Rare Events in Heavy-tailed Stochastic Systems: Algorithms and Analysis

A Thesis

Submitted to the
Tata Institute of Fundamental Research, Mumbai
for the Doctor of Philosophy in Computer and Systems Sciences

by

Karthyek Rajhaa Annaswamy Murthy

School of Technology and Computer Science
Tata Institue of Fundamental Research
Mumbai

April, 2015
Final Version Submitted in August, 2015

DECLARATION

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions. The work was done under the guidance of Professor Sandeep Juneja, at the Tata Institute of Fundamental Research, Mumbai.

Karthyek Rajhaa Annaswamy Murthy

In my capacity as supervisor to the candidate's thesis, I certify that the above statements are true to the best of my knowledge.

Sandeep Juneja

Date:

Abstract

Events that occur with low probability are called rare events. In the design of probabilistic models, it is important to ensure that certain undesirable events, for example, bankruptcy of financial institutions (or) data loss in communication networks, do not happen (or) if at all they happen, they happen with low probability. Conventional theory of large deviations, which is usually used to explain the behaviour of such rare events, is inapplicable when the random variables involved are heavy-tailed. We call a random variable to be heavy-tailed if its tail probabilities decay slowly at some polynomial rate. Simulation of rare events, in particular, has been traditionally considered difficult when heavy-tailed random variables are involved. In this work, we analyse asymptotic behaviour of certain rare events involving heavy-tailed random variables and develop simulation algorithms that are provably efficient. The major contributions of this dissertation are as follows:

1) Based on the "big jump principle", we develop an entirely new methodology for the estimation of various large deviations and level crossing probabilities that arise in the context of heavy-tailed sums. Our key contribution has been to question the prevailing view that one needs to resort to state-dependent methods when a large number of heavy-tailed random variables are involved. The algorithms we develop are provably efficient, follow a general template, and are easy to generalize, as we shall see, to settings more general than sums of independent random variables. In addition to simulation involving finite sums, we tackle the problem of estimating tail probabilities of sums of infinite sums, where bias is generally a problem. Apart from eliminating bias, we show that our proposed algorithm for this problem is able to compute tail probabilities with uniformly bounded computational effort.

2) Building up on the theory of rare events in heavy-tailed sums, we attempt to explain how large delays in service happen in multi-server queues when incoming jobs are heavy-tailed (in size) and traffic intensity is an integer. In the current literature, characterizations of steady-state delay exist only when the traffic intensity is not an integer. There are qualitative reasons, as we shall see, that makes the integer case more delicate to analyse. Our main contribution is that we develop the first known tail asymptotic for steady-state delay in multi-server queues when the traffic intensity is an integer. Specifically, we consider a two-server queue and identify interesting transitions in the tail behaviour of steady-state delay via a careful analysis that is not typical in the analysis of queuing systems.

Acknowledgements

First, I thank my advisor Professor Sandeep Juneja for taking me as his student, for choosing excellent textbooks to teach, and for considerately mentoring me. Research has been very enjoyable, primarily because of the discussions I have had with him and Professor Jose Blanchet. This thesis would not have been possible without their direction, encouragement and enthusiasm to discuss.

Next, I thank my thesis and synopsis committee members and anonymous journal referees who have reviewed the material that has gone into this thesis at various stages. I also register my thanks to ICERM at Brown University for the wonderful thematic semester on computational probability. I gratefully acknowledge the support of IBM through its International Fellowship for PhD students.

I thank Professor Prahladh Harsha for always being available for friendly advice, and Professor Jaikumar Radhakrishnan for teaching me how to teach. There could not have been a better way to begin life in TIFR than doing a course on probability under Jaikumar.

While my time in TIFR started well with the courses, it continued to be great and enjoyable mainly because of friends. Of several great friends I happened to make in TIFR, special mention to 'mera bhai' Sarat, 'perennial office mate' Sagnik, Rakesh, and 'gym instructor' Swapnil, for I have learnt a lot from them. My visits to Columbia University and ICERM would not have been so memorable had I not had the pleasure of spending time with Arjun and Arnab.

I sincerely thank my surgeon Dr. Parag Munshi for me to be able to walk again after the debilitating ACL tear I sustained during my final year in TIFR. I thank my mother, all my friends, Sandeep, and TIFR Medical Section for helping me navigate through those painful 6 months.

I express my heartfelt thanks to my parents for letting me pursue my goals despite their hardships. When short of inspiration, I have not had a need to look beyond my mother. I thank my Akka, Athimber and Kishore for their support.

And finally, I reserve my special thanks to TIFR, for being the wonderful place it is, to sit and think Mathematics.

To

Ammā and Appā

Contents

1 Introduction **2**

2 Simulation of Rare Events: Mathematical Preliminaries **9**

2.1 Simulation algorithms for estimation of rare probabilities 10

2.2 Regular variation and heavy tails 14

3 Estimation of Large Deviation Probabilities **18**

3.1 Limit theorems for sums of regularly varying random variables 18

3.2 The simulation problem 20

3.3 Simulation of $\{S_n > b\}$: An importance sampling algorithm 23

3.4 Simulation of $\{S_n > b\}$: Conditional Monte Carlo 28

3.5 A numerical example 30

3.6 Proofs of auxiliary results 30

4 Estimation of Level Crossing Probabilities **36**

4.1 Simulation Methodology for $\{\tau_b < \infty\}$ 38

4.2 Proofs of key theorems 49

4.3 Simulation of $\{\tau_b < \tau\}$ 59

4.4 Numerical Experiments 62

4.5 Proofs of auxiliary results 62

5 Estimation of Tail Probabilities for Infinite Series **71**

5.1 Simulation Methodology 72

5.2 Analysis of Variance of $Z(b)$ 79

5.3 A note on computational complexity of the simulation procedure 88

5.4 A numerical example 88

5.5 Proofs of auxiliary results 89

6 Concluding Remarks on Simulation of Rare Events in Heavy-tailed Sums **91**

6.1 Summary of techniques 91
6.2 An example with non-i.i.d. sums 93

7 Tail probabilities for large delays in half-loaded two-server queues 103
7.1 The main result and its intuition 106
7.2 Proof of lower bound 113
7.3 Proof of upper bound 117
7.4 Lyapunov bound techniques for a uniform bound on $\mathbb{P}_{\mathbf{w}}\{\tau_b^{(2)} < \tau_0\}$ 129
7.5 Another proof of the upper bound 134
7.6 Proofs of auxiliary results 144
7.7 Concluding remarks 149

Bibliography 150

1 Introduction

Recent developments in various branches of science and technology are witnessing an increasing use of probabilistic models. These models happen to be of varied levels of sophistication based on how representative they are of the phenomenon they attempt to model. Usually, probabilistic models are specified by probability distributions over various stochastic elements in the system. For example, if the variables $X_1, X_2, \ldots, X_n$ denote the collection of stochastic components in a system, then the joint probability distribution given by $\mathbb{P}\{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}$ for every $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ completely determines its stochastic behaviour. It turns out that the rate at which the tail probabilities $\mathbb{P}\{X_j > x\}$ of these random variables decay (with respect to x) dictate the behaviour of the system, often to an unexpected extent. The following example serves as an illustration.

Example 1.1. Let $X_1, \ldots, X_n$ be independent copies of a generic random variable X having zero mean. Let

$$S_n = X_1 + \ldots + X_n$$

denote their sum. Then under mild regularity conditions on the distribution of X, it is well-known due to central limit theorem that the scaled sum S_n/n^β, for some $\beta > 0$, converges in distribution to a non-degenerate random variable, as $n \to \infty$. Here the scaling constant β depends on the rate at which the tail probabilities $\bar{F}(x) := \mathbb{P}\{X > x\}$ decay:

1) If X has an exponentially decaying tail, it is well-known that β equals 0.5. This corresponds to the traditional convergence of $S_n/\sqrt{n}$ to a normal distribution. An implication of this statement is that, given $\delta > 0$, there exists a positive constant C_δ such that S_n lies between $-C_\delta\sqrt{n}$ and $C_\delta\sqrt{n}$ with probability larger than $1 - \delta$.

2) On the other hand, if

$$\bar{F}(x) = \begin{cases} (x+3)^{-\frac{3}{2}} & \text{if } x \geq -2, \\ 1 & \text{otherwise} \end{cases}$$

it can be shown that no constant C_δ satisfying

$$\mathbb{P}\{S_n \in (-C_\delta\sqrt{n}, C_\delta\sqrt{n})\} \geq 1-\delta, \text{ for every } n,$$

exists. Additionally, it is known that β equals $2/3$ in this case. In other words, for every $\delta > 0$, one can find a positive constant c_δ such that the sum S_n lies in the interval $(-c_\delta n^{\frac{2}{3}}, c_\delta n^{\frac{2}{3}})$ with probability at least $1-\delta$.

To understand the possible implications of Example 1.1, consider the stylized example of a bridge which gets damaged due to n independent risk factors that are roughly identical to each other. The bridge under consideration will collapse if the cumulative effect of these risks is larger than a threshold value which the bridge can withstand. Here, while trying to estimate the cumulative effect of the risks if the engineers involved in the design of the bridge ignore the tail probabilities of the risk factors and use the traditional convergence to normal distribution, there is a possibility of gross underestimation of total risk to be $O(\sqrt{n})$, whereas the actual risk could be $O(n^{2/3})$ as in Example 1.1. The central limit theorem is one of the fundamental results in probability theory, and it describes how large the cumulative effect of large number of independent random variables tend to be. It is rather surprising that the rate of decay of tail probabilities play a deciding role in a result as fundamental as the central limit theorem. Similar to this example, there has been a sharp contrast in the nature of results obtained and techniques used based on how heavy the tail probabilities of random variables involved are.

A random variable X is said to be (right) heavy-tailed if $\mathbb{E}[\exp(\theta X)]$ is infinite for every $\theta > 0$; otherwise X is considered light-tailed. In other words, X is heavy-tailed if $\bar{F}(x) = \mathbb{P}\{X > x\}$ decreases to 0 at a sub-exponentially slower rate (with respect to x). Statistical analysis reveals that heavy-tailed distributions are very common in practice: for example, they have become indispensable in the analysis of computer and communication networks, financial risk analysis, social and collaborative network graphs, epidemics, physics and economics. In particular, heavy-tailed increments with infinite variance are a convenient means to explain the long-range dependence observed in tele-traffic data, and to model highly variable claim sizes in insurance settings. Popular references to this strand of literature include Embrechts et al. [1997], Resnick [1997] and Adler et al. [1998].

This dissertation broadly deals with tail behaviours of stochastic systems that involve heavy-tailed random variables whose tail probabilities decay at a polynomial rate. To be specific, we are concerned with developing computational algorithms and deriving analytical bounds for tail probabilities in various settings including heavy-tailed sums and multi-server queues.

In many stochastic models, it often happens that the effects of random variables $X_1, \ldots, X_n$ are cumulative in nature, and hence sums of random variables (or) convolutions of distributions turn out to be a common feature in their analysis. However, exact analysis of n-fold convolutions involve evaluation of n-dimensional integrals which are intractable (both analytically and computationally) when n is large. In such cases, one resorts to approximate evaluation of integrals, and one convenient way of doing this is via Monte Carlo simulation. For example, if the interest is in computing the probability $\mathbb{P}\{S_n > b\}$ for some $b \in \mathbb{R}$ within a relative accuracy of $\pm\epsilon$, then one can simply perform the following two steps:

1) Generate N independent realizations $S_n^{(1)}, \ldots, S_n^{(N)}$ of S_n. A sample of S_n, for example, can be generated by summing n independent copies of X.

2) Return the average $N^{-1} \sum_{i=1}^{N} \mathbb{I}\left(S_n^{(i)} > b\right)$ as an estimate for $\mathbb{P}\{S_n > b\}$.

It follows from the law of large numbers that the average returned by the above procedure converges to $\mathbb{P}\{S_n > b\}$ as the number of simulation runs N grow to infinity. Further, given $\delta > 0$, it can be shown that if one takes

$$N > \frac{1 - \mathbb{P}\{S_n > b\}}{\delta \epsilon^2 \mathbb{P}\{S_n > b\}},$$

then the estimate returned by the above procedure remains within $\pm\epsilon$ relative precision of $\mathbb{P}\{S_n > b\}$ with probability as high as $1 - \delta$. The scaling of computational effort needed here with respect to ϵ and δ is quite standard in the context of Monte Carlo algorithms. However, the quantity $\mathbb{P}\{S_n > b\}$ appearing in the denominator can be extremely small for large values of b: for example, if $\mathbb{P}\{S_n > b\}$ decays exponentially fast with respect to b as $b \to \infty$, then the number of simulation runs N required scales exponentially. Events that occur with low probability are called rare events. Here, $\{S_n > b\}$ is a rare event for large values of b. The appearance of the probability of rare event in the denominator is a reminder of the fact that one would have do at least a million experiments to witness a rare event that is of order 10^{-6}. Therefore, it is natural to expect the computational effort required in the estimation of such probabilities to scale likewise. One of the primary goals in this dissertation is to develop algorithms that estimate the probabilities of such rare events with uniformly bounded number of realizations

irrespective of how rare the events are. The rare events we consider are the ones that arise commonly in the analysis of stochastic systems involving heavy-tailed random variables.

Efficient simulation of rare events has gained a lot of attention in stochastic operations research because of its applications in the context of computer and communication networks, building highly reliable systems, actuarial sciences, etc. In particular, systems involving random variables possessing light tails are considered pleasing to analyse for the simulation of rare events because there is a rich theory of large deviations (see, for example, Dembo and Zeitouni [1998] and Dupuis and Ellis [1997]) that studies the limiting behaviour of rare events under light-tailed assumptions. However, the conventional theory of large deviations ceases to explain the behaviour of rare events when the random variables involved are heavy-tailed. As a result, there is a huge gap in the rare event simulation literature in terms of development of efficient simulation algorithms for light-tailed and heavy-tailed random variables. However, over the past decade, there has been considerable interest in the research community on simulation of rare events in heavy-tailed setup, primarily because of their importance and applications (see Blanchet and Lam [2012] for a survey).

As mentioned earlier, our goal in this dissertation is to derive asymptotic behaviours and develop efficient simulation algorithms for some of the important rare events involving heavy-tailed stochastic processes. To be specific, we consider how certain heavy-tailed stochastic processes assume values which are much larger than what they are expected to assume. For example, if $S_n = X_1 + \ldots + X_n$ is the sum of n independent random variables with zero mean and unit variance, it is well-known that for large values of n, S_n exceeds n with negligibly low probability. Here, the collection $(\{S_n > n\} : n \geq 1)$ constitutes a family of rare events satisfying

$$\mathbb{P}\{S_n > n\} \to 0, \text{ as } n \to \infty.$$

Simulation of rare events in heavy-tailed settings is considered difficult because of what is known as "the big jump phenomenon": When the random variables $X_1, \ldots, X_n$ are heavy-tailed, it is often the case that S_n is unusually large because one of the increments $X_1, \ldots, X_n$ is unusually large while the others assume usual values. It is instructive to contrast this behaviour with light-tailed sums where all the increments are known to"conspire" to take a value larger than they normally do, in order for the sum to be large. The following example, used often in lectures, illustrates the difference:

Example 1.2. If the sum of the heights of 100 people randomly chosen in a city is larger than 600 ft, it is likely that many of them are taller than 6 ft. Here, several components in the sum conspire to achieve a large value for the sum. On the other hand, if the sum of the number of

twitter followers of the same 100 people is larger than a million, it is likely that one of them is a celebrity who has about a million followers, and everyone else has a few hundreds of followers (which is the average number of followers). Here, the sum has become large because one of the components has "misbehaved", while the others "behave normally." In fact, it has been empirically verified that the number of twitter followers in the web follows a power law (which is heavy-tailed).

After considering a wide range of problems involving efficient simulation of rare probabilities in heavy-tailed sums, we move to an interesting framework involving multi-server queues. In particular, we study tail behaviour of steady-state waiting time in multi-server queues processing arrivals with heavy-tailed job sizes. This problem has received substantial attention in the last 15 years (see Scheller-Wolf and Sigman [1997], Whitt [2000], Foss and Korshunov [2006], Scheller-Wolf and Vesilo [2011] and Foss and Korshunov [2012]). However in the current literature, characterizations of steady-state delay exist only when the traffic intensity* is not an integer. There are qualitative reasons, as we shall see, that make the integer case significantly more delicate to analyse. Our main contribution in this part is that we provide the first tail asymptotics of steady-state delay when the traffic intensity is an integer.

Organization of the dissertation

After providing required mathematical background in Chapter 2, we present our results on simulation of rare events in Part I of the dissertation. This part, comprising Chapters 3, 4, 5 and 6, deals with various interesting large deviations and level crossing probabilities that arise in the heavy-tailed sums. The rare events considered in Chapters 3 and 4, as we shall discuss, have received immense attention in the last decade. Apart from devising efficient algorithms for these rare event probabilities, our key contribution has been to question the prevailing view that one needs to resort to state-dependent methods for efficient computation of rare event probabilities involving "large number" of heavy-tailed random variables. We develop an entirely new methodology for the simulation of rare events in heavy-tailed random walks by partitioning the events based on the "big jump principle". After suitably partitioning, we devise intuitive changes of measure which are easy to draw samples from.

In Chapter 5, we consider the problem of estimation of tail probabilities of infinite sum of heavy-tailed random variables. Several linear processes, important time series and stochastic recurrence equations that arise in economics, finance and network science can be written in the

*Traffic intensity is the ratio of rate of arrivals to the rate of service. It is a quantitative measure of how busy the servers tend to be under the assumed stochastic dynamics.

form of infinite sums we consider. Unlike the simulation problems considered in Chapters 3 and 4, any algorithm that stops after generating only finitely many increments is likely to introduce bias. Apart from eliminating bias, the algorithm we develop estimates tail probabilities just with uniformly bounded number of realizations.

The algorithms developed in Chapters 3, 4 and 5 are provably "efficient", follow a general template, and are easy to generalize, as we shall see in Chapter 6, to settings more general than sums of independent random variables. To prove results on efficiency of these algorithms, we develop asymptotic estimates for various probabilities involved in these settings which are interesting in their own right.

In Part II of the dissertation, comprising only Chapter 7, we develop asymptotic tail bounds for steady-state waiting time in multi-server queues. As mentioned earlier, such tail bounds are known only when the traffic intensity is not an integer. The case of integer traffic intensity, as we shall identify, exhibits interesting transitions in system behaviour, which are not present when the traffic intensity is not an integer. To be specific about our contribution, let V denote the generic service requirement of a job arriving to a first come first serve (FCFS) two-server queue. Let $\bar{B}(x) = \mathbb{P}\{V > x\}$ be regularly varying[†] and let the jobs arrive independently at a rate equal to the rate of service. In this case the traffic intensity ρ equals 1. Further, let W denote the amount of time a job arriving in steady-state has to wait. Our first result is that

$$\mathbb{P}\{W > b\} = \Theta\left(b^2\bar{B}(b^2) + b^2\bar{B}^2(b)\right), \text{ as } b \to \infty, \tag{1.1}$$

where $f(b) = \Theta(g(b))$ means that $f(b) \leq c_1 g(b)$ and $g(b) \leq c_2 f(b)$ for some positive constants c_1, c_2 independent of b. Let us contrast this asymptotic result with that derived in Foss and Korshunov [2006]. For the case $\rho < 1$, it was found that if the job sizes have finite variance,

$$\mathbb{P}\{W > b\} = \Theta\left(b^2\bar{B}^2(b)\right),$$

whereas for the case $\rho \in (1, 2)$, Foss and Korshunov [2006] obtained that

$$\mathbb{P}\{W > b\} = \Theta\left(b\bar{B}(b)\right),$$

as $b \to \infty$. Since there is a sharp transition in the behaviour between the cases $\rho < 1$ and $\rho \in (0, 1)$, it has been of great interest to identify what happens when ρ equals 1. We resolve this in our work by noting that (1.1) is much closer to the case $\rho < 1$ than to the case $\rho > 1$. The case of job sizes having infinite variance introduces another sharp transition in behaviour which is also reported in Chapter 7. In the development of (1.1), we identify that large delays

[†]Refers to probability distributions with polynomially decaying tails; see Section 2.2 for a definition

happen when $\rho = 1$ because of arrival of one or two big jobs and consequent buildup of workload due to effects that occur at time scales governed by both law of large numbers and central limit theorem, which is not common in the asymptotic analysis of multi-server queues.

In order to streamline presentation, proofs of some of the results that are of auxiliary nature are collected and presented in a separate section at the end of the corresponding chapters.

Bibliographics

The results presented in Chapters 3 and 4, on simulation of large deviations and level crossing probabilities, are joint with Sandeep Juneja and Jose Blanchet, and have appeared in Murthy et al. [2014]. An example of extending simulation of rare events in i.i.d. sums to that of Markov modulated random walks has been done in Murthy et al. [2013]. The results on simulation of tail probabilities for infinite sums of regularly varying random variables in Chapter 5 are obtained jointly with Sandeep Juneja and Henrik Hult. The work on tail asymptotics of multi-server queues which is described in Chapter 7 is joint with Jose Blanchet, and has been accepted to appear in Queuing Systems.

2 Simulation of Rare Events: Mathematical Preliminaries

In this chapter, we describe the theoretical framework for analysis of algorithms that aim to estimate rare probabilities. In addition, we explain the concept of regular variation which is the form of heavy-tailed behaviour the stochastic processes we consider exhibit.

To begin with, consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We call a family of events $(A_x : x > 0)$ as *rare* if $\mathbb{P}\{A_x\} \to 0$ as $x \to \infty$. It is common to refer the event A_x as rare event, x as rarity parameter and $\mathbb{P}\{A_x\}$ as rare probability. To make this clear, consider the following examples:

Example 2.1. Take any proper random variable X. The tail probabilities $\mathbb{P}\{X > x\}$ for large values of x are small and are considered rare. Here x is the rarity parameter.

Example 2.2. Consider the random walk defined by

$$S_0 := 0, \quad S_n := X_1 + \ldots + X_n \text{ for } n \geq 1.$$

The increment random variables $X_1, X_2, \ldots$ are independent copies of a random variable X with zero mean and unit variance. The famous central limit theorem for sums of independent and identically distributed (i.i.d.) random variables states that the sequence of probability distributions corresponding to $S_n/\sqrt{n}$ converges to standard normal distribution. Observe that the limiting standard normal distribution does not assign positive probability to ∞. As a consequence, for any $\kappa > 1/2$, the probabilities

$$\mathbb{P}\left\{S_n > n^{\kappa}\right\} \to 0 \text{ as } n \to \infty,$$

and hence are rare for large values of n. In this example, the rarity parameter is n.

2.1 Simulation algorithms for estimation of rare probabilities

As mentioned in Chapter 1, we estimate the rare probabilities $\mathbb{P}\{A_x\}$ mainly by doing Monte-Carlo simulations.

> Given $\epsilon, \delta > 0$, our objective is to compute an estimate for $\mathbb{P}\{A_x\}$ that lies in the interval $((1-\epsilon)\mathbb{P}\{A_x\}, (1+\epsilon)\mathbb{P}\{A_x\})$ with probability at least $1-\delta$.

For every $x > 0$, let us assume, for the time being, that we have access to an algorithm which returns a realization of a random variable Z_x satisfying $\mathbb{E}[Z_x] = \mathbb{P}\{A_x\}$. Further, for $N \geq 1$, let $\hat{Z}_x(N)$ denote the arithmetic mean of N realizations of Z_x. Then $\mathrm{Var}[\hat{Z}_x(N)] = \mathrm{Var}[Z_x]/N$. Due to a simple application of Markov's inequality,

$$\mathbb{P}\left\{\frac{\left|\hat{Z}_x(N) - \mathbb{P}\{A_x\}\right|}{\mathbb{P}\{A_x\}} > \epsilon\right\} \leq \frac{\mathrm{Var}\left[Z_x\right]}{N\epsilon^2\mathbb{P}\{A_x\}^2} = \frac{\mathrm{CV}^2[Z_x]}{N\epsilon^2},$$

where $\mathrm{CV}[Z_x] = \sqrt{\mathrm{Var}[Z_x]}/\mathbb{E}[Z_x]$ is the coefficient of variation of Z_x. Therefore, if we choose

$$N > \frac{\mathrm{CV}^2[Z_x]}{\delta\epsilon^2}, \tag{2.1}$$

our objective is achieved: that is, we have an estimate $\hat{Z}_x(N)$ lying within ϵ-relative precision of $\mathbb{P}\{A_x\}$ with probability at least $1-\delta$. It remains to understand how to arrive at Z_x. Consider the following examples:

1) *The naive estimator:* Take $Z_x = \mathbb{I}(A_x)$. Then $\mathrm{Var}[Z(x)]$ equals $\mathbb{P}\{A_x\}(1-\mathbb{P}\{A_x\})$. As a consequence of (2.1), N must be chosen as large as

$$\frac{\mathbb{P}\{A_x\}\left(1-\mathbb{P}\{A_x\}\right)}{\delta\epsilon^2\mathbb{P}\{A_x\}^2} \sim \frac{1}{\delta\epsilon^2\mathbb{P}\{A_x\}} \to \infty,$$

as $x \to \infty$. The problem with this approach is that for small probabilities $\mathbb{P}\{A_x\}$, the number of realizations of Z_x required grows like $\mathbb{P}\{A_x\}^{-1}$. For example, if $\mathbb{P}\{A_x\}$ is exponentially decaying with respect to x, then we have to generate exponentially growing number of realizations of Z_x, which is computationally prohibitive for large values of x.

2) *Importance sampling estimators:* Given $x > 0$, consider a probability measure $\mathbb{Q}_x(\cdot)$ such that $\mathbb{P}(\cdot)$ when restricted to the event A_x is absolutely continuous with respect to $\mathbb{Q}_x(\cdot)$. Further, let $\mathbb{E}_{\mathbb{Q}_x}[\cdot]$ denote the expectation operator with respect to the measure $\mathbb{Q}_x(\cdot)$. Then

$$\mathbb{P}\{A_x\} = \int_{A_x} \mathbb{P}(d\omega) = \int_{A_x} \frac{d\mathbb{P}}{d\mathbb{Q}_x}(\omega)\mathbb{Q}_x(d\omega) = \mathbb{E}_{\mathbb{Q}_x}\left[L_x; A_x\right], \tag{2.2}$$

where $L_x = d\mathbb{P}/d\mathbb{Q}_x$ is the Radon-Nikodym derivative, which we refer to as likelihood ratio of $\mathbb{P}(\cdot)$ with respect to $\mathbb{Q}_x(\cdot)$. Now instead of obtaining samples from $\mathbb{P}(\cdot)$ like how we did in naive estimation, one can obtain samples from $\mathbb{Q}_x(\cdot)$ and use the estimator $Z_x = L_x\mathbb{I}(A_x)$. It follows from (2.2) that such an estimator will be unbiased, that is, $\mathbb{E}_{\mathbb{Q}_x}[Z_x] = \mathbb{P}\{A_x\}$. The advantage with this approach is that one can choose the measure $\mathbb{Q}_x(\cdot)$ in a way that the rare event under consideration is emphasized more in simulation, and hence it is witnessed more often than under the original measure $\mathbb{P}(\cdot)$. The choice $\mathbb{Q}_x(\cdot)$ is referred in the literature as *importance sampling measure* and the resulting estimator $Z_x = L_x\mathbb{I}(A_x)$ as *importance sampling estimator*. The collection $((\mathbb{Q}_x, Z_x) : x > 0)$ is collectively referred to as *importance sampling algorithm*. Due to (2.1), the number of realizations of Z_x that are required is at least

$$N_x := \frac{\mathrm{CV}^2[Z_x]}{\delta\epsilon^2} = \frac{\mathrm{Var}_{\mathbb{Q}_x}[Z_x]}{\delta\epsilon^2\mathbb{P}\{A_x\}^2}.$$

As we shall see later in this dissertation, one can often exploit the structure of the simulation problem in hand and choose importance sampling measures $\mathbb{Q}_x(\cdot)$ in a way that the resulting estimators Z_x have low variance, at times, even smaller than $\mathbb{P}\{A_x\}^2$.

3) *Zero-variance estimators:* For every $x > 0$, consider the importance sampling choice $Q_x(\cdot) = \mathbb{P}(\cdot|A_x)$. In other words, we obtain samples from the conditional measure $\mathbb{P}(\cdot|A_x)$ and use the resulting Radon-Nikodym derivative $Z_x = d\mathbb{P}/d\mathbb{Q}_x$ as the estimator. It can be easily verified that the Radon-Nikodym derivative $d\mathbb{P}/d\mathbb{Q}_x$ restricted to event A_x equals $\mathbb{P}\{A_x\}$. As a result the estimator Z_x has zero variance, and hence the choice $Q_x(\cdot) = \mathbb{P}(\cdot|A_x)$ (also referred to as *zero-variance measure*[*]) is optimal. However, the explicit dependence of the estimator on $\mathbb{P}\{A_x\}$, the quantity which we want to compute, makes this method impractical. Though the zero variance estimator described here is unimplementable, it turns out that one can attempt to gain an understanding of the asymptotic structure of zero-variance measure $\mathbb{P}(\cdot|A_x)$ and use it as a guidance to derive importance sampling estimators with low variance. See Example 2.3 for an application of this idea. Choosing importance sampling measures based on asymptotic behaviour of conditional measures, as we shall see, is a recurrent theme in the design of importance sampling algorithms presented in this dissertation.

4) *Conditional Monte Carlo estimators:* For certain simulation problems, it is possible to obtain an alternate representation for the quantity of interest by conditioning on suitable

[*]The terminology "zero-variance measure" does not mean that the measure is degenerate; the name is commonly used in simulation literature because the estimator associated with the corresponding change of measure has zero variance.

information. In such cases, the alternate representation can be used to come up with simulation estimators. For example, consider the i.i.d. sum $S_n = X_1 + \ldots + X_n$ and let us say our aim is to estimate $\mathbb{P}\{S_n > b\}$ for some $b > 0$. Here assume that $X_1, X_2, \ldots$ are i.i.d. copies of some random variable X and let $\bar{F}(x) := \mathbb{P}\{X > x\}$ denote the tail probabilities of X. See that $\mathbb{P}\{S_n > b\}$ can be alternatively written as

$$\mathbb{P}\{S_n > b\} = \mathbb{E}\left[\mathbb{P}\left\{X_n > b - S_{n-1} \;\middle|\; S_{n-1}\right\}\right] = \mathbb{E}\left[\bar{F}(b - S_{n-1})\right].$$

Following the above representation, one can simulate the first $n-1$ increment random variables $X_1, \ldots, X_{n-1}$ from measure $\mathbb{P}(\cdot)$ and use $Z = \bar{F}(b - S_{n-1})$ as the estimator for $\mathbb{P}\{S_n > b\}$. This type of estimators where simulation is carried out on the basis of an alternate representation resulting from conditioning on a suitable collection of random variables are called conditional Monte Carlo estimators.

Performance measures for rare event simulation algorithms

As explained in the previous section, one can have a variety of estimators Z_x satisfying $\mathbb{E}[Z_x] = \mathbb{P}\{A_x\}$. To compare between them, one obvious candidate for performance measure could be the number of realizations of Z_x that are required to achieve the desired relative precision. Due to (2.1), this quantity, in turn, is proportional to the coefficient of variation of Z_x. The following are some of the highly desirable performance measures that are defined based on coefficient of variation of Z_x.

Definition 2.1. *A family of estimators $(Z_x : x > 0)$ is said to be strongly efficient in the estimation of $(\mathbb{P}\{A_x\} : x > 0)$ if their coefficient of variation is bounded: that is, if*

$$\sup_x \mathrm{CV}[Z_x] < \infty.$$

Definition 2.2. *A family of estimators $(Z_x : x > 0)$ is said to possess asymptotically vanishing relative error property in the estimation of $(\mathbb{P}\{A_x\} : x > 0)$ if*

$$\mathrm{CV}[Z_x] \to 0 \text{ as } x \to \infty.$$

The significance of these definitions are as follows: If a family of estimators $(Z_x : x > 0)$ is strongly efficient, then as a consequence of (2.1), the number of samples of Z_x required to achieve the desired relative precision stays bounded irrespective of how rare A_x is. Similarly, if

the family $(Z_x : x > 0)$ has vanishing relative error property, then just with a uniformly bounded number of realizations of Z_x, the estimators possess negligible relative error for large values of x. As is evident from the definitions, vanishing relative error property is a slightly stronger criterion compared to strong efficiency. Though algorithms that satisfy these performance criteria are appealing because of their computational advantage, it has to be remembered that for several rare event simulation problems it is difficult to develop algorithms that satisfy these criteria. For such problems, it is considered good to have estimators which necessitate the number of samples of Z_x required to grow, as the event A_x becomes rarer, but at a rate much smaller than that of naive simulation. One such performance criteria is as follows:

Definition 2.3. *A family of estimators* $(Z_x : x > 0)$ *is said to be weakly efficient if*

$$\overline{\lim_{x\to\infty}} \frac{\mathrm{Var}[Z_x]}{\mathbb{P}\{A_x\}^{2-\epsilon}} < \infty$$

for every $\epsilon > 0$.

Example 2.3. (**Importance sampling via exponential twisting**) Consider the random walk in Example 2.2. The rare event of interest is $A_n := \{S_n > na\}$ for some $a > 0$. Since $\kappa = 1$ is larger than $1/2$, $\mathbb{P}\{A_n\} \to 0$ as $n \to \infty$ (follows from the explanation in Example 2.2). Therefore, A_n is rare for large values of n. Let $F(\cdot)$ denote the probability distribution of increment random variable X. Let $\Lambda(\theta) := \log \mathbb{E}[\exp(\theta X)]$ denote the log-moment generating function of X and assume that there exists θ larger than 0 for which $\Lambda(\theta)$ is finite and $\Lambda'(\theta) = a$. Then it can be shown that (see, for example, Juneja and Shahabuddin [2006]) for any fixed m,

$$\mathbb{P}\left\{X_{m+1} \in dx \;\middle|\; S_m = s, S_n > na\right\} \to \exp\left(\theta x - \Lambda(\theta)\right) F(dx) \tag{2.3}$$

as $n \to \infty$. Recall that $\mathbb{P}\{\cdot | S_n > na\}$ is the zero-variance measure corresponding to our simulation problem. The above asymptotics indicate that for large values of n, under the zero-variance measure, the increments $X_1, X_2, \ldots$ roughly behave like i.i.d. copies obtained from the distribution

$$F_\theta(dx) := \exp\left(\theta a - \Lambda(\theta)\right) F(dx).$$

Since $F_\theta(dx) \propto \exp(\theta x) F(dx)$, the distribution $F_\theta(\cdot)$ is called the exponentially twisted version of $F(\cdot)$. Let $\mathbb{P}_\theta(\cdot)$ denote the probability measure induced when the i.i.d. increments $X_1, X_2, \ldots$ follow the exponentially twisted distribution $F_\theta(\cdot)$. Let $\mathbb{E}_\theta[\cdot]$ denote the expectation operator associated with $\mathbb{P}_\theta(\cdot)$. Since $\Lambda'(\theta) = a$, it is immediate that $\mathbb{E}_\theta[X] = a$. Therefore, if the increments $X_1, X_2, \ldots$ are drawn from the distribution $F_\theta(\cdot)$, as a consequence of law of large numbers,

$$\frac{S_n}{n} \to a \quad \mathbb{P}_\theta(\cdot) \text{ a.s.},$$

and hence the rare event $\{S_n > na\}$ happens with non-vanishing probability. The intuition behind the limiting result (2.3) is that the most likely way for the sum S_n to attain a value larger than na is by having the individual increments $X_1, X_2, \ldots, X_n$ all of them conspire to take values higher than usual by following distribution $F_\theta(\cdot)$. See that unlike the zero-variance measure, the distribution $F_\theta(\cdot)$ is amenable for obtaining samples for simulation. Now the strategy for estimation of $\mathbb{P}\{S_n > na\}$ should be obvious. Since we cannot obtain samples of increments $X_1, X_2, \ldots$ from the zero-variance measure, we instead obtain samples for $X_1, X_2, \ldots$ from $F_\theta(\cdot)$ (which is close to the zero-variance measure for large values of n because of (2.3)). It is shown in Sadowsky and Bucklew [1990] that obtaining samples for increments i.i.d. from $F_\theta(\cdot)$ instead of $F(\cdot)$ indeed works and the resulting family of importance sampling estimators

$$Z_n(\theta) := \exp\left(-\theta S_n + n\Lambda(\theta)\right) \mathbb{I}\left(S_n > na\right)$$

is weakly efficient (see Definition 2.3).

2.2 Regular variation and heavy tails

In Example 2.3, it is assumed that the increment random variable X used to define the random walk $(S_n : n \geq 1)$ satisfies the following property: there exists a $\theta > 0$ such that $\mathbb{E}[\exp(\theta X)]$ is finite. In general, random variables for which origin belongs to the interior of the set $\{\theta : \mathbb{E}[\exp(\theta X)] < \infty\}$ are called light-tailed random variables. As mentioned in Chapter 1, random variables possessing light tails are considered pleasing to analyse for the simulation of rare events because there is a rich theory of large deviations (see, for example, Dembo and Zeitouni [1998] and Dupuis and Ellis [1997]) that studies the limiting behaviour of rare events under light-tailed assumptions. However, the conventional theory of large deviations ceases to explain the behaviour of rare events when $\mathbb{E}[\exp(\theta X)]$ is infinite for every positive θ, that is, when the distribution of X is *heavy-tailed*. As a result, there is a huge gap in the rare event simulation literature in terms of development of efficient simulation algorithms for light-tailed and heavy-tailed random variables. However, over the past decade, there has been a considerable interest in the research community on rare event simulation of heavy-tailed stochastic processes, primarily because of their importance and applications (see Blanchet and Lam [2012] for a survey). As mentioned in Chapter 1, our interest in this dissertation is to develop efficient simulation algorithms for some of the important rare events involving heavy-tailed stochastic processes. For accomplishing this, we need to precisely identify the kind of heavy-tail behaviour that we shall be dealing with. The rest of this chapter is devoted for this purpose.

Definition 2.4. *A random variable X is said to be heavy-tailed if $\mathbb{E}[\exp(\theta X)]$ is infinite for every positive θ.*

Example 2.4. Consider a random variable X distributed according to the following Pareto distribution:

$$\bar{F}(x) := \mathbb{P}\{X > x\} = \begin{cases} x^{-3} & \text{if } x > 1, \\ 1 & \text{otherwise.} \end{cases}$$

Since the tail probabilities $\bar{F}(x)$ is polynomially decaying, it is immediate that $\mathbb{E}[\exp \theta X]$ is infinite for every $\theta > 0$. Therefore, X is a heavy-tailed random variable. On the other hand, consider an exponential random variable Y with mean 1. It is easy to check that the moment-generating function $\mathbb{E}[\exp(\theta Y)]$ equals $(1-\theta)^{-1}$ for every θ smaller than 1. Therefore, Y is a light-tailed random variable.

Definition 2.5. *A (measurable) function $L : \mathbb{R} \to \mathbb{R}^+$ is said to be slowly varying at infinity if*

$$\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1 \quad \textit{for every } t > 0.$$

Some examples of slowly varying functions include constant functions, $|\log x|^\beta$ for any $\beta \in \mathbb{R}, 1 - e^{-x}$, etc.

Definition 2.6. *A function $V : \mathbb{R} \to \mathbb{R}^+$ is said to be regularly varying with index α if $V(x) = x^\alpha L(x)$ for some slowly varying function $L(\cdot)$.*

Definition 2.7. *A random variable X is said to be regularly varying if its tail distribution $\mathbb{P}\{X > x\} = x^{-\alpha} L(x)$ for some slowly varying function $L(\cdot)$ and $\alpha > 0$.*

Alternatively, we call a random variable X to be regularly varying if its tail distribution $\bar{F}(x) = \mathbb{P}\{X > x\}$ is a regularly varying function (in terms of x) with an index that is negative. In Example 2.4, the tail distribution $\bar{F}(\cdot)$ is regularly varying with index -3.

Regularly varying distribution functions capture the concept of polynomially decaying tails, and form an important class of heavy-tailed distributions. The heavy-tailed random variables we consider in this dissertation will always be regularly varying unless specified explicitly. The following properties of regularly varying functions will be useful. A proof of these elementary facts can be found, for example, in Chapter VIII of Feller [1971].

1) *Any slowly varying function $L(x)$ is just $x^{o(1)}$[†]:* If $L(\cdot)$ is a slowly varying function at infinity then for any fixed $\epsilon > 0$, there exists x_ϵ large enough such that for all $x > x_\epsilon$,

$$x^{-\epsilon} < L(x) < x^{\epsilon}. \tag{2.4}$$

Consequently, any regularly varying function $V(x)$ with index α is just $x^{\alpha+o(1)}$.

2) *Regularly varying distributions are subexponential:* Let X be a regularly varying random variable and $X_1, \ldots, X_m$ be i.i.d. copies of X. Then

$$\mathbb{P}\{X_1 + \ldots + X_m > x\} \sim m\mathbb{P}\{X > x\}, \tag{2.5}$$

as $x \to \infty$[‡]. Any random variable X that satisfies (2.5) is called a subexponentially distributed random variable.

3) *Regularly varying distributions possess long tails:* If $V(\cdot)$ is a regularly varying function, then for any fixed $c > 0$,

$$\lim_{x\to\infty} \frac{V(x+c)}{V(x)} = 1. \tag{2.6}$$

Verification of this property is straightforward from the definition of regularly varying and slowly varying functions. As a consequence, if a random variable X is regularly varying, then

$$\lim_{x\to\infty} \mathbb{P}\{X > x + c | X > x\} = \lim_{x\to\infty} \frac{\mathbb{P}\{X > x + c\}}{\mathbb{P}\{X > x\}} = 1.$$

In fact, it is true that

$$\lim_{x\to\infty} \mathbb{P}\{X > x + h(x) | X > x\} = \lim_{x\to\infty} \frac{\mathbb{P}\{X > x + h(x)\}}{\mathbb{P}\{X > x\}} = 1$$

for any $h(x) = o(x)$. This property can be checked from the following bounds, which are referred to as *Potter's bounds* in the literature: If $L(\cdot)$ is a slowly varying function, then as in Theorem 1.1.4 of Borovkov and Borovkov [2008], for any $\delta > 0$, there exists a $t_\delta > 0$ such that for all t and v satisfying $t \geq t_\delta$ and $vt \geq t_\delta$,

$$(1-\delta)\min\{v^\delta, v^{-\delta}\} \leq \frac{L(vt)}{L(t)} \leq (1+\delta)\max\{v^\delta, v^{-\delta}\}$$

As a consequence, for any regularly varying function $V(x) = x^{-\alpha}L(x)$, we obtain

$$(1-\delta)\min\{v^{-\alpha+\delta}, v^{-\alpha-\delta}\} \leq \frac{V(vt)}{V(t)} \leq (1+\delta)\max\{v^{-\alpha+\delta}, v^{-\alpha-\delta}\}. \tag{2.7}$$

[†]We follow Landau's notation for describing asymptotic behaviour of functions: for given functions $f : \mathbb{R}^+ \to \mathbb{R}^+$ and $g : \mathbb{R}^+ \to \mathbb{R}^+$, we say $f(x) = O(g(x))$ if there exists $c_1 > 0$ and x_1 large enough such that $f(x) \leq c_1 g(x)$ for all $x > x_1$; and $f(x) = \Omega(g(x))$ if there exists $c_2 > 0$ and x_2 large enough such that $f(x) \geq c_2 g(x)$ for all $x > x_2$. We use $f(x) = o(g(x))$ if $f(x)/g(x) \to 0$, and $f(x) \sim g(x)$ if $f(x)/g(x) \to 1$, as $x \to \infty$.

[‡]Recall that we use $f(x) \sim g(x)$ if $f(x)/g(x) \to 1$, as $x \to \infty$.

4) *Finiteness of moments:* Let X be a non-negative regularly varying random variable with tail index $-\alpha$ (for some $\alpha > 0$). Then

$$\mathbb{E}X^\beta < \infty \quad \text{if } \beta < \alpha,$$
$$\mathbb{E}X^\beta = \infty \quad \text{if } \beta > \alpha.$$

A proof of this fact and several other interesting properties of regularly varying functions can be found, for example, in Bingham et al. [1989].

5) *Karamata's theorem*: For any regularly varying function $V(\cdot)$ with index $-\alpha$, if β is such that $\alpha - \beta > 1$, then

$$\int_x^\infty u^\beta V(u)du \sim \frac{x^{\beta+1}V(x)}{\alpha - \beta - 1}, \text{ as } x \to \infty. \tag{2.8}$$

On the other hand, if $\alpha - \beta < 1$, then

$$\int_0^x u^\beta V(u)du \sim \frac{x^{\beta+1}V(x)}{1 - \alpha + \beta}, \text{ as } x \to \infty. \tag{2.9}$$

These results, popularly known as Karamata's theorem (cf. Theorem 1 in Chapter VIII.9 of Feller [1971]), provide asymptotic characterizations of integrated tails of regularly varying functions.

As mentioned earlier, all the heavy-tailed random variables considered in this dissertation will be regularly varying unless specified explicitly.

3 Estimation of Large Deviation Probabilities

Let X be a regularly varying random variable and $(X_n : n \geq 1)$ be i.i.d. copies of X. Let

$$S_0 := 0, \text{ and } S_n := X_1 + \ldots + X_n \text{ for } n \geq 1$$

denote the random walk associated with the i.i.d. collection $(X_n : n \geq 1)$.

Given $\epsilon, \delta > 0$, the objective of this chapter to devise efficient simulation algorithms that estimate large deviation probabilities $\mathbb{P}\{S_n > b\}$ within ϵ-relative precision with probability at least $1 - \delta$.

After providing a brief summary of important results on sums of heavy-tailed random variables in Section 3.1, the simulation probem in hand is explained in Section 3.2 along with a brief literature survey. While Sections 3.3 and 3.4 explain our efficient algorithms in full detail, Section 5.4 is devoted for a numerical experiment that compares the performance of the proposed algorithms with those existing in the literature. In order to streamline presentation, proofs of certain results are collected and presented separately in Section 3.6.

3.1 Limit theorems for sums of regularly varying random variables

We begin with a discussion of basic limit theorems for sums of regularly varying random variables. A proof of these results can be found, for example, in Feller [1971] or Bingham et al. [1989].

1) *The Strong law of large numbers:* If $\mathbb{E}|X|$ is finite, then

$$\frac{S_n}{n} \to \mathbb{E}X \text{ a.s.}$$

as $n \to \infty$.

One can say that the value $n\mathbb{E}X$ is a first-order approximation to the sum S_n.

2) *The central limit theorem:* Let Z denote a normal random variable with zero mean and unit variance. If $\mathbb{E}X^2$ is finite, then

$$\frac{S_n - n\mathbb{E}X}{\sqrt{n\text{Var}[X]}} \Rightarrow Z \quad \text{as } n \to \infty. \tag{3.1}$$

The symbol $\Rightarrow$ denotes convergence in distribution. In other words, for every $x \in \mathbb{R}$,

$$\mathbb{P}\left\{\frac{S_n - n\mathbb{E}X}{\sqrt{n\text{Var}[X]}} > x\right\} \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} \exp(-u^2/2)du$$

as $n \to \infty$. One can think of $n\mathbb{E}X + Z\sqrt{n\text{Var}[X]}$ as a second-order approximation to the sum S_n.

3) *Convergence to stable distributions:* Consider the case where $\mathbb{E}X$ is finite but the variance $\text{Var}[X]$ is infinite. Let

$$G(x) := \mathbb{P}\{X > x\} + \mathbb{P}\{X \leq -x\}, \qquad x > 0.$$

be a regularly varying function with index $-\alpha$. If α lies between 1 and 2, then it happens that $\mathbb{E}X$ is finite and $\text{Var}[X]$ is infinite (see Property 4 of regularly varying functions in Section 2.2 in Chapter 2). Under this assumption on $G(\cdot)$,

$$\frac{S_n - n\mathbb{E}X}{n^{\frac{1}{\alpha}}L_1(n)} \Rightarrow Z_\alpha \quad \text{as } n \to \infty \tag{3.2}$$

for some slowly varying function $L_1(n)$ and random variable Z_α following an α-stable distribution. In this case, the value $n\mathbb{E}X + n^{1/\alpha}L_1(n)Z_\alpha$ corresponds to a second-order approximation to the sum S_n. Compare this with the case of X having finite variance where the traditional central limit theorem is applicable.

4) *Tail Asymptotics:* Let us consider random variable S_m for any fixed m. Since S_m is a proper random variable, it is evident that $\mathbb{P}\{S_m > b\} \to 0$ as $b \to \infty$. Further, since regularly varying distributions are also subexponential (see (2.5)),

$$\mathbb{P}\{S_m > b\} \sim m\mathbb{P}\{X > b\}, \tag{3.3}$$

as $b \to \infty$. To gain further understanding, if we assume that the random variable X is non-negative, it follows from Proposition 1 of Blanchet and Lam [2012] that

$$\mathbb{P}\left\{\max_{1\le j\le m} X_j > b \,\middle|\, S_m > b\right\} \to 1, \tag{3.4}$$

as $b \to \infty$. That is, the most likely way for the sum S_m to become large is by having at least one of the increments $X_1, \ldots, X_m$ (and hence the maximum of $X_1, \ldots, X_m$) as large.

3.2 The simulation problem

Let X be a zero mean random variable with distribution $F(\cdot)$ satisfying the following:

Assumption 3.1. *The tail probabilities of X are given by $\bar{F}(x) := \mathbb{P}\{X > x\} = x^{-\alpha}L(x)$ for some slowly varying function $L(\cdot)$ and $\alpha > 1$. Further if $Var[X] = \infty$, the tail probabilities of X satisfy the following condition:*

$$\overline{\lim_{x\to\infty}} \frac{\mathbb{P}\{X < -x\}}{\mathbb{P}\{X > x\}} < \infty.$$

Let $\beta := (\alpha \wedge 2)^{-1}$. Then, as a consequence of central limit theorem (or) convergence to appropriate stable distribution (discussed in Section 3.1), we have that $S_n/n^{\beta+\epsilon} \to 0$ for every $\epsilon > 0$, as $n \to \infty$. Therefore, the event $\{S_n > b\}$ for $b > n^{\beta+\epsilon}$ is rare for large values of n. Our objective in this chapter is to efficiently estimate the probabilities $\mathbb{P}\{S_n > b\}$ when $b > n^{\beta+\epsilon}$.

In Section 3.1, we saw that for fixed n, $\mathbb{P}\{S_n > b\} \sim n\bar{F}(b)$ as $b \to \infty$. Proposition 3.1, whose proof is provided later in Section 3.6, asserts that a similar asymptotics hold even in the limiting regime where $n \to \infty$ and $b > n^{\beta+\epsilon}$.

Proposition 3.1. *Under Assumption 3.1, for every $\epsilon > 0$,*

$$\sup_{b>n^{\beta+\epsilon}} \left|\frac{\mathbb{P}\{S_n > b\}}{n\bar{F}(b)} - 1\right| \to 0 \quad \text{as } n \to \infty.$$

Remark 3.1. A simple application of Bonferroni inequalities will yield

$$\mathbb{P}\{\max\{X_1, \ldots, X_n\} > b\} = n\bar{F}(b)\left(1 - \frac{1}{2}\bar{F}(b) + \frac{\theta}{6}\left(n\bar{F}(b)\right)^2\right), \tag{3.5}$$

for some θ in $(0, 1)$. This indicates that the tail asymptotics of maximum and the sum of increments $\{X_1, \ldots, X_n\}$ match asymptotically as $n \to \infty$.

Relevant literature

Sums of random variables are ubiquitous, and one of the basic challenging problem that arise in the context of sums of random variables is the estimation of $\mathbb{P}\{S_n > b\}$ when the probability $\mathbb{P}\{S_n > b\}$ is small. In addition, estimation of large deviation probabilities $\mathbb{P}\{S_n > b\}$ is important because they form building blocks to many more complex rare event problems involving combination of renewal processes: for examples in queueing, see Parekh and Walrand [1989] and in financial credit risk modeling, see Glasserman and Li [2005] and Bassamboo et al. [2008]. We shall see in the later chapters that the techniques developed for simulation of these large deviation problems can be used as a guideline in the simulation of more complex rare events.

As in Example 2.3, in the light-tailed settings, large deviations analysis can be used to show that exponential twisting based importance sampling well approximates the zero-variance measure for efficient estimation of large deviations probabilities (see Sadowsky and Bucklew [1990]). However, development of simulation algorithms for these rare probabilities is considered harder in heavy-tailed settings. Asmussen et al. [2000] provide an account of failure of simple large deviations based simulation methods that approximate zero-variance measure in heavy-tailed systems. Research on efficient simulation of rare events involving heavy-tailed variables first focussed on probabilities such as $\mathbb{P}\{S_N > b\}$ in the simpler asymptotic regime where N is fixed or geometrically distributed and $b \to \infty$. Some of the notable algorithms for this problem include Asmussen et al. [2000], Juneja and Shahabuddin [2002], Asmussen and Kroese [2006] and Juneja [2007]. In particular, the conditional Monte Carlo estimators developed in Asmussen and Kroese [2006] have proved to be simple to implement and extremely effective in the simulation of tail probabilities of sum S_N when N is a fixed positive integer.

The idea behind the conditional Monte Carlo estimators developed in Asmussen and Kroese [2006] is as follows: For simulating tail probabilities of the sum $S_N = X_1 + \ldots + X_N$, let $M_k = \max\{X_1, \ldots, X_k\}$ for every $k \leq N$ and see that for any given N and b,

$$\begin{aligned}\mathbb{P}\left\{S_N > b, M_N = X_N \;\middle|\; X_1, \ldots, X_{N-1}\right\} &= \mathbb{P}\left\{X_N > (b - S_{N-1}) \vee M_{N-1} \;\middle|\; X_1, \ldots, X_{N-1}\right\} \\ &= \bar{F}\left((b - S_{N-1}) \vee M_{N-1}\right).\end{aligned}$$

Due to the i.i.d. nature of increments $X_1, \ldots, X_n$, the quantity of interest $\mathbb{P}\{S_N > b\}$ can be alternatively written as below:

$$\begin{aligned}\mathbb{P}\{S_N > b\} &= N\mathbb{P}\left\{S_N > b, M_N = X_N\right\} \\ &= N\mathbb{E}\left[\mathbb{P}\left\{S_N > b, M_N = X_N \;\middle|\; X_1, \ldots, X_{N-1}\right\}\right] \\ &= N\mathbb{E}\left[\bar{F}\left((b - S_{N-1}) \vee M_{N-1}\right)\right]. \qquad (3.6)\end{aligned}$$

In every simulation run, the idea is to simply obtain samples of increments $X_1, \ldots, X_{N-1}$ from their distribution (no importance sampling involved) and return

$$Z_N(b) := N\bar{F}\left((b - S_{N-1}) \vee M_{N-1}\right) \tag{3.7}$$

as an unbiased estimator for $\mathbb{P}\{S_N > b\}$. The conditional Monte-Carlo estimators $Z_N(b)$ in (3.7) are simple to implement, efficient in practice and are popularly referred to as Asmussen-Kroese estimators. One can easily check that

$$(b - S_{N-1}) \vee M_{N-1} \geq \frac{b}{N}$$

with equality holding when all the increments $X_1, \ldots, X_{N-1}$ equal b/N. As a result, $Z_N(b)$ is at least $N\bar{F}(b/N)$ which is small for N fixed and b large, and hence $Z_N(b)$ has low variance in the limiting regime where N is fixed and $b \to \infty$. However, the same argument does not hold when N is also large, which is the case with the simulation problem that we are tackling in this chapter.

Importance sampling (State-dependence vs state-independence): When it comes to importance sampling in the context of sums of random variables, the importance sampling algorithms can be either state-dependent or state-independent: State-dependence essentially means that the sampling distribution for generating the increment X_k depends on the realized values of $X_1, \ldots, X_{k-1}$ (typically, through S_{k-1}). State-independence on the other hand implies that samples of $X_1, \ldots, X_n$ can be drawn independently. State-independent methods often enjoy advantages over state-dependent ones in terms of complexity of generating samples and ease of implementation. However, Bassamboo et al. [2007] prove that any importance sampling change of measure that prescribes increments to be drawn in a state-independent i.i.d. fashion cannot efficiently estimate the level crossing probabilities within a regenerative cycle of a heavy-tailed random walk. In addition to the perceived difficulties in the simulation of rare event $\{S_n > b\}$ when n and b are large, the zero variance measure for the estimation of $\mathbb{P}\{S_n > b\}$ turns out to be state-dependent (see Juneja and Shahabuddin [2006]). These considerations have motivated research over the last few years in development of complex and elegant state-dependent algorithms to efficiently estimate these probabilities (see, e.g., [Dupuis et al., 2007, Blanchet and Glynn, 2008a, Blanchet and Liu, 2008, 2012, Chan et al., 2012]). The first efficient algorithm for the simulation of $\mathbb{P}\{S_n > b\}$ (when n and b are large), developed in Blanchet and Liu [2008], is state-dependent. The strategy for simulation in Blanchet and Liu [2008] is to restrict the candidate importance sampling distributions to a parametric family that is rich enough to mimic the behaviour of zero-variance measure $\mathbb{P}\{\cdot|S_n > b\}$. The parameters are chosen in a way that a suitable Lyapunov inequality is satisfied. Then the Lyapunov

inequality is used to derive desirable bounds on second moments of estimator. Though the algorithm is provably efficient, the inherent difficulties in choosing appropriate parameters for state-dependent sampling leaves room for more research to come up with estimators that are simple and intuitive (similar to (3.7)). We also have a sequential importance sampling and resampling algorithms due to Chan et al. [2012]. In Murthy and Juneja [2012], it is shown that a variant capped exponential twisting based state-independent importance sampling, which does not involve any decomposition, provides a strongly efficient estimator for the large deviations probabilities that we consider in this chapter.

3.3 Simulation of $\{S_n > b\}$: An importance sampling algorithm

In this section we present an importance sampling algorithm for the estimation of large deviation probabilities $\mathbb{P}\{S_n > b\}$ for $b > n^{\beta+\epsilon}$ as $n \to \infty$. The algorithm is state-independent, simple to implement and expends only $O(n)$ computational effort. We show that the proposed estimator possesses the desirable vanishing relative error property, and perform at least as well as the existing state-dependent algorithms. Thus our key contribution has been to question the prevailing view that one needs to resort to state-dependent methods for efficient computation of rare event probabilities involving 'large number' of heavy-tailed random variables. A key idea to be exploited is the fact that the corresponding rare event occurrence is governed by the "single big jump" principle, that is, the most likely paths leading to the occurrence of the rare event have one of the increments taking large value (see, for e.g., Foss et al. [2011] and the references therein).

