Extremes and fluid queues

Extremes and fluid queues / Antonius Bernardus Dieker, 2006
Proefschrift Universiteit van Amsterdam
Met lit. opg. – Met samenvatting in het Nederlands

Omslagontwerp door Tobias Baanders
Gedrukt door Ponsen & Looijen BV
ISBN 90-5776-151-3

Thomas Stieltjes Institute
for Mathematics

Nederlandse Organisatie voor Wetenschappelijk Onderzoek

Dit onderzoek kwam tot stand met steun van NWO

Extremes and fluid queues

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus prof. mr. P. F. van der Heijden
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Aula der Universiteit
op donderdag 9 maart 2006, te 10.00 uur

door

Antonius Bernardus Dieker

geboren te Amsterdam

Promotiecommissie:

Promotor:	prof. dr. M. Mandjes
Overige leden:	dr. A. A. Balkema prof. dr. ir. O. J. Boxma prof. dr. C. A. J. Klaassen prof. dr. I. Norros prof. dr. T. Rolski dr. ir. W. R. W. Scheinhardt dr. P. J. C. Spreij

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Dankwoord

Dit boekje is het resultaat van vier jaar zweten. Toen ik aan dit project begon kon ik me nauwelijks voorstellen wat het schrijven ervan allemaal zou behelzen, maar het uiteindelijke resultaat vervult mij met trots. Hoewel alleen mijn naam op de omslag staat, had ik het niet kunnen schrijven zonder de hulp van velen. Graag richt ik het woord tot enkele personen in het bijzonder, en haal enkele speciale herinneringen naar boven.

Michel, jij bent zonder twijfel de persoon die de meeste invloed heeft gehad op mijn werk; de hoofdstukken 8, 9, 10 en 14 hebben wij samen geschreven. Niet alleen deze samenwerking heb ik als zeer prettig ervaren, maar ook het vertrouwen dat jij altijd in mij gesteld hebt. De deur stond en staat altijd open voor een uitgebreide discussie over uiteenlopende onderwerpen. Veel dank hiervoor.

Dit proefschrift heeft ook erg geprofiteerd van vele discussies met collega's uit binnen- en buitenland, waarvan ik hier slechts enkelen kan noemen. Hoofdstuk 13 is gezamenlijk werk met Krzyś en Tomasz; jullie ben ik in het bijzonder erkentelijk. Bert, jij hebt mij altijd van goede adviezen voorzien. Misja, bedankt voor je ongezouten kritiek op een deel van het manuscript. Verder gaat mijn dank uit naar Thyra en Richard voor alle Twentse inspanningen en het instituut EURANDOM voor de Brabantse gastvrijheid.

Ik kan niet aan de afgelopen jaren denken zonder dat mooie herinneringen naar boven komen aan de plaatsen waar dit werk mij heeft gebracht. Ik denk daarbij niet alleen aan de grote conferenties in Brazilië en Canada, maar ook aan de werkbezoeken aan Polen en Ierland (waarvoor dank, Krzyś, Tomasz en Neil). Een langere tijd heb ik in Parijs doorgebracht, op deze mogelijkheid geattendeerd door Peter. In Parijs heb ik de basis gelegd voor een artikel met Marc [109].

René, van alle kamergenoten neem jij een speciale plaats in. Vier jaar lang wist jij het uit te houden tussen mijn ongelooflijke troep, waarvoor respect op z'n plaats is. Na ruim twee jaar werd onze kamer aangevuld met Regina; enigszins verrast door de saaie werkomgeving van twee mannen leukte jij de kamer al snel op met een paar planten. Loïc, jou wil ik bedanken voor jouw geduld als ik Frans probeerde te praten.

Mijn familie heeft mij van heel dichtbij zien worstelen aan dingen die ik hen nooit precies heb kunnen uitleggen. Pap, mam, Jos, Miranda en Marjon, heel erg bedankt voor alle steun. Jasper, jij weet altijd een lach op m'n gezicht te toveren. Jullie weten dat de afgelopen vier jaar niet altijd zo makkelijk was als dit gelikte boekje doet vermoeden.

De wiskundige stellingen en talloze formules uit dit proefschrift verbleken bij mijn grootste ontdekking van de afgelopen vier jaar: Cyndi. Lieverd, ik weet niet hoeveel punten ik je moet geven voor alles wat je voor me hebt gedaan! Ik ben blij dat we samen van Ierland kunnen gaan genieten.

Amsterdam, januari 2006
Ton

Contents

CHAPTER 1

Introduction

In queueing theory, the main entity is a station where service is provided; examples include counters, call centers, elevators, and traffic lights. Customers seeking this service arrive at the system, where they wait if the service facility cannot immediately allocate the required amount of service. They depart after being served. In most of these applications, it is intrinsically uncertain at what time customers arrive and how long they need to be served. This explains why probability theory plays an important role in the analysis of queueing systems.

In order to design these systems optimally, it is desirable to know how this randomness influences the *performance* of the system. Therefore, it is crucial to study system characteristics that reflect this performance. For instance, one may wish to analyze the probability distributions of the queue length or the waiting time of a customer, as a function of the random arrival process and service requirements.

The investigation of such a performance measure starts with a reformulation of the problem into mathematical terms. On an appropriate level of abstraction, the analysis no longer involves queues; rather a problem of (applied) probability theory needs to be solved.

The system characteristics that we encounter in this thesis are all related to so-called *extremes* as a result of this translation. An inherent advantage of investigating queues through examining extremes is the wide applicability of the resulting theory. For instance, many of our results are also relevant for risk theory and financial mathematics; even though it may be unclear upfront how these fields relate to customers waiting in a line, extremes also play a pivotal role in these theories. Since our results are often illustrated with queueing examples, we first briefly discuss how queues are related to extremes.

Extremes

To see how the connection between queues and extremes arises, we set $V_0 := 0$, and consider *Lindley's recursion*

$$V_{n+1} := \max(0, V_n + B_n - r), \tag{1.1}$$

for $n \geq 0$, where $r > 0$ is given and the B_n are nonnegative. This recursion plays a role in a variety of queueing situations, for instance when analyzing waiting times and the remaining amount of work in a queueing system. In those situations, the sequence $B := \{B_n : n \geq 0\}$ is *random*, and $V := \{V_n : n \geq 0\}$ inherits this property; we then say that B and V are *stochastic processes*.

In the absence of the maximum in (1.1), the solution of this recursion can immediately be

given. Indeed, if $\tilde{V}_{n+1} := \tilde{V}_n + B_n - r$ for $n \geq 0$ while $\tilde{V}_0 = 0$, then we have for $n \geq 1$,

$$\tilde{V}_n = \sum_{\ell=0}^{n-1} B_\ell - rn. \tag{1.2}$$

We call $\tilde{V} := \{\tilde{V}_n : n \geq 0\}$ the *free process*. Before solving the recursion *with* the maximum, it is insightful to interpret this maximum in terms of the free process. As long as the constrained process $V := \{V_n : n \geq 0\}$ is strictly positive, it behaves exactly in the same way as the free process. However, if the free process becomes negative, the constrained process is 'pushed' back to zero. The process $\{V_n : n \geq 0\}$ is called the *reflected process*; occasionally, the term *regulated process* is encountered in the literature.

In Section 1.1.2, we apply the recursion in (1.1) n times to obtain the identity

$$V_n = \tilde{V}_n - \min_{0 \leq k \leq n} \tilde{V}_k. \tag{1.3}$$

For many underlying stochastic processes $\{B_n\}$, $\tilde{V}_n - \min_{0 \leq k \leq n} \tilde{V}_k$ has the same distribution as $\max_{0 \leq k \leq n} \tilde{V}_k$. This constitutes the remarkable (and perhaps even counterintuitive) fact that the reflected process is *equal in distribution* to the running maximum of the free process. In other words, the probability that V_n exceeds x equals the probability that $\tilde{V}$ reaches x no later than time n.

As a result, there are essentially two ways to analyze the solution of Lindley's recursion. A 'direct' approach, which is based on the queueing interpretation, studies the evolution of the reflected process V. Alternatively, one can analyze the running maximum of $\tilde{V}$; this approach is based on the *extremes* of the free process $\tilde{V}$. This shows that extremes and queues are closely related.

Throughout, we use the term 'extreme' to refer to the maximum (or minimum) of a random function. We are also interested in other quantities, such as the epoch at which the maximum is attained. These objects, as well as the closely related hitting and first-passage times, have been investigated ever since the foundations of modern probability theory. In this broader context, the term *fluctuation theory* is sometimes used.

Fluid queues

In several applications of queueing theory, individual customers are so small that they can hardly be distinguished. Instead of customers, it is then easier to imagine a continuous stream of work that flows into the system. The resulting queueing models are called *fluid queues*.

Fluid queues are closely related to *dams*. A dam can be modeled as a reservoir, in which water builds up due to rainfall, is temporarily stored, and then released according to some release rule. Consequently, a fluid queue can be viewed as a dam in which work is buffered until enough capacity becomes available. Alternatively, the water in the dam can be interpreted as the amount of goods stored, and the name *storage model* is therefore sometimes used.

Modern communication networks provide further motivation for studying fluid queues. In such networks, small data packets are sent to routers (or switches), where they are queued up, subsequently inspected, and then forwarded to (other routers closer to) their destination. For network design purposes it is desirable to gain insight into the amount of work that builds up in router buffers, rather than the individual waiting times of the packets. Instead of interpreting packets as customers, the aggregate packet flow can thus be viewed as fluid.

