\
&\leq e^{Q^{(d)}\theta^* + \varepsilon N},
\end{aligned}$$

where the first inequality follows since for all $\mathbf{z} \in D^N_{J_N-\kappa}$ we have $U(\mathbf{z}) = W(\mathbf{0}) - W(\mathbf{z}) \geq \kappa \log R/N$ and the second inequality follows since $\bar{z}_d - \bar{z}_0 \leq Q^{(d)}/N$.

Using (7.22) we find that for the given N we have

$$\begin{aligned}
\frac{1}{N}\log w(N) &\leq \frac{1}{N}\log\left(Re^{Q^{(d)}\theta^*+\varepsilon N}\sum_{\kappa=0}^{J_N}(\kappa+1)\right) \\
&= \frac{1}{N}\log\left(Re^{Q^{(d)}\theta^*+\varepsilon N}\frac{1}{2}(J_N+1)(J_N+2)\right).
\end{aligned}$$

Since J_N increases linearly in N, we find

$$\limsup_{N\to\infty} \frac{1}{N} \log w(N) \leq \varepsilon,$$

for all $\varepsilon > 0$. This concludes the proof. □

As a result of the previous two lemmas, we immediately have the following main result.

Theorem 7.11. *Consider a d-node tandem queue satisfying Assumptions 2.3 and 5.1. Under Conjecture 7.8, the splitting scheme based on $U(\mathbf{x}) = W(\mathbf{0}) - W(\mathbf{x})$ is asymptotically efficient in the sense of Definition 1.3.*

Proof. To show that the splitting scheme is asymptotically efficient, we need to show that

$$\limsup_{N\to\infty} \frac{1}{N} \log\left(w(N)R^{-2J_N}\mathbb{E}\left[T^2\right]\right) \leq -2\gamma(\mathbf{0}).$$

As a result of Lemmas 7.9 and 7.10 it remains to show that

$$\limsup_{N\to\infty} \frac{1}{N} \log\left(R^{-2J_N}\right) \leq -2\gamma(\mathbf{0}),$$

which clearly holds since $J_N \geq \frac{\gamma(\mathbf{0})N+\theta^* Q^{(d)}}{\log R}$. □

Following the same steps that lead to the proof of Theorem 7.11, we can also prove the following.

Theorem 7.12. *Consider a d-node tandem queue satisfying Assumptions 2.3 and 5.1. Under Conjecture 7.8, the splitting scheme based on $U(\mathbf{x}) = \gamma(\mathbf{0}) - \gamma(\mathbf{x})$ is asymptotically efficient in the sense of Definition 1.3.*

Proof. In the proof of Lemmas 7.9 and 7.10 we can now skip the steps where we bound the difference of $-\gamma(\mathbf{z})$ and $W(\mathbf{z})$. All other steps follow similarly, where in Lemma 7.10 we need to use that $\max_j\{(1-x_1-\ldots x_j)\Lambda_A(-\theta^{(j)})\} \leq (1-x_1-\ldots-x_d)\Lambda_A(-\theta^*)$. □

Alternatively, we can define the following **alternative splitting scheme**, with importance function $\widetilde{U}(\mathbf{x})$, as follows:

1. Choose the splitting rate R being some integer.

2. Let J_N be the smallest integer such that $J_N \frac{\log R}{N} \geq \widetilde{U}(\mathbf{x})$ for all $\mathbf{x}$ in the domain. Obviously, J_N depends on N.

3. Let $C_j^N = \{\mathbf{x} \in \mathcal{D} : \widetilde{U}(\mathbf{x}) \leq j\log(R)/N\}$ be the level sets. Therefore, the levels are $c_j^N = \{\mathbf{x} \in \mathcal{D} : \widetilde{U}(\mathbf{x}) = j\log(R)/N\}$.

Similar to the proofs of Theorem 7.11 and 7.12 we can show the following two statements.

Theorem 7.13. *Consider a d-node tandem queue satisfying Assumptions 2.3 and 5.1. Under Conjecture 7.8, the alternative splitting scheme based on $\widetilde{U}(\mathbf{x}) = W(\mathbf{x})$ is asymptotically efficient in the sense of Definition 1.3.*

Theorem 7.14. *Consider a d-node tandem queue satisfying Assumptions 2.3 and 5.1. Under Conjecture 7.8, the alternative splitting scheme based on $\widetilde{U}(\mathbf{x}) = \gamma(\mathbf{x})$ is asymptotically efficient in the sense of Definition 1.3.*

To illustrate the difference between all these splitting schemes, we consider the following example. Consider a single $D|U|1$ queue with $A = 1.1$, $B^{(1)} \sim U[0, 2]$, $R = 3$ and $N = 5$. Then the levels c_j^N are illustrated in Figure 7.2. Recall that no splitting occurs upon crossing threshold J_N.

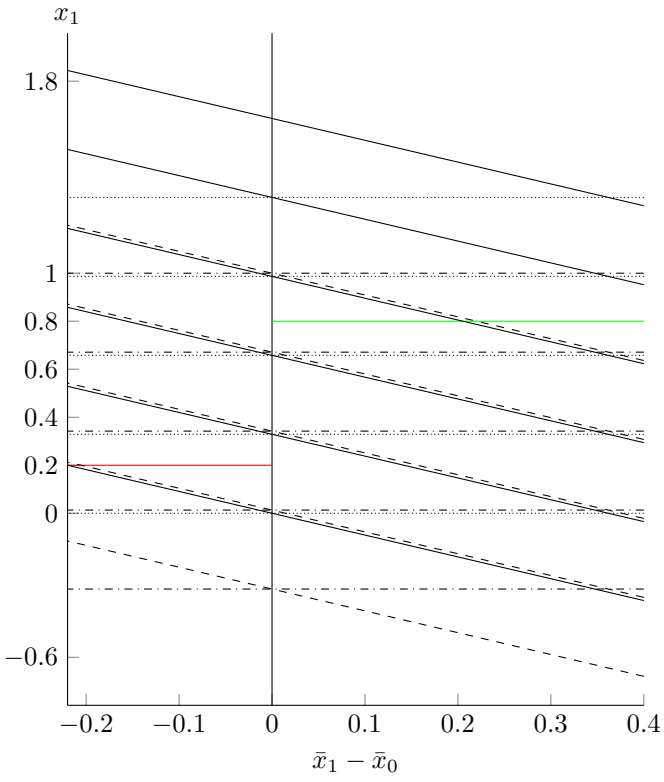


Figure 7.2 All level sets c_j^N, $j = 0, \ldots, J_N$, from top to bottom, for different splittings schemes for a single $D|U|1$ queue, with $A = 1.1$, $B^{(1)} \sim U[0, 2]$, $R = 3$ and $N = 5$. The solid lines are for $U(\mathbf{x}) = W(\mathbf{0}) - W(\mathbf{x})$, the dashed lines are for $\widetilde{U}(\mathbf{x}) = W(\mathbf{x})$, the dotted lines are for $U(\mathbf{x}) = \gamma(\mathbf{0}) - \gamma(\mathbf{x})$ and the dashed dotted lines are for $\widetilde{U}(\mathbf{x}) = \gamma(\mathbf{x})$. The red line is the taboo set, since for $x_1 = 0.2$ (that is, 1 customer in the system) and $\bar{x}_1 < \bar{x}_0$ the queue is sure to be empty, and the green line is the target set, since for $x_1 = 0.8$ (that is, 4 customers in the system) and $\bar{x}_1 > \bar{x}_0$ the queue is sure to reach level 5.

From this figure we see that when starting at the point $\mathbf{0}$, splitting occurs earlier for $\widetilde{U}(\mathbf{x})$ than for $U(\mathbf{x})$. Obviously, the difference between the use of the subsolution and the (positive) decay rate as importance function differs since the latter one does not contain the residual inter-arrival time and residual service time. For large enough N, these residuals do not have any influence. In the next section, we show some numerical results for all of these different importance

functions for this $D|U|1$ queue, as well as for four tandem systems.

7.4 Numerical results

In this section, we present numerical results for the single $GI|GI|1$ queue and the 2-node $GI|GI|1$ tandem queue. In each of the examples, we explicitly write which importance function we use. Recall that the estimator for p_N can be found in (1.5).

In all tables below, RE is the relative error, which for splitting is formally defined as

$$\mathrm{RE} = \frac{\sqrt{\frac{1}{S-1}\sum_{i=1}^{S}\left(\frac{T(i)}{R^{J_N}} - \hat{p}_N\right)^2}}{\sqrt{S}\hat{p}_N},$$

where we recall that $T(i)$ denotes the number of particles that reach the target set in simulation i, and S is the total number of simulations. Furthermore, we define AE as

$$\mathrm{AE} = \frac{\frac{1}{N}\log\frac{1}{S}R^{-2J_N}\sum_{i=1}^{S}T(i)^2}{\frac{1}{N}\log\hat{p}_N}.$$

Thus, if the splitting scheme is asymptotically efficient, the value of AE should go to 2 when N tends to infinity. Furthermore, we present columns with the decay of the expected computational effort ($\frac{1}{N}\log w(N)$), the maximum number of particles simulated at the same time (Max), and the (rounded) simulation time in seconds (Time).

7.4.1 The single $GI|GI|1$ queue

In Tables 7.1–7.4 we consider the same $D|U|1$ queue and we use the different importance functions as discussed at the end of the previous section. That is, in Table 7.1 we use the importance function $U(\mathbf{x}) = W(\mathbf{0}) - W(\mathbf{x})$, in Table 7.2 we use $U(\mathbf{x}) = \gamma(\mathbf{0}) - \gamma(\mathbf{x})$, in Table 7.3 we use $\widetilde{U}(\mathbf{x}) = W(\mathbf{x})$ and in Table 7.4 we use $\widetilde{U}(\mathbf{x}) = \gamma(\mathbf{x})$.

Clearly, all of these tables support asymptotic efficiency of the estimator: the relative error does not increase too fast, the value of AE tends to 2 and the decay of the expected computational effort decreases towards zero.

Comparing Tables 7.1 and 7.2, we see that the computation time for the same number of simulations is higher in Table 7.2. On the other hand, the relative error is slightly lower in that table. It is likely that both of these observations happen due to the fact that splitting occurs at slightly different places in the state space. In particular, when using the importance function as in Table 7.1, splitting occurs later than when using the importance function in Table 7.2 when $\bar{x}_0 - \bar{x}_1 \geq 0$, that is, when the next event is a departure. Thus, using the importance function

as in Table 7.2, we get more particles and hence the simulation time is longer. On the other hand, these additional particles also result in a lower relative error.

For most values of N, we see similar behavior in Tables 7.3 and 7.4. Whenever the simulation time is longer, it appears that the relative error is slightly lower. Again, when we compare Tables 7.1 and 7.3 or Tables 7.2 and 7.4, we see the same behavior. In these comparisons it is clear that splitting occurs earlier when considering the importance functions as in Tables 7.2 and 7.4, see also Figure 7.2, and thus we have to consider more particles. This in turn results in a lower relative error, for the same number of simulations, and a larger simulation time.

7.4.2 The 2-node $GI|GI|1$ tandem queue

For the 2-node tandem queue, we only consider the importance function $U(\mathbf{x}) = W(\mathbf{0}) - W(\mathbf{x})$. In particular, we consider some of the cases from Section 5.3 in order to compare importance sampling and splitting based on subsolutions.

First of all, note that Tables 7.5, 7.6, 7.7 and 7.8 consider the same model as in Tables 5.6, 5.7, 5.8 and 5.9 respectively. In particular, Tables 7.6–7.8 consider unbounded inter-arrival times and hence we did not formally prove that the estimator is asymptotically efficient.

All tables from this section suggest that the splitting estimator is asymptotically efficient. However, importance sampling shows a much better result when comparing simulation time with the relative error, that is, both the relative error and the simulation time are lower (and the number of simulations is higher) for importance sampling. This may have several causes. First of all, importance sampling is implemented in C++, while splitting is implemented in Matlab. Furthermore, it may be that the splitting algorithm that was implemented is not yet optimized, which causes a higher simulation time. Also, the results from importance sampling and splitting were determined on different computers. It seems likely that an implementation of RESTART would decrease the simulation time.

Moreover, it is important to note that for importance sampling we have used 10^6 simulations, whereas for splitting we have used either 10^3 or $5{\cdot}10^4$ simulations. Even though this does not explain the high simulation times for splitting, it does explain that the relative errors from splitting are still higher than those from importance sampling. In particular, when we consider the results from Tables 5.6 and 7.5, we expect that when performing 10^6 simulations for splitting, the relative errors decrease by a factor $\sqrt{1000}$. This would suggest that, when performing 10^6 simulations for splitting, splitting performs better than importance sampling in the sense that it gives a lower relative error with the same number of simulations.