Our approach for estimating the large deviations probability $\mathbb{P}\{S_n > b\}$ relies on decomposing $\mathbb{P}\{S_n > b\}$ into a dominant and a residual component, and developing efficient estimation techniques for both. Recall (from Section 3.1) that the asymptotics for $\mathbb{P}\{S_n > b\}$ and $\mathbb{P}\{M_n > b\}$ match asymptotically as $n \to \infty$. The strategy for simulation is to partition the event $\{S_n > b\}$ based on whether the maximum of the increments $\{X_1, \ldots, X_n\}$ (denoted by M_n) exceeds the large value b or not:

$$A_{\text{dom}}(n, b) := \{S_n > b, M_n \geq b\} \text{ and}$$
$$A_{\text{res}}(n, b) := \{S_n > b, M_n < b\}.$$

Such a partition is considered in Juneja [2007] for the simulation of $\{S_n > b\}$ when n is fixed. We prove the following result in Section 3.6

Proposition 3.2. *Under Assumption 3.1, given any* $\epsilon > 0$,

$$\sup_{b > n^{\beta+\epsilon}} \frac{\mathbb{P}\left(S_n > b, M_n < b\right)}{n\bar{F}(b)} \to 0 \text{ as } n \to \infty.$$

It is immediate from Propositions 3.1 and 3.2, the probability of the event A_{res} is vanishingly small compared to the probability of A_{dom} as $n \to \infty$; the suffixes stand to indicate that A_{dom} is the dominant way of occurrence of $\{S_n > b\}$ for large n, and A_{res} has only residual contributions. We estimate $\mathbb{P}(A_{\text{dom}})$ and $\mathbb{P}(A_{\text{res}})$ independently via different changes of measure that typify the way in which the respective events occur, and add the individual estimates to arrive at a final estimator for $\mathbb{P}\{S_n > b\}$.

Simulating A_{dom}

For the simulation of A_{dom}, we follow the two-step procedure outlined in Chan et al. [2012]:

1. Choose an index I uniformly at random from $\{1, \ldots, n\}$
2. For $k = 1, \ldots, n$, generate a realization of X_k from $F(\cdot | X \geq b)$ if $k = I$; otherwise, generate X_k from $F(\cdot)$

Let $\mathbb{P}_1(\cdot)$ denote the measure induced when the increments are generated according to the above procedure. For brevity we have chosen not to highlight the dependence of the importance sampling change of measure $\mathbb{P}_1(\cdot)$ on n and b in the notation. Note that the probability measure $\mathbb{P}(\cdot)$ is absolutely continuous with respect to $\mathbb{P}_1(\cdot)$ when restricted to A_{dom}. We have,

$$d\mathbb{P}_1(x_1, \ldots, x_n) = \sum_{k=1}^{n} \frac{1}{n} \cdot \frac{dF(x_1) \ldots dF(x_n)}{\bar{F}(b)} \mathbf{1}(x_k \geq b).$$

Therefore the likelihood ratio on the set A_{dom} is given by,

$$\frac{d\mathbb{P}}{d\mathbb{P}_1}(X_1, \ldots, X_n) = \frac{n\bar{F}(b)}{\#\{X_i \geq b : 1 \leq i \leq n\}},$$

and the resulting unbiased estimator for the evaluation of $\mathbb{P}(A_{\text{dom}})$ is,

$$Z_{\text{dom}}(n, b) := \frac{n\bar{F}(b)}{\#\{X_i \geq b : 1 \leq i \leq n\}} \mathbb{I}(A_{\text{dom}}). \tag{3.8}$$

Generate N independent realizations of Z_{dom} and take their sample mean as an estimator of $\mathbb{P}(A_{\text{dom}})$. To evaluate how large N should be chosen so that the computed estimate satisfies the given relative error specification, we need to obtain bounds on variance of Z_{dom}. Since $\#\{X_i \geq b : 1 \leq i \leq n\}$ is at least 1, when the increments are drawn following the measure $\mathbb{P}_1(\cdot)$, we have $Z_{\text{dom}}(n, b) \leq n\bar{F}(b)$, and hence,

$$\mathbb{E}_1\left[Z_{\text{dom}}^2(n, b)\right] \leq \left(n\bar{F}(b)\right)^2.$$

Also, due to Propositions 3.1 and 3.2, $\mathbb{E}_1\left[Z_{\text{dom}}(n,b)\right] = \mathbb{P}(A_{\text{dom}}(n,b)) \sim \mathbb{P}\{S_n > b\} \sim n\bar{F}(b)$ as $n \to \infty$. Therefore we get,

$$\text{Var}_1\left[Z_{\text{dom}}(n,b)\right] = o\left(\left(n\bar{F}(b)\right)^2\right), \text{ as } n \to \infty. \tag{3.9}$$

Remark 3.2. Since $\mathbb{P}\{S_n > b, M_n > b\} = n\mathbb{P}\{S_n > b, M_n > b, M_n = X_1\}$, one can estimate $\mathbb{P}\{S_n > b, M_n > b, M_n = X_1\}$ efficiently by simulating X_1 from $F(\cdot|X_1 > b)$ and the other increments from $F(\cdot)$. This avoids the simulation of an additional random variable I. However, we have presented the two step procedure above so that the simulation procedures introduced in later chapters appear intuitive.

Remark 3.3. If the increments $X_1, \ldots, X_n$ are not identically distributed, and if at least one of the increments is regularly varying, then it can be verified that the following modification to the simulation of auxiliary random variable I would suffice: Say $X_j \sim F_j(\cdot)$. Then choose $I = i$ from $\{1, \ldots, n\}$ with probability $\bar{F}_i(b)/\sum_{j=1}^n \bar{F}_j(b)$.

Simulating A_{res}

We see that all the increments $\{X_1, \ldots, X_n\}$ are bounded from above by b on the occurrence of event A_{res}. Though the bound on the increments vary with n, we can employ methods similar to exponential twisting for light-tailed random walks (see Example 2.3) to simulate the event A_{res}. For given b, define

$$\Lambda_b(\theta) := \log\left(\int_{-\infty}^{b} \exp(\theta x) F(dx)\right), \quad \theta \geq 0.$$

Since the upper limit of integration is b, $\Lambda_b(\cdot)$ is well-defined for any positive value of θ. For given values of n and b, consider the distribution function $F_\theta(\cdot)$ satisfying,

$$\frac{dF_\theta(x)}{dF(x)} = \exp(\theta_{n,b}x - \Lambda_b(\theta_{n,b}))\mathbf{1}(x < b),$$

for all $x \in \mathbb{R}$ and some $\theta_{n,b} > 0$. Now the prescribed procedure is to just obtain independent samples of the increments $\{X_1, \ldots, X_n\}$ from $F_\theta(\cdot)$ and adjust via the likelihood ratio resulting due to sampling from a different distribution $F_\theta(\cdot)$.

Let $\mathbb{P}_2(\cdot)$ denote the measure induced by sampling increments i.i.d. from $F_\theta(\cdot)$. As before, for brevity, we have chosen not to highlight the dependence on parameters n and b in the notations $F_\theta(\cdot)$ and $\mathbb{P}_2(\cdot)$. For given values of n and b, we have the following unbiased estimator for the computation of $\mathbb{P}(A_{\text{res}})$:

$$Z_{\text{res}}(n,b) := \exp\left(-\theta_{n,b}S_n + n\Lambda_b(\theta_{n,b})\right)\mathbb{I}(A_{\text{res}}). \tag{3.10}$$

Now generate independent replications of Z_{res} and take their sample mean as the computed estimate for $\mathbb{P}(A_{\text{res}})$. However it remains to choose $\theta_{n,b}$. Since S_n is larger than b on A_{res},

$$Z_{\text{res}}(n,b) \leq \exp\left(-\theta_{n,b}b + n\Lambda_b(\theta_{n,b})\right)\mathbb{I}(A_{\text{res}}).$$

If we choose

$$\theta_{n,b} := -\frac{\log\left(n\bar{F}(b)\right)}{b}, \text{ then} \tag{3.11}$$

$$Z_{\text{res}}(n,b) \leq n\bar{F}(b)\exp\left(n\Lambda_b(\theta_{n,b})\right)\mathbb{I}(A_{\text{res}}). \tag{3.12}$$

We use Lemma 3.1, which is proved in Section 3.6, to obtain an upper bound on the second moment of the estimator Z_{res}.

Lemma 3.1. *Under Assumption 3.1, for the choice of $\theta_{n,b}$ as in (3.11),*

$$\exp\left(\Lambda_b(\theta_{n,b})\right) \leq 1 + \frac{1}{n}(1 + o(1)),$$

as $n \to \infty$, uniformly for $b > n^{\beta+\epsilon}$.

Therefore there exists a constant c such that

$$\exp\left(n\Lambda_b(\theta_{n,b})\right) \leq c,$$

for all admissible values of n and b. We evaluate the second moment of the estimator Z_{res} through the equivalent expectation operation corresponding to the original measure $\mathbb{P}(\cdot)$ as below:

$$\mathbb{E}_2\left[Z_{\text{res}}^2(n,b)\right] = \mathbb{E}\left[Z_{\text{res}}(n,b)\right] \leq cn\bar{F}(b)\mathbb{P}(A_{\text{res}}),$$

where the last inequality follows from (3.12) and Lemma 3.1. From Proposition 3.2, we have that $\mathbb{P}(A_{\text{res}}) = o\left(n\bar{F}(b)\right)$. Therefore,

$$\text{Var}_2\left[Z_{\text{res}}(n,b)\right] = o\left((n\bar{F}(b))^2\right), \text{ as } n \to \infty, \tag{3.13}$$

thus arriving at the following theorem:

Theorem 3.1. *If the realizations of the estimators Z_{dom} and Z_{res} are generated respectively from the measures $\mathbb{P}_1(\cdot)$ and $\mathbb{P}_2(\cdot)$, and if we let*

$$Z_{\text{IS}}(n,b) := Z_{\text{dom}}(n,b) + Z_{\text{res}}(n,b),$$

then under Assumption 3.1, the family of estimators $(Z_{\text{IS}}(n,b) : n \geq 1, b > n^{\beta+\epsilon})$ achieves asymptotically vanishing relative error for the estimation of $\mathbb{P}\{S_n > b\}$, as $n \to \infty$; that is,

$$\frac{\text{Var}_{n,b}\left[Z_{\text{IS}}(n,b)\right]}{\mathbb{P}\{S_n > b\}^2} = o(1),$$

as $n \to \infty$, uniformly for $b > n^{\beta+\epsilon}$.

Here $\mathrm{Var}_{n,b}[\cdot]$ denotes the variance operator resulting due to the composite procedure of drawing realizations of Z_{dom} and Z_{res} from the measures $\mathbb{P}_1(\cdot)$ and $\mathbb{P}_2(\cdot)$ respectively.

Proof. Since the realizations of Z_{dom} and Z_{res} are obtained independent of each other, the variance of Z_{IS} is just the sum of variances of Z_{dom} and Z_{res} computed according to the measures from which they are generated; the proof is now evident from (3.9), (3.13) and Proposition 3.1. □

Algorithm 1 below summarizes the entire simulation procedure.

Algorithm 1 Given n and $b > n^{\beta+\epsilon}$, the aim is to efficiently simulate $\mathbb{P}\{S_n > b\}$ via importance sampling

procedure ISEstimator(n, b)
Let $Z_{\mathrm{dom}}(n,b) =$ DominantEstimator(n,b) and
$Z_{\mathrm{res}}(n,b) =$ ResidualEstimator(n,b)
Return $Z_{\mathrm{IS}}(n,b) = Z_{\mathrm{dom}}(n,b) + Z_{\mathrm{res}}(n,b)$

procedure DominantEstimator(n, b)
Initialize $Z_{\mathrm{dom}}(n,b) = 0$ and I to a number drawn uniformly at random from $\{1,\ldots,n\}$
Simulate an i.i.d. realization of $(X_i : 1 \le i \le n, i \ne I)$ from the distribution $F(\cdot)$
For $i = I$, simulate X_i from $F(\cdot|X > b)$.
Assign $S_n = X_1 + \ldots + X_n$, $M_n = \max\{X_1,\ldots,X_n\}$ and $N = \#\{X_i \ge b : 1 \le i \le n\}$
If $S_n > b$ and $M_n \ge b$, assign $Z_{\mathrm{dom}}(n,b) = n\bar{F}(b)/N$
Return $Z_{\mathrm{dom}}(n,b)$

procedure ResidualEstimator(n, b)
Initialize $Z_{\mathrm{res}}(n,b) = 0$ and $\theta_{n,b} = -\log(n\bar{F}(b))/b$
Draw i.i.d. samples for $X_1,\ldots,X_n$ from $F_\theta(dx) = \exp(\theta_{n,b}x - \Lambda_b(\theta_{n,b}))F(dx)\mathbf{1}(x < b)$
Assign $S_n = X_1 + \ldots + X_n$ and $M_n = \max\{X_1,\ldots,X_n\}$
If $S_n > b$ and $M_n < b$, let $Z_{\mathrm{res}}(n,b) = \exp(-\theta_{n,b}S_n + n\Lambda_b(\theta_{n,b}))$
Return $Z_{\mathrm{res}}(n,b)$

Remark 3.4. A consequence of Theorem 3.1 is that, due to (2.1), the number of i.i.d. replications of $Z_{\mathrm{IS}}(n,b)$ required to achieve ϵ-relative precision with probability at least $1-\delta$ grows like $o(\epsilon^{-2}\delta^{-1})$. In our algorithm each call to the procedure ISEstimator(n,b) demands $O(n)$ computational effort, thus requiring an overall computational cost of $O(n)$ as $n \to \infty$.

3.4 Simulation of $\{S_n > b\}$: Conditional Monte Carlo

Let $M_k = \max\{X_1, \ldots, X_k\}$ for $k \leq n$ and recall the representation (3.6):

$$\mathbb{P}\{S_n > b\} = n\mathbb{E}\left[\bar{F}\left((b - S_{n-1}) \vee M_{n-1}\right)\right]$$

for the probability $\mathbb{P}\{S_n > b\}$ derived in Section 3.2. It is well-known that the Asmussen-Kroese estimators (introduced in Section 3.2)

$$Z_{\mathrm{AK}}(n, b) := n\bar{F}\left((b - S_{n-1}) \vee M_{n-1}\right) \tag{3.14}$$

are efficient in the estimation of probabilities $\mathbb{P}\{S_n > b\}$ in the limiting regime n fixed and $b \to \infty$ (see Asmussen and Kroese [2006]). The objective of this section is to prove that contrary to the prevailing understanding, the Asmussen-Kroese estimators $Z_{\mathrm{AK}}(n, b)$ are efficient in the estimation of $\mathbb{P}\{S_n > b\}$ in the limiting regime $b > n^{\beta+\epsilon}, n \to \infty$ as well.

Theorem 3.2. *Under Assumption 3.1, the family of estimators* $(Z_{\mathrm{AK}}(n, b) : n \geq 1, b > n^{\beta+\epsilon})$ *satisfy vanishing relative error property. In other words,*

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[Z^2_{\mathrm{AK}}(n, b)\right]}{\mathbb{P}\{S_n > b\}^2} = 1,$$

uniformly for $b > n^{\beta+\epsilon}$.

Proof. Fix $\gamma \in (0, 1)$ and let

$$I_1(n, b) := \mathbb{E}\left[\bar{F}^2\left((b - S_{n-1}) \vee M_{n-1}\right); (b - S_{n-1}) \vee M_{n-1} \geq \gamma b\right] \text{ and}$$
$$I_2(n, b) := \mathbb{E}\left[\bar{F}^2\left((b - S_{n-1}) \vee M_{n-1}\right); (b - S_{n-1}) \vee M_{n-1} < \gamma b\right].$$

Then $\mathbb{E}\left[Z^2_{\mathrm{AK}}(n, b)\right] = n^2(I_1(n, b) + I_2(n, b))$. From the definition of $I_1(n, b)$, it is immediate that

$$I_1(n, b) \leq \bar{F}^2(b)\mathbb{E}\left[\frac{\bar{F}^2\left(\left(\left(1 - \frac{S_{n-1}}{b}\right) \vee \gamma\right) b\right)}{\bar{F}^2(b)}\right].$$

Since $\bar{F}(x) = x^{-\alpha}L(x)$ for a slowly varying function $L(\cdot)$, given $\delta > 0$, it follows from (2.7) that

$$\frac{\bar{F}\left(\left(\left(1 - \frac{S_{n-1}}{b}\right) \vee \gamma\right) b\right)}{\bar{F}(b)} \leq (1 + \delta)h\left(\frac{S_{n-1}}{b}\right),$$

where $h(x) := ((1 - x) \vee \gamma)^{-(\alpha+\delta)}$. Observe that $h(\cdot)$ is an non-decreasing function. Therefore, for every n and $b > n^{\beta+\epsilon}$,

$$\frac{I_1(n, b)}{\bar{F}^2(b)} \leq (1 + \delta)^2\mathbb{E}\left[h^2\left(\frac{S_{n-1}}{b}\right)\right] \leq (1 + \delta)^2\mathbb{E}\left[h^2\left(\frac{S_{n-1}}{n^{\beta+\epsilon}}\right)\right]. \tag{3.15}$$

Observe that $h(\cdot)$ is a bounded function. Further, because of the convergence to stable distribution as in (3.1) or (3.2), the sequence $S_n/n^{\beta+\epsilon}$ converges to 0 almost surely for every $\epsilon > 0$ as $n \to \infty$. Since $h(0) = 1$, because of bounded convergence,

$$\mathbb{E}\left[h^2\left(\frac{S_{n-1}}{n^{\beta+\epsilon}}\right)\right] \to 1$$

as $n \to \infty$. As a result, it follows from (3.15) that

$$\overline{\lim_{n\to\infty}} \sup_{b>n^{\beta+\epsilon}} \frac{I_1(n,b)}{\bar{F}^2(b)} \leq (1+\delta)^2. \tag{3.16}$$

For the other term $I_2(n,b)$, we proceed as below: Since $\bar{F}(\cdot) \leq 1$,

$$\begin{aligned} I_2(n,b) &:= \mathbb{E}\left[\bar{F}^2\left((b-S_{n-1}) \vee M_{n-1}\right); (b-S_{n-1}) \vee M_{n-1} < \gamma b\right] \\ &\leq \mathbb{P}\left\{S_{n-1} \geq (1-\gamma)b, M_{n-1} < \gamma b\right\}. \end{aligned}$$

It follows from Theorem 4.1.2 (finite variance case) and Thereom 3.1.6 (infinite variance case) that $\mathbb{P}\left\{S_{n-1} \geq x, M_{n-1} < rx\right\} \leq c\left(n\bar{F}(x)\right)^r$ for some positive constant c independent of x. Therefore,

$$\frac{I_2(n,b)}{\bar{F}^2(b)} \leq c\frac{\left(n\bar{F}(b)\right)^{\frac{1-\gamma}{\gamma}}}{\bar{F}^2(b)}$$

for every $b > n^{\beta+\epsilon}$. Since $n < b^{(\beta+\epsilon)^{-1}}$, for a suitably small (but fixed) γ, it is immediate that

$$\overline{\lim_{n\to\infty}} \sup_{b>n^{\beta+\epsilon}} \frac{I_2(n,b)}{\bar{F}^2(b)} = 0$$

This observation, together with (3.16) and the large deviations asymptotics in Proposition 3.1, result in

$$\overline{\lim_{n\to\infty}} \sup_{b>n^{\beta+\epsilon}} \frac{\mathbb{E}\left[Z_{\mathrm{AK}}^2(n,b)\right]}{\mathbb{P}\{S_n > b\}^2} = \overline{\lim_{n\to\infty}} \sup_{b>n^{\beta+\epsilon}} \frac{n^2(I_1(n,b)+I_2(n,b))}{n^2\bar{F}^2(b)} \leq (1+\delta)^2 + 0.$$

Since δ is arbitrary and $\mathbb{E}\left[Z_{\mathrm{AK}}^2(n,b)\right] \geq \mathbb{E}[Z_{\mathrm{AK}}(n,b)]^2 = \mathbb{P}\{S_n > b\}^2$,

$$\lim_{n\to\infty} \sup_{b>n^{\beta+\epsilon}} \frac{\mathbb{E}\left[Z_{\mathrm{AK}}^2(n,b)\right]}{\mathbb{P}\{S_n > b\}^2} = 1.$$

This proves the claim. □

Remark 3.5. A consequence of the Theorem 3.2 is that a uniformly bounded number of realizations of $Z_{\mathrm{AK}}(n,b)$ is enough to estimate $\mathbb{P}\{S_n > b\}$ irrespective of the how large n or b is.

We summarize the simulation procedure below in Algorithm 2:

Algorithm 2 Given n and $b > n^{\beta+\epsilon}$, the aim is to efficiently simulate $\mathbb{P}\{S_n > b\}$ via conditional Monte Carlo

procedure AKestimator(n, b)
Generate an i.i.d. realization of $(X_i : 1 \leq i < n)$ from $F(\cdot)$
Let $M_{n-1} = \max\{X_1, \ldots, X_{n-1}\}$ and $S_{n-1} = X_1 + \ldots + X_{n-1}$
Let $Z_{\mathrm{AK}}(n, b) = n\bar{F}((b - S_{n-1}) \vee M_{n-1})$
Return $Z_{\mathrm{AK}}(n, b)$

3.5 A numerical example

In this section, we present the results of numerical simulation experiments performed on an example previously considered in literature. Take $X = \Lambda R$, where $\mathbb{P}\{\Lambda > x\} = 1 \wedge x^{-4}, R \sim$ Laplace(1), and Λ is independent of R. Let $S_n = X_1 + \ldots + X_n$ where $X_1, \ldots, X_n$ are i.i.d. copies of X. We use $N = 10,000$ simulation runs to estimate $\mathbb{P}\{S_n > n\}$ for $n = 100, 500$ and 1000. In Table 3.2, we compare the numerical estimates obtained by our importance sampling estimator (Z_{IS}) and Asmussen-Kroese estimator (Z_{AK}) with the true values of $\mathbb{P}\{S_n > n\}$ evaluated in Blanchet and Liu [2008] via inverse transform techniques; further, a comparison of performance of our simulation methodologies with Algorithms 1 and 2 in Blanchet and Liu [2008] (referred to as BL1 and BL2) has also been presented. From the columns CV of $\hat{Z}_{\mathrm{IS}}$, CV of $\hat{Z}_{\mathrm{AK}}$, CV of BL1, and CV of BL2, it can be inferred that our state-independent simulation procedures yield estimators with substantially lower coefficient of variation throughout the range of values considered. The state-dependent algorithms in comparison have been proven to be strongly efficient. The numerical performance of our algorithms in Table 3.2 just reflects the vanishing relative error of the estimators (a notion stronger than strong efficiency), which has been verified in Theorem 3.1.

Table 3.1: Numerical result for Example 1 - here CV denotes the empirically observed coefficient of variation based on 10,000 simulation runs

n	$\mathbb{P}\{S_n > n\}$	Estimate $\hat{Z}_{\mathrm{IS}}$	Estimate $\hat{Z}_{\mathrm{AK}}$	CV of $\hat{Z}_{\mathrm{IS}}$	CV of $\hat{Z}_{\mathrm{AK}}$	CV of BL1	CV of BL2
100	2.21×10^{-5}	2.17×10^{-5}	2.17×10^{-5}	1.97	4.37	10.3	4.70
500	1.04×10^{-7}	1.05×10^{-7}	1.04×10^{-7}	0.66	0.40	1.00	4.10
1000	1.25×10^{-8}	1.29×10^{-8}	1.25×10^{-8}	0.53	0.26	1.10	3.80

3.6 Proofs of auxiliary results

In this section, we provide proofs of Propositions 3.1, 3.2 and Lemma 3.1.

Table 3.2: Numerical result for Example 1 - here Std. error denotes the standard deviation of the estimator of $\mathbb{P}\{S_n > n\}$ based on 10,000 simulation runs

n	Std. error of $\hat{Z}_{\text{IS}}$	Std. error of $\hat{Z}_{\text{AK}}$
100	4.31×10^{-7}	9.49×10^{-7}
500	6.91×10^{-10}	4.17×10^{-10}
1000	6.02×10^{-11}	3.29×10^{-11}

Proof of Proposition 3.1 Proposition 3.1 is a direct consequence of Theorem 3.3 of Cline and Hsing [1991]. The content of Theorem 3.3 of Cline and Hsing [1991] is the following: If X is regularly varying with index $-\alpha$ satisfying

$$\sup_{x\geq 0}\frac{\mathbb{P}\{X\leq -x\}}{\mathbb{P}\{X>x\}}<\infty,$$

and if $(x_n : n \geq 1)$ is such that $n\mathbb{P}\{X > x_n\} \to 0$, then one of the following three conditions

(i) $0 \leq \alpha < 1$

(ii) $1 \leq \alpha < 2$ and $\lim_{n\to\infty} \frac{n}{x_n}\mathbb{E}\left[XI\left(|X|\leq x_n\right)\right] = 0$

(iii) $\alpha \geq 2$ and $\lim_{n\to\infty} \frac{n\mathbb{E}X}{x_n} = \lim_{n\to\infty} \frac{n\log x_n}{x_n^2}\mathbb{E}\left[X^2I\left(|X|\leq x_n\right)\right] = 0$

imply

$$\lim_{n\to\infty}\sup_{b\geq x_n}\left|\frac{\mathbb{P}\{S_n>b\}}{n\mathbb{P}\{X>b\}}-1\right|=0.$$

We simply verify that under Assumption 3.1, one of conditions (i)-(iii) listed above is true: When $\alpha \geq 2$, since $\mathbb{E}X = 0$, condition (iii) is trivially satisfied. On the other hand, when $\alpha \in (1, 2)$, as $\mathbb{E}X = 0$, we have that

$$\begin{aligned}\mathbb{E}\left[XI\left(|X|\leq b\right)\right] &= -\mathbb{E}\left[XI\left(|X|>b\right)\right]\\ &= bF(-b)+\int_b^\infty F(-y)dy-b\bar{F}(b)-\int_b^\infty \bar{F}(x)dx\end{aligned}$$

due to integration by parts. As a consequence of Assumption 3.1, we have that $F(-x) \leq \bar{F}(x)$ for all x large enough. Then,

$$\begin{aligned}\left|\frac{n}{b}\mathbb{E}\left[XI\left(|X|\leq b\right)\right]\right| &\leq \frac{2n}{b}\left(b\bar{F}(b)+\int_b^\infty \bar{F}(u)du\right)(1+o(1))\\ &\leq \frac{2n}{b}\left(b\bar{F}(b)+\frac{b\bar{F}(b)}{\alpha-1}\right)(1+o(1))\to 0,\ \text{as } n\to\infty,\end{aligned}$$

uniformly for $b > n^{\beta+\epsilon}$. Then, as a consequence of Theorem 3.3 of Cline and Hsing [1991], for any $\epsilon > 0$, we obtain that

$$\sup_{b>n^{\beta+\epsilon}}\left|\frac{\mathbb{P}\{S_n>b\}}{n\bar{F}(b)}-1\right|\to 0 \text{ as } n\to\infty.$$

This verifies the claim. □

Proof of Proposition 3.2 Let $M_n := \max\{X_1, \ldots, X_n\}$ for $n \geq 1$. We first obtain a lower bound for $\mathbb{P}\{S_n > b, X_n > b, M_{n-1} \leq b\}$:

$$\begin{aligned}\mathbb{P}\{S_n > b, X_n > b, M_{n-1} \leq b\} &\geq \mathbb{P}\left\{S_{n-1} > -b^\gamma, M_{n-1} \leq b, X_n > b + b^\gamma\right\} \\ &= \mathbb{P}\left\{S_{n-1} > -b^\gamma, M_{n-1} \leq b\right\} \bar{F}\left(b + b^\gamma\right) \qquad (3.17)\end{aligned}$$

for some $\gamma < 1$ to be chosen later in the proof. Due to (3.5),

$$\mathbb{P}\{M_{n-1} > b\} \sim (n-1)\bar{F}(b) \to 0$$

uniformly for all $b > n^{\beta+\epsilon}$, as $n \to \infty$. Here recall that $\beta := (\alpha \wedge 2)^{-1}$. Similarly for $\gamma > \beta/(\beta+\epsilon)$, because of the convergence of S_n/n^β to the stable distribution, we have $\mathbb{P}\{S_{n-1} < -b^\gamma\} \to 0$, uniformly for all $b > n^{\beta+\epsilon}$, as $n \to \infty$. Therefore, it follows from union bound that,

$$\mathbb{P}\left\{S_{n-1} \geq -b^\gamma, M_{n-1} \leq b\right\} \geq 1 - o(1),$$

uniformly for all $b > n^{\beta+\epsilon}$, as $n \to \infty$. Since $\gamma < 1$,

$$\frac{\bar{F}(b + b^\gamma)}{\bar{F}(b)} \geq 1 - o(1)$$

because of (2.7). Combining these observations with (3.18), it follows that

$$\mathbb{P}\{S_n > b, X_n > b, M_{n-1} \leq b\} \geq (1 - o(1))\bar{F}(b) \qquad (3.18)$$

uniformly for all $b > n^{\beta+\epsilon}$, as $n \to \infty$.
Since $\mathbb{P}(A_{\text{res}}(n, b)) = \mathbb{P}\{S_n > b\} - \mathbb{P}\{S_n > b, M_n > b\}$,

$$\begin{aligned}\mathbb{P}(A_{\text{res}}(n, b)) &\leq \mathbb{P}\{S_n > b\} - \sum_{j=1}^{n} \mathbb{P}\left\{S_n > b, X_j > b, \max_{i \neq j, i \leq n} X_i \leq b\right\} \\ &= \mathbb{P}\{S_n > b\} - n\mathbb{P}\{S_n > b, X_n > b, M_{n-1} \leq b\} \\ &\leq (1 + o(1))n\bar{F}(b) - (1 - o(1))n\bar{F}(b) = o\left(n\bar{F}(b)\right),\end{aligned}$$

where the last inequality follows from Proposition 3.1 and (3.18). □

To prove Lemma 3.1, we need Lemmas 3.2 and 3.3 stated and proved below.

Lemma 3.2. *For any pair of sequences $\{x_n\}, \{\phi_n\}$ satisfying $x_n \to \infty$ and $\phi_n x_n \to \infty$, the integral,*

$$\int_{-\infty}^{x_n} e^{\phi_n x} F(dx) \leq 1 + c\phi_n^\kappa + e^{2\alpha}\bar{F}\left(\frac{2\alpha}{\phi_n}\right) + e^{\phi_n x_n}\bar{F}(x_n)(1 + o(1)),$$

as $n \to \infty$, for any $1 < \kappa < \alpha \wedge 2$, and some constant c which does not depend on n and b.

Proof. We split the region of integration into $(-\infty, \gamma/\phi_n]$ and $(\gamma/\phi_n, x_n]$ for some constant $\gamma > 0$; the partition is such that the integrand stays bounded in the former region.
Let $I_1 := \int_{-\infty}^{\gamma/\phi_n} e^{\phi_n x} F(dx)$ and $I_2 := \int_{\gamma/\phi_n}^{x_n} e^{\phi_n x} F(dx)$.
For any $\kappa \in (1, 2]$ and $y > 0$, it is easily verified that

$$e^x \leq 1 + x + |x|^\kappa e^y, \quad x \in (-\infty, y].$$

Therefore,

$$\begin{aligned} I_1 &\leq \int_{-\infty}^{\gamma/\phi_n} \left(1 + \phi_n x + \phi_n^\kappa |x|^\kappa \exp(\phi_n \cdot \gamma/\phi_n)\right) F(dx) \\ &\leq \int_{-\infty}^{\gamma/\phi_n} F(dx) + \phi_n \int_{-\infty}^{\gamma/\phi_n} xF(dx) + \phi_n^\kappa e^\gamma \int_{-\infty}^{\gamma/\phi_n} |x|^\kappa F(dx) \\ &\leq \int_{-\infty}^{\infty} F(dx) + \phi_n \int_{-\infty}^{\infty} xF(dx) + \phi_n^\kappa e^\gamma \int_{-\infty}^{\infty} |x|^\kappa F(dx) \\ &= 1 + c\phi_n^\kappa, \end{aligned} \tag{3.19}$$

where $c := e^\gamma \int_{-\infty}^{\infty} |x|^\kappa F(dx) < \infty$ because $\mathbb{E}|X|^\kappa < \infty$; this follows because $\kappa < \alpha$ and from Assumption 3.1. We have also used $\mathbb{E}X = 0$ to arrive at (3.19). Integrating by parts for the second integral I_2 :

$$\begin{aligned} I_2 &= -\int_{\gamma/\phi_n}^{x_n} e^{\phi_n x} \bar{F}(dx) \\ &= e^{\phi_n \gamma/\phi_n} \bar{F}\left(\frac{\gamma}{\phi_n}\right) - e^{\phi_n x_n} \bar{F}(x_n) + \phi_n \int_{\gamma/\phi_n}^{x_n} e^{\phi_n x} \bar{F}(x) dx \\ &\leq e^\gamma \bar{F}\left(\frac{\gamma}{\phi_n}\right) + I_2', \end{aligned} \tag{3.20}$$

where, $I_2' := \phi_n \int_{\gamma/\phi_n}^{x_n} e^{\phi_n x} \bar{F}(x) dx$. Now the change of variable $u = \phi_n(x_n - x)$ results in:

$$\begin{aligned} I_2' &= e^{\phi_n x_n} \int_0^{\phi_n x_n - \gamma} e^{-u} \bar{F}\left(x_n - \frac{u}{\phi_n}\right) du \\ &= e^{\phi_n x_n} \bar{F}(x_n) \int_0^{\phi_n x_n - \gamma} e^{-u} g_n(u) du, \end{aligned} \tag{3.21}$$

where,

$$g_n(u) := \frac{\bar{F}\left(x_n - \frac{u}{\phi_n}\right)}{\bar{F}(x_n)} = \frac{\bar{F}\left(x_n\left(1 - \frac{u}{\phi_n x_n}\right)\right)}{\bar{F}(x_n)}.$$

Since $L(\cdot)$ is slowly varying and $\phi_n x_n \to \infty$, given any $\delta > 0$, it follows from (2.7) that,

$$(1-\delta)\left(1 - \frac{u}{\phi_n x_n}\right)^{-\alpha+\delta} \leq g_n(u) \leq (1+\delta)\left(1 - \frac{u}{\phi_n x_n}\right)^{-\alpha-\delta}.$$

for all n large enough. So for any fixed u, we have $g_n(u) \to 1$ as $n \to \infty$. Now fix $\delta = \frac{\alpha}{2}$. Then for n large enough,

$$g_n(u) \leq \left(1 + \frac{\alpha}{2}\right) \left(1 - \frac{u}{\phi_n x_n}\right)^{-\frac{3\alpha}{2}}. \tag{3.22}$$

Let $h(u) = (1 - u/\phi_n x_n)^{-\frac{3\alpha}{2}}$. Since $\log h(0) = 0$ and $\frac{d}{du}(\log(h(u)) \leq \frac{3\alpha}{2\gamma}$ for $0 \leq u \leq \phi_n x_n - \gamma$, we have $h(u) \leq \exp(3\alpha u/2\gamma)$ on the same interval. Therefore if we choose $\gamma = 2\alpha$, the integrand in I_2' is bounded for large enough n by an integrable function as below:

$$\begin{aligned} \left|e^{-u} g_n(u) \mathbf{1}(0 \leq u \leq \phi_n x_n - \gamma)\right| &\leq \left|e^{-u} \left(1 + \frac{\alpha}{2}\right) h(u) \mathbf{1}(0 \leq u \leq \phi_n x_n - \gamma)\right| \\ &\leq \left(1 + \frac{\alpha}{2}\right) e^{-u + \frac{3\alpha u}{2\gamma}} = \left(1 + \frac{\alpha}{2}\right) e^{-\frac{u}{4}}. \end{aligned}$$

Applying dominated convergence theorem, we get

$$\int_0^{\phi_n x_n - \gamma} e^{-u} g_n(u) du \sim 1 \text{ as } n \to \infty.$$

Since $\int_{-\infty}^{x_n} e^{\phi_n x} F(dx) = I_1 + I_2$, combining this result with (3.19), (3.20) and (3.21), completes the proof. □

Lemma 3.3. *Given any $\epsilon > 0$, uniformly for $b > n^{\beta+\epsilon}$, we have:*

(a) $n\theta_{n,b}^{\kappa} \to 0$ for some $1 < \kappa < \alpha \wedge 2$, and

(b) $\bar{F}(2\alpha/\theta_{n,b}) = o(1/n)$, as $n \to \infty$.

Proof. (a) We have $\bar{F}(x) = x^{-\alpha} L(x)$. Since $L(\cdot)$ is slowly varying, following (2.7) we have that $L(b) = b^{o(1)}$ as $b \to \infty$. Further noting that $b > n^{\beta+\epsilon}$ helps us to write:

$$n\theta_{n,b}^{\kappa} = \frac{n}{b^{\kappa}} \log^{\kappa}\left(\frac{1}{n\bar{F}(b)}\right) \leq n^{1-\kappa(\beta+\epsilon)} \log^{\kappa}\left(\frac{b^{\alpha}}{nL(b)}\right).$$

If we choose $\kappa \in ((\beta+\epsilon)^{-1}, \alpha)$ then $\kappa(\beta+\epsilon) > 1$ and subsequently $n\theta_{n,b}^{\kappa} \to 0$ as $n \to \infty$, uniformly for $b > n^{\beta+\epsilon}$.

(b) We have $\theta_{n,b} := -\log\left(n\bar{F}(b)\right)/b$. Therefore,

$$n\bar{F}\left(\frac{2\alpha}{\theta_n}\right) = n\bar{F}(b) \frac{\bar{F}\left(\frac{2\alpha b}{-\log(n\bar{F}(b))}\right)}{\bar{F}(b)}.$$

Since $\bar{F}(\cdot)$ is regularly varying, given any $\delta > 0$, it follows from (2.7) that

$$\frac{\bar{F}\left(\frac{2\alpha b}{-\log(n\bar{F}(b))}\right)}{\bar{F}(b)} \leq \left(\frac{-\log\left(n\bar{F}(b)\right)}{2\alpha}\right)^{\alpha+\delta},$$

for n large enough. Therefore,

$$n\bar{F}\left(\frac{2\alpha}{\theta_n}\right) \leq n\frac{L(b)}{b^\alpha}\left(\frac{-\log\left(n\bar{F}(b)\right)}{2\alpha}\right)^{\alpha+\delta} = o(1),$$

uniformly for $b > n^{\beta+\epsilon}$ as $n \to \infty$. Here the convergence to 0 is justified because $\alpha > 1$ and $b > n^{\beta+\epsilon}$. □

Now we are ready to provide a proof of Lemma 3.1.

Proof of Lemma 3.1 From the definition of $\Lambda_b(\cdot)$ and Lemma 3.2, we have:

$$\begin{aligned}\exp\left(\Lambda_b(\theta_{n,b})\right) &= \int_{-\infty}^{b} \exp(\theta_{n,b}x)F(dx) \\ &\leq 1 + c\theta_{n,b}^{\kappa} + e^{2\alpha}\bar{F}\left(\frac{2\alpha}{\theta_{n,b}}\right) + \exp(\theta_{n,b})\bar{F}(b)(1+o(1)),\end{aligned}$$

for $\kappa \in ((\beta+\epsilon)^{-1}, \alpha)$. Usage of Lemma 3.2 is justified because $b\theta_{n,b} = -\log\left(n\bar{F}(b)\right) \to \infty$. The last term,

$$\exp(\theta_{n,b}b)\bar{F}(b) = \frac{1}{n\bar{F}(b)}\bar{F}(b) = \frac{1}{n}.$$

From Lemma 3.3, we have $n\theta_{n,b}^{\kappa} = o\left(1\right)$ and $\bar{F}\left(2\alpha/\theta_{n,b}\right) = o\left(1/n\right)$, uniformly for $b > n^{\beta+\epsilon}$. Therefore,

$$\exp\left(\Lambda_b(\theta_n)\right) \leq 1 + \frac{1}{n}\left(1+o(1)\right), \text{ as } n \to \infty.$$

□

4 Estimation of Level Crossing Probabilities

A random walk with negative drift, due to law of large numbers, drifts to $-\infty$ almost surely, and it is unlikely that it exceeds a large positive level b in its course. In this chapter, we shall be interested in estimating such level crossing probabilities. To be precise, let X be a regularly varying random variable with negative mean. Take $(X_n : n \geq 1)$ to be i.i.d. copies of X. As in Chapter 3, let the sequence $(S_n : n \geq 0)$ with

$$S_0 := 0 \text{ and } S_n := X_1 + \ldots + X_n$$

represent the random walk associated with the i.i.d. collection $(X_n : n \geq 1)$. Additionally, let

$$M_n := \max_{k \leq n} S_k \quad \text{and } M := \max_{k \geq 0} S_k.$$

Since $\mathbb{E}X < 0$, the random walk $(S_n : n \geq 0)$ has negative drift, and it drifts towards $-\infty$ as $n \to \infty$. Consequently, the regeneration time

$$\tau := \inf\{n \geq 0 : S_n \leq 0\}$$

is almost surely finite. Further, let

$$\tau_b := \inf\{n \geq 1 : S_n > b\},$$

denote the first instance when the random walk S_n exceeds level b. In this chapter, our aim is to estimate

1) the probability that the negative-drift random walk $(S_n : n \geq 0)$ ever crosses a large positive level b, that is, $\mathbb{P}\{\tau_b < \infty\}$ (which is same as $\mathbb{P}\{M > b\}$), and

2) the probability that the negative-drift random walk $(S_n : n \geq 0)$ crosses a large positive level b within a regenerative cycle, that is, $\mathbb{P}\{\tau_b < \tau\}$ (which is same as $\mathbb{P}\{M_\tau > b\}$).

From here onwards, we refer to the probabilities $\mathbb{P}\{\tau_b < \infty\}$ and $\mathbb{P}\{\tau_b < \tau\}$ as level crossing probabilities. The maximum of random walk M and the maximum value M_τ within a regenerative cycle are both proper* random variables. Therefore, the tail events $\{M > b\}$ and $\{M_\tau > b\}$ (and hence the level crossing events $\{\tau_b < \infty\}$ and $\{\tau_b < \tau\}$) are rare for large values of b.

Given $\epsilon, \delta > 0$, the objective of this chapter to devise efficient simulation algorithms that estimate the level crossing probabilities $\mathbb{P}\{\tau_b < \infty\}$ and $\mathbb{P}\{\tau_b < \tau\}$ within ϵ-relative precision with probability at least $1 - \delta$.

These level crossing probabilities, as we shall discuss, can be identified with probability of ultimate ruin in stochastic insurance settings and probability of large delays in queuing systems. The events $\{\tau_b < \infty\}$ and $\{\tau_b < \tau\}$ correspond to the first passage of the Markov chain $(S_n - n\mu)$ into the set (b, ∞). The algorithms developed in this chapter could be a step towards development of algorithms for first passage probabilities for more general Markov chains involving heavy-tailed transition kernels.

Related literature

Maxima of random walks have always been of interest in stochastic operations research because they arise naturally when reflection maps are involved. For example, one can relate steady-state waiting time of a stable GI/GI/1 queue to maxima of random walks with negative drift via Lindley's recursion. Similarly, ultimate ruin in actuarial settings is related to maxima of random walk in a straightforward manner (see Asmussen [2003a] for these connections). As a result, simulation of tail probabilities of maxima (or level crossing probabilities of random walks) has received notable attention in the literature. Siegmund [1976] provides the first efficient importance sampling algorithm for estimating the level crossing probabilities when the increments X_n are light-tailed using large deviations based exponentially twisted change of measure. Bassamboo et al. [2007] prove that any algorithm that samples increments $X_1, X_2, \ldots$ in an i.i.d. fashion cannot efficiently estimate $\mathbb{P}\{\tau_b < \tau\}$. Thirty years after Siegmund's work, the first provably efficient algorithm for estimating level crossing probabilities in heavy-tailed settings was introduced in Blanchet and Glynn [2008a]. The Lyapunov bound techniques used in Blanchet and Glynn [2008a] are involved, and the sequel Blanchet and Liu [2012] attempt to make the simulation algorithms more comprehensive. While there exist simple state-independent simulation procedures for the estimation of these level crossing probabilities in the light-tailed settings,

*A proper random variable does not assign positive probability to $\pm\infty$.

the common feature of all the algorithms for heavy-tailed random walks mentioned here is that they are state-dependent.

Organisation of the chapter

The simulation methodology for the estimation of $\mathbb{P}\{\tau_b < \infty\}$ along with a variety of results encompassing both the finite and infinite variance (for the distribution of increments) cases are first presented in Section 4.1. Following this, we prove the key theorems in Section 4.2. Section 4.3 deals with simulation algorithms for efficient estimation of the $\mathbb{P}\{\tau_b < \tau\}$. In Section 4.4, we report performance of our algorithms comparing with the existing algorithms in the literature. Section 4.5 is a compendium of proofs of results that have not been proved in the main body of this chapter. Some of these proofs provide error bounds for asymptotic expressions of certain probabilities, and could be of interest in their own right.

4.1 Simulation Methodology for $\{\tau_b < \infty\}$

To precisely describe the problem in hand, let X be a zero mean random variable satisfying the following assumption:

Assumption 4.1. *The tail probabilities of X are given by $\bar{F}(x) := \mathbb{P}\{X > x\} = x^{-\alpha}L(x)$ for some slowly varying function $L(\cdot)$ and $\alpha > 1$. Further if $Var[X] = \infty$, the tail probabilities of X satisfy the following condition:*

$$\overline{\lim_{x\to\infty}} \frac{\mathbb{P}\{X < -x\}}{\mathbb{P}\{X > x\}} < \infty.$$

For the i.i.d. collection $(X_n : n \geq 1)$ which has the same distribution as X, we have the zero-drift random walk S_n defined as usual by

$$S_0 := 0 \text{ and } S_n = X_1 + \ldots + X_n.$$

Given $\mu > 0$, the stochastic process $(S_n - n\mu : n \geq 0)$ will be the negative-drift random walk whose level crossing probabilities we shall be interested in. Therefore, we define the following maxima

$$M_n := \max_{k\leq n}(S_k - k\mu) \quad \text{and} \quad M := \sup_n (S_n - n\mu),$$

and first passage times

$$\tau_b := \inf\{n > 0 : S_n - n\mu > b\} \quad \text{and } \tau := \inf\{n > 0 : S_n - n\mu < 0\}.$$

accordingly. In this section, our objective will be to present an algorithm for the estimation of level crossing probability $\mathbb{P}\{\tau_b < \infty\}$ and prove its efficiency.

Key ideas and a brief summary of results

Naive simulation of $\{\tau_b < \infty\}$ will require generation of all the increments until the partial sum process $S_n - n\mu$ exceeds b. Due to the negative drift of the random walk $(S_n - n\mu : n \geq 0)$, we have $\tau_b \to \infty$ almost surely as $b \to \infty$, and hence this method is not computationally feasible. To counter the prospect of generating uncontrollably large number of increment random variables in simulation, we re-express $\mathbb{P}\{\tau_b < \infty\}$ as below: Consider a strictly increasing sequence of integers $(n_k : k \geq 0)$ with $n_0 = 0$. Additionally, fix $p := (p_k : k \geq 1)$ satisfying $p_k > 0$ for all k and $\sum_k p_k = 1$; the vector p can be seen as a probability mass function on positive integers. Consider an auxiliary random variable K which takes the value of positive integer k with probability p_k. Then

$$\begin{aligned}\mathbb{P}\{\tau_b < \infty\} &= \sum_{k\geq 1} p_k \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}}{p_k} \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{P}\{n_{K-1} < \tau_b \leq n_K\}}{p_K} \,\middle|\, K\right]\right]. \end{aligned} \tag{4.1}$$

Now, in a simulation run, if the realized value of the auxiliary random variable K is k, generate a sample from a probability measure, possibly different from $\mathbb{P}(\cdot)$, of a random variable Z_k that has $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}$ as its expectation under the changed measure. Then equation (4.1) assures that taking the sample mean of i.i.d. replications of Z_K/p_K following the changes of measure (to be explained in the forthcoming sections) for the generation of $\{Z_k : k \geq 1\}$ will yield an unbiased estimator for the quantity $\mathbb{P}\{\tau_b < \infty\}$.

The performance of any importance sampling algorithm following the outlined procedure will depend crucially on the choice of probabilities p_k, and the changes of measure employed to estimate $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}$, for $k \geq 1$. The sequence $(n_k : k \geq 0)$ partitions non-negative integers into 'blocks' $((n_{k-1}, n_k] : k \geq 1)$. For reasons that will be clear later, we choose the blocks $(n_{k-1}, n_k]$ in the following manner: Fix a positive integer $r > 1$ and let,

$$n_0 = 0,\ n_k = r^k,\ \text{for } k \geq 1.$$

As in Chapter 3, the key idea to be exploited in the estimation of probabilities $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}$ is the fact that the corresponding rare event occurrence is governed by the "single big jump" principle, that is, the most likely paths leading to the occurrence of the rare event have one of the increments taking large value. Our approach for estimating the component probabilities

$\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}$ relies on decomposing it into a dominant and further residual components, and developing efficient estimation techniques for each of them.

When the increments X_n have finite variance, we develop unbiased estimators for level crossing probabilities $\mathbb{P}\{\tau_b < \infty\}$ that satisfy vanishing relative error property and whose implementation requires only $O(b)$ computational effort. However, when the variance of increments is infinite, there is an added complexity in estimating the level crossing probabilities $\mathbb{P}\{\tau_b < \infty\}$ as even the well known zero-variance measure is known to have an infinite expected termination time. We modify our algorithms so that this expectation remains finite while the estimators remain strongly efficient although they may no longer have asymptotically vanishing relative error.

Our specific contributions are as follows: For $\alpha > 1$, we develop unbiased estimators for level crossing probabilities $\mathbb{P}\{\tau_b < \infty\}$ that achieve vanishing relative error as $b \to \infty$. These estimators require an overall computational effort that scales as $O(b)$ when the variance of increments X_n is finite. This is similar to the complexity of the Zero variance operator since, as is well known, the latter requires at least $\mathbb{E}[\tau_b | \tau_b < \infty]$ computation in generating a single sample and this is known to be linear in b when the variance of increments is finite. On the other hand, $\mathbb{E}[\tau_b | \tau_b < \infty] = \infty$ for the case of increments having infinite variance. These conclusions on $\mathbb{E}[\tau_b | \tau_b < \infty]$ follow from Theorem 1.1 of Asmussen and Kluppelberg [1996]. Since $\mathbb{E}[\tau_b | \tau_b < \infty] = \infty$ for the case of increments having infinite variance, the zero-variance change of measure might not necessarily be a good benchmark, because from a computational standpoint any useful estimator needs to have finite expected termination time. For random walks with increments having infinite variance, we develop algorithms that satisfy the following:

a) When $\alpha > 1.5$, the associated estimators are strongly efficient and have $O(b)$ expected termination time. As a converse, we also prove that for $\alpha < 1.5$ *no algorithm can be devised in our framework* that has both the variance and expected termination time simultaneously finite. The situation is more nuanced when $\alpha = 1.5$ and depends on the form of the slowly varying function $L(\cdot)$.

b) When $\alpha \leq 1.5$, each replication of the estimator terminates in $O(b)$ time in expectation; also we require only $O(1)$ replications to achieve a given relative error, thus resulting in overall complexity of $O(b)$.

The above results for infinite increment variance, and in particular the bottleneck arising at $\alpha = 1.5$, closely mirror the results proved in Blanchet and Liu [2012] where vastly different

state-dependent algorithms based on Lyapunov inequalities are considered.

Related Asymptotics

In this section, we present asymptotics related to $\mathbb{P}\{\tau_b < \infty\}$ that will be useful in the efficiency analysis of the algorithms that are developed. Recall that

$$\tau_b := \inf\{k : S_k > b + k\mu\} \text{ and } M := \sup_n (S_n - n\mu).$$

The events $\{M > b\}$ and $\{\tau_b < \infty\}$ are the same. The main goal of this section is to precisely list the asymptotics that expose the intuition behind the "big jump principle" in this setting. Due to law of large numbers, the random walk $S_k - k\mu \approx -k\mu$ for large values of k. Therefore, if $X_k > b + k\mu$, it is likely that $S_k - k\mu > b$ for the first time, and hence τ_b equals k. This reasoning of big jump due to k^{th} increment can be used to write

$$\mathbb{P}\{\tau_b = k\} \approx \mathbb{P}\{X_k > b + k\mu\} = \bar{F}(b + k\mu).$$

This intuition can be used as a rough guideline to write

$$\mathbb{P}\{\tau_b < \infty\} = \sum_{k\geq 1} \mathbb{P}\{\tau_b = k\} \approx \sum_{k\geq 1} \bar{F}(b + k\mu) \approx \frac{1}{\mu}\int_0^\infty \bar{F}(b + u)du, \text{ and}$$

$$\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\} = \sum_{n_{k-1}+1}^{n_k} \mathbb{P}\{\tau_b = k\} \approx \sum_{n_{k-1}+1}^{n_k} \bar{F}(b + k\mu).$$

The following results rigorously justify the above approximation: Let $\bar{F}_I(x) := \int_x^\infty \bar{F}(u)du$ denote the integrated tail of $\bar{F}(\cdot)$. Under Assumption 4.1, it is well known (see, for example, Veraverbeke [1977]) that

$$\mathbb{P}\{\tau_b < \infty\} \sim \frac{1}{\mu}\bar{F}_I(b) \quad \text{as } b \to \infty. \tag{4.2}$$

The asymptotics (4.2) hold for level crossing probabilities of random walks under more general increment distributions (see, for example, Korshunov [1997]).

The following finite-horizon asymptotics are also available if we make the following non-restrictive smoothness assumption on the tail probabilities $\bar{F}(\cdot)$:

Assumption 4.2. *There exists a $t_0 > 0$ such that the slowly varying function $L(\cdot)$ in $\bar{F}(x) = x^{-\alpha}L(x)$ is continuously differentiable for all $t \geq t_0$. Further $L(\cdot)$ satisfies,*

$$L'(x) = o\left(\frac{L(x)}{x}\right), \text{ as } x \to \infty.$$

Most of the commonly encountered slowly varying functions, including asymptotically constant functions, logarithmic functions like $(\log x)^{\beta}$, $\log\log x$, all satisfy Assumption 4.2.

If X is such that $\mathrm{Var}[X] < \infty$ and it satisfies Assumptions 4.1 and 4.2, then from Theorem 6 of Borovkov and Borovkov [2002], we have uniformly in n that,

$$\mathbb{P}\{\tau_b \le n\} = \left(\sum_{j=1}^{n} \bar{F}(b+j\mu)\right)\left(1 + O\left(\frac{1}{b}\right)\right) + o\left(\sqrt{b \wedge n}\ \bar{F}(b)\right). \tag{4.3}$$

When $\mathrm{Var}[X] = \infty$, under Assumptions 4.1 and 4.2, it follows from Theorem 2.4 of Borovkov and Boxma [2001] that uniformly for all n, b satisfying $n\bar{F}(b) = o(1)$,

$$\mathbb{P}\{\tau_b \le n\} = \left(\sum_{j=1}^{n} \bar{F}(b+j\mu)\right)\left(1 + O\left(\frac{n^{\frac{1}{\alpha}+\epsilon}}{b}\right)\right) \tag{4.4}$$

for every $\epsilon > 0$.

The following characterization of the zero-variance measure $\mathbb{P}\{\cdot|\tau_b < \infty\}$ (see Theorem 1.1 of Asmussen and Kluppelberg [1996]) sheds light on how the first passage over a level b happens asymptotically: If we use $a(b) := \bar{F}_I(b)/\bar{F}(b)$, then conditional on $\tau_b < \infty$,

$$\left(\frac{\tau_b}{a(b)}, \left(\frac{S_{\lfloor u\tau_b \rfloor}}{\tau_b} : 0 \le u < 1\right), \frac{S_{\tau_b} - b}{a(b)}\right) \Longrightarrow \left(\frac{Y_0}{\mu}, (-u\mu : 0 \le u < 1), Y_1\right) \tag{4.5}$$

in $\mathbb{R} \times D[0,1) \times \mathbb{R}$. Here $D[0,1)$ denotes the space of càdlàg real-valued functions defined on the interval $[0,1)$. The joint law of Y_0, Y_1 is defined as follows: for $y_0, y_1 \ge 0, \mathbb{P}\{Y_0 > y_0, Y_1 > y_1\} = \mathbb{P}\{Y_1 > y_0 + y_1\}$ with $Y_0 \overset{d}{=} Y_1$, and

$$\mathbb{P}\{Y_1 > y_1\} = \frac{1}{(1 + y_1/(\alpha-1))^{\alpha-1}}.$$

Efficient simulation of $\{n_{k-1} < \tau_b \le n_k\}$

In this section we identify importance sampling changes of measure for the efficient computation of the probabilities $\mathbb{P}\{n_{k-1} < \tau_b \le n_k\}$. Define the following events:

$$A_k = \bigcup_{i=n_{k-1}+1}^{n_k} \{X_i > b + i\mu\} \text{ and } B_k = \bigcap_{i=1}^{n_k} \{X_i < b + n_{k-1}\mu\}.$$

The events A_k and B_k are defined in the same spirit as that of A_{dom} and A_{res} in the simulation of $\{S_n > b\}$ in Chapter 3: the event A_k includes sample paths that have at least one "big" jump of appropriate size in one of the increments indexed between n_{k-1} and n_k, whereas on the other set B_k, we have all the increments bounded from above. As in the simulation of large

deviation probabilities of sums of random variables in Chapter 3, we can partition the event $\{n_{k-1} < \tau_b \leq n_k\}$ into:

$$\{n_{k-1} < \tau_b \leq n_k, A_k\}, \{n_{k-1} < \tau_b \leq n_k, B_k\} \text{ and } \{n_{k-1} < \tau_b \leq n_k, \bar{A}_k \cap \bar{B}_k\},$$

and arrive at unbiased estimators for their probabilities separately via different importance sampling measures. Here $\bar{A}$ denotes complement of the set A.

Simulating $\{n_{k-1} < \tau_b \leq n_k, A_k\}$

We prescribe the following two step procedure which is similar to the simulation of event A_{dom} in Chapter 3: Let $q_k(b) := \sum_{i=n_{k-1}+1}^{n_k} \bar{F}(b + i\mu)$.