Technically, shifting from standard queues to fluid queues amounts to working with a continuous-time version of Lindley's recursion. Even though a considerable body of theory is needed to specify this 'continuous recursion', the free and reflected processes are simply continuous-time analogues of (1.2) and (1.3) respectively. This can be intuitively understood

by imagining increasingly smaller time units. Most importantly, a distributional duality again often relates the reflected process to the running maximum of its free counterpart.

This chapter is organized as follows. First, in Section 1.1, we describe fluid queues and their relationship to extremes of stochastic processes in more detail. Then, a brief introduction to fluid networks is given in Section 1.2. Finally, Section 1.3 motivates the investigation of several special fluid queues that are encountered in this thesis.

1.1 The fluid queue

1.1.1 Model description

Let $A := \{A_t : t \geq 0\}$ be a continuous-time stochastic process such that for any $t \geq 0$, A_t is the amount of work offered to the system in the interval $[0, t]$. Throughout this thesis, we use A_t and $A(t)$ interchangeably; this notational convention is employed for all continuous-time stochastic processes. We suppose that A has right-continuous sample paths with left-hand limits, which is often summarized by saying that A has càdlàg sample paths. The buffer can be interpreted as a *fluid reservoir*, to which input is offered according to the *input process* A. The buffer is drained at a constant rate r, i.e., a tap at the bottom of the fluid reservoir releases fluid at rate r as long as the buffer is nonempty. After the fluid is processed, it immediately leaves the system. Throughout, we suppose that the buffer capacity is unlimited.

Although its name suggests that sample paths of the input process are always nondecreasing, we do not impose this condition. Indeed, if the input (over a certain time interval) is negative while the buffer stays nonempty, this input is interpreted as additional outflow. In this way, nonconstant (e.g., random) release rules can be incorporated into the model. On the other hand, since it is impossible to drain fluid from an empty reservoir, nothing happens if the buffer is empty and the system dynamics dictate that fluid be released.

We write W_t for the amount of work in the buffer at epoch t, and call this the *buffer content*. The buffer-content process is also known as a (stochastic) *storage process*. For simplicity, we suppose that there is initially no work in the system, i.e., that $W_0 = 0$.

The fluid queue naturally arises from the standard queueing model with individual customers, since it can be used to represent the unfinished work in such a model. Indeed, the input corresponds to service requirements at arrival epochs of customers, and the unfinished work declines at unit rate when service is provided. In this context, the buffer content is sometimes referred to as *workload* or *virtual waiting time*.

1.1.2 Lindley's recursion

To gain intuition for the general fluid queue, we first consider the special case where work only arrives at the epochs $0, 1, 2, \ldots$. Let B_n be the (nonnegative) amount of work that arrives at epoch n. Therefore, A is a nondecreasing (random) step function with jump B_n at epoch n. It is our aim to study the buffer content just before a new batch of work arrives: we analyze $V_n := W_{n-}$. It follows from the description of the model that V_n satisfies Lindley's recursion (1.1). Since we start with an empty system, we set $V_0 := 0$.

Lindley's recursion shows that V_n can be expressed in terms of the B_n:

$$\begin{aligned} V_1 &= \max(0, B_0 - r), \\ V_2 &= \max(0, V_1 + B_1 - r) = \max(0, B_1 - r, B_1 + B_0 - 2r), \\ &\vdots \\ V_n &= \max(0, B_{n-1} - r, B_{n-1} + B_{n-2} - 2r, \ldots, B_{n-1} + \ldots + B_0 - rn). \end{aligned}$$

The latter equation can be rewritten in terms of the A_n as

$$V_n = A_{n-1} - rn - \min(0, A_0 - r, A_1 - 2r, \ldots, A_{n-1} - rn),$$

which is Equation (1.3).

For a wide class of random step functions A, the distribution of $\{A_0, \ldots, A_{n-1}\}$ equals the distribution of $\{A_{n-1} - A_{n-2}, \ldots, A_{n-1} - A_0, A_{n-1}\}$. In that case, A is called *time-reversible*, and V_n then has the same distribution as

$$\max(0, A_0 - r, A_1 - 2r, \ldots, A_{n-1} - rn). \tag{1.4}$$

In queueing theory, one is often interested in the behavior of a system after initial effects have disappeared, i.e., when the system is in *steady-state*. This amounts to letting n grow large, but further assumptions are required to ensure that V_n does not blow up. In other words, we need assumptions for *stability* of the system. In view of (1.4), this means that $\limsup_{n\to\infty} A_{n-1} - rn < \infty$.

Under this stability assumption, as indicated by (1.4), the random quantity V_n converges in distribution (see Definition 2.9) to

$$V := \sup_{n\geq 0} A_{n-1} - rn, \tag{1.5}$$

where A_{-1} should be interpreted as zero. If A fails to be time-reversible, a related formula can be given by looking backward in time; this concept is often called *Loynes' construction*. Equation (1.5) is the main reason why extremes play an important role in the context of queues.

There is a vast body of literature on queueing theory in general, and Lindley's recursion in particular. Some textbooks (in alphabetical order) are Asmussen [19], Beneš [38], Borovkov [55], Cohen [79], Kleinrock [191], Prabhu [263], Robert [272], Takács [293], and Tijms [295]; a list that is by no means exhaustive.

Baccelli and Brémaud [33] is a basic reference on stochastic recursions in queueing theory; for stability issues, we refer to the survey paper by Foss and Konstantopoulos [134]. We also mention Aldous and Bandyopadhyay [9], who give an overview of max-type stochastic recursions in a more general context.

1.1.3 The buffer content

Motivated by the above representation of V as the maximum of the free process in case the input process A is a step function, we now focus on the distribution of the buffer content W_t as $t \to \infty$ for a general input process A. It is our aim to give a continuous-time analogue of (1.5).

Since the analysis is much more technical than in the discrete-time case, we do not give a derivation of the resulting formula, but refer to Section 13.5 for more details. We suppose that A is time-reversible; in the continuous-time case, this means that the distributions of $\{A_t - \lim_{u\uparrow(t-s)} A_u : 0 \leq s \leq t\}$ and $\{A_s : 0 \leq s \leq t\}$ are equal. The steady-state buffer content then has the same distribution as

$$\sup_{t\geq 0} A_t - rt, \tag{1.6}$$

a representation that is often attributed to Reich [268].

It is interesting to see the connection between Reich's formula and the buffer-content process $\widehat{W} := \{\widehat{W}_t : t \in \mathbb{R}\}$ on the *whole* real line. To give the definition of $\widehat{W}$, suppose that time is indexed by $\mathbb{R}$ instead of $\mathbb{R}_+$, and for $t \leq 0$, let A_t be the amount of work fed into the fluid

reservoir during the time interval $[t, 0]$. The buffer content is then also defined at negative epochs, and $\widehat{W}_t$ can be thought of as the amount of work in the system at time t if the buffer was empty at time $-\infty$. Since $\widehat{W}$ satisfies the continuous-time analogue of Lindley's recursion, it can be shown that

$$\widehat{W}_t = \sup_{-\infty < s \leq t} A_t - A_s - r(t-s). \tag{1.7}$$

Observe that $\widehat{W}_0 = \sup_{s \leq 0} rs - A_s$, so that it has the same distribution as the random variable in (1.6) if A is time-reversible.

Of special interest for this monograph is the case where A has stationary increments, i.e., $A_t - A_s$ has the same distribution as A_{t-s} for $t > s$. In that case, we say that the input is stationary. The distribution of $\widehat{W}_t$ then does not depend on t; in other words, the buffer-content process $\widehat{W}$ is stationary, and it is therefore sometimes called the *stationary buffer-content process*.

1.1.4 The busy period

If the buffer content is strictly positive at some given time t, one may ask when the buffer has last been empty. We call this the *age of the busy period* (straddling t); it is also known as the *backward busy period*. The total and remaining length of the busy period are two characteristics closely related to this age. Similarly, if the buffer is empty, the *age of the idle period* is of interest, i.e., how long the server has been idle. Here, we only focus on busy periods.

Suppose that we are interested in the steady-state age of the busy period, i.e., we let $t \to \infty$. It is convenient to consider the stationary buffer-content process $\widehat{W}$ and look at the last epoch at which the buffer content vanishes before time zero. By definition, this is the largest $t \leq 0$ such that $\sup_{s \leq t} rs - A_s = rt - A_t$. In the time-reversible case, this is equal in distribution to

$$\inf \left\{ t \geq 0 : A_t - rt = \sup_{s \geq 0} A_s - rs \right\}. \tag{1.8}$$

Hence, intuitively, the age of the busy period is the (random) optimizing t in (1.6).

1.1.5 Connection with risk theory and finance

Consider the evolution of the reserves within an insurance company. Let $\{R_t : t \geq 0\}$ be the *risk (reserve) process*, representing the capital of the company at time t. Suppose that $R_0 = x > 0$.

The *ruin probability* is the probability that the reserve ever drops below zero, i.e.,

$$\mathbb{P}\left(\inf_{t \geq 0} R_t < 0 \right) = \mathbb{P}\left(\sup_{t \geq 0} [R_0 - R_t] > x \right).$$

Consequently, studying the steady-state buffer-content distribution in a fluid queue immediately yields the ruin probability in the corresponding risk model. Note that the queue is stable if and only if the ruin probability is strictly smaller than one. This *duality* between queueing and risk models is due to Prabhu [262] for the so-called M/G/1 queue; we refer to Asmussen [19, Sec. XIV.5] for more details.