Table 7.1 $D|U|1$ queue. We choose $A = 1.1$, $B^{(1)} \sim U[0,2]$ and so we find $\theta^* = 0.6073$. We use $R = 3$ and $U(\mathbf{x}) = W(\mathbf{0}) - W(\mathbf{x})$, and as a result $J_N = \lceil \frac{\gamma(\mathbf{0})N + 2\theta^*}{\log R} \rceil$. The number of simulations is 10^5.

N	$\hat{p}_N$	RE	AE	$\frac{1}{N}\log w(N)$	Max	Time (s)
20	1.1767e-06	0.0241	1.7012	0.2118	159	76
60	3.0029e-18	0.0437	1.8697	0.1052	1713	203
100	7.2480e-30	0.0565	1.9140	0.0734	6853	399
140	1.6915e-41	0.0691	1.9343	0.0565	12010	612
180	4.7791e-53	0.0741	1.9476	0.0470	13742	966
220	1.0447e-64	0.0867	1.9550	0.0403	19636	1407
260	3.1555e-76	0.0863	1.9620	0.0355	27519	1911
300	7.8094e-88	0.1011	1.9654	0.0318	49982	2502
340	1.7569e-99	0.1067	1.9691	0.0285	54035	2780
380	4.3368e-111	0.1052	1.9724	0.0262	48376	3619
420	1.1438e-122	0.1133	1.9745	0.0243	81107	4464
460	2.8862e-134	0.1251	1.9761	0.0226	123001	5779
500	7.2861e-146	0.1251	1.9780	0.0211	117094	6528

Table 7.2 The same $D|U|1$ queue as in Figure 7.1. We use $R = 3$ and $U(\mathbf{x}) = \gamma(\mathbf{0}) - \gamma(\mathbf{x})$, and as a result $J_N = \lceil \frac{\gamma(\mathbf{0})N}{\log R} \rceil$. The number of simulations is 10^5.

N	$\hat{p}_N$	RE	AE	$\frac{1}{N}\log w(N)$	Max	Time (s)
20	1.1823e-06	0.0206	1.7237	0.2285	157	106
60	2.8224e-18	0.0364	1.8788	0.1101	1571	274
100	7.6018e-30	0.0481	1.9187	0.0768	5758	550
140	2.0464e-41	0.0550	1.9390	0.0597	9202	942
180	4.6114e-53	0.0644	1.9500	0.0491	18334	1398
220	1.1803e-64	0.0720	1.9575	0.0421	22509	1997
260	2.8993e-76	0.0773	1.9632	0.0367	40257	2594
300	6.4717e-88	0.0786	1.9680	0.0328	34445	3378
340	1.8183e-99	0.0884	1.9707	0.0297	58494	4180
380	4.3207e-111	0.0927	1.9734	0.0271	66808	5133
420	9.6573e-123	0.0934	1.9759	0.0248	67570	5704
460	2.8208e-134	0.1005	1.9775	0.0233	86092	7388
500	6.2859e-146	0.1089	1.9788	0.0216	94286	8126

Table 7.3 The same $D|U|1$ queue as in Figure 7.1. We use $R = 3$ and $\widetilde{U}(\mathbf{x}) = W(\mathbf{x})$, and as a result $J_N = \lceil \frac{\gamma(\mathbf{0})N + 1.1\theta^*}{\log R} \rceil$. The number of simulations is 10^5.

N	$\hat{p}_N$	RE	AE	$\frac{1}{N}\log w(N)$	Max	Time (s)
20	1.1721e-06	0.0155	1.7641	0.2639	218	171
60	2.8288e-18	0.0196	1.9091	0.1335	1978	923
100	7.2411e-30	0.0294	1.9334	0.0863	4979	1366
140	1.7398e-41	0.0403	1.9457	0.0638	11034	1672
180	4.5233e-53	0.0324	1.9613	0.0566	19615	5286
220	1.0290e-64	0.0437	1.9643	0.0463	26290	5148
260	2.7584e-76	0.0549	1.9672	0.0391	36049	4872
300	6.8990e-88	0.0400	1.9747	0.0375	49438	13684
340	1.8357e-99	0.0508	1.9756	0.0329	66178	12495
380	4.6078e-111	0.0631	1.9764	0.0291	62365	10692
420	1.1233e-122	0.0792	1.9771	0.0258	81858	8562
460	2.7601e-134	0.0589	1.9810	0.0256	94518	21622
500	6.3617e-146	0.0754	1.9810	0.0230	129434	17187

Table 7.4 The same $D|U|1$ queue as in Figure 7.1. We use $R = 3$ and $\widetilde{U}(\mathbf{x}) = \gamma(\mathbf{x})$, and as a result $J_N = \lceil \frac{\gamma(\mathbf{0})N}{\log R} \rceil$. The number of simulations is 10^5.

N	$\hat{p}_N$	RE	AE	$\frac{1}{N}\log w(N)$	Max	Time (s)
20	1.1874e-06	0.0135	1.7832	0.2820	182	226
60	2.9183e-18	0.0279	1.8919	0.1206	2052	463
100	7.0029e-30	0.0420	1.9228	0.0787	5121	684
140	1.7860e-41	0.0362	1.9480	0.0663	12697	2351
180	4.6270e-53	0.0470	1.9552	0.0525	19338	2547
220	1.1275e-64	0.0632	1.9593	0.0432	22470	2567
260	2.7537e-76	0.0478	1.9688	0.0405	41756	6857
300	6.7381e-88	0.0595	1.9707	0.0348	47613	6104
340	1.6366e-99	0.0756	1.9721	0.0303	61460	5285
380	4.3689e-111	0.0541	1.9776	0.0298	68177	13963
420	9.2362e-123	0.0732	1.9776	0.0264	77811	10950
460	2.6270e-134	0.0859	1.9785	0.0240	108252	10419
500	6.5829e-146	0.0648	1.9819	0.0238	121062	24896

Table 7.5 $D|U|1-\cdot|U|1$ tandem queue when queue 2 is the bottleneck queue. We choose $A = 1.1$, $B^{(1)} \sim U[0.5, 1.5]$, $B^{(2)} \sim U[0, 2]$. We find $\theta^* = \theta^{(2)} = 0.6073$, $\theta^{(1)} = 2.5215$. We use $R = 3$ and $U(\mathbf{x}) = W(\mathbf{0}) - W(\mathbf{x})$ and as a result $J_N = \lceil \frac{\gamma(\mathbf{0})N + 2\theta^*}{\log R} \rceil$. The number of simulations is 10^3.

N	$\hat{p}_N$	RE	AE	$\frac{1}{N}\log w(N)$	Max	Time (s)
20	7.9313e-05	0.0309	1.9289	0.4121	282	134
60	2.0681e-16	0.0554	1.9612	0.1752	2998	421
100	5.4886e-28	0.0616	1.9750	0.1162	5357	951
140	1.3423e-39	0.0777	1.9782	0.0874	9088	1563
180	2.7429e-51	0.0974	1.9798	0.0702	26168	2149
220	9.6275e-63	0.0889	1.9847	0.0604	28524	3644
260	1.9554e-74	0.1163	1.9842	0.0518	33572	4211
300	4.7646e-86	0.1211	1.9860	0.0456	41483	4928
340	1.3617e-97	0.1256	1.9874	0.0411	63260	6390
380	2.3788e-109	0.1370	1.9881	0.0369	55553	6820
420	6.4064e-121	0.1455	1.9888	0.0340	67737	8694
460	1.6320e-132	0.1694	1.9888	0.0315	148768	10667
500	5.4523e-144	0.1267	1.9914	0.0299	115071	16310

Table 7.6 $M|U|1-\cdot|U|1$ tandem queue when both queues are the bottleneck queue. We choose $A \sim \exp(0.5)$, $B^{(1)} \sim U[0, 3]$ and $B^{(2)} \sim U[0, 3]$. We find $\theta^* = \theta^{(1)} = \theta^{(2)} = 0.2690$. We use $R = 3$ and $U(\mathbf{x}) = W(\mathbf{0}) - W(\mathbf{x})$, and as a result $J_N = \lceil \frac{\gamma(\mathbf{0})N + 3\theta^*}{\log R} \rceil$. The number of simulations is 10^3.

N	$\hat{p}_N$	RE	AE	$\frac{1}{N}\log w(N)$	Max	Time (s)
20	1.9643e-03	0.0630	1.7431	0.2979	100	9
60	1.6207e-10	0.0947	1.8981	0.1493	2779	47
100	1.0814e-17	0.1045	1.9366	0.1060	8302	162
140	4.4208e-25	0.1117	1.9536	0.0820	18276	352
180	1.8467e-32	0.1067	1.9656	0.0682	27634	696
220	9.0993e-40	0.1313	1.9677	0.0588	55685	1256
260	3.0929e-47	0.1595	1.9694	0.0513	122122	1874
300	1.3965e-54	0.1432	1.9753	0.0464	142837	3361
340	5.7194e-62	0.1493	1.9777	0.0421	232899	5463
380	1.5716e-69	0.1674	1.9787	0.0379	235490	6312
420	4.1473e-77	0.2068	1.9785	0.0343	288481	6461
460	2.4657e-84	0.1799	1.9818	0.0328	507913	13820
500	1.0529e-91	0.1877	1.9829	0.0311	1108949	23185

Table 7.7 $M|U|1-\cdot|U|1$ tandem queue when queue 1 is the bottleneck queue. We choose $A \sim \exp(0.5)$, $B^{(1)} \sim U[0,3]$ and $B^{(2)} \sim U[0,2]$. We find $\theta^* = \theta^{(1)} = 0.2690$, $\theta^{(2)} = 0.8966$. We use $R = 3$ and $U(\mathbf{x}) = W(\mathbf{0}) - W(\mathbf{x})$, and as a result $J_N = \lceil \frac{\gamma(\mathbf{0})N + 2\theta^*}{\log R} \rceil$. The number of simulations is $5 \cdot 10^4$.

N	$\hat{p}_N$	RE	AE	$\frac{1}{N}\log w(N)$	Max	Time (s)
20	3.2370e-04	0.0168	1.6624	0.2266	105	159
60	1.0804e-11	0.0294	1.8500	0.1095	860	402
100	3.7576e-19	0.0382	1.8985	0.0756	2147	673
140	1.1115e-26	0.0478	1.9206	0.0581	4695	979
180	3.8789e-34	0.0528	1.9357	0.0484	10368	1483
220	1.3269e-41	0.0577	1.9456	0.0414	13228	2003
260	4.4920e-49	0.0591	1.9536	0.0364	13501	2597
300	1.4468e-56	0.0652	1.9583	0.0324	23041	3219
340	4.5235e-64	0.0745	1.9614	0.0291	22294	3665
380	1.8318e-71	0.0731	1.9657	0.0270	32122	5013
420	5.5515e-79	0.0760	1.9685	0.0248	37553	5624
460	1.6700e-86	0.0852	1.9701	0.0228	51695	6176
500	6.4803e-94	0.0822	1.9728	0.0216	57988	7799

Table 7.8 $M|U|1-\cdot|U|1$ tandem queue when queue 2 is the bottleneck queue. We choose $A \sim \exp(0.5)$, $B^{(1)} \sim U[0,2]$ and $B^{(2)} \sim U[0,3]$. We find $\theta^* = \theta^{(2)} = 0.2690$, $\theta^{(1)} = 0.8966$. We use $R = 3$ and $U(\mathbf{x}) = W(\mathbf{0}) - W(\mathbf{x})$, and as a result $J_N = \lceil \frac{\gamma(\mathbf{0})N + 3\theta^*}{\log R} \rceil$. The number of simulations is $5 \cdot 10^4$.