1. Choose an index $J \in \{n_{k-1} + 1, \ldots, n_k\}$ such that $\mathbb{P}\{J = n\} = \bar{F}(b + n\mu)/q_k(b)$, for $n_{k-1} < n \leq n_k$.
2. Simulate the increment X_n from $F(\cdot | X \geq b + n\mu)$ if $n = J$; otherwise, simulate X_n from $F(\cdot)$, for any $n \leq n_k$.

In this sampling procedure, we induce the 'big' jumps typically responsible for the occurrence of $\{n_{k-1} < \tau_b \leq n_k\}$ with suitable probabilities by sampling from the conditional distribution $F(\cdot | X \geq b + J\mu)$. This sampling procedure results in the importance sampling measure $\mathbb{P}_{k,1}(\cdot)$ characterised by:

$$d\mathbb{P}_{k,1}(x_1, \ldots, x_{n_k}) := \sum_{i=n_{k-1}+1}^{n_k} \frac{\bar{F}(b + i\mu)}{q_k(b)} . \frac{dF(x_1) \ldots dF(x_{n_k})}{\bar{F}(b + i\mu)} \mathbf{1}(x_i \geq b + ia).$$

This in turn yields a likelihood ratio,

$$\frac{d\mathbb{P}}{d\mathbb{P}_{k,1}}(X_1, \ldots, X_{n_k}) = \frac{q_k(b)}{\#\{X_i \geq b + i\mu : n_{k-1} < i \leq n_k\}},$$

on the set A_k. Then we have,

$$Z_{k,1}(b) := \frac{q_k(b)}{\#\{X_i \geq b + i\mu : n_{k-1} < i \leq n_k\}} \mathbb{I}(n_{k-1} < \tau_b \leq n_k, A_k) \tag{4.6}$$

as the unbiased estimator for the quantity $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, A_k\}$.

Simulating $\{n_{k-1} < \tau_b \leq n_k, B_k\}$

On the event B_k, none of the random variables $X_1, \ldots, X_{n_k}$ exceed the level $(b + n_{k-1}\mu)$; since these increments are bounded (on B_k), we can draw their samples from the distribution obtained

by exponentially twisting the distribution of $XI(X < b + n_{k-1}\mu)$, as in Section 3.3, without losing absolute continuity on $\{n_{k-1} < \tau_b \leq n_k, B_k\}$. For estimating $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, B_k\}$, we draw samples of $X_1, \ldots, X_{\tau_b \wedge n_k}$ independently from the distribution $F_k(\cdot)$ satisfying,

$$\frac{dF_k(x)}{dF(x)} = \exp(\theta_k x - \Lambda_k(\theta_k))\mathbf{1}(x < b + n_{k-1}\mu), \quad x \in \mathbb{R};$$

$$\text{here, } \theta_k(= \theta_k(b)) := \frac{-\log(n_k \bar{F}(b + n_{k-1}\mu))}{b + n_{k-1}\mu}, \text{ and} \tag{4.7}$$

$$\Lambda_k(\theta) := \log\left(\int_{-\infty}^{b+n_{k-1}\mu} \exp(\theta_k x) F(dx)\right), \quad \theta \geq 0. \tag{4.8}$$

Let $\mathbb{P}_{k,2}(\cdot)$ be the measure induced by drawing samples as above. Then the resulting likelihood ratio on $\{n_{k-1} < \tau_b \leq n_k, B_k\}$ is:

$$\frac{d\mathbb{P}}{d\mathbb{P}_{k,2}}(X_1, \ldots, X_{n_k}) = \exp\left(-\theta_k S_{\tau_b} + \tau_b \Lambda_k(\theta_k)\right).$$

The associated estimator for computing $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, B_k\}$ is:

$$Z_{k,2}(b) := \exp\left(-\theta_k S_{\tau_b} + \tau_b \Lambda_k(\theta_k)\right) \mathbb{I}(n_{k-1} < \tau_b \leq n_k, B_k). \tag{4.9}$$

Simulating $\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k \cap \bar{B}_k\}$

We draw samples in a two step procedure similar to that in the Section 4.1.

1. Choose an index J uniformly at random from $\{1, \ldots, n_k\}$
2. Simulate the increment X_n from $F(\cdot | X \geq b + n_{k-1}\mu)$, if $n = J$; otherwise, simulate X_n from $F(\cdot)$, for any $n \leq n_k$.

If $\mathbb{P}_{k,3}(\cdot)$ denotes the change of measure induced by drawing samples according to the above procedure, then the likelihood ratio on the set $\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k \cap \bar{B}_k\}$ is:

$$\frac{d\mathbb{P}}{d\mathbb{P}_{k,3}}(X_1, \ldots, X_{n_k}) = \frac{n_k \bar{F}(b + n_{k-1}\mu)}{\#\{X_i \geq b + n_{k-1}\mu : 1 < i \leq n_k\}}.$$

The resulting estimator for the computation of $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k \cap \bar{B}_k\}$ is:

$$Z_{k,3}(b) := \frac{n_k \bar{F}(b + n_{k-1}\mu)}{\#\{X_i \geq b + n_{k-1}\mu : 1 < i \leq n_k\}} \mathbb{I}\left(n_{k-1} < \tau_b \leq n_k, \bar{A}_k \cap \bar{B}_k\right). \tag{4.10}$$

Finally, as in Chapter 3, the estimator for $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}$ can be obtained by summing the estimators of component probabilities $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, A_k\}, \mathbb{P}\{n_{k-1} < \tau_b \leq n_k, B_k\}$, and $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k \cap \bar{B}_k\}$:

$$Z_k(b) := Z_{k,1}(b) + Z_{k,2}(b) + Z_{k,3}(b). \tag{4.11}$$

Simulation of $\{\tau_b < \infty\}$ - the finite variance case

Here we develop on the ideas stated at the beginning of Section 4.1. We have the increasing sequence of integers $(n_k : k \geq 0)$,

$$n_0 = 0, n_k = r^k \text{ for } k \geq 1,$$

for some integer $r > 1$. Further, we have an auxiliary random variable K taking values in positive integers according to the probability mass function $(p_k : k \geq 1)$. As in (4.1), we re-express the quantity of interest as:

$$\mathbb{P}\{\tau_b < \infty\} = \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{P}\{n_{K-1} < \tau_b \leq n_K\}}{p_K} \,\middle|\, K\right]\right].$$

From (4.11), we have estimators $\{Z_k(b) : k \geq 1\}$ that can be used to compute the corresponding probabilities $(\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\} : k \geq 1)$. Consider the following simulation procedure:

1. Draw a sample of K such that $\mathbb{P}\{K = k\} = p_k$.
2. Conditional on the realized value of K,
   - 2a) Generate a realization of $Z_K(b)$ as in Section 4.1.
   - 2b) Return $Z_K(b)/p_K$.

We present the sample mean of the values returned by N independent simulation runs of the above procedure as our final estimate of $\mathbb{P}\{\tau_b < \infty\}$. Let $Q(\cdot)$ denote the probability measure in the path space induced by the generation of increment random variables as a result of this sampling procedure; let $\mathbb{E}^Q[\cdot]$ and $\mathrm{Var}^Q[\cdot]$ be the expectation and variance operator associated with the measure $Q(\cdot)$. Given $b > 0$, the overall unbiased estimator for the computation of $\mathbb{P}\{\tau_b < \infty\}$ is,

$$Z(b) := \frac{Z_K(b)}{p_K}.$$

Note that the number of independent simulation runs needed to achieve a desired relative precision, as in (2.1), is directly related to the sampling variance of $Z(b)$. If $(Z(b) : b > 0)$ offer asymptotically vanishing relative error, we just need $o(\epsilon^{-2}\delta^{-1})$ independent replications of the estimator. However, as pointed in Hammersley and Handscomb [1965], and further justified in Glynn and Whitt [1992], both the variance of an estimator and the expected computational effort required to generate a single sample are important performance measures, and their product can be considered as a 'figure of merit' in comparing performance of algorithms that provide unbiased estimators of $\mathbb{P}\{\tau_b < \infty\}$. For any given b, let ν_b denote the largest index of the

increment random variables $(X_n : n \geq 1)$ considered for simulation in a particular simulation run. The expectation of ν_b, then gives a measure of the expected number of increment random variables generated, and subsequently of the expected computational effort in every simulation run. In particular, the latter may be bounded from above by a constant $C > 0$ times the expectation of ν_b.

In a single run of the above procedure, if the realized value of K is k, we look for estimating $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}$ which does not entail the generation of more than n_k increment random variables, thus ensuring termination. In particular,

$$n_{_{K-1}} \leq \nu_b \leq n_{_K}.$$

The following theorems give a measure of both the variance and the expected computational effort per replication of $Z(b)$ for a specific choice of the probabilities p_k. Recall that $Q(\cdot)$ is the probability measure that governs the law of $Z(b)$ when the random variables $Z_{K,j}(b)$ are generated as explained in previous sections.

In all the theorems that follow it is assumed that the common distribution $F(\cdot)$ of the increments satisfy Assumptions 4.1 and 4.2.

Theorem 4.1. *For*

$$p_k = \frac{\bar{F}_I(b + n_{k-1}\mu) - \bar{F}_I(b + n_k\mu)}{\bar{F}_I(b)}, k \geq 1, \tag{4.12}$$

the family of unbiased estimators $(Z(b) : b > 0)$ *achieves asymptotically vanishing relative error for the computation of* $\mathbb{P}\{\tau_b < \infty\}$, *as* $b \to \infty$; *that is:*

$$\overline{\lim_{b\to\infty}} \frac{\mathrm{Var}^Q\left[Z(b)\right]}{\mathbb{P}\{\tau_b < \infty\}^2} = 0.$$

Theorem 4.2. *If* $\bar{F}(\cdot)$ *is regularly varying with index* $\alpha > 2$, *for the choice of* $p = (p_k : k \geq 1)$ *in* (4.12)*:*

$$\mathbb{E}^Q[\nu_b] \leq \frac{r + o(1)}{\mu(\alpha - 2)} b, \text{ as } b \to \infty.$$

Proofs of both these results are given later in Section 4.2.

Remark 4.1. From Theorem 4.1, we have the vanishing relative error property for computing $\mathbb{P}\{\tau_b < \infty\}$ whenever the increment random variables X_n have finite mean (irrespective of the variance). Therefore we require only $o(\epsilon^{-2}\delta^{-1})$ i.i.d replications of $Z(b)$ to arrive at estimators

that have relative error smaller than ϵ with probability at least $1-\delta$. Now from Theorem 4.2 we conclude that, if the tail index $\alpha > 2$ (in which case the increments have finite variance), our importance sampling methodology estimates $\mathbb{P}\{\tau_b < \infty\}$ in $O(b)$ expected computational effort.

Remark 4.2. From the conditional limit result in (4.5), one can infer that the values p_k as in (4.12) roughly match the zero-variance probability $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k \mid \tau_b < \infty\}$ asymptotically. For tails $\bar{F}(\cdot)$ with regularly varying index $1 < \alpha < 2$, we have that $\mathbb{E}[\tau_b \mid \tau_b < \infty] = \infty$; that is, the zero-variance measure itself has infinite expected termination time! Since the probabilities p_k are assigned a value similar to $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k \mid \tau_b < \infty\}$, one might suspect infinite expected termination time for a single run of Algorithm 1 as well. As we note later in Remark 4.6 after the Proof of Theorem 4.2, for p_ks as in (4.12), this is indeed the case.

Simulation of $\{\tau_b < \infty\}$ - the infinite variance case

As indicated in Remark 4.2, infinite termination time for a simulation algorithm is clearly unacceptable. The following question then is natural: By choosing p_ks differently, even if it means compromising on variance of the estimator, can one achieve finite expected termination time for the procedure in Section 4.1? Before answering this question below, we introduce a family of tail distributions and their integrated counterparts: for any $\beta > 2$, define

$$\bar{G}^{(\beta)}(x) := \frac{\bar{F}(x)}{x^{\beta-\alpha}}, \text{ and } \bar{G}_I^{(\beta)}(x) := \int_x^\infty \bar{G}^{(\beta)}(u)du. \tag{4.13}$$

Theorem 4.3. *If the tail $\bar{F}(\cdot)$ is regularly varying with index $\alpha \in (1.5, 2]$, then for any $\beta \in (2, 2\alpha - 1)$,*

$$p_k = \frac{\bar{G}_I^{(\beta)}(b + n_{k-1}\mu) - \bar{G}_I^{(\beta)}(b + n_k\mu)}{\bar{G}_I^{(\beta)}(b)}, k \geq 1 \tag{4.14}$$

yields a family of unbiased estimators $(Z(b) = Z_{_K}(b)/p_{_K} : b > 0)$ achieving

1. *strong efficiency: $\overline{\lim}_{b\to\infty} \frac{\mathrm{Var}^Q[Z(b)]}{\mathbb{P}\{\tau_b<\infty\}^2} < \infty$, and*
2. *finite expected termination time: $\mathbb{E}^Q[\nu_b] \leq \frac{r+o(1)}{\mu(\beta-2)}b$, as $b \to \infty$.*

Remark 4.3. Because of the strong efficiency, we need just $O(\epsilon^{-2}\delta^{-1})$ i.i.d. replications of $Z(b)$ to achieve the desired relative precision. As in Remark 4.1, due to the bound on $\mathbb{E}[\nu_b]$ in Theorem 4.3, the average computational effort for the entire estimation procedure is just $O(\epsilon^{-2}\delta^{-1}b)$. It is important to see this achievement in the context of Remark 4.2: the induced measure $Q(\cdot)$ deviates from the zero-variance measure such that we get finite expected

termination time, but only at the cost of losing vanishing relative error property to strong efficiency. Thus for the selection of p_ks as in (4.14), the suggested procedure ends up offering superior performance (in terms of computational complexity) compared to the algorithms that tend to just approximate the zero-variance measure.

Given this result, it is difficult not to wonder why the tail index α should be larger than 1.5 in the statement of Theorem 4.3, and what happens when $\alpha \leq 1.5$. The following result shows that it is indeed impossible to have both strong efficiency and finite expected termination time when the tail index $\alpha < 1.5$.

Theorem 4.4. *If the tail index $\alpha < 1.5$, there does not exist an assignment of $(p_k, n_k : k \geq 1)$ such that both $\mathbb{E}^Q[Z^2(b)]$ and $\mathbb{E}^Q[\nu_b]$ are simultaneously finite.*

Remark 4.4. If the tail index $\alpha = 1.5$, the possibility of having both $\mathbb{E}^Q[Z^2(b)]$ and $\mathbb{E}^Q[\nu_b]$ finite will depend on the slowly varying function $L(\cdot)$. As we shall see in the proof of Theorem 4.4,

$$\mathbb{E}^Q[Z^2(b)]\mathbb{E}^Q[\nu_b] = \Omega\left(\int_{b^2}^{\infty} \sqrt{u}\bar{F}(u)du\right),$$

as $b \to \infty$. If $L(x) = O((\log x)^{-m})$, $m \geq 2$, the above integral is finite, whereas if $L(x) = O(\log x)$ it is infinite; and it easily verified that the case of $L(x) = O((\log x)^{-m})$, $m \geq 2$, goes through the proof of Theorem 4.3, thus achieving both strong efficiency and finite expected termination time. This illustrates the subtle dependence on the associated slowly varying function $L(\cdot)$ for the existence of such p_ks and n_ks.

As illustrated by the theorem below, for $\alpha \in (1, 1.5]$, we still have algorithms that demand only $O(b)$ units of expected computer time if we look for less stringent notions of efficiency.

Theorem 4.5. *If the tail $\bar{F}(\cdot)$ is regularly varying with index $\alpha \in (1, 1.5]$, then there exists an explicit selection of $p = (p_k : k \geq 1)$ such that the family of unbiased estimators $(Z(b) : b > 0)$ satisfies both:*

$$\overline{\lim_{b\to\infty}} \frac{\mathbb{E}^Q\left[Z^{1+\gamma}(b)\right]}{\mathbb{P}\{\tau_b < \infty\}^{1+\gamma}} < \infty \text{ for all } \gamma \in \left(0, \frac{\alpha-1}{2-\alpha}\right), \text{ and} \tag{4.15}$$

$$\mathbb{E}^Q[\nu_b] \leq Cb \text{ for some constant } C.$$

In particular, for the following selection of $p = (p_k : k \geq 1)$,

$$p_k = \frac{\bar{G}_I^{(\beta)}(b + n_{k-1}\mu) - \bar{G}_I^{(\beta)}(b + n_k\mu)}{\bar{G}_I^{(\beta)}(b)}, \tag{4.16}$$

and $n_k = r^k$ for $k \geq 1$, if β is chosen in $(2, \alpha + \gamma^{-1}(\alpha - 1))$, both the above inequalities are satisfied.

Remark 4.5. If the estimator $Z(b)$ satisfies (4.15), similar to how we arrived at (2.1), it can be shown that $O(\epsilon^{-(1+\gamma^{-1})}\delta^{-\gamma^{-1}})$ i.i.d. replications of $Z(b)$ are enough to produce estimates having relative error at most ϵ with probability at least $1-\delta$. Now according to Theorem 4.5, the expected termination time in each replication is $O(b)$. Thus with the p_ks chosen as in (4.16), we expend just $O(\epsilon^{-(1+\gamma^{-1})}\delta^{-\gamma^{-1}}b)$ units of computer time on an average, which is still linear in b. The price we pay by not adhering to strong efficiency is the worse dependence on the parameters ϵ and δ.

It is further interesting to note that a vastly different state-dependent methodology developed using Lyapunov inequalities in Blanchet and Liu [2012] also hits identical barriers and provides results similar to ours: They present algorithms that are both strongly efficient and possess $O(b)$ expected termination time for the case of tails having index $\alpha > 1.5$; whereas when $\alpha \in (1, 1.5]$, they provide estimators satisfying (4.15) along with $O(b)$ expected termination time of a simulation run.

4.2 Proofs of key theorems

For proving Theorems 4.1, 4.3 and 4.5, which are on the efficiency of estimators $\{Z(b) : b > 0\}$, we first present a result pertaining to the efficiency of the component estimators $\{Z_k(b) : k \geq 1\}$. Recall from Section 4.1 that

$$Z_k(b) := Z_{k,1}(b) + Z_{k,2}(b) + Z_{k,3}(b)$$

is an unbiased estimator for $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}$, and

$$q_k(b) := \sum_{j=n_{k-1}+1}^{n_k} \bar{F}(b + j\mu).$$

To aid the analysis of second moment of estimators $Z_k(b)$, let $\mathbb{P}_k(\cdot)$ denote the composite measure induced due to the simulation of random variables $Z_{k,j}, j = 1, 2, 3$ independently according to measures $\mathbb{P}_{k,j}, j = 1, 2, 3$, respectively. Let $\mathbb{E}_k[\cdot]$ denote the corresponding expectation operator.

Theorem 4.6. *Under Assumptions 4.1 and 4.2, the family of estimators* $\{Z_k(b) : k \geq 1, b > 0\}$ *satisfies the following as* $b \to \infty$:

$$\sup_{k:n_k<b^\eta} \frac{\mathbb{E}_k\left[Z_k^2(b)\right]}{q_k^2(b)} \leq 1 + o(1) \text{ and } \sup_{k:n_k\geq b^\eta} \frac{\mathbb{E}_k\left[Z_k^2(b)\right]}{q_k^2(b)} \leq c$$

for some $c > 0$ *and* $\eta > 1$.

We prove Theorem 4.6 by analysing the second moment of estimators $Z_{k,1}(\cdot), Z_{k,2}(\cdot)$ and $Z_{k,3}(\cdot)$ separately in the Lemmas 4.1, 4.5 and 4.6 below.

Lemma 4.1. *Under Assumption 4.1,*

$$\sup_k \frac{\mathbb{E}_{k,1}\left[Z_{k,1}^2(b)\right]}{q_k^2(b)} \leq 1.$$

Proof. Recall that $\mathbb{P}_{k,1}(\cdot)$ is the measure resulting due to the simulation of increments as in the two-step procedure specified in Section 4.1. Since the quantity $\#\{X_i \geq b + i\mu : n_{k-1} < i \leq n_k\}$ is at least 1 when the increments are generated from $\mathbb{P}_{k,1}(\cdot)$, we have $Z_{k,1}(b) \leq q_k(b)$. Therefore,

$$\mathbb{E}_{k,1}\left[Z_{k,1}^2(b)\right] \leq q_k^2(b), \tag{4.17}$$

which proves the claim. □

For a similar analysis on the second moment of estimators $Z_{k,2}(b)$ and $Z_{k,3}(b)$, we need the following results which are proved in Section 4.5.

Lemma 4.2. *Under Assumption 4.1, there exists a constant $c_1 > 1$ such that $\exp(n_k\Lambda_k(\theta_k)) \leq c_1$ for all k, b.*

Lemma 4.3. *Under Assumption 4.1, there exists a positive constant c_2 such that,*

$$\sup_{k\geq 1, b>0} \frac{n_k \bar{F}(b + n_{k-1}\mu)}{q_k(b)} \leq c_2.$$

Proposition 4.1. *Under Assumption 4.1,*

$$\sup_{k\geq 1} \left| \frac{\mathbb{P}\left\{n_{k-1} < \tau_b \leq n_k, A_k\right\}}{q_k(b)} - 1 \right| = O\left(b^{-\frac{\alpha-1}{2\alpha}}\right),$$

as $b \to \infty$.

Lemma 4.4. *Under Assumptions 4.1 and 4.2, there exist constants $\eta > 1$ and c_3 such that,*

$$\sup_{k:n_k<b^\eta} \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k\}}{q_k(b)} = o(1) \text{ and}$$
$$\sup_{k:n_k\geq b^\eta} \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k\}}{q_k(b)} \leq c_3,$$

as $b \to \infty$.

A proof of Proposition 4.1 and Lemma 4.4 can be found in Section 4.5. Using Lemmas 4.2, 4.3 and 4.4, we now present an asymptotic analysis on the second moment of estimators $Z_{k,2}(\cdot)$ and $Z_{k,3}(\cdot)$.

Lemma 4.5. *Under Assumptions 4.1 and 4.2, as* $b \to \infty$,

$$\sup_{k:n_k<b^\eta} \frac{\mathbb{E}_{k,2}\left[Z_{k,2}^2(b)\right]}{q_k^2(b)} = o(1) \text{ and } \sup_{k:n_k\geq b^\eta} \frac{\mathbb{E}_{k,2}\left[Z_{k,2}^2(b)\right]}{q_k^2(b)} \leq c_4$$

for some positive constant c_4.

Proof. Since $\tau_b \leq n_k$ on the event $\{n_{k-1} < \tau_b \leq n_k\}$,

$$\exp\left(\tau_b \Lambda_k(\theta_k)\right) \mathbb{I}(n_{k-1} < \tau_b \leq n_k, B_k) \leq c_1,$$

because of Lemma 4.2. Further note that $\theta_k S_{\tau_b} \geq -\log(n_k \bar{F}(b + n_{k-1}\mu))$ on $\{n_{k-1} < \tau_b \leq n_k\}$. Therefore from (4.9),

$$Z_{k,2}(b) \leq c_1 \left(n_k \bar{F}(b + n_{k-1}\mu)\right) \mathbb{I}(n_{k-1} < \tau_b \leq n_k, B_k), \text{ for all } k.$$

Now changing the expectation operator in the evaluation of second moment of the estimator results in the following bound: for all k,

$$\mathbb{E}_{k,2}\left[Z_{k,2}^2(b)\right] = \mathbb{E}\left[Z_{k,2}(b)\right] \leq c_1 \left(n_k \bar{F}(b + n_{k-1}\mu)\right) \mathbb{P}\{n_{k-1} < \tau_b \leq n_k, B_k\}.$$

$$\text{Therefore } \frac{\mathbb{E}_{k,2}\left[Z_{k,2}^2(b)\right]}{q_k^2(b)} \leq c_1 \frac{\left(n_k \bar{F}(b + n_{k-1}\mu)\right)}{q_k(b)} \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k\}}{q_k(b)}.$$

Then it follows from Lemmas 4.3 and 4.4 that, as $b \to \infty$,

$$\sup_{k:n_k<b^\eta} \frac{\mathbb{E}_{k,2}\left[Z_{k,2}^2(b)\right]}{q_k^2(b)} = o(1), \text{ and } \sup_{k:n_k\geq b^\eta} \frac{\mathbb{E}_{k,2}\left[Z_{k,2}^2(b)\right]}{q_k^2(b)} \leq c_1 c_2 c_3 =: c_4 < \infty,$$

thus proving the claim. □

Lemma 4.6. *Under Assumptions 4.1 and 4.2, as* $b \to \infty$,

$$\sup_{k:n_k<b^\eta} \frac{\mathbb{E}_{k,3}\left[Z_{k,3}^2(b)\right]}{q_k^2(b)} = o(1) \text{ and } \sup_{k:n_k\geq b^\eta} \frac{\mathbb{E}_{k,3}\left[Z_{k,3}^2(b)\right]}{q_k^2(b)} \leq c_5$$

for some positive constant c_5.

Proof. When the increments are generated as prescribed in the two-step procedure in Section 4.1, we have $\#\{X_i \geq b + n_{k-1}\mu : 1 < i \leq n_k\} \geq 1$, and hence,

$$Z_{k,3}(b) \leq n_k \bar{F}(b + n_{k-1}\mu)\mathbb{I}\left(n_{k-1} < \tau_b \leq n_k, \bar{A}_k \cap \bar{B}_k\right).$$

Now a bound on the second moment of the estimator can be obtained as before:

$$\mathbb{E}_{k.3}\left[Z_{k,3}^2(b)\right] = \mathbb{E}\left[Z_{k.3}(b)\right] \leq n_k \bar{F}(b + n_{k-1}\mu)\mathbb{P}\left\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k \cap \bar{B}_k\right\}.$$

$$\text{Therefore } \frac{\mathbb{E}_{k,3}\left[Z_{k,3}^2(b)\right]}{q_k^2(b)} \leq \frac{\left(n_k \bar{F}(b + n_{k-1}\mu)\right)}{q_k(b)} \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k\}}{q_k(b)}.$$

Then it follows from Lemmas 4.3 and 4.4 that, as $b \to \infty$,

$$\sup_{k:n_k<b^\eta} \frac{\mathbb{E}_{k,3}\left[Z_{k,3}^2(b)\right]}{q_k^2(b)} = o(1), \text{ and } \sup_{k:n_k\geq b^\eta} \frac{\mathbb{E}_{k,3}\left[Z_{k,3}^2(b)\right]}{q_k^2(b)} \leq c_2 c_3 =: c_5 < \infty,$$

thus establishing the claim. □

Proof of Theorem 4.6 Since $\{Z_{k,i}(b) : i = 1, 2, 3\}$ are independent, for $i \neq 1$,

$$\begin{aligned}\frac{\mathbb{E}_k\left[Z_{k,1}(b)Z_{k,i}(b)\right]}{q_k^2(b)} &= \frac{\mathbb{E}_{k,1}\left[Z_{k,1}(b)\right]}{q_k(b)} \frac{\mathbb{E}_{k,i}\left[Z_{k,i}(b)\right]}{q_k(b)} \\ &\leq \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, A_k\}}{q_k(b)} \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k\}}{q_k(b)}.\end{aligned}$$

Then from Proposition 4.1 and Lemma 4.4, we have that as $b \to \infty$,

$$\sup_{k:n_k<b^\eta} \frac{\mathbb{E}_k\left[Z_{k,1}(b)Z_{k,i}(b)\right]}{q_k^2(b)} = o(1), \text{ and } \sup_{k:n_k\geq b^\eta} \frac{\mathbb{E}_k\left[Z_{k,1}(b)Z_{k,i}(b)\right]}{q_k^2(b)} < \infty.$$

Similarly from Lemma 4.4, as $b \to \infty$,

$$\sup_k \frac{\mathbb{E}_k\left[Z_{k,2}(b)Z_{k,3}(b)\right]}{q_k^2(b)} \leq \sup_k \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k\}^2}{q_k^2(b)} = o(1).$$

Since $Z_k(b) = Z_{k,1}(b) + Z_{k,2}(b) + Z_{k,3}(b)$, we have

$$\mathbb{E}_k\left[Z_k^2(b)\right] = \sum_{i,j=1}^{3} \mathbb{E}_k\left[Z_{k,i}(b)Z_{k,j}(b)\right].$$

Combining above observations with the results of Lemmas 4.1, 4.5 and 4.6, we conclude that as $b \to \infty$,

$$\sup_{k:n_k<b^\eta} \frac{\mathbb{E}_k\left[Z_k^2(b)\right]}{q_k^2(b)} \leq 1 + o(1) \text{ and } \sup_{k:n_k\geq b^\eta} \frac{\mathbb{E}_k\left[Z_k^2(b)\right]}{q_k^2(b)} \leq c$$

for some positive constant c. □

The following uniform bounds will be useful:

Lemma 4.7. *For all $k \geq 1$,*

$$q_k(b) \leq \frac{1}{\mu}\left(\bar{F}_I(b + n_{k-1}\mu) - \bar{F}_I(b + n_k\mu)\right).$$

Further as $b \to \infty$,

$$q_k(b) \geq (1 - o(1))\frac{1}{\mu}\left(\bar{F}_I(b + n_{k-1}\mu) - \bar{F}_I(b + n_k\mu)\right),$$

uniformly in k.

Proof. For any $k \geq 1$,

$$q_k(b) = \sum_{i=n_{k-1}+1}^{n_k} \bar{F}(b + i\mu) \leq \sum_{i=n_{k-1}+1}^{n_k} \int_{i-1}^{i} \bar{F}(b + u\mu)du = \int_{n_{k-1}}^{n_k} \bar{F}(b + u\mu)du.$$

Changing variables from u to $v = b + u\mu$ results in,

$$q_k(b) \leq \frac{1}{\mu}\int_{b+n_{k-1}\mu}^{b+n_k\mu} \bar{F}(v)dv,$$

which establishes the upper bound because $\bar{F}_I(x) := \int_x^\infty \bar{F}(u)du$.

For the lower bound, see that

$$\begin{aligned} q_k(b) &= \sum_{i=n_{k-1}+1}^{n_k} \bar{F}(b + i\mu) \geq \sum_{i=n_{k-1}+1}^{n_k} \int_{i}^{i+1} \bar{F}(b + u\mu)du \\ &= \int_{n_{k-1}+1}^{n_k+1} \bar{F}(b + u\mu)du. \end{aligned}$$

Now after changing variables from u to $v = b + u\mu$, we use the long-tailedness[†] of $\bar{F}_I(\cdot)$ to see that, given $\epsilon > 0$, for large values of b,

$$\begin{aligned} q_k(b) &\geq \frac{1}{\mu}\left(\bar{F}_I(b + (n_{k-1} + 1)\mu) - \bar{F}_I(b + (n_k + 1)\mu)\right) \\ &\geq (1 - \epsilon)\frac{1}{\mu}\left(\bar{F}_I(b + n_{k-1}\mu) - \bar{F}_I(b + n_k\mu)\right) \end{aligned}$$

for all k. □

Proof of Theorem 4.1 Recall that the overall estimator is,

$$Z(b) = \frac{Z_K(b)}{p_K},$$

[†]refer (2.6)

where p_k is as in (4.12). Second moment of the estimator $Z(b)$ is bounded as below:

$$\begin{aligned}\mathbb{E}^Q[Z^2(b)] &= \mathbb{E}^Q\left[\left(\frac{Z_K(b)}{p_K}\right)^2\right] \\ &= \mathbb{E}^Q\left[\mathbb{E}^Q\left[\frac{Z_K^2(b)}{q_K^2(b)}\frac{q_K^2(b)}{p_K^2}; n_K < b^\eta \;\middle|\; K\right]\right] \\ &\quad + \mathbb{E}^Q\left[\mathbb{E}^Q\left[\frac{Z_K^2(b)}{q_K^2(b)}\frac{q_K^2(b)}{p_K^2}; n_K \geq b^\eta \;\middle|\; K\right]\right] \end{aligned} \tag{4.18}$$

From the definition of p_k and Lemma 4.7, we have $q_K^2(b) \leq \bar{F}_I^2(b)p_K^2$. Combining this with Theorem 4.6 it follows that,

$$\begin{aligned}\frac{\mathbb{E}^Q[Z^2(b)]}{\bar{F}_I^2(b)} &\leq \mathbb{E}^Q\left[\mathbb{E}^Q\left[\frac{Z_K^2(b)}{q_K^2(b)}; n_K < b^\eta \;\middle|\; K\right]\right] + \mathbb{E}^Q\left[\mathbb{E}^Q\left[\frac{Z_K^2(b)}{q_K^2(b)}; n_K \geq b^\eta \;\middle|\; K\right]\right] \\ &\leq 1 + o(1) + c\mathbb{P}\{n_K \geq b^\eta\} \\ &\leq 1 + o(1) + O\left(\frac{\bar{F}_I(b+b^\eta)}{\bar{F}_I(b)}\right) = 1 + o(1),\end{aligned}$$

as $b \to \infty$. The last inequality follows from observing that $\mathbb{P}\{n_K \geq b^\eta\} = \sum_{k:n_k \geq b^\eta} p_k$. Since $\eta > 1$, we have the asymptotically vanishing relative error property of the estimators $(Z(b) : b > 0)$. □

Proof of Theorem 4.2 Recall that ν_b denotes the maximum of indices of the increment random variables (X_i's) considered for simulation in a particular simulation run. From the sampling procedures in Section 4.1, it is clear that $\nu_b \leq n_K$. Therefore,

$$\begin{aligned}\mathbb{E}^Q[\nu_b] &\leq \sum_{k\geq 1} p_k n_k \\ &= rp_1 + \sum_{k\geq 2} r^k p_k \\ &= \frac{1}{\bar{F}_I(b)}\left(r\int_b^{b+r\mu}\bar{F}(u)du + \sum_{k\geq 1} r^{k+1}\int_{b+r^k\mu}^{b+r^{k+1}\mu}\bar{F}(u)du\right).\end{aligned} \tag{4.19}$$

$$\begin{aligned}\text{Since } r^k\int_{b+r^k\mu}^{b+r^{k+1}\mu}\bar{F}(u)du &= \frac{b+r^k\mu-b}{\mu}\int_{b+r^k\mu}^{b+r^{k+1}\mu}\bar{F}(u)du \\ &\leq \frac{1}{\mu}\left(\int_{b+r^k\mu}^{b+r^{k+1}\mu} u\bar{F}(u)du - b\int_{b+r^k\mu}^{b+r^{k+1}\mu}\bar{F}(u)du\right),\end{aligned}$$

$$\begin{aligned}
\text{we write } \sum_{k\geq 1} r^{k+1} \int_{b+r^k\mu}^{b+r^{k+1}\mu} \bar{F}(u)du \\
&\leq \frac{r}{\mu} \sum_{k\geq 1} \left(\int_{b+r^k\mu}^{b+r^{k+1}\mu} u\bar{F}(u)du - b \int_{b+r^k\mu}^{b+r^{k+1}\mu} \bar{F}(u)du \right) \\
&= \frac{r}{\mu} \left(\int_{b+r\mu}^{\infty} u\bar{F}(u)du - \int_{b+r\mu}^{\infty} \bar{F}(u)du \right) \qquad (4.20) \\
&\leq \frac{r+o(1)}{\mu} \left(\frac{(b+r\mu)^2}{\alpha-2} - b\frac{b+r\mu}{\alpha-1} \right) \bar{F}(b+r\mu), \\
&= \frac{r+o(1)}{\mu(\alpha-1)(\alpha-2)} b^2\bar{F}(b), \text{ as } b\to\infty.
\end{aligned}$$

where the penultimate step follows from Karamata's theorem (see (2.8)), and the final step just uses long-tailed nature of $\bar{F}(\cdot)$. Also note that: $\int_b^{b+r\mu} \bar{F}(u)du \leq r\mu\bar{F}(b)$, and by application of Karamata's theorem, we have $\bar{F}_I(b) \sim bF(b)/(\alpha-1)$, as $b\to\infty$. Therefore from (4.19),

$$\mathbb{E}^Q[\nu_b] \leq \frac{r+o(1)}{\mu(\alpha-2)} b, \text{ as } b\to\infty,$$

thus yielding the required bound on the expected termination time. □

Remark 4.6. Similar to how we arrived at (4.20), lower bounds can be obtained to show that $\mathbb{E}^Q[\nu_b] = \Omega\left(\int_b^\infty u\bar{F}(u)du\right)$. If the tail index $\alpha < 2$, $\int_b^\infty u\bar{F}(u)du$ turns out to be infinite, and subsequently $\mathbb{E}^Q[\nu_b] = \infty$. Though the assignment of p_ks in (4.12) yields vanishing relative error for any $\alpha > 1$, it fails to provide algorithms which have finite expected termination time when the increment random variables X have infinite variance (e.g., when $\alpha < 2$), thus making this choice of p_k not suitable for practice.

Proof of Theorem 4.3 We obtain upper bounds for both the variance of the estimator $Z(b)$ and the expected termination time.

1. *Variance of $Z(b)$:* Since $Q(K=k) = p_k$,

$$\begin{aligned}
\mathbb{E}^Q[Z^2(b)] = \mathbb{E}^Q\left[\frac{Z_K^2(b)}{p_K^2}\right] &= \sum_k p_k \frac{\mathbb{E}^Q[Z_k^2(b)]}{p_k^2} \qquad (4.21) \\
&= \sum_k \frac{\mathbb{E}^Q[Z_k^2(b)]}{q_k^2(b)} \frac{q_k^2(b)}{p_k}. \qquad (4.22)
\end{aligned}$$

Following Lemma 4.7 and the assignment of p_ks as in (4.14), we can write,

$$\frac{q_k(b)}{p_k} \leq \frac{\bar{F}_I(b+n_{k-1}\mu) - \bar{F}_I(b+n_k\mu)}{\bar{G}_I^{(\beta)}(b+n_{k-1}\mu) - \bar{G}_I^{(\beta)}(b+n_k\mu)} \bar{G}_I^{(\beta)}(b).$$

To obtain an upper bound, we note the following:

$$\begin{aligned}
\bar{F}_I(b+n_{k-1}\mu) - \bar{F}_I(b+n_k\mu) &= \int_{b+n_{k-1}\mu}^{b+n_k\mu} \bar{F}(u)du \\
&\leq (n_k - n_{k-1})\mu \bar{F}(b+n_{k-1}\mu), \\
\bar{G}_I^{(\beta)}(b+n_{k-1}\mu) - \bar{G}_I^{(\beta)}(b+n_k\mu) &= \int_{b+n_{k-1}\mu}^{b+n_k\mu} \bar{G}^{(\beta)}(u)du \\
&\geq (n_k - n_{k-1})\mu \bar{G}^{(\beta)}(b+n_k\mu), \text{ and} \\
\frac{\bar{G}^{(\beta)}(b+n_{k-1}\mu)}{\bar{G}^{(\beta)}(b+n_k\mu)} &\leq r^\beta + o(1), \text{ as } b \to \infty.
\end{aligned}$$

The last inequality follows by observing that $b + n_k\mu \leq r(b + n_{k-1}\mu)$ and subsequently from the regularly varying nature of $\bar{G}^{(\beta)}(\cdot)$. Therefore as $b \to \infty$,

$$\begin{aligned}
\frac{q_k(b)}{p_k} &\leq \frac{\bar{G}^{(\beta)}(b+n_{k-1}\mu)}{\bar{G}^{(\beta)}(b+n_k\mu)} \frac{\bar{F}(b+n_{k-1}\mu)}{\bar{G}^{(\beta)}(b+n_{k-1}\mu)} \bar{G}_I^{(\beta)}(b) \\
&= (r^\beta + o(1))(b+n_{k-1}\mu)^{\beta-\alpha} \bar{G}_I^{(\beta)}(b), \qquad (4.23)
\end{aligned}$$

for all k, because $\bar{F}(x)/\bar{G}^{(\beta)}(x) = x^{\beta-\alpha}$. Combining this with Theorem 4.6, it follows from (4.22) that

$$\begin{aligned}
\mathbb{E}^Q[Z^2(b)] &\leq (cr^\beta + o(1))\bar{G}_I^{(\beta)}(b) \sum_k (b+n_{k-1}\mu)^{\beta-\alpha} q_k(b) \\
&\leq (cr^\beta + o(1))\bar{G}_I^{(\beta)}(b) \sum_k (b+n_{k-1}\mu)^{\beta-\alpha} \int_{b+n_{k-1}\mu}^{b+n_k\mu} \bar{F}(u)du, \\
&\leq (cr^\beta + o(1))\bar{G}_I^{(\beta)}(b) \sum_k \int_{b+n_{k-1}\mu}^{b+n_k\mu} u^{\beta-\alpha} \bar{F}(u)du \\
&\leq (cr^\beta + o(1))\bar{G}_I^{(\beta)}(b) \int_b^\infty u^{\beta-\alpha} \bar{F}(u)du
\end{aligned}$$

as $b \to \infty$. Since $2\alpha - \beta > 1$, it follows from Karamata's theorem (cf. (2.8)) that

$$\mathbb{E}^Q[Z^2(b)] \leq (cr^\beta + o(1))\bar{G}_I^{(\beta)}(b) b^{\beta-\alpha+1} \frac{\bar{F}(b)}{2\alpha - \beta - 1}, \text{ as } b \to \infty.$$

Further $(\alpha-1)\bar{F}_I(b) \sim b\bar{F}(b)$ and $b^{\beta-\alpha}\bar{G}_I^{(\beta)}(b) \sim \bar{F}_I(b)$, as $b \to \infty$. Therefore,

$$\overline{\lim_{b\to\infty}} \frac{\mathbb{E}^Q[Z^2(b)]}{\bar{F}_I^2(b)} \leq \frac{(\alpha-1)cr^\beta + o(1)}{2\alpha - \beta - 1} < \infty.$$

Now since $\mathbb{P}\{\tau_b < \infty\} \sim \mu^{-1}\bar{F}_I(b)$, we have strong efficiency.

2. *Expected termination time:* Since $\nu_b \leq n_{_K}, \mathbb{E}^Q[\nu_b] \leq \mathbb{E}^Q[n_{_K}] = \sum_k p_k n_k$. For the choice of p_k in (4.14), following exactly the same steps in the proof of Theorem 4.2, we arrive at:

$$\mathbb{E}^Q[\nu_b] \leq \frac{r}{\mu}\left(\mu \int_b^{b+r\mu} \bar{G}^{(\beta)}(u)du + \int_{b+r\mu}^{\infty} u\bar{G}^{(\beta)}(u)du - b\int_{b+r\mu}^{\infty} \bar{G}^{(\beta)}(u)du\right).$$

Since $\bar{G}^{(\beta)}(\cdot)$ is regularly varying with tail index larger than 2, by application of Karamata's theorem, we have:

$$\int_{b+r\mu}^{\infty} u\bar{G}^{(\beta)}(u)du \sim \frac{(b+r\mu)^2}{\beta-2}\bar{G}^{(\beta)}(b+r\mu),$$

which would not have been the case if we had persisted with using $\bar{F}_I(\cdot)$ instead of $\bar{G}_I^{(\beta)}(\cdot)$ for p_k. Again following the remaining steps in the proof of Theorem 4.2, we conclude that:

$$\mathbb{E}^Q[\nu_b] \leq \frac{r+o(1)}{\mu(\beta-2)}b, \text{ as } b \to \infty,$$

thus yielding finite termination time even when the zero-variance measure fails to offer this desirable property. □

Proof of Theorem 4.4 Since $Q(K=k) = p_k$, see that:

$$\mathbb{E}^Q[Z^2(b)] = \mathbb{E}^Q\left[\frac{Z_K^2(b)}{p_K^2}\right] = \sum_k \frac{\mathbb{E}^Q[Z_k^2(b)]}{p_k} \geq \sum_k \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}^2}{p_k},$$

because of Jensen's inequality. To arrive at a contradiction, let us assume that both $\mathbb{E}^Q[Z^2(b)]$ and $\mathbb{E}^Q[\nu_b]$ are finite. Then,

$$\begin{aligned}
\mathbb{E}^Q[Z^2(b)]\mathbb{E}^Q[\nu_b] &\geq \left(\sum_k \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}^2}{p_k}\right)\left(\sum_k p_k n_k\right) \\
&\geq \left(\sum_k \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}}{\sqrt{p_k}} \cdot \sqrt{p_k n_k}\right)^2 \\
&= \left(\sum_k \sqrt{n_k}\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}\right)^2. \qquad (4.24)
\end{aligned}$$

where the penultimate step follows from Cauchy-Schwarz inequality. Then from Proposition 4.1 and Lemma 4.7, it is immediate that

$$\begin{aligned}
\sum_k \sqrt{n_k}\ \mathbb{P}\{n_{k-1} < \tau_b \leq n_k\} &\geq (1-o(1))\sum_k \sqrt{n_k}\ q_k(b) \\
&\geq (1-o(1))\sum_k \sqrt{n_k}\int_{n_{k-1}\mu}^{n_k\mu} \bar{F}(b+u)du \\
&\geq \frac{1-o(1)}{\sqrt{\mu}}\sum_k \int_{n_{k-1}\mu}^{n_k\mu} \sqrt{u}\bar{F}(b+u)du \\
&= \frac{1-o(1)}{\sqrt{\mu}}\int_0^{\infty} \sqrt{u}\bar{F}(b+u)du.
\end{aligned}$$

Now it can be seen easily that the RHS is finite only when $\alpha \geq 1.5$, via the following change of variable and the subsequent integration of the resulting regularly varying tail:

$$\begin{aligned}\int_0^\infty \sqrt{u}\bar{F}(b+u)du &= \int_b^\infty \sqrt{u-b}\bar{F}(u)du \\ &\geq \int_{b^2}^\infty \sqrt{u}\cdot\sqrt{1-\frac{b}{u}}\bar{F}(u)du \\ &\geq \sqrt{1-\frac{1}{b}}\int_{b^2}^\infty \sqrt{u}\bar{F}(u)du,\end{aligned}$$

which cannot be finite if $\alpha < 1.5$, thus arriving at the desired contradiction. Therefore from (4.24), we conclude that we cannot have both the second moment of $Z(b)$ and the expected termination time $\mathbb{E}^Q[\nu_b]$ to be simultaneously finite if the tail index $\alpha < 1.5$. □

Proof of Theorem 4.5 The proof is similar to that of Theorem 4.3, and we provide only an outline of the steps involved. Since $Q(K=k) = p_k$,

$$\begin{aligned}\mathbb{E}^Q[Z^{1+\gamma}(b)] = \mathbb{E}^Q\left[\frac{Z_K^{1+\gamma}(b)}{p_K^{1+\gamma}}\right] &= \sum_k \frac{\mathbb{E}_k\left[Z_k^{1+\gamma}(b)\right]}{p_k^{1+\gamma}}p_k \\ &\leq \sum_k \left(\frac{\mathbb{E}_k\left[Z_k^2(b)\right]}{q_k^2(b)}\right)^{\frac{1+\gamma}{2}}\left(\frac{q_k(b)}{p_k}\right)^\gamma q_k(b)\end{aligned}$$

Now from Theorem 4.6 and (4.23) , following the routine calculation in the proof of Theorem 4.3, we deduce that

$$\begin{aligned}\mathbb{E}^Q[Z^{1+\gamma}(b)] &\leq \left(c^{\frac{1+\gamma}{2}}r^{\beta\gamma}+o(1)\right)\left(\bar{G}_I^{(\beta)}(b)\right)^\gamma \sum_k (b+n_{k-1}\mu)^{\gamma(\beta-\alpha)}q_k(b) \\ &\leq \left(c^{\frac{1+\gamma}{2}}r^{\beta\gamma}+o(1)\right)\left(\bar{G}_I^{(\beta)}(b)\right)^\gamma \int_b^\infty u^{\gamma(\beta-\alpha)}\bar{F}(u)du,\end{aligned}$$

as $b\to\infty$. Since β is smaller than $\alpha+\gamma^{-1}(\alpha-1)$ as in the statement of Theorem 4.5, the tail index of the integrand, $\alpha-\gamma(\beta-\alpha) > 1$. Therefore we can apply Karamata's theorem to conclude that

$$\mathbb{E}^Q[Z^{1+\gamma}(b)] \leq \left(c^{\frac{1+\gamma}{2}}r^{\beta\gamma}+o(1)\right)\left(\bar{G}_I^{(\beta)}(b)\right)^\gamma \frac{b^{\gamma(\beta-\alpha)+1}}{\alpha-\gamma(\beta-\alpha)-1}\bar{F}(b), \text{ as } b\to\infty.$$

Now observing that $(\alpha-1)\bar{F}_I(b) \sim b\bar{F}(b), b^{\beta-\alpha}\bar{G}_I^{(\beta)}(b) \sim \bar{F}_I(b)$, and $\mathbb{P}\{\tau_b<\infty\} \sim \mu^{-1}\bar{F}_I(b)$ as $b\to\infty$, we have:

$$\overline{\lim_{b\to\infty}}\frac{\mathbb{E}^Q[Z^{1+\gamma}(b)]}{\mathbb{P}\{\tau_b<\infty\}^{1+\gamma}} \leq \frac{\mu^2(\alpha-1)c^{\frac{1+\gamma}{2}}r^{\beta\gamma}+o(1)}{\alpha-\gamma(\beta-\alpha)-1} < \infty.$$

Since β is ensured to be larger than 2, the same proof for $\mathbb{E}^Q[\nu_b] = O(b)$ goes through. □

4.3 Simulation of $\{\tau_b < \tau\}$

Let $X, X_1, X_2, \ldots$ be an iid collection of random variables satisfying the following assumption:

Assumption 4.3. *The tail probabilities of X are given by $\bar{F}(x) := \mathbb{P}\{X > x\} = x^{-\alpha}L(x)$, for some slowly varying function $L(\cdot)$ and $\alpha > 2$. Further, $\mu := -\mathbb{E}X > 0$.*

As usual, let $S_0 = 0, S_n = X_1 + \ldots + X_n$, for $n \geq 1$. Further, let $M_n = \max_{k\leq n} S_k, \tau = \inf\{n \geq 1 : S_n \leq 0\}$ and $\tau_b = \inf\{n \geq 1 : S_n > x\}$ for given $b > 0$. Our aim is to estimate the probability that the negative drift random walk $(S_n : n \geq 0)$ starting at 0 crosses a large positive level b before it hits 0 again (that is, within regeneration time τ). In other words, we aim to efficiently estimate the tail probabilities of busy cycle maximum M_τ. From the definitions, it is immediate to see that the tail probability $\mathbb{P}\{M_\tau > b\}$ is the same as $\mathbb{P}\{\tau_b < \tau\}$. Under Assumption 4.3, it is well-known that (see, for example, Theorem 2.1 of Asmussen [1998])

$$\mathbb{P}\{\tau_b < \tau\} \sim \mathbb{E}\tau\bar{F}(b), \text{ as } b \to \infty. \tag{4.25}$$

Simulation methodology

As in the simulation of $\{S_n > b\}$ in Chapter 3, we partition the probability of interest into dominant and residual components as below:

$$\mathbb{P}\left\{\tau_b < \tau\right\} = \mathbb{P}\left\{\tau_b < \tau, \max_{k\leq\tau_b} X_k > b\right\} + \mathbb{P}\left\{\tau_b < \tau, \max_{k\leq\tau_b} X_k \leq b\right\}.$$

Since $S_n > 0$ for all $n < \tau$, the first component has a simple representation:

$$\begin{aligned}
\mathbb{P}\left\{\tau_b < \tau, \max_{k\leq\tau_b} X_k > b\right\} &= \mathbb{P}\left\{\tau_b < \tau, X_{\tau_b} > b\right\} \\
&= \sum_{n=1}^{\infty} \mathbb{P}\left\{S_i \in (0, b] \text{ for } i = 1, \ldots, n-1, X_n > b\right\} \\
&= \sum_{n=1}^{\infty} \mathbb{P}\left\{S_i \in (0, b] \text{ for } i = 1, \ldots, n-1\right\} \bar{F}(b) \\
&= \bar{F}(b) \sum_{n=1}^{\infty} \mathbb{P}\left\{\tau_b \wedge \tau > n-1\right\} \\
&= \mathbb{E}\left[\tau_b \wedge \tau\right] \bar{F}(b).
\end{aligned}$$

Therefore to estimate $\mathbb{P}\left\{\tau_b < \tau, \max_{k\leq\tau_b} X_k > b\right\}$, we draw samples of increments X_n naively from the distribution $F(\cdot)$, and compute the following as the estimator:

$$Z_{\text{dom}}(b) := (\tau_b \wedge \tau)\bar{F}(b). \tag{4.26}$$

Now it is straightforward to see that

$$\mathbb{E}\left[Z_{\text{dom}}\right] = \mathbb{P}\left\{\tau_b < \tau, \max_{k\le\tau_b} X_k > b\right\},$$
$$\text{Var}\left[Z_{\text{dom}}\right] = \text{Var}\left[\tau_b \wedge \tau\right] \bar{F}^2(b),$$

and hence, due to (4.25) and monotone convergence,

$$\overline{\lim_{b\to\infty}} \frac{\text{Var}\left[Z_{\text{dom}}\right]}{\mathbb{P}\left\{\tau_b < \tau\right\}^2} = \overline{\lim_{b\to\infty}} \frac{\text{Var}\left[\tau_b \wedge \tau\right]}{\mathbb{E}\left[\tau\right]^2} = \frac{\text{Var}\left[\tau\right]}{\mathbb{E}\left[\tau\right]^2}. \tag{4.27}$$

To estimate the residual probability $\mathbb{P}\left\{\tau_b < \tau, \max_{k\le\tau_b} X_k \le b\right\}$, we perform exponential twisting as in Section 3.3. Draw samples of $\{X_n : n \le \tau_b \wedge \tau\}$ independently from $F_\theta(\cdot)$ given by:

$$\frac{dF_\theta}{dF}(x) = \exp\left(\theta_b x - \Lambda_b(\theta_b)\right) \mathbf{1}(x \le b), \tag{4.28}$$

where

$$\Lambda_b(\theta) := \log\left(\int_{-\infty}^{b} \exp\left(\theta x\right) F(dx)\right) \text{ for } \theta > 0, \text{ and}$$
$$\theta_b := -\frac{\log b\bar{F}(b)}{b}.$$

Then the resulting estimator is given by

$$Z_{\text{res}}(b) := \exp\left(-\theta_b S_{\tau_b} + \tau_b \Lambda_b(\theta_b)\right) \mathbb{I}\left(\tau_b < \tau, \max_{k\le\tau_b} X_k \le b\right). \tag{4.29}$$

For proving efficiency results of $Z_{\text{res}}(b)$, we shall need Proposition 4.2 and Lemma 4.8, whose proofs are presented in Section 4.5.

Proposition 4.2. *Under Assumption 4.3,*

$$\mathbb{P}\left\{\tau_b < \tau, \max_{k\le\tau_b} X_k \le b\right\} = O\left(\frac{\bar{F}(b)}{b}\right), \text{ as } b \to \infty.$$

Lemma 4.8. *Under Assumption 4.3, we have that*

$$\overline{\lim_{b\to\infty}} \sup_{n\ge1} \exp\left(n\Lambda_b(\theta_b)\right) \le 1.$$

Let $\mathbb{P}_\theta(\cdot)$ and $\mathbb{E}_\theta[\cdot]$ denote the probability measure and the corresponding expectation operator when the increments X_n are drawn independent from $F_\theta(\cdot)$. Since $S_{\tau_b} > b$, it follows from the definition of θ_b and (4.29) that

$$\begin{aligned}\mathbb{E}_\theta\left[Z_{\text{res}}^2(b)\right] = \mathbb{E}\left[Z_{\text{res}}(b)\right] &\leq \mathbb{E}\left[\exp\left(-\theta_b b\right)\exp\left(\tau_b \Lambda_b(\theta_b)\right); \tau_b < \tau, \max_{k\leq\tau_b} X_k > b\right] \\ &\leq b\bar{F}(b)\sup_{n\geq 1}\exp\left(n\Lambda_b(\theta_b)\right)\mathbb{P}\left\{\tau_b < \tau, \max_{k\leq\tau_b} X_k > b\right\} \\ &= O\left(\bar{F}^2(b)\right), \text{ as } b\to\infty,\end{aligned}$$

because of Lemma 4.8 and Proposition 4.2. Then due to (4.25), it is immediate that

$$\frac{\mathbb{E}_\theta\left[Z_{\text{res}}^2(b)\right]}{\mathbb{P}\left\{\tau_b < \tau\right\}^2} = O(1), \text{ as } b\to\infty. \tag{4.30}$$

Theorem 4.7. *If the realizations of the estimators $Z_{\text{dom}}(b)$ and $Z_{\text{res}}(b)$ are generated respectively from the measures $\mathbb{P}(\cdot)$ and $\mathbb{P}_\theta(\cdot)$, and if we let*

$$Z(b) := Z_{\text{dom}}(b) + Z_{\text{res}}(b),$$

then under Assumption 4.3, the family of estimators $(Z(b) : b > 0)$ are strongly efficient for the estimation of $\mathbb{P}\{\tau_b < \tau\}$, as $b\to\infty$; that is,

$$\frac{\text{Var}\left[Z(b)\right]}{\mathbb{P}\{\tau_b < \tau\}^2} = O(1), \text{ as } b\to\infty.$$

Proof. Since $Z_{\text{dom}}(b)$ and $Z_{\text{res}}(b)$ are generated independently,

$$\text{Var}\left[Z(b)\right] = \text{Var}\left[Z_{\text{dom}}(b)\right] + \text{Var}\left[Z_{\text{res}}(b)\right].$$

This observation, together with (4.27) and (4.30) proves the claim. □

Remark 4.7. It is instructive to consider Theorem 4.7 in the light of the negative result due to Bassamboo et al. [2007], where it is proved that no probability measure that draws samples of increments in an i.i.d. fashion can efficiently estimate $\mathbb{P}\{\tau_b < \tau\}$. However, by separating the event of interest $\{\tau_b < \tau\}$ into two intuitive components based on the big jump principle, our result establishes that it is easy to efficiently estimate $\mathbb{P}\{\tau_b < \tau\}$ by designing sampling algorithms for each event individually in a way that typifies the event.

