Rare-event simulation for multidimensional stochastic models

Cahen, Ewan Jacov

Accepted/In press: 21/01/2019

Document Version
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Rare-event simulation for multidimensional stochastic models

Ewan Jacov Cahen

A catalogue record is available from the Eindhoven University of Technology Library
ISBN: 978-90-386-4682-4

Typeset by LaTeX

Printing: Ridderprint BV, the Netherlands
Cover: Remco Wetzels

Rare-event simulation for multidimensional stochastic models

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op maandag 21 januari 2019 om 11:00 uur

door

Ewan Jacov Cahen

geboren te Amsterdam

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof. dr. M.A. Peletier
1e promotor:	prof. dr. A.P. Zwart
2e promotor:	prof. dr. M.R.H. Mandjes (Universiteit van Amsterdam)
leden:	prof. dr. D.T. Crommelin (Universiteit van Amsterdam)
	dr. A.A.N. Ridder (Vrije Universiteit Amsterdam)
	prof. dr. I.J.F.B. Adan
	prof. dr. ir. S.C. Borst
	prof. dr. ir. O.J. Boxma

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Contents

1 Introduction **1**
1.1 Rare event simulation 2
1.2 Queueing processes and random walks 8
1.3 Large deviations theory 10
1.4 Complications in multiple dimensions 14
1.5 Contributions 16

2 A multi-dimensional ruin problem **19**
2.1 Introduction 19
2.2 Logarithmic asymptotics under Gärtner-Ellis conditions 22
2.3 Efficient estimation in a fluid model 29
2.3.1 Analysis of the fluid model 29
2.3.2 Simulation: construction of an efficient method 33
2.3.3 Numerical results 38
2.4 Outlook 41
2.5 Proof of Theorem 2.3.1 43

3 Large delay probabilities in two correlated queues **47**
3.1 Introduction 47

3.2 Model description and preliminaries 51
3.2.1 The model 51
3.2.2 Preliminaries from large deviations 52
3.3 Large deviations result 53
3.4 Importance sampling and efficiency 54
3.5 Partitioned importance sampling 58
3.6 Numerical results 61
3.6.1 The model 62
3.6.2 Variable level 64
3.6.3 Variable covariance 66
3.6.4 Variable number of intervals 67
3.7 Proof of Theorem 3.3.1 67
3.8 Concluding remarks 75

4 Linear stochastic fluid networks with Markov modulation 77
4.1 Introduction 77
4.2 Single resource, no modulation 81
4.2.1 Preliminaries 81
4.2.2 Tail probabilities, change of measure 82
4.2.3 Efficiency properties of the IS procedure 86
4.2.4 Simulation experiments 87
4.3 Multi-node systems, no modulation 88
4.3.1 Preliminaries 88
4.3.2 Tail probabilities, change of measure 92
4.3.3 Efficiency properties of the IS procedure 93
4.3.4 Simulation experiments 96
4.4 Multi-node systems under Markov modulation 98
4.4.1 Tail probabilities, change of measure 99
4.4.2 Efficiency properties of the IS procedure 102
4.4.3 Simulation experiments 103
4.5 Discussion and concluding remarks 107
4.6 Appendix A 108

5 The stationary limit of a multidimensional stochastic recursion 111
5.1 Introduction 111
5.2 One dimension 113
5.2.1 Fast estimation 114
5.2.2 Unbiased sampling 119
5.2.3 Numerical results 123
5.3 Multiple dimensions 123
5.3.1 Fast estimation 127
5.3.2 Unbiased sampling 129
5.3.3 Numerical results 132
5.4 Concluding remarks 132

Bibliography **135**

Summary **143**

Dankwoord – Acknowledgments **145**

About the author **147**

1 Introduction

Airplane crashes, earthquakes, stock market collapses, pandemics, large power grid failures... These are some examples of events that are rare, but that have a big impact on society nonetheless. Therefore, it is important to study these phenomena. One would like to know how often these events occur, what the impact is, how to prevent them from happening, how to predict them, etc.

In order to try to answer these questions, one can try to model these phenomena mathematically. These models can then be analysed in various ways. The preferred way is to calculate exactly the quantities of interest, such as the associated probabilities and most likely causes. However, even for relatively simple models, this might be very hard or even impossible. Therefore, other methods to analyse these models have to be used. One method is to perform numerical simulations, which is one of the main focus points of this thesis. Many methods for performing simulations exist, each with their advantages and disadvantages. The first method that comes to mind is (naive) Monte Carlo sampling. Due to its simplicity, this is a popular method. However, as will be demonstrated below, it is not well suited to estimate events that have a small probability, due to the high number of runs that is required. A sampling method that tries to remedy this is importance sampling, a method that is used in multiple chapters in this thesis. Simply put, importance sampling means simulating the model under a different probability measure. The biggest problem often lies in finding a new measure that performs well, in some yet to be defined sense. Often, large

deviations theory can provide a basis for finding a good new measure and as such, this thesis also provides several large deviations results. Furthermore, as one-dimensional models are often too simple to be of practical value, we turn to analysing multidimensional models. Contrary to their one-dimensional counterparts, multidimensional models have not been studied as much and they are often much harder to analyse. The focus in this thesis is on *two*-dimensional models, although many of the results carry over to higher dimensions.

As the reader of this thesis might not be familiar with the mathematical topics on which this thesis is based, we give an introduction to the required knowledge below. More specifically, we can roughly divide the required base knowledge into four topics, which have also been mentioned above: rare event simulation, queueing theory, large deviations theory and multidimensional processes. References to more in-depth literature will be provided as well. We close this chapter with an overview of the contributions made in this thesis.

1.1 Rare event simulation

Many mathematical models are hard, or even impossible, to analyse exactly. This is where numerical simulation comes in. With computers becoming faster every year, simulation is a viable way to obtain reasonably close approximations in a short time frame.

Turning to a probabilistic framework, suppose that in some model we want to estimate the probability of an event E; denote this by $p_E := \mathbb{P}(E)$. One of the most intuitive and easy methods to estimate this is to perform *Monte Carlo simulation*. In order to execute this, one samples N instances of the model. For each instance, one observes if the event E occurred; suppose we see N_s occurrences, or, successes. An unbiased estimator (recall that an estimator $\widehat{A}$ of some quantity A is *unbiased* if $\mathbb{E}\left(\widehat{A}\right) = A$) for p_E, $\widehat{p}_E$, then is

$$\widehat{p}_E := \frac{N_s}{N}. \tag{1.1}$$

A natural question that arises is: how good of an approximation of p_E is $\widehat{p}_E$? In order to answer this question, we follow the setup of [64]. Let z_α, for $0 < \alpha < 1$, be given by the equation $\mathbb{P}(-z_\alpha \leq \mathcal{N} \leq z_\alpha) = \alpha$, where $\mathcal{N}$ has a standard

normal distribution. Using the central limit theorem, it can be shown that an $\alpha\%$-confidence interval for $\widehat{p}_E$ is given by $(\widehat{p}_E \pm z_\alpha \widehat{\sigma} n^{-1/2})$, where $\widehat{\sigma}$ is the square root of the sample variance. The *relative error* of the estimator is defined to be the absolute error (half the size of the confidence interval) divided by the value of the estimator, i.e., it is equal to $z_\alpha \sigma n^{-1/2}/p_E$. We often require that the relative error is smaller than some set *precision*, e.g. 10%. Suppose now that $\widehat{p}_E \ll 1$. We then have

$$RE = z_\alpha \frac{\sqrt{p_E(1-p_E)}}{\sqrt{N}p_E} \approx \frac{z_\alpha}{\sqrt{N p_E}}. \tag{1.2}$$

Consider the case $p_E = 10^{-9}$. If we require that the 95%-confidence interval has a relative error of at most 10%, then we need $N \geq 3.84 \times 10^{11}$, i.e., we need over three hundred billion runs. This big amount of runs is, for many models, too high to be executed on most computers in a reasonable amount of time. In fact, the number of runs needed is *inversely proportional* to p_E. Thus, the need arises for sampling methods that need a lower number of runs to obtain a fixed relative error for $p_E \ll 1$.

In addition to the high number of runs required to estimate small probabilities, the central limit theorem can downright be a very bad approximation of those probabilities. We will show this with a numerical example.

Example 1.1.1. Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables having an exponential distribution with parameter μ (so that its mean is $1/\mu$). Let $\widehat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be its empirical mean. We want to estimate $\mathbb{P}\left(\widehat{S}_n > 2/\mu\right)$ for large n. According to the central limit theorem, we know that, for large n,

$$\sqrt{n}(\widehat{S}_n - \frac{1}{\mu}) \overset{\mathrm{d}}{\approx} \mathcal{N}(0, \frac{1}{\mu^2}).$$

Hence, according to the central limit theorem, for large n, we obtain

$$\begin{aligned}\mathbb{P}\left(\widehat{S}_n > \frac{2}{\mu}\right) &= \mathbb{P}\left(\widehat{S}_n - \frac{1}{\mu} > \frac{1}{\mu}\right) \\ &= \mathbb{P}\left(\sqrt{n}(\widehat{S}_n - \frac{1}{\mu}) > \frac{\sqrt{n}}{\mu}\right)\end{aligned}$$

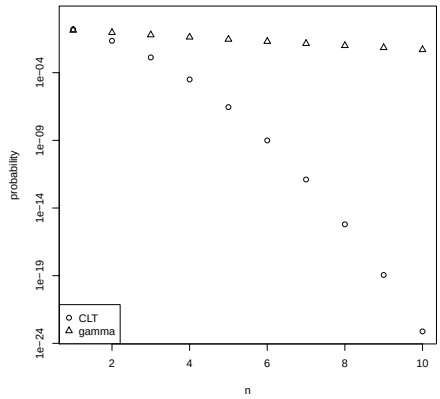

Figure 1.1: Numerical values corresponding to Example 1.1.1, using $\mu = 2$.

$$\begin{aligned} &= \mathbb{P}\left(\mu\sqrt{n}(\widehat{S}_n - \frac{1}{\mu}) > \sqrt{n}\right) \\ &\approx \mathbb{P}\left(\mathcal{N}(0,1) > \sqrt{n}\right). \end{aligned}$$

Note that actually $n\widehat{S}_n$ has a gamma distribution with shape parameter n and scale parameter $1/\mu$. In Figure 1.1 we see how drastic the difference is.

One of those methods that tries to remedy this issue is *importance sampling*. In order to explain this method, let us first generalise the setting. Let the model under consideration be represented by some random variable X. Furthermore, let the quantity of interest we want to estimate be $\psi(X)$ for some function ψ. Assume that X is real-valued and has a density f. Importance sampling is essentially sampling X from a different density $\widetilde{f}$, which has to be absolutely continuous with respect to f, i.e., we require that $\widetilde{f}(x) > 0$ whenever $\psi(x)f(x) > 0$; this is called using a *change of measure*. Observe now that

$$\gamma := \int \psi(x)f(x)\,\mathrm{d}x = \int \psi(x)\frac{f(x)}{\widetilde{f}(x)}\widetilde{f}(x)\,\mathrm{d}x. \tag{1.3}$$

Hence, we can write, using self-explanatory notation, $\gamma = \mathbb{E}_{\widetilde{f}}(\psi(X)L(X))$,

where L is the *likelihood ratio*, which is defined as $L(x) = \frac{f(x)}{\widetilde{f}(x)}$ when $\psi(x)f(x) > 0$ and $L(x) = 0$ otherwise. This gives rise to the following estimator for γ: if we sample N samples from X, using density $\widetilde{f}$, the unbiased estimator is

$$\widetilde{\gamma} := \sum_{i=1}^{N} \frac{\psi(X_i)L(X_i)}{N}. \tag{1.4}$$

In order for this procedure to perform better than Monte Carlo sampling, an appropriate choice for $\widetilde{f}$ has to be made. Suppose we choose $\widetilde{f}(x) = f(x)\psi(x)/\gamma$ and thus $L = \gamma/\psi(x)$. We then have

$$\mathbb{V}\mathrm{ar}\,(\widetilde{\gamma}) = \frac{1}{N}\mathbb{V}\mathrm{ar}_{\widetilde{f}}\,(\psi(X)L(X)) = \frac{1}{N}\mathbb{V}\mathrm{ar}_{\widetilde{f}}\,(\gamma) = 0,$$

i.e., we have a "perfect" estimator in the sense that it has zero variance. Unfortunately, this estimator can not be used in practice, as it requires knowledge of the value of γ, which is exactly the quantity we're trying to estimate in the first place. However, this estimator does provide a benchmark in the sense that a "good" change of measure probably looks closely like this perfect change of measure.

Several definitions of what constitutes a "good" change of measure exist. Generally, the number of runs required to obtain some precision is proportional to the variance of the estimator. Hence, a good change of measure is one that achieves variance reduction. Additionally there exist some optimality notions. Two commonly used notions exist in an asymptotic regime, for which we will provide the setting first. We restrict to the case where ψ is an indicator function of the outcome of some event; we will thus replace ψ with I. Let now $\gamma \equiv \gamma_n$ be indexed by a *rarity parameter* n and assume that $\gamma_n \to 0$ as $n \to \infty$. Suppose that new density $\widetilde{f}$ induces a probability measure $\mathbb{Q}$. We then say that $\mathbb{Q}$ is *asymptotically optimal* (or *asymptotically efficient*) with respect to n if

$$\lim_{n\to\infty} \frac{\ln \mathbb{E}_{\mathbb{Q},n}\left(L^2 I\right)}{\ln \mathbb{E}_{\mathbb{Q},n}\left(LI\right)} = 2. \tag{1.5}$$

Note that, by Jensen's inequality, the limit is always smaller than or equal to 2. In the light-tailed case, i.e., when γ_n decays exponentially as a function of n, asymptotic optimality means that the number of runs needed grows at most

subexponentially as well to obtain a certain precision.

As a second, stronger, notion of optimality, we say that $\mathbb{Q}$ has *bounded relative error* with respect to n if

$$\limsup_{n\to\infty} \frac{\mathbb{V}\mathrm{ar}_{\mathbb{Q},n}\,(LI)}{\gamma_n^2} < \infty. \tag{1.6}$$

It is a stronger notion of optimality in the sense that to obtain a required precision, a *bounded* number of runs is required as n goes to infinity.

We now give an example in which bounded relative error is attained.

Example 1.1.2. Let N have a geometric distribution with success parameter $p \equiv p(x)$, where x is a rarity parameter (e.g., we could let $p(x) = 1/x$ for $x \geq 1$). Let

$$z := \mathbb{P}\,(N \leq m) = 1 - (1-p)^m,$$

be such that $z \to 0$ as $x \to \infty$. We want to use importance sampling in order to efficiently estimate z. We propose that, under $\mathbb{Q}$, N has a geometric distribution with a different success parameter $\tilde{p}$. Some analysis reveals that a good candidate is to take $\tilde{p} = 1/m$. Indeed, we will show that $\mathbb{E}\left(L^2 I\right) \sim z^2(e-1)$ so that bounded relative error is achieved. Using that $1 - p \approx 1$, we get

$$\begin{aligned}
\mathbb{E}\left(L^2 I\right) &= \sum_{k=0}^{m-1} (1-\tilde{p})^k \tilde{p} \left(\frac{(1-p)^k p}{(1-\tilde{p})^k \tilde{p}}\right)^2 = \frac{p^2}{\tilde{p}} \sum_{k=0}^{m-1} (1-\tilde{p})^{-k} (1-p)^{2k} \\
&= \frac{p^2}{\tilde{p}} \frac{1-(1-\tilde{p})^{-m}(1-p)^{2m}}{1-(1-\tilde{p})^{-1}(1-p)^2} = mp^2 \frac{1-(1-\tfrac{1}{m})^{-m}(1-p)^{2m}}{1-(1-\tfrac{1}{m})^{-1}(1-p)^2} \\
&\approx mp^2 \frac{1-(1-\tfrac{1}{m})^{-m}}{1-(1-\tfrac{1}{m})^{-1}} \approx mp^2 \frac{1-e}{1-(1-\tfrac{1}{m})^{-1}} \\
&= mp^2 \frac{e-1}{\frac{1}{m-1}} \sim mp^2 \frac{e-1}{\frac{1}{m}} = m^2 p^2 (e-1) \\
&\sim z^2(e-1).
\end{aligned}$$

This example should convince the reader that importance sampling can indeed perform much better than Monte Carlo sampling. In this example, a bounded number of runs in x is needed to achieve a certain precision, whereas in Monte Carlo sampling, the number of runs is inversely proportional to $z(x)$.

In Section 1.3, we give an example where importance sampling can also perform badly if one is not careful enough in choosing a change of measure.

More methods exist that provide a method to efficiently implement rare event simulation. One of those methods is *splitting*, which is best explained by a simple example. Consider a random walk S_k, $k \geq 0$, with i.i.d. increments and a negative drift. Of interest now is the event of the random walk ever reaching some high level n. As this is a rare event, consider the following idea. Define levels $0 < n_1 < \ldots < n_j < n$. Now, we simulate M_0 independent copies of S_0. We simulate each copy until it either reaches level n_1 or returns to level 0. The idea behind splitting is that, conditioning on reaching level n_1, reaching level n is less rare. Hence we continue by discarding the copies that returned to level 0 and creating M_1 copies of each of the original copies that reached level n_1. We repeat this procedure, until we end up with some copies entering level n. As an estimator, we use the number of copies that reached level n divided by the total number of copies that were created during the process.

In order to optimise this procedure, one has to consider, for example, how to choose the number of levels j, how to choose the levels $n_1, \ldots, n_j$ and how to choose the numbers of copies to create at each level. These are just a few aspects of the splitting method.

Another method for simulating rare events, called *conditional Monte Carlo sampling*, is used in Chapter 5. As the name suggests, it involves conditioning on a random variable or on a sigma-field. In the context of estimating a rare event, recall that we want to estimate the probability of an event E, denoted by p_E. Suppose that this event depends on some random variable X. For conditional Monte Carlo estimation, we replace the estimator $\widehat{p}_E$ (see (1.1)) with the unbiased estimator $\mathbb{E}\left(\widehat{p}_E|X\right)$. Using this estimator always gives a variance reduction. Indeed, using the law of total variance, we get

$$\mathbb{V}\mathrm{ar}\left(\widehat{p}_E\right) = \mathbb{V}\mathrm{ar}\left(\mathbb{E}\left(\widehat{p}_E|X\right)\right) + \mathbb{E}\left(\mathbb{V}\mathrm{ar}\left(\widehat{p}_E|X\right)\right) \geq \mathbb{V}\mathrm{ar}\left(\mathbb{E}\left(\widehat{p}_E|X\right)\right).$$

However, for this method, the hard part is in finding a suitable X so that the conditional expectation is computable. The following example is closely related to Chapter 5.

Example 1.1.3. Let E and X be independent random variables with E having an exponential distribution with unit mean and X having some distribution

from which we can easily sample. We want to calculate $\mathbb{P}(E + X > x)$ for some large x. However, depending on the distribution of X, this might not be possible to do exactly. Furthermore, as this probability is small for large x, performing Monte Carlo sampling also is not a good option, as explained before in this section. Therefore, we resort to conditional Monte Carlo sampling. Note that

$$\mathbb{P}(E + X > x) = \mathbb{E}(\mathbb{1}(E + X > x)) = \mathbb{E}(\mathbb{E}(\mathbb{1}(E + X > x)\,|X)),$$

so an unbiased estimator for $\mathbb{P}(E + X > x)$ is

$$\mathbb{E}(\mathbb{1}(E > x - X)\,|X) = e^{-(x-X)} \wedge 1.$$

Hence, the probability can easily be estimated by sampling from X and substituting its values in the formula above.

We refer the reader to [64] and [3] for a more thorough treatment on importance sampling, splitting, conditional Monte Carlo sampling and other rare event simulation methods.

1.2 Queueing processes and random walks

In this thesis, we look at rare event simulation in the context of queueing processes and random walks in a multidimensional setting. The theoretical background that is required to read the rest of this thesis is treated in this section. This section doesn't aim to be a full introduction to queueing processes and random walks; we provide references to more comprehensive literature below.

Queueing theory is one of the largest and most applicable fields within probability theory. Many books and papers are devoted to this topic. Although we could easily write a whole chapter on introducing the field, we will restrict ourselves to the bare minimum that is required for understanding the rest of this thesis. We refer to, for example, [26] and [24], for a more thorough introduction to queueing theory, and to [44] for a more recent publication.

We will first define a discrete-time queue and show its connection to a random walk. Let $Q_0 = 0$ and let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables. A

queueing process is then recursively defined by

$$Q_n := \max\{Q_{n-1} + X_n, 0\}, \qquad n \geq 1. \tag{1.7}$$

Hence, X_n describes the amount by which the queue in- or decreases between time $n-1$ and time n, where we restrict the content of the queue to be non-negative. If we iterate this equation we get

$$\begin{aligned} Q_n &= \max\{0, X_1 + \cdots + X_n, X_2 + \cdots + X_n, \ldots, X_n\} \\ &\stackrel{\mathrm{d}}{=} \max\{0, X_1, X_1 + X_2, \ldots, X_1 + \cdots + X_n\}. \end{aligned}$$

For the stationary distribution we naturally obtain

$$Q_\infty \stackrel{\mathrm{d}}{=} \sup_{n=0,1,2,\ldots} \sum_{i=1}^{n} X_i,$$

where if $n = 0$ the sum is to be taken equal to zero. We can conclude that the stationary distribution of the workload in a queue is equal to the all-time supremum of a random walk.

This can in some cases be extended to higher dimensions. We will, however, show that in higher dimensions, two different types of events can be considered. For ease of notation, we will only show the two-dimensional case. Let $(X_i)_{i\in\mathbb{N}}$ and $(\tilde{X}_i)_{i\in\mathbb{N}}$ be two sequences of i.i.d. random variables and let the two sequences also be independent of each other (but not necessarily equally distributed). The first type of event is a *ruin probability*, which is

$$\mathbb{P}\left(\exists n : \sum_{i=1}^{n} X_i > q, \sum_{i=1}^{n} \tilde{X}_i > \tilde{q}\right). \tag{1.8}$$

This event is analysed in Chapter 2.

In order to provide the second type of event, we will first define a two-dimensional queueing system:

$$\begin{pmatrix} Q_n \\ \tilde{Q}_n \end{pmatrix} = \begin{pmatrix} \max\{Q_{n-1} + X_n, 0\} \\ \max\{\tilde{Q}_{n-1} + \tilde{X}_n, 0\} \end{pmatrix} = \ldots$$

$$= \begin{pmatrix} \max\{0, X_1 + \cdots + X_n, \ldots, X_n\} \\ \max\{0, \tilde{X}_1 + \cdots + \tilde{X}_n, \ldots, \tilde{X}_n\} \end{pmatrix}$$
$$\stackrel{\mathrm{d}}{=} \begin{pmatrix} \max\{0, X_1, \ldots, X_1 + \cdots + X_n\} \\ \max\{0, \tilde{X}_1, \ldots, \tilde{X}_1 + \cdots + \tilde{X}_n\} \end{pmatrix}.$$

Its stationary distribution is given by

$$\begin{pmatrix} Q_\infty \\ \tilde{Q}_\infty \end{pmatrix} \stackrel{\mathrm{d}}{=} \begin{pmatrix} \sup_{n=0,1,2,\ldots}\{\sum_{i=1}^n X_n\} \\ \sup_{n=0,1,2,\ldots}\{\sum_{i=1}^n \tilde{X}_n\} \end{pmatrix}.$$

This last equality gives rise to the following probability:

$$\mathbb{P}\left(Q_\infty > q, \tilde{Q}_\infty > \tilde{q}\right) = \mathbb{P}\left(\exists n_1 : \sum_{i=1}^{n_1} X_i > q; \exists n_2 : \sum_{i=1}^{n_2} \tilde{X}_i > \tilde{q}\right). \tag{1.9}$$

The analysis of such an event is the main topic of Chapter 3.

We want to emphasise that in the event of Equation (1.8), the two processes have to reach the high levels *at the same time*, while in (1.9), they can reach the high levels at *different times*.

1.3 Large deviations theory

Large deviations theory is an extensive collection of mathematical theories and tools that are developed to analyse rare events. In this thesis, large deviations theory is used for two main reasons. One is that it provides tools and heuristics that allow us to find a good change of measure in designing importance sampling algorithms. The second reason is that it allows us to evaluate the efficiency of a change of measure.

Consider an i.i.d. sequence of random variables. Extensive theory exists on the "regular" behaviour of such a sequence: the law of large numbers provides us with information on its sample mean and the central limit theorem describes the behaviour around the sample mean. But how can events with a very low probability be described? Large deviations theory aims to provide a mathematical framework to answer this question.

Let us illustrate that with an example.

Example 1.3.1. Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables having a normal distribution with mean zero and unit variance. Let $\widehat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be its empirical mean. Note that $\widehat{S}_n$ again has a normal distribution with zero mean and variance $1/n$. Therefore, for any $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|\widehat{S}_n| \geq \delta\right) = 0, \tag{1.10}$$

and for any interval A

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\widehat{S}_n \in A\right) = \frac{1}{2\pi}\int_A e^{-x^2/2}\,\mathrm{d}x. \tag{1.11}$$

Furthermore,

$$\mathbb{P}\left(|\widehat{S}_n| \geq \delta\right) = 1 - \frac{1}{2\pi}\int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} e^{-x^2/2}\,\mathrm{d}x,$$

so that

$$\lim_{n\to\infty} \frac{1}{n}\ln \mathbb{P}\left(\sqrt{n}\widehat{S}_n \geq \delta\right) = -\frac{\delta^2}{2}. \tag{1.12}$$

This last equation is a typical large deviations statement. It says that $|\widehat{S}_n|$ can take relatively large values with a probability of the order $e^{-n\delta^2/2}$.

Note that, from the law of large numbers and the central limit theorem, equations (1.10) and (1.11) still hold for any i.i.d. sequence with finite second moment. One could ask whether equation (1.12) also still holds in the general case. It can be derived that the limit indeed always exists, but that its value depends on the distribution of X_1. This is a simple example of a so-called *large deviations principle*, which is one of the main tools used in large deviations theory. We will give the exact definition below, but first we need to introduce some preliminary definitions.

Most of the definitions and results below are taken from [30]. Let $\{\mu_\epsilon\}$ be a family of probability measures on some topological space $\mathcal{X}$. A *rate function* I then is a non-negative lower semi-continuous function on $\mathcal{X}$, i.e., all level sets are closed sets. We call I a *good* rate function if, in addition, all level sets are compact.

We say that $\{\mu_\epsilon\}$ satisfies a large deviations principle with rate function I

if for all $\Gamma \in \mathcal{B}$,

$$-\inf_{x\in\Gamma^\circ} I(x) \leq \liminf_{\epsilon\to 0} \epsilon \ln \mu_\epsilon(\Gamma) \leq \limsup_{\epsilon\to 0} \epsilon \ln \mu_\epsilon(\Gamma) \leq -\inf_{x\in\bar{\Gamma}} I(x), \tag{1.13}$$

where in the infima, the interior and closure of Γ are meant respectively.

Informally, if $\{\mu_\epsilon\}$ satisfies a large deviations principle with rate function I, it means that

$$\mu_\epsilon(\Gamma) \approx e^{\frac{-1}{\epsilon}\inf_{x\in\Gamma} I(x)},$$

i.e., the probability of interest decays exponentially in $1/\epsilon$ with decay rate $\inf_{x\in\Gamma} I(x)$.

There are several conditions under which a family of probability measures satisfies a large deviations principle. One of the most famous results is Cramér's theorem (originally published in [27]), which is a statement about the sequence of probability measures corresponding to the sample means of an i.i.d. sequence of random variables. Before we can state the theorem, some more definitions are required.

We again consider a sequence of i.i.d. random variables $(X_i)_{i\in\mathbb{N}}$ and its empirical mean $\widehat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Let μ_n denote the probability law of $\widehat{S}_n$. The *logarithmic moment generating function*, or *cumulant generating function*, of X_1 is defined as

$$\Lambda(\lambda) := \ln M(\lambda), \qquad M(\lambda) := \mathbb{E}\left(e^{\lambda X_1}\right),$$

and the *Fenchel-Legendre transform* of $\Lambda(\lambda)$ is defined as

$$\Lambda^\star(x) := \sup_{\lambda\in\mathbb{R}}\{\lambda x - \Lambda(\lambda)\}.$$

Theorem 1.3.2 (Cramér)**.** *The sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ satisfy a large deviations principle with convex rate function $\Lambda^\star(\cdot)$.*

Note that no assumptions on the existence of any moment of X_1 were made; in particular, the theorem holds even when the first moment of X_1 doesn't exist (though in that case, $\Lambda^\star(x)$ would be 0 for all x, and the theorem is not very informative).

Several generalisations of Cramér's theorem exist. It should be no surprise that a multivariate counterpart exists, i.e., in the case where $(X_i)_{i\in\mathbb{N}}$ is a sequence of i.i.d. random vectors in $\mathbb{R}^d$. A generalisation to the non-i.i.d. case can

also be made, which is called the Gärtner-Ellis theorem. We use this theorem in Chapter 2, where it is also stated.

So far, all results have been about the empirical mean of a sequence of random variables. However, sometimes one wants to obtain information about the sample path of such a sequence. For example, the maximum of the random walk can be of interest, or one might want to know what the probability is of the random walk ever attaining a value in some (rare) set. A result indeed exists, in the form of Mogulskii's theorem (see [61] for the original publication). Let's again provide some preliminary definitions first.

Since the theorem is stated in higher dimension, we first state the higher dimensional counterparts of the definitions above. Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$, with $\Lambda(\lambda) = \ln \mathbb{E}\left(e^{\langle \lambda, X_1\rangle}\right)$. Let $|\cdot|$ be the Euclidian norm on $\mathbb{R}^d$ (i.e., $|x| = \sqrt{\langle x, x\rangle}$), let $\|\cdot\|$ be the supremum norm on $L_\infty([0,1])$ and let $\Lambda^\star(x) := \sup_{\lambda\in\mathbb{R}^d}\{\langle \lambda, x\rangle - \Lambda(\lambda)\}$ be the d-dimensional Fenchel-Legendre transform of $\Lambda(\cdot)$.

$$Z_n(t) := \frac{1}{n}\sum_{i=1}^{\lfloor nt\rfloor} X_i, \qquad 0 \le t \le 1,$$

and let μ_n be the probability law of $Z_n(\cdot)$.

Theorem 1.3.3 (Mogulskii)**.** *Let $\Lambda(\lambda) < \infty$ for all $\lambda \in \mathbb{R}^d$. The sequence of probability measures $(\mu_n)_{n\in\mathbb{N}}$ satisfy, in $L_\infty([0,1])$, a large deviations principle with good rate function*

$$I(\phi) = \begin{cases} \int_0^1 \Lambda^\star(\phi'(t))\,\mathrm{d}t, & \text{if } \phi \in \mathcal{AC}, \phi(0) = 0, \\ \infty & \text{otherwise,} \end{cases}$$

where $\mathcal{AC}$ is the set of absolutely continuous functions.

We conclude this part by stating an elementary property called the *principle of the largest term*. We provide the version as stated in [39]; a slightly more general version can be found in [30].

Lemma 1.3.4 (Principle of the largest term)**.** *Let a_n and b_n be sequences in $\mathbb{R}_+$. If $1/n \ln a_n \to a$ and $1/n \ln b_n \to b$ as $n \to \infty$, then*

$$\lim_{n\to\infty} \frac{1}{n}\ln(a_n + b_n) = a \vee b.$$

If a_n and b_n are probabilities, this lemma has the interpretation that rare events tend to happen in the most likely way.

We conclude this section by providing a change of measure that is often used in importance sampling, namely *exponential twisting*. If a random variable X has a density $f(\cdot)$, then we say that we exponentially twist X with parameter λ if we sample using density

$$f_\lambda(x) := f(x)\frac{e^{\lambda x}}{M(\lambda)}.$$

This change of measure is useful in the following setting. Consider the random walk $S_n = \sum_{i=1}^n X_i$, where the X_i are i.i.d. with expected value μ. Suppose that we are interested in the probability $\mathbb{P}\left(S_n/n > a\right)$ for some $a > \mu$. Let $\lambda(a)$ be the optimising argument of $\Lambda^\star(a)$. As shown in [12, Section 5.2.1], when the random walk is twisted with parameter $\lambda(a)$, asymptotic optimality is achieved.

1.4 Complications of rare event estimation in a multidimensional setting

One-dimensional systems have some very useful properties: they are (relatively) easy to analyse, it is easy to design simulation algorithms for them and a lot is known about them already. For example, for single server queues, many results exist on e.g. waiting times, sojourn times and queue length distribution. However, many practical situations are too complicated to model by, e.g., a single queue, a single buffer or a one-dimensional random variable. Therefore, multidimensional models need to be studied.

Although multidimensional models have the potential to be a much closer approximation of reality than their one-dimensional counterparts, they come with a cost: they are relatively hard to analyse and as a result, less results exist for these models. In several special cases, some results do exist. Recent publications include, e.g., [53], where the stationary workload distribution of a fluid tandem queue is analysed in heavy traffic, and [66], where the steady-state marginal workload distribution for a single server polling model with two queues is analysed. Additional results can be found in e.g. [62], [16] and [55].

Because of this, and because of their inherent mathematical value, this thesis

focuses on multidimensional models. In order to illustrate the hardness in analysing multidimensional models in the setting of rare-event simulation, we give some examples of potential problems that they have.

One example, which is shown in Chapter 2 as well, is briefly introduced here. Consider a two-dimensional fluid buffer model in which the net inflow (or outflow) is determined by the state of some finite state space background process. This means that in between two jumps of the background process, the net in- or outflow of both buffers is constant. Suppose that one is interested in the rare event of both buffers attaining a high level simultaneously. If one was to perform simulations in order to estimate the probability of this event, one would simulate the background process and calculate at each jump point what the levels of the buffers must be. An easy trap to fall into is thinking that the event of interest only occurs if, at a jump point, both buffers reached the high level. As this is indeed true in the one-dimensional case, one might think that this extends to the two-dimensional case. However, as Figure 2.2 in Chapter 2 shows, the rare set of interest can also be reached *in between* two jump points of the background process. Naturally, this can easily be overcome by linearly interpolating between two jump points, but this simple example shows that if one is not careful, mistakes can easily be made when transitioning from one dimension to multiple.

In the next example, although the stochastic process is one-dimensional, the rare set of interest is a disjoint union of two sets, in particular, the set is non-convex. This is an example in which one wants to implement an efficient importance sampling algorithm. Consider again a random walk $(X_i)_{i\in\mathbb{N}}$ and its empirical mean $\widehat{S}_n = \frac{1}{n}\sum_{i=1}^n X_i$. We also use the notation $S_n = \sum_{i=1}^n X_i$ and define $\lambda^\star(a)$ to be the optimiser of $\Lambda^\star(a)$. We are interested in the probability

$$p_n(a,b) := \mathbb{P}\left(\frac{\widehat{S}_n}{n} \geq a \text{ or } \frac{\widehat{S}_n}{n} \leq b\right)$$

for $b < \mathbb{E}(X_1) < a$. Suppose that the point a is the most likely point for the rare event to occur, meaning that $p_n(a)/p_n(a,b) \to 1$ as $n \to \infty$, where $p_n(a) := \mathbb{P}\left(\frac{\widehat{S}_n}{n} \geq a\right)$ (this corresponds to the case that $\Lambda^\star(a) \leq \Lambda^\star(b)$). It seems tempting to perform an exponential twist with parameter $\lambda(a)$. Let $\mathbb{Q}$ denote the corresponding change of measure of this exponential twist. As shown

in [12, Section 5.2.3], we have for the second moment of the estimator that

$$\mathbb{E}_{\mathbb{Q},n}\left(L^2 I\right) = \left[\mathbb{E}_{\mathbb{Q},n}\left(e^{-2\lambda^\star(a)S_n}\mathbb{1}_{\{S_n\geq na\}} + \mathbb{E}_{\mathbb{Q},n}\left(e^{-2\lambda^\star(a)S_n}\mathbb{1}_{\{S_n\leq nb\}}\right)\right)\right] \times \left(M(\lambda^\star(a))\right)^{2n}.$$

The decay rates of the first and second term respectively are $-2\Lambda^\star(a)$ and $-\lambda^\star(a))b-\Lambda^\star(b)+\ln M(\lambda^\star(a))$. In order to get a subexponential number of runs, we need that $\mathbb{V}\mathrm{ar}_{\mathbb{Q},n}\left(LI\right)$ needs to decay (decrease on an exponential scale) as fast as $p_n^2(a,b)$, which is the case when the inequality $\Lambda^\star(a) \leq \lambda^\star(a))b+\Lambda^\star(b)-\ln M(\lambda^\star(a))$ must be satisfied. It can be shown that it is satisfied only when $b \leq b^\star$ for some fixed $b^\star$. Hence, when b is too large, at least an exponential number of runs is needed. The heuristic explanation is that, even in importance sampling, it might occur that $S_n/n \leq b$. When that happens, the likelihood becomes very large, contributing to a large variance. We refer to [12] for various methods that circumvent this problem, one of which is to write the probability as the sum of two probabilities and estimating each term separately.

Throughout this thesis, we will make use of several techniques than can be applied in multiple dimensions, e.g., partitioned importance sampling and conditional Monte Carlo sampling. These methods have the advantage that they are easier to implement than some of the methods that can be found in the survey paper [10].

1.5 Contributions

This section highlights the contributions that are made in this thesis.

In Chapter 2, we consider a bivariate stochastic process. The (rare) event of interest is the event of both components exceeding some large level *simultaneously*. Both asymptotic techniques as well as efficient simulation techniques are provided. The asymptotic result concerns various expressions for the decay rate of the probability of interest and is valid under Gärtner-Ellis-type conditions. The simulation result concerns a specific instance of the model under consideration, which is that of two Markov fluid queues driven by the same background process. An asymptotically efficient importance sampling procedure is provided

for this example. Furthermore, several numerical experiments are provided that support the theory.