N	$\hat{p}_N$	RE	AE	$\frac{1}{N}\log w(N)$	Max	Time (s)
20	3.2772e-04	0.0153	1.6836	0.2299	76	160
60	1.1034e-11	0.0269	1.8568	0.1105	710	421
100	3.5799e-19	0.0353	1.9023	0.0760	2142	697
140	1.1696e-26	0.0424	1.9245	0.0591	4923	1077
180	3.9887e-34	0.0472	1.9386	0.0487	8530	1510
220	1.3540e-41	0.0505	1.9484	0.0416	9487	2049
260	4.4884e-49	0.0577	1.9540	0.0365	14020	2641
300	1.3536e-56	0.0605	1.9594	0.0324	15240	3161
340	5.0824e-64	0.0650	1.9632	0.0296	26220	4137
380	1.6376e-71	0.0636	1.9674	0.0271	26752	5094
420	5.2689e-79	0.0700	1.9695	0.0248	28492	5689
460	1.8845e-86	0.0722	1.9718	0.0232	56621	6934
500	6.5122e-94	0.0767	1.9735	0.0217	48277	8027

CHAPTER 8

Long time estimates

In Chapters 5 and 7 we have shown that in order to estimate the probability to reach a large number of customers in a non-Markovian tandem queue, we can use *both* importance sampling and splitting schemes based on subsolutions so that we get an asymptotically efficient estimator. A more general question that remains is which method performs best. We refer to [13] for a discussion on some of the differences and similarities of importance sampling and splitting. In this chapter, we take a first step in order to show that for a particular type of models and splitting thresholds based on (time independent) subsolutions, splitting performs better than importance sampling. Even though non-Markovian tandem queues are not part of these type of models, it is still very interesting that for a particular type of models there *is* a difference between importance sampling and splitting.

To be more precise, in this chapter we study *exit probabilities* for discrete time stochastic processes, where the process escapes from some neighborhood of an attractor prior to a given time. In contrast to existing work on exit probabilities, see for example [26, 37], we allow the time by which the process has to exit to grow polynomially with the large deviations scaling parameter. One motivation for considering this scaling is that when the time interval is large and the escape probability is small, the probability closely approximates the inverse of the mean escape time, and in particular the exponential decay rate for the probability and growth rate for the escape time coincide. Although one can easily conjecture an expression for the decay rate for such probabilities, it does not follow from standard sample path large deviation estimates, which apply to bounded time intervals. The main contribution of this chapter is to apply Freidlin-Wentzell type arguments to rigorously determine the decay rate.

It was shown in [20] that importance sampling has some shortcomings when applied to the problem of estimating probabilities to escape from the neighborhood of an attractor. For example, when the time interval over which escape can occur is large, one is tempted to consider subsolutions to the corresponding time independent Hamilton-Jacobi-Bellman equation for the basis for algorithm design. It is sometimes the case (for example, for reversible systems) that useful subsolutions can be found much more easily for the time independent version. However, as discussed in [20], since the attractor is inside the interior of the domain of interest, importance sampling schemes based on such time independent subsolutions generically degrade as the time interval gets large.

In order to show that such problems do not arise when using splitting, that is, the estimator based on the same subsolution is asymptotically efficient, we need to determine the decay rate of the probability of interest. The latter is the topic of the current chapter, which is based on joint work in [10]. In addition to determining the decay rate, in [10] the decay rate is used to construct a RESTART scheme which results in an asymptotically efficient estimator. We will only briefly discuss these other findings from [10] in this chapter. In [10], RESTART is used, as opposed to ordinary splitting, since RESTART has some computational advantages compared to ordinary splitting, in particular when the time interval gets large (which is the case in the current chapter).

The outline of this chapter is as follows. In Section 8.1 we introduce the model, state conditions that will hold throughout this chapter and state some preliminary results. Since we consider a completely different model than in all previous chapters, the notation is different as well. Then in Section 8.2 we determine the decay rate for the probability to escape from a domain within some time $T(n)$, where n is the large deviation parameter and $T(n)$ is allowed to grow polynomially in n. In Section 8.3 we summarize some of the results from [10], where it is shown that using time independent subsolutions as splitting thresholds for RESTART results in an asymptotically efficient estimator.

8.1 Model and preliminary results

The problem of interest is to estimate exit probabilities of a discrete time process $\{X_i^n\}$ from a bounded open set $\mathcal{D} \subset \mathbb{R}^d$ over a time interval $[0, T(n)]$. The index n serves a dual purpose: we assume that X^n satisfies a large deviations principle (LDP) with rarity parameter n, and we also assume $T(n) \to \infty$ as $n \to \infty$. We will require that the LDP be uniform with respect to the initial condition, as characterized in Definition 8.3.

A large class of processes satisfying our assumptions can be obtained from recursive chains of the form

$$X_{i+1}^n = X_i^n + \frac{1}{n} v_i(X_i^n), \tag{8.1}$$

where $v_i(\cdot)$ are independent and identically distributed random vectors fields whose distribution on $\mathbb{R}^d$ is given by a stochastic kernel $\mu(\cdot|x)$, $x \in \mathbb{R}^d$, having mean $b(x)$ and with $b(\cdot)$ Lipschitz continuous. We will use this as a canonical model for discussion. However, the results will be valid in greater generality, with the key assumption being that an appropriate uniform LDP is available. This covers models such as Markov-modulated processes.

The process X^n naturally induces a family of probability measures $\mathbb{P}_x$, where $\mathbb{P}_x(X^n(0) = x) = 1$. We let $\mathbb{E}_x[\cdot]$ denote the expected value with respect to $\mathbb{P}_x$. Define the cumulant generating function of the vector fields by

$$H(x, \alpha) \doteq \log \int_{\mathbb{R}^d} e^{\langle \alpha, u \rangle} \mu(du|x).$$

We assume that the following conditions are satisfied by H and the stochastic kernel μ.

Condition 8.1. *1. For all $\alpha \in \mathbb{R}^d$, we have $\sup_{x\in\mathbb{R}^d} H(x,\alpha) < \infty$;*

2. The map $x \to \mu(\cdot|x)$ is continuous in the topology of weak convergence.

The cumulant generating function $H(x,\alpha)$ is used to define the local rate associated to the system (8.1) via the Legendre-Fenchel transformation. For $(x,\beta) \in \mathbb{R}^d$, let

$$L(x,\beta) \doteq \sup_{\alpha\in\mathbb{R}^d} \{\langle\alpha,\beta\rangle - H(x,\alpha)\}.$$

We first formulate conditions on the local rate L. For models of the form (8.1), since L is the Legendre-Fenchel transform of a cumulant generating function it is automatic that $L(x,\cdot)$ is convex and that there is a unique $b(x)$ such that $L(x,b(x)) = 0$ for each $x \in \mathbb{R}^d$. We will assume the following properties of L. These can all be related to properties of the kernel $\mu(\cdot|x)$ for models of the form (8.1), see [4], and can also apply to other types of models.

Condition 8.2. *We assume that $L : \mathbb{R}^d \times \mathbb{R}^d \to [0,\infty]$ satisfies the following properties:*

1. *for every $x \in \mathbb{R}^d$ there is a unique $b(x)$ such that $L(x,b(x)) = 0$, and $b : \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous;*
2. *for every $x \in \mathbb{R}^d$ $L(x,\cdot)$ is convex;*
3. *for every compact set K, there is $r \in (0,\infty)$ such that $L(x,\beta) \le 1/r$ if $x \in K$, $|\beta| \le r$.*

The most significant of these is the last property, which is used to establish that the infimum of the rate function (defined below) over the interior and closure of sets of trajectories coincide. This is not strictly necessary, but without it both the statement of the results and their proofs become much more complicated. The second property guarantees the existence and uniqueness of solutions to the initial value problem $\dot{x}(t) = b(x(t)), x(0) = x_0$ for any $x_0 \in \mathbb{R}^d$. We call the ordinary differential equation $\dot{x}(t) = b(x(t))$ the *noiseless dynamics*.

To state the LDP for X_i^n, it is convenient to interpolate it as a piecewise constant process. For fixed $T \in (0,\infty)$, define

$$X^n(t) \doteq X_i^n, \quad t \in [i/n, (i+1)/n),\ i = 0,\dots,\lfloor nT\rfloor,$$

where $\lfloor a\rfloor$ is the integer part of a. We consider X^n as taking values in $D([0,T]) \doteq D([0,T] : \mathbb{R}^d)$, with the usual Skorokhod topology [23]. We could also have considered a piecewise linear interpolation that takes values in $C([0,T]) \doteq C([0,T] : \mathbb{R}^d)$ and obtain an LDP with the same rate function, but then we would have to justify the use of the Markov property at only the interpolation times and this seems to lead to more complicated proofs. We will make

use of the fact that the Skorokhod topology on $D([0,T])$ relativized to $C([0,T])$ coincides with the uniform topology given by the metric

$$d(\varphi_1, \varphi_2) \doteq \|\varphi_1 - \varphi_2\|_\infty = \sup_{0 \le t \le T} |\varphi_1(t) - \varphi_2(t)|.$$

We omit the implicit dependence of the metric d on the time interval $[0,T]$ for notational simplicity, as it will be clear from the context what the interval is. For any set $A \subset C([0,T])$ and $\varphi \in C([0,T])$, we define $d(\varphi, A) \doteq \inf_{\phi \in A} d(\varphi, \phi)$. For objects taking values in $\mathbb{R}^d$ we use absolute values $|\cdot|$ to denote the standard Euclidean norm, and we use $\mathcal{E}_\mu$ to denote a ball of radius $\mu > 0$ around 0: $\mathcal{E}_\mu \doteq \{x \in \mathbb{R}^d : |x| < \mu\}$.

Under the conditions we will assume, the rate function associated with the process X^n over the interval $[0,T]$ is

$$I_T(\varphi) \doteq \int_0^T L(\varphi(s), \dot{\varphi}(s)) ds,$$

if $\varphi(t)$ is absolutely continuous, and $I_T(\varphi) = \infty$ otherwise. Note that, owing to part 1 of Condition 8.2, if $I_T(\varphi) = 0$ then φ satisfies the noiseless dynamics $\dot{\varphi}(t) = b(\varphi(t))$.

We phrase the uniform large deviations principle in terms of level sets of I_T, which are defined by

$$\Phi_{x,T}(s) \doteq \{\varphi \in D([0,T]) : I_T(\varphi) \le s, \varphi(0) = x\}.$$

The formulation of the uniform large deviations principle presented here is taken from [26, p. 74].

Definition 8.3 (Uniform large deviations principle)**.** *The sequence X^n satisfies a uniform large deviations principle if:*

1. *the functional I_T is lower semicontinuous on $D([0,T])$ and for each $T \in [0,\infty)$, compact $K \subset \mathbb{R}^d$ and $s < \infty$, the set $\cup_{x \in K} \Phi_{x,T}(s)$ is compact;*

2. *for any $\delta > 0$, $\gamma > 0$, $s_0 < \infty$ and any compact $K \subset \mathbb{R}^d$, there exists $N \in \mathbb{N}$ such that*
$$\mathbb{P}_x(\|X^n - \phi\|_\infty < \delta) \ge \exp(-n(I_T(\phi) + \gamma)),$$
for all $n \ge N$, all $x \in K$ and all $\phi \in \Phi_{x,T}(s_0)$;

3. *for any $\delta > 0$, $\gamma > 0$, $s_0 < \infty$ and any compact $K \subset \mathbb{R}^d$, there exists $N \in \mathbb{N}$ such that*
$$\mathbb{P}_x(d(X^n, \Phi_{x,T}(s)) \ge \delta) \le \exp(-n(s - \gamma)),$$
for all $n \ge N$, $s \le s_0$ and $x \in K$.

This definition allows us to phrase the final assumption on the process X_i^n.

Condition 8.4. *For each $T \in (0, \infty)$, the sequence $\{X^n, n \in \mathbb{N}\}$ satisfies the uniform LDP in Definition 8.3 with a rate function $I_T(\varphi)$ of the form*

$$I_T(\varphi) = \int_0^T L(\varphi(s), \dot{\varphi}(s))ds,$$

for some L satisfying Condition 8.2.

Having posed all conditions on this process, we return to our original problem. We seek to estimate the probability that the process escapes from a domain $\mathcal{D}$ before some time $T(n)$. We will require the domain $\mathcal{D}$ to satisfy some mild regularity properties which are stated below. We will make use of the following notation throughout this chapter. For any subset A of a topological space, we let $\overline{A}$ denote its closure, A^c its complement, and $\partial A = \overline{A} \cap \overline{A^c}$ its boundary. We denote the first exit time of the process X^n from the set $\mathcal{D}$ by

$$\rho^n \doteq \inf\{t \geq 0 : X^n(t) \notin \mathcal{D}\}.$$

We assume that the process exits the set $\mathcal{D}$ in finite time with probability 1, that is, for all $x \in \mathcal{D}$, $\mathbb{P}_x(\rho^n < \infty) = 1$.