In a *classical risk process*, premiums are received continuously at a constant rate c. Claims arrive according to a Poisson process N, and their sizes are independent and identically distributed (i.i.d.). This leads to the description

$$R_t = x + ct - X_t,$$

where $X := \{X_t : t \geq 0\}$ is a compound Poisson process; X_t represents the total value of the claims that arrive in the interval $[0, t]$. A detailed analysis of this model and its ramifications can be found in the standard texts on insurance risk by Asmussen [18], Embrechts *et al.* [131], or Rolski *et al.* [275].

We now show that extremes also play an important role in finance. For this, consider a risky asset with *price process* S; for instance, S could represent the price process of a share. It could also be the total capital (i.e., the risk process) of the aforementioned insurance company, see Gerber and Shiu [143].

Suppose that we can buy a *perpetual option* on the asset modeled by S. A perpetual call (or put) option is a contract that allows the holder to buy (or sell) the underlying share(s) at some prespecified price K, the *exercise price*. The word 'perpetual' means that the holder can exercise the option at any time (yet only once); the option is then also called *American*. We remark that, as opposed to their American counterparts, *European* options have a fixed expiry date. If the holder of a perpetual call or put option chooses to exercise the option at time t, $t \geq 0$, he receives the amount

$$\max(0, S_t - K), \quad \text{or} \quad \max(0, K - S_t)$$

respectively. For simplicity, we suppose here that the interest rate of a risk-free asset (i.e., the discount rate) is zero, and that the initial price is deterministic, i.e., $S_0 = s_0$ for some $s_0 > 0$.

The widely-used *Black-Scholes model* yields a closed-form expression for the price of European calls and puts. In the Black-Scholes model, the price process S is modeled as a geometric Brownian motion, i.e., $\log S$ evolves (up to normalization) as a Brownian motion with drift.

Mordecki [238] studies the prices of perpetual options if $Z := \log S$ is a *Lévy process*. A Lévy process is characterized by the requirement that its increments are stationary and independent; more details are given in Section 11.2. In particular, since a Brownian motion with drift falls within the class of Lévy processes, the price process underlying Mordecki's model is more general than the Black-Scholes price process. Mordecki shows that the price of a perpetual call option is given by

$$\mathbb{E} \max \left(0, \frac{s_0 e^{\sup_{t \geq 0} Z_t}}{\mathbb{E} e^{\sup_{t \geq 0} Z_t}} - K \right),$$

provided $\mathbb{E} S_1 < s_0$ (or, equivalently, $\mathbb{E} e^{Z_1} < 1$). As a result, if the distribution of $\sup_{t \geq 0} Z_t$ is known, it immediately yields the price of this contract. Mordecki gives a similar formula for the perpetual call option, in which the minimum of Z plays a role.

It is not only possible to investigate Lévy-driven price processes; similar results hold for so-called Markov-additive price processes, see Asmussen *et al.* [20], Jobert and Rogers [176], or Pistorius [255].

We end this section by discussing the relevance of extremes for statistics. Closely related notions such as boundary-crossing times play an important role in *sequential analysis*, in which data is collected sequentially (as opposed to in a fixed sample). Unfortunately, since the main questions are slightly different, queueing results are of limited statistical interest. We refer to Appendix 2 of Siegmund [290] for more details.

1.2 Fluid networks

This section serves as an introduction to studying buffer contents and busy periods in (tree) networks of fluid queues. The analysis of such networks is useful for evaluating the performance of communication systems with a more realistic structure than the single fluid queue.

It is most instructive to analyze the canonical network with two fluid queues in tandem (or series). Our main goal in this section is to show that it is possible to derive formulas reminiscent of (1.6)–(1.8). In the course of this thesis, we also encounter networks with more than two nodes.

1.2.1 The buffer contents

We first look for an analogue of (1.6), Reich's representation for the steady-state buffer content.

A fluid network with two stations in tandem can be described by means of two fluid reservoirs with unlimited capacity. We allow for external input to the second queue. The first input process governs the flow of work into the first reservoir as for a single queue, and fluid is released at some constant rate r_1. As soon as it is drained, it flows *immediately* into the second reservoir, where it is joined by the input from the (second) external input process. There is a tap at the bottom of the second buffer that releases fluid at a constant rate r_2. We denote the external input processes to the first and second station by $A^{(1)}$ and $A^{(2)}$ respectively, again with the understanding that negative input should be interpreted as additional outflow.

A representation for buffer content in the first station follows as in the single fluid queue, but more work is needed for the second station. It can be seen that the steady-state buffer content in this station has the same distribution as

$$\sup_{t\geq 0}\left[\sup_{0\leq s\leq t}\left(A_s^{(2)}+(r_1-r_2)s\right)+A_t^{(1)}-r_1t\right]-\sup_{t\geq 0}\left[A_t^{(1)}-r_1t\right]. \tag{1.9}$$

The presence of multiple suprema in this expression is typical for so-called feedforward networks.

Of special interest is the case where $A^{(2)}$ is a nondecreasing process and $r_1 > r_2$. The second buffer is then *always* nonempty as long as there is content in the first buffer. Such a system is called *work-conserving*. Representation (1.9) becomes

$$\sup_{t\geq 0}\left[A_t^{(1)}+A_t^{(2)}-r_2t\right]-\sup_{t\geq 0}\left[A_t^{(1)}-r_1t\right], \tag{1.10}$$

which can be interpreted as follows. The first term corresponds to the steady-state buffer-content distribution in a single fluid queue with input process $A^{(1)}+A^{(2)}$ and drain rate r_2. Since fluid that flows from the first into the second reservoir does not influence the *aggregate* buffer-content process, the first term can be thought of as the (steady-state) *total* amount of fluid in the system. Therefore, the content in the second buffer is found upon subtracting the buffer content in the first station.

1.2.2 The busy periods

Suppose we are interested in the age of the busy periods in the above two-station tandem fluid network. Clearly, there is a busy period in each of the stations, and their joint distribution is of interest.

Motivated by the relationship between (1.6) and (1.8), it seems natural to expect a connection between busy periods and the 'arguments' of the suprema in (1.9). In the work-conserving case, the interpretation of (1.10) shows that the optimizing t in the first and second term correspond to the age of a busy period in the second and first station respectively. Note that it follows from (1.10) that the age of the busy period in the first station cannot exceed the one in the second station, as expected from the interpretation of a work-conserving system.

1.3 Approximations for the input process

In Section 1.1, we found that the steady-state buffer content in a single fluid queue is equal (in distribution) to the maximum of the free process, cf. (1.6). Unfortunately, for many arrival processes A, it is hard to analyze this quantity rigorously. A natural idea to overcome this problem is to approximate A with a 'simpler' process. This leads to limit theory for stochastic processes, see for instance Whitt [305] or Jacod and Shiryaev [170].

The goal of this section is to introduce two classes of approximating (time-reversible) input processes: Gaussian processes and Lévy processes. To do so, we first discuss a fairly generic input model motivated by telecommunications engineering, the superposition of so-called ON/OFF sources. Despite this specific model choice, the results in this section are typical in the sense that Gaussian processes and Lévy processes also appear as limits of other input processes. A well-studied alternative to the ON/OFF model is the infinite-source Poisson model, sometimes called M/G/∞-input model; see [178, 206, 228] and references therein. A connection with the ON/OFF model has been established in [172].

Two important issues arise when resorting to approximating fluid models based on limit-theory arguments. First, as argued by Wischik [307], it is extremely important to check the appropriateness of scalings. Moreover, even though the limiting process might be a good approximation for the input process, this need not be the case for the corresponding buffer-content distributions. For the ON/OFF model, this desirable property has been verified by Dębicki and Mandjes [91]; in fact, they show that the *whole* stationary buffer-content process $\widehat{W}$ converges. However, as pointed out by Konstantopoulos and Lin [196], the (tail) behavior of the corresponding distributions need not be the same, cf. Boxma and Dumas [59] and Chapter 4. In a different setting, this point has been worked out in detail by Mandjes and Borst [223].

1.3.1 The ON/OFF model

In the ON/OFF model, a router in a communication network is represented as a fluid queue. As seen before, queueing arises when information must be temporarily stored in the buffer of the router. Here, a fluid formulation is particularly suitable, as traffic consists of small packets that together resemble a continuous stream. Willinger *et al.* [306] provide statistical evidence for the applicability of the ON/OFF model in a communication-network context.

An *ON/OFF source* has two alternating states, called ON and OFF. During an ON-period, the source generates fluid (*traffic* in this context), say at unit rate. During an OFF-period, the source remains silent and there is no input to the system. We write $S_t = 1$ (or $S_t = 0$) if the source is active (or inactive) at time t. The lengths of the ON-periods are i.i.d., those of the OFF-periods as well, and the lengths of ON-periods and OFF-periods are independent; denote their distribution functions by $F_{\rm ON}$ and $F_{\rm OFF}$ respectively. We suppose that these lengths have finite mean; then we can choose S_0 and the initial ON-period (or OFF-period) so that S becomes stationary.

The buffer content in a fluid system fed by a superposition of independent ON/OFF sources is well-studied if the lengths of the ON-periods and OFF-periods have an exponential distribution, see Anick *et al.* [11]. Related systems have been studied earlier by Kosten [199] and Cohen [78]. In this exponential case, the model can be viewed as a special case of so-called *fluid-flow models*; see the survey paper by Kulkarni [203]. We examine these models extensively in Chapter 14. For the exposition in the present section, however, it is essential that we allow general distributions for the lengths of the ON-periods and OFF-periods.