Chapter 3 is closely related to Chapter 2. Again, a bivariate stochastic process is analysed, but the event of interest now is that of both components exceeding some large level, but *not* necessarily at the same time. This chapter also provides two results. As in Chapter 2, the first result is an expression for the decay rate of the probability of interest. The second result focuses on efficient simulation techniques. Using a "nearest-neighbour random walk" as an example, we first show that a "naive" implementation of importance sampling, based on the decay rate, is not asymptotically efficient. To remedy this, we introduce a technique which we call *partitioned* importance sampling and prove that this technique, indeed, is asymptotically efficient. We conclude the chapter by providing several simulation results.

In Chapter 4, we look at a linear stochastic fluid network under Markov modulation. The focus is on the probability that the joint storage level will ever attain a value in a rare set. The majority of the chapter is devoted to developing efficient importance sampling algorithms. For linear stochastic fluid networks *without* modulation, we prove that the algorithm presented needs at most a polynomial number of runs, whereas the probability of interest decays exponentially. For linear stochastic fluid networks *with* modulation, we show that the algorithm is asymptotically efficient. Furthermore, we point out how to set up a recursion to evaluate the (transient and stationary) moments of the joint storage level in Markov-modulated linear stochastic fluid networks.

Chapter 5 focuses on the stationary distribution of a stochastic recursion. The goal is to estimate the probability that the stationary process has a large value. We provide a conditional Monte Carlo algorithm that can estimate this probability both efficiently (bounded relative error is attained) and unbiasedly in finite running time. We also provide an expression for the asymptotic tail behaviour.

Together, these chapters show how it is possible to develop efficient simulation techniques for multidimensional models with a variety of approaches, that complement the approaches in the literature which are surveyed in [10].

2

A multi-dimensional ruin problem

This chapter is devoted to analysing the event of both components of a bivariate stochastic process exceeding a large threshold at the same time. This is closely related to Chapter 3. There, the event of both components of a bivariate stochastic process exceeding a large threshold at possibly different times is analysed. This chapter is based on [21].

2.1 Introduction

Let $((A_t, B_t))_{t\geq 0}$ be a bivariate stochastic process, with possibly dependent components. This chapter focuses on techniques to quantify the so-called (bivariate) *ruin probability* over level u, denoted by α_u, being defined as the probability that this process will ever hit the set $S_u := (u, \infty) \times (u, \infty)$, for some $u > 0$. Note that even in the case that the two processes are independent, this probability cannot be evaluated from the corresponding one-dimensional ruin probabilities, since both components have to be bigger than u at the *same time*.

A leading example of such a bivariate model is the *two-dimensional Markov modulated fluid model*, which can be described as follows. Let $(X_t)_{t\geq 0}$ be an irreducible Markov process, taking values on a finite state space. Whenever $X_t = i$, both A_t and B_t change at constant, possibly negative, rates r_i^A and r_i^B

respectively so that

$$(A_t, B_t) = \left(\int_0^t r^A_{X_s} \, \mathrm{d}s, \int_0^t r^B_{X_s} \, \mathrm{d}s \right);$$

the two processes $(A_t)_{t\geq 0}$ and $(B_t)_{t\geq 0}$ thus depend on each other as they react to the same realisation of the *background process* $(X_t)_{t\geq 0}$. The process (A_t, B_t) could be used to model the evolution of two random quantities which are driven by the same environment. Many examples can be thought of: the process can represent data buffers in a wireless network whose dynamics react to the same variations in the channel conditions, or multiple asset prices reacting to the same market fluctuations. The one-dimensional counterpart of this model is well understood. In particular, techniques have been developed to evaluate the ruin probabilities, by setting up a system of linear differential equations which can be solved by imposing the appropriate boundary conditions; see e.g. [35]. Importantly, in the two-dimensional case these methods fail.

Since an exact analysis of α_u has been beyond reach so far, in this chapter we turn to two approximation techniques that are intended to gain insight into the quantitative properties of α_u. The first technique is of an *asymptotic* nature: it characterises the (essentially exponential) tail behaviour of α_u for large u. The second approach is an efficient simulation technique based on importance sampling; it remedies the complication that straightforward, naïve simulation methods are typically slow due to the rarity of the event under consideration.

The research reported on in this chapter is in the tradition of a series of papers on large deviations estimates and importance sampling for queues. For an introduction to importance sampling, we refer to e.g. [64]. Importance sampling is a variance-reduction technique which essentially amounts to sampling under another measure than the actual one, recovering unbiasedness by weighing the simulation data by appropriate likelihood ratios; the complication lies in the selection of the new measure, which should ideally be chosen such that the variance of the resulting estimator is minimised. Part of this chapter relates to relatively general bivariate processes, and part to the specific case of bivariate Markov fluid. Asymptotics and efficient simulation for the one-dimensional model have been studied in detail; see e.g. [60] and [64, Section 5.3.3]. We also mention [52], where the focus lies on the existence of so-called *effective*

bandwidths; as it turns out, despite the fact that this work focuses on one-dimensional Markov fluids, results from this paper are useful in the context of our two-dimensional setup.

There is a vast literature that directly relates to the material presented in this chapter; without aiming to give a complete overview, we mention a number of relevant contributions. In [41], for a broad class of queues the exponential decay rate of the waiting time distribution is given; this result can be translated into the context of ruin probabilities. In [32], it is generalised to continuous time, as well as to non-linear scaling. [25] considers a related result for multi-dimensional discrete time Markov additive processes. In [48], a fluid model is considered as well, but only for a Lévy input process (i.e., without Markov modulation); the main result is an expression for the Laplace-Stieltjes transform of the joint steady-state distribution.

This chapter has two main results. The first one is Theorem 2.2.2, which gives multiple equivalent expressions for the *decay rate*

$$\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\exists t > 0 : A_t > u, B_t > u\right)$$

under mild assumptions. Importantly, these assumptions are satisfied by the bivariate Markov fluid model. In the first part of the proof, we interpret a representation of the decay rate as the solution of a concave optimisation problem with respect to several constraints. The use of this interpretation we believe is novel; for example, using this argumentation would have led to a considerably shorter proof in the one-dimensional case in [39, Lemma 1.7]. A different expression for the decay rate is provided as well, with a proof that is split into two parts. First the lower bound is proven, which focuses on the largest contribution to α_u (in terms of a 'dominant time scale'). For the upper bound, we determine the decay rate of the probability that the bivariate process hits some set T_u which contains S_u; we let T_u be as small as possible so as to still obtain the correct decay rate. This approach is similar to the one used by [54].

The second main result concerns efficiency properties of an importance-sampling-based simulation scheme, which applies to the bivariate Markov fluid model only; it states that the underlying new measure is optimal (in a specific asymptotic sense). The new measure we propose can be regarded as the

two-dimensional analogue of the change of measure that was used for the one-dimensional model in e.g. [60] and [64, Section 5.3.3]. A complication is that the process (A_t, B_t) can attain values in the target set while the 'embedded process' (recording values of (A_t, B_t) only at transition epochs of the background process) does not; we describe a technique to remedy this.

The rest of this chapter is organised as follows. Section 2.2 contains the first main result, namely the decay rate for the general two-dimensional stochastic process under a Gärtner-Ellis-type condition. In Section 2.3, we specifically consider the bivariate Markov fluid model; we first present a number of results for this model, then we develop an efficient simulation algorithm, and finally we present a number of illustrative numerical examples. For readability, the proof of Theorem 2.3.1 is given in Appendix 2.5. The chapter concludes with Section 2.4, in which we discuss two natural extensions of the theory developed in Section 2.2, namely the extension of the theory to higher dimensions and the extension to bivariate processes for which the components may hit level u at *different* times.

2.2 Logarithmic asymptotics under Gärtner-Ellis conditions

Let $((A_t, B_t))_{t\geq 0}$ be a bivariate stochastic process on $\mathbb{R}^2$. We are interested in the probability α_u that the process will ever hit the set $S_u := (u, \infty) \times (u, \infty)$, for $u \gg 0$, when the average movement of the process is directed away from this set. More specifically, we wish to characterise the decay rate of this probability, i.e.,

$$\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\exists t : A_t > u, B_t > u\right). \tag{2.1}$$

We consider the situation that

$$\left(\lim_{t\to\infty} \frac{\mathbb{E}A_t}{t}, \lim_{t\to\infty} \frac{\mathbb{E}B_t}{t}\right) \notin [0, \infty) \times [0, \infty),$$

so that the event of interest is indeed rare. We will also assume that the process can reach the set with a positive probability. Note that we don't have to restrict the event of interest to both components having to reach the *same* level, i.e.,

the results developed in this chapter can also handle the event

$$\{\exists t > 0 : A_t > u, B_t > cu\}$$

for any $c > 0$. This can be done by applying the analysis to the stochastic process $((A_t, B_t/c))_{t\geq 0}$. Hence, we will choose $c = 1$ in the remainder of this chapter.

In order to be able to analyse the above decay rate, we now provide some results from large deviations theory, following the setup of [39]. We denote the *limiting cumulant generating function* of $((A_t, B_t))_{t\geq 0}$ by

$$M(\theta_1, \theta_2) := \lim_{t\to\infty} \frac{1}{t} \ln \mathbb{E}\left(e^{\theta_1 A_t + \theta_2 B_t}\right). \tag{2.2}$$

A function $I : \mathbb{R}^d \to \mathbb{R}^*$ (where $\mathbb{R}^* := \mathbb{R} \cup \{\infty\}$) is a *rate function* if it is non-negative and if it is lower semi-continuous, i.e., all level sets are closed. Furthermore, it is called a *good* rate function if in addition all level sets are compact. We say that $((A_t, B_t))_{t\geq 0}$ satisfies a *large deviations principle* in $\mathbb{R}^2$ with rate function $I : \mathbb{R}^2 \to \mathbb{R}^*$ if for any measurable set $F \subseteq \mathbb{R}^2$

$$-\inf_{x\in F^\circ} I(x) \leq \liminf_{t\to\infty} \frac{1}{t} \ln \mathbb{P}\left((A_t, B_t) \in F\right) \leq \limsup_{t\to\infty} \frac{1}{t} \ln \mathbb{P}\left((A_t, B_t) \in F\right) \leq -\inf_{x\in \bar{F}} I(x),$$

where F° and $\bar{F}$ denote the interior and closure of F respectively. For any function $f : \mathbb{R}^d \to \mathbb{R}^*$, we denote its *convex conjugate* by $f^*(x) := \sup_\theta \langle \theta, x\rangle - f(\theta)$. A function $f : \mathbb{R}^d \to \mathbb{R}^*$ is called *essentially smooth* if the interior of its effective domain (the set on which f is finite-valued) is non-empty, f is differentiable in the interior of its effective domain and f is steep, namely, for any sequence x_n which converges to a boundary point of the effective domain, $\lim_{n\to\infty} |\nabla f(x_n)| = \infty$. The following well-known theorem will be used in the proof of our main result.

Theorem 2.2.1 (Gärtner-Ellis, see [39, Theorem 2.11])**.** *If* (2.2) *exists for all* θ_1, θ_2*, possibly taking value infinity, and if it is essentially smooth, lower semi-continuous and finite in a neighbourhood of the origin, then the process*

$((A_t/t, B_t/t))_{t\geq 0}$ satisfies a large deviations principle in $\mathbb{R}^2$ with good convex rate function M^*.

Our main result gives the aforementioned decay rate in terms of the corresponding limiting cumulant generating function. Aside from assuming Gärtner-Ellis conditions, we also want to ensure that the continuous-time process is locally well-behaved so that we can apply the one-dimensional result from [32]. In order to make this precise, we define for $n \in \mathbb{N}$:

$$(A_n + B_n)^* := \sup_{0\leq r<1} (A_{n+r} + B_{n+r}) .$$

Theorem 2.2.2. *Let $M(\cdot,\cdot)$ satisfy the conditions of the Gärtner-Ellis theorem and let either*

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{E}\left(e^{(\theta_1 A_n+\theta_2 B_n)^* - \theta_1 A_n - \theta_2 B_n}\right) = 0 \tag{2.3}$$

for all $\theta_1, \theta_2 > 0$, or let (2.3) hold for some $\theta_1, \theta_2 > 0$ and let

$$\limsup_{n\to\infty} \frac{1}{n} \ln \mathbb{P}\left((wA_n + (1-w)B_n)^* - wA_n - (1-w)B_n > xn\right) \leq -I(x,x) \tag{2.4}$$

hold for all $x > 0$ and all $w \in [0,1]$, with $I(\cdot,\cdot)$ as in (2.6). Then

$$\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\exists t > 0 : A_t > u, B_t > u\right) = -\inf_{x>0,y>0} \frac{I(x,y)}{\min(x,y)} \tag{2.5}$$

holds, where

$$I(x,y) := \sup_{\theta_1,\theta_2} \left(\theta_1 x + \theta_2 y - M(\theta_1,\theta_2)\right) . \tag{2.6}$$

Furthermore,

$$\inf_{x>0,y>0} \frac{I(x,y)}{\min(x,y)} = \sup_{\theta_1\geq 0,\theta_2\geq 0: M(\theta_1,\theta_2)=0} (\theta_1 + \theta_2) . \tag{2.7}$$

Proof. We begin with the proof of the latter statement, i.e., (2.7). To this end, write

$$\inf_{x,y>0} \frac{I(x,y)}{\min(x,y)} = \inf_{x,y>0} \sup_{\theta_1,\theta_2} \frac{\theta_1 x + \theta_2 y}{\min(x,y)} - \frac{M(\theta_1,\theta_2)}{\min(x,y)}.$$

Setting $p := \min(x, y)$ we obtain

$$\inf_{x,y>0} \sup_{\theta_1,\theta_2} \frac{\theta_1 x + \theta_2 y}{\min(x,y)} - \frac{M(\theta_1,\theta_2)}{\min(x,y)} = \inf_{x,y,p>0, p=\min(x,y)} \sup_{\theta_1,\theta_2} \frac{\theta_1 x}{p} + \frac{\theta_2 y}{p} - \frac{M(\theta_1,\theta_2)}{p}.$$

We may now replace in the infimum $p = \min(x, y)$ by $p \leq \min(x, y)$, so that we find

$$\inf_{x,y,p>0, p=\min(x,y)} \sup_{\theta_1,\theta_2} \frac{\theta_1 x}{p} + \frac{\theta_2 y}{p} - \frac{M(\theta_1,\theta_2)}{p} = \inf_{x,y,p>0, p\leq\min(x,y)} \sup_{\theta_1,\theta_2} \frac{\theta_1 x}{p} + \frac{\theta_2 y}{p} - \frac{M(\theta_1,\theta_2)}{p};$$

this equality holds because

- The '$\geq$' part holds because the infimum on the right-hand side is taken over a larger set.
- The '$\leq$' part holds because the supremum is non-negative (choose $\theta_1 = \theta_2 = 0$), hence p will be taken as large as possible.

The next step is to replace the quantities x/p, y/p and $1/p$ by u, v and q respectively, which is allowed as long as we impose the restrictions $u, v \geq 1$ and $q > 0$ in the infimum. We thus obtain

$$\inf_{x,y,p>0, p\leq\min(x,y)} \sup_{\theta_1,\theta_2} \frac{\theta_1 x}{p} + \frac{\theta_2 y}{p} - \frac{M(\theta_1,\theta_2)}{p} = \inf_{u,v\geq 1, q>0} \sup_{\theta_1,\theta_2} \theta_1 u + \theta_2 v - qM(\theta_1,\theta_2).$$

We can now write $u\theta_1 = (1 + a)\theta_1$ with $a \geq 0$, and similarly for θ_2 in order to obtain the alternative representation

$$\inf_{u,v\geq 1, q>0} \sup_{\theta_1,\theta_2} \theta_1 u + \theta_2 v - qM(\theta_1,\theta_2) = \inf_{a,b\geq 0, q>0} \sup_{\theta_1,\theta_2} (\theta_1 + \theta_2) + a\theta_1 + b\theta_2 - qM(\theta_1,\theta_2).$$

The right-hand side of the previous display can now be seen as the Lagrangian dual of a concave optimisation problem with respect to constraints as given in the infimum; it should be borne in mind that the limiting cumulant generating function $M(\theta_1, \theta_2)$ is convex. As a consequence,

$$\inf_{a\geq 0, b\geq 0, q>0} \sup_{\theta_1,\theta_2} (\theta_1+\theta_2) + a\theta_1 + b\theta_2 - qM(\theta_1,\theta_2) = \sup_{\theta_1\geq 0,\theta_2\geq 0, M(\theta_1,\theta_2)\leq 0} \theta_1+\theta_2,$$

which establishes the proof of (2.7).

We will split the proof of the first statement, i.e., (2.5), into a lower bound and an upper bound. We first give the lower bound. For all $s, u > 0$ we have the obvious bound

$$\mathbb{P}\left(\exists t>0 : A_t > u, B_t > u\right) = \\ \mathbb{P}\left(\exists t>0 : A_{tu} > u, B_{tu} > u\right) \geq \mathbb{P}\left(A_{su} > u, B_{su} > u\right).$$

This means that also, for all $s > 0$,

$$\liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\exists t>0 : A_{tu} > u, B_{tu} > u\right) \geq \\ \liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(A_{su} > u, B_{su} > u\right).$$

As this inequality is uniform in $s > 0$, we can take the supremum on the right-hand side. We thus obtain, after rewriting:

$$\liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\exists t : \frac{A_{tu}}{tu} > \frac{1}{t}, \frac{B_{tu}}{tu} > \frac{1}{t}\right) \geq \\ \sup_{s>0} \liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\frac{A_{su}}{su} > \frac{1}{s}, \frac{B_{su}}{su} > \frac{1}{s}\right). \quad (2.8)$$

Furthermore, the Gärtner-Ellis (GE) theorem gives us the following inequality:

$$\begin{aligned}
&\liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\frac{A_{tu}}{tu} > \frac{1}{t}, \frac{B_{tu}}{tu} > \frac{1}{t}\right) \\
&= t \left[\liminf_{u\to\infty} \frac{1}{tu} \ln \mathbb{P}\left(\frac{A_{tu}}{tu} > \frac{1}{t}, \frac{B_{tu}}{tu} > \frac{1}{t}\right)\right] \\
&\overset{\text{GE}}{\geq} -t \inf_{x>\frac{1}{t}, y>\frac{1}{t}} \left[\sup_{\theta_1,\theta_2} \theta_1 x + \theta_2 y - M(\theta_1,\theta_2)\right].
\end{aligned} \tag{2.9}$$

Upon combining (2.8) and (2.9), we thus conclude

$$\begin{aligned}
&\liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\exists t : A_t > u, B_t > u\right) \\
&= \liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\exists t : \frac{A_{tu}}{tu} > \frac{1}{t}, \frac{B_{tu}}{tu} > \frac{1}{t}\right) \\
&\overset{(2.8)}{\geq} \sup_{t>0} \liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\frac{A_{tu}}{tu} > \frac{1}{t}, \frac{B_{tu}}{tu} > \frac{1}{t}\right) \\
&\overset{(2.9)}{\geq} \sup_{t>0} -t \inf_{x,y>\frac{1}{t}} \left[\sup_{\theta_1,\theta_2} \theta_1 x + \theta_2 y - M(\theta_1,\theta_2)\right] \\
&= \sup_{t>0} -t \inf_{x>\frac{1}{t}, y>\frac{1}{t}} I(x,y) \\
&= -\inf_{t>0} \inf_{x,y>\frac{1}{t}} tI(x,y) = -\inf_{x,y>0} \inf_{t>\max(\frac{1}{x},\frac{1}{y})} tI(x,y) \\
&= -\inf_{x,y>0} \max\left(\frac{1}{x}, \frac{1}{y}\right) I(x,y) = -\inf_{x>0,y>0} \frac{I(x,y)}{\min(x,y)},
\end{aligned}$$

which establishes the lower bound.

For the upper bound, we consider the probability of (A_t, B_t) reaching a set in which $(u,\infty)\times(u,\infty)$ is contained. We evidently have, for all 'weights' $w \in [0,1]$,

$$\mathbb{P}\left(\exists t > 0 : A_t > u, B_t > u\right) \leq \mathbb{P}\left(\exists t > 0 : wA_t + (1-w)B_t > u\right);$$

for the moment we keep w fixed; later in the proof we minimise over w to identify the tightest upper bound. The crucial idea is that $wA_t + (1-w)B_t$ is now a

one-dimensional stochastic process, for which we can apply the result of [32, Corollary 2.3], so as to obtain

$$\begin{aligned}&\limsup_{u\to\infty}\frac{1}{u}\ln\mathbb{P}\left(\exists t>0: A_t>u, B_t>u\right)\\&\leq\lim_{u\to\infty}\frac{1}{u}\ln\mathbb{P}\left(\exists t>0: wA_t+(1-w)B_t>u\right)\leq-\inf_{x>0}\frac{I_w(x)}{x}=-\theta_w^*,\end{aligned}$$

where (i) the rate function $I_w(x)$ is defined by $\sup_\theta(\theta x-M_w(\theta))$, (ii) the limiting cumulant generating function $M_w(\theta)$ by

$$\lim_{t\to\infty}\frac{1}{t}\ln\mathbb{E}\left(e^{\theta wA_t+\theta(1-w)B_t}\right)=0,$$

and (iii) $\theta_w^*>0$ solves $M_w(\theta)=0$. Because the above upper bound on the decay rate holds for any $w\in[0,1]$, we can take the infimum with respect to w on both sides. We thus obtain

$$\limsup_{u\to\infty}\frac{1}{u}\ln\mathbb{P}\left(\exists t>0: A_t>u, B_t>u\right)\leq-\sup_{w\in[0,1]}\theta_w^*.$$

Setting $\theta_1^*:=w\theta_w^*$ and $\theta_2^*:=(1-w)\theta_w^*$, we observe that $M(\theta_1^*,\theta_2^*)=0$. So then

$$\begin{aligned}&\limsup_{u\to\infty}\frac{1}{u}\ln\mathbb{P}\left(\exists t>0: A_t>u, B_t>u\right)\\&\quad\leq-\sup_{\theta_w^*>0,w\in[0,1]:M(w\theta_w^*,(1-w)\theta_w^*)=0}w\theta_w^*+(1-w)\theta_w^*\\&\quad=-\sup_{\theta_1\geq0,\theta_2\geq0:M(\theta_1,\theta_2)=0}\theta_1+\theta_2,\end{aligned}$$

which establishes the upper bound. □

The first part of the proof uses a Lagrange-multiplier argument in order to show that the optimising θ_1 and θ_2 are non-negative. This (seemingly novel) idea can be used more broadly; for instance in the proof of [39, Lemma 1.7]. In the proof of the upper bound, there is a one-to-one correspondence between w,θ_w^* and θ_1,θ_2, namely $\theta_1=w\theta_w^*$ and $\theta_2=(1-w)\theta_w^*$. A similar result is proven in [25] in a discrete time Markov additive setting, though the representation of the decay rate is different.

2.3 Efficient estimation of ruin probability in bivariate fluid model

In this section, as mentioned in the introduction, we let A_t and B_t represent two fluid processes, modulated by the *same* Markov process $(X_t)_{t\geq 0}$ that attains values on a finite state space $\mathcal{N}$. We apply the theory of the previous section to set up an importance-sampling-based efficient simulation procedure for estimating

$$\alpha_u = \mathbb{P}\left(\exists t > 0 : A_t > u, B_t > u\right) = \mathbb{P}\left(\exists t > 0 : (A_t, B_t) \in S_u\right).$$

To make the model precise, let r^A and r^B be two vectors in $\mathbb{R}^{|\mathcal{N}|}$. Whenever $X_t = i$, the net input per time unit of the two components are r_i^A and r_i^B, respectively; note that these numbers are not necessarily positive. A compact representation is

$$\frac{\delta A_t}{\delta t} = r_i^A, \quad \frac{\delta B_t}{\delta t} = r_i^B \qquad \text{if } X_t = i, \tag{2.10}$$

where we set $A_0 = B_0 = 0$.

In the following subsection, we analyse the above model in greater detail, and conclude that it satisfies the conditions of Theorem 2.2.2. In Section 2.3.2 we construct a method in order to estimate the ruin probability efficiently and in Section 2.3.3 we give numerical examples (which also indicate the efficiency gain with respect to naïve simulation approaches).

2.3.1 Analysis of the fluid model

In order for the process to have a positive probability to hit the set S_u, some conditions have to be imposed on r^A and r^B. These conditions can be satisfied in essentially two ways. The first way is that there exists some state i such that $r_i^A > 0$ and $r_i^B > 0$. If such a state does not exist, we need two states i and j such that $r_i^A > 0, r_i^B \leq 0, r_j^A \leq 0, r_j^B > 0$ such that if the Markov process spends time in those states in a correct ratio, both A_t and B_t increase. A geometrical intuition for these conditions is given in Figure 2.1. The following theorem characterises these conditions, and shows that if S_u can be reached, it

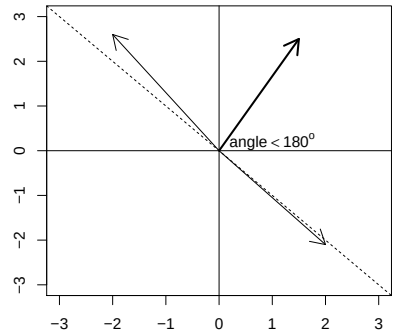

Figure 2.1: This figure illustrates the conditions under which the process can hit the set $S_u = (u, \infty) \times (u, \infty)$. The arrows represent the direction the process is going when X_t remains in some state for one time unit. The first possibility is that there is an arrow in the upper-right quadrant, like the bold arrow. The other possibility is that there are two arrows in the upper-left and lower-right quadrant respectively, of which at least one is above the dashed line and the other makes an angle less than 180° with the first one. An example for this possibility is given by the two non-bold arrows.

can do so by using, indeed, only at most two of the states of the modulating Markov process. Its (algebraic) proof is postponed to Appendix 2.5.

Theorem 2.3.1. *The following three statements are equivalent:*

1. *The joint process can reach the set S_u, in that $\mathbb{P}(\exists t : A_t > u, B_t > u) > 0$.*
2. *There exist $i, j \in \mathcal{N}$ (possibly $i = j$) and $c, d \geq 0$ such that $cr_i^A + dr_j^A > 0$ and $cr_i^B + dr_j^B > 0$.*
3. *There exists $\vec{c} \in \mathbb{R}_+^{|\mathcal{N}|}$ such that $\langle \vec{c}, r^A \rangle > 0$ and $\langle \vec{c}, r^B \rangle > 0$, where $\langle \cdot, \cdot \rangle$ denotes the inner product.*

In order to avoid trivialities, we also assume that both components have a negative drift, i.e., $\langle r^A, \pi \rangle < 0$ and $\langle r^B, \pi \rangle < 0$, where π denotes the equilibrium distribution of X.

It is not clear yet that this process satisfies the conditions of the Gärtner-Ellis theorem. The following theorem, which is a generalisation of [52, page 5], shows that it indeed does. Note also that condition (2.3) is satisfied trivially, since $(\theta_1 A_n + \theta_2 B_n)^* - \theta_1 A_n - \theta_2 B_n$ is uniformly bounded in n. We can thus apply Theorem 2.2.2 to this model.

Theorem 2.3.2. *The value of $M(\theta_1, \theta_2)$ is equal to the largest real eigenvalue of the matrix*

$$Q + \theta_1 R^A + \theta_2 R^B,$$

where R^A and R^B are diagonal matrices with the rate vectors r^A and r^B, respectively, on their diagonals. Furthermore, $M(\cdot, \cdot)$ is differentiable.

Proof. We will first derive an expression for $f(t) := \mathbb{E}(\exp(\theta_1 A_t + \theta_2 B_t)) = \sum_i f_i(t)$, where

$$f_i(t) := \mathbb{E}\left(e^{\theta_1 A_t + \theta_2 B_t} \mathbb{1}(X(t) = i)\right).$$

Relying on standard 'Markovian reasoning', as $\Delta \downarrow 0$,

$$f_i(t) = \sum_{k \neq i} f_k(t - \Delta) q_{ki} \Delta e^{\theta_1 r_{ki}^A \Delta} e^{\theta_2 r_{ki}^B \Delta} + f_i(t - \Delta)(1 - q_i \Delta) e^{\theta_1 r_i^A \Delta} e^{\theta_2 r_i^B \Delta} + o(\Delta).$$

By writing exponentials as power series we straightforwardly obtain:

$$f_i(t) = \sum_{k\neq i} f_k(t-\Delta)q_{ki}\Delta(1+\theta_1 r_{ki}^A\Delta)(1+\theta_2 r_{ki}^B\Delta) \\ +\ f_i(t-\Delta)(1-q_i\Delta)(1+\theta_1 r_i^A\Delta)(1+\theta_2 r_i^B\Delta)+o(\Delta),$$

which simplifies to

$$f_i(t) = \sum_{k\neq i} f_k(t-\Delta)q_{ki}\Delta + f_i(t-\Delta)(1-q_i\Delta+\theta_1 r_i^A\Delta+\theta_2 r_i^B\Delta)+o(\Delta).$$

Rearranging and dividing by Δ gives

$$\frac{f_i(t)-f_i(t-\Delta)}{\Delta} = \\ \sum_{k\neq i} f_k(t-\Delta)q_{ki} + f_i(t-\Delta)(-q_i+\theta_1 r_i^A+\theta_2 r_i^B)+o(1).$$

Now letting $\Delta \downarrow 0$ and realising that $q_i := -q_{ii} = \sum_{k\neq i} q_{ki}$,

$$f_i'(t) = \sum_k f_k(t)q_{ki} + f_i(t)(\theta_1 r_i^A+\theta_2 r_i^B),$$

which is in matrix-vector notation equivalent to $f'(t) = (Q^T+\theta_1 R^A+\theta_2 R^B)f(t)$. This system of linear differential equations is solved by

$$f(t) = \exp((Q^T+\theta_1 R^A+\theta_2 R^B)t)\, f(0).$$

Along the lines of [52, page 5], the first result now follows.

Note that $\exp(Q^T+\theta_1 R^A+\theta_2 R^B)$ has positive entries only. This can be seen by choosing some $a>0$ large enough such that $Q+\theta_1 R^A+\theta_2 R^B+aI\geq 0$ and hence

$$e^{Q+\theta_1 R^A+\theta_2 R^B} = e^{Q+\theta_1 R^A+\theta_2 R^B+aI-aI} = e^{-a}e^{Q+\theta_1 R^A+\theta_2 R^B+aI} > 0.$$

The last strict inequality holds because Q is irreducible, and thus so is $Q+\theta_1 R^A+\theta_2 R^B+aI$, and then according to [67, Lemma 1.3] some power of this

matrix is positive. Note that each entry of $\exp(Q^T+\theta_1 R^A+\theta_2 R^B)$ can be written as a power series in the variables θ_1 and θ_2, so each entry of $\exp(Q^T + \theta_1 R^A + \theta_2 R^B)$ is infinitely many times differentiable with respect to these variables. It thus follows from [8, Proposition 1] that M is differentiable. □

2.3.2 Simulation: construction of an efficient method

As discussed in the introduction, we propose to use importance sampling in order to efficiently estimate α_u. In this section, we identify an appropriate new measure $\mathbb{Q}$, adopting an approach similar to that used in the one-dimensional case, see e.g. [60]. There it is argued that the decay rate can be found by solving the eigensystem $-\theta^* R^A x = Qx$, i.e., by calculating the eigenvectors and eigenvalues of $(R^A)^{-1}Q$, where the non-negative eigenvector that corresponds to the largest negative eigenvalue, is used in the change of measure; such eigenvector/eigenvalue pair exists due to 'Perron-Frobenius'. For the two-dimensional model, where the analogous eigensystem is $-(\theta_1 R^a + \theta_2 R^b)x = Qx$, we can not use this method anymore, since the left-hand side of the eigensystem can not be inverted without knowing θ_1 and θ_2 beforehand. However, we can combine Theorem 2.2.2 and Theorem 2.3.2 in order to find θ_1^* and θ_2^* numerically, by using binary search on the value of $\theta_1 + \theta_2$. Once we have found θ_1^* and θ_2^*, we can use the eigensystem to calculate the appropriate eigenvector.

Theorem 2.3.3. *For all θ_1, θ_2 such that $M(\theta_1, \theta_2) = 0$, there exists $x \in \mathbb{R}_+^{|\mathcal{N}|}$ such that $-(\theta_1 R^A + \theta_2 R^B)x = Qx$.*

Proof. Since the value of $M(\theta_1, \theta_2)$ is equal to the largest real eigenvalue of $Q + \theta_1 R^A + \theta_2 R^B$, it follows from [67, Theorem 2.5(e) (p. 40, 41)] that

$$q_i > \theta_1 r_i^A + \theta_2 r_i^B$$

for all i. We need this for the existence of specific moment generating functions below. Let

$$x_{ii} := \mathbb{E}\left(\exp(\theta_1 \tilde{A}_{ii} + \theta_2 \tilde{B}_{ii})\right),$$

where $\tilde{A}_{ii}$ ($\tilde{B}_{ii}$) denotes the net amount of fluid generated by A_t (B_t) between

two consecutive visits of the Markov process to state i. Likewise, we let

$$x_{ij} := \mathbb{E}\left(e^{\theta_1 \widehat{A}_{ij} + \theta_2 \widehat{B}_{ij}}\right),$$

$i \neq j$, where $\widehat{A}_{ij}$ ($\widehat{B}_{ij}$) denotes the net amount of fluid generated by A_t (B_t) between a visit of the Markov process to state i and the next visit to state $j \neq i$. We can then write, by conditioning on the first state the Markov process visits,

$$x_{ii} \equiv x_{ii}(\theta_1, \theta_2) = \sum_{j \neq i} \frac{\lambda_{ij}}{\lambda_i} \frac{\lambda_i}{\lambda_i - \theta_1 r_i^A - \theta_2 r_i^B} x_{ji}.$$

Using the reasoning of [60], solutions of $M(\theta_1, \theta_2) = 0$ also solve $x_{ii}(\theta_1, \theta_2) = 1$. Cancelling the q_i and multiplying by the denominator of the right-hand side gives us

$$(\lambda_i - \theta_1 r_i^A - \theta_2 r_i^B) x_{ii} = \sum_{j \neq i} q_{ij} x_{ji},$$

which is equivalent to

$$(-\theta_1 r_i^A - \theta_2 r_i^B) x_{ii} = \sum_{j} q_{ij} x_{ji}.$$

In the same way we can also write, for $j \neq i$,

$$x_{ji} = \frac{q_{ji}}{q_j - \theta_1 r_j^A - \theta_2 r_j^B} + \sum_{d \neq j, d \neq i} \frac{q_{jd}}{q_j - \theta_1 r_j^A - \theta_2 r_j^B} x_{di},$$

which can be rewritten as

$$(-\theta_1 r_j^A - \theta_2 r_j^B) x_{ji} = q_{ji} + \sum_{d \neq i} q_{jd} x_{di}.$$

As we have that $x_{ii} = 1$, we obtain, for all i, j that

$$(-\theta_1 r_j^A - \theta_2 r_j^B) x_{ji} = \sum_{d} q_{jd} x_{di}.$$

For fixed i, this can be rewritten as $-(\theta_1 R^A + \theta_2 R^B) x_i = Q x_i$, with $x_i := (x_{ji})_j$. □

The new measure $\mathbb{Q}$ under which we sample $(X_t)_{t\geq 0}$ is then constructed as follows. Let θ_1^* and θ_2^* be the optimising values resulting from Theorem 2.2.2, and let x be the corresponding eigenvector as given in Theorem 2.3.2. We replace $Q = (\lambda_{ij})_{i,j}$ with $\tilde{Q} = (\tilde{q}_{ij})_{i,j}$, where $\tilde{q}_{ij} := q_{ij}x_j/x_i$ for $i \neq j$ and $\tilde{q}_i := -\tilde{q}_{ii} = q_i - r_i^A\theta_1^* - r_i^B\theta_2^*$.

Our objective is to show that this is indeed a good change of measure, in the sense that it is *asymptotically optimal* (see [64, Definition 1, p. 89, 90]). To this end, let L be the likelihood ratio of a path generated under the change of measure for which both $A_t > u, B_t > u$ for some $t > 0$. Denote by J_m the state of the Markov process after the m-th jump and let T_m denote the time spent there. Furthermore, let

$$N := \inf\{n \in \mathbb{N} : \exists t \leq t_n \text{ such that } A_t > u \text{ and } B_t > u\},$$

i.e., N is the smallest number of jumps until there was some t such that both components were bigger than u at time t. As pointed out in [64], the likelihood ratio then reads

$$L = \frac{\pi_{J_0}}{\varrho_{J_0}} \frac{q_{J_0J_1}}{\tilde{q}_{J_0J_1}} \cdots \frac{q_{J_{N-1}J_N}}{\tilde{q}_{J_{N-1}J_N}} \cdot \frac{q_{J_N}}{\tilde{q}_{J_N}} \cdot \exp\left(- \sum_{m=0}^{N} (q_{J_m} - \tilde{q}_{J_m})T_m \right), \tag{2.11}$$

with ϱ the invariant distribution of X_t under $\mathbb{Q}$.

We first point out a 'naïve' implementation, which we denote by $\mathbb{Q}_1$. Start with some initial state X_0, sampled according to ϱ; say we draw j_0. Then sample, according to the corresponding exponential distribution (i.e., with parameter $\tilde{q}_{j_0}$), some time $t_0 > 0$ for which the Markov process remains in this state. We can then update the likelihood and calculate A_{t_0} and B_{t_0}. If both are bigger than u we stop; else we sample the next state, say j_1, using the probabilities $\tilde{q}_{j_0,k}/\tilde{q}_{j_0}$ for $k \neq j_0$. We continue with this procedure until both $A_{t_N} > u$ and $B_{t_N} > u$.