Condition 8.5. *We impose the following conditions on the set $\mathcal{D} \subset \mathbb{R}^d$:*

1. *$\mathcal{D}$ is a bounded open subset of $\mathbb{R}^d$;*
2. *$\mathcal{D}$ satisfies a regularity condition at all points of its boundary: for any $\delta > 0$ and $p \in \partial\mathcal{D}$, there is a point q in the interior of $\mathcal{D}^c$ with $|p - q| < \delta$;*
3. *the noiseless dynamics $\dot{x}(t) = b(x(t))$ possess a unique equilibrium point $O \in \mathcal{D}$ which is asymptotically stable on the whole domain: for any $x_0 \in \mathcal{D}$ and $\mu > 0$, there is $T = T(x_0, \mu) \in [0, \infty)$ such that $\dot{x}(t) = b(x(t))$ for $t \geq 0$ and $x(0) = x_0$ together imply $|x(T) - O| < \mu$. Without loss of generality, we take O to be the origin $0 \in \mathbb{R}^d$.*

The assumption that $\mathcal{D}$ is bounded is not necessary for the decay rate of the probability of interest to hold. Even for the results on the decay rate on the second moment of the RESTART estimator in [10] to hold this assumption is not necessary, since one can always restrict to an appropriately chosen bounded subset of the domain. Such a generalization is straightforward but cumbersome and is therefore omitted in [10]. We also note that the noiseless dynamics may exit the domain when started at certain initial conditions $x_0 \in \mathcal{D}$, though they must eventually re-enter.

With regard to $T(n)$ we assume the following.

Condition 8.6. *$T(n)$ grows polynomially in n, by which we mean*

$$\lim_{n\to\infty} T(n) = \infty, \quad \textit{and} \quad \lim_{n\to\infty} [\log T(n)]/n = 0.$$

Throughout this chapter, several interdependent parameters are used. For instance, a statement which holds for all n greater than some threshold value N that depends on ε. We highlight the important dependencies by explicitly writing the parameters at the outset, for example $N(\varepsilon)$, and thereafter using only N. We highlight only the most relevant dependencies and ignore others. For instance, N might also depend on the domain $\mathcal{D}$, but since $\mathcal{D}$ is fixed throughout we do not include it in the list of dependencies.

8.1.1 Preliminary results

In this section we will state some results that will be used in Section 8.2 to determine the rate of decay of the exit probability from $\mathcal{D}$ by some time $T(n)$. The proofs of the lemmas in this section follow more or less directly from the conditions that were presented in Section 8.1 and hence these proofs can either be found in Section 8.4.1 or are omitted.

The uniform bound on the cumulant generating function $H(x, \alpha)$ allows us to establish an asymptotic bound on the maximum jump size of the process, which is the content of Lemma 8.7. The proof follows from Chebyshev's inequality and the fact that $n^{-1} \log(nT(n)) \to 0$, and is omitted.

Lemma 8.7. *Assume Condition 8.1. Then for all $\delta > 0$,*

$$\limsup_{n\to\infty} \sup_{x\in\mathbb{R}^d} \frac{1}{n} \log \mathbb{P}_x \left(\max_{0\le i\le nT(n)} |X_{i+1}^n - X_i^n| \ge \delta \right) = -\infty.$$

The following lemma is an easy consequence of Condition 8.2, and therefore the proof is omitted.

Lemma 8.8. *Under Condition 8.2, for any compact set K there is $c \in (0, \infty)$ such that for any x and $y \in K$, there exists $\tau \in (0, \infty)$ and a smooth function φ, with $\varphi(0) = x$, $\varphi(\tau) = y$, $\tau = |x - y| / r$ for which $I_\tau(\varphi) < c\,|x - y|$.*

Next, we define a function which will be used throughout this chapter and will be commonly referred to as the "cost to exit," starting at a given point. Let

$$W(x) \doteq \inf \{ I_T(\varphi) : \varphi(0) = x, \varphi(T) \notin \mathcal{D}, T < \infty \},$$

for $x \in \mathcal{D}$ and $W(x) \doteq 0$ otherwise. The function $W(\cdot)$ has the following properties.

Lemma 8.9. *Under Condition 8.2, the function $W(\cdot)$ satisfies*

1. *$W(\cdot)$ is continuous on $\mathbb{R}^d$,*
2. *for every $x \in \mathcal{D}$, $W(x) \le W(0)$.*

The proof of Lemma 8.9 can be found in Section 8.4.1. The last lemma in this section gives an upper bound on the probability that the process does not enter the μ-neighborhood of the origin within some finite time T, starting at some point x that is not in the μ-neighborhood of the origin. The proof of this lemma can be found in [26], Lemma 2.2 in Chapter 4.

Lemma 8.10. *For any $\mu > 0$ and any $M < \infty$, there are $T(\mu, M) < \infty$ and $N(\mu, M) < \infty$ such that for any initial point $x \in \overline{\mathcal{D}} \setminus \mathcal{E}_\mu$, we have*

$$\mathbb{P}_x(X^n(t) \in \overline{\mathcal{D}} \setminus \mathcal{E}_\mu, 0 \le t \le T) \le e^{-nM},$$

for all $n \ge N$.

8.2 Main result

The main result of this chapter is the following estimate.

Theorem 8.11. *Suppose Conditions 8.1, 8.2, 8.4, 8.5 and 8.6 are met. Then for any $x \in \mathcal{D}$ we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}_x(\rho^n \le T(n)) = -W(x).$$

Since $T(n) \to \infty$, this is not covered by an LDP of the form in Definition 8.3, and in fact requires a Freidlin-Wentzell analysis appropriate for large time problems as in [26]. Theorem 8.11 follows directly from Lemmas 8.13 and 8.14 below, which give upper and lower bounds for the decay of $\mathbb{P}_x(\rho^n \le T(n))$. We remark that in fact Theorem 8.11 holds uniformly in x, but it is not stated as such since only the upper bound in Lemma 8.13 is proved in its uniform version as it is required in order to prove that the RESTART algorithm gives an asymptotically efficient estimator. The lower bound is easier and holds uniformly as well, but since it is not used in the analysis of RESTART we only establish the pointwise estimate in the interest of saving space.

We start with the upper bound when the process starts in a small neighborhood of the attracting point 0.

Lemma 8.12. *Assume the conditions of Theorem 8.11 are met. For any $\varepsilon > 0$, there is a $\mu(\varepsilon) > 0$ and $N(\varepsilon) < \infty$ such that for all $n \ge N$ and all $y \in \mathcal{E}_\mu = \{x : |x| < \mu\}$,*

$$\frac{1}{n} \log \mathbb{P}_y(\rho^n \le T(n)) \le -W(y) + \varepsilon. \tag{8.2}$$

Proof. Let $\varepsilon > 0$. The function $y \to W(y)$ is continuous on $\mathcal{D}$ by Lemma 8.9, hence uniformly continuous on any bounded neighborhood of 0. For the given $\varepsilon > 0$, there is $\mu > 0$ such that

$$|W(x) - W(y)| < \varepsilon/8 \quad \text{whenever } |x - y| < 2\mu. \tag{8.3}$$

We choose such a $\mu > 0$ and any $\eta \in (0, \mu/2)$. We also assume μ small enough that $\mathcal{E}_\mu \subset \mathcal{D}$. Both μ and η will remain fixed for the rest of the proof. For readability we have divided the proof into 5 steps.

Step 1. *A priori estimates.* Define the sets

$$\Gamma_\mu^\eta \doteq \{x : \mu - \eta \le |x| \le \mu\},$$
$$\gamma_\mu^\eta \doteq \left\{x : \frac{\mu}{2} - \eta \le |x| \le \frac{\mu}{2}\right\}.$$

We will choose a time T_1 such that the probability that the process does not enter the $\mu/2$-neighborhood of the attracting point 0 is superexponentially small. By Lemma 8.10, there are $T_1(\mu)$ and $N_1(\mu)$ such that $T \ge T_1$ and $n \ge N_1$ together imply, for any $x \in \mathcal{D} \setminus \mathcal{E}_{\mu/2}$, that

$$\mathbb{P}_x(X^n(t) \in \mathcal{D} \setminus \mathcal{E}_{\mu/2}, t \in [0, T_1]) \le e^{-n(\bar{W}+1)}, \tag{8.4}$$

where $\bar{W} \doteq \inf_{x\in\Gamma_\mu^\eta} W(x)$. Without loss of generality, we may assume that $\bar{W} > \varepsilon$.

We claim there is $\delta(\varepsilon) > 0$ such that any path φ which satisfies $\varphi(0) \in \Gamma_\mu^\eta \subset \mathcal{D}$ and $\varphi(t) \notin \mathcal{D}$ for some $t \in [0, T_1]$ also satisfies

$$d(\varphi, \Phi_{\varphi(0),T_1}(\bar{W} - \varepsilon/4)) \ge \delta > 0.$$

If not, for any $\delta > 0$ we could find a path φ^δ with $\varphi^\delta(0) \in \Gamma_\mu^\eta$ and $\varphi^\delta(t_*) \notin \mathcal{D}$ for some $t_* \in [0, T_1]$, and $d(\varphi^\delta, \Phi_{\varphi^\delta(0),T_1}(\bar{W} - \varepsilon/4)) < \delta$. We construct this path as follows. By Lemma 8.8, on the compact set $K = \{x : \inf_{z\in\mathcal{D}} |x - z| \le 1\}$ there is $c \in (0, \infty)$ such that for all $x, y \in K$, we can construct φ_{xy} satisfying $\varphi_{xy}(0) = x$, $\varphi_{xy}(\tau) = y$ with $\tau \in (0, \infty)$ and $I_\tau(\varphi_{xy}) \le c|x - y|$.

Choose $\delta \in (0, \varepsilon/8c)$, and without loss of generality assume $\delta < 1$. Under the assumption that $d(\varphi^\delta, \Phi_{\varphi^\delta(0),T_1}(\bar{W} - \varepsilon/4)) < \delta$, we can find $\phi \in \Phi_{\varphi^\delta(0),T_1}(\bar{W} - \varepsilon/4)$ such that $\|\phi - \varphi^\delta\|_\infty < \delta$. Since $\phi(0) = \varphi^\delta(0) \in \Gamma_\mu^\eta$, the definition of $\bar{W}$ implies that $\phi(t_*) \in \mathcal{D} \subset K$, while $\delta < 1$ ensures $\varphi^\delta(t_*) \in K$. Applying Lemma 8.8 to the points $x = \phi(t_*)$ and $y = \varphi^\delta(t_*)$, we obtain a path φ_{xy} connecting x and y in time τ with cost $I_\tau(\varphi_{xy}) \le c\delta < \varepsilon/8$. Concatenating ϕ on $[0, t_*]$ with φ_{xy} on $(t_*, t_* + \tau]$, we obtain a path starting in Γ_μ^η and ending outside of $\mathcal{D}$ at time $t_* + \tau < \infty$, with cost less than $I_{t_*}(\phi) + I_\tau(\varphi_{xy}) \le \bar{W} - \varepsilon/8$. This contradicts the definition of $\bar{W}$. Thus $\delta > 0$ must exist as claimed.

As a consequence of the above discussion, for any $y \in \Gamma_\mu^\eta$,

$$\{\rho^n \le T_1, X^n(0) = y\} \subset \{d(X^n, \Phi_{y,T_1}(\bar{W} - \varepsilon/4)) \ge \delta\}.$$

By the uniform large deviations upper bound on the interval $[0, T_1]$, there is $N_2(\varepsilon, T_1) < \infty$ for which $n \ge N_2$ implies, for any $y \in \Gamma_\mu^\eta$,

$$\mathbb{P}_y(\rho^n \le T_1) \le \mathbb{P}_y(d(X^n, \Phi_{y,T_1}(\bar{W} - \varepsilon/4)) \ge \delta) \le e^{-n(\bar{W} - \frac{\varepsilon}{2})}. \tag{8.5}$$

With T_1 chosen, we introduce the stopping times $\sigma_0 \doteq 0$, and for $j \ge 1$,

$$\tau_j^n \doteq \inf\{t > \sigma_{j-1}^n : X^n(t) \in (\gamma_\mu^\eta \cup \mathcal{D}^c)\} \wedge (T_1 + \sigma_{j-1}^n),$$
$$\sigma_j^n \doteq \inf\{t > \tau_j^n : X^n(t) \in \Gamma_\mu^\eta\}.$$

The stopping times τ_j^n and σ_j^n also depend on η and μ, but we do not include this dependence to avoid an overload of notation. Figure 8.1 shows a sample path with its associated stopping times.