Let $A^N := \{A_t^N : t \geq 0\}$ be the output process of N i.i.d. ON/OFF sources, i.e., the fluid

generated in the interval $[0, t]$ is

$$A_t^N := \int_0^t \sum_{i=1}^N S_s^{(i)} ds,$$

where the $S^{(i)}$ are i.i.d. copies of S. The process A^N can serve as an input process for a fluid queue or possibly a fluid network. In the remainder of this section, we are interested in approximations for A^N. Our presentation closely follows Taqqu *et al.* [294], to whom we refer for further details and proofs.

1.3.2 Gaussian processes as limits

A first approximation for A^N arises by letting N grow large. In that case, a natural connection arises with the so-called heavy-traffic asymptotic regime, see Dębicki and Palmowski [95].

As a result of the central limit theorem, the process

$$\frac{A_t^N - Nt\mathbb{E}S_0}{\sqrt{N}} \tag{1.11}$$

tends to a centered Gaussian process A_* as $N \to \infty$. The mathematically vague term 'tends to' can be made precise with the terminology of Chapter 4, but we do not address this issue here. A *Gaussian process* is characterized by the requirement that the finite-dimensional distributions are Gaussian (i.e., that they have a multivariate normal distribution). In particular, these distributions are completely specified by the mean and covariance structure; see Chapter 3 for more details.

Since the centered Gaussian process A_* has stationary increments, its covariance structure is determined by its variance function σ^2 given by

$$\sigma^2(t) := \mathbb{V}\mathrm{ar}\left(\int_0^t S_s ds\right) = 2\int_0^t \left(\int_0^s R(u)du\right) ds,$$

where $R : u \mapsto \mathbb{E}S_u S_0 - (\mathbb{E}S_0)^2$ is the covariance function of the stationary process S. It is easy to check that A_* is distributed as $\int_0^{\cdot} Z_s ds$, where Z is a stationary centered Gaussian process with covariance function R. Therefore, A_* is known as a *Gaussian integrated process*. An important consequence of this representation is that the sample paths of A_* are always continuous.

The behavior of the process A_* critically depends on the distributions F_{ON} and F_{OFF}. For instance, if both F_{ON} and F_{OFF} have finite second moments, the increments on intervals which are 'far enough apart' behave almost independently: $\sigma^2(t) \sim Ct$ for some constant C, like a Brownian motion [271]. This is called the *short-range dependent* case. However, the behavior of A_* is different if $1 - F_{\mathrm{ON}}(x) \sim x^{-\alpha}$ for some $1 < \alpha < 2$ while F_{OFF} has a finite second moment: then $\sigma^2(t) \sim Ct^{3-\alpha}$ for some (different) constant C. The so-called heavy tail of F_{ON} (see Section 2.4) then causes a significant long-term influence of initial behavior. Therefore, this is called the *long-range dependent* case. It also arises if $1 - F_{\mathrm{OFF}}(x) \sim x^{-\alpha'}$ for some $1 < \alpha' < 2$, see [294].

These qualitative remarks are made precise by scaling A_* in time and space. In the short-range dependent case, up to a normalizing constant, the process $A_*(Tt)/\sqrt{T}$ tends to a Brownian motion as $T \to \infty$. However, in the long-range dependent case, a different limiting process appears: again up to a normalizing constant, $A_*(Tt)/T^{(3-\alpha)/2}$ tends to a *fractional Brownian motion* with Hurst index $H := (3 - \alpha)/2 > 1/2$, i.e., a continuous Gaussian process with stationary increments and variance function $t \mapsto t^{2H}$. Note that Brownian motion is a special case of fractional Brownian motion ($H = 1/2$), but that it is excluded by the assumption that

$1 < \alpha < 2$. As an aside, we point out that a fractional Brownian motion cannot be written as an integral of a stationary Gaussian process.

A fractional Brownian motion B enjoys the following *self-similarity* property:

$$\{B_{at} : t \geq 0\} = \{a^H B_t : t \geq 0\}, \tag{1.12}$$

where the equality should be interpreted as an equality in the sense of finite-dimensional distributions. After the pioneering work of Leland *et al.* [217], much work has been done to provide statistical evidence for this property in different communication-network settings.

1.3.3 Lévy processes as limits

As an alternative to studying (1.11) for large N, one could analyze the process A^N_{Tt} for large T and fixed N. The limiting process $A^N_\circ$ that results from this scaling need not be Gaussian. In fact, there is only one case in which the limit is Gaussian: when $F_{\rm ON}$ and $F_{\rm OFF}$ have a finite second moment. In that case, up to a normalizing constant, $A^N_\circ$ is a Brownian motion. For more details on the use of Lévy processes as an approximation for network traffic, we refer to Konstantopoulos and Lin [196].

As in the previous section, let us suppose that $1 - F_{\rm ON}(x) \sim x^{-\alpha}$ for some $1 < \alpha < 2$, while $F_{\rm OFF}$ has a finite second moment. The limiting process $A^N_\circ$ then behaves completely differently from the Gaussian processes encountered in Section 1.3.2; its sample paths are not even continuous. In fact, $A^N_\circ$ is a so-called α-stable Lévy process with positive jumps, see Section 11.2. Since Lévy processes have stationary, independent increments and a Gaussian distribution is 2-stable, Brownian motion is a special case of an α-stable Lévy process. However, $\alpha = 2$ is excluded from the parameter range. The process $A^N_\circ$ is self-similar with index $1/\alpha$.

In view of the previous section, where we obtained a fractional Brownian motion by first letting $N \to \infty$ and then $T \to \infty$, it is natural to investigate $A^N_\circ$ as $N \to \infty$. Interestingly, it turns out that $A^N_\circ / N^{1/\alpha}$ does not tend to a fractional Brownian motion as $N \to \infty$, but to $CA^1_\circ$, where C is some constant.

Mikosch *et al.* [228] let $N, T \to \infty$ simultaneously, and show that it depends on the growth of N relative to T whether convergence occurs to an α-stable process or to a fractional Brownian motion. For the infinite-source Poisson model, recent progress has been made by Kaj and Taqqu [178]: they show that in certain regimes non-Gaussian and non-Lévy limiting processes appear.

1.4 Outline of the thesis

Chapter 2 surveys four techniques that play an important role in parts of this thesis: regular variation, weak convergence, large deviations, and tail asymptotics. Chapters 3–14 can be roughly classified according to the *method* of studying extremes. We distinguish three methods, and indicate how they can be used to gain insight into the probability $p_x := \mathbb{P}(\sup_{t\geq 0} A_t - rt > x)$ for $x > 0$, cf. (1.6):

- Tail asymptotics: an approximation for p_x is found that works well for large x. A disadvantage of this approach is that it is not known how large x should be chosen for the approximation to be 'satisfactory'.
- Simulation: sample paths of A are generated, and p_x is approximated by the fraction of these paths for which $\sup_{t\geq 0} A_t - rt$ exceeds x. However, if p_x is small, there are (too) few of these paths. This can be resolved by simulating a different free process under which x is exceeded rather frequently, and by then correcting afterwards for the 'error' that has been introduced.

- Transforms: suppose that we have the so-called Laplace transform

$$\beta \in \mathbb{R}_+ \mapsto \mathbb{E} \exp\left(-\beta \sup_{t \geq 0}(A_t - rt)\right) \in \mathbb{R}_+$$

  at our disposal. Since this transform completely characterizes the distribution of $\sup_{t \geq 0} A_t - rt$, it determines in principle p_x for all $x > 0$. Numerical inversion [1, 102] can be used to convert the transform to these probabilities.

This thesis consists of three parts. In Part A, we study *Gaussian queues*: motivated by Section 1.3.2, we suppose that the input process is Gaussian. We focus on the steady-state buffer content and the steady-state (total) length of the busy period. First, we restrict ourselves to so-called logarithmic tail asymptotics and qualitative behavior of the queue. After that, we establish the (exact) tail asymptotics for the buffer content. The latter results are applied to examine reduced-load equivalence for Gaussian queues, i.e., the question when a subset of M independent Gaussian input processes dominates the tail asymptotics for the buffer content.

Part B is motivated by the need to *simulate* the buffer content for Gaussian queues, as analytic results are often hard to obtain. Since the buffer-content distribution can be written as a so-called large-deviation probability, we first study how large-deviation probabilities can be simulated in general. To this end, we formulate sharp conditions under which a widely-used method, *exponential twisting*, works. These conditions are then applied to a random-walk setting, before we turn to the buffer content in a Gaussian queue.

In Part C, we study *Lévy-driven fluid systems*, relying on path decompositions (so-called splitting properties). First, these are applied to analyze the transform of the buffer content in a queue with Lévy input and a special jump structure. Furthermore, splitting is an effective concept to investigate perturbed risk processes, a variant of the classical risk process discussed in Section 1.1.5. We also show that splitting is not only useful to obtain the exact tail asymptotics for the buffer content in a single fluid queue, but that it is also a powerful method to study fluid networks driven by Lévy processes. For these networks, we find (joint) transforms of busy periods, idle periods, and buffer contents. Finally, some of these results are extended to queueing networks in a random environment, including the fluid-flow models of Section 1.3.1. This relies on an extensive analysis of Markov-additive processes.

Each of the three parts starts with an introductory chapter, where fundamental results from the literature are discussed to put the material into the right context.

CHAPTER 2

Techniques

In this chapter, we discuss four topics of analysis and probability theory: regular variation, weak convergence, large deviations, and tail asymptotics.