There is a complication, however. With the above procedure we only check at transition epochs of the modulating Markov process whether or not S_u has been reached. However, this poses a problem which does not occur in the one-dimensional process: it could happen that at two consecutive transition epochs the process did not reach the desired set, but at some time epoch *in between* these jumps, it did. This scenario is illustrated in Figure 2.2. As a result, using

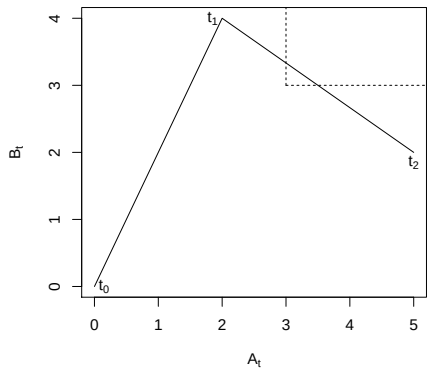

Figure 2.2: The process hits the desired set, bounded by the dashed lines, between t_1 and t_2, but is not in the set at t_1 or t_2.

this procedure we do not estimate α_u, but rather, with $U_n := T_0 + \dots T_n$,

$$\bar{\alpha}_u := \mathbb{P}\left(\exists n \in \mathbb{N} : A_{U_n} > u, B_{U_n} > u\right).$$

Clearly $\bar{\alpha}_u < \alpha_u$, creating a bias.

Whether or not the scenario of Figure 2.2 has occurred can, however, easily be checked from subsequent pairs of the form $(A_{t_{N-1}}, B_{t_{N-1}})$ and (A_{t_N}, B_{t_N}). If this happens, we propose to replace the factor

$$\frac{q_{J_N}}{\tilde{q}_{J_N}} \cdot \exp\left(-(q_{J_N} - \tilde{q}_{J_N})T_N\right) \tag{2.12}$$

in (2.11) by

$$\exp\left(-(q_{J_N} - \tilde{q}_{J_N})\tau\right), \tag{2.13}$$

where τ is the length of the interval between t_{N-1} and the first time epoch at which the process hit the desired set. The theorem below states that this adapted version of the naïve implementation, denoted by $\mathbb{Q}_2$, estimates α_u in an unbiased and asymptotically optimal way.

Theorem 2.3.4. *The implementation $\mathbb{Q}_1$ yields an unbiased, asymptotically op-*

timal estimate of $\bar{\alpha}_u$. *The implementation* $\mathbb{Q}_2$ *yields an unbiased, asymptotically optimal estimate of* α_u.

Proof. We start by proving the claim regarding $\mathbb{Q}_1$. Directly from the definition of the new rates $\tilde{q}_{ij}$ and the stopping time T_N,

$$\begin{aligned} L &= \frac{\pi_{J_0}}{\varrho_{J_0}} \frac{x_{J_0}}{x_{J_N}} \frac{q_{J_N}}{\tilde{q}_{J_N}} \cdot \exp\left(- \sum_{m=0}^{N} (q_{J_m} - \tilde{q}_{J_m}) T_m \right) = \\ &= \frac{\pi_{J_0}}{\varrho_{J_0}} \frac{x_{J_0}}{x_{J_N}} \frac{q_{J_N}}{\tilde{q}_{J_N}} \cdot \exp\left(- \sum_{m=0}^{N} (r^A_{J_m} \theta_1^* + r^B_{J_m} \theta_2^*) T_m \right) \\ &\leq k \cdot \exp\left(-(\theta_1^* + \theta_2^*) u \right) \end{aligned}$$

with

$$k := \max_{i,j} \frac{\pi_i}{\varrho_i} \frac{x_i}{x_j} \frac{q_j}{\tilde{q}_j};$$

observe that

$$\sum_{m=0}^{N} r^A_{J_m} T_m > u, \quad \sum_{m=0}^{N} r^B_{J_m} T_m > u.$$

From this upper bound on the likelihood it follows that

$$\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}_{\mathbb{Q}_1} \left(L^2 \mathbb{1} \left(\exists n \in \mathbb{N} : A_{U_n} > u, B_{U_n} > u \right) \right) \leq \lim_{u\to\infty} \frac{1}{u} \ln \left[k \cdot e^{-2(\theta_1^* + \theta_2^*) u} \right] = -2(\theta_1^* + \theta_2^*).$$

From this and Theorem 2.2.2, asymptotic optimality follows.

Regarding implementation $\mathbb{Q}_2$, observe that, because of the definition of τ,

$$\sum_{m=0}^{N-1} r^A_{J_m} T_m + r^A_{J_N} \tau > u, \quad \sum_{m=0}^{N-1} r^B_{J_m} T_m + r^B_{J_N} \tau > u.$$

In this case

$$L = \frac{\pi_{J_0}}{\varrho_{J_0}} \frac{x_{J_0}}{x_{J_N}} \frac{q_{J_N}}{\tilde{q}_{J_N}} \cdot \exp\left(-\theta_1^* \left(\sum_{m=0}^{N-1} r^A_{J_m} T_m + r^A_{J_N} \tau \right) \right)$$

$$\cdot \exp\left(-\theta_2^*\left(\sum_{m=0}^{N-1} r_{J_m}^B T_m + r_{J_N}^B \tau\right)\right)$$
$$\leq k \cdot \exp\left(-(\theta_1^* + \theta_2^*)u\right).$$

Asymptotic optimality follows as before. To show that $\mathbb{Q}_2$ indeed yields an unbiased estimate of α_u, we need to show that (2.12) and (2.13) have the same expectation under $\mathbb{Q}_2$ whenever $T_N > \tau$. Note that

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}_2}\left(e^{-(\lambda_{J_N}-\mu_{J_N})\tau}\mathbb{1}\left(T_N>\tau\right)\right) &= e^{-(\lambda_{J_N}-\mu_{J_N})\tau}\mathbb{P}_{\mathbb{Q}}\left(T_N>\tau\right)\\ &= e^{-(\lambda_{J_N}-\mu_{J_N})\tau}e^{-\mu_{J_N}\tau} = e^{-\lambda_{J_N}\tau},\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}_2}\left(\frac{\lambda_{J_N}}{\mu_{J_N}}\cdot e^{-(\lambda_{J_N}-\mu_{J_N})T_N}\mathbb{1}\left(T_N>\tau\right)\right)\\ &= \frac{\lambda_{J_N}}{\mu_{J_N}}\mathbb{E}_{\mathbb{Q}_2}\left(e^{-(\lambda_{J_N}-\mu_{J_N})T_N}|\mathbb{1}\left(T_N>\tau\right)\right)\mathbb{P}_{\mathbb{Q}_2}\left(\mathbb{1}\left(T_N>\tau\right)\right)\\ &= \frac{\lambda_{J_N}}{\mu_{J_N}}e^{-(\lambda_{J_N}-\mu_{J_N})\tau}\mathbb{E}_{\mathbb{Q}_2}\left(e^{-(\lambda_{J_N}-\mu_{J_N})(T_N-\tau)}|\mathbb{1}\left(T_N>\tau\right)\right)\\ &\quad\times\mathbb{P}_{\mathbb{Q}_2}\left(\mathbb{1}\left(T_N>\tau\right)\right)\\ &= \frac{\lambda_{J_N}}{\mu_{J_N}}e^{-(\lambda_{J_N}-\mu_{J_N})\tau}\mathbb{E}_{\mathbb{Q}_2}\left(e^{-(\lambda_{J_N}-\mu_{J_N})T_N}\right)\mathbb{P}_{\mathbb{Q}_2}\left(\mathbb{1}\left(T_N>\tau\right)\right)\\ &= \frac{\lambda_{J_N}}{\mu_{J_N}}e^{-(\lambda_{J_N}-\mu_{J_N})\tau}\frac{\mu_{J_N}}{\mu_{J_N}+(\lambda_{J_N}-\mu_{J_N})}e^{-\mu_{J_N}\tau} = e^{-\lambda_{J_N}\tau}.\end{aligned}$$

□

2.3.3 Numerical results

We now consider some numerical examples. For the first example, we consider 20 on-off processes feeding into a two-dimensional reservoir.

Each on-off process generates, while on, traffic at constant rate 3 (4) into the first (second) reservoir. The first reservoir has a constant leak rate of 30.5, where for the second reservoir this is equal to 47.5. Note that the reservoirs have an equal net input when 17 of the sources are turned on. When a source

is off, it will turn on at rate 2, while a working source turns off at rate 3. Since all sources behave identically, it suffices to take as state space $\mathcal{N} = \{0, \ldots, 20\}$, where $X_t = i$ means that at time t there are i sources turned on. The rate matrix Q is then

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ \vdots \\ 19 \\ 20 \end{array} \begin{array}{c} \begin{array}{ccccccccccc} 0 & 1 & 2 & 3 & 4 & \cdots & 18 & 19 & 20 \end{array} \\ \begin{pmatrix} -40 & 40 & & & & & & & \\ 3 & -41 & 38 & & & & & & \\ & 6 & -42 & 36 & & & & & \\ & & 9 & -43 & 34 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 57 & -59 & 2 \\ & & & & & & & 60 & -60 \end{pmatrix} \end{array},$$

and r^A and r^B are

$$r^A = \begin{pmatrix} -30.5 & -27.5 & -24.5 & \cdots & 26.5 & 29.5 \end{pmatrix},$$
$$r^B = \begin{pmatrix} -47.5 & -43.5 & -39.5 & \cdots & 28.5 & 32.5 \end{pmatrix}.$$

The results can be found in Table 2.1.

The second example has the same structure as the first example. We now consider 5 on-off sources. The first reservoir has a constant service rate of 8.5, where for the second reservoir this is equal to 12.5. The other numbers are the same as for the first example. Note that the reservoirs have an equal net input when 4 of the sources are turned on. A key difference is that under the change of measure, in this example both queues have the same drift, whereas in the first example A_t has a higher drift than B_t. Informally this means that for the second example the two processes reach level u roughly simultaneously, while for the first example the joint process will hit the set S_u when B_t does. The results can be found in Table 2.2. In both examples, the number of runs needed when using importance sampling is significantly lower than when ordinary Monte Carlo sampling is used. Furthermore, when using Monte Carlo sampling, the number of runs needed increases rapidly as a function of u, while there is only a slight increase when using importance sampling.

In our next example the two processes are positively correlated. We let r^A

	importance sampling			Monte Carlo sampling	
u	$\widehat{\alpha}$	$\widehat{\alpha}e^{\theta^* u}$	# samples	$\widehat{\alpha}$	# samples
0.25	2.95E-02	3.79E-02	4884	2.78E-02	13457
0.50	1.76E-02	2.90E-02	5667	1.70E-02	22212
1.00	7.97E-03	2.18E-02	6846	8.22E-03	46369
1.50	4.94E-03	2.23E-02	10987	4.78E-03	80170
2.00	2.24E-03	1.66E-02	5113	-	-
3.00	7.61E-04	1.55E-02	8244	-	-
4.00	2.83E-04	1.57E-02	14962	-	-
5.00	9.01E-05	1.36E-02	8493	-	-
6.00	3.42E-05	1.42E-02	9893	-	-
7.00	1.14E-05	1.28E-02	9886	-	-
8.00	4.63E-06	1.43E-02	10354	-	-
9.00	1.66E-06	1.39E-02	12600	-	-
10.00	5.87E-07	1.35E-02	12777	-	-

Table 2.1: Simulation results of the first example. We denote by $\widehat{\alpha}$ the estimated probabilities for both importance sampling and Monte Carlo sampling. Furthermore, we denote by θ^* the value of (2.7)*. This table also shows the number of samples needed to get a* 95% *confidence interval with* 10% *precision. The missing values took more than* 100000 *samples in order to give the desired precision. In this case* $\theta_1^* = 0, \theta_2^* = 1.004$.

and r^B be

$$r^A = \begin{pmatrix} -8 & -6 & 8 & 2 \end{pmatrix}, \quad r^B = \begin{pmatrix} -5 & -9 & 3 & 9 \end{pmatrix}.$$

We let Q contain only ones off the diagonal, so that all values on the diagonal are -3. We want to focus on the quality of the importance sampling results, so we did not do Monte Carlo sampling. The results can be found in Table 2.3. It can be seen that the logarithmic result is very accurate as u grows large, as the fraction of the logarithmic result and θ^* tends to unity. However, from the next column we can see that by using the logarithmic asymptotics we lose some information, as the values in that column do not tend to a constant.

In the fourth example, we let the two processes be *negatively* correlated. We let r^A and r^B be

$$r^A = \begin{pmatrix} -8 & -6 & 8 & 2 \end{pmatrix}, \quad r^B = \begin{pmatrix} 3 & 9 & -5 & -9 \end{pmatrix}.$$

	importance sampling			Monte Carlo sampling	
u	$\widehat{\alpha}$	$\widehat{\alpha}e^{\theta^* u}$	# samples	$\widehat{\alpha}$	# samples
0.25	1.37E-01	1.85E-01	685	1.38E-01	2407
0.50	9.91E-02	1.81E-01	794	1.03E-01	3339
1.00	4.85E-02	1.61E-01	770	4.88E-02	7507
1.50	2.49E-02	1.51E-01	825	2.69E-02	13908
2.00	1.45E-02	1.60E-01	1001	1.14E-02	33320
3.00	4.08E-03	1.50E-01	856	4.04E-03	94731
4.00	1.21E-03	1.47E-01	855	-	-
5.00	3.51E-04	1.42E-01	723	-	-
6.00	1.04E-04	1.40E-01	992	-	-
7.00	3.30E-05	1.47E-01	1517	-	-
8.00	1.02E-05	1.52E-01	932	-	-
9.00	2.71E-06	1.34E-01	1035	-	-
10.00	7.68E-07	1.26E-01	1075	-	-

Table 2.2: Simulation results of the second example. We denote by $\widehat{\alpha}$ the estimated probabilities for both importance sampling and Monte Carlo sampling. Furthermore, we denote by θ^* the value of (2.7)*. This table also shows the number of samples needed to get a* 95% *confidence interval with* 10% *precision. The missing values took more than* 100000 *samples in order to give the desired precision. In this case* $\theta_1^* = 0.128, \theta_2^* = 1.072$.

Note that r^A is the same as in the third example and that r^B is a permutation of r^B as in the third example. The results can be found in Table 2.4. It can be seen that the hitting probability decreases much faster than in the previous example. This can be explained by the fact that the Markov process in the previous example just needs to spend enough time in one of the two up-states for long enough, while in this example the process needs to visit *two* states long enough, and also has to do this in a correct ratio. Note that in this example it will happen frequently that the process will hit the set $(u, \infty) \times (u, \infty)$ in the way as depicted in Figure 2.2.

2.4 Outlook

There are some natural extensions of the theory developed in Section 2.2 that can be investigated, of which we briefly comment on two. The first one is

	importance sampling			
u	$\widehat{\alpha}$	$\frac{-1}{u\theta^*}\ln(\widehat{\alpha})$	$\widehat{\alpha}e^{\theta^* u}$	# samples
1	6.37E-01	4.748	7.00E-01	73
2	6.14E-01	2.572	7.42E-01	40
3	5.23E-01	2.273	6.96E-01	67
4	4.46E-01	2.125	6.52E-01	84
5	4.72E-01	1.580	7.59E-01	54
6	3.62E-01	1.784	6.40E-01	96
7	3.28E-01	1.677	6.37E-01	100
8	2.98E-01	1.592	6.38E-01	89
9	3.11E-01	1.366	7.31E-01	70
10	2.58E-01	1.428	6.66E-01	80
20	8.60E-02	1.291	5.75E-01	129
50	4.80E-03	1.124	5.54E-01	181
100	3.39E-05	1.084	4.52E-01	260
250	2.00E-11	1.038	4.09E-01	344

Table 2.3: Simulation results of the third example. We denote by $\widehat{\alpha}$ the estimated probabilities. Furthermore, we denote by θ^* the value of (2.7)*. This table also shows the number of samples needed to get a* 95% *confidence interval with* 10% *precision. In this case* $\theta_1^* = 0.094, \theta_2^* = 0.001$.

whether the two-dimensional theory can be extended to arbitrary higher dimensions. At the expense of introducing additional notation and loss of transparency, Theorem 2.2.2 can indeed be extended to a multi-dimensional result. For proving (2.5), the proof of the lower bound holds true trivially in multiple dimensions. In proving the upper bound, instead of using a single weight $w \in [0, 1]$, one has to consider non-negative weights $w_1, w_2, \ldots, w_d$ that add up to unity. The proof of (2.7) does not use any specific two-dimensional argument, so is also valid for multiple dimensions.

The second extension considers a bivariate process for which the two components have to be larger than u but can do so at *different* times. Of course, this probability is greater than the probability of the event which is considered in this chapter. A significant difference also occurs in simulating, because the simulation can now be split in two parts. First, one simulates under a change-of-measure until one of the components has hit level u. After this has occurred, one can use a *different* change of measure until the other component hits level

	importance sampling			
u	$\widehat{\alpha}$	$\frac{-1}{u\theta^*}\ln(\widehat{\alpha})$	$\widehat{\alpha}e^{\theta^* u}$	# samples
1	1.34E-01	3.002	2.62E-01	657
2	6.05E-02	2.098	2.30E-01	766
3	2.95E-02	1.757	2.19E-01	1179
4	1.32E-02	1.619	1.91E-01	1180
5	6.60E-03	1.502	1.87E-01	1531
6	3.26E-03	1.428	1.80E-01	1881
7	1.59E-03	1.377	1.72E-01	1858
8	7.60E-04	1.343	1.60E-01	5005
9	3.84E-04	1.307	1.58E-01	3319
10	1.79E-04	1.291	1.43E-01	4241
20	1.77E-07	1.163	1.13E-01	8366
50	2.16E-16	1.079	7.10E-02	6000
100	5.11E-31	1.043	5.54E-02	26001
250	8.85E-75	1.020	3.43E-02	22857

Table 2.4: Simulation results of the fourth example. We denote by $\widehat{\alpha}$ the estimated probabilities. Furthermore, we denote by θ^* the value of (2.7)*. This table also shows the number of samples needed to get a* 95% *confidence interval with* 10% *precision. In this case* $\theta_1^* = 0.344, \theta_2^* = 0.325$.

u; under this different change of measure, the first component to have hit level u may now also have a negative drift, since it has already hit level u. This extension is the topic of Chapter 3.

2.5 Proof of Theorem 2.3.1

Proof. This is shown by proving a number of implications.
◦ "2 $\Rightarrow$ 1": Suppose there exist $i, j \in \mathcal{N}$ (possibly $i = j$) and $c, d \geq 0$ such that $cr_i^A + dr_j^A > 0$ and $cr_i^B + dr_j^B > 0$. Assume without loss of generality that a direct jump of the Markov process from state i to state j is possible. We may do this, because if this jump is not possible, we can jump from state i to state j in arbitrarily small time (because we assumed the Markov process is irreducible). Although this may have a very small probability, of importance is only that it

has a *positive* probability. Let

$$T := \frac{(c+d)u}{cr_i^A + dr_j^A} \vee \frac{(c+d)u}{cr_i^B + dr_j^B}.$$

If $X_t = i$ for $0 \leq t < \frac{c}{c+d}T$ and $X_t = j$ for $\frac{c}{c+d}T \leq t \leq T$, then

$$A_T = \frac{1}{c+d}(cr_i^A + dr_j^A)T \geq u;$$
$$B_T = \frac{1}{c+d}(cr_i^B + dr_j^B)T \geq u.$$

◦ "3 ⇒ 2": This follows from Carathéodory's theorem for convex cones, see [37, Theorem 2.4], but we give a direct proof below. Define the sets PP, PM, MP and MM (P stands for plus and M for minus) as follows: PP := $\{k \in \mathcal{N} : r_k^A > 0, r_k^B > 0\}$, PM := $\{k \in \mathcal{N} : r_k^A > 0, r_k^B < 0\}$, etc. If there exists $k \in$ PP, then we are done (choose $i = j = k$ and $c = d = 1$). Assume now without loss of generality that each $k \in \mathcal{N}$ is either in PM or MP. We can do this, because we can change the entries of $\vec{c}$ belonging to states in MM to zero, which makes both inner products even larger. Note now that neither PM nor MP are empty. We let

$$i := \arg\max_{k \in \text{PM}} \frac{r_k^A}{-r_k^B}, \quad j := \arg\max_{k \in \text{MP}} \frac{r_k^B}{-r_k^A},$$

i.e., i and j are the states which are the "closest" to the first quadrant in a geometrical sense. We now want to shift "mass" from the entries of $\vec{c}$ to c_i and c_j in such a way that both inner products do not decrease. A correct way of doing that is by constructing c, d as follows:

$$c := c_i + \sum_{k \in \text{PM}, k \neq i} c_k \frac{r_k^B}{r_i^B}, \qquad d := c_j + \sum_{k \in \text{MP}, k \neq j} c_k \frac{r_k^A}{r_j^A}.$$

In this way, the negative contributions stay the same, while the positive contributions do not decrease. We have thus found $i, j \in \mathcal{N}$ and $c, d \geq 0$ with the desired requirements.

◦ "1 ⇒ 3": Suppose we have a sample path such that $A_t > u, B_t > u$ for some

t. We can then write

$$A_t = \int_0^t r^A_{X_s}\,\mathrm{d}s = \sum_{i=1}^{|\mathcal{N}|} c_i r_i^A, \quad B_t = \int_0^t r^B_{X_s}\,\mathrm{d}s = \sum_{i=1}^{|\mathcal{N}|} c_i r_i^B,$$

for some $c_i \geq 0$. Choose $\vec{c} = (c_1, c_2, \ldots, c_{|\mathcal{N}|})$. □

3

Large delay probabilities in two correlated queues

Whereas in Chapter 2 we analyse the event of the components of a bivariate process exceeding a high threshold *simultaneously*, this chapter focuses on the variant where the high threshold can be exceeded at *different* times. This chapter is based on [22].

3.1 Introduction

Model Consider a two-dimensional random walk $((A_s, B_s))_{s\in\mathbb{N}}$ with i.i.d. increments and with the partial sum processes denoted by

$$A_s := \sum_{i=1}^{s} X_i, \quad B_s := \sum_{i=1}^{s} Y_i.$$

The focus is on the probability $\pi(u) := \mathbb{P}\left(\exists s \in \mathbb{N} : A_s \geq u, \exists t \in \mathbb{N} : B_t \geq u\right)$, i.e., the probability of the event that both components will ever exceed some large level u, but not necessarily at the same time. We allow that X_i and Y_i are dependent; note that if they were independent, the probability of interest would simply be the product of the marginal probabilities. An exact analysis of $\pi(u)$, however, seems possible only in special cases. In all of them, the model

is such that the components are ordered: $A_s \leq B_s$ for all s, which implies that the epochs that the two components achieve their respective maximum values are almost surely ordered. These special cases cover tandem systems of $M/D/1$ queues [59] and a tandem Brownian queue [58]. In both models, let D_s denote the amount of work that has arrived up to time s and let c_1 and c_2 denote the constant service rate of the upstream and downstream queue respectively (assume that $c_1 > c_2$). The stationary workloads of the upstream and the sum of the upstream and downstream queues are then respectively distributed as the supremum of $A_s = D_s - c_1 s$ and the supremum of $B_s = D_s - c_2 s$. In the above-mentioned papers the analysis relies on the availability of the distribution of the maximum over a finite interval, given that we know the position at the end of the interval (in the Brownian case this is a Brownian bridge, in the $M/D/1$ case it can be dealt with using ballot theorems); the fact that such results are not generally available complicates the extension to more general models.

In the transform domain, results for the joint distribution of $\sup_{s\in\mathbb{N}} A_s$ and $\sup_{s\in\mathbb{N}} B_s$ have been established under more general conditions, but still an 'ordering property' of the type mentioned above needs to be imposed; see for instance [50, 48, 29]. Having expressions for such multivariate transforms, one still needs to perform numerical inversion to obtain numerical output, which tends to be challenging in the tail of the multivariate distribution. Therefore, we resort in this chapter to large deviations and to rare-event simulation. We consider the rare-event regime in which both $\mathbb{E}(X_i) < 0$ and $\mathbb{E}(Y_i) < 0$.

This model has several applications. First, it can model two correlated queues. A queue is essentially a stochastic process reflected at zero. Consider now two queues fed by the (possibly correlated) input processes A_s and B_s. Then their stationary versions obey the distributional equalities, see e.g. [39, Section 1.1],

$$Q_1 \stackrel{\mathrm{d}}{=} \sup_s A_{-s}, \quad Q_2 \stackrel{\mathrm{d}}{=} \sup_s B_{-s},$$

which follow as a direct application from Lindley's recursion. The steady-state probability of both queues having more than an amount u of work is therefore equal to

$$\mathbb{P}(Q_1 > u, Q_2 > u) = \mathbb{P}(\exists s : A_{-s} > u, \exists t : B_{-t} > u),$$

which is, after reversing time, precisely the probability of our interest. Another application is in risk management, where we can use the model to study rare

events in the context of two correlated portfolios; see e.g. [6] and references therein.

Literature There is a substantial literature on efficient estimation of rare-event probabilities for queueing systems, see e.g. the surveys [45], [47], and [12]. We here provide an account of this literature (without aiming at being exhaustive), focusing on multivariate rare events. In addition, we briefly comment on how these results relate to ours.

Over the past decades, different techniques have been developed, the most prominent being importance sampling (based on a change of measure) and splitting; our present study falls in the former category. Building on the ideas of e.g. [69], [65] focuses on estimating overflow-related quantities in a stable $GI/GI/m$ queue using importance sampling. Later attention shifted to more sophisticated queueing systems. In [28] it is assessed to what extent state-independent change of measures can lead to asymptotically efficient performance in two-node tandem Jackson networks. The probability of interest in that paper is of an overflow event in a two-node tandem Jackson network, whereas we focus on the two components of a two-dimensional random walk *both* ever reaching a high level. In [33], where the focus is on two-node tandem Jackson networks too, the authors consider, contrary to this chapter, state-*dependent* changes of measure. A generalisation to arbitrary Jackson networks is treated in [34]. In both papers, the so-called subsolution method is used, which is also briefly discussed in this chapter. Improved results are given in [9], where the author focuses on optimal simulation algorithms for overflow probabilities during a busy period. Instead of using exponential twisting forward in time, the author proposes a method that goes backwards in time.

This chapter is a logical continuation of our previous work, see the previous chapter or [21]. In that chapter, we study a similar model but there the event of interest corresponds to both components exceeding a large level *at the same time* (whereas in the current chapter these epochs can be different).

Decay rate The first result of this chapter is Theorem 3.3.1, which provides an expression for the *decay rate*

$$\lim_{u\to\infty} \frac{1}{u} \ln \pi(u).$$

The proof of this result uses two important theorems in large deviations theory, namely Cramér's theorem and Mogulskii's theorem. In the proof, the lower bound is attained by conditioning where the "slower" component (i.e. the component that hits level u second) of the process is when the "fast" component hits level u for the first time. For the upper bound, we first show that the decay rate can be bounded by the decay rate of the probability of the event of interest occurring on a bounded time-interval. We then use this bounded interval to apply Mogulskii's theorem. Then we apply a "linear geodesics" type of argument to show that the obtained rate function over general sample paths is the same as over some set of piecewise linear sample paths.

Importance sampling and challenges The second result of this chapter is the construction of an efficient simulation method in order to estimate $\pi(u)$. Since Monte Carlo simulation is slow due to the rarity of the event of interest when u gets large, we resort to importance sampling (IS). Importance sampling is a method to simulate stochastic systems using a different underlying probability measure, such that the (rare) event of interest is not rare anymore; the simulation output is weighted by appropriate likelihood ratios to recover unbiasedness. Each probability measure leads to a particular variance performance, and it is therefore crucial to identify the one that is, according to some specific definition, optimal. There exist various performance metrics; the metric we use is called *asymptotic optimality*. We refer to Section 3.4 for the definition. For our IS procedure, we first propose a "naive" change of measure based on the decay rate given in Theorem 3.3.1. We will show, however, that this approach does not necessarily perform well. More specifically, we show that using this new measure for a "nearest-neighbour random walk" results in a procedure that is not asymptotically optimal. The underlying problem with this procedure is that at the moment when the "fast" component of the process hits level u, we don't have any control over the position of the "slow" component.

Partitioned IS In order to solve this problem, we introduce *partitioned importance sampling*. This approach is based on conditioning where the "slow" component of the process has to be when the "fast" component hits level u for the first time. More specifically, we partition the event of interest into disjoint events and perform simulations to estimate the probabilities corresponding to

those events. For more details we refer to Section 3.5. We show that this approach is indeed asymptotically optimal. It is pointed out how the method's inherent bias, arising from the need to truncate the infinite sum obtained through this method, can be made arbitrarily small.

Numerical results The results above are illustrated through various numerical experiments. In order to carry out the simulations, we chose a specific instance of the model, namely a model in which the increments (X_i, Y_i) have a bivariate normal distribution. We investigate how the performance of the three simulation methods described above (i.e., Monte Carlo, naive IS and partitioned IS) depend on various factors, e.g., the level u, the covariance of the two components and the number of partitions used in partitioned importance sampling.

Organisation of this chapter The rest of this chapter is organised as follows. Section 3.2 gives a detailed description of the model and a brief overview of large-deviations theory. In Section 3.3 we state the first main result of this chapter, namely an expression for the decay rate of the probability of the event of interest. In Section 3.4 we give a first naive importance sampling-based simulation scheme and we show that this method is not asymptotically optimal. This is remedied in Section 3.5, where we introduce partitioned importance sampling. Moreover, we show that this new approach is indeed asymptotically optimal. These findings are illustrated in Section 3.6, where we give numerical results of various simulation experiments. The proof of the result in Section 3.3 is given in Section 3.7.

3.2 Model description and preliminaries

3.2.1 The model

Consider the bivariate random walk $((A_s, B_s))_{s\in\mathbb{N}}$ where the partial sum process is denoted by, for $s \in \mathbb{N}$,

$$A_s := \sum_{i=1}^{s} X_i, \quad B_s := \sum_{i=1}^{s} Y_i,$$

with (X_i, Y_i) i.i.d. bivariate random vectors (whose components are not necessarily independent). We also introduce the events

$$\mathscr{A}_s(u) \equiv \mathscr{A}_s := \{A_s \geq u\}, \quad \mathscr{B}_s(u) \equiv \mathscr{B}_s := \{B_s \geq u\}.$$

The main object of study of this chapter is the probability $\pi(u)$ that both A and B exceed some (large) threshold u, but not necessarily at the same time:

$$\pi(u) := \mathbb{P}\left(\exists s \in \mathbb{N} : A_s \geq u, \exists t \in \mathbb{N} : B_t \geq u\right) = \mathbb{P}\left(\left(\bigcup_{s=1}^{\infty} \mathscr{A}_s\right) \cap \left(\bigcup_{t=1}^{\infty} \mathscr{B}_t\right)\right).$$

It is assumed throughout that both $\mathbb{E}(X_1)$ and $\mathbb{E}(Y_1)$ are negative, such that $\pi(u)$ is a rare-event probability, for u large.

Since an exact analysis of this probability seems not possible in general, we will look at the so-called *decay rate* of this probability:

$$\lim_{u \to \infty} \frac{1}{u} \ln \pi(u). \tag{3.1}$$

Our main result, which is an expression for the decay rate and which is stated in the next section, depends on both Cramér's and Mogulskii's theorem. Before we state Cramér's theorem below, a quick recap of some large deviation theory is first given. We follow the setup of [39].

3.2.2 Preliminaries from large deviations

We define the *limiting cumulant generating function* of $((A_s, B_s))_{s \geq 0}$ as

$$\ln \Lambda(\theta, \eta) := \lim_{s \to \infty} \frac{1}{s} \ln \mathbb{E}\left(e^{\theta A_s + \eta B_s}\right) = \ln \mathbb{E}\left(e^{\theta X + \eta Y}\right). \tag{3.2}$$

A function $I : \mathbb{R}^2 \to \mathbb{R}^*$ (where $\mathbb{R}^* := \mathbb{R} \cup \{\infty\}$) is a *rate function* if it is non-negative and if it is lower semi-continuous, i.e., all level sets are closed (level sets of some function f are sets of the form $\{x : f(x) \leq \alpha\}$, $\alpha \in \mathbb{R}$). Furthermore, it is called a *good* rate function if in addition all level sets are compact. We say that $((A_s, B_s))_{s \geq 0}$ satisfies a *large deviations principle* in $\mathbb{R}^2$ with rate function $I : \mathbb{R}^2 \to \mathbb{R}^*$ if for any measurable set $F \subseteq \mathbb{R}^2$

$$-\inf_{x\in F^\circ} I(x) \le \liminf_{s\to\infty} \frac{1}{s}\ln \mathbb{P}\left((A_s,B_s)\in F\right) \le \limsup_{s\to\infty} \frac{1}{s}\ln \mathbb{P}\left((A_s,B_s)\in F\right) \le -\inf_{x\in F^{\mathrm{c}}} I(x),$$

where F° and F^{c} denote the interior and closure of F respectively. For any function $f:\mathbb{R}^d\to\mathbb{R}^*$, we denote its *convex conjugate* by $f^\star(x) := \sup_\theta \langle\theta,x\rangle - f(\theta)$.

3.3 Large deviations result

This section presents the first main result of this chapter, namely Theorem 3.3.1, which provides an expression for the decay rate (3.1). This result is based on both Cramér's theorem and Mogulskii's theorem. It requires the definition of joint (bivariate, that is) moment generating functions and Legendre transforms. Recall that

$$\Lambda(\theta,\eta) = \mathbb{E}\left(e^{\theta X+\eta Y}\right);$$

we define its univariate counterparts through

$$\Lambda_1(\zeta) := \Lambda(\zeta,0), \qquad \Lambda_2(\zeta) := \Lambda(0,\zeta),$$

which we assume to satisfy the condition of Mogulskii's theorem, i.e., we assume (3.2) to be finite everywhere. In addition, the bivariate Legendre transform is given through

$$I(x,y) := \sup_{\theta,\eta}\left(\theta x+\eta y-\ln\Lambda(\theta,\eta)\right),$$

and its univariate counterparts through

$$I_1(x) := \sup_{\zeta}\left(\zeta x-\ln\Lambda_1(\zeta)\right), \qquad I_2(x) := \sup_{\zeta}\left(\zeta x-\ln\Lambda_2(\zeta)\right).$$

Define, for $i=1,2$,

$$\alpha_i := \inf_{z>0}\frac{I_i(z)}{z}. \tag{3.3}$$

The main result of this section gives an expression for the decay rate (3.1) in terms of a variational problem.

Theorem 3.3.1. *If* (3.2) *exists and is finite everywhere, then*

$$\begin{aligned}&\lim_{u\to\infty}\frac{1}{u}\ln\pi(u)\\&=-\min\left\{\inf_{x>0,y\le x}\left(\frac{I(x,y)}{x}+\left(1-\frac{y}{x}\right)\alpha_2\right),\\&\inf_{y>0,x<y}\left(\frac{I(x,y)}{y}+\left(1-\frac{x}{y}\right)\alpha_1\right)\right\}.\end{aligned}\tag{3.4}$$

Proof. Due to the complexity and length, the proof is postponed to Section 3.7. Instead, we will give a short summary of the main ideas here.

We will prove this as a lower and an upper bound. For the lower bound, we first look at specific times s and t such that $A_s \geq u$ and $B_t \geq u$. We condition on where the slower process is, call this position y, when the faster process hits level u. We then first use Cramér's theorem and subsequently optimise over the position y. Taking the supremum over s and t then gives us, after some rewriting, the correct decay rate. For the upper bound, we use the union bound to again condition where the slow process is when the fast process hits level u. We then argue that we only need to look at the probability of the event occurring in a bounded time interval. This allows us to use Mogulskii's theorem to arrive at the claimed expression. □

In the next section, we will focus on estimating $\pi(u)$ numerically for u large. In that section, we propose an implementation of importance sampling which is based on the ideas behind the decay rate presented in Theorem 3.3.1. We will show, however, that this method, although natural, is not necessarily asymptotically efficient.

3.4 Importance sampling and efficiency

In this and the next sections, we focus on estimating $\pi(u)$ numerically. We will show that a naive importance sampling procedure, based on the decay rate given by Theorem 3.3.1, is not asymptotically optimal in general; recall that a

simulation procedure is called *asymptotically optimal* if, in self-evident notation,

$$\lim_{u\to\infty} \frac{\ln \mathbb{E}_{\mathbb{Q},u}\left(L^2 I\right)}{\ln \mathbb{E}_{\mathbb{Q},u}\left(LI\right)} = 2, \tag{3.5}$$

where $\mathbb{Q}$ is the new measure and L is the likelihood ratio, or Radon-Nikodym derivative, between $\mathbb{P}$ and $\mathbb{Q}$, i.e., $L = \frac{\delta \mathbb{P}}{\delta \mathbb{Q}}$. We refer to e.g. [12, Definition 1] for background information on this optimality concept, or [57] for an in-depth account of various performance metrics. Note that by Jensen's inequality, the limit in (3.5) is always smaller than or equal to 2, so it is left to prove that it is larger than or equal to 2. We say that we *exponentially twist* a random variable X, having density $f_{\mathbb{P}}(\cdot)$, with parameter θ if under $\mathbb{Q}$, the density of X equals (in self-evident notation) $f_{\mathbb{Q}}(x) = f_{\mathbb{P}}(x)e^{\theta x}/\mathbb{E}_{\mathbb{P}}\left(e^{\theta X}\right)$.

As an instance of the model, we will consider a "nearest-neighbour random walk". More specifically, we let

$$(X_i, Y_i) = \begin{cases} (1,0) & \text{w.p. } p_1; \\ (1,1) & \text{w.p. } p_2; \\ (-1,-1) & \text{w.p. } p_3; \\ (-1,1) & \text{w.p. } p_4. \end{cases} \tag{3.6}$$

We will assume that the probabilities add up to unity.

The naive importance sampling procedure, which will yield the change of measure $\mathbb{Q}$, is as follows. Each simulation run consists of (up to) two sequential exponential twists; the first twist is used to bring one of the components up to level u, whereas the second twist (if still necessary) is used to bring the other (slower) component up to level u. Denote by $\theta^\star$ and $\eta^\star$ the optimising parameters of $I(\cdot,\cdot)$ in (3.4). We first exponentially twist the joint process with parameter $(\theta^\star, \eta^\star)$ until one of the components hits level u. If process A (B) hits level u first, then we exponentially twist process B (A) with parameter $\zeta^\star$, which is defined as the optimising parameter of $I_2(\cdot)$ ($I_1(\cdot)$), until this component hits level u. The following will be used in the proofs op Property 3.4.2 and Theorem 3.5.1: Let $S \equiv S(u)$ be the first passage time of level u for process A, i.e., $S := \inf\{s : A_s \geq u\}$. Furthermore, let T be the analogous counterpart for process B.