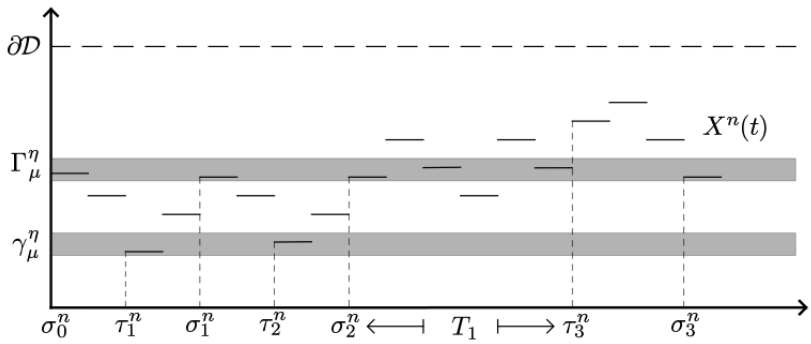


Figure 8.1 The piecewise constant process $X^n(t)$ moving between the sets Γ_μ^η and γ_μ^η. The process in the figure has not exited the domain $\mathcal{D}$ nor hit γ_μ^η at time τ_3^n, since it wandered around for time longer than T_1. Thus, σ_3^n is triggered when the process returns to Γ_μ^η.

Step 2. *The probability of an excursion to $\partial\mathcal{D}$ within time $T(n)$ is approximately the probability of excursion over a single time interval $(\sigma_{j-1}^n, \tau_j^n)$ times the expected number of such intervals.* Define the random variables

$$B_j^n \doteq \begin{cases} 1 & \text{if} \quad X^n(\tau_j^n) \notin \mathcal{D} \quad \text{or} \quad \tau_j^n - \sigma_{j-1}^n \geq T_1, \\ 0 & \text{else.} \end{cases}$$

For any $y \in \Gamma_\mu^\eta$ it holds that

$$\begin{aligned} &\mathbb{P}_y(\rho^n \leq T(n)) \\ &\quad \leq \mathbb{P}_y\left(B_j^n = 1 \text{ for some } j \text{ with } \tau_j^n \leq T(n)\right) \\ &\quad = \mathbb{P}_y\left(B_j^n = 1 \text{ for some } j \text{ with } \left(\tau_1^n - \sigma_0^n + \textstyle\sum_{i=2}^j \tau_i^n - \tau_{i-1}^n\right) \leq T(n)\right) \\ &\quad \leq \mathbb{P}_y\left(B_j^n = 1 \text{ for some } j \text{ with } \left(\textstyle\sum_{i=1}^j \tau_i^n - \sigma_{i-1}^n\right) \leq T(n)\right) \\ &\quad \leq \mathbb{E}_y\left[\textstyle\sum_{i=1}^{M^n} B_i^n\right], \end{aligned} \tag{8.6}$$

where $M^n \doteq \inf\left\{j \geq 1 : \sum_{i=1}^j \tau_i^n - \sigma_{i-1}^n > T(n)\right\}$. We can bound (8.6) by

$$\mathbb{E}_y\left[\textstyle\sum_{i=1}^{M^n} B_i^n\right] \leq \mathbb{E}_y\left[M^n\right] \sup_{x \in \Gamma_\mu^\eta} \mathbb{E}_x\left[B_1^n\right], \tag{8.7}$$

for any $y \in \Gamma_\mu^\eta$. The proof of this statement can be found in Section 8.4.2, Lemma 8.16.

Step 3. *The expected number of intervals $(\sigma_{j-1}^n, \tau_j^n)$ in time $T(n)$ is approximately $T(n)$ divided by the expected duration of each interval. A lower bound on the expected duration gives an upper bound on the expected number of intervals.* For any $y \in \Gamma_\mu^\eta$, we have

$$\begin{aligned}\mathbb{E}_y[M^n] \inf_{x\in\Gamma_\mu^\eta} \mathbb{E}_x\left[\tau_1^n\right] &\le \mathbb{E}_y\left[\textstyle\sum_{i=1}^{M^n} \tau_i^n - \sigma_{i-1}^n\right] \\ &\le T(n) + \tau_{M^n}^n - \sigma_{M^n-1}^n \\ &\le T(n) + T_1,\end{aligned}$$

see Lemma 8.16 in Section 8.4.2 for the proof of the first inequality.

It takes some positive time to travel from Γ_μ^η to either γ_μ^η or to escape $\mathcal{D}$, so $\mathbb{E}_x[\tau_1^n] \ge s > 0$ for some fixed s independent of $y \in \Gamma_\mu^\eta$ and all sufficiently large n. A precise proof can be found in Lemma 8.17 in Section 8.4.2. Hence, for all $y \in \Gamma_\mu^\eta$,

$$\mathbb{E}_y[M^n] \le \frac{T(n) + T_1}{s}. \tag{8.8}$$

Step 4. *Combine estimates for $y \in \Gamma_\mu^\eta$.* From equations (8.6) and (8.7), we find for any $y \in \Gamma_\mu^\eta$

$$\mathbb{P}_y(\rho^n \le T(n)) \le \mathbb{E}_y[M^n] \left(\sup_{x\in\Gamma_\mu^\eta} \mathbb{E}_x[B_1^n] \right). \tag{8.9}$$

The first factor in (8.9) was bounded above in (8.8). For the second factor, note that for any $y \in \Gamma_\mu^\eta$,

$$\mathbb{E}_y\left[B_1^n\right] = \mathbb{P}_y(B_1^n = 1) = \mathbb{P}_y(\{X^n(\tau_1) \notin \mathcal{D}, \tau_1 \le T_1\} \cup \{\tau_1 \ge T_1\}). \tag{8.10}$$

We consider the two events separately and then bound the probability of the union by the sum of the probabilities. The probability of the first event is upper bounded in (8.5) for $n \ge N_2$. For the second event, note that

$$\{\tau_1 \ge T_1\} \subset \{X^n(t) \in \mathcal{D} \setminus \mathcal{E}_{\mu/2}, t \in [0, T_1]\} \cup \left\{ \max_{0\le i\le nT(n)} \left|X_{i+1}^n - X_i^n\right| \ge \eta \right\},$$

for if the process did enter the $\mu/2$-neighborhood of the origin but did not trigger the event τ_1, then at some time index $i = 0, \ldots, nT(n)$ it must have jumped over the set γ_μ^η. From Lemma 8.7, we can choose $N_3(\mu, \eta) < \infty$ for which $n \ge N_3$ implies

$$\mathbb{P}_y\left(\max_{0\le i\le nT(n)} \left|X_{i+1}^n - X_i^n\right| \ge \eta \right) \le e^{-n(\bar{W}+1)}. \tag{8.11}$$

The event $\{X^n(t) \in \mathcal{D} \setminus \mathcal{E}_{\mu/2}, t \in [0, T_1]\}$ is also superexponentially small, as the constant T_1 was originally chosen so that any $y \in \Gamma_\mu^\eta \subset (\mathcal{D} \setminus \mathcal{E}_{\mu/2})$ satisfies (8.4) for all $n \ge N_1$.

Set $N_4 = \max(N_1, N_2, N_3)$. Then, by (8.4), (8.5), (8.10) and (8.11), for any $n \geq N_4$ and any $y \in \Gamma_\mu^\eta$

$$\begin{aligned}\mathbb{E}_y\left[B_1^n\right] &\leq \mathbb{P}_y(\rho^n \leq T_1) + \mathbb{P}_y(X^n(t) \in \mathcal{D} \setminus \mathcal{E}_{\mu/2}, t \in [0, T_1]) \\ &\quad + \mathbb{P}_y\left(\max_{0 \leq i \leq nT(n)} \left|X_{i+1}^n - X_i^n\right| \geq \eta\right) \\ &\leq e^{-n\left(\bar{W} - \frac{\varepsilon}{2}\right)} + 2e^{-n(\bar{W}+1)}.\end{aligned}$$

From (8.8) and (8.9) it follows that

$$\mathbb{P}_y(\rho^n \leq T(n)) \leq \frac{T(n) + T_1}{s}(e^{-n(\bar{W} - \frac{\varepsilon}{2})} + 2e^{-n(\bar{W}+1)}).$$

We can now take logarithms and scale by n. By using Condition 8.6 we can choose $N_5 \in (N_4, \infty)$, so that $n \geq N_5$ implies

$$\frac{1}{n} \log \mathbb{P}_y(\rho^n \leq T(n)) \leq -\bar{W} + \frac{\varepsilon}{2} + \frac{\varepsilon}{8}.$$

The additional error term of $\varepsilon/8$ appears to account for the polynomial factor and the additional exponential term which decays at rate $\bar{W} + 1$.

Since $|x - y| < 2\mu$ whenever $x, y \in \Gamma_\mu^\eta$, (8.3) ensures that $-\bar{W} \leq -W(y) + \varepsilon/8$. We conclude that whenever $n \geq N_5$ and $y \in \Gamma_\mu^\eta$, we have

$$\frac{1}{n} \log \mathbb{P}_y(\rho^n \leq T(n)) \leq -\bar{W} + \frac{\varepsilon}{2} + \frac{\varepsilon}{8} \leq -W(y) + \frac{3\varepsilon}{4}. \tag{8.12}$$

In particular, (8.2) holds for $y \in \Gamma_\mu^\eta$.

Step 5. *Extend estimate to $y \in \mathcal{E}_\mu$.* The estimate (8.12), which we have just shown holds for all $y \in \Gamma_\mu^\eta$, can be harnessed to obtain (8.2) for all $y \in \mathcal{E}_\mu$. Let $z \in \mathcal{E}_\mu \setminus \Gamma_\mu^\eta$, and define the stopping time

$$H_\Gamma^n \doteq \inf\{t \geq 0 : X^n(t) \in \Gamma_\mu^\eta\}.$$

In order to escape $\mathcal{D}$, a trajectory starting at z must pass through Γ_μ^η, or else jump over it. In the latter case we have

$$\mathbb{P}_z(\rho^n < H_\Gamma^n) \leq \mathbb{P}_z\left(\max_{i=0,\ldots,nT(n)} |X_{i+1}^n - X_i^n| \geq \eta\right).$$

Owing to Lemma 8.7, the probability of jumping over Γ_μ^η is superexponentially small: we can choose $N_6(\mu, \eta) < \infty$ so large that

$$\mathbb{P}_z\left(\max_{i=0,\ldots,nT(n)} |X_{i+1}^n - X_i^n| \geq \eta\right) \leq e^{-n(\bar{W}+1)},$$

for all $n \geq N_6$. By the law of total probability,

$$\begin{aligned}\mathbb{P}_z(\rho^n \leq T(n)) &= \mathbb{P}_z(\rho^n \leq T(n) \mid \rho^n < H_\Gamma^n)\, \mathbb{P}_z(\rho^n < H_\Gamma^n) \\ &\quad + \mathbb{P}_z(\rho^n \leq T(n) \mid \rho^n > H_\Gamma^n)\, \mathbb{P}_z(\rho^n > H_\Gamma^n) \\ &\leq e^{-n(\bar{W}+1)} + \mathbb{P}_{X^n(H_\Gamma^n)}(\rho^n \leq T(n) - H_\Gamma^n), \end{aligned} \tag{8.13}$$

where the last step follows from the strong Markov property applied at H_Γ^n. Since $X^n(H_\Gamma^n) \in \Gamma_\mu^\eta$ by definition of H_Γ^n, and since $H_\Gamma^n \geq 0$, it also holds that

$$\mathbb{P}_{X^n(H_\Gamma^n)}(\rho^n \leq T(n) - H_\Gamma^n) \leq \sup_{y \in \Gamma_\mu^\eta} \mathbb{P}_y(\rho^n \leq T(n)).$$

Recall that the choice of $\mu > 0$ in (8.3) ensures $W(z) \leq W(y) + \varepsilon/8$. Using (8.12) in (8.13), we obtain for the given $z \in \mathcal{E}_\mu \setminus \Gamma_\mu^\eta$ and $n \geq \max(N_5, N_6)$ that

$$\mathbb{P}_z(\rho^n \leq T(n)) \leq e^{-n(\bar{W}+1)} + e^{-n(W(y)-\frac{3\varepsilon}{4})} \leq 2e^{-n(W(z)-7\varepsilon/8)}.$$

By choosing n large enough and applying the logarithmic scaling, we can absorb the coefficient of 2 into the remaining $\varepsilon/8$ of room for error. This establishes the desired estimate for $z \in \mathcal{E}_\mu \setminus \Gamma_\mu^\eta$ and thus for all $y \in \mathcal{E}_\mu$. □

The next lemma extends the asymptotic upper bound to all points in $\mathcal{D}$. It is stated in its uniform version as required for the analysis of RESTART.

Lemma 8.13. *Under the conditions of Theorem 8.11,*

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}_z(\rho^n \leq T(n)) \leq -W(z), \tag{8.14}$$

uniformly in $z \in \mathcal{D}$.