4.4 Numerical Experiments

In this section, we present the results of numerical simulation experiments performed for estimation of level crossing probabilities $\mathbb{P}\{\tau_b < \infty\}$. To facilitate comparison with existing methods, we use the following example from Blanchet and Glynn [2008a]: Consider an M/G/1 queue with traffic intensity $\rho = 0.5$ and Pareto service times having tail $\mathbb{P}\{V > t\} = (1+t)^{-2.5}$. The aim is to estimate the probability that this queue develops a waiting time b in stationarity by equivalently estimating the level crossing probabilities $\mathbb{P}\{\tau_b < \infty\}$ of the associated negative drift random walk. For this example, we use the simulation procedures discussed in Section 4.1 and compare the results with that of the existing algorithms in literature in Table 4.1. While Algorithms AK (in Asmussen and Kroese [2006]) and DLW (in Dupuis et al. [2007]) restrict the arrivals to be Poisson, the schemes BGL, BG and BL referring to the algorithms, respectively, in Blanchet et al. [2007a], Blanchet and Glynn [2008a] and Blanchet and Liu [2012] do not impose any such restriction.

In our implementation, r has been chosen to be 2 to keep the expected termination time low, as suggested by Theorem 4.2. The results reported in Table 4.1 correspond to the simulation estimates of $\mathbb{P}\{\tau_b < \infty\}$ for values of $b = 10^2, 10^3$ and 10^4 using $N = 10,000$ simulation runs. From Table 4.1, it can be inferred that the error offered by the estimates of our simpler state-independent procedure is much smaller when compared with other existing algorithms. Table 4.2 gives a comparison of coefficient of variation of the estimators empirically observed for different values of r, and a fixed $b = 10^3$. It can be seen from Table 4.2 as well that choosing $r = 2$ helps in keeping the relative error low.

4.5 Proofs of auxiliary results

In this section, we provide proofs of Propositions 4.1, 4.2, and Lemmas 4.2, 4.3, 4.4 and 4.8.

Proof of Proposition 4.1 The upper bound follows simply by applying union bound as below:

$$\begin{aligned}\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, A_k\} &\leq \mathbb{P}\left\{\bigcup_{j=n_{k-1}+1}^{n_k} \{X_j > b + j\mu\}\right\} \\ &\leq \sum_{j=n_{k-1}+1}^{n_k} \bar{F}(b+j\mu) = q_k(b). \end{aligned} \tag{4.31}$$

Table 4.1: Numerical result for Example 2 - here Std. error denotes the standard deviation of the estimator of $\mathbb{P}\{\tau_b < \infty\}$ based on 10,000 simulation runs; CV denotes the empirically observed coefficient of variation

Estimation Std. error CV	$b = 10^2$	$b = 10^3$	$b = 10^4$
	9.75×10^{-4}	3.15×10^{-5}	9.98×10^{-7}
Proposed	4.11×10^{-6}	7.89×10^{-8}	1.39×10^{-9}
method	0.42	0.25	0.14
	1.20×10^{-3}	3.15×10^{-5}	9.98×10^{-7}
AK	1.48×10^{-5}	2.19×10^{-7}	6.95×10^{-9}
	1.23	0.70	0.70
	1.05×10^{-3}	3.16×10^{-5}	9.91×10^{-7}
DLW	5.20×10^{-6}	1.69×10^{-7}	2.99×10^{-9}
	0.50	0.53	0.30
	1.02×10^{-3}	3.17×10^{-5}	1.13×10^{-6}
BGL	3.84×10^{-5}	1.60×10^{-6}	7.28×10^{-8}
	3.76	5.05	6.44
	1.08×10^{-3}	3.15×10^{-5}	9.98×10^{-7}
BG	5.97×10^{-6}	9.73×10^{-8}	2.07×10^{-9}
	0.55	0.31	0.21
	1.05×10^{-3}	3.18×10^{-5}	9.88×10^{-7}
BL	3.76×10^{-5}	2.60×10^{-7}	8.19×10^{-9}
	3.58	0.82	0.83

Table 4.2: Comparison of relative errors for different choices of r in Example 2 with $b = 1000$; here Std. error denotes the standard deviation of the estimator of $\mathbb{P}\{\tau_b < \infty\}$ based on 10,000 simulation runs; CV denotes the empirically observed coefficient of variation

r	Estimate	Std. error	CV
2	3.15×10^{-5}	7.89×10^{-8}	0.25
10	3.16×10^{-5}	1.03×10^{-7}	0.33
100	3.16×10^{-5}	1.55×10^{-7}	0.49

For obtaining a lower bound, see that

$$\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, A_k\} = \sum_{j=n_{k-1}+1}^{n_k} \mathbb{P}\{\tau_b = j, A_k\}$$

is bounded from below by

$$\sum_{j=n_{k-1}+1}^{n_k} \mathbb{P}\left\{\tau_b = j, S_i > -(b+i\mu)^\gamma \text{ for all } i < j, X_j > b + j\mu + (b+j\mu)^\gamma\right\},$$

for some $\gamma < 1$ to be chosen later in this proof. Let $M_n := \max_{k\le n}(S_k - k\mu)$ and $M := \sup_k(S_k - k\mu)$. Then $\mathbb{P}\{n_{k-1} < \tau_b \le n_k, A_k\}$ is lower bounded by

$$\begin{aligned}
&\sum_{j=n_{k-1}+1}^{n_k} \mathbb{P}\left\{M_{j-1} \le b, S_i > -(b+i\mu)^\gamma \text{ for all } i<j, X_j > b+j\mu+(b+j\mu)^\gamma\right\} \\
&= \sum_{j=n_{k-1}+1}^{n_k} \mathbb{P}\left\{M_{j-1} \le b, S_i > -(b+i\mu)^\gamma \text{ for all } i<j\right\} \bar{F}\left(b+j\mu+(b+j\mu)^\gamma\right) \\
&\ge \mathbb{P}\left\{M \le b, \min_{i<n_k} S_i > -(b+n_{k-1}\mu)^\gamma\right\} \sum_{j=n_{k-1}+1}^{n_k} \bar{F}\left(b+j\mu+(b+j\mu)^\gamma\right) \qquad (4.32)
\end{aligned}$$

From (4.2) we have that $\mathbb{P}\{M > b\} \sim \mu^{-1}\bar{F}_I(b)$ as $b \to \infty$. Recall that $\beta = (\alpha \wedge 2)^{-1}$. If $\gamma > \beta$, then under the lighter left tail assumption formally stated in Assumption 4.1,

$$\mathbb{P}\left\{\min_{i<n_k} S_i < -(b+n_{k-1}\mu)^\gamma\right\} = O\left(n_k(b+n_{k-1}\mu)^{-\frac{\gamma}{\beta}+o(1)}\right), \text{ as } b\to\infty.$$

This follows from the well-known large deviation asymptotic that

$$\mathbb{P}\left\{\max_{i\le n} S_i > x\right\} \sim n\bar{F}(x)$$

uniformly for $x > n^{\beta+\epsilon}$; this can be found, for example, in Theorem 2.2 of Borovkov and Boxma [2001] and Theorem 5 of Borovkov and Borovkov [2002]. Therefore, by union bound,

$$\begin{aligned}
&\mathbb{P}\left\{M \le b, \min_{i<n_k} S_i > -(b+n_{k-1}\mu)^\gamma\right\} \\
&\qquad \ge 1 - \mathbb{P}\{M > b\} - \mathbb{P}\left\{\min_{i<n_k} S_i < -(b+n_{k-1}\mu)^\gamma\right\} \\
&\qquad \ge 1 - \bar{F}_I(b)(1-o(1)) - O\left(n_k(b+n_{k-1}\mu)^{-\frac{\gamma}{\beta}+o(1)}\right), \qquad (4.33)
\end{aligned}$$

as $b \to \infty$. Further because of (2.7),

$$\begin{aligned}
\sum_{j=n_{k-1}+1}^{n_k} \bar{F}\left(b+j\mu+(b+j\mu)^\gamma\right) &\ge \sum_{j=n_{k-1}+1}^{n_k} \left(1+\frac{(b+j\mu)^\gamma}{b+j\mu}\right)^{-\alpha+o(1)} \bar{F}\left(b+j\mu\right) \\
&\ge \left(1-\frac{c}{(b+n_{k-1}\mu)^{1-\gamma}}\right) \sum_{j=n_{k-1}+1}^{n_k} \bar{F}\left(b+j\mu\right)
\end{aligned}$$

for some positive constant c. If we choose $\gamma = (\alpha+1)/2\alpha$, then

$$\sum_{j=n_{k-1}+1}^{n_k} \bar{F}\left(b+j\mu+(b+j\mu)^\gamma\right) \ge \left(1-\frac{c}{(b+n_{k-1}\mu)^{\frac{\alpha-1}{2\alpha}}}\right) q_k(b).$$

Combining this with (4.32) and (4.33), we see that

$$\mathbb{P}\{n_{k-1} < \tau_b \le n_k, A_k\} \ge \left(1-\frac{c+o(1)}{b^{\frac{\alpha-1}{2\alpha}}}\right) q_k(b), \qquad (4.34)$$

as $b \to \infty$. Along with (4.31), we have that

$$\sup_k \left| \frac{\mathbb{P}\{n_{k-1} < \tau_b \le n_k, A_k\}}{q_k(b)} - 1 \right| = O\left(b^{-\frac{\alpha-1}{2\alpha}}\right), \text{ as } b \to \infty.$$

□

Proof of Proposition 4.2 Consider

$$\begin{aligned} P_1 &:= \mathbb{P}\left\{\tau_b < \tau, \max_{k \le \tau_b} X_k \le b, S_{\tau_b - 1} < \frac{b}{2}\right\} \\ &\le \mathbb{P}\left\{S_n \in \left(0, \frac{b}{2}\right), S_{n+1} > b, X_{n+1} < b \text{ for some } n < \tau\right\} \\ &\le \mathbb{E}\left[\sum_{n=0}^{\tau-1} \mathbb{I}\left(S_n \in \left(0, \frac{b}{2}\right), S_{n+1} > b, X_{n+1} < b\right)\right] \\ &\le \mathbb{E}\left[\sum_{n=0}^{\tau-1} \mathbb{I}\left(S_n \in \left(0, \frac{b}{2}\right)\right) \mathbb{P}\left\{b - S_n < X < b\right\}\right]. \end{aligned}$$

Let $\pi(B) = \mathbb{P}\{\sup_n S_n \in B\}$. Then according to the regenerative ratio representation,

$$\frac{1}{\mathbb{E}\tau}\mathbb{E}\left[\sum_{n=0}^{\tau-1} \mathbb{I}\left(S_n \in \left(0, \frac{b}{2}\right)\right) \mathbb{P}\left\{b - S_n < X < b\right\}\right] = \int_0^{\frac{b}{2}} \mathbb{P}\left\{b - u < X < b\right\} \pi(du).$$

Therefore,

$$\begin{aligned} P_1 &\le \mathbb{E}\tau \int_0^{\frac{b}{2}} \left(\bar{F}(b-u) - \bar{F}(b)\right) \pi(du) \\ &= \mathbb{E}\tau \bar{F}(b) \int_0^{\frac{b}{2}} \left(\frac{\bar{F}(b-u)}{\bar{F}(b)} - 1\right) \pi(du). \end{aligned}$$

From Potter's bounds (see (2.7)), we have that for all $u < \frac{b}{2}$,

$$\frac{\bar{F}(b-u)}{\bar{F}(b)} \le \left(1 - \frac{u}{b}\right)^{-\alpha-\delta} \le 1 + (\alpha+\delta)2^{\alpha+\delta+1}\frac{u}{b},$$

for any $\delta > 0$ and all b large enough. The last inequality follows from Taylor's theorem. Hence

$$P_1 \le (\alpha+\delta)2^{\alpha+\delta+1}\mathbb{E}\tau \frac{\bar{F}(b)}{b} \int_0^{\frac{b}{2}} u\pi(du).$$

Recall that $\pi((x,\infty)) \sim \int_x^\infty \bar{F}(u)du$. Since $\alpha > 2$, $\int_0^\infty u\pi(du) < \infty$. Therefore,

$$P_1 = O\left(\frac{\bar{F}(b)}{b}\right), \text{ as } b \to \infty. \tag{4.35}$$

Now consider the complementary event $\{\tau_b < \tau, \max_{k\le\tau_b} X_k \le b, S_{\tau_b-1} > b/2\}$:

$$\begin{aligned}
P_2 &:= \mathbb{P}\left\{\tau_b < \tau, \max_{k\le\tau_b} X_k \le b, S_{\tau_b-1} > \frac{b}{2}\right\} \\
&= \sum_{n=1}^{\infty} \mathbb{P}\left\{\tau > n, \max_{k\le n} X_k \le b, S_{n-1} > \frac{b}{2}, \tau_b = n\right\} \\
&= \sum_{n=1}^{\infty} \mathbb{P}\left\{\tau > n, \max_{k\le n} X_k \le b, S_{n-1} > \frac{b}{2}, M_{n-1} \le b, S_n > b\right\} \\
&= \sum_{n=1}^{\infty} \int_{\frac{b}{2}}^{b} \mathbb{P}\left\{\tau > n, \max_{k\le n} X_k \le b, S_{n-1} \in dy, M_{n-1} \le b, S_n > b\right\} \\
&\le \sum_{n=1}^{\infty} \int_{\frac{b}{2}}^{b} \mathbb{P}\left\{\tau > n, S_{n-1} \in dy, M_{n-1} \le b\right\} F((b-y, b]),
\end{aligned}$$

where the notation $F((x, y])$ stands for $\mathbb{P}\{x < X \le y\}$. Consider the taboo renewal function $H_x(\cdot)$ defined below:

$$H_x(B) := \sum_{n=0}^{\infty} \mathbb{P}\left\{\tau > n, M_n \le x, S_n \in B\right\}.$$

Then it is immediate that

$$P_2 \le \int_{\frac{b}{2}}^{b} H_b(dy) F((b-y, b]).$$

From Theorem 2 of Denisov and Shneer [2007], given $\epsilon > 0$, we have a y_0 large enough such that, for all x and y with $y \in (y_0, x - y_0)$,

$$(1-\epsilon)\frac{\mathbb{E}\tau}{\mu} F((y, x])dy \le H_x((y, y+dy)) \le (1+\epsilon)\frac{\mathbb{E}\tau}{\mu} F((y, x])dy.$$

Therefore, for a fixed $\epsilon > 0$, we have

$$H_{b+c}((y, y+dy)) \le (1+\epsilon)\frac{\mathbb{E}\tau}{\mu} F((y, b+c])dy$$

in the interval $(b/2, b)$, for some constant c and all b large enough. Since $H_b(\cdot) \le H_{b+c}(\cdot)$,

$$\begin{aligned}
P_2 &\le (1+\epsilon)\frac{\mathbb{E}\tau}{\mu} \int_{\frac{b}{2}}^{b} F((y, b+c])F((b-y, b])dy \\
&= (1+\epsilon)\frac{\mathbb{E}\tau}{\mu} \int_{0}^{\frac{b}{2}} F((b-y, b+c])F((y, b])dy \\
&\le (1+\epsilon)\frac{\mathbb{E}\tau}{\mu} \left(\bar{F}(b) \int_{0}^{\frac{b}{2}} \frac{F((b-y, b])}{\bar{F}(b)} F((y, b])dy + \int_{0}^{\frac{b}{2}} F((b, b+c])F((y, b])dy\right)
\end{aligned}$$

For a fixed $\delta > 0$, it follows from (2.7) that

$$\begin{aligned}\frac{F((b-y,b])}{\bar{F}(b)} &= \bar{F}(b)\left(\frac{\bar{F}(b-y)}{\bar{F}(b)} - 1\right)\\ &\leq \bar{F}(b)\left(\left(1-\frac{y}{b}\right)^{-\alpha-\delta} - 1\right)\\ &\leq (\alpha+\delta)2^{\alpha+\delta+1}\frac{y}{b}\bar{F}(b)\end{aligned}$$

for all $y < b/2$ and b large enough. The last inequality is a consequence of Taylor's theorem. Then,

$$P_2 \leq (1+\epsilon)\frac{\mathbb{E}\tau}{\mu}\left((\alpha+\delta)2^{\alpha+\delta+1}\frac{\bar{F}(b)}{b}\int_0^{\frac{b}{2}} y\bar{F}(y)dy + F((b,b+c])\int_0^{\frac{b}{2}} F((y,b])dy\right). \tag{4.36}$$

Since $\bar{F}(\cdot)$ is regularly varying, given $\gamma > 0$ it follows from (2.7) that for all b large enough,

$$\begin{aligned}F((b,b+c]) &= \bar{F}(b)\left(1-\frac{\bar{F}(b+c)}{\bar{F}(b)}\right)\\ &\leq \bar{F}(b)\left(1-\left(1+\frac{c}{b}\right)^{-\alpha-\gamma}\right)\\ &\leq \bar{F}(b)\left((\alpha+\gamma)\frac{c}{b}\right).\end{aligned}$$

Therefore

$$F((b,b+c]) = O\left(\frac{\bar{F}(b)}{b}\right), \text{ as } b\to\infty.$$

Further, when $\alpha > 2$,

$$\int_0^\infty \bar{F}(u)du < \infty \text{ and } \int_0^\infty u\bar{F}(u)du < \infty.$$

Using these in (4.36), we obtain that for tails with regularly varying index $\alpha > 2$,

$$P_2 = O\left(\frac{\bar{F}(b)}{b}\right), \tag{4.37}$$

as $b\to\infty$. Therefore, from (4.35) and (4.37), we obtain

$$\mathbb{P}\left\{\tau_b < \tau, \max_{k\leq\tau_b} X_k < b\right\} = P_1 + P_2 = O\left(\frac{\bar{F}(b)}{b}\right),$$

as $b\to\infty$. This proves the claim. □

Proof of Lemma 4.2 Consider $\theta : \mathbb{R}^+ \to \mathbb{R}^+$. From Lemma 3.2, we have that: for given $\epsilon > 0$, if $x\theta(x)\to\infty$, then there exists x_ϵ such that for all $x > x_\epsilon$,

$$\int_{-\infty}^x e^{\theta(x)u}F(du) \leq 1 + c\theta^{1+\delta}(x) + e^{2\alpha}\bar{F}\left(\frac{2\alpha}{\theta(x)}\right) + e^{\theta(x)x}\bar{F}(x)(1+\epsilon),$$

for some $\delta > 0$. By definition of $\theta_k(b)$ in (4.7), we have $(b + n_{k-1}\mu) \cdot \theta_k(b) \to \infty$, either if b or k grows to infinity. Writing $\theta_k(b)$ as θ_k, for values of b and k satisfying $b + n_{k-1}\mu > x_\epsilon$, we have,

$$\begin{aligned}\exp\left(\Lambda_k(\theta_k)\right) &\leq 1 + c\theta_k^{1+\delta} + e^{2\alpha}\bar{F}\left(\frac{2\alpha}{\theta_k}\right) + e^{\theta_k\cdot(b+n_{k-1}\mu)}\bar{F}(b+n_{k-1}\mu)(1+\epsilon)\\ &\leq \exp\left(c\theta_k^{1+\delta} + e^{2\alpha}\bar{F}\left(\frac{2\alpha}{\theta_k}\right) + \frac{1}{n_k}(1+\epsilon)\right),\end{aligned}$$

because $1 + x \leq e^x$ and $e^{\theta_k\cdot(b+n_{k-1}\mu)}\bar{F}(b + n_{k-1}\mu) = 1/n_k$. Then,

$$\exp\left(n_k\Lambda_k(\theta_k)\right) \leq \exp\left(cn_k\theta_k^{1+\delta} + e^{2\alpha}n_k\bar{F}\left(\frac{2\alpha}{\theta_k}\right) + 1 + \epsilon\right). \tag{4.38}$$

Also see that,

$$n_k\theta_k^{1+\delta} = \frac{n_k}{(b+n_{k-1}\mu)^{1+\delta}}\left(\log\left(\frac{1}{n_k\bar{F}(b+n_{k-1}\mu)}\right)\right)^{1+\delta} < \epsilon, \tag{4.39}$$

if b and k are such that $(b + n_{k-1}\mu)$ is large enough. Similarly for given $\delta > 0$, there exists x_δ such that if $b + n_{k-1}\mu > x_\delta$, then

$$\begin{aligned}\frac{\bar{F}\left(\frac{2\alpha}{\theta_k}\right)}{\bar{F}(b+n_{k-1}\mu)} &= \frac{\bar{F}\left(\frac{2\alpha(b+n_{k-1}\mu)}{-\log\left(n_k\bar{F}(b+n_{k-1}\mu)\right)}\right)}{\bar{F}(b+n_{k-1}\mu)}\\ &\leq \left(\frac{1}{2\alpha}\log\left(\frac{1}{n_k\bar{F}(b+n_{k-1}\mu)}\right)\right)^{\alpha+\delta}.\end{aligned}$$

Then for values of b and k such that $(b + n_{k-1}\mu)$ is large enough,

$$\begin{aligned}n_k\bar{F}\left(\frac{2\alpha}{\theta_k}\right) &\leq n_k\bar{F}(b+n_{k-1}\mu)\left(\frac{1}{2\alpha}\log\left(\frac{1}{n_k\bar{F}(b+n_{k-1}\mu)}\right)\right)^{\alpha+\delta}\\ &= \frac{n_kL(b+n_{k-1}\mu)}{(b+n_{k-1}\mu)^\alpha}\left(\frac{1}{2\alpha}\log\left(\frac{1}{n_k\bar{F}(b+n_{k-1}\mu)}\right)\right)^{\alpha+\delta} < \epsilon,\end{aligned}$$

because $\alpha > 1$. Combining this with (4.38) and (4.39), for b and k such that $b + n_{k-1}\mu$ is sufficiently large,

$$\exp\left(n_k\Lambda_k(\theta_k)\right) \leq \exp(1+3\epsilon),$$

thus establishing the claim. □

Proof of Lemma 4.3 Since $n_k = rn_{k-1}$,

$$\begin{aligned}\sup_k \frac{n_k\bar{F}(b+n_{k-1}\mu)}{q_k(b)} &= \sup_k \frac{n_k\bar{F}(b+n_{k-1}\mu)}{\sum_{j=n_{k-1}+1}^{n_k}\bar{F}(b+j\mu)}\\ &\leq \frac{n_k\bar{F}(b+\frac{n_k}{r}\mu)}{(1-r^{-1})n_k\bar{F}(b+n_k\mu)} < \infty,\end{aligned}$$

because of (2.7). □

Proof of Lemma 4.4 Recall that $n_k = rn_{k-1}$ for some constant r. Therefore, for any $k \geq 1$,

$$1 \leq \frac{\sum_{j=1}^{n_k} \bar{F}(b+j\mu)}{\sum_{j=1}^{n_{k-1}} \bar{F}(b+j\mu)} = 1 + \frac{\sum_{j=n_{k-1}+1}^{n_k} \bar{F}(b+j\mu)}{\sum_{j=1}^{n_{k-1}} \bar{F}(b+j\mu)} \leq 1 + \frac{n_{k-1}\bar{F}(b+n_{k-1}\mu)}{n_{k-1}\bar{F}(b+n_{k-1}\mu)} = 2.$$

When $\mathrm{Var}[X] < \infty$, see from (4.3) and Proposition 4.1 that $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k\}$ equals

$$\begin{aligned}
&\mathbb{P}\{\tau_b \leq n_k\} - \mathbb{P}\{\tau_b \leq n_{k-1}\} - \mathbb{P}\{n_{k-1} < \tau_b \leq n_k, A_k\} \\
&= \sum_{j=1}^{n_k} \bar{F}(b+j\mu)\left(1 + O\left(\frac{1}{b}\right)\right) + o\left(\sqrt{n_k \wedge b}\ \bar{F}(b)\right) \\
&\qquad - \sum_{j=1}^{n_{k-1}} \bar{F}(b+j\mu)\left(1 + O\left(\frac{1}{b}\right)\right) - \left(1 - O\left(b^{-\frac{\alpha-1}{2\alpha}}\right)\right) \sum_{j=n_{k-1}+1}^{n_k} \bar{F}(b+j\mu), \\
&= \sum_{j=1}^{n_k} \bar{F}(b+j\mu)\left(O\left(b^{-\frac{\alpha-1}{2\alpha}}\right)\right) + o\left(\sqrt{n_k \wedge b}\ \bar{F}(b)\right) \qquad (4.40)
\end{aligned}$$

as $b \to \infty$. Similarly when $\mathrm{Var}[X] = \infty$, for k such that $n_k\bar{F}(b) = o(1)$, see from (4.4) and Proposition 4.1 that $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k\}$ equals

$$\begin{aligned}
&\sum_{j=1}^{n_k} \bar{F}(b+j\mu)\left(1 + O\left(\frac{n_k^{\frac{1}{\alpha}+\epsilon}}{b}\right)\right) - \sum_{j=1}^{n_{k-1}} \bar{F}(b+j\mu)\left(1 + O\left(\frac{n_{k-1}^{\frac{1}{\alpha}+\epsilon}}{b}\right)\right) \\
&\qquad - \left(1 - O\left(b^{-\frac{\alpha-1}{2\alpha}}\right)\right) \sum_{j=n_{k-1}+1}^{n_k} \bar{F}(b+j\mu) \\
&= \sum_{j=1}^{n_k} \bar{F}(b+j\mu)\left(O\left(\frac{n_k^{\frac{1}{\alpha}+\epsilon}}{b}\right) + O\left(b^{-\frac{\alpha-1}{2\alpha}}\right)\right) \qquad (4.41)
\end{aligned}$$

for every $\epsilon > 0$. Since $n_k = rn_{k-1}$ for some constant r, it follows from (2.7) that for small enough ϵ and suitably chosen $\eta > 1$,

$$\sup_{k:n_k<b^\eta} \frac{b^{-\frac{\alpha-1}{2\alpha}} \sum_{j=1}^{n_k} \bar{F}(b+j\mu)}{q_k(b)} \leq \sup_{k:n_k<b^\eta} \frac{b^{-\frac{\alpha-1}{2\alpha}} n_k\bar{F}(b)}{n_{k-1}\bar{F}(b+n_k\mu)} = o(1),$$

$$\sup_{k:n_k<b^\eta} \frac{n_k^{\frac{1}{\alpha}+\epsilon}}{b} \frac{\sum_{j=1}^{n_k} \bar{F}(b+j\mu)}{q_k(b)} \leq \sup_{k:n_k<b^\eta} \frac{n_k^{\frac{1}{\alpha}+\epsilon}}{b} \frac{n_k\bar{F}(b)}{n_{k-1}\bar{F}(b+n_k\mu)} = o(1), \text{ and}$$

$$\sup_{k:n_k<b^\eta} \frac{\sqrt{n_k \wedge b}\ \bar{F}(b)}{\sum_{j=1}^{n_k} \bar{F}(b+j\mu)} \leq \sup_{k:n_k<b^\eta} \frac{\sqrt{n_k \wedge b}\ \bar{F}(b)}{n_k\bar{F}(b+n_k\mu)} = o(1),$$

as $b \to \infty$. Therefore from (4.40) and (4.41), for some $\eta > 1$,

$$\sup_{k:n_k<b^\eta} \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k, \bar{A}_k\}}{q_k(b)} = o(1), \text{ as } b \to \infty. \qquad (4.42)$$

For k such that $n_k > b^\eta$, we obtain a loose bound that suffices for our purposes:

$$\begin{aligned}\mathbb{P}\{n_{k-1} < \tau_b \le n_k\} &= \mathbb{P}\left\{n_{k-1} < \tau_b \le n_k, S_{n_{k-1}} > \frac{b + n_{k-1}\mu}{2}\right\} \\ &\quad + \mathbb{P}\left\{n_{k-1} < \tau_b \le n_k, S_{n_{k-1}} \le \frac{b + n_{k-1}\mu}{2}\right\} \\ &\le \mathbb{P}\left\{S_{n_{k-1}} > \frac{b + n_{k-1}\mu}{2}\right\} + \mathbb{P}\left\{\tau_{\frac{b+n_{k-1}\mu}{2}} < \infty\right\} \\ &\le (1+\epsilon) n_{k-1} \bar{F}\left(\frac{b + n_{k-1}\mu}{2}\right) + \frac{(1+\epsilon)}{\mu} \bar{F}_I\left(\frac{b + n_{k-1}\mu}{2}\right), \end{aligned} \tag{4.43}$$

for all k, b large enough. While the final inequality is due to the asymptotics (4.2) and Proposition 3.1, the second term in the penultimate step follows by observing that whenever the event $\{n_{k-1} < \tau_b \le n_k, S_{n_{k-1}} \le (b + n_{k-1}\mu)/2\}$ happens, the random walk $(Z_n : n \ge 0)$ defined by

$$Z_n := S_{n+n_{k-1}} - S_{n_{k-1}} - n\mu$$

crosses the level $(b + n_{k-1}\mu)/2$ at some finite $n \le n_k - n_{k-1}$. Here recall that $\tau_x := \inf\{k \ge 1 : S_k > x + k\mu\}$.

Further, since $n_k = r n_{k-1}$ for some constant r, from (2.8) and (2.7), we have that

$$\sup_{k:n_k > b^\eta} \frac{n_{k-1}\bar{F}\left(\frac{b+n_{k-1}\mu}{2}\right)}{q_k(b)} < \infty \text{ and } \sup_{k:n_k > b^\eta} \frac{\bar{F}_I\left(\frac{b+n_{k-1}\mu}{2}\right)}{q_k(b)} < \infty.$$

Therefore from (4.43),

$$\sup_{k:n_k > b^\eta} \frac{\mathbb{P}\{n_{k-1} < \tau_b \le n_k, \bar{A}_k\}}{q_k(b)} < \infty,$$

which along with (4.42) establishes the claim. □

Proof of Lemma 4.8 Since $\theta_b b \to \infty$ and $\mathbb{E}X \ne 0$, similar to Lemma 3.2, we have:

$$\begin{aligned}\exp\left(\Lambda_b(\theta_b)\right) &\le 1 + \theta_b \mathbb{E}X + c\theta_b^2 + \exp(2\alpha)\bar{F}\left(\frac{2\alpha}{\theta_b}\right) + \exp\left(\theta_b b\right)\bar{F}(b)(1 + o(1)) \\ &= 1 - \theta_b\mu + c\theta_b^2 + \exp(2\alpha)\bar{F}\left(\frac{2\alpha}{\theta_b}\right) + \frac{1}{b}(1 + o(1)).\end{aligned}$$

It follows from the definition of θ_b and a simple application of (2.7) that

$$\frac{1}{b} = o\left(\theta_b\right) \text{ and } \bar{F}\left(\frac{2\alpha}{\theta_b}\right) = o\left(\theta_b\right), \text{ as } b \to \infty.$$

Therefore,

$$\exp\left(\Lambda_b(\theta_b)\right) \le 1 - \theta_b\mu(1 + o(1)),$$

as $b \to \infty$. Then

$$\overline{\lim_{b\to\infty}} \sup_{n\ge 1} \exp\left(n\Lambda_b(\theta_b)\right) \le \inf_y \sup_n \sup_{b>y} (1 - \theta_b\mu(1 + o(1)))^n \le 1,$$

which proves the claim. □

5 Estimation of Tail Probabilities for Infinite Series

In this chapter our goal is to estimate tail probabilities of infinite series involving regularly varying random variables just with uniformly bounded computational effort. To precisely introduce the problem, let X be a zero mean random variable satisfying the following condition:

Assumption 5.1. *The distribution function of X denoted by $F(\cdot)$ is such that the tail probabilities $\bar{F}(x) := 1 - F(x) = x^{-\alpha}L(x)$ for some slowly varying function $L(\cdot)$ and $\alpha > 2$.*

Let $(X_n : n \geq 1)$ be a sequence of i.i.d. copies of X. Our aim is to efficiently estimate the tail probabilities of

$$S := \sum_n a_n X_n,$$

where $(a_n : n \geq 1)$ satisfies the following condition:

Assumption 5.2. *The sequence $(a_n : n \geq 1)$ is such that a_n lies in the interval $(0, 1)$ for every n and $\sum_n n a_n < \infty$.*

The random variable S is proper because $\sum_n a_n^2 < \infty$ (follows from Kolmogorov's three-series theorem). The assumptions that X has zero mean and $a_n < 1$ have been made just for the ease of exposition. If X has non-zero mean or if $a_n > 1$ for any n, then the corresponding problem of estimating $\mathbb{P}\{S > b\}$ can be translated to a problem instance satisfying Assumptions 5.1 and 5.2 by letting $\tilde{a}_n := a_n / \sup_n a_n$ and by instead simulating the right hand side of the equation below:

$$\mathbb{P}\left\{\sum_n a_n X_n > b\right\} = \mathbb{P}\left\{\sum_n \tilde{a}_n (X_n - \mathbb{E}X) > \frac{b - (\sum_n a_n)\,\mathbb{E}X}{\sup_n a_n}\right\}.$$

Here note that $\sup_n a_n$ exists because we require $\sum_n a_n < \infty$.

The problem considered is of importance because of its applications in economics, finance, network science, etc. In particular, several linear processes, important time series and stochastic recurrence equations can be written in the form of infinite sums we consider (see Hult and Samorodnitsky [2008] for a discussion). Developing efficient simulation algorithms for the estimation of tail probabilities of S in a computer is non-trivial because the definition of S involves countably infinite number of random variables. Unlike previously considered simulation problems, any algorithm that stops after generating only finitely many increments is likely to introduce a bias. Apart from the task of eliminating bias, we are faced with an additional challenge of estimating the rare probability just with bounded number of realizations.

This chapter is organized as follows: We first present the simulation algorithm and prove that the resulting estimator is unbiased for the estimation of $\mathbb{P}\{S > b\}$ in Section 5.1. Following this, we analyse the variance of the estimator in Section 5.2 where we prove that the algorithm satisfies vanishing relative error property. After a note on the expected running time in Section 5.3, we conclude the chapter by presenting some proofs in Section 5.5.

5.1 Simulation Methodology

Given $b > 0$, we aim to compute $\mathbb{P}\{S > b\}$. Due to the absence of closed form expressions, in general, for the probability distribution of sums of random variables, we resort to simulation algorithms for computing $\mathbb{P}\{S > b\}$. If S is a sum of, say, for example, k i.i.d. random variables $X_1, \ldots, X_k$, then one can simply generate an i.i.d. realization of $X_1, \ldots, X_k$ and check whether their sum is larger than b or not. However, the countably infinite number of random variables involved in the definition of S makes the task of obtaining a sample of S via its increments, at least at a preliminary look, appear computationally infeasible. To overcome this difficulty, we introduce an auxiliary random variable N and re-express the probability $\mathbb{P}\{S > b\}$ below in (5.1) in a form that gives computational tractability: Let

$$S_0 := 0 \text{ and } S_n := \sum_{i=1}^{n} a_i X_i \text{ for } n \geq 1.$$

Further, let $p_n := \mathbb{P}\{N = n\}$ be positive for every $n \geq 1$. Then,

$$\begin{aligned} \mathbb{P}\{S > b\} &= \lim_n \mathbb{P}\{S_n > b\} \\ &= \sum_{n \geq 1} p_n \frac{\mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\}}{p_n} \\ &= \mathbb{E}\left[\frac{\mathbb{P}\{S_N > b \mid N\} - \mathbb{P}\{S_{N-1} > b \mid N\}}{p_N}\right] \end{aligned} \tag{5.1}$$

In Section 5.1, we aim to develop unbiased estimators $(Z_{\text{loc}}(n,b) : n \geq 1, b > 0)$ satisfying the following desirable properties:

(1) The expectation of $Z_{\text{loc}}(n,b)$ is $\mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\}$ for every n and b.

(2) The computational effort required to generate a realization of $Z_{\text{loc}}(n,b)$ is bounded from above by Cn, for some constant $C > 0$, uniformly for all b.

(3) The estimators $Z_{\text{loc}}(n,b)$ have low variance, uniformly in n and b.

Now, in a simulation run, if the realized value of N is n, we generate an independent realization of estimator $Z_{\text{loc}}(n,b)$ and use

$$Z(b) := \frac{Z_{\text{loc}}(N,b)}{p_N}$$

as an estimator for $\mathbb{P}\{S > b\}$. The fact that $Z(b)$ yields estimates of $\mathbb{P}\{S > b\}$ without any bias follows from (5.1). Thus by introducing an auxiliary random variable N, in every simulation run, we are faced with the task of generating only finitely many random variables, as opposed to the naive approach which requires generation of countably infinite random variables. The random variables $Z_{\text{loc}}(n,b)$ which are instrumental in estimating the tail probabilities of S will be referred hereafter as 'local' estimators.

Local estimators

As mentioned before, in this section, we present estimators for quantities

$$(\mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\} : n \geq 1)$$

that have low variance, uniformly in n, as $b \to \infty$. These form building blocks to serve our initial aim of estimating the tail probabilities of S. It is well-known that the sum of heavy-tailed random variables attain a large value typically because one of the increments (and hence the maximum of the increments) attain a large value. Therefore, we focus our attention on identifying the maximum of the increments

$$M_n := \max\{a_i X_i : 1 \leq i \leq n\}$$

in a manner that is reflective of the way in which the rare event under consideration happens. For this, we partition the sample space based on which of the n increments $\{a_1 X_1, \ldots, a_n X_n\}$ is the maximum. Let Max_n denote the index of the increment $a_i X_i$ which equals the maximum

M_n. In case of many increments having the same value as the maximum, we take the largest (index) of them to be Max_n. That is,

$$\text{Max}_n := \max\{\text{argmax}\{a_i X_i : 1 \le i \le n\}\}.$$

See that the quantity $\mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\}$ can be alternatively expressed as

$$\mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\} = p_1(n,b) + p_2(n,b) \tag{5.2}$$

where

$$\begin{aligned} p_1(n,b) &= \mathbb{P}\left\{S_n > b, \text{Max}_n = n\right\} - \mathbb{P}\left\{S_{n-1} > b, \text{Max}_n = n\right\} \text{ and} \\ p_2(n,b) &= \mathbb{P}\left\{S_n > b, \text{Max}_n \neq n\right\} - \mathbb{P}\left\{S_{n-1} > b, \text{Max}_n \neq n\right\}. \end{aligned}$$

We develop alternative representations for quantities $p_1(n,b)$ and $p_2(n,b)$ and use them to separately estimate $p_1(n,b)$ and $p_2(n,b)$ in the following sections.

Estimator for $p_1(n,b)$

Observe that $\mathbb{P}\left\{S_{n-1} > b, S_n \le b, \text{Max}_n = n\right\} = 0$ because whenever $S_n \le b$ and $S_{n-1} > b$, it is necessary that X_n be negative, and in which case S_n also needs to be negative (since $M_n = a_n X_n$). Therefore,

$$\begin{aligned} p_1(n,b) &= \mathbb{P}\left\{S_n > b, S_{n-1} \le b, \text{Max}_n = n\right\} - \mathbb{P}\left\{S_{n-1} > b, S_n \le b, \text{Max}_n = n\right\} \\ &= \mathbb{P}\left\{S_n > b, S_{n-1} \le b, \text{Max}_n = n\right\}. \end{aligned}$$

Further,

$$\begin{aligned} \mathbb{P}\left\{S_n > b, \text{Max}_n = n \;\middle|\; X_1, \ldots, X_{n-1}\right\} &= \mathbb{P}\left\{a_n X_n > b - S_{n-1}, a_n X_n > M_{n-1} \;\middle|\; X_1, \ldots, X_{n-1}\right\} \\ &= \bar{F}\left(\frac{1}{a_n}\left((b - S_{n-1}) \vee M_{n-1}\right)\right). \end{aligned}$$

Therefore, it is immediate that

$$\mathbb{E}\left[\bar{F}\left(\frac{1}{a_n}\left((b - S_{n-1}) \vee M_{n-1}\right)\right)\mathbb{I}(S_{n-1} \le b)\right] = \mathbb{P}\left\{S_n > b, S_{n-1} \le b, \text{Max}_n = n\right\}.$$

If we let

$$Z_1(n,b) := \bar{F}\left(\frac{1}{a_n}\left((b - S_{n-1}) \vee M_{n-1}\right)\right)\mathbb{I}(S_{n-1} \le b),$$

then it follows from the above discussion that $\mathbb{E}\left[Z_1(n,b)\right]$ equals $p_1(n,b)$. We note this observation below as Lemma 5.1.

Lemma 5.1. *For every $n > 1$ and $b > 0$, $\mathbb{E}\left[Z_1(n,b)\right] = p_1(n,b)$.*

In a simulation run, one can generate samples of $X_1, \ldots, X_{n-1}$ simply from the distribution $F(\cdot)$ and plug it in the expression of $Z_1(n,b)$ to arrive at an unbiased estimator for $p_1(n,b)$. Since $Z_1(n,b)$ is just the probability that the event of interest $\{S_n > b, S_{n-1} \leq b, \text{Max}_n = n\}$ happens conditional on the observed values of $X_1, \ldots, X_{n-1}$, $Z_1(n,b)$ is said to belong to a family of estimators called conditional Monte Carlo estimators (see, for example, Asmussen and Glynn [2007]). Estimators of the form $Z_1(n,b)$, also referred to as Asmussen-Kroese estimators, are shown to be extremely effective in the simulation of tail probabilities of sums of fixed number of heavy-tailed random variables in Asmussen and Kroese [2006] (see Chapters 2 and 3 for more details).

Estimator for $p_2(n,b)$

Similar to $p_1(n,b)$, one can develop conditional Monte Carlo estimators for the simulation of $p_2(n,b)$ as well. To accomplish this, we need some more notation: For any $j \leq n$, let

$$S_n^{(-j)} := \sum_{i=1, i\neq j}^{n} a_i X_i \text{ and } M_n^{(-j)} := \max_{i\leq n, i\neq j} a_i X_i.$$

Further, for any $n > 1$, let $(q(j,n) : 0 < j < n)$ be a probability mass function that assigns positive probability to every integer in $\{1, \ldots, n-1\}$. Let J_n be an auxiliary random variable which takes values in $\{1, \ldots, n-1\}$ such that $\mathbb{P}\{J_n = j\} = q(j,n)$. Aided with this notation, define the estimator for $p_2(n,b)$ as

$$Z_2(n,b) := \frac{Z_{2,1}(n,b) - Z_{2,2}(n,b)}{q(J_n,n)},$$

where

$$Z_{2,1}(n,b) := \bar{F}\left(\frac{1}{a_{J_n}}\left(\left(b - S_n^{(-J_n)}\right) \vee M_n^{(-J_n)}\right)\right) \text{ and}$$
$$Z_{2,2}(n,b) := \bar{F}\left(\frac{1}{a_{J_n}}\left(\left(b - S_{n-1}^{(-J_n)}\right) \vee M_n^{(-J_n)}\right)\right).$$

Lemma 5.2 below verifies that $Z_2(n,b)$ is an unbiased estimator for $p_2(n,b)$.

Lemma 5.2. *For every $n > 1$ and $b > 0$, $\mathbb{E}[Z_2(n,b)] = p_2(n,b)$.*

Proof. For any n and $j < n$, observe that

$$\mathbb{P}\left\{S_n > b, \text{Max}_n = j \;\middle|\; S_n^{(-j)}, M_n^{(-j)}\right\} = \mathbb{P}\left\{a_j X_j > b - S_n^{(-j)}, a_j X_j > M_n^{(-j)} \;\middle|\; S_n^{(-j)}, M_n^{(-j)}\right\}$$
$$= \bar{F}\left(\frac{1}{a_j}\left(\left(b - S_n^{(-j)}\right) \vee M_n^{(-j)}\right)\right), \tag{5.3}$$

and similarly,

$$\mathbb{P}\left\{S_{n-1} > b, \mathrm{Max}_n = j \;\middle|\; S_{n-1}^{(-j)}, M_n^{(-j)}\right\} = \bar{F}\left(\frac{1}{a_j}\left(\left(b - S_{n-1}^{(-j)}\right) \vee M_n^{(-j)}\right)\right). \tag{5.4}$$

Recall that J_n takes values only in $\{1, \ldots, n-1\}$. Therefore, it follows from the definition of $Z_{2,1}(n,b)$ and $Z_{2,2}(n,b)$ that

$$\begin{aligned} Z_{2,1}(n,b) &= \mathbb{P}\left\{S_n > b, \mathrm{Max}_n = J_n \;\middle|\; S_n^{(-J_n)}, M_n^{(-J_n)}, J_n\right\} \text{ and} \\ Z_{2,2}(n,b) &= \mathbb{P}\left\{S_{n-1} > b, \mathrm{Max}_n = J_n \;\middle|\; S_n^{(-J_n)}, M_n^{(-J_n)}, J_n\right\}. \end{aligned}$$

Then it is immediate that

$$\begin{aligned} \mathbb{E}\left[Z_{2,1}(n,b) \mid J_n\right] &= \mathbb{P}\left\{S_n > b, \mathrm{Max}_n = J_n \;\middle|\; J_n\right\} \text{ and} \\ \mathbb{E}\left[Z_{2,2}(n,b) \mid J_n\right] &= \mathbb{P}\left\{S_{n-1} > b, \mathrm{Max}_n = J_n \;\middle|\; J_n\right\}. \end{aligned}$$

Since $\mathbb{P}\{J_n = j\} = q(j,n)$, it follows that

$$\begin{aligned} \mathbb{E}\left[\frac{Z_{2,1}(n,b)}{q(J_n,n)}\right] &= \sum_{j=1}^{n-1} \mathbb{P}\{J_n = j\} \frac{\mathbb{E}\left[Z_{2,1}(n,b) | J_n = j\right]}{q(j,n)} \\ &= \sum_{j=1}^{n-1} \mathbb{P}\left\{S_n > b, \mathrm{Max}_n = j\right\} \\ &= \mathbb{P}\left\{S_n > b, \mathrm{Max}_n \neq n\right\}. \end{aligned} \tag{5.5}$$

Similarly one can derive that

$$\mathbb{E}\left[\frac{Z_{2,2}(n,b)}{q(J_n,n)}\right] = \mathbb{P}\left\{S_{n-1} > b, \mathrm{Max}_n \neq n\right\} \tag{5.6}$$

Since $Z_2(n,b) = (Z_{2,1}(n,b) - Z_{2,2}(n,b))/q(J_n,n)$, it is immediate from (5.5) and (5.6) that $\mathbb{E}[Z_2(n,b)] = p_2(n,b)$. □

To summarize the simulation procedure, we present Algorithm 3 here, which returns a realization of

$$Z_{\mathrm{loc}}(n,b) := Z_1(n,b) + Z_2(n,b)$$

for given values of n and b. It follows from Lemmas 5.1 and 5.2 that $Z_{\mathrm{loc}}(n,b)$ is indeed an unbiased estimator for the quantity $\mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\}$.

Algorithm 3 Given n and b, the aim is to efficiently simulate $\mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\}$

procedure LocalSimulation(n, b)
Let $Z_1(n,b) =$ Estimator1(n,b) and $Z_2(n,b) =$ Estimator2(n,b)
Return $Z_{\text{loc}}(n,b) = Z_1(n,b) + Z_2(n,b)$

procedure Estimator1(n, b)
Initialize $Z_1(n,b) = 0$
Simulate a realization of $(X_i : 1 \le i \le n-1)$ independently from the distribution $F(\cdot)$
Let $S_{n-1} = \sum_{i=1}^{n-1} a_i X_i$ and $M_{n-1} = \max\{a_i X_i : 1 \le i \le n-1\}$
if $S_{n-1} \le b$ **then**
Let

$$Z_1(n,b) = \bar{F}\left(\frac{1}{a_n}\left((b - S_{n-1}) \vee M_{n-1}\right)\right)$$

Return $Z_1(n,b)$

procedure Estimator2(n, b)
Generate a sample of J_n such that for $j = 1, \ldots, n-1, \mathbb{P}\{J_n = j\} = q(j,n) := a_j / \sum_{i=1}^{n-1} a_i$
For $1 \le i \le n, i \ne J_n$ simulate X_i independently from the distribution $F(\cdot)$
Let $S_n^{(-J_n)} = \sum_{i=1, i\ne J_n}^{n} a_i X_i$, $M_n^{(-J_n)} = \max\{a_i X_i : i \le n, i \ne J_n\}$,

$$Z_{2,1}(n,b) = \bar{F}\left(\frac{1}{a_{J_n}}\left(\left(b - S_n^{(-J_n)}\right) \vee M_n^{(-J_n)}\right)\right),$$

$$Z_{2,2}(n,b) = \bar{F}\left(\frac{1}{a_{J_n}}\left(\left(b - S_{n-1}^{(-J_n)}\right) \vee M_n^{(-J_n)}\right)\right) \text{ and}$$

$$Z_2(n,b) = \frac{Z_{2,1}(n,b) - Z_{2,2}(n,b)}{q(J_n,n)}.$$

Return $Z_2(n,b)$

Simulation of $\mathbb{P}\{S > b\}$

As outlined in the beginning of Section 5.1, we use an auxiliary random variable N to estimate the tail probabilities of the infinite series $S = \sum_n a_n X_n$. Recall that LocalSimulation(n,b) is a simulation procedure introduced in Algorithm 3 in Section 5.1, which for given values of n and b, returns realizations of random variable $Z_{\text{loc}}(n,b)$ that has $\mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\}$ as its expectation. Given $b > 0$, we present below Algorithm 4 that makes a call to LocalSimulation procedure of Algorithm 3 and returns

$$Z(b) := \frac{Z_{\text{loc}}(N,b)}{p_N}$$

which is the estimator we propose for computing the probability $\mathbb{P}\{S > b\}$.

Algorithm 4 Given $b > 0$, the aim is to efficiently simulate $\mathbb{P}\{S > b\}$

Generate a sample of N such that $\mathbb{P}\{N = n\} = p_n$, for $n \geq 1$
Let $Z_{\text{loc}}(N, b) = \text{LocalSimulation}(N, b)$
Let

$$Z(b) = \frac{Z_{\text{loc}}(N, b)}{p_N}$$

Return $Z(b)$

Theorem 5.1. *The estimators $(Z(b) : b > 0)$ are unbiased: that is, for every $b > 0$,*

$$\mathbb{E}[Z(b)] = \mathbb{P}\{S > b\}.$$

Proof. Since $\mathbb{E}[Z_{\text{loc}}(n, b)] = \mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\}$ for every n and b,

$$\begin{aligned}
\mathbb{E}[Z(b)] &= \mathbb{E}\left[\mathbb{E}\left[\frac{Z_{\text{loc}}(N, b)}{p_N} \,\middle|\, N\right]\right] \\
&= \mathbb{E}\left[\frac{\mathbb{P}\{S_N > b \mid N\} - \mathbb{P}\{S_{N-1} > b \mid N\}}{p_N}\right] \\
&= \sum_n \mathbb{P}\{N = n\} \frac{\mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\}}{p_n}.
\end{aligned}$$

Since $\mathbb{P}\{N = n\} = p_n$, it is immediate that,

$$\mathbb{E}[Z(b)] = \sum_n [\mathbb{P}\{S_n > b\} - \mathbb{P}\{S_{n-1} > b\}] = \lim_n \mathbb{P}\{S_n > b\}.$$

Since $S_n \to S$ almost surely, as $n \to \infty$, $\lim_n \mathbb{P}\{S_n > b\}$ equals $\mathbb{P}\{S > b\}$. Thus, we have that the estimators $Z(b)$ are unbiased. □

Theorem 5.1 above re-emphasizes the fact that $Z(b)$ returned by Algorithm 4 is unbiased in the estimation of $\mathbb{P}\{S > b\}$ for every choice of $(p_n : n \geq 1)$ satisfying $p_n > 0$ and $\sum_n p_n = 1$. However, for our simulation procedure, we take

$$p_n := c_b \left(a_n^\alpha + \frac{a_n}{b^r}\right), \tag{5.7}$$

for some $r \geq 1$. The reasoning behind this choice is as follows: For the second moment of the estimator

$$\mathbb{E}\left[Z^2(b)\right] = \frac{\mathbb{E}\left[Z_{\text{loc}}^2(N, b)\right]}{p_N^2}$$

to be low, it is desirable that p_n be chosen such that p_n^2 roughly resembles $\mathbb{E}[Z_{\text{loc}}(n,b)]^2$, which is the lowest possible value for $\mathbb{E}[Z_{\text{loc}}^2(n,b)]$. However, part of the problem is that we do not know the value of $\mathbb{E}[Z_{\text{loc}}(n,b)]$. To guess an estimate for the same, recall that

$$Z_{\text{loc}}(n,b) = Z_1(n,b) + Z_2(n,b).$$

Next, we set $S_{n-1} = O(1)$ and $M_{n-1} = O(1)$ in the expression for $Z_1(n,b)$ and take

$$\bar{F}\left(\frac{b - O(1)}{a_n}\right) \approx a_n^{\alpha} \bar{F}\,(b)$$

as a rough guess for $\mathbb{E}[Z_1(n,b)]$ for large values of b. Additionally, recall that $Z_2(n,b) = (Z_{2,1}(n,b) - Z_{2,2}(n,b))/a_{J_n}$. Under modest continuity assumptions on $\bar{F}(\cdot)$ and taking all the random variables involved in the expression $M_n^{(-j)}, S_n^{(-j)}, X_n$ to be $O(1)$, one can take

$$\begin{aligned}\bar{F}\left(\frac{b - S_n^{(-j)}}{a_j}\right) - \bar{F}\left(\frac{b - S_{n-1}^{(-j)}}{a_j}\right) &\approx \frac{S_{n-1}^{(-j)} - S_n^{(-j)}}{a_j}\frac{\bar{F}(b)}{b} \\ &= \frac{a_n X_n}{a_j}\frac{\bar{F}(b)}{b} = O\left(\frac{a_n}{b}\bar{F}(b)\right)\end{aligned}$$

as a guess for $\mathbb{E}\left[Z_2(n,b)\right]$. Combining these two approximate estimates, we propose to take

$$p_n \propto a_n^{\alpha} + \frac{a_n}{b}$$

as a choice for p_n. It is important to remember that the estimators $Z(b)$ are unbiased for any choice of probabilities $(p_n : n \geq 1)$. As we proceed, it shall become clear that having an additional parametrization in terms of r, as in (5.7), is useful. The choice (5.7), as we shall see in Section 5.2, enables us to prove useful results on the variance of $Z(b)$.

5.2 Analysis of Variance of $Z(b)$

The aim of this section is to prove the following theorem when Assumptions 5.1 and 5.2 are in force:

Theorem 5.2. *For the choice of probabilities $(p_n : n \geq 1)$ as in (5.7), if r is taken larger than 1, the family of estimators $(Z(b) : b > 0)$ returned by Algorithm 4 has vanishing relative error, asymptotically, as $b \to \infty$. In other words,*

$$\lim_{b\to\infty} \frac{\mathbb{E}\left[Z^2(b)\right]}{\mathbb{P}\{S > b\}^2} = 1.$$

To prove that the estimators $Z(b)$ have low variance asymptotically as in the statement of Theorem 5.2, we need to establish that $\mathbb{E}[Z^2_{\text{loc}}(n,b)]$ is comparable to that of $p_n^2\bar{F}^2(b)$, which is challenging because proving such a proposition will have to establish that $\mathbb{E}[Z^2_{\text{loc}}(n,b)]$ is low with respect to two rarity parameters n and b. We accomplish this in the following section.

Uniform bounds on variance of local estimators

To obtain bounds on variance of estimators $Z_{\text{loc}}(n,b)$, we separately analyse the second moments of $Z_1(n,b)$ and $Z_2(n,b)$ (defined in Algorithm 3) below. Proposition 5.1 which is stated below and proved in Section 5.5 will be useful in the analysis.

Proposition 5.1. *Under Assumptions 5.1 and 5.2,*

$$\mathbb{P}\left\{S_n^{(-j)} > b,\ M_n^{(-j)} \le \frac{b}{k}\right\} \le \exp\left(k + o(1)\right)\left(\frac{\sum_i a_i^\alpha}{k}\bar{F}\left(\frac{b}{k}\right)\right)^k,\ \text{as } b \to \infty$$

uniformly in n, for every $j \le n$ and $k > 1$.

Remark 5.1. For large values of b, Proposition 5.1 roughly captures the idea that when the maximum of the increments are constrained, for example, to be smaller than $b/2$, the likely way for a heavy-tailed sum to become larger than b is by having two large increments roughly of size $b/2$. Though k being an integer helps in understanding the upper bound in Proposition 5.1 in terms of the number of jumps, one can check from the proof of Proposition 5.1 that the upper bound holds true for k being any real number larger than 1.

Analysis of $Z_1(n,b)$

Recall that

$$Z_1(n,b) := \bar{F}\left(\frac{1}{a_n}\left((b - S_{n-1}) \vee M_{n-1}\right)\right).$$

To upper bound second moment of $Z_1(n,b)$, we consider the following two quantities:

$$I_1(n,b) := \mathbb{E}\left[Z_1^2(n,b); (b - S_{n-1}) \vee M_{n-1} \ge \gamma b\right] \text{ and}$$
$$I_2(n,b) := \mathbb{E}\left[Z_1^2(n,b); (b - S_{n-1}) \vee M_{n-1} < \gamma b\right]$$

for $\gamma \in (0,1)$.

Lemma 5.3. *Under Assumptions 5.1 and 5.2,*

$$\varlimsup_{b\to\infty} \sup_{n>1} \frac{I_1(n,b)}{(a_n^{\alpha-\delta}\bar{F}(b))^2} \le (1+\delta)^2$$

for every $\delta > 0$ and $\gamma \in (0,1)$.

Proof. From the definition of $Z_1(n,b)$, it is immediate that

$$I_1(n,b) \leq \bar{F}^2(b)\mathbb{E}\left[\frac{\bar{F}^2\left(\frac{b}{a_n}\left(\left(1-\frac{S_{n-1}}{b}\right)\vee\gamma\right)\right)}{\bar{F}^2(b)}\right].$$

Since $\bar{F}(x) = x^{-\alpha+o(1)}$, given $\delta > 0$, for b large enough, because of (2.7), we have that for every n,

$$\frac{\bar{F}\left(\frac{b}{a_n}\left(\left(1-\frac{S_{n-1}}{b}\right)\vee\gamma\right)\right)}{\bar{F}(b)} \leq (1+\delta)a_n^{\alpha-\delta}h\left(\frac{S_{n-1}}{b}\right),$$

where $h(x) = ((1-x)\vee\gamma)^{-(\alpha+\delta)}$. Therefore,

$$\sup_{n\geq 1}\frac{I_1(n,b)}{\left(a_n^{\alpha-\delta}\bar{F}(b)\right)^2} \leq (1+\delta)^2 \sup_{n\geq 1}\mathbb{E}\left[h^2\left(\frac{S_{n-1}}{b}\right)\right].$$

Since $h(\cdot)$ is a non-decreasing function, it is immediate that

$$\sup_{n\geq 1}\frac{I_1(n,b)}{\left(a_n^{\alpha-\delta}\bar{F}(b)\right)^2} \leq (1+\delta)^2\mathbb{E}\left[h^2\left(\frac{\sum_n a_n X_n^+}{b}\right)\right],$$

where $x^+ := \max\{x,0\}$ for $x\in\mathbb{R}$. The following observations are in order:

1) $h(\cdot)$ is bounded

2) The random variable $\sum_n a_n X_n^+$ is proper (this is because $\sum_n a_n < \infty$ and hence a consequence of Kolmogorov's three-series theorem). Therefore, $b^{-1}\sum_n a_n X_n^+ \to 0$ almost surely, as $b\to\infty$.