2.1 Regular variation

The concept of regular variation plays an important role in various branches of mathematics. Whereas it is essentially a chapter in classical real-variable theory, interest in the subject has been stimulated predominantly by probabilists, for example Feller [132, Sec. VIII.8].

In this section, we give some elements of the theory of regular variation, which we mainly use in Part A. Our treatment is based on the standard reference book by Bingham *et al.* [51]. Other good introductions to the theory are Geluk and de Haan [142], Resnick [269], and Seneta [286]. See also Embrechts *et al.* [131] for a concise treatment.

We start with the definition of regular variation.

Definition 2.1 *A nonnegative measurable function f on $[0,\infty)$ is said to be* regularly varying *at infinity with index $\rho \in \mathbb{R}$, written as $f \in \mathcal{R}_\rho$, if for all $t > 0$,*

$$\lim_{\alpha\to\infty} \frac{f(\alpha t)}{f(\alpha)} = t^\rho. \tag{2.1}$$

If $f \in \mathcal{R}_0$, then f is called slowly varying.

This definition implies that $f \in \mathcal{R}_\rho$ always has the form $f(x) = x^\rho \ell(x)$, where $\ell \in \mathcal{R}_0$. Typical examples of functions in $\mathcal{R}_\rho$ are x^ρ, $x^\rho \log(1+x)$, $(x\log(1+x))^\rho$, $x^\rho \log\log(e+x)$.

Occasionally, we make use of regular variation at the origin; a function f is regularly varying at the origin with index $\rho \in \mathbb{R}$ if $x \mapsto f(1/x)$ is regularly varying at infinity with index $-\rho$. Unless stated otherwise, we consider regular variation at infinity.

We first discuss two of the most fundamental results in the theory of regular variation, the *representation theorem* and the *uniform convergence theorem*. Interestingly, it is known that these two theorems are equivalent.

Lemma 2.2 (Representation theorem) *The function f is regularly varying with index ρ ($f \in \mathcal{R}_\rho$) if and only if*

$$f(x) = c(x)x^\rho \exp\left(\int_z^x \frac{\varepsilon(u)}{u} du\right), \quad x \geq z,$$

for some $z > 0$, where c and ε are measurable functions, $c(x) \to c_0 \in (0, \infty)$, $\varepsilon(x) \to 0$ as $x \to \infty$.

The representation theorem yields important information on the behavior of f near infinity. Indeed, if $f \in \mathcal{R}_\rho$, then the representation for the function $x \mapsto \log f(e^x)$ shows that

$$\lim_{x\to\infty} f(x) = \begin{cases} \infty & \text{if } \rho > 0; \\ 0 & \text{if } \rho < 0. \end{cases} \tag{2.2}$$

We remark that the limit of a slowly varying function ℓ at infinity does not necessarily exist; one can have $\limsup_{x\to\infty} \ell(x) = \infty$ while $\liminf_{x\to\infty} \ell(x) = 0$.

The next fundamental theorem for regularly varying functions, the uniform convergence theorem, is used extensively in Chapter 5; for notational convenience, we abbreviate it as UCT.

Theorem 2.3 (Uniform convergence theorem) *Let $0 < a \le b < \infty$. If $f \in \mathcal{R}_\rho$ (in case $\rho > 0$, assuming f bounded on each interval $(0, \cdot]$), then (2.1) holds uniformly for t*

(i) in the interval $[a, b]$ if $\rho = 0$,

(ii) in the interval $(0, b]$ if $\rho > 0$, and

(iii) in the interval $[a, \infty)$ if $\rho < 0$.

Before giving an important corollary of the UCT, we first introduce some notation that is used throughout this thesis. We say that a function g is *asymptotically equivalent* to f as $x \to \zeta$, where $\zeta \in [-\infty, \infty]$, if $f(x) = g(x)(1 + o(1))$ as $x \to \zeta$, i.e., $f(x)/g(x) \to 1$ as $x \to \zeta$. In that case, the function g is a reasonable approximation for f near ζ. The symbol '$\sim$' is shorthand for asymptotic equivalence.

Corollary 2.4 *If $f \in \mathcal{R}_\rho$ be locally bounded on $[x_0, \infty)$ for some $\rho > 0$, then we have for $x \to \infty$,*

$$\sup\{f(t) : x_0 \le t \le x\} \sim \inf\{f(t) : t \ge x\} \sim f(x).$$

If $f \in \mathcal{R}_\rho$ for some $\rho < 0$, then we have for $x \to \infty$,

$$\sup\{f(t) : t \ge x\} \sim \inf\{f(t) : x_0 \le t \le x\} \sim f(x).$$

The representation theorem and the uniform convergence theorem are key to proving other useful properties of regularly varying functions. For instance, when integrating, slowly varying functions can be treated in the same way as constants; the following lemma makes this precise.

Lemma 2.5 (Karamata's theorem; direct half) *Let $f \in \mathcal{R}_\rho$ be locally bounded on $[x_0, \infty)$. Then for any $\sigma \ge -(\rho + 1)$, as $x \to \infty$,*

$$\int_{x_0}^{x} t^\sigma f(t)dt \sim \frac{x^{\sigma+1} f(x)}{\sigma + \rho + 1}.$$

Moreover, for any $\sigma < -(\rho + 1)$, as $x \to \infty$,

$$\int_{x}^{\infty} t^\sigma f(t)dt \sim -\frac{x^{\sigma+1} f(x)}{\sigma + \rho + 1}.$$

In a sense, regularly varying functions are the only functions with this property, see the 'converse half' of Karamata's theorem [51, Thm. 1.6.1]. There also exist theorems on the behavior of regularly varying functions after differentiation. Typically, additional monotonicity assumptions are then needed, see [51, Thm. 1.7.2].

Let $f \in \mathcal{R}_\rho$, $\rho > 0$ be locally bounded on some interval $[x_0, \infty)$. One can then define the generalized inverse

$$\overleftarrow{f}(x) := \inf\{y \geq x_0 : f(y) > x\}.$$

Interestingly, functions that are asymptotically equivalent to $\overleftarrow{f}$ share an important property, as the following lemma shows. These functions are known as *asymptotic inverses*.

Lemma 2.6 *There exists some $g \in \mathcal{R}_{1/\rho}$ such that, as $x \to \infty$,*

$$g(f(x)) \sim f(g(x)) \sim x. \tag{2.3}$$

Here g is determined uniquely within asymptotic equivalence, and one version is $\overleftarrow{f}$.

Similarly, g is an asymptotic inverse near zero if (2.3) holds for $x \to 0$. It becomes clear from the context whether $x \to \infty$ or $x \to 0$.

Second-order regular variation: de Haan theory

If $f \in \mathcal{R}_\rho$, then $f(\alpha t)$ is of the same order as $f(\alpha)$. De Haan theory studies the difference $f(\alpha t) - f(\alpha)$, and can therefore be regarded as a second-order theory for regularly varying functions. Our exposition is based on the treatment in Chapter 3 of [51].

We start with the definition.

Definition 2.7 *A measurable function f on $[0, \infty)$ is said to lie in the* de Haan class *of $g \in \mathcal{R}_\rho$ with index c if for all $t > 0$,*

$$\lim_{\alpha \to \infty} \frac{f(\alpha t) - f(\alpha)}{g(\alpha)} = \begin{cases} c(t^\rho - 1)/\rho & \text{if } \rho \neq 0; \\ c \log t & \text{if } \rho = 0. \end{cases} \tag{2.4}$$

Interestingly, de Haan theory can not only be regarded as a second-order theory for regularly varying functions, but it also generalizes the first-order theory. Indeed, if $f \in \mathcal{R}_\rho$, then as $\alpha \to \infty$,

$$\log f(\alpha t) - \log f(\alpha) \to \rho \log t,$$

so that de Haan theory applies to $\log f$ with $g \equiv 1$. Although many results for (first-order) regularly varying functions have analogues in de Haan theory, we only need the uniform convergence theorem (Theorem 3.1.16 of [51]).

Theorem 2.8 *Let $0 < a \leq b < \infty$. If f lies in the de Haan class of $g \in \mathcal{R}_\rho$ with index c, then (2.4) holds uniformly for t in the interval $[a, b]$.*

2.2 Weak convergence

Weak convergence deals with the convergence of probability measures. Such convergence occurs, for instance, in the central limit theorem. We use the theory predominantly in Chapter 4.

Modern theory of weak convergence, which was originally promoted in the 1968 seminal work of Billingsley [49], supposes that these measures are defined on some metric space. Our presentation is based on the 1999 update [50] of Billingsley's work. Other expositions are in Resnick [270], Stroock [291], and Whitt [305].

Let $\mathcal{X}$ be a complete, separable metric space with metric d and let $\mathcal{B}$ be the Borel σ-field of subsets of $\mathcal{X}$ generated by the open sets. Recall that $\mathcal{X}$ is complete if every Cauchy sequence has a limit in $\mathcal{X}$, i.e., for any sequence $\{x_n\}$ with $\lim_{n\to\infty} \sup_{i,j\geq n} d(x_i, x_j) = 0$, there exists some $x \in \mathcal{X}$ such that $\lim_{n\to\infty} d(x_n, x) = 0$. Also recall that $\mathcal{X}$ is separable if it contains a countable, dense subset. A complete, separable metric space is often called a Polish space.

Suppose that $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space. A random element X in $\mathcal{X}$ is a measurable mapping from $(\Omega, \mathcal{A})$ into $(\mathcal{X}, \mathcal{B})$. For instance, X is a random variable if $\omega \in \Omega$ is mapped into $\mathcal{X} = \mathbb{R}$, and a random vector if $\mathcal{X} = \mathbb{R}^d$. X can also be a random continuous function, or even a discontinuous function.