Proof. Let $\varepsilon > 0$. To establish the lemma, we claim it suffices to show that for any $z \in \mathcal{D}$, there exist $\alpha(z, \varepsilon) > 0$ and $N(z, \varepsilon) < \infty$ such that

$$\frac{1}{n} \log \mathbb{P}_y(\rho^n \leq T(n)) \leq -W(y) + \varepsilon, \tag{8.15}$$

for all $n \geq N(z, \varepsilon)$ and $y \in B_{\alpha(z,\varepsilon)}(z) \doteq \{y : |z - y| < \alpha(z, \varepsilon)\}$.

To justify the claim, first observe that for any choice of $\alpha(z, \varepsilon) > 0$, we have $\overline{\mathcal{D}} \subset \cup_{z \in \mathcal{D}} B_{\alpha(z,\varepsilon)}(z)$. Since $\overline{\mathcal{D}}$ is compact, there are $K < \infty$ and $\{z_i\}_{i=1}^K \subset \mathcal{D}$ such that $\overline{\mathcal{D}} \subset \cup_{i=1}^K B_{\alpha(z_i,\varepsilon)}(z_i)$. If $N \doteq \max_{i=1}^K N(z_i, \varepsilon) < \infty$, then (8.15) holds for any $n \geq N$ and any $y \in \mathcal{D}$. Since $\varepsilon > 0$ is arbitrary, (8.14) follows.

We now establish (8.15). The particular case of $z = 0$ is covered by Lemma 8.12, for which one can simply take $\alpha = \mu$, so it suffices to consider all other points.

Next, we eliminate the case where $W(z) = 0$. Since W is continuous on $\mathbb{R}^d$, there is $\alpha(z, \varepsilon) > 0$ such that $0 < -W(y) + \varepsilon$ for all $y \in B_\alpha(z)$. Since $\mathbb{P}_y(\rho^n \leq T(n)) \leq 1$, it automatically follows that (8.15) holds for all $y \in B_\alpha(z)$,

as required. Furthermore, we observe that $W(z) = 0$ for any $z \notin \mathcal{D}$, and so we may focus on points $z \in \mathcal{D} \setminus \{0\}$.

Let then $z \in \mathcal{D} \setminus \{0\}$ with $W(z) > 0$, and let $\varepsilon > 0$ be given. Without any loss of generality, we may assume that $0 < \varepsilon < W(z)$. By Lemma 8.12, there is $\mu_0(\varepsilon) > 0$ and $N_1(\varepsilon) < \infty$ such that $n \geq N_1$ implies, for all $y \in \mathcal{E}_{\mu_0} = \{x : |x| < \mu_0\}$, that

$$\frac{1}{n} \log \mathbb{P}_y(\rho^n \leq T(n)) \leq -W(y) + \frac{\varepsilon}{4}. \tag{8.16}$$

We claim that we can choose $\mu(z, \varepsilon) \in (0, \mu_0)$ and $\alpha(z, \varepsilon)$ such that: (1) $B_\alpha(z) \subset \mathcal{D} \setminus \mathcal{E}_\mu$ and (2) we have the inequality

$$W(y) \leq \inf_{x \in \mathcal{E}_\mu} W(x) + \frac{\varepsilon}{2}, \tag{8.17}$$

for all $y \in B_\alpha(z)$.

We first choose μ. By part 2 of Lemma 8.9 we have $W(z) \leq W(0)$, and by the first part of the same lemma W is continuous at 0. Thus we can choose $\mu > 0$ such that $W(0) < W(x) + \varepsilon/4$ for all $x \in \mathcal{E}_\mu$, and without loss of generality $\mu < \min(|z|/2, \mu_0)$. In particular,

$$W(z) \leq \inf_{x \in \mathcal{E}_\mu} W(x) + \frac{\varepsilon}{4}. \tag{8.18}$$

Next, since $\mu < |z|/2$ we can take $\alpha(z, \varepsilon, \mu) > 0$ small enough that $B_\alpha(z) \cap \mathcal{E}_\mu = \emptyset$ and moreover we may insist that $B_\alpha(z) \subset \mathcal{D}$ because $z \in \mathcal{D}$ and $\mathcal{D}$ is open. Since W is continuous at z, α can be taken even smaller to guarantee

$$|W(y_1) - W(y_2)| < \varepsilon/4, \tag{8.19}$$

for all $y_1, y_2 \in B_\alpha(z)$. Note that μ depends only on z and ε, so α inherently depends only on z and ε as well. Combined with (8.18), an application of the triangle inequality shows that this choice of α guarantees (8.17).

Define

$$\zeta_\mu^n \doteq \inf\{t > 0 : X^n(t) \notin \mathcal{D} \setminus \mathcal{E}_\mu\},$$

which is the first time the process enters the μ-neighborhood of the attracting point 0 or escapes the set $\mathcal{D}$. By Lemma 8.10, there are $T_1(\mu) < \infty$ and $N_2(\mu) < \infty$ so large that $n \geq N_2$ implies, for any $x \notin \mathcal{E}_\mu$

$$\mathbb{P}_x(\zeta_\mu^n > T_1) \leq e^{-n(W(x)+1)}. \tag{8.20}$$

Then, for any $y \in B_\alpha(z)$,

$$\begin{aligned} \mathbb{P}_y(\rho^n \leq T(n)) = \mathbb{P}_y(\rho^n \leq T_1) + \mathbb{P}_y(T_1 < \rho^n \leq T(n), \zeta_\mu^n \leq T_1) \\ + \mathbb{P}_y(T_1 < \rho^n \leq T(n), \zeta_\mu^n > T_1). \end{aligned} \tag{8.21}$$

We estimate each term separately. For the first term, we use the finite time uniform large deviations upper bound, part 3 of Definition 8.3. Let

$\bar{W} \doteq \inf_{y \in B_\alpha(z)} W(y)$. Since the closure of $B_\alpha(z)$ is compact, an argument similar to the one in Step 1 of Lemma 8.12 shows that for the given $\varepsilon > 0$ there is $\delta(\varepsilon) > 0$ such that, for all $y \in B_\alpha(z)$,

$$\{\rho^n \le T_1, X^n(0) = y\} \subset \{d(X^n, \Phi_{y,T_1}(\bar{W} - \varepsilon/8)) \ge \delta\}.$$

Applying the large deviations upper bound on the compact set $\overline{B_\alpha(z)}$ with $s_0 = \bar{W} - \varepsilon/8$, $\gamma = \varepsilon/8$ and $\delta > 0$ as above, we find $N_3(\varepsilon, T_1) < \infty$ for which $n \ge N_3$ implies, for all $y \in B_\alpha(z)$, $\mathbb{P}_y(\rho^n \le T_1) \le e^{-n(\bar{W}-\varepsilon/4)}$. By the choice of α in (8.19) we have $W(y) \le \bar{W} + \varepsilon/4$, and so

$$\mathbb{P}_y(\rho^n \le T_1) \le e^{-n(\bar{W}-\varepsilon/4)} \le e^{-n(W(y)-\varepsilon/2)}, \tag{8.22}$$

for all $y \in B_\alpha(z)$.

For the second term, note that on the event $\{T_1 < \rho^n \le T(n)\} \cap \{\zeta_\mu^n \le T_1\}$, the process has entered the μ-neighborhood of 0, which is contained in the μ_0-neighborhood of 0 by the choice of μ. After entering $\mathcal{E}_\mu$ at time ζ_μ^n, it has $T(n) - \zeta_\mu^n$ time remaining to exit $\mathcal{D}$. By allowing $T(n)$ time we increase the probability of exiting. Thus, by (8.16) and the strong Markov property, $n \ge N_1$ implies

$$\begin{aligned}\mathbb{P}_y(T_1 < \rho^n \le T(n), \zeta_\mu^n < T_1) &\le \sup_{x \in \mathcal{E}_\mu} \mathbb{P}_x(\rho^n \le T(n)) \\ &\le e^{-n(\inf_{x \in \mathcal{E}_\mu} W(x) - \varepsilon/4)}.\end{aligned} \tag{8.23}$$

For the last term in (8.21), we use the estimate (8.20),

$$\mathbb{P}_y(T_1 < \rho^n \le T(n), \zeta_\mu^n > T_1) \le \mathbb{P}_y(\zeta_\mu^n > T_1) \le e^{-n(W(y)+1)}, \tag{8.24}$$

which holds for all $y \in B_\alpha(z) \subset \mathcal{D} \setminus \mathcal{E}_\mu$ whenever $n \ge N_2$.

Set $N_4 \doteq \max(N_1, N_2, N_3)$. Using (8.22), (8.23) and (8.24) in (8.21), we find for all $n \ge N_4$ and $y \in B_\alpha(z)$

$$\mathbb{P}_y(\rho^n \le T(n)) \le e^{-n(W(y)-\varepsilon/2)} + e^{-n(\inf_{x \in \mathcal{E}_\mu} W(x) - \varepsilon/4)} + e^{-n(W(y)+1)}. \tag{8.25}$$

From (8.17), for any $y \in B_\alpha(z)$, $e^{-n(\inf_{x \in \mathcal{E}_\mu} W(x) - \varepsilon/4)} \le e^{-n(W(y)-3\varepsilon/4)}$ for all $n \ge N_4$. It therefore follows from (8.25) that, for all $y \in B_\alpha(z)$,

$$\frac{1}{n} \log \mathbb{P}_y(\rho^n \le T(n)) \le -W(y) + \frac{3\varepsilon}{4} + \frac{3}{n},$$

and choosing $N_4 \le N(z, \varepsilon) < \infty$ large enough that $3/n < \varepsilon/4$, we obtain (8.15) for all $y \in B_\alpha(z)$, as required. □

Next we prove the lower bound for the decay of $\mathbb{P}_x(\rho^n \le T(n))$. We remind the reader that Lemma 8.14 also holds uniformly in x, but in the interest of space we only establish the pointwise version.

Lemma 8.14. *Assume the conditions of Theorem 8.11. For any $x \in \mathcal{D}$, we have*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}_x(\rho^n \le T(n)) \ge -W(x).$$

Proof. It suffices to show that for any $x \in \mathcal{D}$ and any $\varepsilon > 0$, there is $N(\varepsilon) < \infty$ such that $n \ge N$ guarantees

$$\frac{1}{n} \log \mathbb{P}_x(\rho^n \le T(n)) \ge -W(x) - \varepsilon.$$

To establish the estimate, we will take a trajectory which is within ε of the infimum in the definition of $W(x)$, and use part 3 of Condition 8.2 in conjunction with part 2 of Condition 8.5 to extend the trajectory a little past the boundary. This will allow us to apply the large deviations lower bound, which is stated in its uniform version as part 2 of Definition 8.3.

Fix $x \in \mathcal{D}$ and let $\varepsilon > 0$. By the variational definition of $W(x)$, there is a trajectory φ and $T < \infty$ such that $\varphi(0) = x$, $\varphi(T) \notin \mathcal{D}$ and $I_T(\varphi) \le W(x) + \varepsilon/4$. Define the compact set $K = \{y : \inf_{z\in\mathcal{D}} |y - z| \le 1\}$. By Lemma 8.8, there is $c \in (0, \infty)$ such that for all $p, q \in K$, we can find φ_{pq} and τ which satisfy $\varphi_{pq}(0) = p$, $\varphi_{pq}(\tau) = q$, $\tau \in (0, \infty)$ and $I_\tau(\varphi_{pq}) \le c|p - q|$. Let $a \in (0, \varepsilon/4c)$ and assume without loss of generality that $a < 1$.