This procedure follows naturally from Theorem 3.3.1 as it tries to mimic the most likely path given in the theorem: first twist both processes until the fastest reaches level u and then twist the slower process.

Property 3.4.1. *For any x and y in the infimum of Theorem 3.3.1 (even the non-optimal values), if we perform the exponential twist as described above using the optimal $\theta^\star, \eta^\star$ corresponding to these x and y, then*

$$\mathbb{E}_{\mathbb{Q}}(X_i) = x, \qquad \mathbb{E}_{\mathbb{Q}}(Y_i) = y.$$

Proof. We only prove this for $\mathbb{E}_{\mathbb{Q}}(X_i)$, since $\mathbb{E}_{\mathbb{Q}}(Y_i)$ can be dealt with analogously. In the supremum of I, the first-order conditions are

$$x - \frac{\partial \ln \Lambda(\theta,\eta)}{\partial\theta} = x - \frac{\frac{\partial\Lambda(\theta,\eta)}{\partial\theta}}{\Lambda(\theta,\eta)} = 0, \qquad y - \frac{\partial \ln \Lambda(\theta,\eta)}{\partial\eta} = y - \frac{\frac{\partial\Lambda(\theta,\eta)}{\partial\eta}}{\Lambda(\theta,\eta)} = 0.$$

Note now that by twisting, we have

$$\mathbb{E}_{\mathbb{Q}}(X_i) = \frac{\int z e^{\theta^\star z} f(z)\,\mathrm{d}z}{\int e^{\theta^\star z} f(z)\,\mathrm{d}z} = \frac{\frac{\partial\Lambda(\theta,\eta)}{\partial\theta}}{\Lambda(\theta,\eta)} = x,$$

where the first equality comes from the definition of an exponential twist, the second equality comes from the definition of the moment generating function and the final identity comes from the display above. □

Under any exponential twist, the model will be as follows:

$$(X_i, Y_i) = \begin{cases} (1,0) & \text{w.p. } \tilde{p}_1; \\ (1,1) & \text{w.p. } \tilde{p}_2; \\ (-1,-1) & \text{w.p. } \tilde{p}_3; \\ (-1,1) & \text{w.p. } \tilde{p}_4. \end{cases} \tag{3.7}$$

Property 3.4.2. *The naive importance sampling procedure (using probability measure $\mathbb{Q}$) as described above is* not *asymptotically optimal.*

Proof. Without loss of generality assume that A hits level u first. In the case that B hits level u first, the proof is analogous. The proof indicates that the

spread in the vertical direction, i.e., the position of B_S, at time S causes the procedure to be not asymptotically optimal. Observe that we can write

$$B_{S(u)} = \sum_{i=1}^{u} V_i,$$

where the V_i are i.i.d. and $V_i \stackrel{\mathrm{d}}{=} V$ corresponding to the vertical position at the moment the horizontal position (i.e., corresponding to A) attains the value 1 for the first time. Hence,

$$\mathbb{E}\left(z^{B_{S(u)}}\right) = \left(\mathbb{E}\left(z^{V}\right)\right)^u = \phi(z)^u,$$

where the right-hand side can be seen as the u'th power of some probability generating function. From this it follows that, for all $\theta > 0$,

$$\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}\left(e^{\theta B_{S(u)}}\right) = \ln \mathbb{E}\left(e^{\theta V}\right) = \theta \mathbb{E}\left(V\right) + G(\theta), \tag{3.8}$$

with $G(\theta) := \ln \mathbb{E}\left(e^{\theta(V-\mathbb{E}(V))}\right)$ such that $G(0) = G'(0) = 0$ and $G(\cdot)$ is strictly convex.

Note that the likelihood ratio can be written as

$$L(u) = \Lambda(\theta^\star, \eta^\star) \exp\left(-\theta^\star A_S - \eta^\star B_S\right) \cdot \Lambda(0, \zeta^\star) \exp\left(-\zeta^\star (B_T - B_S)\right).$$

It is an elementary exercise to verify that $\Lambda(\theta^\star, \eta^\star) = \Lambda(0, \zeta^\star) = 1$ (which can be checked e.g. by working out the first order conditions). Moreover, in this model we know that $A_S = B_T = u$. Hence, the likelihood ratio reads

$$L(u) = \exp\left(-\theta^\star u - \eta^\star B_S\right) \cdot \exp\left(-\zeta^\star (u - B_S)\right).$$

From this, we obtain,

$$\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}_{\mathbb{Q},u}\left(LI\right) = -(\theta^\star + \zeta^\star) + \lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}_{\mathbb{Q},u}\left(\exp(-(\eta^\star - \zeta^\star) B_S)\right),$$

and in the same way, we get for the second moment that

$$\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}_{\mathbb{Q},u}\left(L^2 I\right) = -2(\theta^\star + \zeta^\star) + \lim_{u\to\infty} \frac{1}{u} \ln \mathbb{E}_{\mathbb{Q},u}\left(\exp(-2(\eta^\star - \zeta^\star) B_S)\right).$$

From the above properties of $G(\cdot)$, it follows that $G(2\theta) > 2G(\theta)$ for all $\theta \neq 0$. From this, it follows that

$$\ln \mathbb{E}_{\mathbb{Q},u}\left(e^{-2(\eta^\star-\zeta^\star)(V-\mathbb{E}(V))}\right) > 2\ln \mathbb{E}_{\mathbb{Q},u}\left(e^{-(\eta^\star-\zeta^\star)(V-\mathbb{E}(V))}\right),$$

or equivalently

$$\begin{aligned} &-2(\theta^\star+\zeta^\star) + \lim_{u\to\infty}\frac{1}{u}\ln \mathbb{E}_{\mathbb{Q},u}\left(\exp(-2(\eta^\star-\zeta^\star)B_S)\right) \\ >&-2(\theta^\star+\zeta^\star) + 2\lim_{u\to\infty}\frac{1}{u}\ln \mathbb{E}_{\mathbb{Q},u}\left(\exp(-(\eta^\star-\zeta^\star)B_S)\right). \end{aligned}$$

This reduces to

$$\lim_{u\to\infty}\frac{1}{u}\ln \mathbb{E}_{\mathbb{Q},u}\left(L^2 I\right) > 2\lim_{u\to\infty}\frac{1}{u}\ln \mathbb{E}_{\mathbb{Q},u}\left(LI\right),$$

which indeed proves that the procedure is *not* asymptotically optimal. □

The proof above indicates that the naive importance sampling procedure cannot be guaranteed to be asymptotically efficient because of the spread in the position of the slower component when the faster component hits level u for the first time. In particular, there is no lower bound for this position. In the next section, we try to overcome this complication by introducing a method called "partitioned importance sampling". This is a method that possesses the desired control over the position of the slower process.

3.5 Partitioned importance sampling

In the previous section we have seen that ordinary importance sampling does not yield an asymptotically optimal procedure in general. In this section, we give a procedure that *is* asymptotically optimal, namely partitioned importance sampling. The method's inherent bias can be made arbitrarily small.

The procedure consists of intersecting the set of interest by a partitioning in terms of the value of B_S. More concretely, we consider a decomposition into probabilities of disjoint events. First, let

$$\pi_1(u) := \mathbb{P}\left(\exists s,t\in\mathbb{N}, s\le t : \forall r<s : A_s \ge u, B_r < u, B_t \ge u\right),$$

$$\pi_2(u) := \mathbb{P}\left(\exists s, t \in \mathbb{N}, s > t : \forall r \leq t : B_t \geq u, A_r < u, A_s \geq u\right),$$

i.e., $\pi_1(u)$ is the probability that A hits level u before B, and $\pi_2(u)$ is its analogous counterpart. Note that $\pi(u) = \pi_1(u) + \pi_2(u)$. It is sufficient to show how $\pi_1(u)$ can be estimated efficiently, since the method for estimating $\pi_2(u)$ can be set up analogously. The decomposition we consider is:

$$\pi_1(u) = \sum_{k=-\infty}^{\infty} \pi_{1,k}(u),$$

where the probabilities $\pi_{1,k}(u)$ are defined by

$$\pi_{1,k}(u) := \mathbb{P}\left(\exists s, t \in \mathbb{N}, t \geq s : \forall r < s : A_s \geq u, B_r < u, B_t \geq u, B_s \in s_k\right),$$

with $s_k := [kf(u), (k+1)f(u))$, where $f(\cdot)$ is a positive function (on which we impose some conditions below). We let $m \equiv m(u), M \equiv M(u)$ be a suitably chosen truncation, possibly dependent on u, and hence we estimate

$$\pi_1^{(\mathrm{app})}(u) := \sum_{k=m}^{M} \pi_{1,k}(u);$$

clearly, by choosing m sufficiently small and M sufficiently large the error made is negligible. We furthermore, in order to guarantee asymptotic optimality, need to impose that $M(u) - m(u)$ grows subexponentially as a function of u, i.e., we require that $\lim_{u\to\infty} \frac{1}{u} \ln(M(u) - m(u)) = 0$. We also require that $kf(u)/u \leq 1$ for each $k \in \{m(u), \ldots, M(u)\}$, and impose the related property that $f(u)$ grows sublinearly in u, for reasons that will become clear soon. In our simulation procedure we perform a separate simulation run for each $k \in \{m(u), \ldots, M(u)\}$, as follows.

We start by solving, for each $k \in \{m, \ldots, M\}$, using definition (3.3),

$$J_k(u) = \inf_x \frac{I(x, x\,k\frac{f(u)}{u})}{x} + \left(1 - k\frac{f(u)}{u}\right)^+ \alpha_2; \tag{3.9}$$

bear in mind that if $f(u)$ had been allowed to grow superlinearly, the second term would be identical 0 eventually (for positive k). Denote the optimisers by

$\theta_k^\star$ and $\eta_k^\star$ (from the definition of $I(\cdot,\cdot)$) and $\zeta \equiv \zeta_k^\star$ (from the definition of $I_2(\cdot)$; use (3.3), and conclude that this twist does not depend on k). Observe that the first element in the minimum of (3.4) can now be majorised as follows: (3.9), and in particular

$$J^{(1)} := \inf_{x>0, y\le x} \left(\frac{I(x,y)}{x} + \left(1 - \frac{y}{x}\right) \alpha_2 \right) \le \lim_{u\to\infty} \inf_k J_k(u);$$

the reason for the inequality is that in the left hand side the minimum is taken over a larger set than in the right hand side. In our refined algorithm, when estimating $\pi_{1,k}(u)$, we first twist the (X_s, Y_s) by $(\theta_k^\star, \eta_k^\star)$ until A exceeds u, and from that point on twist the (Y_s) by $\zeta^\star$ until B exceeds u (if needed). The simulation output of a single run is $L_k I_k$, with $L_k \equiv L_k(u)$ again the likelihood ratio, and the indicator function $I_k \equiv I_k(u)$ which equals 1 iff the path is such that both A and B exceed u, but now in addition that (i) A is required to exceed u before B does (or simultaneously), and (ii) when A exceeds u, B is in the interval s_k; it is this latter requirement that gives us control on the variance of the estimator, as will turn out below.

Theorem 3.5.1. *The simulation procedure described above is asymptotically optimal.*

Proof. Observe that the likelihood reads

$$L_k(u) = \left(\Lambda(\theta_k^\star, \eta_k^\star)\right)^S \exp\left(-\theta_k^\star A_S - \eta_k^\star B_S\right) \cdot \left(\Lambda(0, \zeta_k^\star)\right)^{T-S} \exp\left(-\zeta^\star (B_T - B_S)\right).$$

As in the proof of Property 3.4.2, we have $\Lambda(\theta_k^\star, \eta_k^\star) = \Lambda(0, \zeta^\star) = 1$, so that we can simplify $L_k(u)$ to

$$L_k(u) = \exp\left(-\theta_k^\star A_S - \eta_k^\star B_S\right) \cdot \exp\left(-\zeta^\star (B_T - B_S)\right).$$

Now note that, on the event that $I_k(u) = 1$, it holds that (i) $A_S \ge u$, (ii) $B_S \in s_k$, and (iii) $B_T - B_S \ge (1 - (k+1)\, f(u)/u) \cdot u \ge 0$. Therefore, uniformly in u, using the definition of $J_k(u)$ and $\Lambda(\theta_k^\star, \eta_k^\star) = \Lambda(0, \zeta^\star) = 1$,

$$\frac{1}{u} \ln L_k(u) I_k(u) \le -\theta_k^\star - \min\left\{ \eta_k^\star\, k \frac{f(u)}{u}, \eta_k^\star\, (k+1) \frac{f(u)}{u} \right\}$$

$$
\begin{aligned}
&- \zeta^\star \left(1 - k\frac{f(u)}{u}\right)^+ + \zeta^\star \frac{f(u)}{u} \\
&\leq -\theta_k^\star - \eta_k^\star\, k\frac{f(u)}{u} - \zeta^\star \left(1 - k\frac{f(u)}{u}\right)^+ + \zeta^\star \frac{f(u)}{u} \\
&= -J_k(u) + \zeta^\star \frac{f(u)}{u};
\end{aligned}
$$

the final equality in the above display can be seen by observing that

$$
\alpha_2 = \inf_{z>0} \frac{I_i(z)}{z} = \inf_{z>0} \frac{\sup_\zeta \zeta z - \ln \Lambda_2(\zeta)}{z} = \frac{\zeta^\star z^\star - \ln \Lambda_2(\zeta^\star)}{z^\star} = \frac{\zeta^\star z^\star - 0}{z^\star} = \zeta^\star
$$

and similarly for the first term.

We estimate our probability by evaluating sample averages of independent random variables that are distributed as

$$
Z_{m,M}(u) := \sum_{k=m}^{M} L_k(u) I_k(u),
$$

with m, M a suitably chosen truncation. Note that $Z_{m,M}(u)$ is bounded from above by

$$
Z_{m,M}(u) \leq (M - m + 1) \exp\left(- \min_{k \in \{m,\ldots,M\}} J_k(u)u + \zeta^\star f(u) \right).
$$

We thus see that, using that $M - m$ grows subexponentially and using that $f(u)$ is sublinear,

$$
\limsup_{u\to\infty} \frac{1}{u} \ln \mathbb{E}\left((Z_{m,M}(u))^2 \right) \leq -2J^{(1)} = 2 \lim_{u\to\infty} \frac{1}{u} \ln \pi_1(u),
$$

so the procedure is asymptotically optimal. □

3.6 Numerical results

In this section, we show and discuss several numerical experiments. Throughout this section, we use a specific instance of the general model as described in

Section 3.2, which, in particular, satisfies the conditions of Theorem 3.3.1.

3.6.1 The model

In the examples considered we assume that the (X_i, Y_i) are i.i.d. vectors with a bivariate normal distribution with mean vector and covariance matrix given by

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} (\sigma_1)^2 & \rho \\ \rho & (\sigma_2)^2 \end{pmatrix},$$

respectively. One of the reasons that we chose this model is the following nice property.

Property 3.6.1. *If we exponentially twist (X_i, Y_i) with parameter (θ_1, θ_2), the twisted process again has a bivariate normal distribution with mean vector*

$$\tilde{\mu} = \begin{pmatrix} \mu_1 + \theta_1(\sigma_1)^2 + \theta_2\rho \\ \mu_2 + \theta_1\rho + \theta_2(\sigma_2)^2 \end{pmatrix}$$

and covariance matrix $\tilde{\Sigma} = \Sigma$.

Proof. This follows from elementary calculations. □

Below we perform a number of different experiments, in which we test the influence of various parameters on the performance of the simulations. The simulations were carried out in `R`. Anytime in this section we refer to (non-partitioned) importance sampling (IS), *naive* importance sampling is meant. Unless otherwise stated, in all numerical experiments, we ran simulations until a 95% confidence interval with 10% precision was obtained. We tested both the number of runs and the running time (CPU time) required to obtain the confidence interval. In most cases, the CPU time shows the same quantitative behaviour as the number of runs. Therefore, the CPU times are only shown in Section 3.6.4, where this is not the case.

Remark. As already stated before, the truncation used in partitioned importance sampling inherently produces a biased estimator. It is, however, possible to obtain an estimator with *vanishing relative bias* as $u \to \infty$. This can be accomplished by choosing the lower bound $m(u)f(u)$ and upper bound $M(u)f(u)$

such that the expected value of the slower process, at the moment the faster process hits level u, is in between these bounds. This expected value can be numerically determined by combining Theorem 3.3.1 and Property 3.4.1. We indeed chose m, M and f such that in the experiments in Figures 3.1, 3.2 and 3.3 we have vanishing relative bias.

Remark. There exist more methods of sampling than naive and partitioned importance sampling. One method makes use of a so-called *subsolution method*, see e.g. [34]. We implemented such a state-dependent importance sampling scheme, in the following way. In Example 3 of [10, pp. 47–48] it is explained how to set up a subsolution-based scheme for estimating the probability that *at least one of the components* reaches a rare set. The procedure is then:

- In each simulation run, we used this scheme until one of the components exceeds level u;
- subsequently, we use a single exponential twist for the remainder of the run (with the relevant α_i, as given in (3.3)), until the other component has exceeded level u as well.

We compared this method with the partitioned importance sampling method proposed in this chapter. We tested both procedures extensively, both in terms of the number of runs and the CPU time. In the simulations we performed, we observed that the partitioned IS method performs better. We suspect that this may be due to the same problem as in the naïve IS scheme, namely the random fluctuations of the 'second component' (for instance, the fluctuations of the vertical component at the epoch that the horizontal component first exceeds u) — the way we set up the subsolution-based scheme the fluctuations of the second component are apparently not sufficiently controlled (as in the first part of the run the focus is only on the event that one of the two components exceeds u).

Clearly, the partitioned importance sampling will become less attractive when considering problems of higher dimensions. Considering the counterpart of our problem but then in dimensions higher than 2, one could again come up with a partitioning such that $Z_k(u)$ can be written as sum (over all k) of $L_k(u)I_k(u)$, but the number of k to be included will increase (which will slow down the simulation). It is therefore anticipated that in higher dimensions subsolution-based schemes will become advantageous.

3.6.2 Variable level

In this section we look at how the level to reach u influences both the probability $\pi(u)$ and the number of simulations needed. We ran two different experiments of which the results can be found in Figures 3.1 and 3.2. The experiments differ in the sign of the covariance that was used; we refer to the respective caption for specific details. The results below clearly show that Monte Carlo sampling is much slower than both importance sampling and partitioned importance sampling. Furthermore, when comparing Figure 3.1 and 3.2 we see that a negative correlation between the two components negatively influences both the probability $\pi(u)$ and the number of samples needed to get the desired precision.

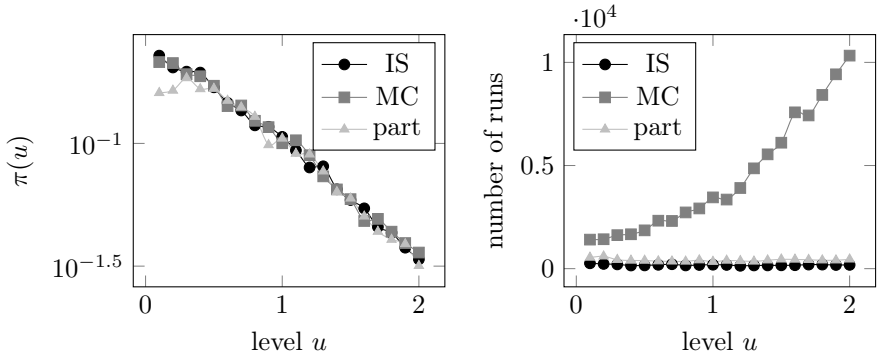

Figure 3.1: These plots show the probability $\pi(u)$ and the number of runs needed to get the desired accuracy respectively. The parameters that were used were $\mu = \binom{-1}{-0.5}$, $\Sigma = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, $m(u) = -40u$, $M(u) = 20u - 1$ and $f(u) = 0.05$. The results show that importance sampling gives a significant efficiency improvement over Monte Carlo sampling.

It should be noted that the event of interest in the experiments as described shown in Figures 3.1 and 3.2 above can hardly be called "rare". The reason that we kept the level to reach (u) relatively low is that Monte Carlo sampling quickly took over 10000 runs, which takes a long time in `R`. Therefore, in Figure 3.3 we performed some simulations for larger u, though only for importance sampling and partitioned importance sampling. The results clearly show that even for extremely small probabilities (around 10^{-52}), both importance sampling and

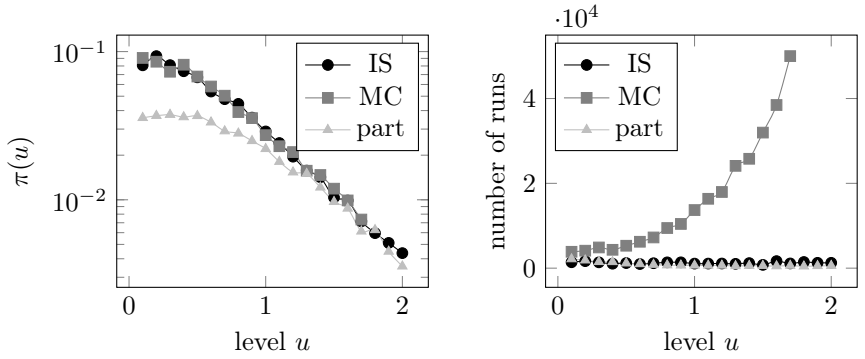

Figure 3.2: These plots show the probability $\pi(u)$ and the number of runs needed to get the desired accuracy respectively. We stopped simulations when 50 000 runs were needed, hence the missing values for Monte Carlo sampling. The parameters that were used were $\mu = \begin{pmatrix} -1 \\ -0.5 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$, $m(u) = -40u$, $M(u) = 20u - 1$ and $f(u) = 0.05$. The results show that importance sampling gives a significant efficiency improvement over Monte Carlo sampling.

partitioned importance sampling need a modest amount of runs.

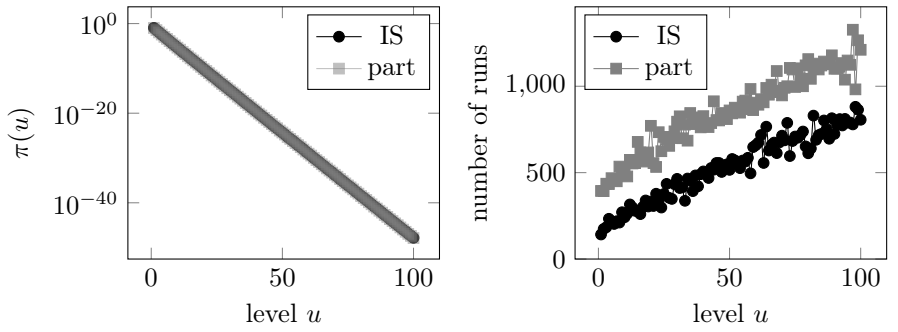

Figure 3.3: These plots show the probability $\pi(u)$ and the number of runs needed to get the desired accuracy respectively. The parameters that were used were $\mu = \begin{pmatrix} -1 \\ -0.5 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, $m(u) = -10$, $M(u) = u - 1$ and $f(u) = 1$. The results show that even for extremely small probabilities, the number of runs needed for IS and partitioned IS is modest.

3.6.3 Variable covariance

Having convinced ourselves that Monte Carlo sampling is prohibitively slow, we will restrict the experiments now to only importance sampling and partitioned importance sampling. In the previous experiments, both IS and partitioned IS performed roughly the same: the number of runs required to get the desired confidence interval did not show any significant differences. We will now identify cases where partitioned IS behaves much better than IS. In the next simulations we look at how the covariance influences both $\pi(u)$ and the number of trials needed. Figures 3.4 and 3.5 give the results of two experiments; the difference is the level u that has to be reached (5 and 2 respectively). The results indicate that a negative covariance slows down both methods, but also show that partitioned importance sampling is faster. When the correlation becomes positive, the opposite seems to hold.

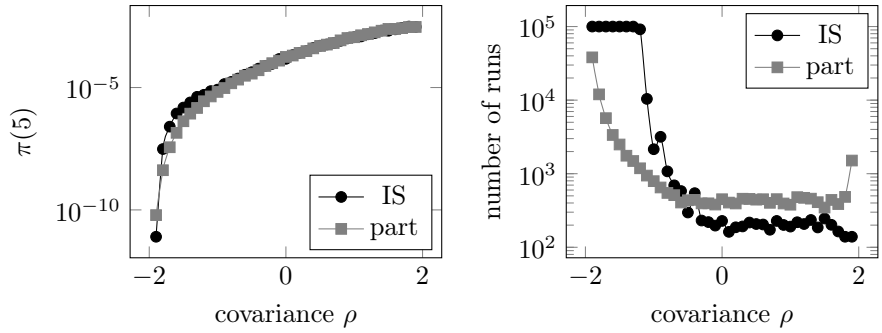

Figure 3.4: These plots show the probability $\pi(5)$ and the number of runs (with a maximum of 100 000) needed to get the desired accuracy respectively. The parameters that were used were $\mu = \binom{-1}{-0.5}$ and $\Sigma = \left(\begin{smallmatrix} 2 & \rho \\ \rho & 2 \end{smallmatrix}\right)$. We fixed the number of intervals to 10, the lower bound of those intervals to -5 and the upper bound to 5. The results show that, when the two components are strongly negatively correlated, partitioned importance sampling behaves much better than ordinary importance sampling. When the components are positively correlated, the opposite holds.

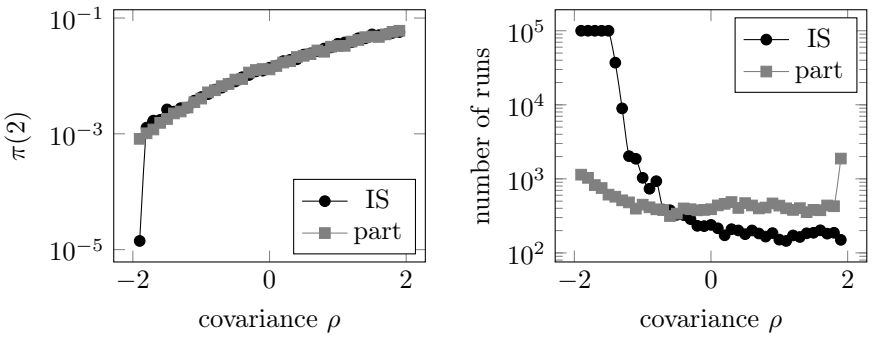

Figure 3.5: The precise same experiment as reported on in Figure 3.4, except that $u = 2$ now and that the upper bound for the intervals also equals 2.

3.6.4 Variable number of intervals

In this part we restrict ourselves to partitioned importance sampling only. The goal is to find how the number of intervals affects the performance. From Figure 3.6 we can conclude that the number of intervals does not seem to have an effect on the number of trials needed. The total running time, however, does seem to suffer from a high number of intervals, though only linearly.

3.7 Proof of Theorem 3.3.1

This section presents the proof of Theorem 3.3.1. The right-hand side of (3.4) will be proved first as a lower bound and then as an upper bound for the left-hand side of (3.4).

Lower bound First observe that $\pi(u) \geq \mathbb{P}\left(A_{su} > u, B_{tu} > u\right)$ for all s, t (where we allow ourselves, here and elsewhere, the imprecise notation su and tu when we mean their respective rounded-off values). It thus follows, for all s and t,

$$\liminf_{u\to\infty} \frac{1}{u} \ln \pi(u) \geq \liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\mathscr{A}_{su}, \mathscr{B}_{tu}\right).$$

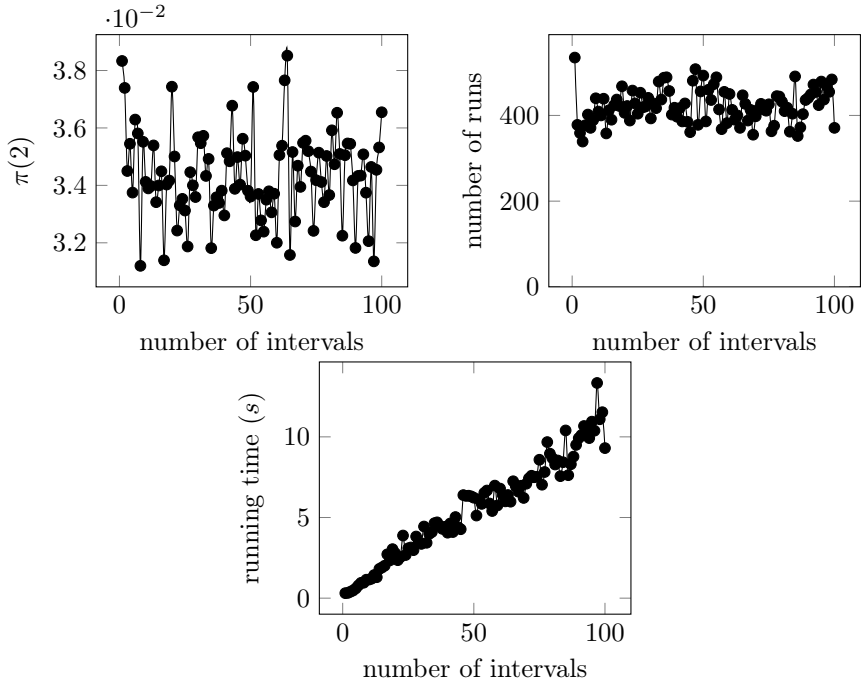

Figure 3.6: These plots show the probability $\pi(2)$, the number of runs (with a maximum of 100 000) needed to get the desired accuracy, and the running time, respectively. The parameters that were used were $\mu = \begin{pmatrix} -1 \\ -0.5 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. The lower and upper bound for the intervals were -5 and 2 respectively. The results show that the number of intervals doesn't seem to have any impact on the number of runs needed. The running time, however, seems to increase linearly as a function of the number of intervals.

Since this holds for all s and t, we take the supremum over all s and t:

$$\liminf_{u\to\infty} \frac{1}{u} \ln \pi(u) \geq \sup_{s,t} \liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\mathscr{A}_{su}, \mathscr{B}_{tu}\right).$$

Suppose, without loss of generality, that $s < t$ and let $\Delta > 0$. Then for all y, using the independence,

$$\begin{aligned}
\mathbb{P}\left(\mathscr{A}_{su}, \mathscr{B}_{tu}\right) &\geq \mathbb{P}\left(\frac{A_{su}}{su} > \frac{1}{s}, \frac{B_{su}}{su} \in [y, y+\Delta], \frac{B_{tu}}{tu} > \frac{1}{t}\right) \\
&\geq \mathbb{P}\left(\frac{A_{su}}{su} > \frac{1}{s}, \frac{B_{su}}{su} \in [y, y+\Delta], B_{tu} - B_{su} > u - suy\right) \\
&= \mathbb{P}\left(\frac{A_{su}}{su} > \frac{1}{s}, \frac{B_{su}}{su} \in [y, y+\Delta]\right) \cdot \mathbb{P}\left(B_{tu} - B_{su} > u - suy\right) \\
&= \mathbb{P}\left(\frac{A_{su}}{su} > \frac{1}{s}, \frac{B_{su}}{su} \in [y, y+\Delta]\right) \cdot \mathbb{P}\left(\frac{B_{(t-s)u}}{(t-s)u} > \frac{1-sy}{t-s}\right).
\end{aligned}$$

By Cramér's theorem, we have that the decay rate of this expression equals

$$-s \inf_{p > \frac{1}{s}, q \in [y, y+\Delta]} I(p,q) - (t-s) \inf_{q > \frac{1-sy}{t-s}} I_2(q).$$

As this relation holds for any Δ, using the continuity of I, we thus find that

$$\liminf_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\mathscr{A}_{su}, \mathscr{B}_{tu}\right) \geq -s \inf_{p > \frac{1}{s}} I(p,y) - (t-s) \inf_{q > \frac{1-sy}{t-s}} I_2(q).$$

This relation holds for any y, so in particular for all $y < 1/s$; as $I_2(q)$ increases in q for q positive, we obtain

$$\inf_{q > \frac{1-sy}{t-s}} I_2(q) = I_2\left(\frac{1-sy}{t-s}\right),$$

and as a consequence

$$\begin{aligned}&\sup_{0<s<t}\liminf_{u\to\infty}\frac{1}{u}\ln\mathbb{P}\left(\mathscr{A}_{su},\mathscr{B}_{tu}\right)\geq\\&-\inf_{0<s<t,y<1/s}\left(s\inf_{p>\frac{1}{s}}I(p,y)+(t-s)I_2\left(\frac{1-sy}{t-s}\right)\right).\end{aligned}\tag{3.10}$$

Now we put $v:=t-s$, so as to obtain that the right-hand side of the previous display equals

$$\begin{aligned}-\inf_{s>0,y<1/s}s\left(\inf_{p>1/s}I(p,y)+(1-sy)\inf_{v>0}\frac{v}{1-sy}I_2\left(\frac{1-sy}{v}\right)\right)&\\=-\inf_{s>0,y<1/s}\left(s\inf_{p>\frac{1}{s}}I(p,y)+(1-sy)\alpha_2\right)&\\=-\inf_{x>0,y<x}\left(\inf_{p>x}\frac{I(p,y)}{x}+\left(1-\frac{y}{x}\right)\alpha_2\right)&\\\geq-\inf_{x>0,y<x}\left(\frac{I(x,y)}{x}+\left(1-\frac{y}{x}\right)\alpha_2\right).&\end{aligned}$$

Upper bound Now turn to the upper bound. We split the event in multiple sub-events. Fix some $s^\star$ and $t^\star$. Then the union bound implies that, for any $a>0$,

$$\begin{aligned}\pi(u)&\leq\mathbb{P}\left(\exists s<(1+a)s^\star u:\mathscr{A}_s,\exists t<(1+a)t^\star u:\mathscr{B}_t\right)\\&\quad+\mathbb{P}\left(\exists s\geq(1+a)s^\star u:\mathscr{A}_s,\exists t\in\mathbb{N}:\mathscr{B}_t\right)\\&\quad+\mathbb{P}\left(\exists s\in\mathbb{N}:\mathscr{A}_s,\exists t\geq(1+a)t^\star u:\mathscr{B}_t\right)\\&\leq\mathbb{P}\left(\exists s<(1+a)s^\star u:\mathscr{A}_s,\exists t<(1+a)t^\star u:\mathscr{B}_t\right)\\&\quad+\mathbb{P}\left(\exists s\geq(1+a)s^\star u:\mathscr{A}_s\right)+\mathbb{P}\left(\exists t\geq(1+a)t^\star u:\mathscr{B}_t\right).\end{aligned}$$

We now show that the latter two terms have a higher decay rate (i.e., decay *faster*) than the first term. First we use the union bound to get

$$\mathbb{P}\left(\exists s\geq(1+a)s^\star u:\mathscr{A}_s\right)\leq\sum_{s=(1+a)s^\star u}^{\infty}\mathbb{P}\left(A_s>u\right).$$

We now focus on the individual terms of the right-hand side. Using Markov's inequality, we get for all s:

$$\mathbb{P}\left(A_s > u\right) \leq \mathbb{E}\left(e^{\theta A_s}\right) e^{-\theta u}.$$

Note that, since the process is a random walk,

$$\mathbb{E}\left(e^{\theta A_s}\right) = \left(\mathbb{E}\left(e^{\theta X}\right)\right)^s = \Lambda_1(\theta)^s = e^{s \ln \Lambda_1(\theta)}.$$

Since this holds for all θ, we take the infimum:

$$\mathbb{P}\left(A_s > u\right) \leq \inf_{\theta} e^{s \ln \Lambda_1(\theta)} e^{-\theta u} = e^{-\sup_\theta(\theta u - s \ln \Lambda_1(\theta))} = e^{-sI_1(u/s)} \leq e^{-sI_1(0)};$$

in the second inequality we have used that $\mathbb{E}\left(X_1\right) < 0$. Now we return to the sum again. Using the display above, we get

$$\mathbb{P}\left(\exists s \geq (1+a)s^\star u : \mathscr{A}_s\right) \leq \sum_{s=(1+a)s^\star u}^{\infty} e^{-sI_1(0)} = \frac{e^{-I_1(0)\cdot(1+a)s^\star u}}{1 - e^{-I_1(0)}},$$

so that for the decay rate of the probability above we get

$$\limsup_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\exists s \geq (1+a)s^\star u : \mathscr{A}_s\right) \leq -(1+a)s^\star I_1(0).$$

We conclude that the decay rate can be made arbitrarily large by letting $a \to \infty$. Obviously, the same procedure can be followed for $\mathbb{P}\left(\exists t \geq (1+a)t^\star u : \mathscr{B}_t\right)$. It thus follows that the first term has the lowest decay rate, and therefore the Principle of the Largest Term (Lemma 1.3.4) gives

$$\limsup_{u\to\infty} \frac{1}{u} \ln \pi(u) \leq \lim_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\exists s < (1+a)s^\star u : \mathscr{A}_s, \exists t < (1+a)t^\star u : \mathscr{B}_t\right). \tag{3.11}$$

Define $T := \max\{s^\star, t^\star\}$ and $\alpha := (1+a)T$. We introduce the scaled processes $\bar{A}_u(s) := \frac{1}{\alpha u} A_{\alpha u s}$ and $\bar{B}_u(t) := \frac{1}{\alpha u} B_{\alpha u t}$. The probability on the right-hand side above is then smaller than or equal to

$$\mathbb{P}\left(\sup_{s\leq 1} \bar{A}_u(s) \geq 1/\alpha, \sup_{t\leq 1} \bar{B}_u(t) \geq 1/\alpha\right).$$

Let f be the sample path of a 2 dimensional function. Also, let $\phi_1(f) := \sup_{s\leq 1} f_1(s)$ and likewise let $\phi_2(f) := \sup_{t\leq 1} f_2(t)$. We now invoke Mogulskii's theorem (Theorem 1.3.3). This gives us

$$\begin{aligned}
&- \inf_{\{f:\phi_1(f)\geq 1/\alpha, \phi_2(f)\geq 1/\alpha\}^o} J(f) \\
&\leq \liminf_{u\to\infty} \frac{1}{\alpha u} \ln \mathbb{P}\left(\sup_{s\leq 1} \bar{A}_u(s) \geq 1/\alpha, \sup_{t\leq 1} \bar{B}_u(t) \geq 1/\alpha\right) \\
&\leq \limsup_{u\to\infty} \frac{1}{\alpha u} \ln \mathbb{P}\left(\sup_{s\leq 1} \bar{A}_u(s) \geq 1/\alpha, \sup_{t\leq 1} \bar{B}_u(t) \geq 1/\alpha\right) \\
&\leq - \inf_{\{f:\phi_1(f)\geq 1/\alpha, \phi_2(f)\geq 1/\alpha\}^c} J(f),
\end{aligned}$$

with

$$J(f) = \int_0^1 I(f'(t))\,\mathrm{d}t;$$

note that the conditions imposed allow that Mogulskii's theorem can be applied. Note that the set over which the infimum is taken is closed, hence we will drop the closure operator. So an upper bound for the right-hand side of (3.11) is

$$-\alpha \inf_{\{f:\phi_1(f)\geq 1/\alpha, \phi_2(f)\geq 1/\alpha\}} J(f). \tag{3.12}$$

In the calculations below, we will drop the factor α in front of the infimum; we will later see that it cancels.