Given a sequence $\{X_n\}$ of mappings from $(\Omega, \mathcal{A})$ to $(\mathcal{X}, \mathcal{B})$, there is a corresponding sequence of distributions on $\mathcal{X}$,

$$\nu_n = \mathbb{P} \circ X_n^{-1} = \mathbb{P}(X_n \in \cdot).$$

The measure ν_n is called the *distribution* of X_n.

Definition 2.9 *We say that* ν_n converges weakly *to* ν *(written* $\nu_n \Rightarrow \nu$*), or that* X_n converges in distribution *to* X *(written* $X_n \Rightarrow X$*), if whenever* $f \in C(\mathcal{X})$*, the class of bounded, continuous real-valued functions on* $\mathcal{X}$*, we have*

$$\lim_{n\to\infty} \int_{\mathcal{X}} f(x)\nu_n(dx) = \int_{\mathcal{X}} f(x)\nu(dx).$$

Note that the metric d defines the topology on $\mathcal{X}$, and hence determines the class $C(\mathcal{X})$. Therefore, if one has two metrics d, d' on the same space $\mathcal{X}$ and convergence in the d'-metric implies convergence in the d-metric, then there are at least as many d'-open sets as d-open sets. The d'-topology is then called *stronger* (or *finer*) than the d-topology, and the class $C(\mathcal{X}, d')$ is larger than $C(\mathcal{X}, d)$. Hence, a d-weakly convergent sequence of probability measures need not be d'-weakly convergent. In Section 4.3.3, we encounter an important example of this observation.

One of the key results in the theory of weak convergence is Prokhorov's theorem. A family Π of probability measures is called *relatively compact* if every sequence $\{\nu_n\} \subset \Pi$ contains a weakly convergent subsequence (the limit need not be in Π). The family Π is called *tight* if for each $\epsilon > 0$ there exists a compact set K such that $\nu(K) > 1 - \epsilon$ for all $\nu \in \Pi$. It can be checked that singletons are tight. The following lemma is a generalization of this fact; we point out that it relies on the assumption that $\mathcal{X}$ is complete and separable.

Lemma 2.10 (Prokhorov) *A family* Π *is tight if and only it is relatively compact.*

There are several equivalent characterizations of weak convergence, which are summarized in Billingsley's portmanteau theorem. To state this theorem, we need the notion of *ν-continuity sets*. Write the interior and closure of a set $B \subseteq \mathcal{X}$ as B^o and $\overline{B}$ respectively. Given some measure ν on $(\mathcal{X}, \mathcal{B})$, we call a set $B \in \mathcal{B}$ a ν-continuity set if its boundary $\partial B = \overline{B} \backslash B^o$ satisfies $\nu(\partial B) = 0$.

Theorem 2.11 (Portmanteau) *The following five conditions are equivalent:*

(i) $\nu_n \Rightarrow \nu$,

(ii) $\lim_n \int f d\nu_n = \int f d\nu$ *for all bounded, uniformly continuous* f,

(iii) $\limsup_n \nu_n(F) \leq \nu(F)$ *for all closed* F,

(iv) $\liminf_n \nu_n(G) \geq \nu(G)$ *for all open* G*, and*

(v) $\nu_n(B) \to \nu(B)$ *for all ν-continuity sets* B.

Another important result entails that weak convergence is preserved by continuous mappings. Let h be a continuous mapping from $\mathcal{X}$ into another metric space $\mathcal{X}'$, with metric d' and Borel σ-field $\mathcal{B}'$. Each probability measure ν on $(\mathcal{X}, \mathcal{B})$ induces a probability measure $\nu \circ h^{-1}$ on $(\mathcal{X}', \mathcal{B}')$ (namely, $\nu \circ h^{-1}(B) = \nu(h^{-1}(B))$ for $B \in \mathcal{B}'$). The following corollary to Theorem 2.11 is a straightforward consequence of the change-of-variable formula for integrals. It is called the *continuous mapping theorem*.

Corollary 2.12 *If $\nu_n \Rightarrow \nu$, then $\nu_n \circ h^{-1} \Rightarrow \nu \circ h^{-1}$.*

It is useful to know that the continuity assumption can be relaxed; see [49, Thm. 1.5.5] for a version where a discontinuous h is also allowed to depend on n.

We end this section by remarking that weak convergence is not the only convergence concept for probability measures or random elements. Stronger forms of convergence are convergence in probability and almost sure convergence; see, e.g., Chung [75, Ch. 4] or Resnick [270, Ch. 6] for an overview. Moreover, it is possible to define $X_n \Rightarrow X$ for nonmeasurable X_n and measurable X; this situation arises in statistical applications, see van der Vaart and Wellner [301].

2.3 Large deviations

There is no unambiguous mathematical meaning of the term 'large deviations'; in fact, two approaches can be distinguished.

The first considers the asymptotic behavior of the *logarithm* of probabilities, which is sometimes referred to as *rough* (or *coarse*) asymptotic behavior. It is merely a set of tools, known as *large-deviation techniques* (or *large-deviation theory*), that allows an analysis under fairly general conditions. For instance, relatively explicit results can be obtained for dependent sequences of random variables, or for variables with values in an infinite-dimensional space.

The second class of large-deviation probabilities deals with *exact asymptotics*; these are much more powerful than logarithmic asymptotics. However, results in this domain are often derived on a case-by-case basis, rather than from general principles. To give a (small) selection of the relevant literature, we mention Petrov [251], Nagaev [240], and Borovkov and Mogul′skiĭ [57].

The rest of this section focuses on the first class of large deviations, with the notable exception of Lemma 2.17. This theory is used extensively in Part B. There is a vast body of textbooks on large-deviation techniques. We mention two books by pioneers in the field: Varadhan [303] and Ellis [129]. The books by Ganesh *et al.* [140], Shwartz and Weiss [288], and Bucklew [63] focus on queueing theory and performance evaluation. Dembo and Zeitouni [100], den Hollander [101], Deuschel and Stroock [104], and Feng and Kurtz [133] are general introductions to the theory. Finally, Dupuis and Ellis [126] and Puhalskii [266] develop the theory from two alternative points of view.

Large-deviation techniques are closely related to the theory of weak convergence as discussed in the previous section. It is no coincidence that many of the statements below look somewhat similar to those in Section 2.2. In fact, one can set up the theory of large deviations completely analogously to the theory of weak convergence, see Dupuis and Ellis [126]. It is even possible to unify both theories by working with *capacities*, see O'Brien [246] or Puhalskii [265] for details.

Despite the similarities with the theory of weak convergence, we do not take the above approach here. Instead, this section is based on the textbook of Dembo and Zeitouni [100], since we need large deviations for measures on so-called topological vector spaces (see, e.g., Conway [81]). These spaces cannot be compared to Polish spaces: there are Polish spaces that are not topological vector spaces and vice versa.

Therefore, throughout this section, we let $(\mathcal{X}, \mathcal{B})$ be a regular topological space with its Borel σ-field, which includes both the possibility that $\mathcal{X}$ is a metric space and a (Hausdorff)

topological vector space. Suppose that we are given a family of probability measures on this space (for instance, as constructed from random functions as in Section 2.2). This family is either indexed by a continuous parameter $\epsilon > 0$ or by a discrete parameter $n \in \mathbb{N}$.

A function I is said to be *lower semicontinuous* if the level sets $\mathcal{L}_I(\alpha) := \{x : I(x) \leq \alpha\}$ are closed subsets of $\mathcal{X}$ for all $\alpha \in [0, \infty)$. A function $I : \mathcal{X} \to [0, \infty]$ is called a *rate function* if it is lower semicontinuous. If $\mathcal{L}_I(\alpha)$ is compact for every $\alpha \geq 0$, I is called a *good rate function.*

The central notion in large-deviation theory is the *large-deviation principle*, which consists of a large-deviation upper and lower bound. Here, and throughout this thesis, we use the convention that the infimum over an empty set is ∞.

Definition 2.13 *A family of probability measures $\{\nu_\epsilon : \epsilon > 0\}$ on $(\mathcal{X}, \mathcal{B})$ satisfies the* large-deviation principle *(LDP) with a rate function I if for all $B \in \mathcal{B}$,*

$$-\inf_{x \in B^o} I(x) \leq \liminf_{\epsilon \to 0} \epsilon \log \nu_\epsilon(B) \leq \limsup_{\epsilon \to 0} \epsilon \log \nu_\epsilon(B) \leq -\inf_{x \in \overline{B}} I(x).$$

Similarly, a sequence of probability measures $\{\nu_n : n \in \mathbb{N}\}$ on $(\mathcal{X}, \mathcal{B})$ satisfies the LDP with a rate function I and scale sequence $\{\lambda_n : n \in \mathbb{N}\}$ if for all $B \in \mathcal{B}$,

$$-\inf_{x \in B^o} I(x) \leq \liminf_{n \to \infty} \frac{1}{\lambda_n} \log \nu_n(B) \leq \limsup_{n \to \infty} \frac{1}{\lambda_n} \log \nu_n(B) \leq -\inf_{x \in \overline{B}} I(x).$$

If the random element X_n has law ν_n and $\{\nu_n\}$ satisfies the LDP with some scale sequence $\{\lambda_n\}$, we say that the family $\{X_n\}$ satisfies the LDP with this scale sequence. In the next two lemmas, we suppose that the family of probability measures is indexed by $\epsilon > 0$; the formulation of the discrete analogues is left to the reader.