$I_T(\varphi) < \infty$ implies φ must be absolutely continuous, so there is a time $s \in (0, T]$ such that $\varphi(s) \in \partial\mathcal{D}$. Set $p \doteq \varphi(s)$ and note that $p \in K$. By the regularity property in part 2 of Condition 8.5, there is a point q in the interior of $\mathcal{D}^c$ which satisfies $|p - q| < a$, and since $a < 1$ we also have $q \in K$. From Lemma 8.8 we obtain a trajectory φ_{pq} and $\tau > 0$ as described above with cost

$$I_\tau(\varphi_{pq}) \le ca < \varepsilon/4.$$

Let ϕ denote the concatenation of the paths φ on $[0, s]$ and φ_{pq} on $(s, s+\tau]$, that is,

$$\phi(t) = \begin{cases} \varphi(t) & t \in [0, s], \\ \varphi_{pq}(t - s) & t \in (s, s + \tau], \end{cases}$$

and note that ϕ is continuous at s. Moreover,

$$I_{s+\tau}(\phi) \le I_T(\varphi) + I_\tau(\varphi_{pq}) \le W(x) + \varepsilon/2.$$

Finally, since q lies in the interior of $\mathcal{D}^c$, there is $\delta > 0$ such that $|q - z| < \delta$ implies $z \in \mathcal{D}^c$. In particular, if $\|\phi' - \phi\|_\infty < \delta$, then $\phi'(s+\tau) \in \mathcal{D}^c$. If n is large enough to guarantee $T(n) \ge s + \tau$, then

$$\{\|X^n - \phi\|_\infty < \delta\} \subset \{\rho^n \le T(n)\}.$$

Applying part 2 of Definition 8.3 with $s_0 = W(x) + 1$, $\gamma = \varepsilon/2$, and the given $\delta > 0$, we find $N < \infty$ such that for all $n \ge N$,

$$\mathbb{P}_x(\|X^n - \phi\|_\infty < \delta) \ge \exp(-n(I_T(\phi) + \varepsilon/2)).$$

Taking logarithms and using $I_T(\phi) \le W(x) + \varepsilon/2$, we obtain the desired result. □

8.3 RESTART

Before presenting the result in [10] regarding RESTART, we first point out that the increasing time intervals $[0, T(n)]$ can indeed ruin asymptotic efficiency when using importance sampling instead of splitting. Let $\mathfrak{S}^n(\bar{V})$ denote the second moment of the estimator when using some subsolution $\bar{V}(x)$, see Definition 5.4 for the definition of a subsolution. It has been shown in [20], where stochastic differential equation models are considered, that when importance sampling is based on a time-independent subsolution, the following lower bound on the second moment holds:

$$\mathfrak{S}^n(\bar{V}) \geq e^{C_1(T-K)} e^{-nC_2}.$$

Here T is the time interval for escape, C_1 and C_2 are positive constants, and $K < T$ is a fixed constant. If T were n dependent, for example, $T(n) = n^2$, then

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathfrak{S}^n(\bar{V}) = \infty.$$

Not only is the estimator not asymptotically efficient, the decay rate of the second moment of the estimator of this importance sampling scheme is worse than ordinary Monte Carlo simulation.

For one-dimensional stochastic differential equations, it is possible to construct time-dependent subsolutions which ensure that the importance sampling schemes do not degrade over long time intervals, see [20]. These time-dependent subsolutions are difficult to construct and it is not known how to construct them in any generality for higher dimensions.

In [10] it is shown that these problems do not arise with splitting. As discussed in [20], the lower decay rate of importance sampling is due to the exponential growth of the likelihood ratio when the trajectory stays near the origin, which it can do with large enough probability over a long time interval. In contrast, such trajectories do not affect the decay rate of splitting schemes since they are not multiplied by a likelihood ratio; all that matters is that on average, out of R particles born in a threshold, one of them reaches a threshold with a lower index.

The analysis in [10] is carried out for the RESTART scheme, but for ordinary splitting the result will be the same and the analysis will be easier. By using techniques similar to the ones in [16] it is straightforward to get estimates on the second moment. The main result from [10] is stated in the following theorem. It places a lower bound on the asymptotic decay rate of the second moment of the estimator in terms of the value of the subsolution at the origin. For a precise definition of subsolution and the piecewise constant importance function, we refer to [10].

Theorem 8.15. *Suppose Conditions 8.1, 8.2, 8.4, 8.5 and 8.6 are met. Let $\bar{V}$ be a subsolution, and let $\mathfrak{S}^n(\bar{V})$ denote the second moment of the estimator when $\bar{V}$ is used to determine the piecewise constant importance functions. We have*

the following lower bound:

$$\liminf_{n\to\infty} -\frac{1}{n}\log \mathfrak{S}^n(\bar{V}) \geq \bar{V}(0) + W(0).$$

Clearly, if $\bar{V}(0) = W(0)$, then the estimator is asymptotically efficient. This is in contrast to importance sampling, which performs worse when using the same time-independent subsolution. Numerical results in [10] support these findings.

8.4 Appendix

8.4.1 Proofs from Section 8.2

In this section, we present the proof of Lemma 8.9, which we copy for convenience.

Lemma 8.9. *Under Condition 8.2, the function $W(\cdot)$ satisfies*

1. *$W(\cdot)$ is continuous on $\mathbb{R}^d$,*
2. *for every $x \in \mathcal{D}$, $W(x) \leq W(0)$.*

Proof. First we prove that W is continuous at any point $x \in \mathcal{D}$. By Lemma 8.8 applied to the compact set $\bar{\mathcal{D}}$, there is $c \in (0,\infty)$ such that for any $x, y \in \mathcal{D}$ satisfying $|x - y| < \varepsilon/2c$, the cost of going from x to y in some finite time $\tau \in (0,\infty)$ is less than $\varepsilon/2$. By concatenating this path from x to y with a near-minimizing one from y to some point not in $\mathcal{D}$, and using additivity of I_T on disjoint time segments, we obtain $|W(x) - W(y)| < \varepsilon$.

If $x \in \overline{\mathcal{D}} \setminus \mathcal{D}$ then $W(x) = 0$, and the above proof can be repeated on a neighborhood N of x, separating into the cases $y \in N \cap \mathcal{D}$ and $y \notin N \cap \mathcal{D}$. Finally, W is constant on the open set $\overline{\mathcal{D}}^c$ and hence continuous there as well.

Next we show that for any $\varepsilon > 0$ and any $x \in \overline{\mathcal{D}}$, $W(x) \leq W(0) + \varepsilon$, which implies that $W(x) \leq W(0)$ for all $x \in \mathcal{D}$. Note that we might have $W(x) = 0$, if the noiseless dynamics starting at x pass through the complement of $\mathcal{D}$ before reaching any sufficiently small neighborhood of the origin, in which case $W(x) \leq W(0)$ automatically since $W(0) \geq 0$. Otherwise, since W is continuous at 0, there is $\delta > 0$ such that any $y \in \mathcal{E}_\delta = \{x : |x| < \delta\}$ satisfies $W(y) < W(0) + \varepsilon$. Since 0 is the unique attracting fixed point, for any $x \in \mathcal{D}$ there is $T > 0$ and a trajectory $\varphi \in C([0,T])$ with $\varphi(0) = x$ such that $I_T(\varphi) = 0$ and $\varphi(T) \in \mathcal{E}_\delta$. Thus,

$$W(x) \leq I_T(\varphi) + W(\varphi(T)) < W(0) + \varepsilon,$$

as claimed. □

8.4.2 Proofs from Section 8.3

In this section, we present the proof of two lemmas that were used in the proof of Lemma 8.12.

Lemma 8.16. *For all $y \in \Gamma_\mu^\eta$,*

$$\mathbb{E}_y\left[\sum_{i=1}^{M^n} B_i^n\right] \le \mathbb{E}_y\left[M^n\right] \sup_{x\in\Gamma_\mu^\eta} \mathbb{E}_x\left[B_1^n\right], \tag{8.26}$$

and

$$\mathbb{E}_y\left[\sum_{i=1}^{M^n} \tau_i^n - \sigma_{i-1}^n\right] \ge \mathbb{E}_y\left[M^n\right] \inf_{x\in\Gamma_\mu^\eta} \mathbb{E}_x\left[\tau_1^n\right]. \tag{8.27}$$

Proof. Consider

$$S_k^n = \sum_{i=1}^{k} \tau_i^n - \sigma_{i-1}^n.$$

Let $\mathcal{F}_i^n = \sigma(X_j^n, 0 \le j \le i)$, $\mathcal{F} = \sigma(\cup_{i=1}^\infty \mathcal{F}_i^n)$ and $\mathcal{G}_k^n = \mathcal{F}_{\sigma_k^n}^n$, where

$$\mathcal{F}_{\sigma_k^n}^n = \{A \in \mathcal{F} : A \cap \{\sigma_k^n \le m\} \in \mathcal{F}_m^n, m \ge 1\},$$

is the sigma-algebra of the stopping time σ_k^n. Since $\sigma_i^n \ge \tau_i^n$, S_k^n is $\mathcal{G}_k^n$-measurable. Let

$$M^n \doteq \inf\{k \ge 1 : S_k^n > T(n)\}.$$

Then M^n is a stopping time with respect to the filtration generated by S_k^n and hence $\mathcal{G}_k^n$. In particular, $\{M^n \ge k\} = \{M^n \le k-1\}^c$ is $\mathcal{G}_{k-1}^n$-measurable. We have

$$\begin{aligned}
\mathbb{E}_y\left[\sum_{i=1}^{M^n} \mathbb{E}_{X^n_{\sigma^n_{i-1}}} B_1^n\right] &= \mathbb{E}_y\left[\sum_{i=1}^{M^n} \mathbb{E}_y[B_i^n | X^n_{\sigma^n_{i-1}}]\right] = \mathbb{E}_y\left[\sum_{i=1}^{M^n} \mathbb{E}_y[B_i^n | \mathcal{G}_{i-1}^n]\right] \\
&= \sum_{i=1}^{\infty} \mathbb{E}_y[1_{\{M^n \ge i\}} \mathbb{E}_y[B_i^n | \mathcal{G}_{i-1}^n]] = \sum_{i=1}^{\infty} \mathbb{E}_y[1_{\{M^n \ge i\}} B_i^n] \\
&= \mathbb{E}_y\left[\sum_{i=1}^{M^n} B_i^n\right].
\end{aligned}$$

In the above display, the first two equalities follow from the strong Markov property conditioned on $X^n_{\sigma^n_{i-1}}$, and pulling out the infinite sum from the expectation sign is permitted by Tonelli's theorem because all terms in the sum are positive. The result in (8.26) immediately follows. Similarly, we have

$$\mathbb{E}_y\left[\sum_{i=1}^{M^n} \tau_i^n - \sigma_{i-1}^n\right] = \mathbb{E}_y\left[\sum_{i=1}^{M^n} \mathbb{E}_{X^n_{\sigma^n_{i-1}}} \tau_1^n\right],$$

and thus (8.27) follows. □

Lemma 8.17. *There exist $s > 0$ and $N < \infty$ such that for all $n \ge N$ and any $y \in \Gamma_\mu^\eta$, we have $\mathbb{P}_y(\tau_1^n \le s) \le \frac{1}{2}$.*

Proof. We may assume that $s \leq 1$ without any loss, since the probability is decreased by decreasing s. The probability $\mathbb{P}_y(\tau_1^n \leq s)$ is the probability that X^n exits $\mathcal{D}$ or enters γ_μ^η before time s. If instead of the piecewise constant interpolation of X^n we had taken the piecewise continuous one, it would not be possible for X^n to jump over the set γ_μ^η and the probability $\mathbb{P}_y(\tau_1^n \leq s)$ would be increased. Consequently, it suffices to establish the desired inequality when X^n is the piecewise continuous interpolation. For the purposes of this proof it is convenient to do so, and by an abuse of notation we use X^n to denote this process, so that $X^n \in C([0,1])$ $\mathbb{P}_y$-almost surely for the remainder of the proof.

Let $\Delta = \min(\mu/2 - \eta, \operatorname{dist}(\mathcal{E}_\mu, \partial\mathcal{D})) > 0$ where $\mathcal{E}_\mu = \{x : |x| < \mu\}$ and dist is the Euclidean distance. Any trajectory which starts in Γ_μ^η and exits $\mathcal{D}$ or enters γ_μ^η must travel a distance of at least $\Delta > 0$ at some point in time. Let

$$A_s = \{\varphi \in C([0,s]) : |\varphi(t) - \varphi(0)| \geq \Delta \text{ for some } t \in [0,s], \varphi(0) \in \Gamma_\mu^\eta\}.$$

Then $\{\tau_1^n \leq s\} \subset \{X^n \in A_s\}$. To get an upper bound on the probability of $\{X^n \in A_s\}$ we place a lower bound on the cost of any $\varphi \in A_s$.