Assume that $\bar{A}_u(\cdot)$ hits u for the first time at time $\bar{s}$ and $\bar{B}_u(\cdot)$ hits u for the first time at time $\bar{t}$, i.e., $\bar{s} \equiv \bar{s}(f) = \inf_{s\in[0,1]}\{s : f_1(s) \geq 1/\alpha\}$ and likewise for $\bar{t}$. We can then rewrite the upper bound in two cases:

$$\begin{aligned}
&\inf_{\{f:\phi_1(f)\geq 1/\alpha, \phi_2(f)\geq 1/\alpha\}} J(f) \\
&= \inf_{\{f:\bar{s}(f)\leq 1, \bar{t}(f)\leq 1\}} J(f) \\
&= \min\left\{ \inf_{\{f:\bar{s}(f)\leq 1, \bar{t}(f)\leq 1, \bar{s}\leq\bar{t}\}} J(f), \inf_{\{f:\bar{s}(f)\leq 1, \bar{t}(f)\leq 1, \bar{s}>\bar{t}\}} J(f) \right\}.
\end{aligned}$$

We will now focus on the first entry of the minimum above; the second entry

can be treated analogously. It can be rewritten as

$$\inf_{v\le 1/\alpha}\ \inf_{\{f:\bar{s}(f)\le 1,\bar{t}(f)\le 1,\bar{s}\le\bar{t},f_2(\bar{s})=v\}} J(f).$$

Pick a fixed but arbitrary f which confines to the restrictions in the infima above. We will now rewrite $J(f)$ as a sum of three integrals:

$$J(f)=\int_0^1 I(f'(t))(f'(t))\,\mathrm{d}t = \int_0^{\bar{s}} I(f'(t))\,\mathrm{d}t+\int_{\bar{s}}^{\bar{t}} I(f'(t))\,\mathrm{d}t+\int_{\bar{t}}^1 I(f'(t))(f'(t))\,\mathrm{d}t.$$

In the spirit of [39], we will construct a straightened path $\tilde{f}$ and then show that the upper bound in the LDP is the same as $-J(\tilde{f})$. Let

$$(\tilde{f}_1'(\tau),\tilde{f}_2'(\tau))=\begin{cases}(\frac{1}{\alpha\bar{s}},\frac{v}{\bar{s}}) & \text{if } 0\le\tau\le\bar{s};\\ (\frac{c^\star}{\bar{t}-\bar{s}},\frac{1/\alpha-v}{\bar{t}-\bar{s}}) & \text{if } \bar{s}<\tau\le\bar{t};\\ (\mu,\nu) & \text{if } \bar{t}<\tau\le 1,\end{cases}$$

where $c^\star := \arg\min_c I(\frac{c}{\bar{t}-\bar{s}},\frac{1/\alpha-v}{\bar{t}-\bar{s}})$, $\mu := \mathbb{E}\,(X_1)$ and $\nu := \mathbb{E}\,(Y_1)$. Now note that

1. using Jensen's inequality,

$$\int_0^{\bar{s}} I(\tilde{f}'(t))\,\mathrm{d}t=\bar{s}I\left(\frac{1}{\alpha\bar{s}},\frac{v}{\bar{s}}\right)=\bar{s}I\left(\frac{1}{\bar{s}}\int_0^{\bar{s}} f'(t)\,\mathrm{d}t\right)\le\int_0^{\bar{s}} I(f'(t))\,\mathrm{d}t;$$

2. using the minimisation gives us

$$\begin{aligned}\int_{\bar{s}}^{\bar{t}} I(\tilde{f}'(t))\,\mathrm{d}t&=(\bar{t}-\bar{s})I\left(\frac{c^\star}{\bar{t}-\bar{s}},\frac{1/\alpha-v}{\bar{t}-\bar{s}}\right)\\ &\le(\bar{t}-\bar{s})I\left(\frac{1}{\bar{t}-\bar{s}}\int_{\bar{s}}^{\bar{t}} f'(t)\,\mathrm{d}t\right)\\ &\le\int_{\bar{s}}^{\bar{t}} I(f'(t))\,\mathrm{d}t;\end{aligned}$$

3. using [39, Lemma 2.6.iv] gives us

$$\int_{\bar{t}}^{1} I(\tilde{f}'(t))\,\mathrm{d}t = 0 \leq \int_{\bar{t}}^{1} I(f'(t))\,\mathrm{d}t,$$

hence, $J(f) \geq J(\tilde{f})$. Furthermore,

$$\inf_{v\leq 1/\alpha} \inf_{\{f:\bar{s}(f)\leq 1,\bar{t}(f)\leq 1,\bar{s}\leq\bar{t},f_2(\bar{s})=v\}} J(f) \leq \inf_{v\leq 1/\alpha} \inf_{\{\tilde{f}:\bar{s}(\tilde{f})\leq 1,\bar{t}(\tilde{f})\leq 1,\bar{s}\leq\bar{t},\tilde{f}_2(\bar{s})=v\}} J(\tilde{f}),$$

where on the right-hand side we restrict the infimum to straightened paths as described above. Hence, this inequality is in fact an equality.

We will now focus on $J(\tilde{f})$. Note that it is equal to

$$\bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + (\bar{t}-\bar{s})I\left(\frac{c^\star}{\bar{t}-\bar{s}}, \frac{1/\alpha - v}{\bar{t}-\bar{s}}\right).$$

When we now bring back the infimum we get

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1} \bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + (\bar{t}-\bar{s})I\left(\frac{c^\star}{\bar{t}-\bar{s}}, \frac{1/\alpha - v}{\bar{t}-\bar{s}}\right),$$

or, using the definition of $c^\star$

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1} \bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + (\bar{t}-\bar{s}) \inf_{c} I\left(\frac{c}{\bar{t}-\bar{s}}, \frac{1/\alpha - v}{\bar{t}-\bar{s}}\right).$$

This is larger than or equal to

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1} \bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + (\bar{t}-\bar{s}) \inf_{c\in\mathbb{R}, z>0} \frac{I\left(\frac{c}{\bar{t}-\bar{s}}, z\right)}{z\frac{\bar{t}-\bar{s}}{1/\alpha - v}},$$

and cancelling the factors gives us

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1} \bar{s}I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + \left(\frac{1}{\alpha} - v\right) \inf_{c\in\mathbb{R}, z>0} \frac{I\left(\frac{c}{\bar{t}-\bar{s}}, z\right)}{z}.$$

This is again larger than or equal to

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1} \bar{s} I\left(\frac{1}{\alpha\bar{s}}, \frac{v}{\bar{s}}\right) + \left(\frac{1}{\alpha} - v\right) \inf_{z>0} \frac{I_2(z)}{z}.$$

Now define $x := 1/\alpha\bar{s}$ and $y := \alpha v x$. Then the expression above is equal to

$$\inf_{v\leq 1/\alpha, 0\leq\bar{s}\leq\bar{t}\leq 1, x=1/\alpha\bar{s}, y=\alpha v x} \frac{1}{\alpha x} I(x,y) + \left(\frac{1}{\alpha} - \frac{y}{\alpha x}\right) \inf_{z>0} \frac{I_2(z)}{z},$$

which equals

$$\frac{1}{\alpha}\left(\inf_{x\geq 1/\alpha, y\leq x} \frac{1}{x} I(x,y) + (1-\frac{y}{x}) \inf_{z>0} \frac{I_2(z)}{z}\right),$$

which majorises

$$\frac{1}{\alpha}\left(\inf_{x>0, y\leq x} \frac{1}{x} I(x,y) + (1-\frac{y}{x}) \inf_{z>0} \frac{I_2(z)}{z}\right).$$

Recall that the factor $\frac{1}{\alpha}$ cancels against the factor α of (3.12), hence we have proven the upper bound.

3.8 Concluding remarks

In this chapter, we studied both logarithmic asymptotics and several numerical methods to study large delay probabilities of two correlated queues. In the first part of the chapter, the first main result, Theorem 3.3.1, was given, which provided an expression for the decay rate of the probability of both components ever reaching some high level. The second part consisted of analysing two numerical procedures, namely a naive importance sampling procedure and partitioned importance sampling. It was shown that the former method is not necessarily asymptotically optimal. This is caused by the undershoot of the slowest component. This problem is overcome by the second method, partitioned importance sampling. It was indeed shown that this procedure is asymptotically optimal. Subsequently, numerical results of simulation experiments were shown, confirming the theory.

4

Linear stochastic fluid networks with Markov modulation

In the previous chapters, the focus was on the probability of ever hitting a rare set. The current chapter, in contrast, treats the event of a multidimensional stochastic process attaining a value in a rare set at a *fixed* time. This chapter is based on [18].

4.1 Introduction

Linear stochastic fluid networks, as introduced in [51], can be informally described as follows. Consider a network consisting of L stations. Jobs, whose sizes are i.i.d. samples from some general L-dimensional distribution, arrive at the stations according to a Poisson process. At each of the nodes, in between arrivals the storage level decreases exponentially. Processed traffic is either transferred to the other nodes or leaves the network (according to a given routing matrix). In addition to this basic version of the linear stochastic fluid network, there is also its *Markov modulated* counterpart [49], in which the arrival rate, the distribution of the job sizes, and the routing matrix depend on the state of an external, autonomously evolving finite-state continuous-time Markov chain (usually referred to as the *background process*).

Linear stochastic fluid networks can be seen as natural fluid counterparts of

corresponding infinite-server queues. As such, they inherit several nice properties of those infinite-server queues. In particular, separate infinitesimally small fluid particles, moving through the network, do not interfere, and are therefore mutually independent. Essentially due to this property, linear stochastic fluid networks allow explicit analysis; in particular, the joint Laplace transform of the storage levels at a given point in time can be expressed in closed form as a function of the arrival rate, the Laplace transform of the job sizes and the routing matrix [51, Thm. 5.1].

When Markov modulation is imposed, the analysis becomes substantially harder. Conditional on the path of the background process, again explicit expressions can be derived, cf. [49, Thm. 1]. Unconditioning, however, cannot be done in a straightforward manner. As a consequence the results found are substantially less explicit than for the non-modulated linear stochastic fluid network. In [49] also a system of ordinary differential equations has been set up that provides the transform of the stationary storage level; in addition, conditions are identified that guarantee the existence of such a stationary distribution.

In this chapter we focus on rare events for Markov-modulated linear stochastic fluid networks. More specifically, in a particular scaling regime (parameterised by n) we analyse the probability p_n that at a given point in time the network storage vector is in a given rare set. By scaling the arrival rate as well as the rare set (which amounts to multiplying them by a scaling parameter n), the event of interest becomes increasingly rare. More specifically, under a Cramér-type assumption on the job-size distribution, application of large-deviations theory yields that p_n decays (roughly) exponentially. As p_n can be characterised only asymptotically, one could consider the option of using simulation to obtain precise estimates. The effectiveness, however, of such an approach is limited due to the rarity of the event under consideration: in order to get a reliable estimate, one needs sufficiently many runs in which the event occurs. This is the reason why one often resorts to simulation using *importance sampling* (or: *change of measure*). This is a variance reduction technique in which one replaces the actual probability measure by an alternative measure under which the event under consideration is *not* rare; correcting the simulation output with appropriate likelihood ratios yields an unbiased estimate.

The crucial issue when setting up an importance sampling procedure con-

cerns the choice of the alternative measure: one would like to select one that provides a substantial variance reduction, or is even (in some sense) optimal. The objective of this chapter is to develop a change of measure which performs provably optimally.

Our ultimate goal is to obtain an efficient simulation procedure for Markov-modulated linear stochastic fluid networks. We do so by (i) first considering a single node without modulation, (ii) then multi-node systems, still without modulation, and (iii) finally modulated multi-node systems. There are two reasons for this step-by-step setup:

- For the non-modulated models we have more refined results than for the modulated models. More specifically, for the non-modulated models we have developed estimates for the number of runs Σ_n required to obtain an estimate with predefined precision (showing that Σ_n grows polynomially in the rarity parameter n), whereas for modulated models we can just prove that Σ_n grows subexponentially.

- In addition, this approach allows the reader to get gradually familiar with the concepts used in this chapter.

The construction and analysis of our importance sampling methodology is based on the ideas developed in [15]; there the focus was on addressing similar issues for a single-node Markov modulated infinite-server system. In line with [15], we consider the regime in which the background process is 'slow': while we (linearly) speed up the driving Poisson process, we leave the rates of the Markovian background process unaltered.

A traditional, thoroughly examined, importance sampling problem concerns the sample mean S_n of n i.i.d. light-tailed random variables $X_1, \ldots, X_n$; the objective there is to estimate $\mathbb{P}(S_n \geqslant a)$ for $a > \mathbb{E}X_1$ and n large. As described in [3, Section VI.2], in this situation importance sampling (i.e., sampling under an alternative measure, and translating the simulation output back by applying appropriate likelihood ratios) works extremely well. To this end, the distribution of the X_i s should be *exponentially twisted*. As it turns out, in our setup, the probability of our interest can be cast in terms of this problem. Compared to the standard setup of sample means of one-dimensional random variables, however, there are a few complications: (i) in our case it is not a priori clear

how to sample from the exponentially twisted distributions, (ii) we consider multi-dimensional distributions (i.e., rare-event probabilities that concern the storage levels of all individual buffers in the network), (iii) we impose Markov modulation. We refer to e.g. [40, 56] for earlier work on similar problems.

We refer to [18, Section 4.1] for additional results on how to set up a recursion to evaluate the (transient and stationary) moments of the joint storage level in Markov-modulated linear stochastic fluid networks (where the results in [49] are restricted to just the first two stationary moments at epochs that the background process jumps).

The single-node model without modulation falls in the class of (one-dimensional) *shot-noise* models, for which efficient rare-event simulation techniques have been developed over the past, say, two decades. Asmussen and Nielsen [5] and Ganesh *et al.* [38] consider the probability that a shot-noise process decreased by a linear drift ever exceeds some given level. Relying on sample-path large deviations results, an asymptotically efficient importance sampling algorithm is developed, under the same scaling as the one we consider in our paper. The major difference with our model (apart from the fact that we deal with considerably more general models, as we focus on networks and allow modulation) is that we focus on a rare-event probability that relates to the position of the process at a fixed point in time; in this setting we succeed in finding accurate estimates of the number of runs needed to get an estimate of given precision.

There is a vast body of literature related to the broader area of rare-event simulation for queueing systems. We refer to the literature overviews [12, 47]; interesting recent papers include [4, 21, 68].

This chapter is organised as follows. In Section 4.2 the focus is on a single-node network, without Markov modulation (addressing complication (i) above), Section 4.3 addresses the extension to multi-node systems (addressing complication (ii)), and in Section 4.4 the feature of modulation is added (addressing complication (iii)). In each of these three sections, we propose a change of measure, quantify its performance, and demonstrate its efficiency through simulation experiments. A discussion and concluding remarks are found in Section 4.5.

4.2 Single resource, no modulation

To introduce the concepts we work with in this chapter, we analyse in this section a linear stochastic fluid network consisting of a single node, in which the input is just compound Poisson (so no Markov modulation is imposed). More precisely, in the model considered, jobs arrive according to a Poisson process with rate λ, bring along i.i.d. amounts of work (represented by the sequence of i.i.d. random variables $(B_1, B_2, \ldots)$), and the workload level decays exponentially at a rate $r > 0$. This model belongs to the class of *shot-noise processes*. As mentioned in the introduction, we gradually extend the model in the next sections.

4.2.1 Preliminaries

We first present a compact representation for the amount of work in the system at time t, which we denote by $X(t)$, through its moment generating function. To this end, let $N(t)$ denote a Poisson random variable with mean λt, and $(U_1, U_2, \ldots)$ i.i.d. uniformly distributed random variables (on the interval $[0, t]$). Assume in addition that the random objects $(B_1, B_2, \ldots)$, $N(t)$, and $(U_1, U_2, \ldots)$ are independent. Then it is well-known that the value of our shot-noise process at time t can be expressed as

$$X(t) = \sum_{j=1}^{N(t)} B_j \mathrm{e}^{-r(t-U_j)} \stackrel{\mathrm{d}}{=} \sum_{j=1}^{N(t)} B_j \mathrm{e}^{-rU_j}, \tag{4.1}$$

where the distributional equality is a consequence of the fact that the distribution of U is symmetric on the interval $[0, t]$. It is easy to compute the moment generating function (MGF) of $X(t)$, by conditioning on the value of $N(t)$:

$$\begin{aligned} M(\vartheta) := \mathbb{E}\,\mathrm{e}^{\vartheta X(t)} &= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \left(\mathbb{E}\exp(\vartheta B\,\mathrm{e}^{-rU})\right)^k \\ &= \exp\left(\lambda \int_0^t \left(\beta(\mathrm{e}^{-ru}\,\vartheta) - 1\right) \mathrm{d}u\right), \end{aligned} \tag{4.2}$$

where $\beta(\cdot)$ is the MGF corresponding to B (throughout assumed to exist). By differentiating and inserting $\vartheta = 0$, it follows immediately that

$$m(t) := \mathbb{E}\, X(t) = \frac{\lambda}{r}(1 - \mathrm{e}^{-rt})\, \mathbb{E}\, B.$$

Higher moments can be found by repeated differentiation. We note that, as t is held fixed throughout the current chapter, we often write N rather than $N(t)$.

4.2.2 Tail probabilities, change of measure

The next objective is to consider the asymptotics of the random variable $X(t)$ under a particular scaling. In this scaling we let the arrival rate be $n\lambda$ rather than just λ, for $n \in \mathbb{N}$. The value of the shot-noise process is now given by

$$Y_n(t) := \sum_{i=1}^{n} X_i(t),$$

with the vector $(X_1(t), \ldots, X_n(t))$ consisting of i.i.d. copies of the random variable $X(t)$ introduced above; here the infinite divisibility of a Compound Poisson distribution is used.

Our goal is to devise techniques to analyse the tail distribution of $Y_n(t)$. Standard theory now provides us with the asymptotics of

$$p_n(a) = \mathbb{P}(Y_n(t) \geqslant na)$$

for some $a > m(t)$; we are in the classical 'Cramér setting' [30, Section 2.2] if it is assumed that $M(\vartheta)$ is finite in a neighbourhood around the origin (which requires that the same property is satisfied by $\beta(\cdot)$). Let $I(a)$ and $\vartheta^\star \equiv \vartheta^\star(a)$, respectively, be defined as

$$I(a) := \sup_{\vartheta} \big(\vartheta a - \log M(\vartheta)\big), \quad \vartheta^\star := \arg\sup_{\vartheta} \big(\vartheta a - \log M(\vartheta)\big),$$

with $M(\cdot)$ as above. Using 'Cramér', we obtain that, under mild conditions,

$$\lim_{n\to\infty} \frac{1}{n} \log p_n(a) = -I(a) = -\vartheta^\star a + \log M(\vartheta^\star).$$

More refined asymptotics are available as well; we get back to this issue in Section 4.2.3.

As these results apply in the regime that n is large, a relevant issue concerns the development of efficient techniques to estimate $p_n(a)$ through simulation. An important rare-event simulation technique is importance sampling, relying on the commonly used exponential twisting technique. We now investigate how to construct the exponentially twisted version $\mathbb{Q}$ (with twist $\vartheta^\star$) of the original probability measure $\mathbb{P}$. The main idea is that under $\mathbb{Q}$ the $X_i(t)$ have mean a, such that under the new measure the event under study is not rare anymore.

More concretely, exponential twisting with parameter $\vartheta^\star$ means that under the new measure $\mathbb{Q}$, the $X_i(t)$ should have the MGF

$$\mathbb{E}_{\mathbb{Q}}\, \mathrm{e}^{\vartheta X(t)} = \frac{\mathbb{E}\, \mathrm{e}^{(\vartheta+\vartheta^\star)X(t)}}{\mathbb{E}\, \mathrm{e}^{\vartheta^\star X(t)}} = \frac{M(\vartheta+\vartheta^\star)}{M(\vartheta^\star)}; \tag{4.3}$$

under this choice the random variable has the desired mean:

$$\mathbb{E}_{\mathbb{Q}}\, X(t) = \frac{M'(\vartheta^\star)}{M(\vartheta^\star)} = a.$$

The question is now: how to sample a random variable that has this MGF? To this end, notice that $M(\vartheta) = \exp(-\lambda t + \lambda t\, \mathbb{E} \exp(\vartheta B \mathrm{e}^{-rU}))$ and

$$M(\vartheta+\vartheta^\star) = \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t\, \mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU}))^k}{k!} \left(\frac{\mathbb{E} \exp((\vartheta+\vartheta^\star) B\, \mathrm{e}^{-rU})}{\mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU})} \right)^k,$$

such that the expression in (4.3) equals

$$\sum_{k=0}^{\infty} \exp(-\lambda t\, \mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU})) \frac{(\lambda t\, \mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU}))^k}{k!} \times \left(\frac{\mathbb{E} \exp((\vartheta+\vartheta^\star) B\, \mathrm{e}^{-rU})}{\mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU})} \right)^k. \tag{4.4}$$

From this expression we can see how to sample the $X_i(t)$ under $\mathbb{Q}$, as follows. In the first place we conclude that under $\mathbb{Q}$ the number of arrivals becomes Poisson

with mean

$$\lambda t\, \mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU}) = \lambda \int_0^t \beta(\mathrm{e}^{-ru}\, \vartheta^\star) \mathrm{d}u, \tag{4.5}$$

rather than λt (which is an increase). Likewise, it entails that under $\mathbb{Q}$ the distribution of the $B_j \mathrm{e}^{-rU_j}$ should be twisted by $\vartheta^\star$, in the sense that these random variables should have under $\mathbb{Q}$ the MGF

$$\mathbb{E}_{\mathbb{Q}} \exp((\vartheta + \vartheta^\star) B\, \mathrm{e}^{-rU}) = \frac{\mathbb{E} \exp((\vartheta + \vartheta^\star) B\, \mathrm{e}^{-rU})}{\mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU})}.$$

We now point out how such a random variable should be sampled. To this end, observe that

$$\mathbb{E} \exp((\vartheta + \vartheta^\star) B\, \mathrm{e}^{-rU}) = \int_0^t \frac{\beta(\mathrm{e}^{-ru}(\vartheta + \vartheta^\star))}{\beta(\mathrm{e}^{-ru}\, \vartheta^\star)} \frac{1}{t}\, \beta(\mathrm{e}^{-ru}\, \vartheta^\star) \mathrm{d}u,$$

so that

$$\mathbb{E}_{\mathbb{Q}} \exp((\vartheta + \vartheta^\star) B\, \mathrm{e}^{-rU}) = \int_0^t \frac{\beta(\mathrm{e}^{-ru}(\vartheta + \vartheta^\star))}{\beta(\mathrm{e}^{-ru}\, \vartheta^\star)} \frac{\beta(\mathrm{e}^{-ru}\, \vartheta^\star)}{\displaystyle\int_0^t \beta(\mathrm{e}^{-rv}\, \vartheta^\star) \mathrm{d}v} \mathrm{d}u.$$

From this representation two conclusions can be drawn. In the first place, supposing there are k arrivals, then the arrival epochs $U_1, \ldots, U_k$ are i.i.d. under $\mathbb{Q}$, with the density given by

$$f_U^{\mathbb{Q}}(u) = \frac{\beta(\mathrm{e}^{-ru}\, \vartheta^\star)}{\displaystyle\int_0^t \beta(\mathrm{e}^{-rv}\, \vartheta^\star)\, \mathrm{d}v}.$$

In the second place, given that the k-th arrival occurs at time u, the density of the corresponding job size B_k should be exponentially twisted by $\mathrm{e}^{-ru}\, \vartheta^\star$ (where each of the job sizes is sampled independently of everything else).

Now that we know how to sample from $\mathbb{Q}$ it is straightforward to implement the importance sampling. Before we describe its complexity (in terms of the number of runs required to obtain an estimate with given precision), we first provide an example in which we demonstrate how the change of measure can be performed.

Example 4.2.1. *In this example we consider the case that the B_i are exponentially distributed with mean μ^{-1}. Applying the transformation $w := \mathrm{e}^{-ru}\,\vartheta/\mu$, it is first seen that*

$$\begin{aligned}\int_0^s \beta(\mathrm{e}^{-ru}\,\vartheta)\mathrm{d}u &= \int_0^s \frac{\mu}{\mu-\mathrm{e}^{-ru}\,\vartheta}\mathrm{d}u = \frac{1}{r}\int_{\mathrm{e}^{-rs}\vartheta/\mu}^{\vartheta/\mu}\frac{1}{1-w}\frac{1}{w}\mathrm{d}w\\ &= \frac{1}{r}\left[\log\frac{w}{1-w}\right]_{\mathrm{e}^{-rs}\vartheta/\mu}^{\vartheta/\mu} = \frac{1}{r}\log\left(\frac{\mu\mathrm{e}^{rs}-\vartheta}{\mu-\vartheta}\right).\end{aligned}$$

As $\vartheta^\star$ solves the equation $M'(\vartheta^\star)/M(\vartheta^\star) = a$, we obtain the quadratic equation

$$m(t) = a\left(1-\frac{\vartheta}{\mu}\right)\left(1-\frac{\vartheta}{\mu}\mathrm{e}^{-rt}\right),$$

leading to

$$\vartheta^\star = \frac{\mu\mathrm{e}^{rt}}{2}\left((1+\mathrm{e}^{-rt}) - \sqrt{(1-\mathrm{e}^{-rt})^2 + 4\mathrm{e}^{-rt}\frac{m(t)}{a}}\right)$$

(where it is readily checked that $\vartheta^\star \in (0,\mu)$).

Now we compute what the alternative measure $\mathbb{Q}$ amounts to. In the first place, the number of arrivals should become Poisson with parameter

$$\frac{\lambda}{r}\log\left(\frac{\mu\mathrm{e}^{rt}-\vartheta^\star}{\mu-\vartheta^\star}\right)$$

(which is larger than λt). In addition, we can check that

$$F_U^{\mathbb{Q}}(u) := \mathbb{Q}(U \leqslant u) = \log\left(\frac{\mu\mathrm{e}^{ru}-\vartheta^\star}{\mu-\vartheta^\star}\right)\Big/\log\left(\frac{\mu\mathrm{e}^{rt}-\vartheta^\star}{\mu-\vartheta^\star}\right)$$

(rather than u/t). The function $F_U^{\mathbb{Q}}(u)$ has the value 0 for $u = 0$ and the value 1 for $u = t$, and is concave. This concavity reflects that the arrival epochs of the shots tend to be closer to 0 under $\mathbb{Q}$ than under $\mathbb{P}$. This is because we identified each of the U_i with t minus the actual corresponding arrival epoch in (4.1); along the most likely path of $Y_n(t)$ itself the shots will be typically closer to t under $\mathbb{Q}$. Observe that one can sample U under $\mathbb{Q}$ using the classical inverse distribution

function method [3, Section II.2a]: with H denoting a uniform number on $[0,1)$, we obtain such a sample by

$$\frac{1}{r}\log\left(\left(\mathrm{e}^{rt}-\frac{\vartheta^\star}{\mu}\right)^H\left(1-\frac{\vartheta^\star}{\mu}\right)^{1-H}+\frac{\vartheta^\star}{\mu}\right).$$

Also, conditional on a U_i having attained the value u, the jobs B_i should be sampled from an exponential distribution with mean $(\mu-\mathrm{e}^{-ru}\,\vartheta^\star)^{-1}$.

Remark. In the model we study in this section, the input of the linear stochastic fluid network is a compound Poisson process. As pointed out in [51] the class of inputs can be extended to the more general class of increasing Lévy processes in a straightforward manner.

4.2.3 Efficiency properties of the IS procedure

In this subsection we analyse the performance of the procedure introduced in the previous section. The focus is on a characterisation of the number of runs needed to obtain an estimate with a given precision (at a given confidence level).

In every run $Y_n(t)$ is sampled under $\mathbb{Q}$, as pointed out above. As $\mathbb{Q}$ is an implementation of an exponential twist (with twist $\vartheta^\star$), the likelihood ratio (of sampling $Y_n(t)$ under $\mathbb{P}$ relative to $\mathbb{Q}$) is given by

$$L=\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}=\mathrm{e}^{-\vartheta^\star Y_n(t)}\mathrm{e}^{n\,\log M(\vartheta^\star)}.$$

In addition, define I as the indicator function of the event $\{Y_n(t)\geqslant na\}$. Clearly, $\mathbb{E}_{\mathbb{Q}}(LI)=p_n(a)$. We keep generating samples LI (under $\mathbb{Q}$), and estimate $p_n(a)$ by the corresponding sample mean, until the ratio of the half-width of the confidence interval (with critical value T) and the estimator drops below some predefined ε (say, 10%). Under $\mathbb{P}$ the number of runs needed is effectively inversely proportional to $p_n(a)$, hence exponentially increasing in n. We now focus on quantifying the reduction of the number of runs when using the importance sampling procedure we described above, i.e., the one based on the measure $\mathbb{Q}$.

Using a Normal approximation, it is a standard reasoning that when performing N runs the ratio of the half-width of the confidence interval and the

estimator is approximately

$$\frac{1}{p_n(a)} \cdot \frac{T}{\sqrt{N}} \sqrt{\mathbb{V}\mathrm{ar}_{\mathbb{Q}}(L^2 I)},$$

and hence the number of runs needed is roughly

$$\Sigma_n := \frac{T^2}{\varepsilon^2} \frac{\mathbb{V}\mathrm{ar}_{\mathbb{Q}}(L^2 I)}{(p_n(a))^2}.$$

We now analyse how Σ_n behaves as a function of the 'rarity parameter' n. Due to the Bahadur-Rao result [7], with $f_n \sim g_n$ denoting $f_n/g_n \to 1$ as $n \to \infty$,

$$p_n(a) = \mathbb{E}_{\mathbb{Q}}(LI) \sim \frac{1}{\sqrt{n}} \frac{1}{\vartheta^\star \sqrt{2\pi\tau}} \mathrm{e}^{-nI(a)}, \quad \tau := \frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2} \log M(\vartheta) \Big|_{\vartheta=\vartheta^\star}. \tag{4.6}$$

Using the same proof technique as in [7], it can be shown that

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \frac{1}{\sqrt{n}} \frac{1}{2\vartheta^\star \sqrt{2\pi\tau}} \mathrm{e}^{-2nI(a)}; \tag{4.7}$$

see Appendix A for the underlying computation. It also follows that $\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \mathbb{V}\mathrm{ar}_{\mathbb{Q}}(L^2 I)$.

We can use these asymptotics, to conclude that under $\mathbb{Q}$ the number of runs required grows slowly in n. More specifically, Σ_n is essentially proportional to $\sqrt{n}$ for n large. This leads to the following result; cf. [11, Section 2] for related findings in a more general context.

Proposition 4.2.2. *As $n \to \infty$,*

$$\Sigma_n \sim \alpha \sqrt{n}, \quad \alpha = \frac{T^2}{\varepsilon^2} \vartheta^\star \cdot \frac{1}{2} \sqrt{2\pi\tau}. \tag{4.8}$$

4.2.4 Simulation experiments

In this subsection we present numerical results for the single-node model without Markov modulation. We focus on the case of exponential jobs, as in Example 4.2.1. We simulate until the estimate has reached the precision $\varepsilon = 0.1$, with confidence level 0.95 (such that the critical value is $T = 1.96$). The parameters chosen are: $t = 1$, $r = 1$, $\lambda = 1$, and $\mu = 1$. We set $a = 1$ (which is

larger than $m(1) = 1 - e^{-1}$). As it turns out, $\vartheta^\star = 0.2918$ and

$$\tau = \frac{\lambda}{r}\left(\frac{1}{(\mu - \vartheta^\star)^2} - \frac{1}{(\mu \mathrm{e}^{rt} - \vartheta^\star)^2}\right) = 1.8240.$$

The top-left panel of Fig. 4.1 confirms the exponential decay of the probability of interest, as a function of n. In the top-right panel we verify that the number of runs indeed grows proportionally to $\sqrt{n}$; the value of α, as defined in (4.8), is 198.7, which is depicted by the horizontal line. The bottom-left panel shows the density of the arrival epochs, which confirms that the arrival epochs tend to be closer to 0 under $\mathbb{Q}$ than under $\mathbb{P}$; recall that under $\mathbb{P}$ these epochs are uniformly distributed on $[0, t]$. Recall that we reversed time in (4.1): for the actual shot-noise system that we are considering, it means that in order to reach the desired level at time t, the arrival epochs tend to be closer to t under $\mathbb{Q}$ than under $\mathbb{P}$. The bottom-right panel presents the rate of the exponential job sizes as a function of u. Using (4.5), the arrival rate under $\mathbb{Q}$ turns out to be 1.2315.

4.3 Multi-node systems, no modulation

In this section we consider multi-node linear stochastic fluid networks, of the type analysed in the work by Kella and Whitt [51]. It is instructive to first consider the simplest multi-node system: a tandem network without external input in the downstream node and no traffic leaving after having been served by the upstream node (and rate r_ℓ for node ℓ, $\ell = 1, 2$); later we extend the ideas developed to general linear stochastic fluid networks.

4.3.1 Preliminaries

As mentioned above, we first consider the two-node tandem. The content of the first node is, as before,

$$X^{(1)}(t) = \sum_{j=1}^{N} B_j \mathrm{e}^{-r_1(t-U_j)}$$

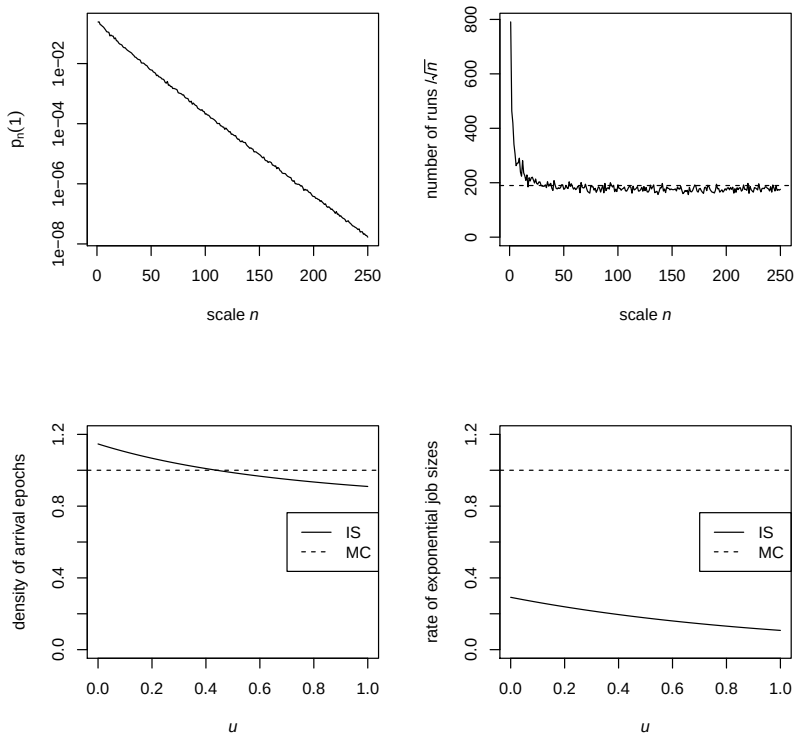

Figure 4.1: Numerical results for Section 4.2.4.

(with N having a Poisson distribution with mean λt), but it can be argued that the content of the second node satisfies a similar representation. More specifically, using the machinery developed in [51], it turns out that

$$\begin{aligned} X^{(2)}(t) &= \sum_{j=1}^{N} B_j \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2(t-U_j)} - \mathrm{e}^{-r_1(t-U_j)} \right) \\ &\stackrel{\mathrm{d}}{=} \sum_{j=1}^{N} B_j \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 U_j} - \mathrm{e}^{-r_1 U_j} \right). \end{aligned} \tag{4.9}$$

As before, perform the scaling by n, meaning that the arrival rate λ is inflated by a factor n. It leads to the random vectors $(X_1^{(1)}(t), \ldots, X_n^{(1)}(t))$ and $(X_1^{(2)}(t), \ldots, X_n^{(2)}(t))$. With these vectors we can define $Y_n^{(1)}(t)$ and $Y_n^{(2)}(t)$, analogously to how this was done in the single-node case; these two random quantities represent the contents of the upstream resource and the downstream resource, respectively.