Then because of bounded convergence,

$$\mathbb{E}\left[h^2\left(\frac{\sum_n a_n X_n^+}{b}\right)\right] \to 1, \text{ as } b\to\infty.$$

Thus, for every $\delta > 0$, we have that

$$\overline{\lim_{b\to\infty}}\sup_{n\geq 1}\frac{I_1(n,b)}{\left(a_n^{\alpha-\delta}\bar{F}(b)\right)^2} \leq (1+\delta)^2.$$

□

Lemma 5.4. *Under Assumptions 5.1 and 5.2, there exists γ in $(0,1)$ such that*

$$\overline{\lim_{b\to\infty}}\sup_{n>1}\frac{\mathbb{E}\left[I_2(n,b)\right]}{(p_n\bar{F}(b))^2} = 0.$$

Proof. Observe that $(b - S_{n-1}) \vee M_{n-1}$ is at least b/n, and this is achieved with equality when $a_i X_i = b/n$ for every $i < n$. Therefore,

$$\begin{aligned} I_2(n,b) &:= \mathbb{E}\left[Z_1^2(n,b); S_{n-1} > (1-\gamma)b, M_{n-1} \le \gamma b\right] \\ &\le \bar{F}^2\left(\frac{b}{na_n}\right) \mathbb{P}\left\{S_{n-1} > (1-\gamma)b, M_{n-1} \le \gamma b\right\} \end{aligned} \tag{5.8}$$

Since $\sum_n na_n < \infty$, $\sup_n n^2 a_n$ exists. Additionally, since $\bar{F}(x) = x^{-\alpha}L(x) = x^{-\alpha+o(1)}$, one can write

$$\bar{F}^2\left(\frac{b}{na_n}\right) \le (1+o(1))\left(\frac{na_n}{b}\right)^{2(\alpha+o(1))} \le (1+o(1))\left(\sup_n n^2 a_n\right)^{\alpha+o(1)}\left(\frac{a_n}{b^2}\right)^{\alpha+o(1)}$$

uniformly in n, as $b \to \infty$. Further, it follows from Proposition 5.1 that for every n,

$$\mathbb{P}\left\{S_{n-1} > (1-\gamma)b, M_{n-1} \le \gamma b\right\} \le C_\gamma \bar{F}^{\frac{1-\gamma}{\gamma}}(b),$$

for some suitable constant $C_\gamma > 0$ and all b large enough. Recall the definition of p_n in (5.7). Since $p_n \ge c_b a_n b^{-r}$, it follows from (5.8) that

$$\frac{I_2(n,b)}{\left(p_n \bar{F}(b)\right)^2} \le C_\gamma \left(\sup_n n^2 a_n\right)^{\alpha+o(1)}\left(\frac{a_n}{b^2}\right)^{\alpha+o(1)} \frac{b^{2r}}{c_b^2 a_n^2} \frac{\bar{F}^{\frac{1-\gamma}{\gamma}}(b)}{\bar{F}^2(b)},$$

uniformly in n, as $b \to \infty$. Since $\alpha > 2$ and $c_b \sim 1/\sum_n a_n^\alpha$ as $b \to \infty$, it follows that

$$\overline{\lim_{b\to\infty}} \sup_{n\ge 1} \frac{I_2(n,b)}{\left(p_n \bar{F}(b)\right)^2} = 0$$

for any choice of $\gamma < 1/3$. □

Recall that $p_n \ge c_b a_n^\alpha$. Since $\mathbb{E}[Z_1^2(n,b)]$ is the sum of $I_1(n,b)$ and $I_2(n,b)$,

$$\frac{\mathbb{E}\left[Z_1^2(n,b)\right]}{\left(p_n^{1-\delta}\bar{F}(b)\right)^2} \le \frac{I_1(n,b)}{\left((c_b a_n^\alpha)^{1-\delta}\,\bar{F}(b)\right)^2} + \frac{I_2(n,b)}{\left(p_n \bar{F}(b)\right)^2}$$

for every n and b. Further, we have that $c_b \sim 1/\sum_n a_n^\alpha$ as $b \to \infty$. Then the following is a simple consequence of Lemmas 5.3 and 5.4:

$$\overline{\lim_{b\to\infty}} \sup_{n\ge 1} \frac{\mathbb{E}\left[Z_1^2(n,b)\right]}{\left(p_n^{1-\delta}\bar{F}(b)\right)^2} \le \lim_{b\to\infty} \frac{1}{c_b^{2(1-\delta)}} \times (1+\delta)^2 + 0 = (1+\delta)^2 \left(\sum_n a_n^\alpha\right)^{2(1-\delta)}. \tag{5.9}$$

Analysis of $Z_2(n,b)$

Recall that

$$Z_2(n,b) = \frac{1}{q(J_n,n)}\left[\bar{F}\left(\frac{\xi_1}{a_{J_n}}\right) - \bar{F}\left(\frac{\xi_2}{a_{J_n}}\right)\right],$$

where

$$\xi_1 := \left(b - S_n^{(-J_n)}\right) \vee M_n^{(-J_n)} \text{ and } \xi_2 := \left(b - S_{n-1}^{(-J_n)}\right) \vee M_n^{(-J_n)}.$$

To upper bound the second moment of $Z_2(n,b)$, we need the following non-restrictive smoothness assumption on $\bar{F}(\cdot)$:

Assumption 5.3. *There exists a t_0 such that the slowly varying function $L(\cdot)$ in $\bar{F}(x) = L(x)x^{-\alpha}$ is continuously differentiable for all $t > t_0$. Further, $F(\cdot)$ is absolutely continuous, the corresponding probability density function $f(\cdot)$ is bounded, and there exists a constant $c > 0$ such that*

$$\bar{F}(x) - \bar{F}(y) \leq c(y-x)\frac{\bar{F}(x)}{x} \tag{5.10}$$

for all $y > x \geq t_0$

One sufficient condition for (5.10) to hold is that the slowly varying function $L(\cdot)$ in $\bar{F}(x) = L(x)x^{-\alpha}$ satisfies

$$L'(t) = o\left(\frac{L(t)}{t}\right) \quad \text{as } t \to \infty.$$

Similar to the analysis of second moment $Z_1(n,b)$, we upper bound $\mathbb{E}[Z_2^2(n,b)]$ via the following two terms: Let

$$J_1(n,b) := \mathbb{E}\left[Z_2^2(n,b); \xi_1 \wedge \xi_2 \geq \left(a_{J_n}^{\eta} \wedge \gamma\right) b\right] \text{ and}$$
$$J_2(n,b) := \mathbb{E}\left[Z_2^2(n,b); \xi_1 \wedge \xi_2 < \left(a_{J_n}^{\eta} \wedge \gamma\right) b\right]$$

for some fixed η and γ in $(0,1)$.

Lemma 5.5. *Under Assumptions 5.1, 5.2 and 5.3,*

$$\overline{\lim_{b\to\infty}} \sup_n \frac{J_1(n,b)}{\left(p_n \bar{F}(b)\right)^2} = 0$$

for every γ in $(0,1)$ and some η in $(0,1)$.

Proof. Observe that $|\xi_1 - \xi_2| \le a_n|X_n|$. Therefore, whenever both ξ_1/a_{J_n} and ξ_2/a_{J_n} are larger than t_0, due to (5.10),

$$Z_2^2(n,b) \le \frac{c^2}{q^2(J_n,n)} \frac{a_n^2 X_n^2}{a_{J_n}^2} \frac{a_{J_n}^2}{(\xi_1 \wedge \xi_2)^2} \bar{F}^2\left(\frac{\xi_1 \wedge \xi_2}{a_{J_n}}\right).$$

As a consequence, we have for every n,

$$J_1(n,b) \le \mathbb{E}\left[Z_2^2(n,b); \xi_1 \wedge \xi_2 \ge \gamma a_{J_n}^{\eta} b\right] \le c^2 a_n^2 \mathbb{E}\left[X_n^2\right] \mathbb{E}\left[\frac{1}{q^2(J_n,n) a_{J_n}^2} \frac{a_{J_n}^{2(1-\eta)}}{\gamma^2 b^2} \bar{F}^2\left(\frac{\gamma b}{a_{J_n}^{1-\eta}}\right)\right].$$

Then given $\delta > 0$, for large values of b, due to (2.7),

$$\bar{F}\left(\frac{\gamma b}{a_{J_n}^{1-\eta}}\right) \le (1+\delta)\left(\frac{a_{J_n}^{1-\eta}}{\gamma b}\right)^{\alpha-\delta}.$$

Further, since $q(j,n) = a_j / \sum_{i=1}^n a_i$,

$$J_1(n,b) \le (1+\delta)^2 \frac{c^2 a_n^2}{\gamma^{2(\alpha-\delta+1)}} \left(\sum_{i=1}^n a_i\right)^2 \mathbb{E}\left[X_n^2\right] \mathbb{E}\left[\frac{a_{J_n}^{\nu}}{b^{2(\alpha-\delta+1)}}\right],$$

where $\nu := 2(1-\eta)(\alpha-\delta+1) - 4$. If we choose $\eta < (\alpha-\delta-1)/(\alpha-\delta+1)$, then ν is positive. Additionally, since $p_n \ge c_b a_n b^{-r}$ (for some $r < 1$),

$$\sup_n \frac{J_1(n,b)}{\left(p_n \bar{F}(b)\right)^2} \le (1+\delta)^2 \frac{c^2}{c_b^2 \gamma^{2(\alpha-\delta+1)}} \left(\sum_{i=1}^{\infty} a_i\right)^2 \mathbb{E}\left[X^2\right] \frac{b^{-2(\alpha-\delta-r+1)}}{\bar{F}^2(b)}.$$

As $\bar{F}(x) \ge (1-\delta)x^{-\alpha-\delta}$ for large values of x, it follows that

$$\overline{\lim_{b\to\infty}} \sup_n \frac{J_1(n,b)}{p_n^2 \bar{F}^2(b)} = 0$$

for any δ smaller than $(1-r)/2$, and this proves the claim. □

For the analysis of $J_2(n,b)$, we define

$$\kappa := \sup\left\{k : \overline{\lim_n} n^k a_n < \infty\right\}$$

and separately analyse the cases $\kappa < \infty$ and $\kappa = \infty$. If a_n is, for example, polynomially decaying with respect to n, then κ happens to be finite. Whereas if a_n is exponentially decaying with respect to n, then κ is infinite. The analysis for the two cases differ, and are presented below in Lemmas 5.6 and 5.7.

Lemma 5.6. *If $\kappa = \infty$, then under Assumptions 5.1, 5.2 and 5.3,*

$$\overline{\lim_{b\to\infty}} \sup_n \frac{J_2(n,b)}{\left(n^{\frac{2}{\eta}} p_n \bar{F}(b)\right)^2} = 0$$

for some γ in $(0,1)$ and every η in $(0,1)$.

Proof. Due to mean value theorem,

$$Z_2(n,b) = \frac{1}{q(J_n,n)} \frac{\xi_1 - \xi_2}{a_{J_n}} f\left(\frac{\zeta}{a_{J_n}}\right)$$

for some ζ between ξ_1 and ξ_2. Here recall that $f(\cdot)$ is the probability density corresponding to the distribution $F(\cdot)$. Since $|\xi_1 - \xi_2| \leq a_n|X_n|$, it follows from the definition of $J_2(n,b)$ that

$$J_2(n,b) \leq \mathbb{E}\left[\frac{1}{q^2(J_n,n)} \frac{a_n^2 X_n^2}{a_{J_n}^2} f^2\left(\frac{\zeta}{a_{J_n}}\right); \xi_1 \wedge \xi_2 < \left(a_{J_n}^{\eta} \wedge \gamma\right)\right].$$

Recall that $q(j,n) = a_j / \sum_{i=1}^n a_i$. Then, due to Hölder's inequality,

$$\frac{J_2(n,b)}{a_n^2} \leq \left(\sum_{i=1}^n a_i\right)^2 \mathbb{E}\left[X_n^{2p} f^{2p}\left(\frac{\zeta}{a_{J_n}}\right)\right]^{\frac{1}{p}} \mathbb{E}\left[\frac{1}{a_{J_n}^{4q}}; \xi_1 \wedge \xi_2 < \left(a_{J_n}^{\eta} \wedge \gamma\right) b\right]^{\frac{1}{q}} \tag{5.11}$$

for some $p, q > 1$ satisfying $p^{-1} + q^{-1} = 1$ and $\mathbb{E}[X^{2p}] < \infty$. See that, as in the proof of Lemma 5.4, $\xi_1 \wedge \xi_2$ is at least b/n. Therefore,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{a_{J_n}^{4q}}; \xi_1 \wedge \xi_2 < \left(a_{J_n}^{\eta} \wedge \gamma\right) b\right] &= \mathbb{E}\left[\left(\frac{b}{\xi_1 \wedge \xi_2}\right)^{\frac{4q}{\eta}}; \xi_1 \wedge \xi_2 < \left(a_{J_n}^{\eta} \wedge \gamma\right) b\right] \\ &\leq n^{\frac{4q}{\eta}} \mathbb{P}\left\{\xi_1 \wedge \xi_2 < \left(a_{J_n}^{\eta} \wedge \gamma\right) b\right\}.\end{aligned}$$

From the definition of ξ_1 and ξ_2, it is immediate that for every n,

$$\begin{aligned}\mathbb{P}\left\{\xi_1 \wedge \xi_2 < \left(a_{J_n}^{\eta} \wedge \gamma\right) b \,\middle|\, J_n\right\} &\leq \mathbb{P}\left\{S_n^{(-J_n)} \vee S_{n-1}^{(-J_n)} > (1-\gamma)b, M_n^{(-J_n)} \leq \gamma b \,\middle|\, J_n\right\} \\ &\leq c_\gamma \bar{F}^{\frac{1-\gamma}{\gamma}}(b)\end{aligned} \tag{5.12}$$

for some constant c_γ and all b large enough, because of union bound and Proposition 5.1. Further, recall that $p_n \geq c_b a_n b^{-r}, \mathbb{E}[X^{2p}]$ is finite, and $f(\cdot)$ is bounded. These observations, in conjunction with (5.11), result in

$$\sup_{n \geq 1} \frac{J_2(n,b)}{\left(n^{\frac{2}{\eta}} p_n \bar{F}(b)\right)^2} = O\left(\frac{b^{2r} \bar{F}^{\frac{1-\gamma}{\gamma q}}(b)}{\bar{F}^2(b)}\right), \text{ as } b \to \infty.$$

Given $r < 1$ and q, one can choose γ suitably so that $b^{2r} \bar{F}^{\frac{1-\gamma}{\gamma q}}(b)$ vanishes as $b \to \infty$. This proves the claim. □

Lemma 5.7. *If $\kappa < \infty$, then under Assumptions 5.1, 5.2 and 5.3,*

$$\overline{\lim_{b \to \infty}} \sup_n \frac{J_2(n,b)}{\left(p_n \bar{F}(b)\right)^2} = 0$$

for some γ in $(0,1)$ and every η in $(0,1)$.

Proof. Observe that the argument leading to (5.11) in the proof of Lemma 5.6 holds irrespective of whether κ is finite or not. To proceed further, see that

$$\mathbb{E}\left[\frac{1}{a_{J_n}^{4q}};\xi_1\wedge\xi_2<\left(a_{J_n}^{\eta}\wedge\gamma\right)b\right]=\mathbb{E}\left[\frac{1}{a_{J_n}^{4q}}\mathbb{P}\left\{\xi_1\wedge\xi_2<\left(a_{J_n}^{\eta}\wedge\gamma\right)b\;\middle|\;J_n\right\}\right]. \tag{5.13}$$

It follows from the definition of ξ_1 and ξ_2 that

$$\mathbb{P}\left\{\xi_1\wedge\xi_2<\left(a_{J_n}^{\eta}\wedge\gamma\right)b\;\middle|\;J_n\right\}\le\mathbb{P}\left\{M_n^{(-J_n)}<a_{J_n}^{\eta}\,b\;\middle|\;J_n\right\}=\prod_{i=1,i\neq J_n}^{n}F\left(\frac{a_{J_n}^{\eta}\,b}{a_i}\right).$$

For any fixed $k>\kappa$, there exists a positive constant $\tilde{c}_k$ such that $n^k a_n\ge\tilde{c}_k$ for all n. Then

$$\mathbb{P}\left\{\xi_1\wedge\xi_2<\left(a_{J_n}^{\eta}\wedge\gamma\right)b\;\middle|\;J_n\right\}\le\prod_{i=1,i\neq J_n}^{n}F\left(\frac{i^k a_{J_n}^{\eta}\,b}{\tilde{c}_k}\right)\le F(1)^{\left(\frac{\tilde{c}_k}{a_{J_n}^{\eta}\,b}\right)^{\frac{1}{k}}-2},$$

where we have simply excluded the last $n-\lceil(\tilde{c}_k/(a_{J_n}^{\eta}\,b))^{1/k}\rceil$ terms in the product to get an upper bound. This inequality, along with (5.12), results in the following loose bound which is enough for our purposes:

$$\mathbb{P}\left\{\xi_1\wedge\xi_2<\left(a_{J_n}^{\eta}\wedge\gamma\right)b\;\middle|\;J_n\right\}\le cF(1)^{\frac{1}{2}\left(\frac{\tilde{c}_k}{a_{J_n}^{\eta}\,b}\right)^{\frac{1}{k}}-1}\bar{F}^{\frac{1-\gamma}{2\gamma}}(b),$$

for some constant $c>0$. Using this in (5.13), we have that

$$\mathbb{E}\left[\frac{1}{a_{J_n}^{4q}};\xi_1\wedge\xi_2<\left(a_{J_n}^{\eta}\wedge\gamma\right)b\right]\le cb^{\frac{4q}{\eta}}\mathbb{E}\left[\frac{1}{(a_{J_n}^{\eta}\,b)^{\frac{4q}{\eta}}}F(1)^{\frac{1}{2}\left(\frac{\tilde{c}_k}{a_{J_n}^{\eta}\,b}\right)^{\frac{1}{k}}-1}\right]\bar{F}^{\frac{1-\gamma}{2\gamma}}(b)$$

Since $x^{\frac{4qk}{\eta}}F(1)^{x-1}$ is bounded for positive values of x, the expectation term in the right hand side of the above equation is finite. Further, $p_n\ge c_b a_n b^{-r}$. As a consequence, we have from (5.11) that

$$\sup_{n\ge1}\frac{J_2(n,b)}{\left(p_n\bar{F}(b)\right)^2}=O\left(b^{\frac{4}{\eta}+2r}\frac{\bar{F}^{\frac{1-\gamma}{2q\gamma}}(b)}{\bar{F}^2(b)}\right),$$

which, for suitably chosen γ, vanishes to 0 as $b\to\infty$. This concludes the proof. □

Since $\mathbb{E}[Z_2^2(n,b)]$ is the sum of $J_1(n,b)$ and $J_2(n,b)$, when $\kappa=\infty$, due to Lemmas 5.5 and 5.6, one can choose η and γ in $(0,1)$ such that

$$\overline{\lim_{b\to\infty}}\sup_{n\ge1}\frac{\mathbb{E}\left[Z_2^2(n,b)\right]}{\left(n^{\frac{2}{\eta}}p_n\bar{F}(b)\right)^2}=0. \tag{5.14}$$

Similarly, when $\kappa<\infty$, due to Lemmas 5.5 and 5.7,

$$\overline{\lim_{b\to\infty}}\sup_{n\ge1}\frac{\mathbb{E}\left[Z_2^2(n,b)\right]}{\left(p_n\bar{F}(b)\right)^2}=0. \tag{5.15}$$

Proof of Theorem 5.2

Recall that

$$Z(b) = \frac{Z_{\text{loc}}(N,b)}{p_N} = \frac{Z_1(N,b) + Z_2(N,b)}{p_N}.$$

Therefore,

$$\frac{\mathbb{E}\left[Z^2(b)\right]}{\bar{F}^2(b)} = \mathbb{E}\left[\frac{Z_1^2(N,b)}{p_N^2\bar{F}^2(b)}\right] + \mathbb{E}\left[\frac{Z_2^2(N,b)}{p_N^2\bar{F}^2(b)}\right] + \mathbb{E}\left[\frac{Z_1(N,b)}{p_N\bar{F}(b)}\right]\mathbb{E}\left[\frac{Z_2(N,b)}{p_N\bar{F}(b)}\right].$$

Then due to Jensen's inequality,

$$\frac{\mathbb{E}\left[Z^2(b)\right]}{\bar{F}^2(b)} \le \mathbb{E}\left[\frac{Z_1^2(N,b)}{p_N^2\bar{F}^2(b)}\right] + \mathbb{E}\left[\frac{Z_2^2(N,b)}{p_N^2\bar{F}^2(b)}\right] + \sqrt{\mathbb{E}\left[\frac{Z_1^2(N,b)}{p_N^2\bar{F}^2(b)}\right]}\sqrt{\mathbb{E}\left[\frac{Z_2^2(N,b)}{p_N^2\bar{F}^2(b)}\right]}. \tag{5.16}$$

Now consider, for example, the first term in the right hand side of the above inequality. Due to the uniform convergence result on $\mathbb{E}[Z_1^2(n,b)]$ in (5.9), there exists a constant c_1 such that

$$\frac{\mathbb{E}\left[Z_1^2(N,b) \mid N\right]}{\left(p_N\bar{F}(b)\right)^2} \le c_1(1+\delta)^2 p_N^{-2\delta}\left(\sum_n a_n^\alpha\right)^{2(1-\delta)}$$

for every δ and b. Since $\sum_n n a_n$ exists, $\mathbb{E}p_N^{-2\delta} < \infty$ for all δ small enough. As δ can be arbitrarily small, due to reverse Fatou's lemma, it follows from (5.9) that

$$\overline{\lim_{b\to\infty}}\,\mathbb{E}\left[\frac{Z_1^2(N,b)}{p_N^2\bar{F}^2(b)}\right] \le \mathbb{E}\left[\overline{\lim_{b\to\infty}}\,\frac{\mathbb{E}\left[Z_1^2(N,b) \mid N\right]}{\left(p_N\bar{F}(b)\right)^2}\right] \le \left(\sum_n a_n^\alpha\right)^2. \tag{5.17}$$

Similarly, one can conclude from (5.14) and (5.15) that for every b,

$$\frac{\mathbb{E}\left[Z_2^2(N,b) \mid N\right]}{\left(p_N\bar{F}(b)\right)^2} \le \begin{cases} c_2 N^{\frac{4}{\eta}} & \text{if } \kappa = \infty \\ c_2 & \text{if } \kappa < \infty. \end{cases}$$

for some constant c_2. Observe that $\mathbb{E}N^{\frac{4}{\eta}} < \infty$ for any fixed η because when $\kappa = \infty$, p_n is exponentially decaying with respect to n. Then as a consequence of (5.14) and (5.15), due to dominated convergence,

$$\lim_{b\to\infty}\mathbb{E}\left[\frac{Z_2^2(N,b)}{p_N^2\bar{F}^2(b)}\right] = \mathbb{E}\left[\lim_{b\to\infty}\frac{\mathbb{E}\left[Z_2^2(N,b) \mid N\right]}{\left(p_N\bar{F}(b)\right)^2}\right] = 0.$$

This conclusion, along with (5.16) and (5.17), results in

$$\lim_{b\to\infty}\frac{\mathbb{E}\left[Z^2(b)\right]}{\bar{F}^2(b)} \le \left(\sum_n a_n^\alpha\right)^2.$$

Further, $\mathbb{P}\{S > b\} \sim \sum_n a_n^\alpha \bar{F}(b)$ as $b \to \infty$. Therefore,

$$\overline{\lim_{b\to\infty}}\,\frac{\mathbb{E}\left[Z^2(b)\right]}{\mathbb{P}\{S>b\}^2} \le 1.$$

Additionally, since $Z(b)$ is an unbiased estimator of $\mathbb{P}\{S > b\}$, $\mathbb{E}[Z^2(b)]$ must be larger than $\mathbb{P}\{S > b\}^2$ because of Jensen's inequality. This proves the theorem. □

5.3 A note on computational complexity of the simulation procedure

Given $b > 0$, our objective has been to devise an algorithm that returns a number in the interval $((1-\epsilon)\mathbb{P}\{S > b\}, (1+\epsilon)\mathbb{P}\{S > b\})$ with probability at least $1-\delta$. In Section 5.1, we proposed to take average of values returned by several runs of Algorithm 4 as the estimate of $\mathbb{P}\{S > b\}$. Assuming that tasks like performing basic arithmetic operations, generating uniform random numbers, evaluating $F(x)$ at specified x, all require unit computational effort, it is immediate that each call to the procedure LocalSimulation(n, b) expends at most Cn computational effort, for some positive constant C, irrespective of the value of b. Given $b > 0$, if one makes N_b calls to Algorithm 4 and returns the average of returned values of $Z(b)$ as the overall estimate,

1) it can be shown by the same reasoning as in (2.1) that the estimate lies within the desired interval with probability at least $\epsilon^{-2}\mathrm{CV}^2[Z(b)]/N_b$, where $\mathrm{CV}[Z_b] = \mathrm{Var}[Z_b]/\mathbb{E}[Z_b]^2$ is the coefficient of variation of Z_b, and

2) the overall computational effort is at most CNN_b, where N is the auxiliary random variable drawn according to the probability mass function $(p_n : n \geq 1)$ in Algorithm 4.

Due to Theorem 5.2, we have that $\mathrm{CV}[Z(b)] = o(1)$, as $b \to \infty$. Therefore, it is enough to choose $N_b = c\epsilon^{-2}\delta^{-1}$ for some positive constant c. Further, note that

$$\mathbb{E}[N] = \sum_n np_n = c_b \sum_n n\left(a_n^\alpha + \frac{a_n}{b^r}\right).$$

First, observe that $\sum_n a_n < \infty$ because of Assumption 5.2. Additionally, since $c_b \sim \sum_n a_n^\alpha$ as $b \to \infty$, we have $\mathbb{E}N = O(1)$ as $b \to \infty$. Therefore, the overall computational effort is just $O(1)$ as $b \to \infty$. Thus, despite the difficulties that the definition of S involves infinitely many random variables and $\mathbb{P}\{S > b\}$ is arbitrarily small for large values of b, our work establishes that one can compute $\mathbb{P}\{S > b\}$ without any bias by expending only a computational effort that is uniformly bounded in b.

5.4 A numerical example

In this section, we present the results of a numerical simulation experiment that demonstrates the efficiency of our estimator. Take $(X_n : n \geq 1)$ to be iid copies of a Pareto random variable X satisfying $\mathbb{P}\{X > x\} = 1 \wedge x^{-4}$. Additionally, take $a_n = 0.9^n$ and let $S = \sum_n a_n X_n$. We use $N = 10,000$ simulation runs to estimate $\mathbb{P}\{S > b\}$ for various values of b listed in Table

5.1. The parameter r in the choice of probabilities p_n in the expression 5.7 is taken to be 1. The values listed in Column 3 correspond to the estimate obtained from 10,000 runs of our simulation algorithm. It is instructive to compare the simulation estimates in Column 3 with the crude asymptotic $\bar{F}(b)\sum_n a_n^\alpha$ listed in Column 2. The empirically observed coefficient of variation of our simulation estimators is listed in Column 5. Although it is required in the proof of Theorem 5.2 that $r > 1$, it can be inferred from Column 5 that the choice $r = 1$ yields estimators that have coefficient of variation that decreases to 0 as b is increased.

Table 5.1: Numerical result for the simulation of $\mathbb{P}\{S > b\}$- here CV denotes the empirically observed coefficient of variation based on 10,000 simulation runs

b	Asymptotic $\bar{F}(b)\sum_n a_n^\alpha$	Estimate for $\mathbb{P}\{S > b\}$	Standard Error	CV
200	1.19×10^{-9}	1.49×10^{-9}	1.61×10^{-11}	1.08
500	3.05×10^{-11}	3.32×10^{-11}	1.54×10^{-13}	0.47
1000	1.91×10^{-12}	1.97×10^{-12}	8.43×10^{-15}	0.42

5.5 Proofs of auxiliary results

In this section, we present proof of Proposition 5.1.

Proof of Proposition 5.1. Observe that for any n and j,

$$\left\{M_n^{(-j)} \leq \frac{b}{k}\right\} = \bigcap_{i=1, i\neq j}^{n} \left\{X_i \leq \frac{b}{ka_i}\right\}.$$

Then for any $\theta > 0$,

$$\mathbb{P}\left\{S_n^{(-j)} > b, M_n^{(-j)} \leq \frac{b}{k}\right\} \leq \exp(-\theta b) \prod_{i=1, i\neq j}^{n} \mathbb{E}\left[\exp(\theta a_i X_i); X_i \leq \frac{b}{ka_i}\right]$$

because of a simple application of Markov's inequality. If θ is chosen such that $\theta b \to \infty$ as $b \to \infty$, from Lemma 3.2, we have

$$\mathbb{E}\left[\exp(\theta a_i X_i); X_i \leq \frac{b}{ka_i}\right] \leq 1 + c\theta^2 a_i^2 + e^{2\alpha}\bar{F}\left(\frac{2\alpha}{\theta a_i}\right) + \exp\left(\theta\frac{b}{k}\right)\bar{F}\left(\frac{b}{ka_i}\right)(1 + o(1)),$$

uniformly in i, as $b \to \infty$. Since $1 + x \le \exp(x)$,

$$\begin{aligned}
&\mathbb{P}\left\{S_n^{(-j)} > b, M_n^{(-j)} \le \frac{b}{k}\right\} \\
&\quad \le \exp(-\theta b) \prod_{i=1, i\neq j}^{n} \exp\left(c\theta^2 a_i^2 + e^{2\alpha}\bar{F}\left(\frac{2\alpha}{\theta a_i}\right) + \exp\left(\theta\frac{b}{k}\right)\bar{F}\left(\frac{b}{ka_i}\right)(1+o(1))\right) \\
&\quad \le \exp\left(-\theta b + c\theta^2 \sum_i a_i^2 + e^{2\alpha}\sum_i \bar{F}\left(\frac{2\alpha}{\theta a_i}\right) + \bar{F}\left(\frac{b}{k}\right)\exp\left(\theta\frac{b}{k}\right)\sum_i a_i^{\alpha-\epsilon}(1+o(1))\right),
\end{aligned} \tag{5.18}$$

for any given $\epsilon > 0$, due to (2.7), uniformly in j and n, as $b \to \infty$. Observe that

$$\theta_b := -\frac{k}{b}\log\left(\frac{\sum_i a_i^\alpha}{k}\bar{F}\left(\frac{b}{k}\right)\right)$$

is the minimizer of $-\theta b + \sum_i a_i^\alpha \bar{F}(b/k)\exp(\theta b/k)$, and it approximately minimizes the right hand side of (5.18). Since $\theta_b \searrow 0$ and $\sum_i a_i^{\alpha-\epsilon} < \infty$ for small enough ϵ, it follows from (2.7) that

$$\begin{aligned}
\sum_i \bar{F}\left(\frac{2\alpha}{\theta_b a_i}\right) \le (1+\epsilon)\sum_i \left(\frac{a_i}{2\alpha}\right)^{\alpha-\epsilon}\bar{F}\left(\frac{1}{\theta_b}\right) = o(\theta_b), \\
\theta_b^2 = o\left(\theta_b\right), \text{ and } \bar{F}\left(\frac{b}{k}\right)\exp\left(\theta_b\frac{b}{k}\right)\sum_i a_i^\alpha = k,
\end{aligned}$$

as $b \to \infty$. Therefore, uniformly for every n and $j \le n$,

$$\begin{aligned}
\mathbb{P}\left\{S_n^{(-j)} > b, M_n^{(-j)} \le \frac{b}{k}\right\} &\le \exp\left(k\log\left(\frac{\sum_i a_i^\alpha}{k}\bar{F}\left(\frac{b}{k}\right)\right) + o(1) + k(1+o(1))\right) \\
&= \exp(k+o(1))\left(\frac{\sum_i a_i^\alpha}{k}\bar{F}\left(\frac{b}{k}\right)\right)^k,
\end{aligned}$$

as $b \to \infty$. This proves the claim. □

6 Concluding Remarks on Simulation of Rare Events in Heavy-tailed Sums

The aim of this chapter is to summarize insights behind various simulation techniques introduced in previous chapters. Using these insights, it will be easy to tackle new problems, as we shall see in Section 6.2, in settings more general than i.i.d. sums.

6.1 Summary of techniques

In Chapters 3, 4 and 5, our focus has been in devising efficient algorithms for the simulation of various types of exceedance probabilities in heavy-tailed sums.

1) *Partitioning based on the big jump principle:* It is well-known that heavy-tailed sums exceed a large value typically because one of the increments become large (larger than b), while other increments behave "as usual". Based on this principle (often referred as big jump principle), our first step has been to partition the event of interest into a dominant component (where at least one of the increments is larger than b) and further residual components. For example, in the simulation of $\mathbb{P}\{S_n > b\}$ in Chapter 3, we split the quantity of interest as below:

$$\mathbb{P}\left\{S_n > b\right\} = \mathbb{P}\left\{S_n > b, \max_{k\leq n} X_k > b\right\} + \mathbb{P}\left\{S_n > b, \max_{k\leq n} X_k \leq b\right\}.$$

This same partitioning yields efficient algorithms, as we shall see later in Section 6.2 in this chapter, even if the increments cease to be i.i.d. Similar partitioning has been successfully employed in the estimation of $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}$ in Chapter 4 as well.

2) *Randomization for infinite horizon problems:* For problems involving infinitely many random variables, a simulation procedure that terminates after generating only finitely many of them is likely to introduce bias. However, by introducing an auxiliary random variable, as in the estimation of level crossing probability $\mathbb{P}\{\tau_b < \infty\}$ in Chapter 4 (or) in the estimation of tail probability $\mathbb{P}\{S > b\}$ in Chapter 5, we show that it is possible to remove bias in such settings. In the estimation of $\mathbb{P}\{\tau_b < \infty\}$, we first re-express $\mathbb{P}\{\tau_b < \infty\}$ as the following infinite sum:

$$\mathbb{P}\{\tau_b < \infty\} = \sum_{k\geq 1} \mathbb{P}\{n_{k-1} < \tau_b \leq n_k\},$$

where $(n_k : k \geq 0)$ is an increasing sequence of integers. Following this representation, one can introduce an auxiliary random variable K, and re-express the above infinite sum as below:

$$\mathbb{P}\{\tau_b < \infty\} = \sum_{k\geq 1} \mathbb{P}\{K = k\} \frac{\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}}{\mathbb{P}\{K = k\}} = \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{P}\{n_{K-1} < \tau_b \leq n_K\}}{p_K} \;\middle|\; K\right]\right],$$

where $p_k = \mathbb{P}\{K = k\}, k \geq 1$. In a simulation run, if the realized value of K is k, then one can just return an unbiased estimator of $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}/p_k$ to obtain an estimate of $\mathbb{P}\{\tau_b < \infty\}$ without bias. The key to achieve low variance for such estimators is to choose the probabilities $(p_k : k \geq 1)$ with asymptotic structure similar to that of $(\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\} : k \geq 1)$. This aspect has been clearly illustrated by considering various cases in Chapters 4 and 5.

3) *Randomization in the induction of big jump.* In the estimation of the dominant component

$$\mathbb{P}\left\{S_n > b, \max_{k\leq n} X_k > b\right\} = \sum_{i=1}^{n} \mathbb{P}\left\{S_n > b, \max_{k\leq n} X_k > b, \max_{k\leq n} X_k = X_i\right\},$$

one can consider estimating each term in the summation in the right hand side separately. However, when the problem is symmetric with respect to the increments, that is, when

$$\mathbb{P}\left\{S_n > b, \max_{k\leq n} X_k > b, \max_{k\leq n} X_k = X_i\right\} = \mathbb{P}\left\{S_n > b, \max_{k\leq n} X_k > b, \max_{k\leq n} X_k = X_j\right\}$$

for every $i, j \leq n$ as in Chapter 3, one can simply estimate

$$n\mathbb{P}\left\{S_n > b, \max_{k\leq n} X_k = X_1, X_1 > b\right\}$$

by considering the importance sampling measure where X_1 is drawn from $\mathbb{P}\{X \in \cdot \,|X > b\}$ and the rest of the increments $X_2, \ldots, X_n$ are drawn from $\mathbb{P}\{X \in \cdot\ \}$. If the problem is not symmetric, (that is, if there exists $i \neq j$ such that the contribution of X_i towards the

occurrence of rare event is not the same as that of X_j), then one can use the randomization technique explained in the estimation of $\mathbb{P}\{n_{k-1} < \tau_b \leq n_k\}$: The conditional measure $\mathbb{P}\{\cdot \mid n_{k-1} < \tau_b \leq n_k\}$ is understood to force the increment X_k to be big with probability asymptotically proportional to $\mathbb{P}\{X > b + k\mu\}$. Using this as a guideline, we choose an importance sampling measure that assigns the same probability for increment X_k to be big. This is ensured in Chapter 4 by randomly picking the increment $X_k, k = n_{k-1}+1, \ldots, n_k$ with probability proportional to $\mathbb{P}\{X > b+k\mu\}$. A similar randomization technique has been employed in the development of "local estimators" in Chapter 5 as well. An additional example is presented at the end of this chapter.

In addition to the importance sampling estimators exploiting the above techniques, one can also use conditional Monte Carlo estimators whenever the quantity to estimate admits an alternate representation that is amenable for estimation via conditional Monte Carlo. While the large deviation probabilities of the form $\mathbb{P}\{S_n > b\}$ are suitable for such alternate representations, the level crossing probabilities are not, and in such contexts, one can resort to importance sampling algorithms as in Chapter 4.

6.2 An example with non-i.i.d. sums

In this section, we consider estimation of large deviations probabilities in heavy-tailed sums where the increments are neither identical, nor independent. In particular, we consider the case where the increments have regularly varying heavy tails and are modulated by an independent Markov chain. To precisely explain the simulation problem in hand, consider a discrete-time ergodic Markov chain $\xi = (\xi_n : n \geq 1)$ taking values in a finite state space $\chi = \{1, 2, \ldots, s\}$. It has a stationary distribution π; that is, for any $i, j \in \chi$, there exists limits:

$$\lim_{n\to\infty} \mathbb{P}\{\xi_n = j | \xi_1 = i\} = \pi_j > 0, \text{ satisfying } \sum_{j=1}^{s} \pi_j = 1.$$

We consider the random walk $(S_n : n \geq 1)$ in $\mathbb{R}$ with

$$S_0 := 0 \text{ and } S_n := X_1 + \ldots + X_n, \text{ for } n \geq 1$$

satisfying the following properties:

(a) The increments $(X_n : n \geq 1)$ are conditionally independent given the Markov chain ξ,

(b) There exists a family of distribution functions $\{F_k(\cdot) : k \in \chi\}$, whose tail probabilities $\bar{F}_k(x) = 1 - F_k(x)$ are regularly varying at infinity, and they satisfy,

$$\mathbb{P}\{X_n > x \mid \xi\} = \mathbb{P}\{X_n > x \mid \xi_n\} = \bar{F}_k(x).$$

We allow the regularly varying functions to have different indices and slowly varying functions: that is, for $k \in \chi$, the tails $\bar{F}_k(\cdot)$ are of the form $\bar{F}_k(x) = x^{-\alpha_k} L_k(x),\ x \in \mathbb{R}$ for slowly varying functions $L_k(\cdot)$ and indices $\alpha_k > 1$.

For $k \in \chi$, name the expectations $\int_{-\infty}^{\infty} xF_k(dx)$ as μ_k and define the steady-state mean

$$\mu_\pi := \sum_{j=1}^{s} \mu_j \pi_j.$$

Under these conditions we have, $S_n/n \to \mu_\pi$ a.s. We are interested in the rare event probabilities $\mathbb{P}\{S_n > n\mu_\pi + b\}$, for large values of n and b. To be precise: Define $\alpha := \min\{\alpha_1, \ldots, \alpha_s\}$ and $\beta := (\alpha \wedge 2)^{-1}$. While studying these probabilities, without loss of generality, we can take μ_π to be zero. From Borovkov and Borovkov [2008], we have the following asymptotics: for any $b > n^{\beta+\epsilon}$,

$$\mathbb{P}\{S_n \geq b\} \sim n\bar{F}_\pi(b), \text{ as } n \to \infty, \tag{6.1}$$

where $\bar{F}_\pi(\cdot) = \pi_1 \bar{F}_1(\cdot) + \ldots + \pi_s \bar{F}_s(\cdot)$ is a regularly function with index $\alpha > 1$. Given any $\epsilon > 0$, we are interested in developing important sampling algorithms for efficiently computing the moderate and large deviation probabilities $\mathbb{P}\{S_n > b\}$, for values of $b > n^{\beta+\epsilon}$, as $n \to \infty$. We make the following non-restrictive technical assumption:

Assumption 6.1. *If* $\int_{-\infty}^{\infty} x^2 \bar{F}_\pi(dx) = \infty$, *then* $\overline{\lim}_{x\to\infty} \frac{F_\pi(-x)}{\bar{F}_\pi(x)} < \infty$.

This assumption just encodes that in case of increments having infinite variance, the heaviest of the left tails is lighter than the heaviest of the right tails in the family $\{F_k(\cdot) : k \in \chi\}$. If this is not the case, we just take $-\alpha$ to be the the maximum of the indices of regular variation of all the tails (both left and right) in the family, set $\beta = \alpha^{-1}$, and look for estimating probabilities $\mathbb{P}\{S_n > b\}$ with $b > n^{\beta+\epsilon}$. For example, if the heaviest tail index in the family is $-3/2$, then we consider values of b larger than $n^{2/3+\epsilon}$.

Related literature

Stochastic models involving random walks with heavy-tailed increments have received substantial attention in queuing and insurance risk theories because of their ability to explain long-range dependence in tele-traffic data, and highly variable claim sizes in insurance (see Resnick [1997], Embrechts et al. [1997], and Adler et al. [1998]). Because of their relative simplicity, random walks with independent increments have been analysed well in the literature to obtain various tail asymptotics and efficient simulation algorithms. On the other hand, it has been observed that the less studied case of random walks with dependent increments, particularly the kind

of dependence where we have an underlying stochastic process that modulates the increments, are of importance in the study of risk in changing economic environments (see Delbaen and Haezendonck [1987], Paulsen [1993] and Paulsen and Gjessing [1997]) and in the analysis of certain queuing networks (see Baccelli et al. [1999] and Huang and Sigman [1999]). As explained in Chapter 3, algorithms for the estimation of the large deviation probabilities $\mathbb{P}\{S_n > b\}$ are particularly important because they act as building block to many complex rare event problems involving combinations of renewal processes: for examples in queueing, see Parekh and Walrand [1989] and in financial credit risk modeling, see Glasserman and Li [2005] and Bassamboo et al. [2008].

Simulation methodology

We perform importance sampling to estimate the large deviation probabilities. One interesting aspect of our importance sampling changes of measure is that it does not involve any change in transition probabilities of the modulating Markov chain, unlike the case of modulated walks with light-tailed increments where exponential twisting is performed to favour certain states over others. As in Chapter 3 and in the beginning of this chapter, we partition the event of interest into a dominant and residual component,

$$A_{\mathrm{dom}}(n,b) := \left\{ S_n > b, \max_{k\le n} \ge b \right\} \text{ and } A_{\mathrm{res}}(n,b) := \left\{ S_n > b, \max_{k\le n} < b \right\},$$

and estimate their probabilities separately.

Simulation of A_{dom}

When one of the increments $X_1, \ldots, X_n$ is large, according to the big jump principle, the sum is also large with high probability. For efficiently simulating the event A_{dom}, we choose an importance sampling measure that induces large positive jumps in the sum with appropriate probabilities that reflect the large deviations behaviour of S_n. Consider the sampling procedure described in Algorithm 5. Let $\mathbb{P}_1(\cdot)$ be the probability measure induced in the path space by sampling according to Algorithm 5, and let $\mathbb{E}_1[\cdot]$ denote the associated expectation operator; for brevity, the dependence on n and b has not been highlighted in the notation. Note that $\mathbb{P}(\cdot)$ is absolutely continuous with respect to the importance sampling measure $\mathbb{P}_1(\cdot)$ when restricted to A_{dom}.

Algorithm 5 might be surprising in the sense that we simulate the modulating Markov chain according to the original dynamics, and not from some other importance sampling transition probabilities that favour sampling the increments often from the heaviest of the tails. But we will show that sampling according to the original dynamics will suffice. We need to argue that

Algorithm 5 Given n and b, the aim is to efficiently simulate $\mathbb{P}\{A_{\text{dom}}(n,b)\}$ via importance sampling

Simulate a realization $\{\xi_1,\ldots,\xi_n\}$ according to the original dynamics of the Markov chain ξ
Set $q \longleftarrow \sum_{i=1}^n \bar{F}_{\xi_i}(b)$
Choose $I \in \{1,2,\ldots,n\}$ such that $\mathbb{P}\{I=i\} = \bar{F}_{\xi_i}(b)/q$.
For $j \in \{1,2,\ldots,n\} - \{I\}$, generate independent samples X_j from $F_{\xi_j}(\cdot)$
Generate sample X_I following distribution $\frac{F_{\xi_I}(dx)}{\bar{F}_{\xi_I}(b)}\mathbf{1}(x \geq b)$
Set $S \longleftarrow X_1 + \ldots + X_n$ and COUNT $\longleftarrow \#\{i \leq i \leq n : X_i \geq b\}$
Return $q \cdot \mathbf{1}(S > b)/\text{COUNT}$

the values returned by the Algorithm 5 are unbiased, and have low variance. The sampling procedure induces the following probabilities conditional on the realization of the modulating Markov chain ξ :

$$\begin{aligned}\mathbb{P}_1\{X_1 \in dx_1,\ldots,X_n \in dx_n \mid \xi_1,\ldots,\xi_n\} &= \sum_{i=1}^n \mathbb{P}\{I=i\} \cdot \frac{\prod_{j=1}^n \mathbb{P}\{X_j \in dx_j\}}{\bar{F}_{\xi_i}(b)}\mathbf{1}(x_i > b) \\ &= \sum_{i=1}^n \frac{\bar{F}_{\xi_i}(b)}{q(\xi_1,\ldots,\xi_n)} \cdot \frac{\mathbb{P}\{X_1 \in dx_1,\ldots,X_n \in dx_n\}}{\bar{F}_{\xi_i}(b)}\mathbf{1}(x_i > b),\end{aligned}$$

where $q(\xi_1,\ldots,\xi_n) := \sum_{i=1}^n \bar{F}_{\xi_i}(b)$. Therefore conditional on the Markov chain realization $(\xi_1,\ldots,\xi_n)$, the likelihood ratio of $\mathbb{P}(\cdot)$ with respect to $\mathbb{P}_1(\cdot)$ on the set A_{dom} is

$$\frac{q(\xi_1,\ldots,\xi_n)}{\sum_{i=1}^n \mathbb{I}(X_i > b)}.$$

Thus for given n and $b > n^{\beta+\epsilon}$, the random variable

$$Z_{\text{dom}}(n,b) := \frac{q(\xi_1,\ldots,\xi_n)}{\#\{1 \leq i \leq n : X_i > b\}}\mathbb{I}(A_{\text{dom}}) \tag{6.2}$$

is an unbiased estimator of $\mathbb{P}(A_{\text{dom}})$ under measure $\mathbb{P}_1(\cdot)$. Here Lemma 6.1 establishes that the estimator has low variance.

Lemma 6.1. *Given $\epsilon > 0$, uniformly for $b > n^{\beta+\epsilon}$,*

$$\text{Var}[Z_{\text{dom}}(n,b)] = o\left(\mathbb{P}\{S_n > b\}^2\right), \text{ as } n \to \infty.$$

Proof. For every $k \in \chi$, let $N_k := \#\{1 \leq i \leq n : \xi_i = k\}$. Then $q(\xi_1,\ldots,\xi_n) = \sum_{i=1}^n \bar{F}_{\xi_i}(b)$ can be alternatively be written as: $q(\xi_1,\ldots,\xi_n) = \sum_{k=1}^s N_k \bar{F}_k(b)$. Further define,

$$Y_n := \frac{\sum_{k=1}^s \frac{N_k}{n}\bar{F}_k(b)}{\bar{F}_\pi(b)}.$$

Since $\#\{1 \leq i \leq n : X_i > b\}$ is at least 1 on A_{dom}, we have $Z_{\text{dom}}(n,b) \leq q(\xi_1, \ldots, \xi_n)$, and hence,

$$\frac{Z_{\text{dom}}(n,b)}{n\bar{F}_\pi(b)} \leq Y_n. \tag{6.3}$$

Recall that $\bar{F}_\pi(b) = \sum_{k=1}^s \pi_k \bar{F}_k(b)$. Therefore,

$$Y_n = \frac{\sum_{k=1}^s \frac{N_k}{n} \bar{F}_k(b)}{\sum_{k=1}^s \pi_k \bar{F}_k(b)} \leq \frac{\sum_{k=1}^s \bar{F}_k(b)}{\sum_{k=1}^s \pi_{\min} \bar{F}_k(b)} \leq \frac{1}{\pi_{\min}},$$

where $\pi_{\min} := \min\{k \in \chi : \pi_k\}$, which is non-zero because of the ergodicity of the Markov chain ξ. For the same reason, the occupation measures of the Markov chain converge to the stationary distribution π as in: $N_k/n \to \pi_k$, as $n \to \infty$, for all $k \in \{1, 2, \ldots, s\}$. Thus uniformly for all b, $Y_n \to 1$, as $n \to \infty$; and since $|Y_n|$ is uniformly bounded, because of bounded convergence theorem, we have $\mathbb{E}_1[Y_n^2] \sim 1$, as $n \to \infty$. Then given $\epsilon > 0$, from (6.3), for large enough values of n, we have:

$$\mathbb{E}_1\left[Z_{\text{dom}}^2(n,b)\right] \leq (1+\epsilon)(n\bar{F}_\pi(b))^2.$$

Since $\mathbb{E}_1\left[Z_{\text{dom}}(n,b)\right] = \mathbb{P}\{S_n > b\} \sim n\bar{F}_\pi(b)$ because of (6.1), we have the desired result on variance of $Z_{\text{dom}}(n,b)$:

$$\text{Var}[Z_{\text{dom}}(n,b)] = o\left(\mathbb{P}\{S_n > b\}^2\right), \text{ as } n \to \infty.$$

□

Simulation of A_{res}

All the increments $X_1, \ldots, X_n$ are bounded (by b) in the set A_{res}. Therefore, as in Chapter 3, we can apply ideas similar to exponential twisting. Given $b > 0$, for every $k \in \chi$, define

$$\Lambda_k(\theta) := \log\left(\int_{-\infty}^b \exp(\theta x) F_k(dx)\right), \text{ for } \theta \geq 0.$$

Consider the following family of 'truncated' and 'exponentially titled' distributions:

$$\hat{F}_k(dx) = \exp\left(\theta_{n,b} x - \Lambda_k(\theta_{n,b})\right) F_k(dx)\mathbf{1}(x \leq b), \quad x \in \mathbb{R}, k \in \chi, \tag{6.4}$$

with $\theta_{n,b}$ given by,

$$\theta_{n,b} := \frac{-\log\left(n\bar{F}_\pi(b)\right)}{b}. \tag{6.5}$$

The sampling procedure will involve simulation of the Markov chain ξ according to its original dynamics as in the simulation of A_{dom}. Now given the realization $\{\xi_1, \ldots, \xi_n\}$ of the modulating

Algorithm 6 Given n and b, the aim is to efficiently simulate $\mathbb{P}\{A_{\text{res}}(n,b)\}$ via importance sampling

Simulate a realization $\{\xi_1,\ldots,\xi_n\}$ according to the original dynamics of the Markov chain ξ
Set $\theta_{n,b} \longleftarrow -\log(n\bar{F}_\pi(b))/b$ and $C \longleftarrow \Lambda_{\xi_1}(\theta_{n,b}) + \ldots + \Lambda_{\xi_n}(\theta_{n,b})$
For $j = 1,\ldots,n$, generate independent samples X_j from $\hat{F}_{\xi_j}(\cdot)$
Actualize $S \longleftarrow X_1 + \ldots + X_n$
Return $\exp\left(-\theta_{n,b}S + C\right)\mathbf{1}(S \geq b)$

process, generate the increments X_i independently according to the exponentially twisted law $\hat{F}_{\xi_i}(\cdot)$, for $i = 1,\ldots,n$. This will result in the following unbiased estimator for $\mathbb{P}(A_{\text{res}})$:

$$Z_{\text{res}}(n,b) := \exp\left(-\theta_{n,b}S_n + \sum_{i=1}^{n}\Lambda_{\xi_i}(\theta_{n,b})\right)\mathbb{I}(A_{\text{res}}). \tag{6.6}$$

Let $\mathbb{P}_1(\cdot)$ be the measure induced by drawing samples according to the Algorithm 6 and let $\mathbb{E}_1[\cdot]$ be the expectation operator associated with $\mathbb{P}_1(\cdot)$. Lemma 6.2 gives an upper bound on the normalizing constants $\exp(\Lambda_k(\theta_{n,b}))$.

Lemma 6.2. *Given $\epsilon > 0$, uniformly for $b > n^{\beta+\epsilon}$ and $k \in \chi$,*

$$\exp(\Lambda_k(\theta_{n,b})) \leq 1 + \theta_{n,b}\mu_k + \frac{\bar{F}_k(b)}{n\bar{F}_\pi(b)}(1+o(1)), \text{ as } n \to \infty.$$

To prove Lemma 6.2, we need the following result:

Lemma 6.3. *Given any $\epsilon > 0$, for $\theta_{n,b}$ as in (6.5), uniformly for $b > n^{\beta+\epsilon}$ and $k \in \chi$, (a) $n\theta_{n,b}^{\kappa} \to 0$ for some $\kappa \in (0, \alpha \wedge 2)$, and (b) $\bar{F}_k\left(\frac{2\alpha}{\theta_{n,b}}\right) = O\left(\frac{1}{n}\right)$, as $n \to \infty$.*

Proof. (a) We have $\bar{F}_\pi(x) = x^{-\alpha}L_\pi(x)$, for some slowly varying function $L_\pi(\cdot)$. Given any $\delta > 0$ for sufficiently large values of b, we have $b^{-\delta} \leq L_\pi(b) \leq b^{\delta}$ because of the slowly varying nature of $L_\pi(\cdot)$. Therefore we have $L_\pi(b) = b^{o(1)}$ as $b \to \infty$. Further noting that $b > n^{\beta+\epsilon}$ helps us to write:

$$n\theta_{n,b}^{\kappa} = \frac{n}{b^{\kappa}}\log^{\kappa}\left(\frac{1}{n\bar{F}(b)}\right) \leq n^{1-\kappa(\beta+\epsilon)}\log^{\kappa}\left(\frac{b^{\alpha}}{nL_\pi(b)}\right).$$

$$\text{Take } \kappa := \begin{cases} 2, \text{ if } \alpha > 2 \\ (1+\epsilon)/(\frac{1}{\alpha}+\epsilon), \text{ if } 1 < \alpha \leq 2. \end{cases} \tag{6.7}$$

Then $\kappa < \alpha$, and $\kappa(\beta+\epsilon) \geq 1 + \epsilon/2$. Then $n\theta_{n,b}^{\kappa} \to 0$ as $n \to \infty$, uniformly for $b > n^{\beta+\epsilon}$.
(b) Since $\theta_{n,b} = -\log\left(n\bar{F}_\pi(b)\right)/b$, and $\bar{F}_k(\cdot)$ is regularly varying, given any $\delta > 0$, for n large enough,

$$\frac{\bar{F}_k\left(\frac{2\alpha}{\theta_{n,b}}\right)}{\bar{F}_k(b)} = \frac{\bar{F}_k\left(\frac{2\alpha b}{-\log(n\bar{F}_\pi(b))}\right)}{\bar{F}_k(b)} \leq \left(\frac{-\log\left(n\bar{F}(b)\right)}{2\alpha}\right)^{\alpha+\delta}.$$

The above inequality is just an application of (2.7). Therefore,

$$n\bar{F}_k\left(\frac{2\alpha}{\theta_{n,b}}\right) \leq n\frac{L_k(b)}{b^{\alpha_k}}\left(\frac{-\log\left(n\bar{F}(b)\right)}{2\alpha}\right)^{\alpha+\delta} = o(1), \text{ uniformly for } b > n^{\beta+\epsilon} \text{ as } n \to \infty.$$

Here the convergence to 0 is justified because $\alpha_k \geq \alpha$ and $b > n^{\beta+\epsilon}$. □

Proof of Lemma 6.2. From the definition of $\Lambda_k(\cdot)$ and Lemma 3.2, we have:

$$\begin{aligned}\exp\left(\Lambda_k(\theta_{n,b})\right) &= \int_{-\infty}^{b} \exp(\theta_{n,b}x)F_k(dx) \\ &\leq 1 + \theta_{n,b}\mu_k + c\theta_{n,b}^{\kappa} + e^{2\alpha}\bar{F}_k\left(\frac{2\alpha}{\theta_{n,b}}\right) + \exp(\theta_{n,b}b)\bar{F}_k(b)(1+o(1)),\end{aligned}$$

for κ as in (6.7). Usage of Lemma 3.2 is justified because $\theta_n b = -\log\left(n\bar{F}(b)\right) \to \infty$. From Lemma 6.3, we have $n\theta_{n,b}^{\kappa} = o(1)$ and $\bar{F}_k\left(\frac{2\alpha}{\theta_{n,b}}\right) = o\left(\frac{1}{n}\right)$, uniformly for $b > n^{\beta+\epsilon}$. Therefore,

$$\exp\left(\Lambda_b(\theta_n)\right) \leq 1 + \theta_{n,b}\mu_k + \frac{\bar{F}_k(b)}{n\bar{F}(b)}\left(1+o(1)\right), \text{ as } n \to \infty,$$

thus proving the result. □

Finally, Lemma 6.4 below gives a bound on the variance of estimators $Z_{\text{res}}(n,b)$.