By Condition 8.1, $\bar{H}(\alpha) = \sup_{x \in \mathbb{R}^d} H(x,\alpha)$ is finite for all $\alpha \in \mathbb{R}^d$. It is also convex, and its Legendre transform $\bar{L}(\beta) = \sup_{\alpha \in \mathbb{R}^d}\{\langle \alpha, \beta\rangle - \bar{H}(\alpha)\}$ satisfies $\bar{L}(\beta) \leq L(x,\beta)$ for all $x \in \mathbb{R}^d$. The finiteness of $\bar{H}$ implies that $\bar{L}$ is superlinear:

$$\lim_{c \to \infty} \inf_{\beta : |\beta| \geq c} \frac{1}{c} \bar{L}(\beta) = \infty.$$

Choose $c \in (0,\infty)$ such that $\inf_{\beta:|\beta|\geq c} \frac{1}{c}\bar{L}(\beta) \geq 1$, and then choose $s > 0$ such that $\Delta/s = c$. Then for any $\varphi \in A_s$ we have either $I_s(\varphi) = \infty$ (when φ is not absolutely continuous) or

$$\int_0^s \bar{L}(\dot{\varphi}(r))dr \geq \int_0^t \bar{L}(\dot{\varphi}(r))dr \geq t\bar{L}\left(\frac{1}{t}\int_0^t \dot{\varphi}(r)dr\right) \geq t\frac{\Delta}{t} = \Delta,$$

where $t \in (0,s]$ is a point such that $|\varphi(t) - \varphi(0)| \geq \Delta$ and therefore $|\int_0^t \dot{\varphi}(r)dr/t| \geq \Delta/t$. Since $\bar{L}(\beta) \leq L(x,\beta)$ for all $x \in \mathbb{R}^d$, it follows that $I_s(\varphi) \geq \Delta$. According to Theorem 1.1 in [17] applied to the closed set A_s, for any $\varepsilon > 0$ there is $N(\varepsilon) < \infty$ such for any $x \in \Gamma_\mu^\eta$ and any $n \geq N$

$$P_x(X^n \in A_s) \leq \exp\left(-n\left(I_s(\varphi) - \varepsilon\right)\right).$$

If we choose $\varepsilon = \Delta/2$ and use $I_s(\varphi) \geq \Delta$, then for n large enough we find $P_x\{X^n \in A_s\} \leq 1/2$. □

Bibliography

[1] R.R. Bahadur. A note on the fundamental identity of sequential analysis. *Annals of Mathematical Statistics*, 29 (2): 534–543, 1958.

[2] D. Bertsimas, I.C. Paschalidis, and J.N. Tsitsiklis. On the large deviations behavior of acyclic networks of $G|G|1$ queues. *Annals of Applied Probability*, 8 (4): 1027–1069, 1998.

[3] D. Bertsimas, I.C. Paschalidis, and J.N. Tsitsiklis. Large deviations analysis of the generalized processor sharing policy. *Queueing Systems*, 32 (4): 319–349, 1999.

[4] A. Budhiraja and P. Dupuis. Analysis and Approximation of Rare Events: Representations and Weak Convergence Methods, 2019. To appear.

[5] A. Buijsrogge, P.T. de Boer, K. Rosen, and W.R.W. Scheinhardt. Large deviations for the total queue size in non-Markovian tandem queues. *Queueing Systems*, 85 (3): 305–312, 2017.

[6] A. Buijsrogge, P.T. de Boer, and W.R.W. Scheinhardt. Importance sampling for Markovian tandem queues using subsolutions: exploring the possibilities. *Submitted*.

[7] A. Buijsrogge, P.T. de Boer, and W.R.W. Scheinhardt. Importance sampling for non-Markovian tandem queues using subsolutions. *Submitted*.

[8] A. Buijsrogge, P.T. de Boer, and W.R.W. Scheinhardt. On state-independent importance sampling for the $GI|GI|1$ tandem queue. *Accepted for publication in Probability in the Engineering and Informational Sciences*.

[9] A. Buijsrogge, P.T. de Boer, and W.R.W. Scheinhardt. Splitting for non-Markovian tandem queues using subsolutions. *Working paper*.

[10] A. Buijsrogge, P. Dupuis, and M. Snarski. Splitting algorithms for rare event simulation over long time intervals. *Submitted*.

[11] P.T. de Boer. Analysis and efficient simulation of queueing models of telecommunication systems. PhD thesis, University of Twente, 2000.

[12] 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.

[13] P.T. de Boer. Some observations on importance sampling and RESTART. In *Proc. of the 6th International Workshop on Rare Event Simulation*, 2006.

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

[15] T. Dean and P. Dupuis. The design and analysis of a generalized RESTART/DPR algorithm for rare event simulation. *Annals of Operations Research*, 189: 63–102, 2011.

[16] T. Dean and P. Dupuis. Splitting for rare event simulation: A large deviation approach to design and analysis. *Stochastic processes and their applications*, 119 (2): 562–587, 2009.

[17] P. Dupuis, R. Ellis, and A. Weiss. Large deviations for Markov processes with discontinuous statistics, I: General upper bounds. *Annals of Probability*, 19 (3): 1280–1297, 1991.

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

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

[20] P. Dupuis, K. Spiliopoulos, and X. Zhou. Escaping from an attractor: importance sampling and rest points I. *Annals of Applied Probability*, 25 (5): 2909–2958, 2015.

[21] P. Dupuis and H. Wang. Subsolutions of an Isaacs equation and efficient schemes for importance sampling. *Mathematics of Operations Research*, 32 (3): 723–757, 2007.

[22] P. Dupuis and H. Wang. Importance sampling for Jackson networks. *Queueing Systems*, 62 (1): 113–157, 2009.

[23] S. Ethier and T. Kurtz. Markov Processes: Characterization and Convergence. *Wiley*, New York, 1986.

[24] S.G. Foss. On the exact asymptotics for the stationary sojourn time distribution in a tandem of queues with light-tailed service times. *Problems of Information Transmission*, 43 (4): 353–366, 2007.

[25] M.R. Frater and B.D.O. Anderson. Fast simulation of buffer overflows in tandem networks of $GI|GI|1$ queues. *Annals of Operations Research*, 49: 207–220, 1994.

[26] M.I. Freidlin and A.D. Wentzell. Random perturbations of dynamical systems, volume 260 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. *Springer-Verlag, New York*, 1998.

[27] A.J. Ganesh. Large deviations of the sojourn time for queues in series. *Annals of Operations Research*, 79: 3–26, 1998.

[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. A large deviations perspective on the efficiency of multilevel splitting. *IEEE Transactions on Automatic Control*, 43 (12): 1666–1679, 1998.

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

[31] P.W. Glynn and W. Whitt. The asymptotic efficiency of simulation estimators. *Operations research*, 40 (3): 505–520, 1992.

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

[33] 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.

[34] D.I. Miretskiy, W.R.W. Scheinhardt, and M.R.H. Mandjes. On efficiency of multilevel splitting. *Communications in Statistics-Simulation and Computation*, 41 (6): 890–904, 2012.

[35] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control*, 34 (1): 54–66, 1989.

[36] 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.

[37] A. Shwartz and A. Weiss. Large deviations for performance analysis: queues, communication and computing. *CRC Press*, 1995.

[38] S. van den Heuvel. Simulatie van zeldzame gebeurtenissen. BSc thesis, University of Twente, 2017.

[39] J. Villén-Altamirano. RESTART vs Splitting: A comparative study. *Performance Evaluation*, 121: 38–47, 2018.

[40] M. Villen-Altamirano and J. Villen-Altamirano. RESTART: A method for accelerating rare event simulations. In *Proc. of the 13th International Teletraffic Congress, Queueing, Performance and Control in ATM*, 71–76, Amsterdam, 1991. Elsevier.

[41] R.R. Weber. The interchangeability of $\cdot|M|1$ queues in series. *Journal of Applied Probability*, 16 (3): 690–695, 1979.

Summary

The main focus of this thesis is rare event simulation for non-Markovian tandem queues with i.i.d. arrival process and light-tailed service processes that are independent of each other. We are interested in developing simulation schemes in order to estimate the probability that the total number of customers reaches some high number during a busy cycle of the system. As ordinary simulation does not give the desired estimate in reasonable time, we consider two methods in order to decrease the simulation time: *importance sampling* and *splitting*.

In order to use either of these methods, we first determine the *large deviations behavior* of the probability of interest. In addition, we show that the large deviations behavior of the probability of having a high number of customers in the system in stationarity, as well as upon arrival of a customer, are the same.

Secondly, we consider importance sampling. With importance sampling, we change the underlying probability distributions of the arrival process and service processes to speed up the simulation. The change of probability distributions is also called a *change of measure*. First, we discuss how to find a *state-independent* change of measure and we show that this change of measure is the only exponential state-independent change of measure that may give an asymptotically efficient estimator. Moreover, we provide necessary conditions for this change of measure to give an asymptotically efficient estimator. In order to find a change of measure that *always* gives an asymptotically efficient estimator, we consider *state-dependent* importance sampling. Based on the subsolution approach, we present a state-dependent change of measure and we prove that our proposed change of measure indeed gives an asymptotically efficient estimator.

Thirdly, we consider splitting. With splitting, we keep the underlying probability distributions the same, but we 'encourage' the process when it tends into the right direction, that is, when the process crosses some predetermined *thresholds*, the process *splits* into several independent processes that all continue according to the same dynamics as the original process. We develop splitting thresholds using the same underlying methods as for designing the importance sampling schemes and we prove that the proposed splitting thresholds result in an asymptotically efficient estimator.

In the course of the analysis, we also consider three different, but related topics. The first is to study the large deviations behavior of the probability of interest when customer sizes differ in each of the queues. Secondly, we study state-dependent importance sampling for Markovian tandem queues and we explore the possibilities for a state-dependent change of measure using subsolutions.

Lastly, we determine for a rather general class of stochastic processes the large deviations behavior of the probability that the process leaves a neighborhood of a metastable point during some long time interval $[0, T]$ when the time interval depends on the rarity parameter.

Samenvatting

In dit proefschrift bestuderen we het simuleren van zeldzame gebeurtenissen in niet-Markovse wachtrijen in serie, waarbij we aannemen dat het aankomstproces en alle bedieningsprocessen onafhankelijk en identiek verdeeld, en onderling onafhankelijk zijn. Hierbij beschouwen we alleen dunstaartige bedieningsprocessen. We zijn geïnteresseerd in het schatten van de kans dat het totaal aantal klanten in deze wachtrijen groot wordt gedurende een periode dat het systeem niet leeg is. Dit doen we met behulp van computersimulatie. Aangezien gewone simulatie niet het gewenste resultaat binnen redelijk tijd oplevert, beschouwen we twee verschillende methoden die ervoor zorgen dat het gewenste resultaat mogelijk wél binnen redelijke tijd wordt bereikt: *importance sampling* en *splitting*.

Om gebruik te maken van deze methoden, bepalen we eerst het zogeheten *large deviations* gedrag van de kans waarin we geïnteresseerd zijn. Bij het bepalen van dit large deviations gedrag laten we ook zien dat het large deviations gedrag van twee gerelateerde kansen hetzelfde is, namelijk die van de kans om een groot totaal aantal klanten in het systeem te hebben in stationariteit, en bij aankomst van een klant.

Ten tweede beschouwen we importance sampling. Bij importance sampling veranderen we de kansmaten van het aankomstproces en de bedieningsprocessen om zo de simulatie sneller te laten verlopen. Als eerste bekijken we een *toestandsonafhankelijke* verandering van de kansmaat en laten we zien dat deze verandering van de kansmaat de enige exponentiële toestandsonafhankelijke verandering is die mogelijkerwijs een asymptotisch efficiënte schatter oplevert. Bovendien geven we noodzakelijke voorwaarden waaronder deze verandering van de kansmaat een asymptotisch efficiënte schatter zou kunnen opleveren. Om een verandering van de kansmaat te vinden die *altijd* een asymptotisch efficiënte schatter oplevert, ontwikkelen we vervolgens een *toestandsafhankelijke* verandering van de kansmaat. Gebaseerd op zogeheten *subsolutions* stellen we zo'n verandering van de kansmaat voor en laten we zien dat onze voorgestelde verandering van de kansmaat inderdaad een asymptotisch efficiënte schatter oplevert.

Ten derde beschouwen we splitting. Met splitting worden de onderliggende kansmaten niet veranderd, maar stimuleren we het proces wanneer het de 'juiste' kant op gaat. Hiermee bedoelen we dat wanneer het proces vooraf bepaalde drempels doorkruist, het zich splitst in meerdere onafhankelijke processen die allemaal weer verder gaan met dezelfde kansmaten als het originele proces. We ontwikkelen splittingdrempels waarbij we gebruik maken van soortgelijke methoden als wanneer we importance sampling gebruiken en we laten zien dat onze

voorgestelde splittingdrempels resulteren in een asymptotisch efficiënte schatter.

In dit proefschrift bekijken we ook nog drie andere, maar gerelateerde problemen. Het eerste is dat we het large deviations gedrag van de kans waarin we geïnteresseerd zijn onderzoeken wanneer de grootte van de klanten verandert als zij in een andere wachtrij zitten. Ten tweede bekijken we toestandsafhankelijke veranderingen van de kansmaat voor Markovse wachtrijen in serie en onderzoeken we de mogelijkheden voor zo'n verandering van de kansmaat wanneer we gebruik maken van subsolutions. Tot slot bepalen we voor een tamelijk algemene klasse van stochastische processen het large deviations gedrag van de kans dat het proces ontsnapt uit de buurt van een metastabiel punt gedurende een lang tijdsinterval $[0, T]$ wanneer het tijdsinterval afhangt van de zeldzaamheidsparameter.