A set $B \in \mathcal{B}$ is called an *I-continuity set* if $\inf_{x \in B^o} I(x) = \inf_{x \in B} I(x) = \inf_{x \in \overline{B}} I(x)$. Note that then $\lim_{\epsilon \to 0} \epsilon \log \nu_\epsilon(B) = -\inf_{x \in B} I(x)$; this is a large-deviation version of the fifth item of Theorem 2.11.

There is also a notion of tightness in the large-deviation context. A family $\{\nu_\epsilon\}$ of probability measures is said to be *exponentially tight* if for each $\alpha < \infty$, there exists a compact set K such that

$$\limsup_{\epsilon \to 0} \epsilon \log \nu_\epsilon(K^c) < -\alpha,$$

where K^c denotes the complement of K in $\mathcal{X}$. An analogue of Prokhorov's theorem holds, provided additional conditions on $\mathcal{X}$ are met; see Theorem 2.3 of O'Brien and Vervaat [247].

The following lemma captures the most important implication (for us) of Theorem 2.11 in a large-deviation framework.

Lemma 2.14 (Varadhan) *If $\{\nu_\epsilon\}$ satisfies the LDP with some good rate function I, then for any $f \in C(\mathcal{X})$, we have*

$$\lim_{\epsilon \to 0} \epsilon \log \int_{\mathcal{X}} e^{f(x)/\epsilon} \nu_\epsilon(dx) = \sup_{x \in \mathcal{X}} \left[f(x) - I(x) \right].$$

The condition that f be bounded can be relaxed: it suffices to have either

$$\lim_{M \to \infty} \limsup_{\epsilon \to 0} \epsilon \log \int_{x : f(x) \geq M} e^{f(x)/\epsilon} \nu_\epsilon(dx) = -\infty,$$

or the exponential-moment condition that for some $\gamma > 1$,

$$\limsup_{\epsilon \to 0} \epsilon \log \int_{\mathcal{X}} e^{\gamma f(x)/\epsilon} \nu_\epsilon(dx) < \infty, \tag{2.5}$$

see Lemma 4.3.8 of [100].

We end our exposition of large-deviation tools with an analogue of the continuous mapping theorem, which is known as the *contraction principle.*

Lemma 2.15 *Let $\mathcal{X}$ and $\mathcal{X}'$ be Hausdorff topological spaces and $h : \mathcal{X} \to \mathcal{X}'$ a continuous function. If $\{\nu_\epsilon\}$ satisfies the LDP in $\mathcal{X}$ with the good rate function I, then $\{\nu_\epsilon \circ h^{-1}\}$ satisfies the LDP in $\mathcal{X}'$ with the good rate function*

$$I'(x) = \inf\{I(y) : y \in \mathcal{X}, x = h(y)\}.$$

In the rest of this section, we discuss two of the most well-known instances of the large-deviation principle.

2.3.1 Cramér's LDP

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. zero-mean random variables taking values in $\mathbb{R}^d$, with distribution $\mathbb{P}_X$. For $n \geq 1$, we set $S_n := \sum_{i=1}^n X_i$. The stochastic process $\{S_n\}$ is called a *random walk*; see Section 11.1 for a detailed discussion of the one-dimensional case.

Cramér's LDP describes how S_n deviates from its mean. A key role is played by the cumulant-generating function of X_1, which is defined as $\Lambda_X(\xi) := \log \mathbb{E}(e^{\xi X_1})$ for $\xi \in \mathbb{R}$. This function assumes values in $\mathbb{R} \cup \{\infty\}$ and is convex by Hölder's inequality. We write $\operatorname{dom} \Lambda_X = \{\xi \in \mathbb{R}^d : \Lambda_X(\xi) < \infty\}$ for its domain. The *Fenchel-Legendre transform* of Λ_X is defined as $\Lambda_X^*(x) := \sup_{\xi \in \mathbb{R}^d}[\xi' x - \Lambda_X(\xi)]$, see Rockafellar [273] (or any of the aforementioned textbooks on large-deviation theory) for more details.

Proposition 2.16 (Cramér) *Suppose that $0 \in (\operatorname{dom} \Lambda_X)^o$. Then $\{S_n/n\}$ satisfies the LDP with the (convex) good rate function Λ_X^* and scale sequence $\{n\}$.*

Cramér's LDP yields the logarithmic asymptotics of $\mathbb{P}(S_n/n \in A)$, but some sharper results are also known for specific choices of A. Here, we focus on the one-dimensional case with $A = [x, \infty)$; the multidimensional case is significantly more difficult, as can be seen from the work of Iltis [168].

To establish the exact asymptotics of $\mathbb{P}(S_n \geq nx)$, one needs to distinguish between the lattice and nonlattice case. For simplicity, we suppose that $\mathbb{P}_X$ is nonlattice, and let $\xi^*(x)$ be $\arg\sup_{\xi \in \mathbb{R}}[\xi x - \Lambda_X(\xi)]$. The following lemma is taken from Höglund [159]. We denote the complementary distribution function of a standard normal random variable by Ψ, see (3.1). Moreover, $\ddot{\Lambda}_X$ stands for the second derivative of Λ_X.

Lemma 2.17 *As $n \to \infty$, we have*

$$\mathbb{P}(S_n \geq nx) \sim e^{n\ddot{\Lambda}_X(\xi^*(x))[\xi^*(x)]^2/2} \Psi\left(\xi^*(x)\sqrt{n\ddot{\Lambda}_X(\xi^*(x))}\right) e^{-n\Lambda_X^*(x)},$$

uniformly in x when $\xi^*(x)$ stays within compact subsets of $[0, \infty) \cap \operatorname{dom} \Lambda_X$.

Note that this lemma is more general than the Bahadur-Rao theorem as presented in Theorem 3.7.4 of Dembo and Zeitouni [100]. Indeed, the lemma includes a statement on so-called moderate deviations (i.e., when $x \to 0$ as $n \to \infty$ but $x\sqrt{n} \to \infty$), and also on normal deviations (i.e., when $x\sqrt{n} = O(1)$). In particular, it includes the central limit theorem for one-dimensional random walks.

2.3.2 Mogul′skiĭ's LDP

We now return to logarithmic asymptotics, and discuss a sample-path version of Cramér's LDP, which is known as Mogul′skiĭ's LDP. To state such a sample-path version, we first introduce a path space with a certain topology, so that open and closed sets are defined on this space.

For $0 \le t \le 1$, define the scaled polygonal approximation for the partial sums of X_i as

$$S_n(t) := \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} X_i + \left(t - \frac{\lfloor nt \rfloor}{n} \right) X_{\lfloor nt \rfloor + 1},$$

where $\lfloor t \rfloor$ denotes the largest integer smaller than or equal to t. Note that $t \mapsto S_n(t)$ is a continuous function on $[0,1]$.

This observation enables us to consider $S_n(\cdot)$ on the path space $C([0,1])$ of continuous functions on $[0,1]$. It is a Polish space with the metric

$$d(x,y) := \sup_{t \in [0,1]} \|x(t) - y(t)\|.$$

In fact, if one defines $\|x\| := d(x,0)$, then $C([0,1])$ is also a Banach space.

The subspace of absolutely continuous functions $\mathcal{AC}$ plays an important role in the sample-path LDP that we shortly formulate. It is defined as

$$\mathcal{AC} := \left\{ x : \sum_{\ell=1}^{k} |t_\ell - s_\ell| \to 0, s_\ell < t_\ell \le s_{\ell+1} < t_{\ell+1} \Longrightarrow \sum_{\ell=1}^{k} |x(t_\ell) - x(s_\ell)| \to 0 \right\}. \tag{2.6}$$

In particular, for any $x \in \mathcal{AC}$ we have $x(t) = \int_0^t \dot{x}(s) ds$ for some measurable $\dot{x}$. Note that $\dot{x}$ is the derivative of x in case it is differentiable. We can now describe the large deviations of the paths $S_n(\cdot)$; more details can be found in Section 5.1 of Dembo and Zeitouni [100].

Proposition 2.18 (Mogul′skiĭ) *Suppose that $\Lambda_X(\xi) < \infty$ for every $\xi \in \mathbb{R}^d$. Then $\{S_n(\cdot)\}$ satisfies the LDP in $C([0,1])$ with the scale sequence $\{n\}$ and good rate function*

$$I(x) := \begin{cases} \int_0^1 \Lambda_X^*(\dot{x}(t)) dt & \text{if } x \in \mathcal{AC}, x(0) = 0; \\ \infty & \text{otherwise.} \end{cases}$$

Comparing the assumptions of this proposition with the assumptions of Cramér's LDP (Proposition 2.16), it is most striking that we require Λ_X to be finite everywhere. However, this assumption is not necessary to prove an LDP in the spirit of Proposition 2.18, but the statement then needs to be slightly reformulated. In Section 9.3, we discuss this in more detail.

2.4 Tail asymptotics

When explicit results are unavailable, one may resort to finding an asymptotically equivalent expression of a somewhat more tractable form. Typically, we are interested in the behavior of the tail probability $\overline{F}(x) := \mathbb{P}(Y > x)$ near ∞, where Y is a given random variable with values in $\mathbb{R}$. It is the aim of this section to introduce some definitions and discuss basic properties related to such *tail asymptotics*. The techniques discussed in this section are mainly used in Chapter 12. Throughout, let $F = 1 - \overline{F}$ be the distribution function of Y.