The state of this tandem system can be uniquely characterised in terms of its (bivariate) moment generating function. The technique to derive an explicit expression is by relying on the above distributional equality (4.9). Again, the key step is to condition on the number of shots that have arrived in the interval $[0, t]$: with $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)$,

$$\begin{aligned} M(\boldsymbol{\vartheta}) &:= \mathbb{E}\, \mathrm{e}^{\vartheta_1 X^{(1)}(t) + \vartheta_2 X^{(2)}(t)} \\ &= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \\ &\quad \times \left(\mathbb{E} \exp\left(\vartheta_1 B \mathrm{e}^{-r_1 U} + \vartheta_2 B \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 U} - \mathrm{e}^{-r_1 U} \right) \right) \right)^k \\ &= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \\ &\quad \times \left(\int_0^t \frac{1}{t} \mathbb{E} \exp\left(\vartheta_1 B \mathrm{e}^{-r_1 u} + \vartheta_2 B \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 u} - \mathrm{e}^{-r_1 u} \right) \right) \mathrm{d}u \right)^k \\ &= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \end{aligned}$$

$$\times \left(\int_0^t \frac{1}{t} \beta \left(\mathrm{e}^{-r_1 u} \vartheta_1 + \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 u} - \mathrm{e}^{-r_1 u} \right) \vartheta_2 \right) \mathrm{d}u \right)^k$$
$$= \exp \left(\lambda \int_0^t \left(\beta \left(\mathrm{e}^{-r_1 u} \vartheta_1 + \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 u} - \mathrm{e}^{-r_1 u} \right) \vartheta_2 \right) - 1 \right) \mathrm{d}u \right). \tag{4.10}$$

The above computation is for the two-node tandem system, but the underlying procedure can be extended to the case of networks with more than 2 nodes, and external input in each of the nodes. To this end, we consider the following network consisting of L nodes. Jobs are generated according to a Poisson process. At an arrival epoch, an amount is added to the content of each of the resources $\ell \in \{1, \ldots, L\}$, where the amount added to resource ℓ is distributed as the (non-negative) random variable $B^{(\ell)}$; $\beta(\boldsymbol{\vartheta})$, with $\boldsymbol{\vartheta} \in \mathbb{R}^L$, is the joint MGF of $B^{(1)}$ up to $B^{(L)}$ (note that the components are not assumed independent). In addition, let the traffic level at node ℓ decay exponentially with rate r_ℓ (i.e., the value of the output rate is linear in the current level, with proportionality constant r_ℓ). A deterministic fraction $p_{\ell\ell'} \geqslant 0$ ($\ell \neq \ell'$) is then fed into node ℓ', whereas a fraction $p_{\ell\ell} \geqslant 0$ leaves the network (with $\sum_{\ell'=1}^L p_{\ell\ell'} = 1$). We denote $r_{\ell\ell'} := r_\ell p_{\ell\ell'}$. As an aside we mention that this general model covers models in which some arrivals (of the Poisson process with parameter λ) actually lead to arrivals at only a subset of the L queues (since the job sizes $B^{(1)}, \ldots, B^{(L)}$ are allowed to equal 0).

We now point out how the joint buffer content process can be analysed. Again our objective is to evaluate the moment generating function. Define the matrix R as follows: its (ℓ, ℓ)-th entry is $r_{\ell\ell} + \sum_{\ell' \neq \ell} r_{\ell\ell'}$, whereas its (ℓ, ℓ')-th entry (with $\ell \neq \ell'$) is $-r_{\ell\ell'}$. We have, according to Kella and Whitt [51], with N again Poisson with mean λt, the following distributional equality: for any $\ell \in \{1, \ldots, L\}$,

$$X^{(\ell)}(t) = \sum_{\ell'=1}^{L} \sum_{j=1}^{N} B_j^{(\ell')} (\mathrm{e}^{-R(t-U_j)})_{\ell'\ell}.$$

It means we can compute the joint MGF of $X^{(1)}(t)$ up to $X^{(L)}(t)$ as follows, cf.

[51, Thm. 5.1]:

$$\begin{aligned}
M(\boldsymbol{\vartheta}) &:= \mathbb{E}\,\exp\left(\sum_{\ell=1}^{L}\vartheta_\ell X^{(\ell)}(t)\right)\\
&= \sum_{k=0}^{\infty}\mathrm{e}^{-\lambda t}\frac{(\lambda t)^k}{k!}\left(\mathbb{E}\exp\left(\sum_{\ell=1}^{L}\vartheta_\ell\sum_{\ell'=1}^{L}B^{(\ell')}(\mathrm{e}^{-R(t-U)})_{\ell'\ell}\right)\right)^k\\
&= \sum_{k=0}^{\infty}\mathrm{e}^{-\lambda t}\frac{(\lambda t)^k}{k!}\left(\int_0^t\frac{1}{t}\mathbb{E}\exp\left(\sum_{\ell=1}^{L}\vartheta_\ell\sum_{\ell'=1}^{L}B^{(\ell')}(\mathrm{e}^{-Ru})_{\ell'\ell}\right)\mathrm{d}u\right)^k\\
&= \sum_{k=0}^{\infty}\mathrm{e}^{-\lambda t}\frac{(\lambda t)^k}{k!}\\
&\quad\times\left(\int_0^t\frac{1}{t}\beta\left(\sum_{\ell=1}^{L}(\mathrm{e}^{-Ru})_{1\ell}\vartheta_\ell,\ldots,\sum_{\ell=1}^{L}(\mathrm{e}^{-Ru})_{L\ell}\vartheta_\ell\right)\mathrm{d}u\right)^k\\
&= \exp\left(-\lambda t+\lambda\int_0^t\beta\left(\sum_{\ell=1}^{L}(\mathrm{e}^{-Ru})_{1\ell}\vartheta_\ell,\ldots,\sum_{\ell=1}^{L}(\mathrm{e}^{-Ru})_{L\ell}\vartheta_\ell\right)\mathrm{d}u\right)\\
&= \exp\left(\lambda\int_0^t\left(\beta\left(\mathrm{e}^{-Ru}\,\boldsymbol{\vartheta}\right)-1\right)\mathrm{d}u\right),
\end{aligned}$$

which is the matrix/vector-counterpart of the expression (4.2) that we found in the single-node case; for the two-node case the special form (4.10) applies.

4.3.2 Tail probabilities, change of measure

In this subsection we introduce the change of measure that we use in our importance sampling approach. Many of the concepts are analogous to concepts used for the single-node case in Section 2.

Define (in self-evident notation)

$$p_n(\boldsymbol{a}) := \mathbb{P}\left(Y_n^{(1)}(t)\geqslant na_1,\ldots,Y_n^{(L)}(t)\geqslant na_L\right).$$

Due to the multivariate version of Cramér's theorem, with $A := [a_1,\infty)\times\cdots\times$

$[a_L, \infty)$,

$$\lim_{n\to\infty} \frac{1}{n} \log p_n(\boldsymbol{a}) = -\inf_{\boldsymbol{b}\in A} I(\boldsymbol{b}), \text{ where } I(\boldsymbol{b}) := \sup_{\boldsymbol{\vartheta}} \left(\langle \boldsymbol{\vartheta}, \boldsymbol{b}\rangle - \log M(\boldsymbol{\vartheta})\right). \quad (4.11)$$

More refined asymptotics than the logarithmic asymptotics of (4.11) are available as well, but these are not yet relevant in the context of the present subsection; we return to these 'exact asymptotics' in Section 4.3.3.

We assume that the set A is 'rare', in the sense that

$$\boldsymbol{m}(t) \notin A, \quad \text{with} \quad m_i(t) := \left.\frac{\partial M(\boldsymbol{\vartheta})}{\partial \vartheta_i}\right|_{\boldsymbol{\vartheta}=\mathbf{0}}.$$

Let us now construct the importance sampling measure. Let $\boldsymbol{\vartheta}^\star$ be the optimising $\boldsymbol{\vartheta}$ in the decay rate of $p_n(\boldsymbol{a})$. Mimicking the reasoning we used in the single-node case, the number of arrivals becomes Poisson with mean

$$\lambda \int_0^t \beta\left(\mathrm{e}^{-Ru}\,\boldsymbol{\vartheta}^\star\right)\,\mathrm{d}u.$$

The arrival epochs should be drawn using the density

$$f_U^{\mathbb{Q}}(u) = \frac{\beta\left(\mathrm{e}^{-Ru}\,\boldsymbol{\vartheta}^\star\right)}{\displaystyle\int_0^t \beta\left(\mathrm{e}^{-Rv}\,\boldsymbol{\vartheta}^\star\right)\mathrm{d}v}.$$

Given an arrival at time u, $(B^{(1)}, \ldots, B^{(L)})$ should be exponentially twisted by

$$\left((\mathrm{e}^{-Ru}\,\boldsymbol{\vartheta}^\star)_1, \ldots, (\mathrm{e}^{-Ru}\,\boldsymbol{\vartheta}^\star)_L\right).$$

4.3.3 Efficiency properties of the IS procedure

We now consider the efficiency properties of the change of measure proposed in the previous subsection. To this end, we first argue that the vector $\boldsymbol{\vartheta}$ generally has some (at least one) strictly positive entries, whereas the other entries equal 0; i.e., there are *no* negative entries. To this end, we first denote by $\boldsymbol{b}^\star$ the 'most likely point' in A:

$$\boldsymbol{b}^\star := \arg\inf_{\boldsymbol{b}\in A} I(\boldsymbol{b}),$$

so that $\boldsymbol{\vartheta}^\star = \boldsymbol{\vartheta}(\boldsymbol{b}^\star)$. It is a standard result from convex optimisation that

$$\frac{\partial I(\boldsymbol{b})}{\partial b_i} = \vartheta_i(\boldsymbol{b}). \tag{4.12}$$

Suppose now that $\vartheta_i(\boldsymbol{b}^\star) < 0$. Increasing the i-th component of the $\boldsymbol{b}^\star$ (while leaving all other components unchanged) would lead to a vector that is still in A, but that by virtue of (4.12) corresponds to a lower value of the objective function $I(\cdot)$, thus yielding that $\boldsymbol{b}^\star$ was not the optimiser; we have thus found a contradiction. Similarly, when $\vartheta_i(\boldsymbol{b}^\star) = 0$ we have that $b_i^\star > a_i$ (as otherwise a reduction of the objective function value would be possible, which contradicts $\boldsymbol{b}^\star$ being minimiser).

Now define Θ as the subset of $i \in \{1, \dots, L\}$ such that $\vartheta_i > 0$, and let $D \in \{1, \dots, L\}$ the number of elements of Θ. We now argue that the number of runs needed to obtain an estimate of predefined precision scales as $n^{D/2}$. Relying on the results from [23] (in particular their Thm. 3.4), it follows that $p_n(\boldsymbol{a})$ behaves (for n large) proportionally to $n^{-D/2}\exp(-nI(\boldsymbol{b}^\star))$; using the same machinery, $\mathbb{E}_{\mathbb{Q}}(L^2 I)$ behaves proportionally to $n^{-D/2}\exp(-2nI(\boldsymbol{b}^\star))$. Mimicking the line of reasoning of Section 4.2.3, we conclude that the number of runs needed is essentially proportional to $n^{D/2}$. The formal statement is as follows; in Appendix A we comment on the underlying computations.

Proposition 4.3.1. *As $n \to \infty$,*

$$\Sigma_n \sim \alpha\, n^{D/2}, \quad \alpha = \frac{T^2}{\varepsilon^2}\left(\prod_{i\in D}\vartheta_i^\star\right)\cdot\frac{1}{2^D}\left(\sqrt{2\pi}\right)^D\sqrt{\tau}, \tag{4.13}$$

where τ is the determinant of the Hessian of $\log M(\boldsymbol{\vartheta})$ in $\boldsymbol{\vartheta}^\star$.

We further illustrate the ideas and intuition behind the qualitative result described in the above proposition by considering the case $L = 2$. It is noted that three cases may arise: (i) $\Theta = \{1, 2\}$, (ii) $\Theta = \{1\}$, (iii) $\Theta = \{2\}$; as case (iii) can be dealt with in the same way as case (ii), we concentrate on the cases (i) and (ii) only. In case (i), where $D = 2$, the necessary condition [23, Eqn. (3.4)] is fulfilled as $\boldsymbol{\vartheta} > 0$ componentwise. As in addition the conditions A–C of [23] are in place, it is concluded that [23, Thm. 3.4] can be applied, leading to

$\boldsymbol{b}^\star = \boldsymbol{a}$, and

$$p_n(\boldsymbol{a}) \sim \frac{1}{n} \frac{1}{\vartheta_1^\star \vartheta_2^\star \cdot 2\pi\sqrt{\tau}} \mathrm{e}^{-nI(\boldsymbol{a})},$$

where τ is the determinant of the Hessian of $\log M(\boldsymbol{\vartheta})$ in $\boldsymbol{\vartheta}^\star$. Along the same lines, it can be shown that

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \frac{1}{n} \frac{1}{4\vartheta_1^\star \vartheta_2^\star \cdot 2\pi\sqrt{\tau}} \mathrm{e}^{-2nI(\boldsymbol{a})}.$$

It now follows that Σ_n is roughly linear in n: with ε and T as introduced in Section 4.2.3,

$$\Sigma_n = \alpha\, n, \quad \alpha := \frac{T^2}{\varepsilon^2} \vartheta_1^\star \vartheta_2^\star \cdot \frac{\pi\sqrt{\tau}}{2}. \tag{4.14}$$

In case (ii), we do not have that $\boldsymbol{\vartheta} > 0$ componentwise, and hence [23, Thm. 3.4] does not apply; in the above terminology, $D = 1 < 2 = L$. Observe that in this case the exponential decay rate of the event $\{Y_n^{(1)}(t) \geqslant na_1, Y_n^{(2)}(t) < na_2\}$ strictly majorises that of $\{Y_n^{(1)}(t) \geqslant na_1\}$ (informally: the former event is substantially less likely than the latter). It thus follows that $b_1^\star = a_1$ and $b_2^\star > a_2$, and

$$\begin{aligned} p_n(\boldsymbol{a}) &= \mathbb{P}\left(Y_n^{(1)}(t) \geqslant na_1\right) - \mathbb{P}\left(Y_n^{(1)}(t) \geqslant na_1, Y_n^{(2)}(t) < na_2\right) \\ &\sim \mathbb{P}\left(Y_n^{(1)}(t) \geqslant na_1\right) \\ &\sim \frac{1}{\sqrt{n}} \frac{1}{\vartheta_1^\star \sqrt{2\pi\tau}} \mathrm{e}^{-2nI(\boldsymbol{b}^\star)}, \qquad \tau := \frac{\mathrm{d}}{\mathrm{d}\vartheta^2} \log M(\vartheta, 0)\Big|_{\vartheta=\vartheta_1^\star}, \end{aligned}$$

and in addition

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \frac{1}{\sqrt{n}} \frac{1}{2\vartheta_1^\star \sqrt{2\pi\tau}} \mathrm{e}^{-2nI(\boldsymbol{b}^\star)}.$$

As a consequence in this regime Σ_n grows essentially proportional to $\sqrt{n}$ for n large:

$$\Sigma_n \sim \alpha\sqrt{n}, \quad \alpha := \frac{T^2}{\varepsilon^2} \vartheta_1^\star \cdot \frac{1}{2} \sqrt{2\pi\tau}.$$

In case (iii) Σ_n behaves proportionally to $\sqrt{n}$ as well.

4.3.4 Simulation experiments

We conclude this section by providing a few numerical illustrations. In the first set we focus on the downstream queue only (i.e., we analyse $p_n(0, a_2)$), whereas in the second set we consider the joint exceedance probability $p_n(\boldsymbol{a})$. The precision and confidence have been chosen as in Example 4.2.1. Throughout we take $t = 1$, $r_1 = 2$, $r_2 = 1$, $\lambda = 1$, and $\mu = 1$.

In the first set of experiments we take $a_1 = 0$ and $a_2 = 1$. Elementary numerical analysis yields that $\vartheta^\star = 0.8104$ and $\tau = 1.4774$, leading to α, as defined in (4.14), equalling 474.3. For graphs on the behaviour of $p_n(1)$ as a function of n and the number of runs needed, we refer to [17, Fig. 2]. The two panels of Fig. 4.2 should be interpreted as the bottom panels of Fig. 4.1. Interestingly, the left panel indicates that in the tandem system it does not pay off to let jobs arrive right before t (as they first have to go through the first resource to end up in the second resource), as reflected by the shape of the density of the arrival epochs under $\mathbb{Q}$; to this end, recall that we reversed time in (4.9), so that a low density at $u = 0$ in the graph corresponds to a high density at $u = t$ in the actual system. The arrival rate under $\mathbb{Q}$ is 1.5103.

In the second set of experiments we take $a_1 = 1.2$ and $a_2 = 1.1$; all other parameters are the same as in the first set. As mentioned above, we now consider the joint exceedance probability. As it turns out, $\vartheta_1^\star = 0.1367$ and $\vartheta_2^\star = 0.2225$. For graphs describing the behaviour of $p_n(1.2, 1.1)$ as a function of n and the number of runs needed, we refer to [17, Fig. 3]; the latter graph reveals that for this specific parameter setting Σ_n/n converges to the limiting constant rather slowly. Concerning the left panel of Fig. 4.3, note that in Section 4.2 we saw that to make sure the first queue gets large it helps to have arrivals at the end of the interval, whereas above we observed that to make the second queue large arrivals should occur relatively early. We now focus on the event that *both* queues are large, and consequently the arrival distribution becomes relatively uniform again, as shown in the left panel of Fig. 4.3. The arrival rate under $\mathbb{Q}$ is 2.3478.

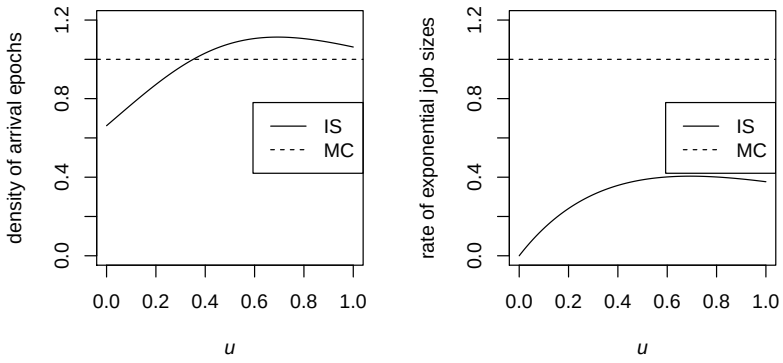

Figure 4.2: Numerical results for Section 4.3.4: downstream queue only.

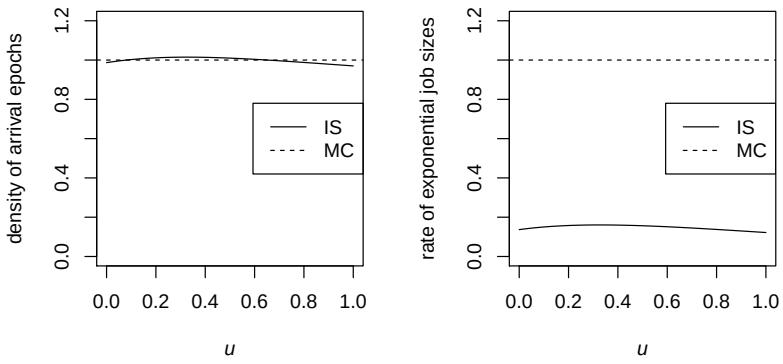

Figure 4.3: Numerical results for Section 4.3.4: both queues.

4.4 Multi-node systems under Markov modulation

In this section consider the networks analysed in the previous section, but now in a random environment. More specifically, the type of random environment we focus on here is known as *Markov modulation*: the system dynamics are affected by the state of an external finite-state irreducible Markov process $J(\cdot)$ with generator matrix $Q = (q_{jj'})_{j,j'=1}^{d}$. When this Markov process (usually referred to as the *background process*) is in state j, arrivals occur according to a Poisson process with rate λ_j, the MGF of the job size is $\beta_j(\boldsymbol{\vartheta})$, and the routing matrix is R_j. Analogously to the definitions used in the case without Markov modulation, this routing matrix' (i,i)-th entry is

$$(R_j)_{ii} := r_{ii}^{(j)} + \sum_{i' \neq i} r_{ii'}^{(j)},$$

which can be interpreted as the rate at which fluid leaves server i when the background process is in j. Likewise, for $i \neq i'$,

$$(R_j)_{ii'} := -r_{ii'}^{(j)},$$

which is the rate at which fluid flows from server i to i' when the background process is in j.

Below we assume that $J(0) = j_0$ for a fixed state $j_0 \in \{1, \ldots, d\}$; it is seen that all results generalise to an arbitrary initial distribution in a straightforward manner.

The structure of the section is as follows: we propose an importance sampling measure, establish efficiency properties of the corresponding estimator, and present a number of numerical experiments.

For the models covered in Sections 4.2 and 4.3, already detailed explicit analysis is available; see e.g. the results in terms of transforms and moments in [51]. Such a complete analysis is lacking for the model featuring in the present section; we refer to [18, Section 4.1] for results on exact expressions for moments.

4.4.1 Tail probabilities, change of measure

We now characterise the decay rate of the rare-event probability under study, and we propose a change of measure to efficiently estimate it. In the notation we have been using so far, we again focus on

$$p_n(\boldsymbol{a}) := \mathbb{P}\left(Y_n^{(1)}(t) \geqslant na_1, \ldots, Y_n^{(L)}(t) \geqslant na_L\right) = \mathbb{P}\left(\boldsymbol{Y}_n(t) \in A\right),$$

where $\boldsymbol{Y}_n(t) = (Y_n^{(1)}(t), \ldots, Y_n^{(L)}(t))$. It is stressed that, following [15], we consider the regime in which the background process is 'slow'. In concrete terms, this means that we linearly speed up the driving Poisson process (i.e., we replace the arrival rates λ_j by $n\lambda_j$), but leave the rates of the Markovian background process unaltered.

First we find an alternative characterisation of the state of the system at time t. Let $\mathscr{F}_t$ denote the set of all functions from $[0,t]$ onto the states $\{1,\ldots,d\}$. Consider a path $f \in \mathscr{F}_t$. Let f have $K(f)$ jumps between 0 and t, whose epochs we denote by $t_1(f)$ up to $t_{K(f)}(f)$ (and in addition $t_0(f) := 0$ and $t_{K(f)+1}(f) := t$). Let

$$j_i(f) := \lim_{t \downarrow t_i(f)} f(t)$$

(i.e., the state of f immediately after the i-th jump). We also introduce

$$\begin{aligned} D_i(u,f) &:= \exp\left(-(t_{i+1}(f) - u)\, R_{j_i(f)}\right), \\ D_i(f) &:= \exp\left(-(t_{i+1}(f) - t_i(f))\, R_{j_i(f)}\right). \end{aligned}$$

Suppose now that the Markov process $J(\cdot)$ follows the path $f \in \mathscr{F}_t$. Then the contribution to the MGF of $\boldsymbol{X}(t)$ due to shots that arrived between $t_i(f)$ and $t_{i+1}(f)$ is, mimicking the arguments that we used in Section 4.3.2 for non-modulated networks,

$$\psi_i(f,\boldsymbol{\vartheta}) := \exp\left(\lambda_{j_i(f)} \int_{t_i(f)}^{t_{i+1}(f)} \left(\beta_{j_i(f)}\left(D_i(u,f) D_{i+1}(f) \cdots D_{K(f)}(f)\, \boldsymbol{\vartheta}\right) - 1\right) \mathrm{d}u\right).$$

As a consequence, the MGF of $X(t)$ given the path f is

$$M_f(\boldsymbol{\vartheta}) := \prod_{i=0}^{K(f)} \psi_i(f, \boldsymbol{\vartheta}).$$

First conditioning on the path of $J(\cdot) \in \mathscr{F}_t$ between 0 and t and then unconditioning, it then immediately follows that the MGF of $\boldsymbol{X}(t)$ is given by

$$M(\boldsymbol{\vartheta}) = \mathbb{E}\, M_J(\boldsymbol{\vartheta}).$$

Then, precisely as is shown in [15] for a related stochastic system, the decay rate can be characterised as follows:

$$\lim_{n\to\infty} \frac{1}{n} \log p_n(\boldsymbol{a}) = - \inf_{f\in\mathscr{F}_t} \mathbb{I}_f(\boldsymbol{a}), \quad \mathbb{I}_f(\boldsymbol{a}) := \inf_{\boldsymbol{b}\in A} \sup_{\boldsymbol{\vartheta}} \left(\langle \boldsymbol{\vartheta}, \boldsymbol{b} \rangle - \log M_f(\boldsymbol{\vartheta}) \right). \tag{4.15}$$

The argumentation to show this is analogous to the one in [15, Thm. 1], and can be summarised as follows. In the first place, let $f^\star$ be the optimising path in (4.15). Then, as $J(\cdot)$ does not depend on n, we can choose a 'ball' $\mathscr{B}_t(f^\star)$ around $f^\star$ such that the decay rate of the probability of $J(\cdot)$ being in that ball is 0. The lower bound follows by only taking into account the contribution due to paths in $\mathscr{B}_t(f^\star)$. The upper bound follows by showing that the contribution of all $f \in \mathscr{F}_t \setminus \mathscr{B}_t(f^\star)$ is negligible.

Informally, the path $f^\star$ has the interpretation of the most likely path of $J(\cdot)$ given that the rare event under consideration happens. To make sure that the event under consideration is rare, we assume that for all $f \in \mathscr{F}_t$

$$\left(\frac{\partial}{\partial \vartheta_1} M_f(\boldsymbol{\vartheta}) \bigg|_{\boldsymbol{\vartheta}=\mathbf{0}}, \ldots, \frac{\partial}{\partial \vartheta_L} M_f(\boldsymbol{\vartheta}) \bigg|_{\boldsymbol{\vartheta}=\mathbf{0}} \right) \notin A.$$

The change of measure we propose is the following. In every run we first sample the path $J(s)$ for $s \in [0,t]$ *under the original measure* $\mathbb{P}$ (i.e., with $J(0) = j_0$, and then using the generator matrix Q). We call the resulting path $f \in \mathscr{F}_t$. For this path, define $\boldsymbol{\vartheta}_f^\star \geqslant \mathbf{0}$ as the optimising $\boldsymbol{\vartheta}$ in the definition of $\mathbb{I}(f)$ in (4.15); $\boldsymbol{b}_f^\star \in A$ is the optimising $\boldsymbol{b}$.

Conditional on the path f of the background process, under the new measure

$\mathbb{Q}$ the number of external arrivals between $t_i(f)$ and $t_{i+1}(f)$ is Poisson with parameter

$$\int_{t_i(f)}^{t_{i+1}(f)} \lambda_{j_i(f)} \beta_{j_i(f)} \left(P_i(u,f)\, \boldsymbol{\vartheta}_f^\star \right) \mathrm{d}u,$$

where $P_i(u,f) := D_i(u,f) D_{i+1}(f) \cdots D_{K(f)}(f)$. The arrival epochs between $t_i(f)$ and $t_{i+1}(f)$ should be drawn using the density

$$f_U^{\mathbb{Q}}(u) = \frac{\beta_{j_i(f)} \left(P_i(u,f)\, \boldsymbol{\vartheta}_f^\star \right)}{\displaystyle\int_{t_i(f)}^{t_{i+1}(f)} \beta_{j_i(f)} \left(P_i(v,f)\, \boldsymbol{\vartheta}_f^\star \right) \mathrm{d}v}.$$

Given an arrival at time u between $t_i(f)$ and $t_{i+1}(f)$, the job sizes $(B^{(1)}, \ldots, B^{(L)})$ should be sampled from a distribution with MGF $\beta_{j_i(f)}(\boldsymbol{\vartheta})$, but then exponentially twisted by

$$\left(\left(P_i(u,f)\, \boldsymbol{\vartheta}_f^\star \right)_1, \ldots, \left(P_i(u,f)\, \boldsymbol{\vartheta}_f^\star \right)_L \right).$$

Remark. As mentioned above, the background process is sampled under the original measure, whereas an alternative measure is used for the number of arrivals, the arrival epochs, and the job sizes. The intuition behind this, is that the rare event under consideration is caused by two effects:

- *In the first place, samples of the background process J should be close to $f^\star$. Under $\mathbb{P}$ a reasonable fraction ends up close to $f^\star$ — more precisely, the event of J being close to $f^\star$ does not become increasingly rare when n grows. As a consequence, no change of measure is needed here.*
- *In the second place, given the path of the background process, the $Y_n^{(\ell)}(t)$ should exceed the values na_ℓ, for $\ell = 1, \ldots, L$. This event does become exponentially rare as n grows, so importance sampling is to be applied here.*

4.4.2 Efficiency properties of the IS procedure

We now analyse the speed up realised by the change of measure introduced in the previous subsection. Unlike our results for the non-modulated systems, now we cannot find the precise rate of growth of Σ_n. What *is* possible though, is proving *asymptotic efficiency* (also sometimes referred to as *logarithmic efficiency*), in the sense that we can show that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{Q}}(L^2 I) = \lim_{n\to\infty} \frac{2}{n} \log p_n(\boldsymbol{a}) = -2 \inf_{f\in\mathscr{F}_t} \inf_{\boldsymbol{b}\in A} \sup_{\boldsymbol{\vartheta}} \left(\langle \boldsymbol{\vartheta}, \boldsymbol{b}\rangle - \log M_f(\boldsymbol{\vartheta})\right)$$

(where the second equality is a consequence of (4.15)). This equality is proven as follows. As by Jensen's inequality $\mathbb{E}_{\mathbb{Q}}(L^2 I) \geqslant (\mathbb{E}_{\mathbb{Q}}(LI))^2 = (p_n(\boldsymbol{a}))^2$, we are left to prove the upper bound:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{Q}}(L^2 I) \leqslant \lim_{n\to\infty} \frac{2}{n} \log p_n(\boldsymbol{a}).$$

If the path of $J(\cdot)$ equals $f \in \mathscr{F}_t$, it follows by an elementary computation that we have constructed the measure $\mathbb{Q}$ such that

$$L = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \prod_{\ell=1}^{L} \exp\left(-\langle \boldsymbol{\vartheta}_f^\star, \boldsymbol{Y}_n(t)\rangle + n \, \log M_f(\boldsymbol{\vartheta}_f^\star)\right).$$

The fact that $\boldsymbol{\vartheta}_f^\star$ is componentwise non-negative, in combination with the fact that $\boldsymbol{Y}_n(t) \geqslant \boldsymbol{a}$ when $I = 1$, entails that

$$LI \leqslant \exp\left(-n \, \langle \boldsymbol{\vartheta}_f^\star, \boldsymbol{a}\rangle + n \, \log M_f(\boldsymbol{\vartheta}_f^\star)\right) = \exp\left(-n \, \langle \boldsymbol{\vartheta}_f^\star, \boldsymbol{b}_f^\star\rangle + n \, \log M_f(\boldsymbol{\vartheta}_f^\star)\right) = \mathrm{e}^{-n\,\mathbb{I}_f(\boldsymbol{a})},$$

noting that $\boldsymbol{a}$ and $\boldsymbol{b}_f^\star$ may only differ if the corresponding entry of $\boldsymbol{\vartheta}_f^\star$ equals 0 (that is, $\langle \boldsymbol{a} - \boldsymbol{b}_f^\star, \boldsymbol{\vartheta}_f^\star\rangle = 0$). The upper bound thus follows: with $f^\star$ the minimising path in (4.15), recalling that $J(\cdot)$ is sampled under $\mathbb{P}$,

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \leqslant \mathbb{E}\, \mathrm{e}^{-2n\,\mathbb{I}_J(\boldsymbol{a})} \leqslant \mathrm{e}^{-2n\,\mathbb{I}_{f^\star}(\boldsymbol{a})}.$$

We have established the following result.

Proposition 4.4.1. *As $n \to \infty$, the proposed importance sampling procedure is asymptotically efficient. This means that the number of runs needed grows subexponentially:*

$$\lim_{n\to\infty} \frac{1}{n} \log \Sigma_n = 0.$$

Remark. In the scaling considered, for both the logarithmic asymptotics of $p_n(\boldsymbol{a})$ and our importance sampling algorithm, the precise transition rates q_{ij} do not matter; the only crucial element is that the background process is irreducible. Observe that, even though the logarithmic asymptotics of $p_n(a)$ do not depend on the actual values of the transition rates q_{ij}, the probability $p_n(a)$ itself and its exact asymptotics do *depend on those rates. We refer to [14] for the exact asymptotics of a related infinite-server model; it is noted that the derivation of such precise asymptotics is typically highly involved.*

The above reasoning indicates that the proposed procedure remains valid under more general conditions: the ideas carry over to any situation in which the rates are piecewise constant along the most likely path.

4.4.3 Simulation experiments

We performed experiments featuring a single-node system under Markov modulation. In our example the job sizes stem from an exponential distribution. When the background process is in state i, the arrival rate is λ_i, the job-size distribution is exponential with parameter μ_i, and the rate at which the storage level decays is r_i, for $i \in \{1, \ldots, d\}$.

The change of measure is then implemented as follows. As pointed out in Section 4.4.1, per run a path f of the background process is sampled under the original measure $\mathbb{P}$. Suppose along this path there are K transitions (remarking that, for compactness, we leave out the argument f here), say at times t_1 up to t_K; with $t_0 = 0$ and $t_{K+1} = t$, the state between t_i and t_{i+1} is denoted by j_i, for $i = 0, \ldots, K$. Per run a specific change of measure is to be computed, parametrised by the t_i and j_i, as follows.

We define

$$P_i(u) := \bar{P}_i e^{r_{j_i} u}, \quad \bar{P}_i := e^{-r_{j_i} t_{i+1}} \prod_{i'=i+1}^{K} e^{-r_{j_{i'}} (t_{i'+1} - t_{i'})};$$

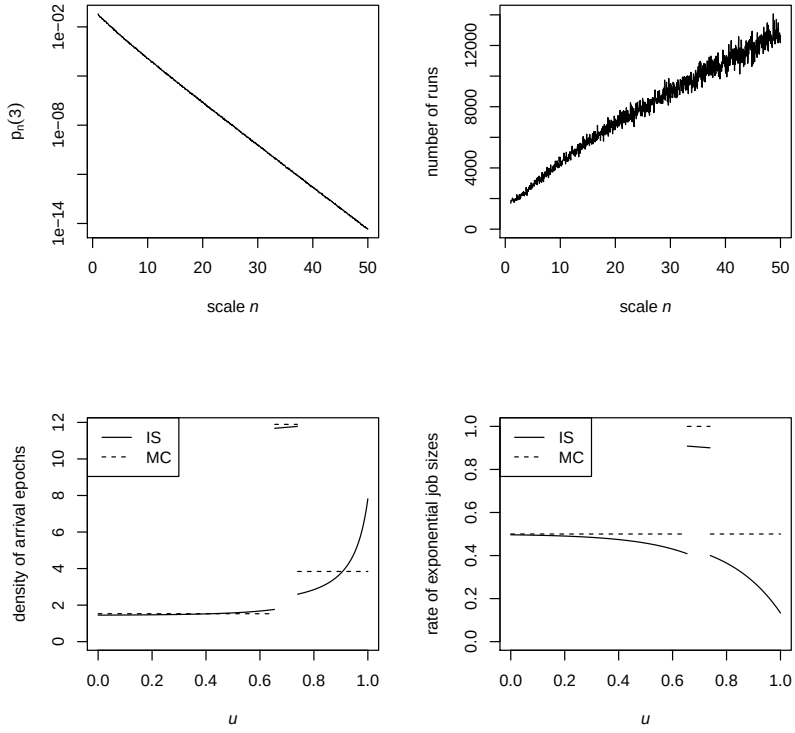

Figure 4.4: Numerical results for Section 4.4.3: first example.

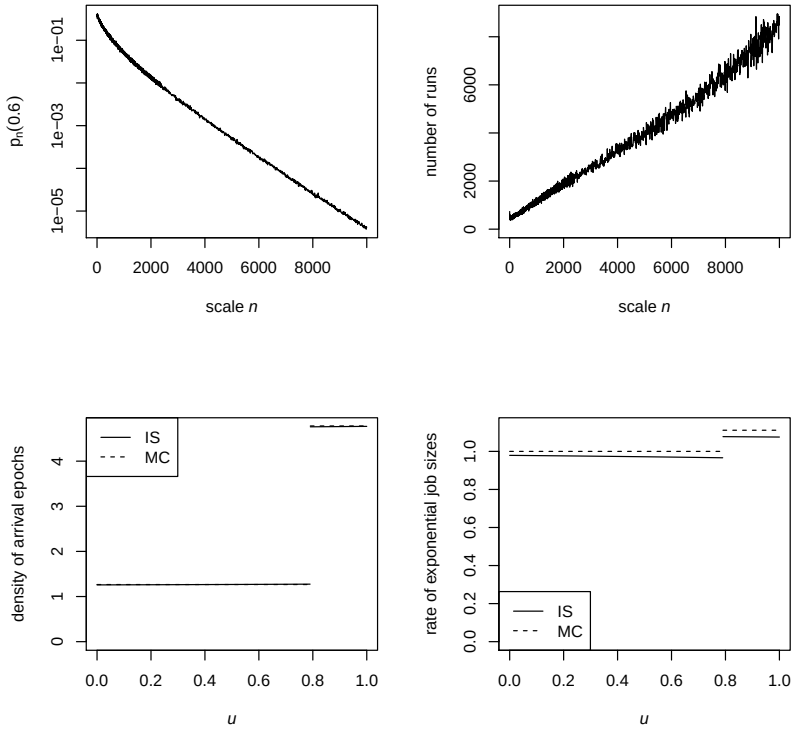

Figure 4.5: Numerical results for Section 4.4.3: second example.

the product in this expression should be interpreted as 1 if $i+1 > K$. It is readily checked that

$$M(\vartheta) = \prod_{i=0}^{K} \exp\left(\lambda_{j_i} \int_{t_i}^{t_{i+1}} \frac{P_i(u)\,\vartheta}{\mu_{j_i} - P_i(u)\,\vartheta} \mathrm{d}u\right).$$

Let $\vartheta^\star$ be the maximising argument of $\vartheta a - \log M(\vartheta)$.