Lemma 6.4. *Given $\epsilon > 0$, uniformly for $b > n^{\beta+\epsilon}$,*

$$\text{Var}[Z_{\text{res}}(n,b)] = o\left(\mathbb{P}\{S_n > b\}^2\right), \text{ as } n \to \infty.$$

Proof. Since $S_n > b$ on A_{res} and $b\theta_{n,b} = -\log(n\bar{F}_\pi(b))$,

$$Z_{\text{res}}(n,b) \leq \exp\left(-\theta_{n,b}b + \sum_{i=1}^{n}\Lambda_{\xi_i}(\theta_{n,b})\right)\mathbb{I}(A_{\text{res}}) \leq n\bar{F}_\pi(b)\exp\left(\sum_{i=1}^{n}\Lambda_{\xi_i}(\theta_{n,b})\right)\mathbb{I}(A_{\text{res}}).$$

The second moment can be evaluated as below:

$$\mathbb{E}_1\left[Z_{\text{res}}^2(n,b)\right] = \mathbb{E}\left[Z_{\text{res}}(n,b)\right] \leq n\bar{F}_\pi(b)\mathbb{E}\left[\exp\left(\sum_{i=1}^{n}\Lambda_{\xi_i}(\theta_{n,b})\right); A_{\text{res}}\right]. \tag{6.8}$$

As before, define $N_k := \sum_{i=1}^n \mathbb{I}(\xi_i = k)$, for $k \in \chi$. Then $\sum_{i=1}^n \Lambda_{\xi_i}(\theta_{n,b}) = \sum_{k=1}^s N_k \Lambda_k(\theta_{n,b})$. Now due to Jensen's inequality,

$$\exp\left(\sum_{i=1}^n \Lambda_{\xi_i}(\theta_{n,b})\right) = \exp\left(n \sum_{k=1}^s \frac{N_k}{n} \Lambda_k(\theta_{n,b})\right) \le \left(\sum_{k=1}^s \frac{N_k}{n} \exp(\Lambda_k(\theta_{n,b}))\right)^n.$$

Given $\delta > 0$, from Lemma 6.2 we have,

$$\exp(\Lambda_k(\theta_{n,b})) \le 1 + \theta_{n,b}\mu_k + (1+\delta)\frac{\bar{F}_k(b)}{n\bar{F}_\pi(b)},$$

for large values of n, and $k \in \chi$. We have also used that $b\theta_{n,b} = -\log(n\bar{F}_\pi(b))$. Since $N_1 + \ldots + N_s = n$,

$$\begin{aligned}\sum_{k=1}^s \frac{N_k}{n} \exp(\Lambda_k(\theta_{n,b})) &\le 1 + \theta_{n,b} \sum_{k=1}^s \frac{N_k}{n}\mu_k + \frac{1+\delta}{n\bar{F}_\pi(b)} \sum_{k=1}^s \frac{N_k}{n}\bar{F}_k(b). \\ &\le 1 + \theta_{n,b} \sum_{k=1}^s \left(\frac{N_k}{n} - \pi_k\right)\mu_k + \frac{1+\delta}{n\bar{F}_\pi(b)} \sum_{k=1}^s \left(\frac{N_k}{n} - \pi_k\right)\bar{F}_k(b) + \frac{1+\delta}{n},\end{aligned}$$

because $\sum_{k=1}^s \pi_k\mu_k = \mu_\pi = 0$ and $\sum_{k=1}^s \pi_k \bar{F}_k(b) =: \bar{F}_\pi(b)$. Since $1 + x \le \exp(x)$,

$$\left(\sum_{k=1}^s \frac{N_k}{n} \exp(\Lambda_k(\theta_{n,b}))\right)^n \le \exp\left(n\theta_{n,b} \sum_{k=1}^s \left(\frac{N_k}{n} - \pi_k\right)\mu_k + \frac{1+\delta}{\bar{F}_\pi(b)} \sum_{k=1}^s \left(\frac{N_k}{n} - \pi_k\right)\bar{F}_k(b) + 1 + \delta\right).$$

For $\delta' > 0$, define sets

$$B_0 := \bigcup_{k\in\chi} \{N_k \le n\pi_k + n^{1/2+\delta'}\} \text{ and } B_j := \bigcup_{k\in\chi} \{2^{j-1}n^{1/2+\delta'} < N_k - n\pi_k \le +2^j n^{1/2+\delta'}\}, j \ge 1.$$

We also have, $\bar{F}_\pi(b) \ge \pi_{\min} \sum_k \bar{F}_k(b)$. Then from (6.8),

$$\mathbb{E}_1\left[Z^2_{\text{res}}(n,b)\right] \le n\bar{F}_\pi(b) \exp\left((1+\delta)\left(1 + \frac{1}{\pi_{\min}}\right)\right) \sum_{j\ge 0} \exp\left(2^j n^{1/2+\delta'}\theta_{n,b}\right) \mathbb{P}(B_j \cap A_{\text{res}}). \quad (6.9)$$

Since the ergodic chain ξ takes values in a finite state space, we have for any $j \ge 1$, $\mathbb{P}(B_j) \le c\exp(-2^{2j-1}n^{2\delta'})$ for some constant c (Hoeffding-type inequality). Since $b > n^{\beta+\epsilon} \ge n^{1/2+\epsilon}$, we have $n^{1/2+\delta'}\theta_{n,b} = o(1)$ if $\delta' < \epsilon$. Therefore,

$$\begin{aligned}&\sum_{j\ge 0} \exp\left(2^j n^{1/2+\delta'}\theta_{n,b}\right) \mathbb{P}(B_j \cap A_{\text{res}}) \\ &\qquad \le \exp\left(n^{1/2+\delta'}\theta_{n,b}\right)\mathbb{P}(A_{\text{res}}) + c\sum_{j\ge 1} \exp\left(-2^{2j-1}n^{2\delta'} + 2^j n^{1/2+\delta'}\theta_{n,b}\right) \\ &\qquad \le \mathbb{P}(A_{\text{res}})(1 + o(1)) + O(\exp(-n^{2\delta'})).\end{aligned}$$

Combining this with (6.9) and the asymptotics that $\mathbb{P}\{S_n > b\} \sim \mathbb{P}\{S_n > b, M_n \ge b\} \sim n\bar{F}_\pi(b)$ as $n \to \infty$, we have: $\text{Var}[Z_{\text{res}}(n,b)] = o\left(\mathbb{P}\{S_n > b\}^2\right)$. □

Now we have Theorem 6.1 that comments about the efficiency of the overall estimation procedure.

Theorem 6.1. *If the realizations of the estimators Z_{dom} and Z_{res} are generated respectively from the measures $\mathbb{P}_1(\cdot)$ and $\mathbb{P}_1(\cdot)$, and if we let,*

$$Z(n,b) := Z_{\text{dom}}(n,b) + Z_{\text{res}}(n,b),$$

then under Assumption 6.1, the family of estimators $(Z(n,b) : n \geq 1, b > n^{\beta+\epsilon})$ achieves asymptotically vanishing relative error for the estimation of $\mathbb{P}\{S_n > b\}$, as $n \to \infty$; that is,

$$\frac{\text{Var}\left[Z(n,b)\right]}{\left(\mathbb{P}\{S_n > b\}\right)^2} = o(1),$$

as $n \to \infty$, uniformly for $b > n^{\beta+\epsilon}$.

Proof. Since the realizations of Z_{dom} and Z_{res} are generated independent of each other, the variance of Z is just the sum of variances of Z_{dom} and Z_{res}; the proof is now evident from Lemmas 6.1, 6.4 and Equation (6.1). □

A consequence of the above theorem is that, due to (2.1), the number of i.i.d. replications of $Z(n,b)$ required to achieve ϵ-relative precision with probability at least $1-\delta$ is at most $o(\epsilon^{-2}\delta^{-1})$. In our algorithms, each replication demands $O(n)$ computational effort, thus requiring a overall computational cost of $O(n)$, as $n \to \infty$.

Numerical experiments

Consider the time-homogeneous Markov chain ξ taking values in $\{0,1\}$ with transition probabilities given by:

$$\mathbb{P}\{\xi_{n+1} = 1 - i | \xi_n = i\} = 2/3, \text{ for } i = 0, 1.$$

The tail probabilities of the corresponding increment when the Markov chain is in state i is given by,

$$\mathbb{P}\{X_n > x | \xi_n = i\} = \bar{F}_i(x) = (17/12 + x)^{-\alpha_i}, \text{ for } x \geq -5/12,$$

with $\alpha_0 = 3$ and $\alpha_1 = 4$. Consider the random walk $(S_n : n \geq 0)$ with $S_n = X_1 + \ldots + X_n$, whose increments are modulated by the Markov chain ξ. It can be verified that the stationary distribution π of the Markov chain ξ is $\pi_i = 0.5, i \in \{0,1\}$, and that the steady-state drift of the random walk S_n is 0. In Table 6.1, we present the results of our estimation procedure for $N = 10000$ simulation runs for the computation of $\mathbb{P}\{S_n > n\}$ for values of $n = 100, 500$ and

1000. From Table 6.1, it can be seen that the empirically observed coefficient of variation of the estimates is small, and it decreases as n increases.

Table 6.1: Numerical result for Example 1 - here Std. error denotes the standard deviation of the estimator of $\mathbb{P}\{S_n > n\}$ based on 10,000 simulation runs; CV denotes the empirically observed coefficient of variation

n	Asymptotic expression $n\bar{F}_\pi(b)$	Proposed estimator ($\hat{z}$) for $\mathbb{P}\{S_n > n\}$	Std. error	CV of $\hat{z}$
100	4.84×10^{-5}	4.64×10^{-5}	7.48×10^{-8}	0.16
500	1.99×10^{-6}	1.95×10^{-6}	2.79×10^{-9}	0.14
1000	4.98×10^{-7}	4.92×10^{-7}	6.04×10^{-10}	0.12

7 Tail probabilities for large delays in half-loaded two-server queues

The tail behaviour of the distribution of steady-state delay in multiserver queues processing jobs with heavy-tailed sizes has attracted substantial attention in stochastic operations research. Most of the literature has focused on the case in which the traffic intensity, ρ, (that is, the ratio between the mean service requirement and the mean interarrival time) is not an integer and there are qualitative reasons, as we shall discuss, that make the integer case significantly more delicate to analyze. Our contribution in this work is to provide the first asymptotic upper and lower bounds for the tail distribution, that match up to a constant factor, for the integer case. In that process, we identify the occurrence of a few surprising phenomena that are not common in the asymptotic analysis of multiserver queues. We concentrate on the two server queue because it provides a vehicle to study the qualitative phenomenon that is of interest to us.

As mentioned earlier, most of the literature concentrates on the case in which ρ is not an integer. A series of conjectures relating tail distribution of steady-state delay to the traffic intensity has been made in Whitt [2000]. These conjectures turned out to be basically correct for the case of regularly varying job sizes and were verified for the case of a two-server queue in Foss and Korshunov [2006], where more general asymptotic bounds for subexponential distributions are provided. In Foss and Korshunov [2012], the authors provide bounds (up to constants) that verify the conjecture in Whitt [2000] for general multiserver queues with regularly varying job sizes and non-integer traffic intensity. There is a related body of literature aimed at studying stability properties, such as the existence of the mean steady-state delay, in terms of the traffic intensity of the system and tail properties of the incoming traffic. The relations found in this literature, see Scheller-Wolf and Sigman [1997] and Scheller-Wolf and Vesilo [2011], again are

also derived only for the case of non-integer traffic intensity and are consistent with the relations found for the tail distributions mentioned earlier (which can be used to derive the existence of moments).

In order to discuss our contributions in more detail, let us introduce some notation. Let V denote the amount of time required to service a generic job arriving to the queue and let $\bar{B}(x) = \mathbb{P}\{V > x\}$. We assume that $\bar{B}(\cdot)$ is regularly varying with index $\alpha > 1$, that is,

$$\bar{B}(x) = x^{-\alpha} L(x),$$

for some function $L(\cdot)$ satisfying $\lim_{x\to\infty} L(tx)/L(x) = 1$ for each $t > 0$; such a function $L(\cdot)$ is said to be slowly varying. Jobs are assumed to arrive as a Poisson stream (or, more generally, a renewal stream) with rate equal to $\mathbb{E}V$ and service requirements that are identical copies of V. Under this setting, the traffic intensity ρ equals 1. Let us write W to denote the steady-state waiting time of the two-server queue that processes jobs according to FCFS (first-come-first-serve) discipline. Our first result in this chapter establishes that if $\alpha > 2$, then

$$\mathbb{P}\{W > b\} = \Theta\left(b^2 \bar{B}(b^2) + b^2 \bar{B}^2(b)\right), \tag{7.1}$$

as $b \to \infty$. Here recall that $f(b) = \Theta(g(b))$ if and only if $f(b) \le c_1 g(b)$ and $g(b) \le c_2 f(b)$ for some positive constants c_1 and c_2 that are independent of b. To get a sense of how subtle the difference between the terms appearing in (7.1) are, it is instructive to consider the example $L(x) = \log(1+x)$, where the second term appearing in the right hand side of (7.1) dominates the asymptotic behaviour. On the other hand, if $L(x) = 1/\log(1+x)$, the first term in the right hand side of (7.1) dominates the asymptotic behaviour. Finally, if $L(x) \sim c$ for some $c > 0$ (the asymptotically Pareto case) both terms contribute substantially.

Further, let us discuss the result in (7.1) in view of the exact asymptotics obtained in Foss and Korshunov [2006] under slightly more general assumptions on the distribution of V. For the case $\rho < 1$, the exact asymptotic result in Foss and Korshunov [2006] in particular implies that

$$\mathbb{P}\{W > b\} = \Theta\left(b^2 \bar{B}^2(b)\right), \tag{7.2}$$

whereas for the case $\rho \in (1,2)$, it follows that

$$\mathbb{P}\{W > b\} = \Theta\left(b \bar{B}(b)\right), \tag{7.3}$$

as $b \to \infty$*. Since there is a sharp difference between the cases $\rho < 1$ and $\rho \in (1,2)$ as in (7.2) and (7.3), it has been of great interest to identify what happens when ρ equals 1. We resolve

*From here on, we avoid the quantification $b \to \infty$ whenever it is evident from the context

this in our work by noting that (7.1) is much closer to the case $\rho < 1$ than it is to the case $\rho > 1$. Although, quantitatively, the rates of convergence between the two terms in (7.1) might differ only by a multiplicative function which varies slowly, the qualitative picture behind the mechanism that gives rise to them is dramatically different. The first term in the right hand side of (7.1) arises from the same type of phenomena behind the tail behaviour in the case $\rho < 1$.

In Section 7.1, apart from introducing the notation required to precisely state our results, we discuss at length the intuition behind both the asymptotic results (7.2) and (7.3), as well as our asymptotic expression (7.1). At this point, it suffices to say that the phenomena underlying the development of (7.2) and (7.3) are a combination of two features, first, arrival of large jobs whose effects persist for long time scales, and, second, the impact of such effects, which is measured using the Law of Large Numbers. In contrast, the development of (7.1) involves not only the combination of these two features, but, in addition, one has to account for the impact of effects which occur at the scales governed by the Central Limit Theorem.

We identify another interesting phenomenon when the job sizes have infinite variance: If $\rho = 1$ and $\alpha \in (1, 2)$, it turns out that the asymptotics are governed by

$$\mathbb{P}\{W > b\} = \Theta\left(b^{\alpha}\bar{B}(b^{\alpha})\right),$$

suggesting that the tail behaviour is closer to the case $\rho > 1$ than to the case $\rho < 1$. This is a sharp transition from the system behaviour when $\mathrm{Var}[V] < \infty$, where the tail asymptotic is closer to the $\rho < 1$ case. Such surprising transitions in system behaviour seem to be unique to the integer traffic intensity case.

In summary, the qualitative development behind our asymptotic bounds introduces a combination of elements that are not typical in the asymptotic analysis of multiserver queues. After developing necessary intuition behind the results (7.1), (7.2) and (7.3) in Section 7.1, we derive the respective lower and upper bounds in (7.1) in Sections 7.2 and 7.3. Apart from unraveling surprising transitions in the system behaviour that seem to happen only when the traffic intensity is an integer, an important contribution is in the use of regenerative ratio representation and Lyapunov bound techniques to characterize tail behaviour of steady-state delay in multiserver queues. An alternate proof for the upper bound, that takes inspiration from a completely different approach due to Foss and Korshunov [2006], is presented in Section 7.5 at the end of this chapter. However, in Foss and Korshunov [2006], it is crucial to have ρ not equal to an integer so that certain upper bound processes might be defined. So, we believe that our alternate approach presented in Section 7.5 might add useful ideas to the traditional

techniques used in the asymptotic analysis of multiserver queues. We conclude in Section 7.7 after providing proofs of certain results in Section 7.6.

7.1 The main result and its intuition

We consider a two-server queue that processes incoming jobs under the first-come-first-serve discipline. Jobs are indexed by the order of arrival. Job 0 arrives at time 0, and for $n \geq 1$, job n arrives at time $T_1 + \ldots + T_n$. Job n requires service for time V_n. Here the sequence of interarrival times $(T_n : n \geq 1)$ and service times $(V_n : n \geq 0)$ are taken to be i.i.d. copies, respectively, of the generic interarrival and service time variables T and V. As mentioned in the Introduction, we assume that $\mathbb{E}T = \mathbb{E}V$, and hence the traffic intensity ρ, which is the ratio between $\mathbb{E}V$ and $\mathbb{E}T$, equals 1. To make the computations easier, we assume, without loss of generality, that $\mathbb{E}T = 1$ (otherwise, time can always be rescaled to make this hold). Additionally, we make the following assumptions on the distributions of V and T.

Assumption 7.1. *The tail distribution of V admits the representation,*

$$\bar{B}(x) := \mathbb{P}\{V > x\} = x^{-\alpha} L(x),$$

for some $\alpha > 1$ and a function $L(\cdot)$ slowly varying at infinity, that is, $\lim_{x\to\infty} L(tx)/L(x) = 1$ for every $t > 0$.

Assumption 7.2. $\mathbb{P}\{T > x\} = o(\bar{B}(x))$, *as* $x \to \infty$.

Assumption 7.2 is quite natural given that typically one models interarrival times as exponentially distributed random variables. We also use the notation

$$X_{n+1} = V_n - T_{n+1} \text{ for } n \geq 0.$$

Since T is non-negative, the right-tail of $X := V - T$ is asymptotically similar to that of V (see, for example, Corollary 1.11 in Chapter IX of Asmussen [2000]). In other words,

$$\mathbb{P}\{X > x\} \sim \bar{B}(x) \text{ as } x \to \infty. \tag{7.4}$$

The ordered workload vector of the servers as seen by the n^{th} job during its arrival, denoted by $\mathbf{W}_n = \left(W_n^{(1)}, W_n^{(2)}\right)$, satisfies the well-known Kiefer-Wolfowitz recursion (see Kiefer and Wolfowitz [1955]):

$$W_{n+1}^{(1)} = \left(W_n^{(1)} + V_n - T_{n+1}\right)^+ \wedge \left(W_n^{(2)} - T_{n+1}\right)^+ \text{ and} \tag{7.5a}$$

$$W_{n+1}^{(2)} = \left(W_n^{(1)} + V_n - T_{n+1}\right)^+ \vee \left(W_n^{(2)} - T_{n+1}\right)^+. \tag{7.5b}$$

Here, $W_n^{(1)}$ denotes the minimum of the remaining workload of the servers when the n^{th} job arrives, and $W_n^{(2)}$ denotes the respective maximum. As a result, $W_n^{(1)} \leq W_n^{(2)}$. To gain an intuitive understanding of the recursions (7.5a) and (7.5b), first see that due to FCFS service discipline, the n-th job will join the server which has minimum remaning workload at the time of its arrival. Therefore the ordered workload immediately after the arrival of n-th job becomes

$$\left(\left(W_n^{(1)} + V_n\right) \wedge W_n^{(2)},\ \left(W_n^{(1)} + V_n\right) \vee W_n^{(2)}\right).$$

Then, the recursions (7.5a) and (7.5b) are immediate once we observe that the next job arrives after T_{n+1} units of time after the arrival of n-th job.

Since $\rho < 2$, the queue is stable in the sense that the weak limit (limit in distribution) of $\mathbf{W}_n$, denoted by $\mathbf{W}_\infty$, exists and we are interested in deriving bounds for the tail probabilities of the steady-state waiting time

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = \lim_{n\to\infty} \mathbb{P}\left\{W_n^{(1)} > b\right\},$$

for large values of b. Our main result is the following.

Theorem 7.1. *Suppose that $\rho = 1$ and Assumptions 7.1 and 7.2 are in force. If $\alpha > 2$, then*

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = \Theta\left(b^2\bar{B}(b^2) + b^2\bar{B}^2(b)\right),\ \text{as } b \to \infty. \tag{7.6}$$

If $\alpha \in (1,2)$, under the additional assumption that $\bar{B}(x) \sim cx^{-\alpha}$ for some $c > 0$, we have that

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = \Theta\left(b^\alpha\bar{B}(b^\alpha)\right),\ \text{as } b \to \infty. \tag{7.7}$$

We now proceed to discuss how this result contrasts with what is known in the literature and thereby expose the intuition behind it.

Discussion of earlier results in the literature

As indicated in the Introduction, the tail asymptotics of steady-state delay is known depending on the case $\rho < 1$ (or) $\rho \in (1,2)$, and is given by (7.2) and (7.3), respectively. In order to see the mechanism behind these two asymptotics, let us assume without loss of generality that $\mathbb{E}T = 1$ (if not, time can be rescaled to make this assumption hold). Additionally, let us assume that the generic interarrival time T has unbounded support (for example, T is exponentially distributed), and consider the regenerative ratio representation

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = \frac{\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right]}{\mathbb{E}_{\mathbf{0}}\left(\tau_0\right)}, \tag{7.8}$$

where $\tau_0 = \inf\{n \geq 1 : W_n^{(2)} = 0\}$ denotes the first time when the Kiefer-Wolfowitz process $\mathbf{W}_n$ enters the set $\{(0,0)\}$. Since $\rho < 2$ and T has unbounded support, the state $(0,0)$ is recurrent, thus leading to the regenerative ratio representation (7.8) For simplicity, throughout our discussions, we shall assume that T has an unbounded support. This assumption is merely technical. It can be relaxed at the price of using a slightly more complicated regenerative representation. For further details on the representation (7.8) and details on relaxing the assumption on support of T, see, for example, Borovkov [1984], Foss [1986], Kalashnikov [1980], Kalashnikov and Rachev [1990], or Foss and Kalashnikov [1991]. Moreover, our alternate proof of the upper bound presented in Section 7.5 does not rely on this assumption.

In order to study (7.8), define the stopping times

$$\tau_b^{(i)} := \inf\{n \geq 0 : W_n^{(i)} > b\}, \quad i = 1, 2.$$

First, let us consider the event $\{\tau_b^{(1)} < \tau_0\}$, which is the event that there is at least one customer who waits more than b units of time in a busy period. Moreover, since $\tau_b^{(2)} < \tau_b^{(1)}$, it is instructive to first consider the event $\{\tau_b^{(2)} < \tau_0\}$, which can be seen, intuitively, to be caused by the arrival of a big job of size larger than b within the initial $O(1)$ units of time in the busy period. Due to this reasoning, one can write

$$\begin{aligned} \mathbb{P}_{\mathbf{0}}\left\{\tau_b^{(2)} < \tau_0\right\} &= \Theta\left(\mathbb{P}\{V > b\}\right) \text{ and} \\ \mathbb{P}\left\{W^{(2)}_{\tau_b^{(2)}} > x \,\middle|\, \tau_b^{(2)} < \tau_0\right\} &\approx \mathbb{P}\left\{V > x \,\middle|\, V > b\right\}, \; x > b. \end{aligned} \tag{7.9}$$

Therefore, one can approximately characterize the process $\mathbf{W}$, immediately after the arrival of the first big job of size larger than b, as below:

$$\frac{1}{b}\mathbf{W}_{\tau_b^{(2)}} = \frac{1}{b}\left(W^{(1)}_{\tau_b^{(2)}}, W^{(2)}_{\tau_b^{(2)}}\right) \approx (0, Z), \tag{7.10}$$

where Z satisfies $\mathbb{P}\{Z > x\} = \lim_{b\to\infty} \mathbb{P}\left\{V > bx \,\middle|\, V > b\right\} = x^{-\alpha}$ for $x \geq 1$. As per recursions (7.5a) and (7.5b), the server that gets to process this big job cannot process any new arrivals until both the workloads become comparable again at some time in the future, which we refer as τ_{eq}. During this period where one of the servers is effectively blocked from processing new arrivals (call it the blocked server and the other server as active server), the dynamics of the queue is given by:

$$\mathbf{W}_n = \left(\left(W_{n-1}^{(1)} + V_{n-1} - T_n\right)^+, W_{n-1}^{(2)} - T_n\right), \quad \tau_b^{(2)} < n < \tau_{eq}.$$

The dynamics of the active server matches with that of the single server queue, and hence the waiting time experienced by the k^{th} job after the big jump can be be roughly approximated, in

distribution, by maximum of k steps of a random walk with increments that are i.i.d. copies of $V - T$. Observe that the aforementioned random walk has drift equal to $\rho - 1$, which can be positive, zero (or) negative, respectively, based on whether $\rho > 1, \rho = 1$ (or) $\rho < 1$. As a consequence, the maximum of the random walk, in the respective cases, can be of magnitude $O(k), O(\sqrt{k})$ (or) $O(1)$ in k units of time (this can be seen by invoking Law of Large Numbers and Central Limit Theorem for i.i.d. sums). Therefore, due to (7.10), the workload until time τ_{eq} can be approximately written as

$$W^{(1)}_{\tau_b^{(2)}+k} \approx \begin{cases} c_1 k & \text{if } \rho > 1, \\ O\left(\sqrt{k}\right) & \text{if } \rho = 1, \\ O(1) & \text{if } \rho < 1 \end{cases} \quad \text{and } W^{(2)}_{\tau_b^{(2)}+k} \approx bZ - k \tag{7.11}$$

for some positive constant c_1. Because of this clear difference in behaviour of $W^{(1)}$ based on the value of ρ, we need to consider cases $\rho \in (1,2), \rho < 1$ and $\rho = 1$ separately. We once again stress that our discussion in this section is completely heuristic, aiming to emphasize the intuition behind the results. While cases $\rho \in (1,2)$ and $\rho < 1$ are treated rigorously in Foss and Korshunov [2006], future sections in this chapter are devoted to the rigorous treatment of the case $\rho = 1$.

Case 1: $\rho \in (1,2)$

If $\rho \in (1,2)$, then one server is not enough to keep the system stable. As a result, when one server is blocked for $O(bZ)$ units of time due to the arrival of a big job, the active server effectively becomes a single server processing all the arrivals, and hence the workload $W^{(1)}$ gradually increases with time as in (7.11). Recall that τ_{eq} is the time where both the servers have roughly equal workload, and therefore due to (7.11), we solve for τ_{eq} by setting

$$c_1\left(\tau_{eq} - \tau_b^{(2)}\right) \approx bZ - \left(\tau_{eq} - \tau_b^{(2)}\right).$$

As a result, $W^{(1)}$ increases roughly up to time

$$\tau_{eq} \approx \tau_b^{(2)} + \frac{bZ}{c_1 + 1},$$

when both $W^{(1)}$ and $W^{(2)}$ become comparable, after which both the servers jointly process incoming arrivals according to (7.5a) and (7.5b), resulting in a total decrease of workload at rate $2 - \rho$. In this mechanism, for any job to be delayed by more than b units of time, it must happen that $c_1 k \geq b$ for some $k \leq bZ/(c_1 + 1)$, and therefore,

$$\lim_{b\to\infty} \mathbb{P}_{\mathbf{0}}\left\{\tau_b^{(1)} < \tau_0 \;\middle|\; \tau_b^{(2)} < \tau_0\right\} = \lim_{b\to\infty} \mathbb{P}\left\{c_1 \frac{bZ}{c_1+1} \geq b\right\} = \mathbb{P}\left\{Z > 1 + \frac{1}{c_1}\right\} > 0. \tag{7.12}$$

If we let N_1 to denote the number of jobs that experience at least b units of delay up to time τ_{eq} and N_2 to denote the respective count after τ_{eq}, then the above heuristics suggest that

$$N_1 = \frac{\left(W^{(1)}_{\tau_{eq}} - b\right)^+}{c_1} = \left(\frac{bZ}{c_1+1} - \frac{b}{c_1}\right)^+ \text{ and}$$
$$N_2 = \frac{\left(W^{(1)}_{\tau_{eq}} - b\right)^+}{2-\rho} = \frac{1}{2-\rho}\left(\frac{c_1 bZ}{c_1+1} - b\right)^+.$$

Therefore, due to (7.12), we obtain that

$$\begin{aligned}\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W^{(1)}_k > b\right)\right] &= \mathbb{E}_0\left[\sum_{k=0}^{\tau_0-1} I\left(W^{(1)}_k > b\right) \;\middle|\; \tau^{(1)}_b < \tau_0\right] \times \mathbb{P}\left\{\tau^{(1)}_b < \tau_0\right\} \\ &\approx \mathbb{E}\left[N_1 + N_2 \;\middle|\; Z > 1 + \frac{1}{c_1}\right] \times \mathbb{P}\left\{Z > 1 + \frac{1}{c_1}\right\} \times \Theta\left(\mathbb{P}\left\{V > b\right\}\right) \\ &= \Theta\left(b \times \mathbb{P}\left\{Z > 1 + \frac{1}{r}\right\} \times \bar{B}(b)\right).\end{aligned}$$

As a result, from (7.8), we obtain that

$$\mathbb{P}\left\{W^{(1)}_\infty > b\right\} = \Theta\left(b\bar{B}(b)\right),$$

which is precisely same as (7.3). This final form of asymptotic is rigorously established in Foss and Korshunov [2006] with exact evaluation of the constants, albeit, using a different reasoning.

Case 2: $\rho < 1$

If $\rho < 1$, conditional on the occurrence of $\{\tau^{(2)}_b < \tau_0\}$, it is no longer true that the event $\{\tau^{(1)}_b < \tau_0\}$ happens with positive probability as $b \to \infty$ (compare this with (7.12) when $\rho \in (1,2)$). The reason is that if $\rho < 1$, the system is stable and the workload remains $O(1)$, as in (7.11), even if one removes one server and force it to operate as a single server system. As a result, we need to invoke heavy-tailed large deviations behaviour, which dictates that arrival of one more job of size larger than b is required, typically, to experience waiting time larger than b. This requirement is dealt as follows: Conditional on the occurrence of $\{\tau^{(2)}_b < \tau_0\}$, as in (7.11), we have

$$W^{(1)}_{\tau^{(2)}_b + k} = O(1) \text{ and } W^{(2)}_{\tau^{(2)}_b + k} \approx bZ - k.$$

Here, the workload $W^{(2)}$ becomes smaller than b if $k > b(Z-1)$, and therefore, the cheapest way to observe large delays (of duration at least b) is to have a $K \leq b(Z-1)$ such that the $(\tau^{(2)}_b + K)^{th}$ job requires service for duration larger than b. Following the same line of reasoning

behind (7.10), we approximate the size of the second big job by $b\hat{Z}$, where $\hat{Z}$ is an independent copy of Z. As a result, we arrive at the following distributional approximation :

$$\mathbf{W}^{(1)}_{\tau_b^{(1)}} \approx \min\left(bZ - K_1, b\hat{Z}\right) \text{ and } \mathbf{W}^{(2)}_{\tau_b^{(1)}} \approx \max\left(bZ - K_1, b\hat{Z}\right).$$

Next, the number of jobs that get delayed by more than b units of time (which depends on K) is approximately given by

$$N(K) := \frac{\min(bZ - K, b\hat{Z}) - b}{2 - \rho},$$

where $K \leq b(Z-1)$. As a result,

$$\begin{aligned}
\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right] &\approx \mathbb{E}\left[N(K)\, I\left(0 \leq K \leq b(Z-1)\right) \;\middle|\; \tau_b^{(2)} < \tau_0\right] \times \mathbb{P}\left\{\tau_b^{(2)} < \tau_0\right\} \\
&= \mathbb{E}\left[\sum_{k=1}^{b(Z-1)} \min\left(b(Z-1) - k, b(\hat{Z}-1)\right) I\left(V_{\tau_b^{(2)}+k} > b\right) \;\middle|\; \tau_b^{(2)} < \tau_0\right] \times \Theta\left(\mathbb{P}\{V > b\}\right) \\
&= \Theta\left(b^2 \mathbb{P}\{V > b\}^2\right),
\end{aligned}$$

and therefore, $\mathbb{P}\{W_\infty^{(1)} > b\} = \Theta(b^2 \bar{B}^2(b))$, which coincides with (7.2).

Intuitive discussion of Theorem 1: The case $\rho = 1$

Our goal in this discussion is to communicate the following insights:

1) Contrary to Case 1 and Case 2, the conditional distribution of the Kiefer-Wolfowitz vector $\mathbf{W}$ given that $\{\tau_b^{(1)} < \tau_0\}$ does not fully explain the mechanism behind the asymptotic results in Theorem 1.

2) Unlike Cases 1 and 2, it is not enough to account for the impact of the large service times using linear dynamics which evolve according to the Law of Large Numbers.

We shall first concentrate on the situation where the job sizes V have finite variance, more precisely, the case $\alpha > 2$. The case $\alpha \in (1, 2)$ can be understood using similar ideas. We shall leverage off the type of arguments that were given for Case 1 and Case 2. Since $\rho = 1$ sits right in the middle we shall consider two mechanisms, one involving two jumps (analogous to Case 2), and one involving one jump (analogous to Case 1).

Delays due to two jumps: Conditional on $\{\tau_b^{(2)} < \tau_0\}$, similar to cases 1 and 2, the dynamics of the active server and the blocked server, as in (7.11), are given respectively by

$$W^{(1)}_{\tau_b^{(2)}+k} \approx O\left(\sqrt{k}\right) \text{ and } W_{\tau_b^{(2)}+k} \approx bZ - k, \tag{7.13}$$

for k such that $\tau_b^{(2)}+k \leq \tau_{eq}$. As discussed previously, fluctuations of order $\sqrt{k}$ arise in workload due to the Central Limit Theorem, and this phenomenon, as we shall see below, gains relevance only when $\rho = 1$. Since our interest here is in studying delays due to the occurrence of two big jumps, as in Case 2, if there exists a $K < b(Z-1)$ such that $(K+\tau_b^{(2)})$-th customer brings a job of size $b - O(\sqrt{b})$ or larger, then at least one job gets delayed by b units or more. The contribution to $\mathbb{P}\{\tau_b^{(1)} < \tau_0\}$ due to the occurrence of 2 jumps can be calculated as below:

$$P_{2 \text{ jumps}}(b) := \mathbb{P}\left\{\tau_b^{(2)} < \tau_0\right\} \times \Theta\left(\mathbb{P}\left\{V_{\tau_b^{(2)}+k} > b - \sqrt{b} \text{ for some } k \leq b(Z-1) \;\middle|\; \tau_b^{(2)} < \tau_0\right\}\right) \tag{7.14}$$

$$= \Theta\left(\mathbb{P}\left\{V > b\right\} \times \sum_{k=1}^{\infty} \mathbb{P}\left\{bZ > k, V_k > b - \sqrt{b}\right\}\right)$$

$$= \Theta\left(\mathbb{P}\left\{V > b\right\}^2 \times \sum_{k=1}^{\infty} \mathbb{P}\left(bZ > k\right)\right) = \Theta\left(b\bar{B}^2(b)\right).$$

Following the same line of reasoning as in Case 2, we obtain the following contribution to $\mathbb{E}_{\mathbf{0}}[\sum_{k=0}^{\tau_0-1} I(W_k^{(1)} > b)]$ due to 2 jumps:

$$Q_{2 \text{ jumps}}(b) = \Theta\left(b^2\bar{B}^2(b)\right). \tag{7.15}$$

Delay due to 1 jump: Similar to Case 1, when $\rho = 1$, the active server accumulates work, albeit at a slower rate, as given in (7.13). Since the workload of a critically loaded single server queue grows like $O(\sqrt{k})$ in k units of time, it is intuitive to expect that if there is a big jump of size exceeding b^2 in the first $O(1)$ units of time of the busy period, subsequently one of the servers gets blocked for more than b^2 units of time, and the active server which faces all the incoming traffic accumulates workload of size larger than b, with non-vanishing probability, in those b^2 units of time. Therefore, similar to (7.12), we have that

$$\lim_{b\to\infty} \mathbb{P}\left\{\tau_b^{(1)} < \tau_0 \;\middle|\; \tau_{b^2}^{(2)} < \tau_0\right\} > 0.$$

Therefore, due to (7.9), the contribution to $\mathbb{P}\{\tau_b^{(1)} < \tau_0\}$ due to the arrival of only one big job is given by

$$P_{1 \text{ jump}}(b) \approx \mathbb{P}\left\{\tau_b^{(1)} < \tau_0 \;\middle|\; \tau_{b^2}^{(2)} < \tau_0\right\} \times \mathbb{P}\left\{\tau_{b^2} < \tau_0\right\} = \Theta\left(\bar{B}(b^2)\right),$$

which is negligible compared to the right hand side of (7.14). As a result, we have that

$$\mathbb{P}\left\{\tau_b^{(1)} < \tau_0\right\} \sim P_{2 \text{ jumps}}(b) = \Theta\left(b\bar{B}^2(b)\right).$$

However, accounting for the number of jobs that experience at least b units of delay dramatically changes the contribution of this single huge jump in the computation of steady-state delay

probabilities. In particular, a single jump of size exceeding b^2 blocks one of the servers for $V \mid V > b^2 \approx b^2 Z$ units of time, and if we perform calculations similar to Case 1, we shall obtain that $\Theta(b^2)$ jobs experience delays larger than b. As a consequence, we have the following contribution in the single, huge jump regime:

$$Q_{1 \text{ jump}} := \mathbb{E}\left[\sum_{i=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right) \;\middle|\; \tau_{b^2}^{(2)} < \tau_0\right] \times \mathbb{P}\left\{\tau_b^{(2)} < \tau_0\right\} = \Theta\left(b^2 \mathbb{P}\left\{V > b^2\right\}\right),$$

which might not be negligible to the corresponding contribution due to 2 jumps derived in (7.15). In fact, as demonstrated in an example in the Introduction, this contribution due to single huge jump could be larger than its counterpart for 2 jumps based on the slowly varying function $L(\cdot)$ (consider the example $L(x) = \log(1+x)$). As a result, we have two competing components in the expression for steady-state probability of delay in (7.6).

We conclude with a heuristic explanation of the mechanism involving one jump for the case $\alpha \in (1,2)$ if $\rho = 1$. In this case, once a server is blocked for k units of time, the active server operates as a critical single-server queue, processing jobs requiring services with infinite variance, and due to the generalized Central Limit Theorem, the workload of the critical queue exhibits fluctuations of order $O(k^{1/\alpha})$. Therefore, if the initial huge jump, which occurs within $O(1)$ units of time at the beginning of the busy period, is of size larger than b^α, then this huge job blocks one of the servers for more than b^α units of time, and as a result, $\Theta(b^\alpha)$ jobs wait for a duration larger than b. Reasoning as in the finite variance case, the contribution to steady-state delay due to the arrival of one huge job is $\Theta(b^\alpha \mathbb{P}\{V > b^\alpha\}) = \Theta(b^\alpha b^{-\alpha^2} L(b^\alpha))$. On the other hand, the contribution arising from two jumps as in Case 2, namely, according to (7.2), remains $\Theta(b^{2-2\alpha} L(b)^2)$, which is negligible compared to $\Theta(b^\alpha b^{-\alpha^2} L(b^\alpha))$ because $\alpha \in (1,2)$ implies $2\alpha - 2 > \alpha^2 - \alpha$. Hence, we arrive at the estimate (7.7) in Theorem 1. Results of similar nature, although in a different setup, has been obtained in Zwart et al. [2005] and Debicki et al. [2013].

7.2 Proof of lower bound

The objective of this section is to prove the following result.

Proposition 7.1. *Suppose that Assumption 7.2 holds, and that $\rho = 1$. Then, if Assumption 7.1 holds with $\alpha > 2$, there exists $c_1 > 0$ and $b_0 > 0$ such that for all $b > b_0$.*

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} \geq c_1 \left(b^2 \bar{B}(b^2) + b^2 \bar{B}^2(b)\right).$$

On the other hand, if $\bar{B}(x) \sim cx^{-\alpha}$ *as* $x \to \infty$ *for some* $c > 0$ *and* $\alpha \in (1,2)$*, then there exists* $c_1 > 0$ *and* $b_0 > 0$ *such that for all* $b > b_0$,

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} \geq c_1 \left(b^\alpha \bar{B}(b^\alpha)\right).$$

We now provide the proof of Proposition 7.1.

Case 1: (Under the assumption that $\alpha > 2$). We first derive a lower bound based on a single big jump of size exceeding b^2. Let $N_A(t)$ denote the number of jobs that arrive in the interval $(0,t]$. Let $b > 2$ and consider the event, D_1, with the following properties:

1) The coordinate $W_1^{(2)} > 6b^2$ (that is, Job 0 blocks one of the servers for $\Omega\left(b^2\right)$ time units).
2) The total amount of work brought by all the jobs that arrive in the time interval $(0, 2b^2]$ does not exceed $3b^2$. In other words, $V_1 + \ldots + V_{N_A(2b^2)} \leq 3b^2$.
3) Every job that arrives in the time interval $[b^2, 2b^2]$ experiences delay for at least b units of time before getting processed. That is,

$$\min_{N_A(b^2) \leq n \leq N_A(2b^2)} W_n^{(1)} > b.$$

On the set D_1, the dynamics of the queue described by recursions (7.5a) and (7.5b) reduces to

$$W_n^{(1)} = (W_{n-1} + X_n)^+ \text{ and } W_n^{(2)} = W_1^{(2)} - (T_2 + \ldots + T_n)$$

for $2 \leq n \leq N_A(2b^2)$. Further, if we let $S_1 := W_1^{(1)} = 0$ and $S_n := X_2 + \ldots + X_n$ for $n \geq 2$, then the following holds on the set D_1 :

$$\min_{N_A(b^2) \leq n \leq N_A(2b^2)} W_n^{(1)} \geq \min_{N_A(b^2) \leq n \leq N_A(2b^2)} S_n.$$

As a result,

$$\begin{aligned}
\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right] &\geq \mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right); D_1\right] \\
&\geq \mathbb{E}_{\mathbf{0}}\left[N_A\left(2b^2\right) - N_A\left(b^2\right); W_1^{(2)} > 6b^2, \min_{N_A(b^2) \leq n \leq N_A(2b^2)} S_n > 1.5b, \sum_{i=1}^{N_A(2b^2)} V_i \leq 3b^2\right] \\
&\geq \mathbb{P}\left\{X_1 > 6b^2\right\} \mathbb{E}\left[N_A\left(2b^2\right) - N_A\left(b^2\right); D_1'\right],
\end{aligned}$$

where the event

$$D_1' := \left\{ N_A\left(b^2\right) \geq 0.5b^2,\ N_A\left(2b^2\right) \in \left[1.5b^2, 2.5b^2\right],\ \min_{0.5b^2 \leq n \leq 2.5b^2} S_n > b,\ \sum_{i=1}^{\lceil 2.5b^2 \rceil} V_i \leq 3b^2 \right\}$$

has probability at least

$$\mathbb{P}\left\{ \inf_{0.5 \leq t \leq 2.5} \sigma B(t) > 1 \right\} (1 - o(1))$$

because of functional CLT and the facts that $N_A(x)/x \to 1$ and $(V_1 + \ldots + V_n)/n \to 1$ with probability one. Here $B(\cdot)$ is a standard Brownian motion and σ^2 denotes the variance of X. Additionally, due to the regenerative ratio representation (7.8) and the regularly varying nature of the tail of X (recall that $\mathbb{P}\{X > x\} \sim \bar{B}(x)$ as $x \to \infty$), we conclude that

$$\begin{aligned} \mathbb{P}\left\{ W_\infty^{(1)} > b \right\} &\geq \frac{b^2 \mathbb{P}\left\{X > 6b^2\right\} \times \mathbb{P}\left(D_1'\right)}{\mathbb{E}\tau_0} \\ &\geq c_1 b^2 \bar{B}\left(b^2\right) \end{aligned} \tag{7.16}$$

for some $c_1 > 0$ and all b large enough.

Case 2: (Also under the assumption that $\alpha > 2$). We now derive a lower bound based on the occurrence of two jumps, each of size exceeding b. Let $b > 2$ and consider the event, D_2, with the following properties:

1) The coordinate $W_1^{(2)} > 5b$ (that is, Job 0 blocks one of the servers for $\Omega(b)$ time units).

2) Apart from Job 0, only one of the $N_A(b)$ jobs that arrive in the time interval $(0, b]$ bring a service requirement of size exceeding $5b$.

3) The number of customers who arrive during the time intervals $(0, b]$ and $(b, 2b]$ are numbers between $0.5b$ and $1.5b$. Alternatively, $N_A(b) \in [0.5b, 1.5b]$ and $N_A(2b) - N_A(b) \in [0.5b, 1.5b]$.

So, on the set D_2 we have that at least $N_A(2b) - N_A(b) \geq 0.5b$ jobs experience a waiting time more than b units of time, and hence

$$\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W_k^{(1)} > b \right); D_2 \right] \geq 0.5b \mathbb{P}\left(D_2\right).$$

However, since $N_A(x)/x \to \infty$, we have that

$$\mathbb{P}\left\{ N_A(b) \in [0.5b, 1.5b],\ N_A(2b) - N_A(b) \in [0.5b, 1.5b] \right\} \sim 1,$$

as $b \to \infty$. As a result,

$$\begin{aligned}\mathbb{P}(D_2) &\geq (1-o(1)) \sum_{k\leq 0.5b} \mathbb{P}_{\mathbf{0}}\left\{W_1^{(2)} > 5b,\ V_k > 5b,\ \bigcap_{i\leq 1.5b, i\neq j} \{V_i < 5b\}\right\} \\ &\geq 0.5b\mathbb{P}\{X_1 > 5b\}\bar{B}(5b)\left(1-\bar{B}(5b)\right)^{1.5b}(1-o(1)) \\ &\geq b\bar{B}^2(5b)(1-o(1))\end{aligned}$$

Then, as in Case 1, due to the regenerative ratio representation (7.8) and the regularly varying nature of $\bar{B}(\cdot)$, we conclude that there exists a constant c_2 such that

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} \geq \frac{0.5b \times b\bar{B}^2(5b)}{\mathbb{E}\tau_0}(1-o(1)) \geq c_1 b^2 \bar{B}(b)^2. \tag{7.17}$$

Combining (7.16) and (7.17) we obtain the statement of Proposition 7.1 for the case $\alpha > 2$.

Case 3: We now consider the assumption that $\alpha \in (1,2)$ and $\bar{B}(x) \sim cx^{-\alpha}$ as $x \to \infty$. The strategy is similar to Case 1. Define an event, D_3, satisfying the following properties:

1) The coordinate $W_1^{(2)} > 6b^\alpha$ (that is, Job 0 blocks one of the servers for $\Omega(b^\alpha)$ time units).
2) The total amount of work brought by all the jobs that arrive in the time interval $(0, 2b^\alpha]$ does not exceed $3b^\alpha$. In other words, $V_1 + \ldots + V_{N_A(2b^\alpha)} \leq 3b^\alpha$.
3) Every job that arrives in the time interval $[b^\alpha, 2b^\alpha]$ experiences delay for at least b units of time before getting processed. That is,

$$\min_{N_A(b^2)\leq n\leq N_A(2b^2)} W_n^{(1)} > b.$$

Then, following the same steps as in Case 1, we obtain that

$$\mathbb{E}_{\mathbf{0}}\left[\sum_{k=1}^{\tau_0} I\left(W_k^{(1)} > b\right)\right] \geq \bar{B}(6b^\alpha)\,\mathbb{E}\left[N_A(2b^\alpha) - N_A(b^\alpha); D_3'\right]$$

where the event

$$D_3' := \left\{N_A(b^\alpha) \geq 0.5b^\alpha,\ N_A(2b^\alpha) \in [1.5b^\alpha, 2.5b^\alpha],\ \min_{0.5b^\alpha\leq n\leq 2.5b^\alpha} S_n > b,\ \sum_{i=1}^{\lceil 2.5b^\alpha\rceil} V_i \leq 3b^\alpha\right\}$$

has non-vanishing probability as $b \to \infty$ because $b^{-1}S_{[tb^\alpha]}$ converges weakly in $D[0,\infty)$, to a Stable process $Z(\cdot)$. As a result, we obtain

$$\mathbb{E}_{\mathbf{0}}\left[\sum_{k=1}^{\tau_0} I\left(W_k^{(1)} > b\right)\right] \geq \bar{B}(6b^\alpha) \times \mathbb{P}\left\{\inf_{1\leq t\leq 3} Z(t) > 1\right\}(1-o(1)).$$

This observation, along with the regenerative ratio representation (7.8), concludes the proof of Proposition 7.1.

7.3 Proof of upper bound

The objective of this section is to prove the following proposition.

Proposition 7.2. *Suppose that Assumption 7.2 holds, and that $\rho = 1$. Then, if Assumption 7.1 holds with $\alpha > 2$, there exist $c_1 > 0$ and $b_0 > 0$ such that for all $b > b_0$,*

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} \leq c_1 \left(b^2 \bar{B}(b^2) + b^2 \bar{B}^2(b)\right).$$

On the other hand, if $\bar{B}(x) \sim cx^{-\alpha}$, as $x \to \infty$, for some $c > 0$ and $\alpha \in (1,2)$, then one can find positive constants c_1 and b_0 such that for all $b > b_0$,

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} \leq c_1 \left(b^\alpha \bar{B}(b^\alpha)\right).$$

The rest of this section is devoted to the proof of Proposition 7.2. First, pick $\delta_-, \delta, \delta_+$ such that $0 < \delta_- < \delta < \delta_+ < 1$. In addition to the stopping times

$$\tau_x^{(i)} = \inf\left\{n \geq 0 : W_n^{(i)} > x\right\},$$

which are defined for $x > 0, i = 1$ and 2, let us define

$$\bar{\tau}_{b\delta_+}^{(2)} = \inf\left\{n \geq \tau_{b\delta_-}^{(2)} : W_n^{(2)} \leq b\delta_+\right\}.$$

Additionally, let

$$B_1(b) := \mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right) I\left(\bar{\tau}_{b\delta_+}^{(2)} > \tau_{b\delta}^{(1)}\right)\right] \text{ and}$$

$$B_2(b) := \mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right) I\left(\bar{\tau}_{b\delta_+}^{(2)} \leq \tau_{b\delta}^{(1)}\right)\right].$$

Then, it follows from the regenerative ratio representation (7.8) that

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = \frac{B_1(b) + B_2(b)}{\mathbb{E}_{\mathbf{0}}[\tau_0]}. \tag{7.18}$$

The term $B_1(b)$ corresponds to the case where all the actions happen: once there is a large jump in $W^{(2)}$ which takes it beyond $b\delta_-$, one of the servers gets blocked for a long time, and the other server which faces the entire traffic in that duration piles up work more than $b\delta$. On the other hand, the term $B_2(b)$ corresponds to the case where the first jump is wasted: that is, there is not enough buildup in $W^{(1)}$ after the occurrence of first jump in $W^{(2)}$. The rigorous procedure of obtaining upper bounds for $B_1(b)$ and $B_2(b)$ is divided into several parts:

Part 1) First, we obtain an upper bound for $\mathbb{E}_{\mathbf{w}}[\tau_0]$ uniformly over all initial conditions $\mathbf{w} = (w_1, w_2)$. This shall be useful in obtaining upper bounds for both $B_1(b)$ and $B_2(b)$ because of the simple observation that

$$\mathbb{E}_{\mathbf{w}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right] \leq \mathbb{E}_{\mathbf{w}}[\tau_0].$$

Additionally, in an attempt to obtain a stochastic description of the workload $W^{(2)}$ after it exceeds δb_-, we derive a stochastic domination result for $W^{(2)}_{\tau_b^{(2)}}$ which shall be useful.

Part 2) We reduce the contribution of the first term $B_1(b)$ into a large deviations problem for zero-mean random walks with regularly varying increments. We use the stochastic domination result obtained in Part 1) along with another domination argument, in terms of a suitably defined critically loaded single-server queue, to account for all of what happens after the first jump. In turn, the introduction of the single-server queue sets the stage for the use of uniform large deviations for random walks. The analysis of part 2) emphasizes the convenience of partitioning the numerator in (7.8) into $B_1(b)$ and $B_2(b)$.

Part 3.a) This is the portion of the argument that requires $\alpha > 2$. It invokes classical results for uniform large deviations of regularly varying random walks available due to Nagaev (uniform in the sense that the asymptotics jointly account both the Brownian approximations in the CLT scaling regime and the large deviations approximations in scaling regimes beyond that of CLT). The execution of part 2) involves routine estimations of one dimensional integrals using basic properties of regularly varying distributions. We obtain the required upper bound for $B_1(b)$ after some elementary simplifications.

Part 3.b) The analysis here is entirely parallel to that of part 3.a), except that the uniform estimates involve an approximation using an α-stable process (instead of Brownian motion as in Part 3.a)).

Part 4) is devoted to obtaining an upper bound for the residual term $B_2(b)$. This is accomplished by first performing calculations that result in an intermediate bound for $B_2(b)$ in terms of expected number of jobs that wait for duration longer than b after the first jump. The second calculation involves obtaining a good upper bound for $\mathbb{P}_{\mathbf{w}}\{\tau_b^{(2)} < \tau_0\}$ uniformly over initial conditions $\mathbf{w} \in \{(w_1, w_2) : w_2 < b\delta_+\}$.

In the following subsections we shall estimate the contributions of $B_1(b)$ and $B_2(b)$ following the outline presented above. In order to streamline the presentation, we present proofs of some of the results that are of auxiliary nature in Section 7.6.

Part 1) Some useful upper bounds

Recall our earlier definition $X := V - T$. As mentioned previously, the goal of this subsection is to provide some generic bounds which will be useful in deriving upper bounds for both $B_1(b)$ and $B_2(b)$.

Lemma 7.1. *Suppose that $\rho = 1$. Then there exist positive constants C_1 and C_0 such that for all $\mathbf{w} = (w_1, w_2)$ satisfying $0 \le w_1 \le w_2$,*

$$\mathbb{E}_{\mathbf{w}}\left[\tau_0\right] \le C_1 w_2 + C_0.$$

Remark 7.1. The conclusion of Lemma 7.1 holds true for every $\rho < 2$. Our proof for Lemma 7.1 can be easily modified to accommodate every $\rho < 2$.

Lemma 7.2. *For every $x \ge b$ and $\mathbf{w} = (w_1, w_2)$ with $0 \le w_1 \le w_2 < b$,*

$$\mathbb{P}_{\mathbf{w}}\left\{ W^{(2)}_{\tau_b^{(2)}} > x \;\middle|\; \tau_b^{(2)} < \tau_0 \right\} \le \mathbb{P}\left\{ X + b > x \mid X > b \right\}.$$

In other words, $W^{(2)}_{\tau_b^{(2)}}$ given $\tau_b^{(2)} < \tau_0$ is stochastically dominated by $X + b$ given $X > b$.

If $\rho < 2$, it is intuitive to expect the servers to effectively drain work whenever $W^{(2)}$ is large. Lemma 7.3, whose proof is given in Section 7.6, asserts the same when $\rho = 1$.

Lemma 7.3. *There exist positive constants C and ε such that*

$$\mathbb{E}_{(w_1,w_2)}\left[W_1^{(1)} + W_1^{(2)} \right] < (w_1 + w_2) - \varepsilon$$

as long as $w_2 \ge C$.

Lemma 7.1 follows as a corollary of Lemma 7.3 via a standard Lyapunov argument.