We can now provide the alternative measure $\mathbb{Q}$ for this path of the background process. The number of arrivals between t_i and t_{i+1} (for $i = 0, \ldots, K$) becomes Poisson with parameter

$$\begin{aligned}\int_{t_i}^{t_{i+1}} \lambda_{j_i} \frac{\mu_{j_i}}{\mu_{j_i} - P_i(u)\,\vartheta^\star} \mathrm{d}u &= \frac{\lambda_{j_i}}{r_{j_i}} \log\left(\frac{\mu_{j_i} - \bar{P}_i e^{r_{j_i} t_i} \vartheta^\star}{\mu_{j_i} e^{-r_{j_i}(t_{i+1}-t_i)} - \bar{P}_i e^{r_{j_i} t_i} \vartheta^\star}\right) \\ &= \frac{\lambda_{j_i}}{r_{j_i}} \log\left(\frac{\mu_{j_i} - \bar{P}_i e^{r_{j_i} t_i} \vartheta^\star}{\mu_{j_i} - \bar{P}_i e^{r_{j_i} t_{i+1}} \vartheta^\star}\right) + \lambda_{j_i}(t_{i+1} - t_i),\end{aligned}$$

(where it is noted that this expression is larger than $\lambda_{j_i}(t_{i+1} - t_i)$, which was the parameter under $\mathbb{P}$). The density of each of the arrivals between t_i and t_{i+1} becomes

$$\begin{aligned}&\left(\frac{1}{\mu_{j_i} - P_i(u)\,\vartheta^\star}\right) \Big/ \int_{t_i}^{t_{i+1}} \left(\frac{1}{\mu_{j_i} - P_i(v)\,\vartheta^\star}\right) \mathrm{d}v \\ &\quad = \left(\frac{\mu_{j_i}}{\mu_{j_i} - P_i(u)\,\vartheta^\star}\right) \Big/ \frac{1}{r_{j_i}} \log\left(\frac{\mu_{j_i} - \bar{P}_i e^{r_{j_i} t_i} \vartheta^\star}{\mu_{j_i} e^{-r_{j_i}(t_{i+1}-t_i)} - \bar{P}_i e^{r_{j_i} t_i} \vartheta^\star}\right)\end{aligned}$$

(rather than a uniform distribution, as was the case under $\mathbb{P}$); sampling from this distribution is easy, since the inverse distribution function can be determined in closed form. Given an arrival that takes place at time u between t_i and t_{i+1}, the job size is exponential with parameter $\mu_{j_i} - P_i(u)\,\vartheta^\star$ (rather than exponential with parameter μ_{j_i}).

We now describe two examples in which the dimension of the background process is $d = 2$, $q_{12} = q_{21} = 2$, and $t = 1$. In the first example we fix $a = 3$, $\boldsymbol{\lambda} = (2, 1)$, $\boldsymbol{\mu} = (\frac{1}{2}, 1)$, and $\boldsymbol{r} = (5, 1)$, in the second example $a = 0.8$, $\boldsymbol{\lambda} = (0.9, 1)$, $\boldsymbol{\mu} = (0.9^{-1}, 1)$, and $\boldsymbol{r} = (0.3, 0.6)$. As before, we simulate until the precision of the estimate has reached $\varepsilon = 0.1$. The top two panels in Figs. 4.4–4.5 should be read as those in Figs. 4.1–4.3; the bottom two panels correspond

to the density of the arrival epochs and the rate of the exponential job sizes, respectively, for f the 'empirical maximiser' of $\mathbb{I}_f(a)$ (i.e., the maximiser of $\mathbb{I}_f(a)$ over all paths f of the background process that were sampled in the simulation experiment).

In the first example the thus obtained 'optimal path' successively visits states 1, 2, and 1, where the corresponding jump times are $t_1^\star = 0.654$ and $t_2^\star = 0.739$, and the decay rate is 0.573. The mean numbers of arrivals in the three parts of the optimal path are 1.392, 0.090 and 0.963 respectively, whereas for Monte Carlo sampling these are 1.308, 0.085 and 0.522 respectively.

In the second example the optimal path successively visits states 2 and 1, where the corresponding jump time is $t_1^\star = 0.790$. In this case the decay rate has the value 0.000806. The mean numbers of arrivals in the two parts of the optimal path are 0.812 and 0.195 respectively, which are slightly higher than the corresponding values under Monte Carlo sampling (0.790 and 0.189 respectively). Observe that in this example the difference between the two measures is relatively small, also reflected by the small value of the decay rate; the event under consideration technically qualifies as 'rare' in that $p_n(0.8) \to 0$ as $n \to \infty$, but has a relatively high likelihood (e.g. as compared to the first example). As a consequence of the fact that both measures almost coincide, the two densities in the bottom-left panel can hardly be distinguished.

We observe that the top panels confirm that in both examples (i) $p_n(a)$ decays roughly exponentially in n, (ii) the number of runs needed grows roughly linearly in n (in the first example slightly sublinearly).

4.5 Discussion and concluding remarks

In this chapter we have considered the probability of attaining a value in a rare set A at a fixed point in time t: with $A = [a_1, \infty) \times \cdots \times [a_L, \infty)$,

$$p_n(a) = \mathbb{P}\left(Y_n^{(1)}(t) \geqslant na_1, \ldots, Y_n^{(L)}(t) \geqslant na_L\right).$$

A relevant related quantity is the probability of having reached the set A *before* t:

$$\mathbb{P}\left(\exists s \leqslant t : Y_n^{(1)}(s) \geqslant na_1, \ldots, Y_n^{(L)}(s) \geqslant na_L\right); \tag{4.16}$$

observe that this probability increases to 1 as $t \to \infty$. Alternatively, one could study the probability that all a_ℓ (for $\ell = 1, \ldots, L$) are exceeded before t, *but not necessarily at the same time*:

$$\mathbb{P}\left(\exists s_1 \leqslant t : Y_n^{(1)}(s_1) \geqslant na_1, \ldots, \exists s_L \leqslant t : Y_n^{(L)}(s_L) \geqslant na_L\right). \tag{4.17}$$

Powerful novel sample-path large deviations results by Budhiraja and Nyquist [19], which deal with a general class of multi-dimensional shot-noise processes, may facilitate the development of efficient importance sampling algorithms for non-modulated linear stochastic fluid networks. The results in [19] do not cover Markov modulation, though.

In the current setup of Section 4.4 the speed of the background process is kept fixed, i.e., not scaled by n. For modulated diffusions a sample-path large deviation principle has been recently established in [46] for the case that the background process is sped up by a factor n (which amounts to multiplying the generator matrix Q by n); the rate function decouples into (i) a part concerning the rare-event behaviour of the background process and (ii) a part concerning the rare-event behaviour of the diffusion (conditional on the path of the background process). With a similar result for the Markov-modulated linear stochastic fluid networks that we have studied in this chapter, one could potentially set up an efficient importance sampling procedure for the probabilities (4.16) and (4.17) under this scaling.

4.6 Appendix A

We here point out how (4.7) can be established; the line of reasoning is precisely the same as in the derivation of (4.6) in [30, Thm. 3.7.4]. First write

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) = \mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-2\vartheta^\star Y_n(t)} \mathrm{e}^{2n \log M(\vartheta^\star)} 1_{\{Y_n(t) \geqslant na\}}) = \\ \mathrm{e}^{-2nI(a)}\, \mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-2\vartheta^\star (Y_n(t) - na)} 1_{\{Y_n(t) \geqslant na\}}),$$

which, with $Z_n := (Y_n(t) - na)/\sqrt{n}$, equals

$$\mathrm{e}^{-2nI(a)}\, \mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-2\vartheta^\star Z_n \sqrt{n}} 1_{\{Z_n \geqslant 0\}}).$$

Observe that $\mathbb{E}_{\mathbb{Q}}\, Y_n = na$, due to the very choice of $\mathbb{Q}$. This entails that Z_n converges in distribution to a centred Normal random variable; as can be verified, the corresponding variance is τ (where τ is defined in (4.6)). Using the Berry-Esseen-based justification presented in [30, page 111], we conclude that, as $n \to \infty$,

$$\mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-2\vartheta^\star Z_n\sqrt{n}}1_{\{Z_n\geqslant 0\}}) \sim \int_0^\infty \mathrm{e}^{-2\vartheta^\star\sqrt{n}\,x}\,\frac{1}{\sqrt{2\pi\tau}}\mathrm{e}^{-x^2/(2\tau)}\mathrm{d}x.$$

Completing the square, the right-hand side of the previous display equals, with $\mathscr{N}(\mathrm{M},\mathrm{V})$ a normal random variable with mean M and variance V,

$$\mathrm{e}^{2(\vartheta^\star)^2 n\tau}\,\mathbb{P}\big(\mathscr{N}(-2\vartheta^\star\sqrt{n}\,\tau,\tau)>0\big) = \mathrm{e}^{2(\vartheta^\star)^2 n\tau}\,\mathbb{P}\big(\mathscr{N}(0,1)>2\vartheta^\star\sqrt{n\tau}\big).$$

Now we use the standard equivalence (as $x\to\infty$)

$$\mathbb{P}(\mathscr{N}(0,1)>x) \sim \frac{1}{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2},$$

to obtain

$$\int_0^\infty \mathrm{e}^{-2\vartheta^\star\sqrt{n}\,x}\,\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/(2\tau)}\mathrm{d}x \sim \frac{1}{\sqrt{n}}\,\frac{1}{2\vartheta^\star\sqrt{2\pi\tau}}.$$

Combining the above, we derive the claim:

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \frac{1}{\sqrt{n}}\,\frac{1}{2\vartheta^\star\sqrt{2\pi\tau}}\mathrm{e}^{-2nI(a)}.$$

We now proceed with the computations underlying (4.13). To this end, first observe that

$$L = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \mathrm{e}^{-\langle\boldsymbol{\vartheta}^\star,\boldsymbol{Y}_n(t)\rangle}\mathrm{e}^{n\log M(\boldsymbol{\vartheta}^\star)}.$$

As a consequence, in line with the above computation for the one-dimensional case,

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}(LI) &= \mathrm{e}^{-nI(\boldsymbol{b}^\star)}\mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-\langle\boldsymbol{\vartheta}^\star,\boldsymbol{Y}_n(t)-n\boldsymbol{a}\rangle}1_{\{\boldsymbol{Y}_n(t)\in A\}}),\\ \mathbb{E}_{\mathbb{Q}}(L^2 I) &= \mathrm{e}^{-2nI(\boldsymbol{b}^\star)}\mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-2\langle\boldsymbol{\vartheta}^\star,\boldsymbol{Y}_n(t)-n\boldsymbol{a}\rangle}1_{\{\boldsymbol{Y}_n(t)\in A\}}).\end{aligned}$$

It was proven in [23, Thm. 3.4] that

$$p_n(\boldsymbol{a}) = \mathbb{E}_{\mathbb{Q}}(LI) \sim \frac{1}{\sqrt{\tau}} \left(\prod_{i \in D} \vartheta_i^\star \right)^{-1} \big(2\pi n\big)^{-D/2} \mathrm{e}^{-nI(\boldsymbol{b}^\star)},$$

while at the same time

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \frac{1}{\sqrt{\tau}} \left(\prod_{i \in D} (2\vartheta_i^\star) \right)^{-1} \big(2\pi n\big)^{-D/2} \mathrm{e}^{-2nI(\boldsymbol{b}^\star)}.$$

This immediately leads to (4.13).

5

The stationary limit of a multidimensional stochastic recursion

In the previous chapters we have analysed both the probability of ever hitting a multidimensional rare set as well as the probability of hitting a multidimensional rare set at a fixed time. In this chapter, however, we analyse the stationary behaviour of a multidimensional stochastic process.

5.1 Introduction

Consider the stochastic recursion in dimension d given by

$$X_{n+1} = AX_n + B_n, \tag{5.1}$$

where X_n and B_n are stochastic column vectors of length d and A is a deterministic d-by-d matrix. Its stationary limit satisfies

$$X \stackrel{\mathrm{d}}{=} AX + B \tag{5.2}$$

when it exists. Of interest is the rare event that (a weighted sum of the components of) X exceeds some high level x. More specifically, we consider the case where the components of B have a light tail. One would expect, since A is

deterministic, that the components of X have a light tail as well.

A large amount of research has been dedicated to Equation (5.2), see e.g. [20]. In [13], various simulation problems in one dimension are considered, where the focus is on importance sampling. An older article is [42], where the asymptotic tail behaviour of X is analysed when B has a light tail, again in one dimension. A more recent paper is [2], where the stability of Equation (5.1) is investigated when (A_n, B_n) is modulated by a Markov chain.

It seems that in the multidimensional case, most of the work focuses on heavy tails, whereas work on the light-tailed settings seems to mostly be done in the one-dimensional case. The goal of this chapter is to provide results, for a certain class of cases, for the light-tailed case in higher dimensions.

In Section 5.2, we analyse the relatively simple case of dimension $d = 1$, in which A is a constant between zero and one in order to obtain stability. Furthermore, in an attempt to make the calculations easy to follow, we assume that B_n has an exponential distribution with rate parameter μ (independently of n). When we analyse the multidimensional model, we will weaken this assumption to that of the components of B having an exponentially decaying tail. As an exact analysis of the probability of interest seems prohibitive, it is natural to try to estimate this quantity through simulations. A first approach might be to apply crude Monte Carlo sampling. However, this is not feasible due to two reasons: First since the event of interest is rare, a high number of samples is needed before a good estimation is obtained, see e.g. [3, Chapter VI]. And second, it is not clear if and how a sample from the distribution of X can be generated. Therefore, we propose to use a *conditional* Monte Carlo procedure. We will show that this procedure solves the efficiency problem. However, the second problem of not being able to generate from the distribution of X, still persists. Therefore, we combine the conditional Monte Carlo procedure with the method from [63]. The key in using this method is that X can be written as an infinite stochastic sum.

The chapter continues in Section 5.3 with analysing the multidimensional counterpart of this model. The event of interest is now that of a weighted sum of the components of X exceeding a large threshold. As in the one-dimensional case, the goal is to construct an efficient simulation procedure that produces unbiased samples of the probability at hand. In order to use the method of [63] to obtain unbiased samples, it is needed that all components of X can, independ-

ently from each other, be written as an infinite sum, so that each component of X can be seen as the limit of an approximating sequence. However, in higher dimensions, the components of X might be dependent on each other through the matrix A. Therefore, to ensure that the desired infinite-sum representation still exists, a restriction on the matrix A is needed.

This chapter is organised as follows. Section 5.2 contains the analysis for the one-dimensional case. First, the model is presented. Furthermore, difficulties in analysing and simulating the tail probability of X are described. Section 5.2.1 and 5.2.2 provide solutions for efficient and unbiased simulation respectively. The section ends by presenting numerical results of an implementation of the simulation procedure as developed in the section. Section 5.3 follows the same set-up, but then for the multidimensional counterpart of the model in Section 5.2.

5.2 One dimension

Consider the one-dimensional stochastic recursion defined by $X_0 = 0$ and

$$X_{n+1} = aX_n + E_n, \qquad n \in \mathbb{Z}_+, \tag{5.3}$$

with $0 < a < 1$ and the E_n being i.i.d. copies of a random variable E, having an exponential distribution with rate parameter μ. The stationary distribution of the sequence $(X_n)_{n\in\mathbb{Z}_+}$ is given by letting $n \to \infty$ in Equation (5.3), which yields

$$X \stackrel{\mathrm{d}}{=} aX + E. \tag{5.4}$$

We want to analyse $p_x := \mathbb{P}(X > x)$ for large x. However, there seems to be no closed-form expression to calculate this probability exactly. For the exponential case, though, a lot is known for transforms, see e.g. [43] for a result on the so-called Mellin transform. This does not scale to higher dimensions, though. Therefore, other methods are needed to analyse this event. One method is to estimate p_x using crude Monte Carlo simulations. This method, however, has two problems:

1. For large x, the probability p_x is very small, hence a large number of simulation runs is required to get a good estimate.

2. It is not clear if and how samples can be taken from the distribution of X.

Below, we will first provide solutions to these problems separately. These solutions can then be easily combined into a simulation algorithm that can estimate p_x efficiently and unbiasedly.

5.2.1 Fast estimation

In order to solve the problem of crude Monte Carlo simulations requiring a large amount of runs, we will use a method called *conditional* Monte Carlo sampling. For now, we assume that we can sample from X. As the name suggests, this method involves conditioning on X. Note that

$$\begin{aligned}\mathbb{P}(X > x) = \mathbb{P}(aX + E > x) = \mathbb{P}(E > x - aX) \\ = \mathbb{E}(\mathbb{1}(E > x - aX)) = \mathbb{E}(\mathbb{E}(\mathbb{1}(E > x - aX)|X)).\end{aligned}$$

Hence, an unbiased estimator for p_x is

$$\mathbb{E}(\mathbb{1}(E > x - aX)|X). \tag{5.5}$$

This can be rewritten as

$$\mathbb{E}(\mathbb{1}(E > x - aX)|X) = \bar{F}(x - aX) = e^{-\mu(x-aX)} \wedge 1,$$

where we used that the complementary distribution function of E has a closed-form expression. Hence, in order to estimate p_x, we first take a sample from X and then substitute the value in the formula above.

We now proceed to show that this procedure is indeed "fast" in the sense that bounded relative error is achieved. However, we can only prove this for $a < \frac{1}{2}$. We will later show how this method can be adapted so that bounded relative error is obtained for all $a < 1$.

Theorem 5.2.1. *The simulation procedure described above has bounded relative error only for $a < \frac{1}{2}$.*

Proof. We first look at the first moment of the estimator. We can apply [31, Proposition 4.1], since $\mathbb{E}\left(e^{a\mu E}\right) < \infty$. We write $f(x) \sim g(x)$ if $\lim_{x\to\infty} f(x)/g(x) =$

C for some $C > 0$.

$$\begin{aligned}
\mathbb{P}\left(X > x\right) = \mathbb{E}\left(\bar{F}(x - aX)\right) &= \int_0^\infty \mathbb{P}\left(\bar{F}(x - aX) > z\right) \mathrm{d}z \\
&= \int_0^1 \mathbb{P}\left(e^{-\mu(x-aX)} > z\right) \mathrm{d}z \\
&= \int_0^1 \mathbb{P}\left(X > \frac{\ln(z) + \mu x}{\mu a}\right) \mathrm{d}z \\
&\sim \int_0^1 \mathbb{P}\left(E > \frac{\ln(z) + \mu x}{\mu a}\right) \mathrm{d}z \\
&= e^{-\mu x} + \int_{e^{-\mu x}}^1 e^{-\frac{\ln(z)+\mu x}{a}} \mathrm{d}z \\
&= e^{-\mu x} + e^{-\mu x/a} \int_{e^{-\mu x}}^1 z^{-1/a} \mathrm{d}z \\
&= e^{-\mu x} + e^{-\mu x/a} \frac{a}{-1+a}(1 - e^{-\mu x \frac{-1+a}{a}}).
\end{aligned}$$

Note that for all $0 < a < 1$, this decays like $e^{-\mu x}$. Assume now that $a \neq \frac{1}{2}$. Then, similarly, we get for the second moment

$$\begin{aligned}
\mathbb{E}\left(\bar{F}^2(x - aX)\right) &= \int_0^1 \mathbb{P}\left(e^{-2\mu(x-aX)} > z\right) \mathrm{d}z \\
&\sim \int_0^1 \mathbb{P}\left(E > \frac{\ln(z) + 2\mu x}{2\mu a}\right) \mathrm{d}z \\
&= e^{-2\mu x} + \int_{e^{-2\mu x}}^1 e^{-\frac{\ln(z)+2\mu x}{2a}} \mathrm{d}z \\
&= e^{-2\mu x} + e^{-\mu x/a} \frac{2a}{-1+2a}(1 - e^{-\mu x \frac{-1+2a}{2a}}).
\end{aligned}$$

Note first that $\frac{1}{a} + \frac{-1+2a}{2a} = \frac{1+2a}{2a}$. When $a < \frac{1}{2}$, this is larger than or equal to 2. Therefore, in this case, the second moment decays like $e^{-2\mu x}$. Comparing this to the first moment, as calculated above, we can conclude that in this case, bounded relative error is achieved. When $a > \frac{1}{2}$, the second moment decays like $e^{-\mu x/a}$. Hence, bounded relative error is not achieved.

Let's finally consider the case $a = \frac{1}{2}$. Then

$$\begin{aligned}
\mathbb{E}\left(\bar{F}^2(x - aX)\right) &= \int_0^1 \mathbb{P}\left(e^{-2\mu(x-aX)} > z\right) \mathrm{d}z \\
&\sim \int_0^1 \mathbb{P}\left(E > \frac{\ln(z) + 2\mu x}{2\mu a}\right) \mathrm{d}z \\
&= e^{-2\mu x} + \int_{e^{-2\mu x}}^1 e^{-\frac{\ln(z)+2\mu x}{2a}} \mathrm{d}z \\
&= e^{-2\mu x} + e^{-\mu x/a} 2\mu x = e^{-2\mu x}(1 + 2\mu x).
\end{aligned}$$

The factor $(1 + 2\mu x)$ makes that bounded relative error is not achieved. Note, however, that the weaker optimality notion of asymptotic optimality is achieved. □

We will now show that a similar result holds for all $a < 1$ if we change the procedure in the following way. Note that we can iterate Equation (5.3) so that we obtain

$$\begin{aligned}
X_{n+2} &= aX_{n+1} + E_{n+1} \\
&= a(aX_n + E_n) + E_{n+1} \\
&= a^2 X_n + aE_n + E_{n+1} \\
&= a^2 X_n + E_{n+1}^{(2)}
\end{aligned}$$

Since a $E_{n+1}^{(2)}$ has an exponentially decaying tail, this seems to suggest that we can find a conditional Monte Carlo procedure that attains bounded relative error when $a^2 \leq \frac{1}{2}$. If this is indeed the case, then by iterating Equation (5.3) further, it can be expected that an efficient Monte Carlo procedure can be constructed when $a^n \leq \frac{1}{2}$ for some $n \in \mathbb{N}$. This effectively gives us an efficient procedure for all $a < 1$. Before describing this procedure and proving its efficiency, we first give a lemma that will be used multiple times in this chapter.

Lemma 5.2.2. *Let $E_1, \ldots, E_n$ be independent non-negative random variables with an exponentially decaying tail, i.e., we have*

$$\mathbb{P}\left(E_i > x\right) = c_i e^{-\mu_i x}(1 + o(1)),$$

for some $\mu_i, c_i > 0$. *Define* $\mu := \min_{i=1,\dots,n} \mu_i$. *Then*

$$\mathbb{P}\left(\sum_{i=1}^{n} E_i > x\right) = ce^{-\mu x}(1 + o(1)),$$

for some $c > 0$.

Proof. In order to simplify the proof, we assume that all μ_i are unique. Assume further without loss of generality that the E_i are ordered by their decay parameter. We thus have $\mu = \mu_1 < \mu_2 \cdots < \mu_n$.

We will use an iterative argument. Assume first that we have two terms, i.e., $n = 2$. Define

$$X := e^{E_1}, \qquad Y := e^{\sum_{i=2}^{n} E_i}.$$

This allows us to write

$$\mathbb{P}\left(\sum_{i=1}^{n} E_i > x\right) = \mathbb{P}\left(XY > e^x\right).$$

Note that

$$\mathbb{E}\left(Y^{\mu}\right) < \infty,$$

by uniqueness of the μ_i, and that

$$\begin{aligned}
\mathbb{P}\left(X > x\right) &= \mathbb{P}\left(E_1 > \ln(x)\right) = c_1 x^{-\mu_1}(1 + o(1)), \\
\mathbb{P}\left(Y > x\right) &= \mathbb{P}\left(\sum_{i=2}^{n} E_i > \ln(x)\right) = c_2 x^{-\mu_2}(1 + o(1)).
\end{aligned}$$

Note that $c_1(1+o(1))$ is slowly varying and converges to c_1. And lastly, we have

$$\lim_{x\to\infty} \frac{\mathbb{P}\left(Y > x\right)}{\mathbb{P}\left(X > x\right)} = \lim_{x\to\infty} \frac{\mathbb{P}\left(\sum_{i=2}^{n} E_i > \ln(x)\right)}{\mathbb{P}\left(E_1 > \ln(x)\right)} = 0.$$

Therefore, we can apply Proposition 2.1 of [31] to conclude that

$$\begin{aligned}
\mathbb{P}\left(XY > e^x\right) &\sim \mathbb{E}\left(Y^{\mu}\right)\mathbb{P}\left(X > e^x\right) = \\
&\mathbb{E}\left(e^{\mu\sum_{i=2}^{n} E_i}\right)\mathbb{P}\left(E_1 > x\right) = c_1\mathbb{E}\left(e^{\mu\sum_{i=2}^{n} E_i}\right)e^{-\mu x}(1 + o(1)).
\end{aligned}$$

These steps can be repeated iteratively for $n = 3, 4, \ldots$ so that the conclusion of the lemma follows. □

Let $E^{(k)} \stackrel{\mathrm{d}}{=} \sum_{i=1}^{k} a^{i-1} E_i$. Note that

$$X_{(n+1)k} = a^k X_{nk} + E_n^{(k)}, \qquad n \in \mathbb{Z}_+,\ k \in \mathbb{N}. \tag{5.6}$$

The limit is given by

$$X \stackrel{\mathrm{d}}{=} a^k X + E^{(k)}, \qquad k \in \mathbb{N}. \tag{5.7}$$

Before we give the estimator, note that from Lemma 5.2.2 we can conclude that $\mathbb{P}\left(E^{(k)} > x\right) = e^{-\mu x}(C + o(1))$ for some $C > 0$. In fact, the exact expression can be calculated. The density of $E^{(k)}$, see e.g. [1, Theorem 2.1], is

$$\sum_{i=1}^{k} \frac{\mu_1 \cdots \mu_k}{\prod_{j=1, j\neq i}^{d} (\mu_j - \mu_i)} e^{-x\mu_i},$$

where $\mu_i := \mu / a^{i-1}$ (we use that $\mu_i \neq \mu_j$ for $i \neq j$). In line with the previous case, we use as estimator

$$\mathbb{E}\left(\mathbb{P}\left(X > x\right)|X\right) = \mathbb{E}\left(\mathbb{P}\left(a^k X + E^{(k)} > x\right)|X\right) = \\ \mathbb{E}\left(\mathbb{P}\left(E^{(k)} > x - a^k X\right)|X\right) = \bar{F}^{(k)}(x - a^k X),$$

where $\bar{F}^{(k)}$ is the complementary distribution function of $E^{(k)}$.

Theorem 5.2.3. *Let k be such that $a^k < \frac{1}{2}$. The simulation procedure described above has bounded relative error.*

Proof. Note that, for x large enough, it follows from Lemma 5.2.2 that there exist $0 < m \leq M < \infty$ such that $m e^{-\mu x} \leq \mathbb{P}\left(E^{(k)} > x\right) \leq M e^{-\mu x}$. As we don't want to show anymore that bounded relative error is not achieved when $a \geq \frac{1}{2}$, we present a proof that is more crude than the proof of Theorem 5.2.1. Note that

$$\mathbb{E}\left(\bar{F}^{(k)}(x - a^k X)\right) \geq m e^{-\mu x} \mathbb{E}\left(e^{a^k \mu X}\right).$$

Similarly, for the second moment we have

$$\mathbb{E}\left(\left[\bar{F}^{(k)}(x-a^kX)\right]^2\right) \leq M^2 e^{-2\mu x}\mathbb{E}\left(e^{2a^k\mu X}\right),$$

where we need that $a^k < \frac{1}{2}$ to ensure that $\mathbb{E}\left(e^{2a^k\mu X}\right) < \infty$.

□

It should be noted that, when implementing the described procedure, it is desirable to choose k as the *smallest* integer such that $a^k < \frac{1}{2}$, as each run requires sampling k exponentially distributed random variables. Likewise, for values of a close to 1, k will be high, so that the computational effort will be relatively high as well.

5.2.2 Unbiased sampling

In the previous part, we have seen that, if we can sample from the distribution of X, an efficient simulation algorithm exists. However, it is not clear if and how a sample from X can be obtained. We will remedy this by applying the algorithm from [63], which does not give us a sample from the distribution of X itself, but it will provide us with an unbiased estimator of p_x. Below, we will first briefly explain how the algorithm works in a general setting. Afterwards, we describe how the algorithm can be applied to our setting.

The method of [63] is used for estimating $\alpha := \mathbb{E}(Y)$ for some random variable Y, when samples from Y can not be generated in finite time. This can be resolved if one has access to a sequence $(Y_n)_{n\in\mathbb{Z}_+}$ of L^2-approximations, from which samples can be generated in finite time. We will then use one of the three estimators from [63], called the *single-term estimator*. Let $Y_{-1} := 0$ and let N be a non-negative discrete random variable for which $p_n := \mathbb{P}(N = n) > 0$ for all $n \in \mathbb{Z}_+$. The single-term estimator then is defined as

$$Z = \frac{Y_N - Y_{N-1}}{p_N}.$$

The result from [63] that we will use is Theorem 1, which states that if

$$\sum_{n=1}^{\infty} \frac{\mathbb{E}\left((Y_{n-1}-Y)^2\right)}{\mathbb{P}\left(N \geq n\right)} < \infty, \tag{5.8}$$

then Z is an unbiased estimator for α and

$$\mathbb{E}\left(Z^2\right) = \sum_{n=0}^{\infty} \frac{\mathbb{E}\left((Y_n - Y_{n-1})^2\right)}{p_n} < \infty, \tag{5.9}$$

where this second moment is given on page 1030 of [63].

We will now apply the algorithm to our setting. At the heart of applying the algorithm is iterating (5.3). Using that $X_0 = 0$, we can write

$$X_n = \sum_{k=0}^{n-1} a^{n-1-k} E_k \stackrel{\mathrm{d}}{=} \sum_{k=0}^{n-1} a^k E_k, \qquad n \in \mathbb{N}. \tag{5.10}$$

For the stationary distribution we then get

$$X \stackrel{\mathrm{d}}{=} \sum_{k=0}^{\infty} a^k E_k. \tag{5.11}$$

Using the estimator we used above, we can write

$$\begin{aligned} \mathbb{P}\left(X > x\right) &= \mathbb{E}\left(e^{-\mu(x-aX)} \wedge 1\right) \\ &= \mathbb{E}\left(e^{-\mu(x-a\sum_{k=0}^{\infty} a^k E_k)} \wedge 1\right). \end{aligned}$$

In the setting of [63], we have

$$Y = e^{-\mu(x-a\sum_{k=0}^{\infty} a^k E_k)} \wedge 1$$

and we set

$$Y_n := e^{-\mu(x-a\sum_{k=0}^{n} a^k E_k)} \wedge 1, \qquad n \in \mathbb{Z}_+.$$

By first using the upper bound on Y and by checking the three cases that can occur for the values of Y_n and Y (that is, $Y_n = Y = 1$ or $Y_n < 1$, $Y = 1$ or

$Y_n < 1$, $Y < 1$), it can be verified that

$$\begin{aligned}
\mathbb{E}\left((Y - Y_n)^2\right) &= \mathbb{E}\left(Y^2\left(1 - \frac{Y_n}{Y}\right)^2\right) \\
&\leq \mathbb{E}\left(1 \times \left(1 - \frac{Y_n}{Y}\right)^2\right) \\
&\leq \mathbb{E}\left(\left(1 - \frac{e^{-\mu(x - a\sum_{k=0}^{n} a^k E_k)}}{e^{-\mu(x - a\sum_{k=0}^{\infty} a^k E_k)}}\right)^2\right) \\
&= \mathbb{E}\left(\left(1 - e^{-a\mu\sum_{k=n+1}^{\infty} a^k E_k}\right)^2\right).
\end{aligned}$$

We now use that $(1 - e^{-x})^2 \leq x^2$ for $x \geq 0$ so that an upper bound for $\mathbb{E}\left((Y - Y_n)^2\right)$ is

$$\mathbb{E}\left(\left(a\mu \sum_{k=n+1}^{\infty} a^k E_k\right)^2\right) = \mathbb{E}\left(\left(a^{n+2}\mu^2 \sum_{k=0}^{\infty} a^k E_k\right)^2\right) = a^{2n+4}\mu^2 \mathbb{E}\left(X^2\right).$$

From [36, Proposition 8.4.3] we know that the second moment of X is finite. Hence, $\mathbb{E}\left((Y - Y_n)^2\right)$ converges to zero not slower than Ca^{2n} for some $C > 0$.

This result allows us to choose a suitable distribution for N. Let b such that $a^2 < b < 1$. We define N such that

$$\mathbb{P}\left(N \geq n\right) = b^n, \qquad n \in \mathbb{Z}_+,$$

i.e., N has a geometric distribution with success parameter $1 - b$ and support $\mathbb{Z}_+$. Then (5.8) is satisfied.

In order to analyse the efficiency of the estimator Z, we need to calculate its first two moments. Since the estimator is unbiased, its first moment equals p_x. We will show that the second moment (given in (5.9)) behaves like the second moment of the estimator used in the previous subsection. The steps used are

similar to showing how $\mathbb{E}\left((Y-Y_n)^2\right)$ behaves.

$$\begin{aligned}
\mathbb{E}\left(Z^2\right) &= \sum_{n=0}^{\infty} \frac{\mathbb{E}\left((Y_n - Y_{n-1})^2\right)}{p_n} \\
&= \sum_{n=0}^{\infty} \frac{\mathbb{E}\left(Y_{n-1}^2 \left(\frac{Y_n}{Y_{n-1}} - 1\right)^2\right)}{p_n} \\
&\leq \sum_{n=0}^{\infty} \frac{\mathbb{E}\left(Y_{n-1}^2 \left(\frac{e^{-\mu(x-a\sum_{k=0}^{n} a^k E_k)}}{e^{-\mu(x-a\sum_{k=0}^{n-1} a^k E_k)}} - 1\right)^2\right)}{p_n} \\
&\overset{*}{=} \sum_{n=0}^{\infty} \frac{\mathbb{E}\left(Y_{n-1}^2 \left(e^{a\mu a^n E_n} - 1\right)^2\right)}{p_n} \\
&= \sum_{n=0}^{\infty} \mathbb{E}\left(Y_{n-1}^2\right) \frac{\frac{1}{1-2a^{n+1}} - \frac{2}{1-a^{n+1}} + 1}{b^n(1-b)} \\
&\leq \sum_{n=0}^{\infty} \mathbb{E}\left(Y^2\right) \frac{\frac{1}{1-2a^{n+1}} - \frac{2}{1-a^{n+1}} + 1}{b^n(1-b)}.
\end{aligned}$$

Note that we need that $a < \frac{1}{2}$ to ensure that at the step with the asterisk, all terms are finite. We are done if we can show that the series above converges. The numerator of the fraction is equal to

$$\begin{aligned}
&\frac{1 - a^{n+1} - 2(1-2a^{n+1}) + (1-2a^{n+1})(1-a^{n+1})}{(1-2a^{n+1})(1-a^{n+1})} \\
&= \frac{1 - a^{n+1} - 2(1-2a^{n+1}) + (1-2a^{n+1})(1-a^{n+1})}{1 - 3a^{n+1} + 2a^{2n+2}} \\
&= \frac{2a^{2n+2}}{1 - 3a^{n+1} + 2a^{2n+2}}.
\end{aligned}$$

Since $a^2/b < 1$, we can conclude that

$$\mathbb{E}\left(Z^2\right) \leq C\mathbb{E}\left(Y^2\right),$$

for some C that is constant in x. Hence, from Theorem 5.2.1 we can conclude

that bounded relative error is achieved for $a < \frac{1}{2}$.

The next step would be to try to adapt this method in order to obtain bounded relative error for all $a < 1$. This can be done by combining the current method with the method as described in Section 5.2.1. The arguments used in the current section, although more tedious, carry over to this case.

5.2.3 Numerical results

In Figures 5.1, 5.2 and 5.3, we present the results of some numerical experiments. For the results of "unbiased" we used the combination of the methods described above, so that we get unbiased estimators that have bounded relative error for all $a < 1$. For "biased", we still used the conditional Monte Carlo method. However, we didn't use the method of [63] to obtain unbiased samples. Instead, we took the first hundred terms of the infinite-sum representation (5.11). More importantly, we didn't use the iterative method in order to obtain bounded relative error for all $a < 1$, which is the cause of the method requiring many runs. We stopped a simulation run when the 95% confidence interval had a precision of 10% or less, or when a certain number of runs were used (2000, 2000 and 10000 respectively). Refer to the captions for the numerical values of the parameters.

It can be seen that in the case that $a < \frac{1}{2}$ (i.e., Figure 5.1), our algorithm ("unbiased") performs worse (in number of runs) than the standard algorithm ("biased"). This can be explained by the extra variance that the algorithm of [63] induces. When $a > \frac{1}{2}$ (i.e., Figures 5.2 and 5.3), our algorithm outperforms the standard algorithm (where the iterative method is not used) as x grows, since $a < \frac{1}{2}$ is required to obtain bounded relative error. The variance induced by the method of [63] probably explains why our algorithm performs worse for small values of x.

5.3 Multiple dimensions

In this section, we will analyse the multidimensional counterpart of the one-dimensional model as described in the previous section. Let X_0 be the d-

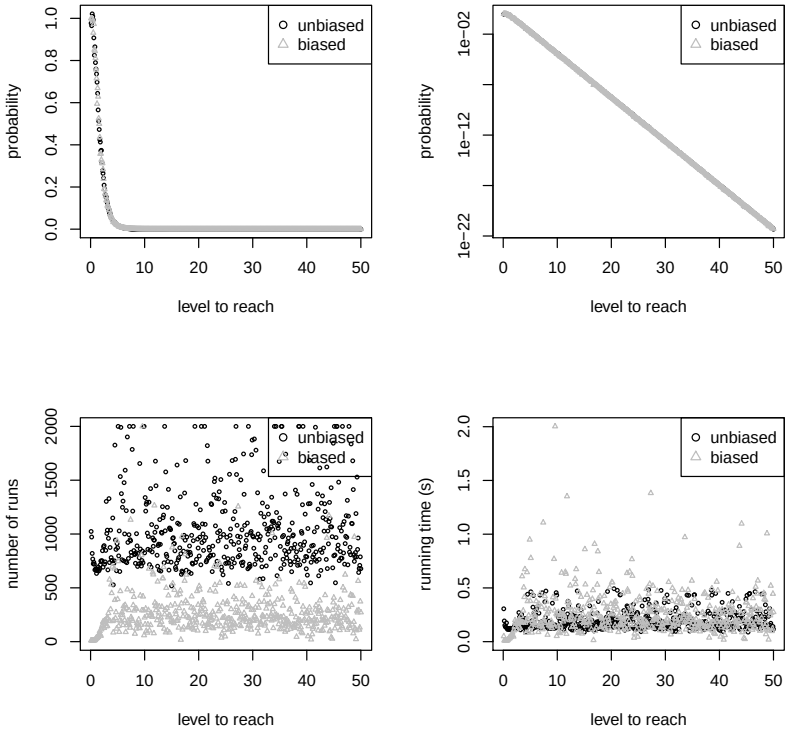

Figure 5.1: We set $a = 0.45$ and $\mu = 1$.