Proof of Lemma 7.1. Let $A := \{(w_1, w_2) : w_1 \le w_2 \le C\}$ and $T_A := \inf\{n \ge 1 : W_n \in A\}$. Additionally, let $V((w_1, w_2)) = (w_1 + w_2)/\varepsilon$ for $0 \le w_1 \le w_2$. Here C and ε are chosen as in Lemma 7.3. It follows from recursions (7.5a) and (7.5b) that

$$\sup_{(w_1,w_2)\in A} \mathbb{E}_{(w_1,w_2)}\left[V\left(W_1^{(1)}, W_2^{(2)} \right) \right] \le \frac{\mathbb{E}\left[(C + V - T)^+ + (C - T)^+ \right]}{\varepsilon} =: C_2 < \infty.$$

This observation, in conjunction with Lemma 7.3 and Theorem 11.3.4 of Meyn and Tweedie [1993], results in

$$\mathbb{E}_{(w_1,w_2)}[T_A] \leq \frac{w_1+w_2}{\epsilon} + C_2 \leq \frac{2}{\epsilon} w_2 + C_2 \tag{7.19}$$

for every $0 \leq w_1 \leq w_2$. Moreover, since $\inf_{\mathbf{w}\in A} \mathbb{P}_{\mathbf{w}}\{W_1^{(2)} = 0\} \geq \mathbb{P}\{T > C\} > 0$, it follows from a simple geometric trials argument that $\sup_{\mathbf{w}\in A} \mathbb{E}_{\mathbf{w}}[\tau_0] < \infty$. This observation, along with (7.19), proves the claim. □

Proof of Lemma 7.2. Note that

$$\mathbb{P}_{\mathbf{w}}\left\{ W^{(2)}_{\tau_b^{(2)}} > x, \tau_b^{(2)} < \tau_0 \right\} = \sum_{k=1}^{\infty} \mathbb{P}_{\mathbf{w}}\left\{ W_k^{(2)} > x, \tau_0 > k, \tau_b^{(2)} = k \right\}.$$

If $\tau_b^{(2)} = k$, it follows from recursion (7.5b) that $W_k^{(2)} = W_{k-1}^{(1)} + X_k$. Here recall that $X_k = V_{k-1} - T_k$. Therefore,

$$\begin{aligned}\mathbb{P}_{\mathbf{w}}\left\{ W^{(2)}_{\tau_b^{(2)}} > x, \tau_b^{(2)} < \tau_0 \right\} &= \sum_{k=1}^{\infty} \mathbb{P}_{\mathbf{w}}\left\{ X_k > x - W_{k-1}^{(1)}, \tau_0 > k-1, \tau_b^{(2)} > k-1 \right\} \\ &= \sum_{k=1}^{\infty} \mathbb{E}_{\mathbf{w}}\left[I\left(\tau_0 > k-1, \tau_b^{(2)} > k-1\right) \bar{F}\left(x - W_{k-1}^{(1)}\right)\right]\end{aligned} \tag{7.20}$$

Observe that $W_{k-1}^{(2)} < b$ whenever $\tau_b^{(2)} > k-1$. Additionally, since x is taken to be larger than b,

$$\frac{\bar{F}\left(x - W_{k-1}^{(1)}\right)}{\bar{F}\left(b - W_{k-1}^{(1)}\right)} \leq \frac{\bar{F}(x-b)}{\bar{F}(b)} \wedge 1 = \mathbb{P}\left\{X + b > x \mid X > b\right\}$$

on the set $\{\tau_b^{(2)} > k-1\}$. Therefore,

$$\begin{aligned}&\mathbb{P}_{\mathbf{w}}\left\{ W^{(2)}_{\tau_b^{(2)}} > x, \tau_b^{(2)} < \tau_0 \right\} \\ &\quad\leq \mathbb{P}\left\{X + b > x \mid X > b\right\} \times \sum_{k=1}^{\infty} \mathbb{E}_{\mathbf{w}}\left[I\left(\tau_0 > k-1, \tau_b^{(2)} > k-1\right) \bar{F}\left(b - W_{k-1}^{(1)}\right)\right] \\ &\quad= \mathbb{P}\left\{X + b > x \,|X > b\right\} \mathbb{P}_{\mathbf{w}}\left\{\tau_b^{(2)} < \tau_0\right\},\end{aligned}$$

where the last expression was obtained by letting $x = b$ in the second line in (7.20). The last inequality is equivalent to the statement of Lemma 7.2, and this concludes the proof. □

Part 2) Reduction to a zero-mean random walk problem

Recall our earlier definition $X_n := V_{n-1} - T_n$ for $n \geq 1$, where $(V_n : n \geq 1)$ are i.i.d. copies of V and $(T_n : n \geq 1)$ are i.i.d. copies of T. Additionally, we had set $V_0 := 0$. Further, define $S_0 := 0$, $S_n := X_1 + \ldots + X_n$, and

$$N_A(t) := \sup\{n \geq 0 : T_1 + \ldots + T_n \leq t\} \vee 0$$

for $t \geq 0$. Here we follow the usual convention that $\sup \emptyset = -\infty$. Therefore, $N_A(0) = 0$. Note that $N_A(t)$ is the number of customers that arrive in the time interval $(0, t]$. In addition to the above definitions, let $X := V - T$ and define

$$B_3(b) := \mathbb{E}\left[I\left(\max_{0 \leq n \leq N_A(X)+1} 2|S_n| > (\delta - \delta_-) b \right) \left(X + \max_{0 \leq n \leq N_A(X)+1} |S_n| \right) \;\middle|\; X > b\delta_+ \right].$$

Our objective in this subsection is to show the following result.

Lemma 7.4. *Suppose that Assumptions 7.1 and 7.2 hold, and that $\rho = 1$. Then,*

$$B_1(b) = O\left(\mathbb{P}_{\mathbf{0}}\left\{ \tau^{(2)}_{b\delta_+} < \tau_0 \right\} \times B_3(b) \right).$$

Let $\mathcal{F}_n$ denote the $\sigma-$algebra generated by the random variables V_k and $T_k, k \leq n$. Then

$$B_1(b) = \mathbb{E}_{\mathbf{0}}\left[I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta} \right) \mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W^{(1)}_k > b \right) \;\middle|\; \mathcal{F}_{\tau^{(1)}_{b\delta}} \right] \right].$$

Since $W^{(1)}_k$ is smaller than b for $k < \tau^{(1)}_{b\delta}$, on the set $\{\tau^{(1)}_b < \tau_0\}$, we have

$$\mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W^{(1)}_k > b \right) \;\middle|\; \mathcal{F}_{\tau^{(1)}_{b\delta}} \right] = \mathbb{E}_{\mathbf{W}_{\tau^{(1)}_{b\delta}}}\left[\sum_{k=0}^{\tau_0 - 1} I\left(W^{(1)}_k > b \right) \right] \leq \mathbb{E}_{\mathbf{W}_{\tau^{(1)}_{b\delta}}}[\tau_0].$$

Then, due to Lemma 7.1,

$$B_1(b) \leq \mathbb{E}_{\mathbf{0}}\left[I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}, \tau^{(1)}_b < \tau_0 \right) \left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0 \right) \right].$$

First, observe that whenever $\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}$, we must also have that $\tau^{(2)}_{b\delta_+} = \tau^{(2)}_{b\delta_-}$. Otherwise, from the definition of $\bar{\tau}^{(2)}_{b\delta_+}$, it follows that $\bar{\tau}^{(2)}_{b\delta_+} = \tau^{(2)}_{b\delta_-}$ which in turn occurs earlier than $\tau^{(1)}_{b\delta}$, and this contradicts our blanket assumption $\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}$. Therefore,

$$\begin{aligned} B_1(b) &\leq \mathbb{E}_{\mathbf{0}}\left[I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}, \tau^{(2)}_{b\delta_+} = \tau^{(2)}_{b\delta_-}, \tau^{(1)}_b < \tau_0 \right) \left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0 \right) \right] \\ &\leq \mathbb{E}_{\mathbf{0}}\left[I\left(\tau^{(2)}_{b\delta_+} \leq \tau^{(1)}_{b\delta_-} \wedge \tau_0 \right) \mathbb{E}_{\mathbf{0}}\left[\left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0 \right) I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta} \right) \;\middle|\; \mathcal{F}_{\tau^{(2)}_{b\delta_+}} \right] \right]. \end{aligned}$$

As a consequence of strong Markov property of $\mathbf{W}$, we have that

$$B_1(b) \leq \mathbb{E}_{\mathbf{0}}\left[I\left(\tau^{(2)}_{b\delta_+} \leq \tau^{(1)}_{b\delta_-} \wedge \tau_0\right)\mathbb{E}_{\mathbf{W}_{\tau^{(2)}_{b\delta_+}}}\left[\left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0\right) I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}\right)\right]\right]$$

If we set $\xi_1 := W^{(1)}_{\tau^{(2)}_{b\delta_+}}$ and $\xi_2 := W^{(2)}_{\tau^{(2)}_{b\delta_+}}$, again due to the Markov property of $\mathbf{W}$,

$$B_1(b) \leq \mathbb{E}\left[I(\xi_1 < b\delta_-)\mathbb{E}_{(\xi_1,\xi_2)}\left[\left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0\right) I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}\right)\right]\right]\mathbb{P}_{\mathbf{0}}\left\{\tau^{(2)}_{b\delta_+} < \tau_0\right\}, \quad (7.21)$$

where ξ_2, by definition, is larger than $b\delta_+$.

Evaluation of the inner expectation

We analyse the inner expectation

$$\chi(\xi_1,\xi_2) := \mathbb{E}_{(\xi_1,\xi_2)}\left[\left(C_1 W^{(2)}_{\tau^{(1)}_{b\delta}} + C_0\right) I\left(\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}\right)\right]$$

in (7.21) by restarting the queuing system with initial conditions $\mathbf{W}_0 = (\xi_1, \xi_2)$. Whenever $\bar{\tau}^{(2)}_{b\delta_+} > \tau^{(1)}_{b\delta}$, due to recursions (7.5a) and (7.5b), the dynamics of the queue until $\tau^{(1)}_{b\delta}$ is described by

$$W^{(1)}_n = \left(W^{(1)}_{n-1} + X_n\right)^+ \text{ and } W^{(2)}_n = W^{(2)}_{n-1} - T_n$$

for $1 \leq n < \tau^{(1)}_{b\delta}$, in conjunction with $W^{(1)}_0 = \xi_1 < b\delta_-$ and $W^{(2)}_0 = \xi_2 > b\delta_+$. As a result,

$$T_1 + \ldots + T_{\tau^{(1)}_{b\delta}-1} = \xi_2 - W^{(2)}_{\tau^{(1)}_{b\delta}-1} \leq \xi_2 - b\delta_+$$

whenever $\bar{\tau}^{(2)}_{b\delta^+} > \tau^{(1)}_{b\delta}$. Therefore, $\tau^{(1)}_{b\delta} \leq N_A(\xi_2 - b\delta^+) + 1$, which in turn implies that

$$\max_{0<n\leq N_A(\xi_2 - b\delta^+)+1} W^{(1)}_n > b\delta.$$

Consequently,

$$\chi(\xi_1,\xi_2) \leq \mathbb{E}_{(\xi_1,\xi_2)}\left[\left(C_1 \max_{0<n\leq N_A(\xi_2 - b\delta^+)+1} W^{(2)}_n + C_0\right) I\left(\max_{0<n\leq N_A(\xi_2 - b\delta^+)+1} W^{(1)}_n > b\delta\right)\right].$$

The following result is verified in Section 7.6.

Lemma 7.5. *Suppose that* $\mathbf{W}_0 = (w_1, w_2)$, *and recall the definitions* $X_n := V_{n-1} - T_n$, $S_0 := 0$ *and* $S_n := X_1 + \ldots + X_n$. *Then, for all* $n \geq 0$,

$$\max_{0<k\leq n} W^{(i)}_k \leq 2 \max_{0\leq k\leq n} |S_k| + w_i, \ i = 1, 2.$$

As a consequence of Lemma 7.5,

$$\chi(\xi_1,\xi_2) \le C_3\mathbb{E}\left[\left(\max_{0\le n\le N_A(\xi_2-b\delta^+)+1} 2S_n+\xi_2\right) I\left(\max_{0\le n\le N_A(\xi_2-b\delta^+)+1} 2S_n+\xi_1 > b\delta\right) \;\middle|\; \xi_1,\xi_2\right].$$

where the constant C_0 been absorbed in another suitable constant C_3. Then it is immediate from (7.21) that

$$B_1(b) \le C_3\mathbb{E}\left[\left(\max_{0\le n\le N_A(\xi_2-b\delta^+)+1} 2S_n+\xi_2\right) I\left(\max_{0\le n\le N_A(\xi_2-b\delta^+)+1} 2S_n > (\delta-\delta_-)\,b\right)\right]\mathbb{P}_{\mathbf{0}}\left\{\tau^{(2)}_{b\delta_+} < \tau_0\right\}.$$

Here, recall that $\xi_2 := W^{(2)}_{\tau^{(2)}_{b\delta_+}}$, which is stochastically dominated by the conditional distribution of $X + b\delta_+$ given that $X > b\delta_+$ (due to Lemma 7.2). Since $B_2(b)$ is a non-decreasing function of ξ_2, we use the above stochastic dominance to yield

$$\begin{aligned} B_1(b) \le{}& C_3\mathbb{P}_{\mathbf{0}}\left\{\tau^{(2)}_{b\delta_+} < \tau_0\right\} \times \\ & \mathbb{E}\left[\left(\max_{0\le n\le N_A(X)+1} 2S_n+X+b\delta_+\right) I\left(\max_{0\le n\le N_A(X)+1} 2S_n > (\delta-\delta_-)\,b\right) \;\middle|\; X > b\delta_+\right]. \end{aligned}$$

Lemma 7.4 follows from the above inequality once we observe that $X+b\delta_+ \le 2X$ when $X > b\delta_+$. This completes the proof of Lemma 7.4. □

Part 3.a) Simplifications using uniform large deviations: the $\alpha > 2$ case

Using classical results borrowed from the literature on large deviations for zero-mean random walks, we aim to prove the following result in this subsection.

Lemma 7.6. *Suppose that Assumptions 7.1 and 7.2 are in force, $\alpha > 2$ and $\rho = 1$. Then,*

$$B_3(b) = O\left(b^2\bar{B}(b) + b^2\frac{\bar{B}(b^2)}{\bar{B}(b)}\right).$$

We begin by recalling results on uniform large deviations for regularly varying random walks. For example, the following large deviations result which holds under Assumptions 7.1 and 7.2 assuming that $\alpha > 2$, is well-known

$$\mathbb{P}\{S_m > b\} = \left(\bar{\Phi}\left(\frac{b}{\sqrt{m}\sigma}\right) + m\mathbb{P}\{X_1 > b\}\right)(1+o(1)), \text{ as } m\to\infty, \tag{7.22}$$

uniformly for $b > \sqrt{m}$, where $\bar{\Phi}(\cdot)$ is the tail of a standard normal distribution. The asymptotic approximation (7.22) is due to A. V. Nagaev (see Theorem 1.9 of Nagaev [1979] or Corollary 7 of Rozovskii [1989]).

For our purposes, we need an extension of (7.22), in which S_n is replaced by $\max_{0\le k\le m}|S_k|$. However, we do not need exact asymptotic results as in (7.22), but only an asymptotic upper

bound. This is the content of the following result, which is proved in Section 7.6 as an immediate consequence of Corollary 1 of Pinelis [1981]. (For related uniform sample path large deviations results see Borovkov and Borovkov [2008], and the related Theorem 5 of Borovkov and Borovkov [2002].)

Lemma 7.7. *Suppose that V satisfies Assumption 7.1 with $\alpha > 2$, and T satisfies Assumption 7.2. Recall that $X_1, X_2, \ldots$ are i.i.d. copies of $X = V - T$. Then, there exists a positive integer m_0 such that for all $x \geq m^{1/2}$ and $m > m_0$*

$$\mathbb{P}\left\{\max_{0\leq k\leq m}|S_k| > x\right\} \leq 3\left(\mathbb{P}\left\{\max_{0\leq t\leq 1}\sigma\left|B\left(t\right)\right| > \frac{x}{m^{1/2}}\right\} + m\mathbb{P}\left\{|X| > x\right\}\right),$$

where $\sigma^2 = Var[X]$ and $B(\cdot)$ is a standard Brownian motion.

We establish Lemma 7.6 in two parts. The first task involves analysing the relatively easier term, which has the running maximum appearing only in the indicator function.

Lemma 7.8. *Under Assumption 7.1 with $\alpha > 2$, and Assumption 7.2,*

$$\mathbb{E}\left[I\left(\max_{0\leq n\leq N_A(X)+1} 2\left|S_n\right| > \left(\delta - \delta_-\right)b\right)X \;\middle|\; X > b\delta_+\right] = O\left(b^2\bar{B}\left(b\right) + b^2\frac{\bar{B}\left(b^2\right)}{\bar{B}(b)}\right).$$

Following this, we estimate the term in which the running maximum appears both multiplying and inside the indicator.

Lemma 7.9. *Under Assumption 7.1 with $\alpha > 2$, and Assumption 7.2,*

$$\mathbb{E}\left[I\left(\max_{0\leq n\leq N_A(X)+1} 2\left|S_n\right| > \left(\delta - \delta_-\right)b\right)\max_{0\leq n\leq N_A(X)+1}|S_n| \;\middle|\; X > b\delta_+\right] = O\left(b^2\bar{B}\left(b\right)\right).$$

Lemma 7.10, whose proof is given in Section 7.6, will be useful in proving Lemmas 7.8 and 7.9.

Lemma 7.10. *If $v(x) = x^{-\alpha}l(x)$ for some $\alpha > 2$ and a function $l(\cdot)$ slowly varying at infinity, then for every $c > 0$,*

$$\int_b^\infty v(t)\exp\left(-c\frac{b^2}{t}\right)dt = O\left(b^2 v\left(b^2\right)\right).$$

Proof of Lemma 7.8. Letting $c = (\delta - \delta_-)/2$, observe that

$$\begin{aligned}
&\mathbb{E}\left[I\left(\max_{0\le n\le N_A(X)} 2\left|S_n\right| > (\delta-\delta_-)\, b,\ N_A(X)+1\le 2X\right) X \;\middle|\; X > b\delta_+\right] \\
&\quad\le \int_{b\delta_+}^{\infty} t\mathbb{P}\left\{\max_{0\le n\le 2t}|S_n| > cb\right\}\frac{\mathbb{P}\{X\in dt\}}{\mathbb{P}\{X>b\delta_+\}} \\
&\quad\le 3\int_{b\delta_+}^{\infty} t\mathbb{P}\left\{\max_{0\le s\le 1}\sigma B(s) > \frac{cb}{\sqrt{2t}}\right\}\frac{\mathbb{P}\{X\in dt\}}{\mathbb{P}\{X>b\delta_+\}} + 3\int_{b\delta_+}^{\frac{c^2b^2}{2}} 2t^2\mathbb{P}\{|X|>cb\}\frac{\mathbb{P}\{X\in dt\}}{\mathbb{P}\{X>b\delta_+\}}
\end{aligned} \tag{7.23}$$

because of the application of the uniform asymptotic presented in Lemma 7.7 in the region $2t \le c^2b^2$ and Central Limit Theorem in the region $2t > c^2b^2$. Recall from (7.4) that $\mathbb{P}\{X > x\} \sim \bar{B}(x)$, and subsequently, due to Karamata's theorem (2.8), we obtain

$$\int_{b\delta_+}^{\infty} t^2\mathbb{P}\{X\in dt\} \le \mathbb{E}\left[X^2 I\left(X > b\delta_+\right)\right] = O\left(\int_{b\delta_+}^{\infty} s\mathbb{P}\{X>s\}\,ds\right) = O\left(b^2\bar{B}\left(b\right)\right)$$

and therefore

$$\frac{\mathbb{P}\left(|X|>cb\right)}{\mathbb{P}\{X>b\delta_+\}}\int_{b}^{\infty} t^2\mathbb{P}\{X\in dt\} = O\left(b^2\bar{B}\left(b\right)\right). \tag{7.24}$$

To deal with the first term in (7.23), we do integration by parts (by taking $u = \mathbb{P}\{\max_{0\le s\le 1} B(s) > cb/\sqrt{2\sigma t}\}$ and $v = \int_t^{\infty}\mathbb{P}\{X>u\}du - t\mathbb{P}\{X>t\}$) to obtain

$$\int_{b\delta_+}^{\infty} t\mathbb{P}\left\{\max_{0\le s\le 1}\sigma B(s) > \frac{cb}{\sqrt{2t}}\right\}\mathbb{P}\{X\in dt\} = O\left(b\int_{b\delta_+}^{\infty}\frac{\mathbb{P}\{X>t\}}{\sqrt{t}}\exp\left(-\frac{cb^2}{4\sigma t}\right)dt\right),$$

which, in turn, is $O(b\times b\bar{B}(b^2))$ because of Lemma 7.10. Therefore, due to (7.23) and (7.24), along with the observation that $\mathbb{P}\{X>b\delta_+\} = \Theta(\bar{B}(b))$ (due to regular variation), we obtain

$$\mathbb{E}\left[I\left(\max_{0\le n\le N_A(X)+1} 2\left|S_n\right| > (\delta-\delta_-)\, b, N_A(X)+1\le 2X\right) X \;\middle|\; X > b\delta_+\right] = O\left(b^2\bar{B}(b) + b^2\frac{\bar{B}(b^2)}{\bar{B}(b)}\right). \tag{7.25}$$

On the other hand, given that $N_A(t)/t \to 1$ as $t\to\infty$, the event $\{N_A(t) > 2t-1\}$ corresponds to a large deviations event with exponentially small probability for large values of t. Therefore, we have that

$$\begin{aligned}
&\mathbb{E}\left[I\left(\max_{0\le n\le N_A(X)} 2\left|S_n\right| > (\delta-\delta_-)\, b, N_A(X)+1> 2X\right) X \;\middle|\; X > b\delta_+\right] \\
&\qquad\le \int_{b\delta_+}^{\infty} t\mathbb{P}\left\{N_A\left(t\right) > 2t-1\right\}\mathbb{P}\left\{X\in dt\right\} = O\left(\exp\left(-\gamma b\right)\right),
\end{aligned}$$

for a suitable $\gamma > 0$. This observation, along with (7.25), concludes the proof of Lemma 7.8. □

The proof of Lemma 7.9, where running maximum appears twice, is similar, but more involved, and is presented in Section 7.6, so that we can continue with central arguments in the main body of the chapter. Before moving to Part 3.b) of the proof, it is important to note that Lemma 7.6 stands proved as an immediate consequence of Lemmas 7.8 and 7.9.

Part 3.b) Simplifications using uniform large deviations: the $\alpha \in (1,2)$ case

We shall leverage much of the reasoning behind Part 3.a) and prove the following result:

Lemma 7.11. *Suppose that* $\bar{B}(x) \sim cx^{-\alpha}$ *as* $x \to \infty$ *for some* $c > 0$ *and* $\alpha \in (1,2)$. *Also, suppose that Assumption 7.2 holds. Then,*

$$B_3(b) = O\left(\frac{b^\alpha \bar{B}(b^\alpha)}{\bar{B}(b)}\right).$$

We begin with a uniform convergence result which is a special case of Theorem 3.8.2 of Borovkov and Borovkov [2008]:

Lemma 7.12. *Suppose that* $\bar{B}(x) \sim cx^{-\alpha}$ *as* $x \to \infty$ *for some* $c > 0$ *and* $\alpha \in (1,2)$. *Also, suppose that Assumption 7.2 holds.Then, there exists a positive integer* m_0 *such that for all* $m \geq m_0$,

$$\mathbb{P}\left\{\max_{0\leq n\leq m} |S_n| > x\right\} \leq 3\mathbb{P}\left\{Z_* > \frac{x}{(cm)^{1/\alpha}}\right\},$$

where $Z_* := \max_{0\leq s\leq 1} Z(s)$ *is the maximum of an* α*-stable process* $(Z(t) : 0 \leq t \leq 1)$ *satisfying* $\mathbb{P}\{Z(1) > x\} \sim x^{-\alpha}$ *as* $x \to \infty$. *Additionally, for such a stable process* $Z(\cdot)$, *we have that*

$$\mathbb{P}\{Z_* > x\} \sim x^{-\alpha} \text{ as } x \to \infty.$$

The adaptation of Theorem 3.8.2 of Borovkov and Borovkov [2008] to the case where maximum of $|S_n|$ appears (instead of maximum of S_n) is similar to the argument in the proof of Lemma 7.7 in Part 3.a), and therefore is omitted. The dominant contribution to $B_3(b)$ is accounted for in the following result:

Lemma 7.13. *Suppose that* $\bar{B}(x) \sim cx^{-\alpha}$ *as* $x \to \infty$ *for some* $c > 0$ *and* $\alpha \in (1,2)$. *Also, suppose that Assumption 7.2 holds. Then,*

$$\mathbb{E}\left[I\left(\max_{0\leq n\leq N_A(X)+1} 2|S_n| > (\delta - \delta_-)b\right) X \;\middle|\; X > b\delta_+\right] = O\left(\frac{b^\alpha \bar{B}(b^\alpha)}{\bar{B}(b)}\right).$$

Proof of Lemma 7.13. As a consequence of Lemma 7.12, we get

$$\begin{aligned}\mathbb{E}\left[I\left(\max_{0\leq n\leq 2X} 2|S_n| > (\delta - \delta_-)b\right) XI(X > b\delta_+)\right] &\leq 3\mathbb{E}\left[\mathbb{P}\left\{Z_* > \frac{(\delta-\delta_-)b}{2(2cX)^{\frac{1}{\alpha}}}\right\} XI(X > b\delta_+)\right] \\ &= 3\mathbb{E}\left[XI\left(X > (b\delta_+) \vee \left(\frac{\bar{c}b}{Z_*}\right)^\alpha\right)\right] \quad (7.26)\end{aligned}$$

where $\bar{c} := (\delta - \delta_-)/(2^{\alpha+1}c)^{1/\alpha}$. Additionally, for all large enough x, there exists a constant C such that $\mathbb{E}[XI(X > x)] \leq Cx^{-(\alpha-1)}$, because of Karamata's theorem (2.8) and the observation that $\mathbb{P}\{X > x\} \sim cx^{-\alpha}$ as $x \to \infty$. Therefore, for all b large enough, we obtain

$$\mathbb{E}\left[XI\left(X > (b\delta_+) \vee \left(\frac{\bar{c}b}{Z_*}\right)^{\alpha}\right)\right] \leq C\mathbb{E}\left[\left((b\delta_+) \vee \left(\frac{\bar{c}b}{Z_*}\right)^{\alpha}\right)^{-(\alpha-1)}\right]$$
$$\leq C\left(b\delta_+\right)^{-(\alpha-1)}\mathbb{P}\left\{Z_* > \frac{\bar{c}b^{1-\frac{1}{\alpha}}}{\delta_+^{\frac{1}{\alpha}}}\right\} + C\left(\bar{c}b\right)^{-\alpha(\alpha-1)}\mathbb{E}\left[Z_*^{\alpha^2-\alpha}\right],$$

which, in turn, is $O(b^{\alpha}\bar{B}(b^{\alpha}))$ because $\mathbb{E}[Z_*^{\alpha^2-\alpha}] < \infty$ when $\alpha \in (1,2)$. Therefore, due to (7.26),

$$\mathbb{E}\left[I\left(\max_{0\leq n\leq 2X} 2\left|S_n\right| > (\delta - \delta_-)\, b,\ N_A(X) + 1 \leq 2X\right) XI\left(X > b\delta_+\right)\right] = O\left(b^{\alpha}\bar{B}\left(b^{\alpha}\right)\right).$$

On the other hand, the event $\{N_A(b) + 1 > 2b\}$ is a large deviations event with probabilities exponentially decaying in b, and as argued in the proof of Lemma 7.8,

$$\mathbb{E}\left[I\left(\max_{0\leq n\leq 2X} 2\left|S_n\right| > (\delta - \delta_-)\, b,\ N_A(X) + 1 > 2X\right) XI\left(X > b\delta_+\right)\right] = O\left(\exp(-\gamma b)\right),$$

for a suitable $\gamma > 0$. These two observations, after adjusting for the conditioning by dividing by $\mathbb{P}\{X > b\delta_+\} = \Theta(\bar{B}(b))$, prove Lemma 7.13. □

Lemma 7.14. *Suppose that $\bar{B}(x) \sim cx^{-\alpha}$ as $x \to \infty$ for some $c > 0$ and $\alpha \in (1,2)$. Also, suppose that Assumption 7.2 holds. Then,*

$$\mathbb{E}\left[I\left(\max_{0\leq n\leq N_A(X)+1} 2\left|S_n\right| > (\delta - \delta_-)\, b\right) \max_{0\leq n\leq N_A(X)+1}\left|S_n\right| \;\middle|\; X > b\delta_+\right] = O\left(b^2\bar{B}\left(b\right)\right).$$

As in Part 3.a), the proof of Lemma 7.14 is furnished in Section 7.6. The main result of this section, Lemma 7.11, which aims to prove that $B_3(b) = O(b^{\alpha}\bar{B}(b^{\alpha})/\bar{B}(b))$ is an immediate consequence of Lemmas 7.13 and 7.14, along with the observation that $b^2\bar{B}^2(b) = o(b^{\alpha}\bar{B}(b^{\alpha}))$ when $\alpha < 2$.

Part 4) Estimation of $B_2(b)$

The objective of this subsection is to prove Lemma 7.15, and subsequently, complete the proof of Proposition 7.2.

Lemma 7.15. *Suppose that Assumptions 7.1 and 7.2 hold, and that $\rho = 1$. Then,*

$$B_2\left(b\right) = O\left(b^2\bar{B}\left(b\right)^2\right).$$

It follows from the definition of $B_2(b)$ that

$$B_2(b) = \mathbb{E}_{\mathbf{0}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right) I\left(\bar{\tau}_{b\delta_+}^{(2)} \le \tau_{b\delta}^{(1)}, \tau_{b\delta_-}^{(2)} < \tau_0\right)\right]$$
$$= \mathbb{E}_{\mathbf{0}}\left[I\left(\bar{\tau}_{b\delta_+}^{(2)} \le \tau_{b\delta}^{(1)}, \tau_{b\delta_-}^{(2)} < \tau_0\right) \mathbb{E}_{\mathbf{0}}\left[\sum_{k=\bar{\tau}_{b\delta_+}^{(2)}}^{\tau_0-1} I\left(W_k^{(1)} > b\right) \;\middle|\; \mathcal{F}_{\bar{\tau}_{b\delta_+}^{(2)}}\right]\right]$$

Then due to the Markov property of $\mathbf{W}$, we get

$$B_2(b) = \mathbb{E}_{\mathbf{0}}\left[I\left(\bar{\tau}_{b\delta_+}^{(2)} \le \tau_{b\delta}^{(1)}, \tau_{b\delta_-}^{(2)} < \tau_0\right) \mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right]\right]. \tag{7.27}$$

Evaluation of inner expectation

Due to a similar conditioning with respect to $\mathcal{F}_{\tau_b^{(2)}}$, we obtain

$$\mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right] = \mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[I\left(\tau_b^{(2)} < \tau_0\right) \mathbb{E}_{\mathbf{W}_{\tau_b^{(2)}}}\left[\sum_{k=\tau_b^{(2)}}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right]\right]$$
$$\le \mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[I\left(\tau_b^{(2)} < \tau_0\right) \mathbb{E}_{\mathbf{W}_{\tau_b^{(2)}}}[\tau_0]\right],$$

which, due to Lemma 7.1, admits the following upper bound:

$$\mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right] \le \mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[I\left(\tau_b^{(2)} < \tau_0\right)\left(C_1 W_{\tau_b^{(2)}}^{(2)} + C_0\right)\right]$$
$$= C_1 \mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[W_{\tau_b^{(2)}}^{(2)} \;\middle|\; \tau_b^{(2)} < \tau_0\right] \mathbb{P}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left\{\tau_b^{(2)} < \tau_0\right\} + C_0 \mathbb{P}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left\{\tau_b^{(2)} < \tau_0\right\}$$
$$\le C_2 b \mathbb{P}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left\{\tau_b^{(2)} < \tau_0\right\} \tag{7.28}$$

for some positive constant C_2, because, due to Lemma 7.2,

$$\mathbb{E}_{\mathbf{W}_{\bar{\tau}_{b\delta_+}^{(2)}}}\left[W_{\tau_b^{(2)}}^{(2)} \;\middle|\; \tau_b^{(2)} < \tau_0\right] \le \mathbb{E}\left[X + b \;\middle|\; X > b\right] = O(b).$$

Next, we obtain a bound for $\mathbb{P}_{\mathbf{w}}\{\tau_b^{(2)} < \tau_0\}$, and use it in (7.28).

Lemma 7.16. *Suppose that Assumptions 7.1 and 7.2 hold, and that $\rho = 1$. Let $\delta_+ < 1/2$, then there exists a constant $C > 0$ such that for all $\mathbf{w} = (w_1, w_2)$ satisfying $w_1 \le w_2 < b\delta_+$ we have*

$$\mathbb{P}_{\mathbf{w}}\left\{\tau_b^{(2)} < \tau_0\right\} \le C\,(w_2 + 1)\,\bar{B}\,(b)\,.$$

The proof of Lemma 7.16 is instructive, as it employs Lyapunov bound techniques to derive the above uniform bound. The arguments involved are different from the rest of the chapter, and the whole of Section 7.4 is dedicated to expose the techniques clearly. Now, we aim to complete the proof of Lemma 7.15. Due to Lemma 7.16 and (7.28),

$$\mathbb{E}_{\mathbf{W}_{\bar{\tau}^{(2)}_{b\delta_+}}}\left[\sum_{k=0}^{\tau_0-1} I\left(W_k^{(1)} > b\right)\right] \leq CC_2 b\left(W^{(2)}_{\bar{\tau}^{(2)}_{b\delta_+}} + 1\right)\bar{B}(b) \leq CC_2 b\,(b\delta_+ + 1)\,\bar{B}(b),$$

and therefore, due to (7.27) and a similar application of Lemma 7.16 with $\mathbf{w} = \mathbf{0}$,

$$B_2(b) \leq CC_2\mathbb{P}_{\mathbf{0}}\left\{\tau^{(2)}_{b\delta_-} < \tau_0\right\} b\,(b\delta_+ + 1)\,\bar{B}(b) = O\left(b^2\bar{B}^2(b)\right).$$

This concludes the proof of Lemma 7.15.

Proof of Proposition 7.2. We apply Lemma 7.16 with $\mathbf{w} = \mathbf{0}$ to the bound for $B_1(b)$ in Lemma 7.4 as well. This, due to Lemmas 7.6 and 7.11, results in

$$B_1(b) = O\left(\bar{B}\left(b\delta_+\right) \times B_3(b)\right) = \begin{cases} O\left(b^2\bar{B}^2(b) + b^2\bar{B}\left(b^2\right)\right) & \text{if } \alpha > 2, \\ O\left(b^\alpha\bar{B}\left(b^\alpha\right)\right) & \text{if } \alpha \in (1,2). \end{cases}$$

Additionally, we have that $B_2(b) = O(b^2\bar{B}(b)^2)$ and $\mathbb{E}_{\mathbf{0}}[\tau_0] < \infty$, respectively, from Lemmas 7.15 and 7.1. Therefore, from (7.18), we arrive at the statement of Proposition 7.2, and this concludes the proof. □

7.4 Lyapunov bound techniques for a uniform bound on $\mathbb{P}_{\mathbf{w}}\{\tau_b^{(2)} < \tau_0\}$

We use the Lyapunov bound technique that has been employed in Blanchet and Glynn [2008b], Blanchet et al. [2007b], and Denisov et al. [2013]. The strategy is to define a Markov kernel $Q_\theta(\mathbf{w}, \cdot)$ (indexed by some parameter θ) and a non-negative function $H_b(w_1, w_2)$ satisfying the following conditions:

(L1) For every $\mathbf{w} = (w_1, w_2)$ such that $w_2 < b$,

$$\mathbb{E}^\theta_{\mathbf{w}}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\, H_b(\mathbf{W}_1)\right] \leq H_b(\mathbf{w}),$$

where $\mathbb{P}_{\mathbf{w}}\{\mathbf{W}_1 \in \cdot\}$ is the nominal transition kernel induced by recursions (7.5a) and (7.5b), $r_\theta(\mathbf{w}, \mathbf{x}) := \mathbb{P}_{\mathbf{w}}\{\mathbf{W}_1 \in d\mathbf{x}\}/Q_\theta(\mathbf{w}, d\mathbf{x})$ is the corresponding Radon-Nikodym derivative with respect to $Q_\theta(\mathbf{w}, \cdot)$, and $\mathbb{E}^\theta_{\mathbf{w}}[\cdot]$ is the expectation associated with the probability measure in path space for the Markov evolution induced by $Q_\theta(\mathbf{w}, \cdot)$.

(L2) Whenever $\mathbf{w} = (w_1, w_2)$ is such that $w_2 > b$, $H_b(w_1, w_2) \geq 1$.

If conditions (L1) and (L2) are satisfied, then following the analysis in Part (iii) of Theorem 2 of Blanchet and Glynn [2008b], we have that

$$\mathbb{P}_{\mathbf{w}}\left\{\tau_b^{(2)} < \tau_0\right\} \leq \mathbb{E}_{\mathbf{w}}^{\theta}\left[\prod_{n=0}^{\tau_b^{(2)}-1} r_\theta(\mathbf{W}_n, \mathbf{W}_{n+1}) H_b\left(\mathbf{W}_{\tau_b^{(2)}}\right) I\left(\tau_b^{(2)} < \tau_0\right)\right] \leq H_b(w_1, w_2). \tag{7.29}$$

The construction of $Q_\theta(\mathbf{w}, \cdot)$ and $H_b(\cdot)$ follows the intuition explained in Blanchet and Glynn [2008b] and Blanchet et al. [2007b]: We wish to select $Q_\theta(\mathbf{w}, \cdot)$ as closely as possible to the conditional distribution of the process $\{\mathbf{W}_n : n \geq 0\}$ given that $\{\tau_b^{(2)} < \tau_0\}$, because in that case, it happens that (7.29) is automatically satisfied with equality. Additionally, we shall find a suitable non-negative function $G_b(\cdot)$ so that $H_b(w_1, w_2) = G_b(w_1 + w_2)$ satisfies the Lyapunov inequality (L1).

For ease of notation, let us write

$$l := w_1 + w_2, \; L := W_1^{(1)} + W_2^{(2)} \text{ and } \Delta := L - l.$$

In order to construct $Q_\theta(\mathbf{w}, \cdot)$ and $G_b(\cdot)$, first define the Markov transition kernel

$$\begin{aligned} Q'(\mathbf{w}, A) = {} & \mathbb{P}_{\mathbf{w}}\{\mathbf{W}_1 \in A \mid X_1 > a(b-l)\} p(\mathbf{w}) \\ & + \mathbb{P}_{\mathbf{w}}\left\{\mathbf{W}_1 \in A \mid X_1 \leq a(b-l), W_1^{(2)} > 0\right\}(1 - p(\mathbf{w})), \end{aligned}$$

where $p(\mathbf{w})$ will be specified momentarily, and the choice $a \in (0,1)$ is arbitrary. On the set $\{\tau_b^{(2)} < \tau_0\}$, given $\mathbf{w} = (w_1, w_2)$ with $w_1 \leq w_2 < b$, we have that the nominal kernel $\mathbb{P}_{\mathbf{w}}\{\mathbf{W}_1 \in \cdot\}$ is absolutely continuous with respect to $Q'(\mathbf{w}, \cdot)$. Now, for $z \geq 0$, define

$$h_b(z) = \int_0^{z+\kappa_0} \mathbb{P}\{X > b - z + t\}\, dt = \int_{b-z}^{b+\kappa_0} \mathbb{P}\{X > u\}\, du.$$

Next, write

$$G_b(l) = \min(\kappa_1 h_b(l), 1)$$

and set

$$p(\mathbf{w}) = \frac{\mathbb{P}\{X > a(b-l)\}}{\kappa_2 h_b(l)}$$

where κ_2 is a number larger than

$$\sup_{x>0} \frac{\mathbb{P}\{X > ax\}}{\int_x^{x+l+\kappa_0} \mathbb{P}\{X > u\}\, du} < \infty.$$

Finally, define $\theta = (\kappa_0, \kappa_1, \kappa_2)$ and write

$$Q_\theta(\mathbf{w}, \cdot) = Q'(\mathbf{w}, \cdot)\, I(G_b(l) < 1) + K(\mathbf{w}, \cdot)\, I(G_b(l) = 1).$$

Recall the notation $l = w_1 + w_2$ and $L = W_1^{(1)} + W_1^{(2)}$. Condition (L1) is verified via the following proposition:

Proposition 7.3. *For every $\mathbf{w} = (w_1, w_2)$ such that $w_2 < b$, we have that*

$$\mathbb{E}_{\mathbf{w}}^{\theta}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\, G_b(L)\right] \leq G_b(l).$$

For proving Proposition 7.3, we consider only the case $G_b(l) < 1$. When $G_b(l) = 1$, the inequality is satisfied trivially. The following results are crucial in the proof of Proposition 7.3.

Lemma 7.17. *There exist positive constants μ and C such that*

$$\mathbb{E}_{(w_1,w_2)}\left[\Delta I\left(W_1^{(2)} > 0\right)\right] < -\mu$$

whenever $w_2 > C$.

Proof. First, observe that

$$\mathbb{E}_{\mathbf{w}}\left[\Delta I\left(W_1^{(2)} > 0\right)\right] = \mathbb{E}_{\mathbf{w}}[\Delta] - \mathbb{E}_{\mathbf{w}}\left[\Delta I\left(W_1^{(2)} = 0\right)\right].$$

Additionally, note that $\Delta = -(w_1 + w_2)$ when $W_1^{(2)} = 0$. Therefore,

$$\mathbb{E}_{(w_1,w_2)}\left[\Delta I\left(W_1^{(2)} = 0\right)\right] = -(w_1 + w_2)\,\mathbb{P}\{w_1 + V - T \leq 0,\ w_2 - T \leq 0\}.$$

Therefore, due to Lemma 7.3,

$$\begin{aligned}\mathbb{E}_{\mathbf{w}}\left[\Delta I\left(W_1^{(2)} > 0\right)\right] &\leq \mathbb{E}_{\mathbf{w}}[\Delta] + (w_1 + w_2)\,\mathbb{P}\{w_2 - T \leq 0\} \\ &\leq -\epsilon + 2w_2\mathbb{P}\{T > w_2\},\end{aligned}$$

where $w_2\mathbb{P}\{T > w_2\}$ can be made arbitrarily small by choosing $C > w_2$ large enough. Hence the claim stands verified. □

Lemma 7.18. *Recall that $l = w_1 + w_2$. The following holds as $(b - l) \to \infty$:*

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}}^{\theta}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\,\frac{G_b(L)}{G_b(l)}\, I(X_1 \leq a(b - l))\right] \\ &\quad\leq \mathbb{P}_{\mathbf{w}}\left\{W_1^{(2)} > 0\right\} + \frac{\mathbb{P}\{X > b - l\}}{h_b(l)}\,\mathbb{E}_{\mathbf{w}}\left[\Delta I\left(W_1^{(2)} > 0\right)\right](1 + o(1)).\end{aligned}$$

Proof. Since

$$G_b(L_1) = G_b(l) + \int_0^1 G_b'(l + u\Delta(1))\,\Delta(1)\,du,$$

we introduce a uniform random variable U, independent of everything else, to write

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}}^{\theta}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)} I\left(X_1 \le a(b-l)\right)\right] \\
&\quad = \mathbb{E}_{\mathbf{w}}\left[\frac{G_b(L)}{G_b(l)} I\left(X_1 \le a(b-l), W_1^{(2)} > 0\right)\right] \\
&\quad = \mathbb{E}_{\mathbf{w}}\left[\left(1 + \frac{G_b'(l+U\Delta)}{G_b(l)}\Delta\right) I\left(X_1 \le a(b-l), W_1^{(2)} > 0\right)\right] \\
&\quad \le \mathbb{P}_{\mathbf{w}}\left\{W_1^{(2)} > 0\right\} + \frac{\mathbb{P}\{X > b-l\}}{h_b(l)}\mathbb{E}_{\mathbf{w}}\left[\frac{\mathbb{P}\{X > b-l-U\Delta\}}{\mathbb{P}\{X > b-l\}}\Delta I\left(X_1 \le a(b-l), W_1^{(2)} > 0\right)\right].
\end{aligned} \tag{7.30}$$

We have used $G_b'(l + U\Delta(1)) = \kappa_1 \mathbb{P}\{X > b - l - U\Delta(1)\}$ to write the last step. Additionally, whenever $X_1 \le a(b-l)$, observe that

$$\Delta = (w_1 + X_1)^+ - w_1 + (w_2 - T_1)^+ - w_2 \le X_1^+ \le a(b-l),$$

and therefore, $\mathbb{P}\{X > b-l-U\Delta\} \le \mathbb{P}\{X > (1-a)(b-l)\} \le m_{1-a}\mathbb{P}\{X > b-l\}$, where

$$m_t := \sup_{x>0} \frac{\mathbb{P}\{X > tx\}}{\mathbb{P}\{X > x\}} < \infty,$$

for every $t > 0$. Here, the finiteness of m_t follows from the regularly varying nature of the tail distribution of X (recall that $\mathbb{P}\{X > x\} \sim \bar{B}(x)$ as $x \to \infty$). As a result, we have the following uniform bound for various values of b and l :

$$\mathbb{E}_{\mathbf{w}}\left[\frac{\mathbb{P}\{X > b-l-U\Delta\}}{\mathbb{P}\{X > b-l\}}\Delta I\left(X_1 \le a(b-l), W_1^{(2)} > 0\right)\right] \le m_{1-a}\mathbb{E}X^+. \tag{7.31}$$

Consequently, due to dominated convergence theorem, we obtain that

$$\mathbb{E}_{\mathbf{w}}\left[\frac{\mathbb{P}\{X > b-l-U\Delta\}}{\mathbb{P}\{X > b-l\}}\Delta I\left(X_1 \le a(b-l), W_1^{(2)} > 0\right)\right] \sim \mathbb{E}_{\mathbf{w}}\left[\Delta I\left(W_1^{(2)} > 0\right)\right],$$

as $(b-l) \to \infty$. Now, the statement of Lemma 7.18 is immediate from (7.30) and the above stated convergence. □

Proof of Proposition 7.3. As mentioned before, we consider $G_b(l) < 1$. First, observe that

$$\begin{aligned}
\mathbb{E}_{\mathbf{w}}^{\theta}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)} I\left(X_1 > a(b-l)\right)\right] &= \mathbb{E}_{\mathbf{w}}\left[\frac{G_b(L)}{G_b(l)} I\left(X_1 > a(b-l)\right)\right] \\
&\le \frac{\mathbb{P}\{X > a(b-l)\}}{\kappa_1 h_b(l)}
\end{aligned} \tag{7.32}$$

because $G_b(\cdot) \leq 1$. For a respective bound on the complementary event $\{X_1 \leq a(b-l)\}$, it is easy to see that our strategy must use Lemmas 7.17 and 7.18 in the following way: Given $\delta > 0$, there exists a constant C_δ large enough such that for all initial conditions $\mathbf{w} = (w_1, w_2)$ satisfying $w_2 > C$ and $b - l > C_\delta$,

$$\mathbb{E}_{\mathbf{w}}^{\theta}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)} I\left(X_1 \leq a(b-l)\right)\right] \leq \mathbb{P}_{\mathbf{w}}\left\{W_1^{(2)} > 0\right\} - (1-\delta)\mu\frac{\mathbb{P}\{X > b-l\}}{h_b(l)}.$$

Combining this bound with (7.32), we obtain

$$\begin{aligned}\mathbb{E}_{\mathbf{w}}^{\theta}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)}\right] &\leq \mathbb{P}_{\mathbf{w}}\left\{W_1^{(2)} > 0\right\} + \frac{\mathbb{P}\{X > a(b-l)\}}{h_b(l)}\left(\frac{1}{\kappa_1} - \frac{(1-\delta)\mu}{m_a}\right) \\ &\leq 1 + \kappa_2 p(\mathbf{w})\left(\frac{1}{\kappa_1} - \frac{(1-\delta)\mu}{m_a}\right)\end{aligned}$$

which is, in turn, smaller than 1 for κ_1 suitably large. In addition to this, in the region $\{(w_1, w_2) : w_2 > C, b - l < C_\delta\}$, we simply let $G_b(l) = 1$ by again choosing κ_1 large enough. This flexibility in the choice of κ_1 yields us

$$\mathbb{E}_{\mathbf{w}}^{\theta}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)}\right] \leq 1 \tag{7.33}$$

for initial conditions $\mathbf{w} = (w_1, w_2)$ satisfying $w_2 > C$. Now, turning our attention to the values of $\mathbf{w}$ such that $w_2 \leq C$, we see that $l = w_1 + w_2 \leq 2C$, and as a consequence of the regularly varying nature of the tail of X, we obtain

$$\frac{\mathbb{P}\{X > b-l\}}{h_b(l)} = \left(\int_0^{l+\kappa_0} \frac{\mathbb{P}\{X > b-l+u\}}{\mathbb{P}\{X > b-l\}} du\right)^{-1} = \frac{1+o(1)}{l+\kappa_0}$$

as $b \to \infty$. Then, it is immediate from (7.30) that whenever $w_2 \leq C$,

$$\mathbb{E}_{\mathbf{w}}^{\theta}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)} I\left(X_1 \leq a(b-l)\right)\right] \leq \mathbb{P}_{(C,C)}\left\{W_1^{(2)} > 0\right\} + \frac{m_{1-a}\mathbb{E}\left[X_1^+\right]}{\kappa_0}(1+o(1)).$$

Combining this bound with the one obtained in (7.32), we get

$$\mathbb{E}_{\mathbf{w}}^{\theta}\left[r_\theta(\mathbf{w}, \mathbf{W}_1)\frac{G_b(L)}{G_b(l)}\right] \leq \frac{p(\mathbf{w})\kappa_2}{\kappa_1} + \mathbb{P}_{(C,C)}\left\{W_1^{(2)} > 0\right\} + \frac{m_{1-a}\mathbb{E}\left[X_1^+\right]}{\kappa_0}(1+o(1)),$$

which can also be made smaller than 1 by picking κ_0 and κ_1 large enough. Thus, for all initial conditions $\mathbf{w}$, we have a consistent choice of parameters $(\kappa_0, \kappa_1, \kappa_2)$ that satisfies (L1). □

Since $G_b(l) = 1$ whenever $w_1 + w_2 \geq b - C_\delta$, we also have $G_b(l) = 1$ if $w_2 > b$. This verifies condition (L2). Since both (L1) and (L2) are satisfied, it follows from (7.29) that if $w_2 < b\delta_+$ for some $\delta_+ < 1/2$, then

$$\begin{aligned}\mathbb{P}_{\mathbf{w}}\left\{\tau_b^{(2)} < \tau_0\right\} &\leq \kappa_1 h_b(l) = \kappa_1 \int_{b-l}^{b+\kappa_0} \mathbb{P}\{X > u\} du \\ &\leq \kappa_1 \mathbb{P}\{X > b-l\}(\kappa_0 + l) \leq \kappa_1 \bar{B}(b(1-2\delta_+))(\kappa_0 + 2w_2)(1+o(1)).\end{aligned}$$

The right hand side of the previous inequality is equivalent to the statement of Lemma 7.16, so we conclude the proof.

7.5 Another proof of the upper bound

In this section, we provide an alternate proof for Proposition 7.2. For simplicity, we consider only the case $\alpha > 2$ here. Let

$$\bar{B}_I(x) := \int_x^\infty \bar{B}(u)du$$

denote the integrated tail of $\bar{B}(\cdot)$. First, we obtain an upper bound for the tail probabilities of $W_\infty^{(2)}$ based on a simple coupling argument.

Lemma 7.19.

$$\mathbb{P}\left\{W_\infty^{(2)} > b\right\} \leq \frac{2}{\mu}\bar{B}_I(b)\left(1+o(1)\right), \text{ as } b \to \infty.$$

Proof. Consider a two server queuing system with cyclic service discipline. As before, the first job is assumed to arrive at time 0. We subject this modified system to the interarrival and service time sequences $(T_n : n \geq 1)$ and $(V_n : n \geq 0)$, same as that of the original system. Due to the optimality of FCFS discipline (see Theorem 1 of Foss [1980]), the workload of queues in the modified system $\tilde{W}_n = \left(\tilde{W}_{n,1}, \tilde{W}_{n,2}\right)$ satisfies

$$W_n^{(2)} \leq_D \max\{\tilde{W}_{n,1}, \tilde{W}_{n,2}\},$$

where $\leq_D$ denotes stochastic dominance. Therefore by union bound,

$$\mathbb{P}\left\{W_\infty^{(2)} > b\right\} \leq \lim_n \left(\mathbb{P}\left\{\tilde{W}_{n,1} > b\right\} + \mathbb{P}\left\{\tilde{W}_{n,2} > b\right\}\right) = 2\mathbb{P}\left\{\tilde{W}_{\infty,1} > b\right\}.$$

Since one of the two servers (here the first server) in the modified system is marginally a $GI/GI/1$ system with service time sequence $(V_{2n} : n \geq 0)$ and interarrival sequence $(T_{2n-1} + T_{2n} : n \geq 1)$, it is well-known that the distribution of stationary waiting time $\tilde{W}_{\infty,1}$ is same as that of the supremum of a random walk with increment distribution $V_0 - (T_1 + T_2)$ and

$$\mathbb{P}\left\{\tilde{W}_{\infty,1} > b\right\} \sim \frac{1}{\mu}\bar{B}_I(b), \text{ as } b \to \infty$$

(see, for example, Asmussen [2003b]). Therefore,

$$\mathbb{P}\left\{W_\infty^{(2)} > b\right\} \leq \frac{2}{\mu}\bar{B}_I(b)\left(1+o(1)\right), \text{ as } b \to \infty. \qquad \square$$

Next, we divide the quantity of interest into two parts as below:

$$\mathbb{P}\left\{W_n^{(1)} > b\right\} = \mathbb{P}\left\{W_n^{(1)} > b, A_n\right\} + \mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n\right\},$$

where

$$A_n := \bigcup_{k=1}^{n-1} A_{nk}$$

is the union of

$$A_{nk} := \{V_{n-k} > \theta b + \beta k\}, \quad 1 \le k < n,$$

and $\bar{A}_n$ is the complement of A_n. The events A_n and $\bar{A}_n$ are considered separately in the following subsections.

Bounds on $\mathbb{P}\{W_n^{(1)} > b, A_n\}$

We initially obtain upper bounds on $\mathbb{P}\{W_n^{(1)} > b, A_{nk}\}$, and subsequently use them to derive bounds for $\mathbb{P}\{W_n^{(1)} > b, A_n\}$. Let

$$S_0 = 0, \text{ and } S_j = S_{j-1} + (V_{n-j} - T_{n-j+1}) \text{ for } 1 \le j \le n. \tag{7.34}$$

Lemma 7.20. *For any $1 \le k \le n$, we have:*

$$W_n^{(1)} \le W_{n-k}^{(2)} + \max_{j<k} S_j.$$

Proof. It follows from the recursion (7.5a) that

$$\begin{aligned} W_{n-j+1}^{(1)} &\le \left(W_{n-j}^{(1)} + V_{n-j} - T_{n-j+1}\right)^+ \\ &= \left(W_{n-j-1}^{(1)} + S_j - S_{j-1}\right)^+. \end{aligned}$$

We obtain the following by repeatedly expanding the above recursive inequality:

$$\begin{aligned} W_n^{(1)} &\le \left(W_{n-1}^{(1)} + S_1 - S_0\right)^+ \\ &\le \max\{0, \max\{0, W_{n-2}^{(1)} + S_2 - S_1\} + S_1\} \\ &\le \max\{0, S_1, W_{n-2}^1 + S_2\} \\ &\quad \vdots \\ &\le \max\{0, S_1, S_2, \ldots, W_{n-k+1}^{(1)} + S_{k-1}\} \\ &\le W_{n-k+1}^{(1)} + \max\{0, S_1, S_2, \ldots, S_{k-1}\} \\ &\le W_{n-k}^{(2)} + \max_{j<k} S_j, \end{aligned}$$

and this verifies the claim. □

Following Lemma 7.20, for any $k \leq n$ and $\delta > 0$, we have that

$$\begin{aligned}\mathbb{P}\left\{W_n^{(1)} > b, A_{nk}\right\} &\leq \mathbb{P}\left\{W_{n-k}^{(2)} + \max_{j<k} S_j > b, A_{nk}\right\} \\ &\leq \mathbb{P}\left\{\max_{j<k} S_j > (1-\delta)b, A_{nk}\right\} + \mathbb{P}\left\{W_{n-k}^{(2)} > \delta b, A_{nk}\right\}.\end{aligned}$$

Note that $W_{n-k}^{(2)}$ is independent of V_{n-k}; similarly $\max_{j<k} S_j$, which involves only the random variables $(V_i, T_{i+1} : n-k < i < n)$ is also independent of V_{n-k}, and hence independent of the event A_{nk}. Therefore,

$$\begin{aligned}\mathbb{P}\left\{W_n^{(1)} > b, A_{nk}\right\} &\leq \mathbb{P}\left\{\max_{j<k} S_j > (1-\delta)b\right\}\mathbb{P}(A_{nk}) + \mathbb{P}\left\{W_{n-k}^{(2)} > \delta b\right\}\mathbb{P}(A_{nk}) \\ &= \bar{B}(\theta b + \beta k)\left(\mathbb{P}\left\{\max_{j<k} S_j > (1-\delta)b\right\} + \mathbb{P}\left\{W_{n-k}^{(2)} > \delta b\right\}\right).\end{aligned}$$

Then for any fixed k,

$$\lim_{n\to\infty} \mathbb{P}\left\{W_n^{(1)} > b, A_{nk}\right\} \leq \bar{B}(\theta b + \beta k)\left(\mathbb{P}\left\{\max_{j<k} S_j > (1-\delta)b\right\} + \mathbb{P}\left\{W_{\infty}^{(2)} > \delta b\right\}\right)$$

Given $\epsilon > 0$, now it follows from Lemma 7.19 that

$$\lim_{n\to\infty} \mathbb{P}\left\{W_n^{(1)} > b, A_{nk}\right\} \leq \bar{B}(\theta b + \beta k)\left(\mathbb{P}\left\{\max_{j<k} S_j > (1-\delta)b\right\} + (1+\epsilon)\frac{2}{\mu}\bar{B}_I(\delta b)\right), \quad (7.35)$$

for all b large enough. Here the last inequality follows from Lemma 7.19.

Recall that the increments in the sum S are independent and identical (in distribution) to $V - T$. Since the service times and interarrival times have the same mean, the random walk $(S_j : j \leq k)$ has zero drift. Let σ^2 denote the variance of $X := V - T$, which is finite because of the finiteness of variance of V and T. Then, due to Lemma 7.7, it follows that

$$\begin{aligned}&\lim_{n\to\infty} \mathbb{P}\left\{W_n^{(1)} > b, A_{nk}\right\} \quad (7.36)\\ &\leq 3\bar{B}(\theta b + \beta k)\left((k-1)\mathbb{P}\left\{X > (1-\delta)b\right\} + \mathbb{P}\left\{\max_{0\leq t\leq 1} \sigma|B(t)| > \frac{(1-\delta)b}{\sqrt{k-1}}\right\} + \frac{1}{\mu}\bar{B}_I(\delta b)\right),\end{aligned}$$

for all $k < b^2$. On the other hand, if $k < b^2$, then it is immediate due to central limit theorem that

$$\lim_{n\to\infty} \mathbb{P}\left\{W_n^{(1)} > b, A_{nk}\right\} \leq 3\bar{B}(\theta b + \beta k)\left(\mathbb{P}\left\{\max_{0\leq t\leq 1} \sigma|B(t)| > \frac{(1-\delta)b}{\sqrt{k-1}}\right\} + \frac{1}{\mu}\bar{B}_I(\delta b)\right), \quad (7.37)$$

for $k \geq b^2$.

Next, because of union bound and monotone convergence, we have from (7.36) and (7.37) that

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left\{W_n^{(1)} > b, A_n\right\} &\leq \sum_{k\geq 1} \lim_{n\to\infty} \mathbb{P}\left\{W_n^{(1)} > b, A_{nk}\right\} \\ &\leq 3\sum_{k\geq 1} \bar{B}(\theta b + \beta k)\left((k-1)\bar{B}\left((1-\delta)b\right) + 4\bar{\Phi}\left(\frac{(1-\delta)b}{\sigma\sqrt{k-1}}\right) + \frac{1}{\mu}\bar{B}_I(\delta b)\right). \quad (7.38)\end{aligned}$$

For an infinite series involving regularly varying terms, we can proceed by approximating sums by integrals, and applying Karamata's theorem (see (2.8)) as below: for example,

$$\begin{aligned}\sum_{k=1}^{\infty} k\bar{B}(\theta b+\beta k) &\leq \sum_{k=1}^{\infty}\int_{k}^{k+1} x\bar{B}(\theta b+\beta(x-1))dx\\ &\leq \frac{1}{\beta^2}\int_{\theta b}^{\infty}(u-\theta b+\beta)\bar{B}(u)du\\ &= O\left(b^2\bar{B}(b)\right).\end{aligned}$$

Similarly by approximating other sum terms by integrals, changing variables and applying Karamata's theorem, we get:

$$\sum_{k=1}^{\infty}\bar{B}(\theta b+\beta k) = O\left(b\bar{B}(b)\right)$$

$$\sum_{k=1}^{\infty}\bar{B}(\theta b+\beta k)\bar{\Phi}\left(\frac{(1-\delta)b}{\sigma\sqrt{k-1}}\right) = O\left(\int_0^{\infty}\bar{B}(b+u)\bar{\Phi}\left(\frac{b}{\sqrt{u}}\right)du\right).$$

Then, because of the regularly varying nature of $\bar{B}(\cdot)$, it follows from (7.38) that

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}\left\{W_n^{(1)}>b, A_n\right\} &= O\left(b^2\bar{B}^2(b)+\int_0^{\infty}\bar{B}(b+u)\bar{\Phi}\left(\frac{b}{\sqrt{u}}\right)du + b\bar{B}(b)\bar{B}_I(b)\right)\\ &= O\left(b^2\bar{B}^2(b)+b^2\bar{B}(b^2)\right). \qquad (7.39)\end{aligned}$$

because of Lemma 7.10.

Bounds on $\mathbb{P}\{W_n^{(1)} > b, \bar{A}_n\}$

The aim of this subsection will be to prove the following result:

$$\lim_{n\to\infty}\mathbb{P}\left\{W_n^{(1)}>b,\bar{A}_n\right\} = o\left(b^2\bar{B}^2(b)\right).$$

For obtaining an upper bound when none of the service times are "big", the strategy is to identify the event of large waiting time with the event of supremum of a suitable negative drift random walk becoming large when none of its increments are big enough. In order to accomplish this, consider the following collection of arrival indices:

$$\mathcal{I} := \left\{i\geq 1 : W_i^{(2)} > W_{i-1}^{(2)}\right\}.$$

Therefore, from (7.5a) and (7.5b), it follows that

$$W_i^{(1)} = \left(W_{i-1}^{(2)}-T_i\right)^+ \text{ and } W_i^{(2)} = \left(W_{i-1}^{(1)}+V_{i-1}-T_i\right)^+, \text{ only if } i\in\mathcal{I}. \qquad (7.40)$$

Further, define the following random times: for given $0 < \delta < 1$,

$$\sigma := \sup\left\{1 \le i < n : W_i^{(2)} \ge \delta b, W_{i-1}^{(2)} < \delta b\right\} \text{ and}$$

$$\bar{\sigma} := \begin{cases} \inf\{\sigma < i \le n : i \in \mathcal{I}\} \wedge n, & \text{if } \sigma \ne -\infty \\ n, & \text{otherwise.} \end{cases}$$

We follow the usual notions that $\sup \emptyset = -\infty$ and $\inf \emptyset = \infty$. Note that σ and $\bar{\sigma}$ are measurable with respect to the sigma algebra generated by the random variables $(V_i, T_{i+1} : 0 \le i < n)$. If $W_n^{(1)}$ is larger than b, then the process $W^{(2)}$ must have become larger than δb before the n^{th} arrival. Therefore,

$$\mathbb{P}\left\{W_n^{(1)} > b, \sigma = -\infty\right\} = 0.$$

The random time σ denotes the last time before n when the maximum workload process $W^{(2)}$ becomes reasonably large and stays large after that. If both the servers have at least δb workload at time $\bar{\sigma}$ and b workload at time n, the servers are less likely to get emptied between times $\bar{\sigma}$ and n. To formally talk about this, define the event

$$B := \left\{W_i^{(1)} > 0 \text{ for all } \bar{\sigma} \le i \le n\right\},$$

and the following iid sum:

$$\hat{S}_k^l := \sum_{i=k}^{l} \left(\frac{1}{2}V_{i-1} - T_i\right), \text{ for } 1 \le k \le l \le n.$$

Lemma 7.21. $\mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n \cap B\right\} \le \mathbb{P}\left\{\hat{S}_\sigma^n + \frac{\beta}{2}(n-\sigma) \ge b\left(1 - \delta - \frac{\theta}{2}\right), \bar{A}_n\right\}.$

Proof. On the set $\{W_n^{(1)} > b, \bar{A}_n \cap B\}$, since the waiting time is positive for all the arrivals between $\bar{\sigma}$ and n, from the recursions (7.5a) and (7.5b),

$$W_{i+1}^{(1)} + W_{i+1}^{(2)} = W_i^{(1)} + W_i^{(2)} + (V_i - T_{i+1}) - T_{i+1}, \text{ for } \bar{\sigma} - 1 \le i < n,$$

which can be used in:

$$\begin{aligned} 2b &\le W_n^{(1)} + W_n^{(2)} \\ &= W_{\bar{\sigma}-1}^{(1)} + W_{\bar{\sigma}-1}^{(2)} + \sum_{i=\bar{\sigma}}^{n}(V_{i-1} - 2T_i) \\ &\le 2W_{\bar{\sigma}-1}^{(2)} + 2\sum_{i=\bar{\sigma}}^{n}\left(\frac{1}{2}V_{i-1} - T_i\right). \end{aligned}$$

Since $\sum_{i=\bar{\sigma}}^{n}\left(\frac{1}{2}V_{i-1} - T_i\right) =: \hat{S}_{\bar{\sigma}}^n$,

$$W_{\bar{\sigma}-1}^{(2)} \ge b - \hat{S}_{\bar{\sigma}}^n \text{ on } \{W_n^{(1)} > b, \bar{A}_n \cap B\}. \tag{7.41}$$

By the definition of σ, we have $W_i^{(2)} > 0$ for all $i \in \{\sigma, \ldots, n\}$. Since $\mathcal{I}$ does not include any further index between σ and $\bar{\sigma}$, $W_{\bar{\sigma}-1}^{(2)} = W_\sigma^{(2)} - \sum_{i=\sigma+1}^{\bar{\sigma}-1} T_i$ on the event under consideration. Further, recall that $\sigma \in \mathcal{I}$. Then it follows from (7.40) that,

$$\begin{aligned} W_{\bar{\sigma}-1}^{(2)} &= W_{\sigma-1}^{(1)} + V_{\sigma-1} - T_\sigma - \sum_{i=\sigma+1}^{\bar{\sigma}-1} T_i \\ &\leq W_{\sigma-1}^{(1)} + \frac{1}{2} V_{\sigma-1} + \hat{S}_\sigma^{\bar{\sigma}-1} \\ &\leq \delta b + \frac{1}{2}(\theta b + \beta(n-\sigma)) + \hat{S}_\sigma^{\bar{\sigma}-1}, \end{aligned}$$

on the set $\{W_n^{(1)} > b, \bar{A}_n \cap B\}$. The last inequality follows from the facts that the service times are restricted on $\bar{A}_n$, that is, $\{V_{n-k} \leq \theta b + \beta k\}$ on $\bar{A}_n$ for any $k < n$, and $W_{\sigma-1}^{(2)} \leq \delta b$ (by the definition of σ). Combining this with (7.41), we get:

$$\hat{S}_\sigma^n + \frac{\beta}{2}(n-\sigma) \geq b\left(1 - \delta - \frac{\theta}{2}\right), \tag{7.42}$$

on $\{W_n^{(1)} > b, \bar{A}_n \cap B\}$. This establishes the claim. □

We shall later argue that for suitably chosen θ, β and δ, the probability of $\hat{S}_\sigma^n$ exceeding such a large value is vanishingly small compared to $\mathbb{P}\{W_n^{(1)} > b, A_n\}$, as $b \to \infty$. Now turning our attention to the complementary event $\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}\}$, we define the following random times:

$$\tau_0 := \sup\{\bar{\sigma} \leq i < n : W_i^{(1)} = 0\}, \text{ and } \tau := \sup\{\bar{\sigma} \leq i \leq \tau_0 : i \in \mathcal{I}\}.$$

It is instructive to check that both τ and τ_0 are well-defined on the set $\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}\}$; that is,

$$\mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}, \tau_0 = -\infty\right\} = 0, \text{ and } \mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}, \tau = -\infty\right\} = 0.$$

Lemma 7.22. $\mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}\right\} \leq \mathbb{P}\left\{\hat{S}_{\tau_0+1}^n \geq \left(1 - \frac{\theta}{2}\right) b + \frac{1}{2}\left(\sum_{i=\tau}^{\tau_0-1} V_i - \beta(n - \tau + 1)\right)\right\}$.

Proof. When τ_0 and τ are well-defined, see that all the customers indexed by $\{\tau+1, \ldots, \tau_0\}$ are routed to the same server, say server $i, i \in \{0, 1\}$. Then the workload of server $1-i$ decreases by $T_{\tau+1} + \ldots + T_{\tau_0}$ during the same period; despite this decrease of the maximum workload process $W^{(2)}$, we have $W_{\tau_0}^{(2)}$ larger than δb, whereas the server i, which processes all the jobs indexed from $\tau+1$ to τ_0, has zero workload at time τ_0. This observation can be used to derive a

lower bound on how large the service requirement of customer $\tau - 1$ must have been. To make all these precise, we note that:

$$W_{\tau_0}^{(1)} \geq W_{\tau}^{(1)} + \sum_{i=\tau+1}^{\tau_0} (V_{i-1} - T_i),$$

due to (7.5a). Since $W_{\tau_0}^{(1)} = 0$,

$$\begin{aligned}
W_{\tau_0}^{(2)} &= W_{\tau_0}^{(2)} - W_{\tau_0}^{(1)} \\
&= \left(W_{\tau}^{(2)} - \sum_{i=\tau+1}^{\tau_0} T_i \right) - W_{\tau_0}^{(1)} \\
&\leq \left(W_{\tau}^{(2)} - \sum_{i=\tau+1}^{\tau_0} T_i \right) - \left(W_{\tau}^{(1)} + \sum_{i=\tau+1}^{\tau_0} (V_{i-1} - T_i) \right) \\
&= W_{\tau}^{(2)} - W_{\tau}^{(1)} - \sum_{i=\tau+1}^{\tau_0} V_{i-1} \\
&= \left(W_{\tau-1}^{(1)} + V_{\tau-1} - T_\tau \right) - \left(W_{\tau-1}^{(2)} - T_\tau \right)^+ - \sum_{i=\tau}^{\tau_0-1} V_i \\
&\leq \left(W_{\tau-1}^{(1)} - W_{\tau-1}^{2)} \right) + V_{\tau-1} - \sum_{i=\tau}^{\tau_0-1} V_i,
\end{aligned}$$

where the penultimate step is due to the application of (7.40) after observing that $\tau \in \mathcal{I}$. Since $W^{(1)} < W^{(2)}$, we obtain the following from the above string of inequalities:

$$V_{\tau-1} \geq W_{\tau_0}^{(2)} + \sum_{i=\tau}^{\tau_0-1} V_i, \tag{7.43}$$

on the set where τ and τ_0 are well-defined. Since $W_i^{(1)} > 0$ for $\tau_0 < i \leq n$, proceeding similar to how we obtained (7.41),

$$\begin{aligned}
2b &\leq W_n^{(1)} + W_n^{(2)} \\
&= W_{\tau_0}^{(1)} + W_{\tau_0}^{(2)} + \sum_{i=\tau_0+1}^{n} (V_{i-1} - T_i) - \sum_{i=\tau_0+1}^{n} T_i \\
&= W_{\tau_0}^{(2)} + 2\hat{S}_{\tau_0+1}^{n}. \\
\therefore W_{\tau_0}^{(2)} &\geq 2(b - \hat{S}_{\tau_0+1}^{n}) \text{ on } \{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}\}.
\end{aligned} \tag{7.44}$$

Further, recall that $V_{\tau-1} \leq \theta b + \beta(n-\tau+1)$ on the set $\bar{A}_n$. Then from (7.43) and (7.44),

$$\begin{aligned}\theta b + \beta(n-\tau+1) &\geq V_{\tau-1} \\ &\geq W_{\tau_0}^{(2)} + \sum_{i=\tau}^{\tau_0-1} V_i \\ &\geq 2(b - \hat{S}_{\tau_0+1}^n) + \sum_{i=\tau}^{\tau_0-1} V_i,\end{aligned}$$

thus yielding,

$$\hat{S}_{\tau_0+1}^n \geq \left(1-\frac{\theta}{2}\right) b + \frac{1}{2}\left(\sum_{i=\tau}^{\tau_0-1} V_i - \beta(n-\tau+1)\right), \tag{7.45}$$

on the set $\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}\}$. This concludes the proof. □

Further, adding $\frac{1}{2}(V_{\tau_0} + \ldots + V_{n-1})$ to both sides of (7.45), we obtain

$$\left\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}\right\} \subseteq \left\{\sum_{i=\tau_0+1}^{n} (V_{i-1} - T_i) \geq \left(1-\frac{\theta}{2}\right) b + \frac{1}{2}\left(\sum_{i=\tau}^{n-1} V_i - \beta(n-\tau+1)\right)\right\}. \tag{7.46}$$

For any $\epsilon > 0$ and $c > 0$, we have the following:

$$\begin{aligned}\mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}\right\} &= \mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}, 0 < n-\tau+1 \leq cb\right\} \\ &\quad + \mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}, \frac{\sum_{i=\tau}^{n-1} V_i}{n-\tau} < \mathbb{E}V_1 - \epsilon, n-\tau+1 > cb\right\} \\ &\quad + \mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}, \frac{\sum_{i=\tau}^{n-1} V_i}{n-\tau} \geq \mathbb{E}V_1 - \epsilon, n-\tau+1 > cb\right\}.\end{aligned}$$

From Lemma 7.22, we have that

$$\mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}, 0 < n-\tau+1 \leq cb\right\} \leq \mathbb{P}\left\{\hat{S}_{\tau_0+1}^n \geq \left(1-\frac{\theta+\beta c}{2}\right) b, \bar{A}_n\right\}.$$

Since $\tau < \tau_0$, it follows from the above inequality and (7.46) that,

$$\begin{aligned}\mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n \cap \bar{B}\right\} &\leq \mathbb{P}\left\{\hat{S}_{\tau_0+1}^n \geq \left(1-\frac{\theta+\beta c}{2}\right) b, \bar{A}_n\right\} \\ &\quad + \mathbb{P}\left\{\frac{\sum_{i=\tau}^{n-1} V_i}{n-\tau} < \mathbb{E}V_1 - \epsilon, n-\tau+1 > cb\right\} \\ &\quad + \mathbb{P}\left\{\sum_{i=\tau_0+1}^{n}\left(V_{i-1} - T_i - \frac{1}{2}(\mathbb{E}V_i - \beta - \epsilon)\right) \geq \left(1-\frac{\theta}{2}\right) b - \frac{\beta}{2}\right\} \\ &=: P_1 + P_2 + P_3.\end{aligned} \tag{7.47}$$

If we denote $\mathbb{P}\left\{\hat{S}^n_\sigma + \frac{\beta}{2}(n-\sigma) \geq b\left(1-\delta-\frac{\theta}{2}\right), \bar{A}_n\right\}$ by P, then from Lemma 7.21 and (7.47), we have:

$$\mathbb{P}\left\{W^{(1)}_n > b, \bar{A}_n\right\} \leq P + P_1 + P_2 + P_3. \tag{7.48}$$

To obtain bounds on P, P_1, P_2 and P_3, we take a small detour to understand tail probabilities of supremum of random walks with negative drift.

A note on the supremum of sums of negative mean regularly varying random variables

Let $(Y_n : n \geq 1)$ be i.i.d. copies of a random variable Y with negative mean, finite variance and a regularly varying tail distribution. Let

$$S_0 = 0 \text{ and } S_n = S_{n-1} + Y_n, \text{ for } n \geq 1.$$

Since $(S_n : n \geq 0)$ is a random walk with negative drift, the weak limit of $\max_{k\leq n} S_k$ exists as $n \to \infty$. The probability that the random walk S_n ever exceeds a large level b is a well-studied quantity. Upper bounds on the above mentioned level crossing probability when the increments are not allowed to take "big" values will be useful in continuing our analysis of the GI/GI/2 queue under consideration.

Lemma 7.23. *Let $0 < \theta, \beta < 1$ satisfy $\beta \leq \frac{-\theta\mathbb{E}Y}{2}$. Then for all n and $b > 0$,*

$$\mathbb{P}\left\{\max_{k\leq n} S_k > b, \bigcap_{k=1}^{n}\{Y_k \leq \theta b + \beta k\}\right\} \leq c_1\left((b\wedge n)\mathbb{P}\{Y > b\}\right)^{\frac{1}{\theta+\beta}},$$

for some constant $c_1 > 0$.

Bounds of above type can be found in Borovkov [2000]. Lemma 7.23 is just a restatement of Theorem 4.2 inBorovkov [2000]; a proof of Lemma 7.23 can be found in the same.

Upper Bounds on $\mathbb{P}\{W^{(1)}_\infty > b\}$ - continued

The events involved in the definition of quantities P_1, P_3 and P in (7.47) and in Lemma 7.21 can be recast to suit the application of Lemma 7.23 as explained below. For example, consider the event in

$$P_1 := \mathbb{P}\left\{\hat{S}^n_{\tau_0+1} \geq \left(1 - \frac{\theta+\beta c}{2}\right) b, \bar{A}_n\right\}$$

For any fixed n, let

$$X^{(n)}_i := \frac{\frac{1}{2}V_{n-i} - T_{n-i+1}}{1-\frac{\theta+\beta c}{2}} = \frac{V_{n-i} - 2T_{n-i+1}}{2-\theta-\beta c} \text{ for } i \geq 1, \text{ and}$$

$$S_0^{(n)} := 0, \qquad S_k^{(n)} := S_{k-1}^{(n)} + X_k^{(n)} \text{ for } k \geq 1.$$

From the above definition of increments $X_i^{(n)}$, it is immediate that

$$\bar{A}_n \subset \bigcap_{k=1}^{n} \left\{ X_k^{(n)} \leq \tilde{\theta} b + \tilde{\beta} k \right\},$$

where $\tilde{\theta} = \theta / (2 - \theta - \beta c)$ and $\tilde{\beta} = \beta / (2 - \theta - \beta c)$. Therefore

$$P_1 := \mathbb{P}\left\{ \hat{S}_{\tau_0+1}^n \geq \left(1 - \frac{\theta + \beta c}{2}\right) b, \bar{A}_n \right\} \leq \mathbb{P}\left\{ \max_{k \leq n} S_k^{(n)} > b, \bigcup_{k=1}^{n} \left\{ X_k^{(n)} \leq \tilde{\theta} b + \tilde{\beta} k \right\} \right\}$$

If we take $c = 1, \beta = \theta\mu/4$ and $\theta < \frac{2}{3}(1 + \mu/4)^{-1}$, then it is easy to verify that

$$\mathbb{E} X_k^{(n)} < 0, \tilde{\beta} \leq \frac{\tilde{\theta} \mathbb{E} X_k^{(n)}}{2}, \text{ and } \tilde{\theta} + \tilde{\beta} < \frac{1}{2},$$

thus satisfying all the conditions in Lemma 7.23. Then by the application of Lemma 7.23, we have

$$P_1 = o\left(\left(b \bar{B}(b) \right)^2 \right), \text{ as } b \to \infty. \tag{7.49}$$

Similarly, for the analysis of

$$P_2 := \mathbb{P}\left\{ \hat{S}_\sigma^n + \frac{\beta}{2}(n - \sigma) \geq b \left(1 - \delta - \frac{\theta}{2}\right), \bar{A}_n \right\},$$

$$\text{let } X_i^{(n)} := \frac{\frac{1}{2} V_{n-i} - T_{n-i+1} + \frac{\beta}{2}}{1 - \delta - \frac{\theta}{2}} = \frac{V_{n-i} - 2T_{n-i+1} + \beta}{2 - 2\delta - \theta} \text{ for } i \geq 1, \text{ and}$$

$$S_0^{(n)} := 0, \qquad S_k^{(n)} := S_{k-1}^{(n)} + X_k^{(n)} \text{ for } k \geq 1.$$

If we choose $\beta = \theta\mu/4$ and $\theta < \frac{2}{3}(1 - \delta)(1 + \mu/6)^{-1}$, then by a similar application of Lemma 7.23 it can be shown that

$$\begin{aligned} P &:= \mathbb{P}\left\{ \hat{S}_\sigma^n + \frac{\beta}{2}(n - \sigma) \geq b \left(1 - \delta - \frac{\theta}{2}\right), \bar{A}_n \right\} \\ &\leq \mathbb{P}\left\{ \max_{k \leq n} S_k^{(n)} > b, \bigcap_{k=1}^{n} \left\{ X_k^{(n)} < \frac{\theta b + \beta k}{2 - 2\delta - \theta} \right\} \right\} = o\left(\left(b \bar{B}(b) \right)^2 \right), \end{aligned} \tag{7.50}$$

as $b \to \infty$. Turning our attention to P_3 in (7.47), if we choose $\beta = \theta\mu/4$ and $\theta < \frac{2}{4}(1 + \mu/5)^{-1}$, then for all $\epsilon < (1 - \theta/2)\frac{\mu}{2}$, the random walk defined by

$$\sum_{i=1}^{n} \left(V_{i-1} - T_i - \frac{1}{2}(\mathbb{E} V_i - \beta - \epsilon) \right)$$

has negative drift. Then, as in (7.49) and (7.50), due to Lemma 7.23, we have

$$P_3 := \mathbb{P}\left\{\sum_{i=\tau_0+1}^{n}\left(V_{i-1} - T_i - \frac{1}{2}(\mathbb{E}V_i - \beta - \epsilon)\right) \geq \left(1 - \frac{\theta}{2}\right)b - \frac{\beta}{2}\right\} = o\left((b\bar{B}(b))^2\right), \quad (7.51)$$

as $b \to \infty$, for any n.

Since the service times V_i are non-negative, $\{V_1 + ... + V_k < k(\mathbb{E}V_1 - \epsilon)\}$ is a large deviations event involving light tails (here the left tail is light) whose probability decays exponentially with k, as $k \to \infty$. Since

$$\mathbb{P}\left\{\frac{\sum_{i=\tau}^{n-1} V_i}{n-\tau} < \mathbb{E}V_1 - \epsilon, n - \tau \geq cb\right\} \leq \sum_{k \geq cb} \mathbb{P}\left\{\frac{\sum_{i=1}^{k} V_i}{k} - \mathbb{E}V_1 < -\epsilon\right\},$$

by application of Cramer's theorem (see, for example, Dembo and Zeitouni [2009]), we say that,

$$-\frac{1}{b}\log\mathbb{P}\left\{\frac{\sum_{i=\tau}^{n-1} V_i}{n-\tau} < \mathbb{E}V_1 - \epsilon, n - \tau \geq cb\right\} = c_2, \text{ as } b \to \infty,$$

for some positive constant $c_2 > 0$. Therefore, combining this with (7.49), (7.50), (7.51) and (7.48), we obtain

$$\mathbb{P}\left\{W_n^{(1)} > b, \bar{A}_n\right\} = o\left((b\bar{B}(b))^2\right),$$

as $b \to \infty$, for any n. Combining this with (7.39) gives an upper bound for the tail probability of stationary waiting time:

$$\mathbb{P}\left\{W_\infty^{(1)} > b\right\} = O\left(b^2\bar{B}^2(b) + b^2\bar{B}(b^2)\right). \quad (7.52)$$

This completes an alternate proof to Proposition 7.2 for the case $\alpha < 2$.

7.6 Proofs of auxiliary results

In this section, we provide proofs for Lemmas 7.3, 7.5, 7.9, 7.10 and 7.14.

Proof of Lemma 7.3. First, observe that

$$\begin{aligned}
\mathbb{E}\left[(w_1 + V - T)^+ - w_1\right] &= \mathbb{E}\left[V - T\right] - \mathbb{E}\left[(V - T)I\left(w_1 + V - T < 0\right)\right] - w_1\mathbb{P}\left\{w_1 + V - T < 0\right\} \\
&= -\mathbb{E}\left[(V - T)I\left(w_1 + V - T < 0\right)\right] - w_1\mathbb{P}\left\{w_1 + V - T < 0\right\}, \text{ and} \\
\mathbb{E}\left[(w_2 - T)^+ - w_2\right] &= -\mathbb{E}T + \mathbb{E}\left[TI\left(w_2 - T < 0\right)\right] - w_2\mathbb{P}\left\{w_2 - T < 0\right\}.
\end{aligned}$$

Then, it follows from the definition of $\mathbf{W}_1$ in recursions (7.5a) and (7.5b) that

$$\begin{aligned}
\mathbb{E}_{(w_1,w_2)}\left[\left(W_1^{(1)} + W_1^{(2)}\right) - (w_1 + w_2)\right] &= \mathbb{E}\left[(w_1 + V - T)^+ - w_1\right] + \mathbb{E}\left[(w_2 + T)^+ - w_2\right] \\
&= -\mathbb{E}\left[VI\left(w_1 + V - T < 0\right)\right] - w_1\mathbb{P}\left\{w_1 + V - T < 0\right\} - \mathbb{E}\left[TI\left(w_1 + V - T \geq 0\right)\right] \\
&\quad - w_2\mathbb{P}\{w_2 - T < 0\} + \mathbb{E}\left[TI\left(w_2 - T < 0\right)\right]
\end{aligned}$$

which is negative if $\mathbb{E}[TI(T > w_2)]$ is small enough, and this can be achieved by choosing $C < w_2$ large enough. This completes the proof. □

Proof of Lemma 7.5. From recursions (7.5a) and (7.5b), it is evident that for every $1 \le k \le n$,

$$W_k^{(i)} \le \left(W_{k-1}^{(i)} + X_k\right)^+, \quad i = 1, 2.$$

We repeatedly expand the recursion, as below, to obtain

$$\begin{aligned} W_k^{(i)} &\le \max\left\{0,\ W_{k-1}^{(i)} + X_k\right\} \\ &\le \max\left\{0,\ X_k,\ W_{k-2}^{(i)} + X_{k-1} + X_k\right\} \\ &\le \max\left\{0,\ X_k,\ X_{k-1} + X_k,\ X_{k-2} + X_{k-1} + X_k, \ldots,\ W_0^{(i)} + X_1 + \ldots + X_{k-1} + X_k\right\} \\ &\le S_k - \min_{0 \le j \le k} S_j + w_i, \end{aligned}$$

where we have used that $S_0 := 0, S_j := X_1 + \ldots + X_j$ and $w_i \ge 0$. Then

$$\max_{0<k\le n} W_k^{(i)} \le \max_{0\le k\le n} S_k + \max_{0<k\le n} \max_{0\le j\le k} (-S_j) + w_i \le 2 \max_{0\le k\le n} |S_k| + w_i,$$

and this proves the result. □

Proof of Lemma 7.7. According to Corollary 1 of Pinelis [1981], we have that

$$\mathbb{P}\left\{\max_{0\le n\le m} S_n > x\right\} = \left(\mathbb{P}\left\{\max_{0\le t\le 1} \sigma B(t) > \frac{x}{m^{1/2}}\right\} + m\mathbb{P}\{X > x\}\right)(1 + o(1)). \tag{7.53}$$

uniformly over $y \ge m^{1/2}$, as $m \to \infty$ (actually, Pinelis [1981] states that the asymptotic is valid assuming $x/m^{1/2} \to \infty$ but the case $x/m^{1/2} = O(1)$ follows from the Central Limit Theorem). Also, from the development in Pinelis [1981], because $\mathbb{P}\{T > x\} = o\left(\bar{B}(x)\right)$, for each $\varepsilon > 0$, there is a positive integer m_ε such that for all $m > m_\varepsilon$,

$$\mathbb{P}\left\{\max_{0\le n\le m} (-S_n) > x\right\} \le (1+\varepsilon)\left(\mathbb{P}\left\{\max_{0\le t\le 1} \sigma B(t) > \frac{x}{m^{1/2}}\right\} + m\mathbb{P}\{-X > x\}\right). \tag{7.54}$$

Additionally, since

$$\mathbb{P}\left\{\max_{0\le n\le m} |S_n| > x\right\} \le \mathbb{P}\left\{\max_{0\le n\le m} S_n > x\right\} + \mathbb{P}\left\{\max_{0\le n\le m} (-S_n) > x\right\}$$

the statement of Lemma 7.7 immediately follows from (7.53) and (7.54). □

Proof of Lemma 7.9. Let

$$I_1(b) := \mathbb{E}\left[I\left(\max_{0\le n\le N_A(X)+1} 2|S_n| > (\delta - \delta_-)\, b, N_A(X) + 1 \le 2X\right) \max_{0\le n\le N_A(X)+1} |S_n| \;\middle|\; X > b\delta_+\right],$$

$$I_2(b) := \mathbb{E}\left[I\left(\max_{0\le n\le N_A(X)+1} 2|S_n| > (\delta - \delta_-)\, b, N_A(X) + 1 > 2X\right) \max_{0\le n\le N_A(X)+1} |S_n| \;\middle|\; X > b\delta_+\right].$$

Then our objective is to show that $I_1(b)+I_2(b) = O(b^2\bar{B}(b))$. This is an immediate consequence of the following two results.

Lemma 7.24. *Under Assumption 7.1 with $\alpha > 2$, and Assumption 7.2,*

$$I_1(b) = O\left(b^2\bar{B}(b)\right).$$

Lemma 7.25. *Under Assumption 7.1 with $\alpha > 2$, and Assumption 7.2,*

$$I_2(b) = O\left(\exp\left(-\nu b\right)\right),$$

for a suitable $\nu > 0$.

Proof of Lemma 7.24. First, observe that

$$I_1(b) \leq \int_{b\delta_+}^{\infty} \mathbb{E}\left[I\left(\max_{0\leq n\leq 2t}|S_n| > \frac{\delta-\delta_-}{2}b\right)\max_{0\leq n\leq 2t}|S_n|\right]\frac{\mathbb{P}\left\{X\in dt\right\}}{\mathbb{P}\left\{X > b\delta_+\right\}}.$$

Additionally, letting $c = (\delta-\delta_-)/2$, observe that

$$\mathbb{E}\left[I\left(\max_{0\leq n\leq 2t}|S_n| > cb\right)\max_{0\leq n\leq 2t}|S_n|\right] = cb\mathbb{P}\left\{\max_{0\leq n\leq 2t}|S_n| > cb\right\} + \int_{cb}^{\infty}\mathbb{P}\left\{\max_{0\leq n\leq 2t}|S_n| > u\right\}du$$

Therefore, due to (7.4),

$$I_1(b) = O\left(\frac{1}{\bar{B}(b\delta_+)}\int_{b\delta_+}^{\infty}\left(b\mathbb{P}\left\{\max_{0\leq n\leq 2t}|S_n| > cb\right\} + \int_{cb}^{\infty}\mathbb{P}\left\{\max_{0\leq n\leq 2t}|S_n| > u\right\}du\right)\mathbb{P}\{X\in dt\}\right). \tag{7.55}$$

Due to the applicability of the uniform asymptotic presented in Lemma 7.7 in the region $2t \leq c^2b^2$, and because of the applicability of Central Limit Theorem in the region $2t > c^2b^2$, we obtain

$$\begin{aligned}
&\int_{b\delta_+}^{\infty}\mathbb{P}\left\{\max_{0\leq n\leq 2t}|S_n| > cb\right\}\mathbb{P}\{X\in dt\} \\
&\quad= O\left(\int_{b\delta_+}^{\infty}\mathbb{P}\left\{\max_{0\leq s\leq 1}\sigma|B(s)| > \frac{cb}{\sqrt{2t}}\right\}\mathbb{P}\{X\in dt\} + \int_{b\delta_+}^{\frac{c^2b^2}{2}} t\mathbb{P}\left\{|X| > cb\right\}\mathbb{P}\{X\in dt\}\right) \\
&\quad= O\left(b\int_{b\delta_+}^{\infty}\frac{\mathbb{P}\left\{X > t\right\}}{\sqrt{t^3}}\exp\left(-\frac{c^2b^2}{4\sigma^2t}\right)dt + \mathbb{P}\{|X| > cb\}\mathbb{E}\left[XI\left(X\in\left[b\delta_+, \frac{c^2b^2}{2}\right]\right)\right]\right)
\end{aligned}$$

due to integration by parts. Now, one can apply Lemma 7.10 to evaluate the first integration, and Karamata's theorem (2.8) for the second integration, together with the observation that $\mathbb{P}\{|X| > x\} = O(\bar{B}(x))$, to obtain

$$\int_{b\delta_+}^{\infty}\mathbb{P}\left\{\max_{0\leq n\leq 2t}|S_n| > cb\right\}\mathbb{P}\{X\in dt\} = O\left(\bar{B}\left(b^2\right) + \bar{B}(b)\times b\bar{B}(b)\right) \tag{7.56}$$

On similar lines of reasoning using Lemma 7.7, again via careful integration by parts and subsequent application of Lemma 7.10 and Karamata's theorem (2.8), one can derive

$$\begin{aligned}
&\int_{b\delta_+}^{\infty}\int_{cb}^{\infty}\mathbb{P}\left\{\max_{0\le n\le 2t}|S_n|>u\right\}du\mathbb{P}\{X\in dt\}\\
&\quad=O\left(\int_{b\delta_+}^{\infty}\int_{cb}^{\infty}\mathbb{P}\left\{\max_{0\le s\le 1}\sigma\,|B(s)|>\frac{u}{\sqrt{2t}}\right\}du\;\mathbb{P}\{X\in dt\}+\int_{b\delta_+}^{\infty}\int_{\sqrt{2t}}^{\infty}t\mathbb{P}\left\{|X|>u\right\}du\;\mathbb{P}\{X\in dt\}\right)\\
&\quad=O\left(\int_{b\delta_+}^{\infty}\frac{\mathbb{P}\{X>t\}}{\sqrt{t}}\exp\left(-\frac{c^2b^2}{4\sigma^2t}\right)dt+\int_{\sqrt{2b\delta_+}}^{\infty}\int_{b\delta_+}^{\frac{u^2}{2}}t\mathbb{P}\{X\in dt\}\mathbb{P}\{|X|>u\}\right)\\
&\quad=O\left(b\bar{B}\left(b^2\right)+b\bar{B}(b)\times b\bar{B}(b)\right).
\end{aligned}$$

This bound, along with (7.55), (7.56) and the observation that $\bar{B}(b\delta_+)=\Theta(\bar{B}(b))$, prove Lemma 7.24. □

Proof of Lemma 7.25. Since $T_1+\ldots+T_{N_A(t)}\le t$ (follows from the definition of $N_A(t)$) and $V_0:=0$,

$$\begin{aligned}
\mathbb{E}\left[I\left(N_A(t)+1>2t\right)\max_{0\le n\le N_A(t)+1}|S_n|\right]&\le\mathbb{E}\left[I\left(N_A(t)>2t-1\right)\sum_{n=1}^{N_A(t)+1}\left(V_{n-1}+T_n\right)\right]\\
&\le\mathbb{E}\left[I\left(N_A(t)>2t-1\right)\left(\sum_{n=1}^{N_A(t)}V_n+t+T_{N_A(t)+1}\right)\right]\\
&\le\mathbb{E}V\times\mathbb{E}\left[N_A(t)I\left(N_A(t)>2t-1\right)\right]+\left(t+\mathbb{E}T\right)\mathbb{P}\left\{N_A(t)>2t-1\right\}\\
&\le C_1t\mathbb{P}\left\{N_A(t)>2t-1\right\}+C_2\int_{2t-1}^{\infty}\mathbb{P}\left\{N_A(t)>s\right\}ds
\end{aligned}$$

for suitable positive constants C_1 and C_2 independent of t. Here, note that the penultimate inequality is simply due to the independence between V_n and T_n for $n\ge 1$. Therefore,

$$\begin{aligned}
I_2(b)&\le\int_{b\delta_+}^{\infty}\mathbb{E}\left[I\left(N_A(t)+1>2t\right)\max_{0\le n\le N_A(t)+1}|S_n|\right]\frac{\mathbb{P}\left\{X\in dt\right\}}{\mathbb{P}\left\{X>b\delta_+\right\}}\\
&\le C_1\int_{b\delta_+}^{\infty}t\mathbb{P}\left\{N_A(t)>2t-1\right\}\frac{\mathbb{P}\left\{X\in dt\right\}}{\mathbb{P}\left\{X>b\delta_+\right\}}+C_2\int_{b\delta_+}^{\infty}\int_{2t-1}^{\infty}\mathbb{P}\left\{N_A(t)>s\right\}ds\frac{\mathbb{P}\left\{X\in dt\right\}}{\mathbb{P}\left\{X>b\delta_+\right\}}\\
&\le\mathbb{P}\left\{N_A\left(b\delta_+\right)>2b\delta_+-1\right\}\left(C_1\int_{b\delta_+}^{\infty}t\frac{\mathbb{P}\left\{X\in dt\right\}}{\mathbb{P}\left\{X>b\delta_+\right\}}+C_2\int_{2b\delta_+-1}^{\infty}\frac{\mathbb{P}\left\{X>\frac{s}{2}\right\}}{\mathbb{P}\left\{X>b\delta_+\right\}}ds\right),
\end{aligned}\tag{7.57}$$

where we have used a simple change of order of integration to arrive at the above conclusion. Since $N_A(x)/x\to 1$ as $x\to\infty$, the event $\{N_A(b\delta_+)>2b\delta_+-1\}$ is a large deviations event with probability exponentially decaying in b, whereas the sum appearing in the parenthesis in (7.57)

is $O(b)$ due to Karamata's theorem (2.8). This proves the claim that $I_2(b) = O(\exp(-\nu b))$ for a suitable constant $\nu > 0$. □

As mentioned earlier, Lemmas 7.24 and 7.25, together complete the proof of Lemma 7.9. □

Proof of Lemma 7.10. Due to Potter's bounds (2.7), given $\varepsilon > 0$, we have

$$\int_b^\infty \frac{v(t)}{v\left(b^2\right)} \exp\left(-\frac{cb^2}{t}\right) dt = O\left(\int_b^\infty \left(\frac{b^2}{t}\right)^{\alpha-\varepsilon} \exp\left(-\frac{cb^2}{t}\right) dt\right)$$

for all suitably large values of b. Changing variables $u = b^2/t$, we obtain

$$\int_b^\infty \frac{v(t)}{v\left(b^2\right)} \exp\left(-\frac{cb^2}{t}\right) dt = O\left(b^2 \int_0^\infty u^{\alpha-2-\epsilon} \exp(-cu) du\right) = O\left(b^2\right)$$

for all ϵ small enough such that $\alpha - 2 - \epsilon > 0$, and this verifies the claim. □

Proof of Lemma 7.14. Letting $\bar{c} = (\delta - \delta_-)/2$, observe that

$$\begin{aligned}
&\mathbb{E}\left[I\left(\max_{0\le n\le 2X} |S_n| > \bar{c}b\right) \max_{0\le n\le 2X} |S_n| \;\middle|\; X > b\delta_+\right] \\
&\quad = \bar{c}b\mathbb{P}\left\{\max_{0\le n\le 2X} |S_n| > \bar{c}b \;\middle|\; X > b\delta_+\right\} + \int_{\bar{c}b}^\infty \mathbb{P}\left\{\max_{0\le n\le 2X} |S_n| > u \;\middle|\; X > b\delta_+\right\} du \\
&\quad \le 3\bar{c}b\mathbb{P}\left\{Z_* > \frac{\bar{c}b}{(2cX)^{\frac{1}{\alpha}}} \;\middle|\; X > b\delta_+\right\} + 3\int_{\bar{c}b}^\infty \mathbb{P}\left\{Z_* > \frac{u}{(2cX)^{\frac{1}{\alpha}}} \;\middle|\; X > b\delta_+\right\} du \quad (7.58)
\end{aligned}$$

because of Lemma 7.12. Since $\mathbb{P}\{X > x\} \sim cx^{-\alpha}$ as $x \to \infty$, after simple integration using Karamata's theorem (2.9), one can show that

$$\bar{c}b\mathbb{P}\left\{Z_* > \frac{\bar{c}b}{(2cX)^{\frac{1}{\alpha}}} \;\middle|\; X > b\delta_+\right\} = \bar{c}b\mathbb{E}\left[\frac{\mathbb{P}\left\{X > \frac{1}{2c}\left(\frac{\bar{c}b}{Z_*}\right)^\alpha\right\}}{\mathbb{P}\left\{X > b\delta_+\right\}} \wedge 1\right] = O\left(b^2\bar{B}(b)\right), \text{ and}$$

$$\int_{\bar{c}b}^\infty \mathbb{P}\left\{Z_* > \frac{u}{(2cX)^{\frac{1}{\alpha}}} \;\middle|\; X > b\delta_+\right\} du = \int_{\bar{c}b}^\infty \left(\frac{\mathbb{P}\left\{X > \frac{u^\alpha}{2cZ_*^\alpha}\right\}}{\mathbb{P}\left\{X > b\delta_+\right\}} \wedge 1\right) du = O\left(b^2\bar{B}(b)\right).$$

Therefore, due to (7.58), we obtain

$$\mathbb{E}\left[I\left(\max_{0\le n\le 2X} |S_n| > \bar{c}b,\ N_A(X) + 1 \le 2X\right) \max_{0\le n\le 2X} |S_n| \;\middle|\; X > b\delta_+\right] = O\left(b^2\bar{B}(b)\right).$$

On the other hand, the component corresponding to the large deviations event $\{N_A(X) + 1 \ge 2X\}$ is handled similar to Lemma 7.25, and this upper bounding procedure results in

$$\mathbb{E}\left[I\left(\max_{0\le n\le 2X} |S_n| > \bar{c}b,\ N_A(X) + 1 > 2X\right) \max_{0\le n\le 2X} |S_n| \;\middle|\; X > b\delta_+\right] = O\left(\exp(-\nu b)\right),$$

for some $\nu > 0$. The last two upper bounds are enough to conclude the statement of Lemma 7.14. □

7.7 Concluding remarks

We considered a two-server queue and developed tail bounds for stationary waiting time when the traffic intensity $\rho = 1$. While doing so, we unravel the following interesting phenomena that seem to be unique to the integer traffic intensity case:

1) Our analysis reveals that when the job sizes have finite variance, there are two competing terms in the tail asymptotics that arise qualitatively due to very different phenomena (arrival of one vs. two big jobs). As mentioned earlier, either of the two terms in (7.1) can be dominant based on the nature of the slowly varying function $L(\cdot)$.

2) When the job sizes have infinite variance, only one effect is dominant (the one involving arrival of one huge job of size b^α).

These transitions in system behaviour do not appear when the traffic intensity is not an integer. Unlike the traditional analysis of multi-server queues processing heavy-tailed jobs, we saw that the development of (7.1) involves not only the combination of these law of large numbers and big jump heuristics, but, in addition, one has to account for the impact of effects which occur at the time scales governed by the Central Limit Theorem.

Bibliography

R. J. Adler, R. E. Feldman, and M. S. Taqqu, editors. *A practical guide to heavy tails*. Birkhäuser Boston Inc., Boston, MA, 1998. ISBN 0-8176-3951-9. Statistical techniques and applications.

S. Asmussen. Subexponential asymptotics for stochastic processes: extremal behavior, stationary distributions and first passage probabilities. *The Annals of Applied Probability*, 8 (2):354–374, 05 1998. doi: 10.1214/aoap/1028903531. URL http://dx.doi.org/10.1214/aoap/1028903531.

S. Asmussen. *Ruin Probabilities*. Advanced series on statistical science & applied probability. World Scientific Publishing Company, Incorporated, 2000. ISBN 9789812779311. URL http://books.google.co.in/books?id=LblaB4XJg9wC.

S. Asmussen. *Applied probability and queues*, volume 51 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003a. ISBN 0-387-00211-1. Stochastic Modelling and Applied Probability.

S. Asmussen. *Applied Probability and Queues*. Applications of mathematics (Springer): Stochastic modelling and applied probability. Springer, 2003b. ISBN 9780387002118. URL http://books.google.co.in/books?id=BeYaTxesKy0C.

S. Asmussen and P. Glynn. *Stochastic simulation: Algorithms and Analysis*, volume 57 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2007. ISBN 978-0-387-30679-7.

S. Asmussen and C. Kluppelberg. Large deviations results for subexponential tails, with applications to insurance risk. *Stochastic Processes and their Applications*, 64(1):103 – 125, 1996. ISSN 0304-4149.

S. Asmussen and D. P. Kroese. Improved algorithms for rare event simulation with heavy tails. *Adv. in Appl. Probab.*, 38(2):545–558, 2006. ISSN 0001-8678.

S. Asmussen, K. Binswanger, and B. Højgaard. Rare events simulation for heavy-tailed distributions. *Bernoulli*, 6(2):303–322, 2000. ISSN 1350-7265.

F. Baccelli, S. Schlegel, and V. Schmidt. Asymptotics of stochastic networks with subexponential service times. *Queueing Systems*, 33(1-3):205–232, 1999. ISSN 0257-0130. doi: 10.1023/A:1019176129224. URL http://dx.doi.org/10.1023/A%3A1019176129224.

A. Bassamboo, S. Juneja, and A. Zeevi. On the inefficiency of state-independent importance sampling in the presence of heavy tails. *Oper. Res. Lett.*, 35(2):251–260, 2007. ISSN 0167-6377.

A. Bassamboo, S. Juneja, and A. J. Zeevi. Portfolio credit risk with extremal dependence: Asymptotic analysis and efficient simulation. *Operations Research*, 56(3):593–606, 2008.

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1989. ISBN 0-521-37943-1.

J. Blanchet and P. Glynn. Efficient rare-event simulation for the maximum of heavy-tailed random walks. *Ann. Appl. Probab.*, 18(4):1351–1378, 2008a. ISSN 1050-5164.

J. Blanchet and P. Glynn. Efficient rare-event simulation for the maximum of heavy-tailed random walks. *Ann. Appl. Probab.*, 18(4):1351–1378, 08 2008b. doi: 10.1214/07-AAP485. URL http://dx.doi.org/10.1214/07-AAP485.

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. ISSN 1876-7354.

J. Blanchet and J. Liu. State-dependent importance sampling for regularly varying random walks. *Adv. in Appl. Probab.*, 40(4):1104–1128, 2008. ISSN 0001-8678.

J. Blanchet and J. Liu. Efficient simulation and conditional functional limit theorems for ruinous heavy-tailed random walks. *Stochastic Processes and their Applications*, 122(8):2994 – 3031, 2012. ISSN 0304-4149.

J. Blanchet, P. Glynn, and J. Liu. Fluid heuristics, Lyapunov bounds and efficient importance sampling for a heavy-tailed G/G/1 queue. *Queueing Systems*, 57(2-3):99–113, 2007a. ISSN 0257-0130.

J. Blanchet, P. Glynn, and J. Liu. Fluid heuristics, Lyapunov bounds and efficient importance sampling for a heavy-tailed G/G/1 queue. *Queueing Systems*, 57(2-3):99–113, 2007b. ISSN 0257-0130. doi: 10.1007/s11134-007-9047-4. URL http://dx.doi.org/10.1007/s11134-007-9047-4.

A. Borovkov. *Asymptotic methods in queuing theory*. Wiley Series in Probability and Statistics: Probability and Statistics Section Series. J. Wiley, 1984. ISBN 9780471902867. URL http://books.google.com/books?id=Ve3uAAAAIAAJ.

A. A. Borovkov. Estimates for the distribution of sums and maxima of sums of random variables without the cramer condition. *Siberian Mathematical Journal*, 41(5):811–848, 2000.

A. A. Borovkov and K. A. Borovkov. On probabilities of large deviations for random walks. I. Regularly varying distribution tails. *Theory of Probability and Its Applications*, 46(2):193–213, 2002. doi: 10.1137/S0040585X97978877. URL http://epubs.siam.org/doi/abs/10.1137/S0040585X97978877.

A. A. Borovkov and K. A. Borovkov. *Asymptotic analysis of random walks*, volume 118 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 2008. ISBN 978-0-521-88117-3. doi: 10.1017/CBO9780511721397. URL http://dx.doi.org/10.1017/CBO9780511721397. Heavy-tailed distributions, Translated from the Russian by O. B. Borovkova.

A. A. Borovkov and O. Boxma. On large deviation probabilities of random walks with heavy tails. *Preprint EURANDOM, Eindhoven*, 2001. ISSN 1389-2355.

H. P. Chan, S. Deng, and T.-L. Lai. Rare-event simulation of heavy-tailed random walks by sequential importance sampling and resampling. *Advances in Applied Probability*, 44(4):1173–1196, 12 2012. doi: 10.1239/aap/1354716593. URL http://dx.doi.org/10.1239/aap/1354716593.

D. B. Cline and T. Hsing. Large deviation probabilities for sums and maxima of random variables with heavy or subexponential tails. *Preprint, Texas A&M University*, 501, 1991.

K. Debicki, I. Sierpiska, and B. Zwart. Asymptotics of hybrid fluid queues with lvy input. *J. Appl. Probab.*, 50(1):103–113, 03 2013. doi: 10.1239/jap/1363784427. URL http://dx.doi.org/10.1239/jap/1363784427.

F. Delbaen and J. Haezendonck. Classical risk theory in an economic environment. *Insurance: Mathematics and Economics*, 6(2):85 – 116, 1987. ISSN 0167-6687. doi:

10.1016/0167-6687(87)90019-9. URL http://www.sciencedirect.com/science/article/pii/0167668787900199.

A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998. ISBN 0-387-98406-2. doi: 10.1007/978-1-4612-5320-4. URL http://dx.doi.org/10.1007/978-1-4612-5320-4.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer, 2009. ISBN 9783642033117. URL http://books.google.co.in/books?id=iT9JRlGPx5gC.

D. Denisov and V. Shneer. Local asymptotics of the cycle maximum of a heavy-tailed random walk. *Advances in Applied Probability*, 39(1):221–244, 03 2007. doi: 10.1239/aap/1175266476. URL http://dx.doi.org/10.1239/aap/1175266476.

D. Denisov, D. Korshunov, and V. Wachtel. Potential analysis for positive recurrent markov chains with asymptotically zero drift: Power-type asymptotics. *Stochastic Processes and their Applications*, 123(8):3027 – 3051, 2013. ISSN 0304-4149. doi: http://dx.doi.org/10.1016/j.spa.2013.04.011. URL http://www.sciencedirect.com/science/article/pii/S0304414913001075.

P. Dupuis and R. S. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1997. ISBN 0-471-07672-4. doi: 10.1002/9781118165904. URL http://dx.doi.org/10.1002/9781118165904. A Wiley-Interscience Publication.

P. Dupuis, K. Leder, and H. Wang. Importance sampling for sums of random variables with regularly varying tails. *ACM Trans. Model. Comput. Simul.*, 17(3), July 2007. ISSN 1049-3301.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. ISBN 3-540-60931-8. For insurance and finance.

W. Feller. *An Introduction to Probability Theory and Its Applications Volume II*. Wiley, 1971. ISBN 9780471257097.

S. Foss. Approximation of multichannel queueing systems. *Siberian Mathematical Journal*, 21 (6):851–857, 1980. ISSN 0037-4466. doi: 10.1007/BF00968472. URL http://dx.doi.org/10.1007/BF00968472.

S. Foss. The method of renovating events and its applications in queueing theory. In *Semi-Markov Models*, pages 337–350. Springer, 1986.

S. Foss and V. V. Kalashnikov. Regeneration and renovation in queues. *Queueing Systems*, 8 (1):211–223, 1991.

S. Foss and D. Korshunov. Heavy tails in multi-server queue. *Queueing Syst.*, 52(1):31–48, 2006. ISSN 0257-0130. doi: 10.1007/s11134-006-3613-z. URL http://dx.doi.org/10.1007/s11134-006-3613-z.

S. Foss and D. Korshunov. On large delays in multi-server queues with heavy tails. *Mathematics of Operations Research*, 37(2):201–218, 2012. doi: 10.1287/moor.1120.0539. URL http://dx.doi.org/10.1287/moor.1120.0539.

S. Foss, D. Korshunov, and S. Zachary. *An introduction to heavy-tailed and subexponential distributions*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2011. ISBN 978-1-4419-9472-1.

P. Glasserman and J. Li. Importance sampling for portfolio credit risk. *Management Science*, 51(11):1643–1656, 2005.

P. Glynn and W. Whitt. The asymptotic efficiency of simulation estimators. *Oper. Res.*, 40(3): 505–520, 1992. ISSN 0030-364X.

J. M. Hammersley and D. C. Handscomb. *Monte Carlo methods*. Methuen & Co. Ltd., London, 1965.

T. Huang and K. Sigman. Steady-state asymptotics for tandem, split-match and other feedforward queues with heavy tailed service. *Queueing Systems*, 33(1-3):233–259, 1999. ISSN 0257-0130. doi: 10.1023/A:1019128213295. URL http://dx.doi.org/10.1023/A%3A1019128213295.

H. Hult and G. Samorodnitsky. Tail probabilities for infinite series of regularly varying random vectors. *Bernoulli*, 14(3):838–864, 08 2008. doi: 10.3150/08-BEJ125. URL http://dx.doi.org/10.3150/08-BEJ125.

S. Juneja. Estimating tail probabilities of heavy tailed distributions with asymptotically zero relative error. *Queueing Syst.*, 57(2-3):115–127, 2007.

S. Juneja and P. Shahabuddin. Simulating heavy tailed processes using delayed hazard rate twisting. *ACM Trans. Model. Comput. Simul.*, 12(2):94–118, Apr. 2002. ISSN 1049-3301.

S. Juneja and P. Shahabuddin. Rare event simulation techniques: An introduction and recent advances. *Simulation, Handbooks in Operations Research and Management Science*, pages 291–350, 2006.

V. Kalashnikov. Stability estimates for renovative processes. *Eng. Cybern*, 17:85–89, 1980.

V. Kalashnikov and S. Rachev. *Mathematical methods for construction of queueing models*. The Wadsworth & Brooks/Cole operations research series. Wadsworth & Brooks/Cole, 1990. ISBN 9780534132545. URL http://books.google.com/books?id=2_V9AAAAIAAJ.

J. Kiefer and J. Wolfowitz. On the theory of queues with many servers. *Transactions of the American Mathematical Society*, 78(1):pp. 1–18, 1955. ISSN 00029947. URL http://www.jstor.org/stable/1992945.

D. Korshunov. On distribution tail of the maximum of a random walk. *Stochastic Processes and their Applications*, 72(1):97 – 103, 1997. ISSN 0304-4149. doi: http://dx.doi.org/10.1016/S0304-4149(97)00060-4. URL http://www.sciencedirect.com/science/article/pii/S0304414997000604.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, 1993. ISBN 3-540-19832-6. doi: 10.1007/978-1-4471-3267-7. URL http://dx.doi.org/10.1007/978-1-4471-3267-7.

K. Murthy and S. Juneja. State-independent importance sampling for estimating large deviation probabilities in heavy-tailed random walks. In *Performance Evaluation Methodologies and Tools (VALUETOOLS), 2012*, pages 127 –135, oct. 2012.

K. Murthy, S. Juneja, and J. Blanchet. Optimal rare event monte carlo for Markov modulated regularly varying random walks. In *Simulation Conference (WSC), 2013 Winter*, pages 564–576, Dec 2013. doi: 10.1109/WSC.2013.6721451.

K. Murthy, S. Juneja, and J. Blanchet. State-independent importance sampling for random walks with regularly varying increments. *Stochastic Systems*, 4, 2014. doi: 10.1214/13-SSY114.

S. V. Nagaev. Large deviations of sums of independent random variables. *The Annals of Probability*, 7(5):745–789, 10 1979. doi: 10.1214/aop/1176994938. URL http://dx.doi.org/10.1214/aop/1176994938.

S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Trans. Automat. Control*, 34(1):54–66, 1989. ISSN 0018-9286.

J. Paulsen. Risk theory in a stochastic economic environment. *Stochastic Processes and their Applications*, 46(2):327 – 361, 1993. ISSN 0304-4149. doi: 10.1016/0304-4149(93)90010-2. URL http://www.sciencedirect.com/science/article/pii/0304414993900102.

J. Paulsen and H. K. Gjessing. Ruin theory with stochastic return on investments. *Advances in Applied Probability*, 29(4):pp. 965–985, 1997. ISSN 00018678. URL http://www.jstor.org/stable/1427849.

I. Pinelis. A problem on large deviations in a space of trajectories. *Theory of Probability and Its Applications*, 26(1):69–84, 1981. doi: 10.1137/1126006. URL http://epubs.siam.org/doi/abs/10.1137/1126006.

S. I. Resnick. Heavy tail modeling and teletraffic data. *Ann. Statist.*, 25(5):1805–1869, 1997. ISSN 0090-5364. With discussion and a rejoinder by the author.

L. Rozovskii. Probabilities of large deviations of sums of independent random variables with common distribution function in the domain of attraction of the normal law. *Theory of Probability and Its Applications*, 34(4):625–644, 1989.

J. S. Sadowsky and J. A. Bucklew. On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE Trans. Inform. Theory*, 36(3):579–588, 1990. ISSN 0018-9448.

A. Scheller-Wolf and K. Sigman. Delay moments for fifo GI/GI/s queues. *Queueing Systems*, 25(1-4):77–95, 1997. ISSN 0257-0130. doi: 10.1023/A:1019152317954. URL http://dx.doi.org/10.1023/A%3A1019152317954.

A. Scheller-Wolf and R. Vesilo. Sink or swim together: Necessary and sufficient conditions for finite moments of workload components in FIFO multiserver queues. *Queueing Syst. Theory Appl.*, 67(1):47–61, Jan. 2011. ISSN 0257-0130. doi: 10.1007/s11134-010-9198-6. URL http://dx.doi.org/10.1007/s11134-010-9198-6.

D. Siegmund. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.*, 4(4):673–684, 1976. ISSN 0090-5364.

N. Veraverbeke. Asymptotic behaviour of Wiener-Hopf factors of a random walk. *Stochastic Processes and their Applications*, 5(1):27 – 37, 1977. ISSN 0304-4149. doi: 10.1016/0304-4149(77)90047-3.

W. Whitt. The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution. *Queueing Systems*, 36(1-3):71–87, 2000. ISSN 0257-0130. doi: 10.1023/A:1019143505968. URL http://dx.doi.org/10.1023/A3A1019143505968.

B. Zwart, S. Borst, and K. Debicki. Subexponential asymptotics of hybrid fluid and ruin models. *Ann. Appl. Probab.*, 15(1A):500–517, 02 2005. doi: 10.1214/105051604000000648. URL http://dx.doi.org/10.1214/105051604000000648.