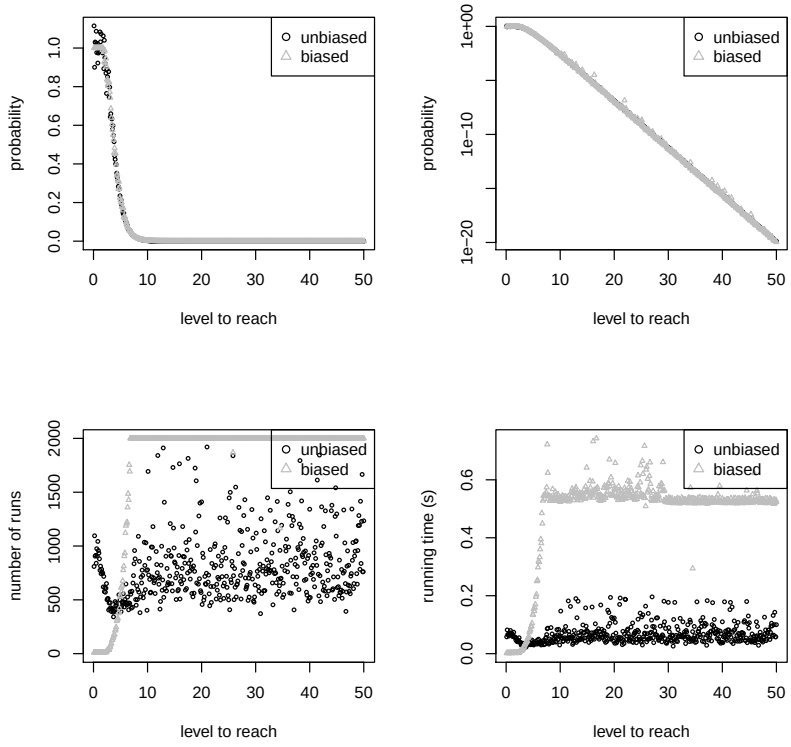

Figure 5.2: We set $a = 0.75$ *and* $\mu = 1$.

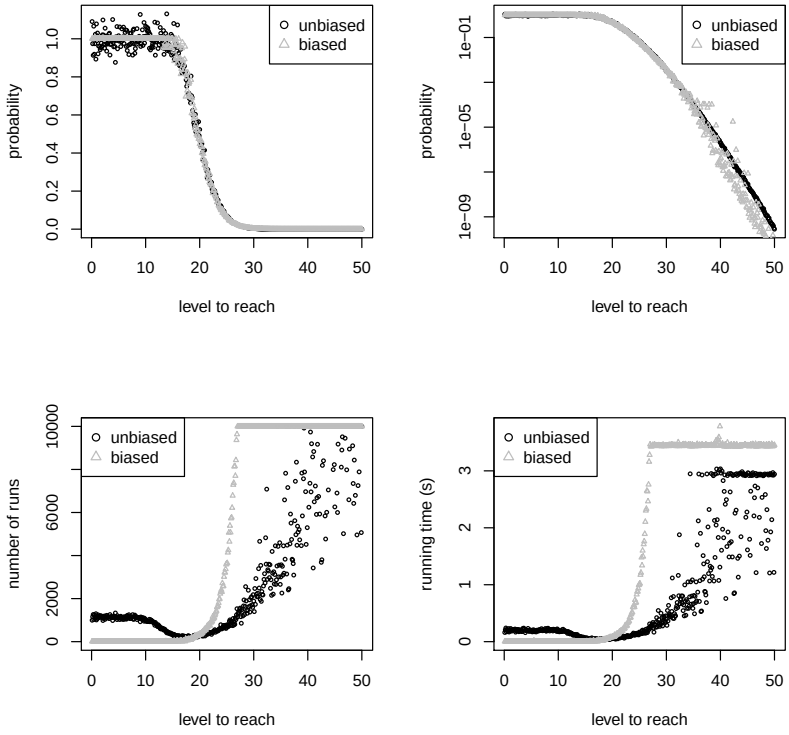

Figure 5.3: We set $a = 0.95$ and $\mu = 1$.

dimensional zero-vector and define recursively

$$X_{n+1} = AX_n + B_n, \qquad n \in \mathbb{Z}_+, \tag{5.12}$$

where A is a d-by-d (constant) matrix and B_n is a d-dimensional stochastic vector. Its stationary limit satisfies

$$X \stackrel{\mathrm{d}}{=} AX + B, \tag{5.13}$$

if it exists. From here on, we will assume that the stationary limit indeed exists, see e.g. [20, Section 4.1.1] for a discussion on necessary and sufficient conditions for existence of the limit. We will assume that the components of B, denoted by $B(1), \ldots, B(d)$, have an exponentially decaying tail, i.e., $\mathbb{P}(B(j) > x) = c_j e^{-\mu_j x}(1 + o(1))$ for $j = 1, \ldots, d$.

Let $u \geq 0$ be a d-dimensional row vector such that $\sum_{j=1}^{d} u(j) = 1$. The probability of interest is $\mathbb{P}(uX > x)$.

As in the one-dimensional case, we will make use of the following representation for X:

$$X \stackrel{\mathrm{d}}{=} \sum_{k=0}^{\infty} A^k B_k. \tag{5.14}$$

However, this representation has two problems that don't exist in the one-dimensional counterpart. First, this representation involves powers of A, which are hard to calculate in general. And second, we can not explicitly write the individual components of X in an infinite-sum representation, since they might depend on each other through A. We will see how this is a problem when trying to construct a method to get an unbiased estimator. To solve both of these problems, we will assume that A is diagonalisable, i.e., we can write $A = S\Lambda S^{-1}$, where Λ is a diagonal matrix containing the eigenvalues of A. This clearly solves the problem that powers of A are hard to calculate. We will see below how this solves the second problem as well. We will have to make one more assumption, namely that all eigenvalues of A are in absolute value smaller than one.

5.3.1 Fast estimation

As in the one-dimensional case, we will first show how bounded relative error can be achieved using a conditional Monte Carlo algorithm, assuming that we

can sample from the distribution of X.

Similarly to the one-dimensional case, the estimator that will be used is

$$\bar{F}\left(x - uAX\right), \tag{5.15}$$

where $\bar{F}(x) := \mathbb{P}\left(uB > x\right)$. Let $\mu^\star := \min_{1\leq j\leq d}\{\frac{\mu_j}{u(j)}\}$. From Lemma 5.2.2 we can conclude that there exists some $M > 0$ such that

$$\bar{F}\left(x\right) \leq Me^{-\mu^\star x}.$$

Hence, for the second moment of the estimator we get

$$\mathbb{E}\left(\bar{F}^2\left(x - uAX\right)\right) \leq M^2 e^{-2\mu^\star x}\mathbb{E}\left(e^{2\mu^\star uAX}\right).$$

If the moment generating function above exists, bounded relative error is indeed achieved. This is the case when the largest element of uA is smaller than $\frac{1}{2}$.

If this is not the case, however, we need to adapt the procedure. We will use the multidimensional counterpart of (5.6). Let $B^{(k)} \overset{\mathrm{d}}{=} \sum_{i=1}^{k} A^{i-1}B_k$. This allows us to write

$$X \overset{\mathrm{d}}{=} A^k X + B^{(k)}. \tag{5.16}$$

Let furthermore

$$\gamma := \min_{k\in\mathbb{Z}_+}\min_{j=1,\ldots,d}\frac{\mu_j}{(uA^k)(j)}. \tag{5.17}$$

Note that γ can be interpreted as the slowest decay rate of the tail probability of a (weighted) component of $B^{(k)}$ for all k.

Define $\bar{F}^{(k)}(x) := \mathbb{P}\left(uB^{(k)} > x\right)$.

Theorem 5.3.1. *Let $k^\star$ be larger than the optimiser in Equation* (5.17) *such that $(uA^{k^\star})(j) < \frac{1}{2}$ for $j = 1, \ldots, d$. Then the estimator*

$$\bar{F}^{(k^\star)}\left(x - uA^{k^\star}X\right)$$

has bounded relative error as $x \to \infty$.

Proof. Note that γ is defined such that there are $m, M > 0$ (see Lemma 5.2.2) such that

$$me^{-\gamma x} \leq \bar{F}^{(k^\star)}(x) \leq Me^{-\gamma x}.$$

Hence, for the first moment of the estimator we get

$$\mathbb{E}\left(\bar{F}^{(k^\star)}\left(x - uA^{k^\star}X\right)\right) \geq me^{-\gamma x}\mathbb{E}\left(e^{\gamma uA^{k^\star}X}\right)$$

and likewise we get for the second moment

$$\mathbb{E}\left((\bar{F}^{(k^\star)})^2\left(x - uA^{k^\star}X\right)\right) \leq M^2e^{-2\gamma x}\mathbb{E}\left(e^{2\gamma uA^{k^\star}X}\right).$$

□

5.3.2 Unbiased sampling

As in the one-dimensional case, we have shown that a fast simulation algorithm exists *if* we can sample from (X, Y). However, it is again not clear if and how this can be done. Therefore, we will show, in two dimensions, how the method of [63] can be applied.

In order to be able to use the estimator given in (5.15), we need to write all the components of X in infinite-sum representation. As mentioned above, this is not possible for general A, since these components might depend on each other through A. Therefore, we imposed the restriction that A needs to be in the set of diagonalisable matrices. Under this restriction, note that (5.13) can be rewritten as

$$X \stackrel{\mathrm{d}}{=} S\Lambda S^{-1}X + B.$$

If we pre-multiply both sides with S^{-1}, this is equivalent to

$$S^{-1}X \stackrel{\mathrm{d}}{=} \Lambda S^{-1}X + S^{-1}B.$$

Note that this can be seen as a stochastic equation for $\tilde{X} := S^{-1}X$, so that we can write

$$\tilde{X} \stackrel{\mathrm{d}}{=} \Lambda\tilde{X} + S^{-1}B.$$

Since Λ is diagonal, the components of $\tilde{X}$ *can* be written in infinite-sum representation. Indeed,

$$\tilde{X}(j) \stackrel{\mathrm{d}}{=} \sum_{n=0}^{\infty}\lambda_j^n(S^{-1}B_n)(j), \qquad j = 1, \ldots, d.$$

Moreover, we can combine this with (5.16). From

$$X \stackrel{\mathrm{d}}{=} S\Lambda^k S^{-1} X + B^{(k)},$$

we obtain

$$\tilde{X} \stackrel{\mathrm{d}}{=} \Lambda^k \tilde{X} + S^{-1} B^{(k)}.$$

We will use this in finding a suitable estimator. Let $k^\star$ be as in Theorem 5.3.1. We can write

$$\begin{aligned}
\bar{F}^{(k^\star)}\left(x - uA^{k^\star}X\right) &= \bar{F}^{(k^\star)}\left(x - uS\Lambda^{k^\star}S^{-1}X\right) \\
&= \bar{F}^{(k^\star)}\left(x - uS\Lambda^{k^\star}\tilde{X}\right) \\
&= \bar{F}^{(k^\star)}\left(x - uS\Lambda^{k^\star}\left(\textstyle\sum_{k=0}^{\infty}(\lambda_j^{k^\star})^k (S^{-1}B_k^{(k^\star)})(j)\right)_{j=1,\ldots,d}\right).
\end{aligned}$$

Define $\tilde{B}_n^{(k^\star)} := \left(\sum_{k=0}^{n}(\lambda_j^{k^\star})^k (S^{-1}B_k^{(k^\star)})(j)\right)_{j=1,\ldots,d}$ for $n \in \mathbb{Z}_+ \cup \{\infty\}$. In the setting of [63] (see Section 5.2.2 for a short overview of the method), we have that $Y = \bar{F}^{(k^\star)}\left(x - uS\Lambda\tilde{B}_\infty^{(k^\star)}\right)$ and we set $Y_n := \bar{F}^{(k^\star)}\left(x - uS\Lambda\tilde{B}_n^{(k^\star)}\right)$. We choose N such that $\mathbb{P}(N \geq n) = b^n$, with $0 < b < 1$ such that $b > \lambda_j^{2k^\star}$ for all j.

Theorem 5.3.2. *The procedure as described above satisfies* (5.8)*. Moreover, the procedure has bounded relative error.*

Proof. We will first show that (5.8) is satisfied. Note that $\bar{F}^{(k^\star)}(x) = Me^{-\gamma x}(1+o(1))$. Using that $Y_n \leq Y \leq 1$, we get

$$\begin{aligned}
\mathbb{E}\left((Y - Y_n)^2\right) &= \mathbb{E}\left(Y^2\left(1 - \frac{Y_n}{Y}\right)^2\right) \\
&\leq \mathbb{E}\left(1 \times \left(1 - \frac{Y_n}{Y}\right)^2\right) \\
&\sim \mathbb{E}\left(\left(1 - \frac{e^{-\gamma(x - uS\Lambda\tilde{B}_n^{(k^\star)})}}{e^{-\gamma(x - uS\Lambda\tilde{B}_\infty^{(k^\star)})}}\right)^2\right)
\end{aligned}$$

$$= \mathbb{E}\left(\left(1 - e^{-\gamma u S \Lambda(\tilde{B}_\infty^{(k^\star)} - \tilde{B}_n^{(k^\star)})}\right)^2\right).$$

Now note that $\tilde{B}_\infty^{(k^\star)} - \tilde{B}_n^{(k^\star)} \stackrel{\mathrm{d}}{=} \Lambda^{n+1}\tilde{B}_\infty^{(k^\star)}$. Hence, the last expression in the display above is equal to

$$\begin{aligned}
&= \mathbb{E}\left(\left(1 - e^{-\gamma u S \Lambda^{n+2}\tilde{B}_\infty^{(k^\star)}}\right)^2\right) \\
&= \mathbb{E}\left(\left(1 - e^{-\gamma u S \Lambda^{n+2}\tilde{X}}\right)^2\right) \\
&= \mathbb{E}\left(\left(1 - e^{-\gamma u S \Lambda^{n+2} S^{-1} X}\right)^2\right) \\
&= \mathbb{E}\left(\left(1 - e^{-\gamma u A^{n+2} X}\right)^2\right).
\end{aligned}$$

Define $\lambda^\star := \max\{|\lambda_j|\}$. Note that there exists some $s > 0$ such that

$$\max_{i,j}\{|S_{ij}|\} \times \max_{i,j}\{|S_{ij}^{-1}|\} \leq s$$

and therefore it follows that, since $u \geq 0$ and sums to unity,

$$(uA^n)_j \leq s(\lambda^\star)^n.$$

Therefore, the second moment above is smaller than or equal to

$$\mathbb{E}\left(\left(1 - e^{-\gamma s \lambda^{n+2} \max_{1\leq j\leq d}\{|X(j)|\}}\right)^2\right).$$

We again use that $(1 - e^{-x})^2 \leq x^2$ for $x \geq 0$, so that the following is an upper bound for the expression above:

$$\mathbb{E}\left(\left(\gamma s \lambda^{n+2} \max_{1\leq j\leq d}\{|X(j)|\}\right)^2\right).$$

It follows, from the choice of b, that (5.8) is indeed satisfied. □

5.3.3 Numerical results

In Figures 5.4 and 5.5, we present the results of some numerical experiments. We stopped a simulation run when the 95% confidence interval had a precision of 10% or less, or when 2000 runs were used. Refer to the captions for the numerical values of the parameters. In both cases, $k^\star = 1$.

In both figures it can be seen that the asymptotic regime is attained for low values of x. The bounded number of runs confirms that the method has bounded relative error.

5.4 Concluding remarks

We have shown that it is possible to develop asymptotic estimates, as well as computational algorithms for a class of multidimensional light-tailed non-Gaussian stochastic recursions. A drawback of our method is that it is restricted to deterministic matrices A, and it is of interest to develop methods to handle the case of random A as well. In addition, we think that the procedure of using conditional Monte Carlo can be useful in other rare event simulation problems as well. The essential feature that seems to make this approach work is that the rare event here is a single big jump of one of the summands, that occurs after $O(1)$ time, and it would be of interest to formalise this for other Markov chains.

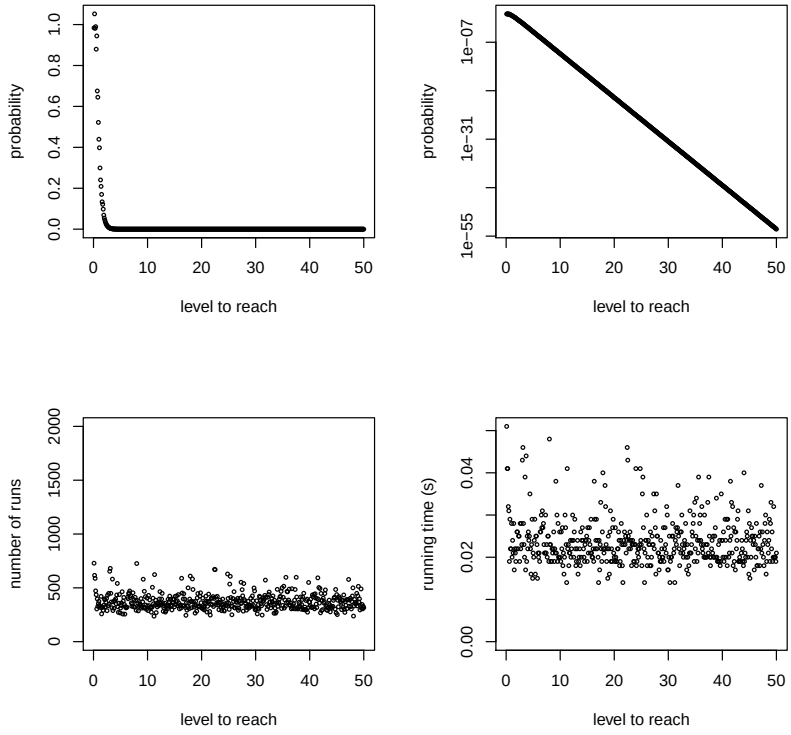

Figure 5.4: We used $A = \left(\begin{smallmatrix} 0.2 & 0 \\ 0.4 & 0.3 \end{smallmatrix}\right)$, $u = 0.4$, $v = 0.6$, $\mu_1 = 1$ *and* $\mu_2 = 2$.

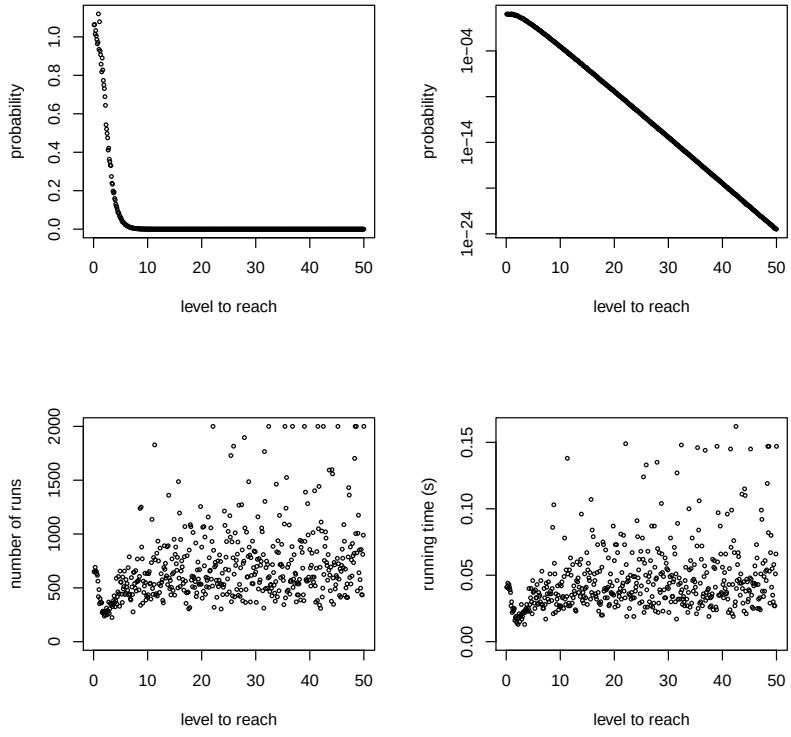

Figure 5.5: We used $A = \left(\begin{smallmatrix} 0.2 & 0 \\ 0.8 & 0.3 \end{smallmatrix}\right)$, $u = 0.4$, $v = 0.6$, $\mu_1 = 0.5$ *and* $\mu_2 = 0.7$.

Bibliography

[1] M. Akkouchi. On the convolution of exponential distributions. *J. Chungcheong Math. Soc*, 21(4):501–510, 2008.

[2] G. Alsmeyer and F. Buckmann. Stability of perpetuities in Markovian environment. *Journal of Difference Equations and Applications*, 23(4):699–740, 2017.

[3] S. Asmussen and P. W. Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer Science & Business Media, 2007.

[4] S. Asmussen and D. Kortschak. Error rates and improved algorithms for rare event simulation with heavy Weibull tails. *Methodology and Computing in Applied Probability*, 17(2):441–461, 2015.

[5] S. Asmussen and H. M. Nielsen. Ruin probabilities via local adjustment coefficients. *Journal of Applied Probability*, 32(3):736–755, 1995.

[6] E. Badila. *Queues and risk models.* PhD thesis, Technische Universiteit Eindhoven, 2015.

[7] R. R. Bahadur and R. R. Rao. On deviations of the sample mean. *Ann. Math. Statist*, 31(4):1015–1027, 1960.

[8] J. Biggins and A. Sani. *Extended Perron-Frobenius Results*. University of Sheffield, Department of Probability and Statistics, 2004.

[9] J. Blanchet. Optimal sampling of overflow paths in Jackson networks. *Mathematics of Operations Research*, 38(4):698–719, 2013.

[10] J. Blanchet and H. Lam. State-dependent importance sampling for rare-event simulation: An overview and recent advances. *Surveys in Operations Research and Management Science*, 17(1):38–59, 2012.

[11] J. Blanchet, K. Leder, and P. Glynn. Strongly efficient algorithms for light-tailed random walks: An old folk song sung to a faster new tune. In *Proceedings of the Eighth International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing (MCQMC 2008). Springer, Berlin*, pages 227–248, 2008.

[12] J. Blanchet and M. Mandjes. Rare event simulation for queues. In G. Rubino and B. Tuffin, editors, *Rare event simulation using Monte Carlo methods*, chapter 5. John Wiley & Sons, 2009.

[13] J. Blanchet and B. Zwart. Importance sampling of compounding processes. In *Simulation Conference, 2007 Winter*, pages 372–379. IEEE, 2007.

[14] J. Blom, K. De Turck, and M. Mandjes. Refined large deviations asymptotics for Markov-modulated infinite-server systems. *European Journal of Operational Research*, 259(3):1036–1044, 2017.

[15] J. Blom and M. Mandjes. A large-deviations analysis of Markov-modulated infinite-server queues. *Operations Research Letters*, 41(3):220–225, 2013.

[16] S. Borst, O. Boxma, and P. Jelenković. Coupled processors with regularly varying service times. In *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, volume 1, pages 157–164. IEEE, 2000.

[17] O. J. Boxma, E. J. Cahen, D. Koops, and M. Mandjes. Linear stochastic fluid networks: Rare-event simulation and Markov modulation. *arXiv preprint arXiv:1705.10273*, 2017.

[18] O. J. Boxma, E. J. Cahen, D. Koops, and M. Mandjes. Linear stochastic fluid networks: Rare-event simulation and Markov modulation. *Methodology and Computing in Applied Probability*, pages 1–29, 2018.

[19] A. Budhiraja and P. Nyquist. Large deviations for multidimensional state-dependent shot-noise processes. *Journal of Applied Probability*, 52(4):1097–1114, 2015.

[20] D. Buraczewski, E. Damek, T. Mikosch, et al. *Stochastic models with power-law tails*. Springer, 2016.

[21] E. J. Cahen, M. Mandjes, and B. Zwart. Rare event analysis and efficient simulation for a multi-dimensional ruin problem. *Probability in the Engineering and Informational Sciences*, 31(3):265–283, 2017.

[22] E. J. Cahen, M. Mandjes, and B. Zwart. Estimating large delay probabilities in two correlated queues. *ACM Trans. Model. Comput. Simul.*, 28(1):2:1–2:19, Jan. 2018.

[23] N. R. Chaganty and J. Sethuraman. Multidimensional strong large deviation theorems. *Journal of statistical planning and inference*, 55(3):265–280, 1996.

[24] J. Cohen. *The Single Server Queue*. North-Holland Series in Applied Mathematics and Mechanics. North-Holland, 1982.

[25] J. Collamore. Hitting probabilities and large deviations. *The Annals of Probability*, 24(4):2065–2078, 1996.

[26] R. B. Cooper. *Introduction to queueing theory*. Elsevier North Holland, Inc., second edition, 1981.

[27] H. Cramér. Sur un nouveau théoreme-limite de la théorie des probabilités. *Actual. Sci. Ind.*, 736:5–23, 1938.

[28] P.-T. de Boer. Analysis of state-independent importance-sampling measures for the two-node tandem queue. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 16(3):225–250, 2006.

[29] K. Dębicki, A. Dieker, and T. Rolski. Quasi-product forms for Lévy-driven fluid networks. *Mathematics of Operations Research*, 32(3):629–647, 2007.

[30] A. Dembo and O. Zeitouni. *Large deviations techniques and applications, Second edition*. Springer-Verlag, 1998.

[31] D. Denisov and B. Zwart. On a theorem of Breiman and a class of random difference equations. *Journal of Applied Probability*, 44(4):1031–1046, 2007.

[32] N. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Mathematical Proceedings of the Cambridge Philosophical Society*, 118(02):363–374, 1995.

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

[34] P. Dupuis and H. Wang. Importance sampling for Jackson networks. *Queueing Systems*, 62(1-2):113–157, 2009.

[35] A. Elwalid and D. Mitra. Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions on Networking (TON)*, 1(3):329–343, 1993.

[36] P. Embrechts, C. Kluppelberg, and T. Mikosch. *Modelling extremal events*. Springer, 1997.

[37] J. Gallier. Notes on convex sets, polytopes, polyhedra, combinatorial topology, Voronoi diagrams and Delaunay triangulations. Book in progress, `http://www.cis.upenn.edu/~jean/gbooks/convexpoly.html`, 2012.

[38] A. Ganesh, C. Macci, and G. L. Torrisi. A class of risk processes with reserve-dependent premium rate: sample path large deviations and importance sampling. *Queueing Systems*, 55(2):83–94, 2007.

[39] A. Ganesh, N. O'Connell, and D. Wischik. *Big queues*. Springer, 2004.

[40] P. Glasserman and S. Juneja. Uniformly efficient importance sampling for the tail distribution of sums of random variables. *Mathematics of Operations Research*, 33(1):36–50, 2008.

[41] P. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability*, 31:131–156, 1994.

[42] C. M. Goldie and R. Grübel. Perpetuities with thin tails. *Advances in Applied Probability*, 28(2):463–480, 1996.

[43] F. Guillemin, P. Robert, and B. Zwart. AIMD algorithms and exponential functionals. *The Annals of Applied Probability*, 14(1):90–117, 2004.

[44] M. Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013.

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

[46] G. Huang, M. Mandjes, and P. Spreij. Large deviations for Markov-modulated diffusion processes with rapid switching. *Stochastic Processes and their Applications*, 126(6):1785–1818, 2016.

[47] S. Juneja and P. Shahabuddin. Rare-event simulation techniques: an introduction and recent advances. *Handbooks in operations research and management science*, 13:291–350, 2006.

[48] O. Kella. Parallel and tandem fluid networks with dependent Lévy inputs. *The Annals of Applied Probability*, 3(3):682–695, 1993.

[49] O. Kella and W. Stadje. Markov-modulated linear fluid networks with Markov additive input. *Journal of Applied Probability*, 39(2):413–420, 2002.

[50] O. Kella and W. Whitt. A tandem fluid network with Lévy input. In U. Bhat and I. Basawa, editors, *Queueing and related models*, chapter 7, pages 112–128. Oxford University Press, 1992.

[51] O. Kella and W. Whitt. Linear stochastic fluid networks. *Journal of Applied Probability*, 36(1):244–260, 1999.

[52] G. Kesidis, J. Walrand, and C.-S. Chang. Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking (TON)*, 1(4):424–428, 1993.

[53] D. T. Koops, O. J. Boxma, and M. R. H. Mandjes. A tandem fluid network with Lévy input in heavy traffic. *Queueing Systems*, 84(3):355–379, Dec 2016.

[54] K. Kosiński and M. Mandjes. Logarithmic asymptotics for multidimensional extremes under nonlinear scalings. *Journal of Applied Probability*, 52(1):68–81, 2015.

[55] E. Krichagina and A. Puhalskii. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems*, 25(1):235–280, 1997.

[56] J. Kuhn, M. Mandjes, and T. Taimre. Exact asymptotics of sample-mean-related rare-event probabilities. *Probability in the Engineering and Informational Sciences*, 32(2):207–228, 2018.

[57] P. L'Ecuyer, J. H. Blanchet, B. Tuffin, and P. W. Glynn. Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 20(1), 2010.

[58] P. Lieshout and M. Mandjes. Tandem Brownian queues. *Mathematical Methods of Operations Research*, 66(2):275–298, 2007.

[59] M. Mandjes. Packet models revisited: tandem and priority systems. *Queueing Systems*, 47(4):363–377, 2004.

[60] M. Mandjes and A. Ridder. Finding the conjugate of Markov fluid processes. *Probability in the Engineering and Informational Sciences*, 9(02):297–315, 1995.

[61] A. Mogulskii. Large deviations for random finite-dimensional walks. *Theory Probab. Appl.*, 21:309–323, 1976.

[62] J. A. Morrison. Two queues with vastly different arrival rates and processor-sharing factors. *Queueing Systems*, 64(1):49–67, 2010.

[63] C.-H. Rhee and P. W. Glynn. Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043, 2015.

[64] G. Rubino and B. Tuffin. *Rare event simulation using Monte Carlo methods*. John Wiley & Sons, 2009.

[65] J. S. Sadowsky. Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *Automatic Control, IEEE Transactions on*, 36(12):1383–1394, 1991.

[66] M. Saxena, O. Boxma, S. Kapodistria, and R. Núñez Queija. Two queues with random time-limited polling. *Probability and Mathematical Statistics*, 37(2):257–289, 2017.

[67] E. Seneta. *Non-negative matrices*. George Allen & Unwin, 1973.

[68] A. D. Sezer. Importance sampling for a Markov modulated queuing network. *Stochastic Processes and Their Applications*, 119(2):491–517, 2009.

[69] D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics*, 4(4):673–684, 1976.

Summary

Rare-event simulation for multidimensional stochastic models

Stochastic models are often used to model real world phenomena. However, many of these models are too complicated to analyse exactly. Therefore, simulation is often used in order to estimate the quantities of interest. As the focus of this thesis is on rare events, naive Monte Carlo sampling is not a good method, as it requires a high number of runs, and thus a high running time. Therefore, we focus on finding efficient rare-event simulation methods, using importance sampling, partitioned importance sampling and conditional Monte Carlo sampling.

Many results already exist for one-dimensional models. However, these results rarely carry over to a multidimensional setting, which is much harder to analyse. Therefore, and because of their ability to model real world processes better, we focus on rare-event estimation for multidimensional models. More specifically, we focus on two-dimensional models, as they are easier to analyse, and the results carry over to higher dimension easier than in the case from one dimension to two.

Several rare-event simulation algorithms are developed. Their efficiency is illustrated both by using theoretical means as well as by showing the results of numerical experiments. Furthermore, a number of large deviations results are presented as well.

This thesis starts by introducing various mathematical concepts that are

required to understand the rest of the content. The introduction contains theory on rare event simulation, queueing processes, random walks and large deviations theory. It also discusses some of the complications that can arise when using rare event simulation techniques in multiple dimensions.

The first results are presented in Chapter 2. There, we analyse a two-dimensional ruin probability, which is the event that both components of a stochastic process exceed a high level at the same time. We provide several expressions for the decay rate of this probability. Furthermore, we give an importance sampling algorithm for a special case of the model at hand, and show that it is asymptotically efficient. This result is supported by several numerical experiments.

Chapter 3 is closely related to Chapter 2. The event of interest is again that of both components of a stochastic process exceed a high level. The difference with the previous chapter is that now the components of the process are allowed to reach the high level at different times. Again, an expression for the decay rate is given. We then show that a 'naive' implementation of importance sampling is not necessarily asymptotically efficient. We introduce a technique, which we call partitioned importance sampling, and we show that it is indeed asymptotically efficient. Also, this result is illustrated by numerical experiments.

In Chapter 4 we consider a linear stochastic fluid network under Markov modulation. We analyse the probability of the event that the joint storage level ever attains a value in some rare set. We develop an importance sampling algorithm and provide its efficiency properties in various settings. Some numerical results are shown as well.

Chapter 5 focuses on the stationary distribution of a stochastic recursion. The goal is to estimate the probability that the stationary process has a large value. We provide a conditional Monte Carlo algorithm that can estimate this probability both efficiently (we show that bounded relative error is attained) and unbiased in finite running time.

Dankwoord – Acknowledgments

Met het schrijven van dit dankwoord komt er een einde aan ruim vier jaar onderzoek. Het maken van een proefschrift is geen gemakkelijke opgave, maar ik kan vooral terugkijken op een heel leuke tijd. Dit komt ook omdat ik de afgelopen vier jaar, direct en indirect, veel hulp heb gekregen van diverse personen. Ik wil dan ook gebruik maken van deze gelegenheid om mijn dank te betuigen aan iedereen die me heeft ondersteund.

Als eerste wil ik mijn begeleiders bedanken. De afgelopen jaren heb ik met veel plezier met jullie samengewerkt. Jullie hebben me, in goede en in minder goede tijden, altijd het vertrouwen gegeven, waarvoor veel dank. Ik herinner me ook de besprekingen die we regelmatig met z'n drieën hadden, waarin jullie vaak tegen het einde, op nuchtere wijze, de ins en outs van de academische wereld bespraken. Bert, jouw kennis en doorzettingsvermogen zijn een grote inspiratie voor mij. Als ik ergens niet uitkwam wist jij me altijd op weg te helpen met een nieuwe aanpak en jouw vermogen om alle literatuur die ook maar enigszins relevant is uit je hoofd te kennen is indrukwekkend. Michel, jouw optimisme en betrokkenheid hebben me altijd weer weten te motiveren. Jouw enthousiasme voor de wiskunde uitte zich vaak in het direct uitproberen van nieuwe ideeën op het bord. Al met al had ik me geen betere begeleiders kunnen wensen.

Ook wil ik alle leden van de promotiecommissie, Daan Crommelin, Ad Ridder, Ivo Adan, Sem Borst en Onno Boxma bedanken voor hun bereidheid om deel te nemen aan de commissie. Dankzij jullie commentaar is dit proefschrift aanzienlijk verbeterd.

My project was part of an NWO TOP-project. I want to thank all the people involved with this project, Onno Boxma, Stella Kapodistria, David Koops, Michel Mandjes, Sindo Núñez-Queija, Mayank Saxena and Bert Zwart, for the many insightful meetings that we had. In particular, I want to thank Onno and David, who are co-authors of the paper on which Chapter 4 is based.

My research took place in the Stochastics group of the CWI, Amsterdam. I want to thank all people who are and were part of this group for the many breaks we spent together, for the countless hours of playing table tennis, for introducing me to bouldering and for their support. Many thanks also go to the colleagues of other groups and institutions for their pleasant company. A special thanks goes to my roommate Bart for our many discussions, for his willingness to listen to my problems and for the many fun moments that we shared.

Muziek is een van de mooiste dingen die er is en misschien wel de mooiste toepassing van wiskunde die bestaat. Daarom wil ik Harmonie Orkest Amstelveen bedanken. De repetities op donderdagavond en de vele optredens boden aangename momenten van onstpanning, in het bijzonder als datgene wat we wisten te produceren "grenst aan het muzikale".

Naast muziek maken zijn er andere bezigheden die het leven een stukje prettiger maker. Daarom wil ik spellenclub Tyche bedanken voor de gezellige maandagavonden en wil ik stichting Kinderland Amstelveen danken voor de fijne zomers.

Ik voel mij vereerd met de fijne personen die ik mijn vrienden mag noemen. Jullie weten me altijd op te vrolijken en ik heb dan ook fijne herinneringen aan de vele spelletjesdagen, de feestjes, de filmavonden, het samen koken (en afwassen) en de vele andere momenten. Jullie worden daarvoor hartelijk bedankt!

I am very grateful for my paranymphs Bohan and Marjolein for their support and friendship throughout the years. We have spent many fun moments together and I have always been able to count on you. I am very happy to have you at my side during my defence.

Tot slot wil ik het woord richten tot mijn broertje, vader en moeder. Jullie hebben me altijd gesteund en stonden altijd voor me klaar met veel liefde, een luisterend oor, advies en chocola. Ik wil jullie daar dan ook voor bedanken en zeggen dat ik heel veel van jullie hou.

Ewan Cahen
Amsterdam, december 2018

About the author

Ewan was born on 17 December 1990 in Amsterdam, the Netherlands. He obtained his atheneum diploma at the Keizer Karel College in Amstelveen in 2009. From 2009 to 2012, Ewan proceeded to study Mathematics at the Vrije Universiteit Amsterdam. For the results of the first year of his study, Ewan was awarded a Young Talent Incentive Award from the Royal Holland Society of Sciences and Humanities. Ewan obtained his bachelor's degree cum laude. Subsequently, he completed the master's programme Mathematics at the same university in 2014. In September 2014, he started a PhD project at Centrum Wiskunde & Informatica in the Stochastics group under the supervision of Michel Mandjes and Bert Zwart. The results of this project are presented in this dissertation.