2.4.1 The classes $\mathcal{L}(\alpha)$ and $\mathcal{S}(\alpha)$

We say that $F \in \mathcal{L}(\alpha)$ for some $\alpha \ge 0$, if

$$\lim_{x \to \infty} \overline{F}(x+y)/\overline{F}(x) = e^{-\alpha y} \tag{2.7}$$

for all $y \in \mathbb{R}$. Note that this requirement cannot be fulfilled for lattice-supported distributions F if $\alpha > 0$; a variant of the above definition is then needed, but this falls outside the scope

of this thesis. It is immediate from this definition that $x \mapsto \overline{F}(\log x)$ is regularly varying with index $-\alpha$, which makes it possible to apply the theorems of Section 2.1. For instance, the UCT shows that the convergence in (2.7) holds locally uniformly (and even more if $\alpha > 0$).

We now investigate the circumstances under which $\mathcal{L}(\alpha)$ is closed under convolutions; recall that the convolution of the distribution functions G and H is the distribution function given by

$$G * H(x) := \int_{\mathbb{R}} G(x-y)H(dy) = \int_{\mathbb{R}} H(x-y)G(dy), \quad x \in \mathbb{R}.$$

The following lemma is a simplification of Lemma 2.1 in Pakes [248]. It entails that the tail of a convolution behaves like the 'heaviest' tail, up to a constant.

Lemma 2.19 *Let $F \in \mathcal{L}(\alpha)$ for some $\alpha \geq 0$, and let G be a distribution function satisfying $\int e^{\beta x} G(dx) < \infty$ for some $\beta > \alpha$. Then $F * G \in \mathcal{L}(\alpha)$ and $\overline{F * G}(x) \sim \int e^{\alpha y} G(dy) \overline{F}(x)$ as $x \to \infty$.*

The requirement $F \in \mathcal{L}(\alpha)$ is often not enough to derive explicit results; the above lemma is an exception. This is usually resolved by replacing $F \in \mathcal{L}(\alpha)$ by the more restrictive condition that $F \in \mathcal{S}(\alpha)$ for some $\alpha \geq 0$. This means that

(i) $F \in \mathcal{L}(\alpha)$,

(ii) $\int_{-\infty}^{\infty} e^{\alpha y} F(dy) < \infty$, and

(iii) $\overline{F^{(2)}}(x) \sim 2 \int_{-\infty}^{\infty} e^{\alpha y} F(dy) \overline{F}(x)$ as $x \to \infty$.

Here, $F^{(2)} = F * F$ is the convolution of F with itself.

If one supposes that $F \in \mathcal{S}(\alpha)$ rather than only $F \in \mathcal{L}(\alpha)$, one gets the following stronger version of Lemma 2.19. It is taken from Braverman and Samorodnitsky [62, Lem. 1.1.(ii)].

Lemma 2.20 *Let $F \in \mathcal{S}(\alpha)$. If $c_i := \lim_{x\to\infty} \overline{G}_i(x)/\overline{F}(x)$ exists and is finite for two distribution functions G_1, G_2, then*

$$\overline{G_1 * G_2}(x) \sim \left(c_1 \int e^{\alpha y} G_2(dy) + c_2 \int e^{\alpha y} G_1(dy) \right) \overline{F}(x), \quad x \to \infty.$$

Moreover, $G_i \in \mathcal{S}(\alpha)$ if $c_i > 0$.

An important special case of this lemma follows by letting G_2 be a degenerate distribution at zero. Then, it shows that $\mathcal{S}(\alpha)$ is closed under tail equivalence: if $F \in \mathcal{S}(\alpha)$ and $\overline{G}(x) \sim c\overline{F}(x)$ for some $c \in (0, \infty)$, then $G \in \mathcal{S}(\alpha)$.

2.4.2 The classes $\mathcal{S}$, $\mathcal{S}^*$, and $\mathcal{S}_\Delta$

The results in the previous subsection hold in particular for $\alpha = 0$, but the classes $\mathcal{L} := \mathcal{L}(0)$ and $\mathcal{S} := \mathcal{S}(0)$ are special in many ways. It is our present aim to study some of their properties. The results that we state here without reference can be found in the survey paper by Goldie and Klüppelberg [152].

If $F \in \mathcal{L}$, we say that F is *long-tailed*. Furthermore, if $F \in \mathcal{S}$, we call F a *subexponential* distribution function. As a result of (2.2) and the fact that $x \mapsto \overline{F}(\log x)$ is slowly varying, we know that if $F \in \mathcal{L}$, then $e^{\epsilon x}\overline{F}(x) \to \infty$ as $x \to \infty$, for each $\epsilon > 0$. Distribution functions with this property are often called *heavy-tailed*, although it merely accounts for the name 'subexponential'.

If F is concentrated on $[0,\infty)$, then $\overline{F^{(2)}}(x) \sim 2\overline{F}(x)$ implies that (2.7) holds with $\alpha = 0$. In fact, $F \in \mathcal{S}$ is equivalent with

$$\overline{F^{(n)}}(x) \sim n\overline{F}(x)$$

for some (and then all) $n \geq 2$, where $F^{(n)}$ is the n-fold convolution of F with itself. That is, the tail probability of the sum of n independent subexponential random variables is asymptotically equivalent to the tail probability of their maximum.

The class $\mathcal{S}$ is a rich class of distribution functions, see Table 3.7 in [152]. Let us single out three important examples with $F \in \mathcal{S}$:

- regularly varying tail: $\overline{F} \in \mathcal{R}_\alpha$ for some $\alpha > 0$;
- lognormal tail: $\overline{F}(x) \sim \Psi(\log x)$; and
- Weibull tail: $\overline{F}(x) \sim e^{-x^\tau}$, $\tau \in (0,1)$.

General criteria for $F \in \mathcal{S}$ can be found in [152, 239].

Another relevant class of distribution functions is $\mathcal{S}^*$. A distribution function F on $[0,\infty)$ is in $\mathcal{S}^*$ if $\int_0^\infty \overline{F}(y)dy < \infty$ and

$$\int_0^x \overline{F}(y)\overline{F}(x-y)dy \sim 2\int_0^\infty \overline{F}(y)dy\overline{F}(x).$$

Like $\mathcal{S}$, this class is closed under tail equivalence, i.e., if $F \in \mathcal{S}^*$ and $\overline{G}(x) \sim c\overline{F}(x)$ for some $c \in (0,\infty)$, then $G \in \mathcal{S}^*$. The distribution functions of all of the above three examples are in $\mathcal{S}^*$.

The following lemma, a special case of Lemma 9 of Denisov *et al.* [103], shows why the class $\mathcal{S}^*$ is interesting.

Lemma 2.21 *If $F \in \mathcal{S}^*$, then $F_H \in \mathcal{S}$, where*

$$\overline{F}_H(x) = \frac{\int_{[0,\infty)} \overline{F}(x+t)H(dt)}{\int_{[0,\infty)} \overline{F}(t)H(dt)},$$

for some measure H with $\sup_t H((t,t+1]) \leq b$, $b \in (0,\infty)$.

In particular, if $F \in \mathcal{S}^*$, then $F \in \mathcal{S}$, and also $F_I \in \mathcal{S}$, where

$$F_I(x) := \int_0^x \overline{F}(y)dy \Big/ \int yF(dy)$$

is the so-called *integrated-tail* distribution function.

We now turn to *local* tail asymptotics for Y. That is, we are interested in the probability $\mathbb{P}(Y \in (x, x+T])$ for some fixed $T > 0$. After setting $\Delta := (0,T]$ and $x+\Delta := \{x+y : y \in \Delta\}$, this can be rewritten as $\mathbb{P}(Y \in x+\Delta)$. If $F \in \mathcal{S}(\alpha)$ for $\alpha > 0$, local tail asymptotics are immediately derived from the asymptotics of $\overline{F}(x)$. However, this cannot be done in the subexponential case; as pointed out by Asmussen *et al.* [22], new classes of distributions are then needed. Recall that Y has distribution function F supported on $\mathbb{R}$.

Write $F \in \mathcal{L}_\Delta$ if $\mathbb{P}(Y \in x+y+\Delta) \sim \mathbb{P}(Y \in x+\Delta)$ as $x \to \infty$, and $F \in \mathcal{S}_\Delta$ if $F \in \mathcal{L}_\Delta$ and

$$\mathbb{P}(Y_1 + Y_2 \in x+\Delta) \sim 2\mathbb{P}(Y \in x+\Delta),$$

where Y_1, Y_2 are independent copies of Y. The class $\mathcal{S}_\Delta$ turns out to be closed under tail equivalence, and many other properties known for the class $\mathcal{S}$ carry over to the $\mathcal{S}_\Delta$-case.

We end this section by pointing out an interesting relationship between $\mathcal{S}^*$ and $\mathcal{S}_\Delta$ [198]: if F is concentrated on $[0,\infty)$, then $F \in \mathcal{S}^*$ is equivalent to $F_I \in \mathcal{S}_\Delta$ for some (and then all) $T > 0$.

Part A

Gaussian queues

CHAPTER 3

Background on Gaussian processes

In Part A of this thesis, we study Gaussian queues. Given a zero-mean Gaussian process Y, Gaussian queues arise if $Y_t + \mu t$ is the amount of 'input' (e.g., internet traffic) offered to the system in the interval $[0, t]$, and $r > \mu$ is the (constant) drain rate. We investigate extremes of Gaussian processes to study tail probabilities for the steady-state buffer content and the length of the busy period.

There exist several tools for Gaussian processes that are useful in this context, and this chapter gives a selection. We first discuss large-deviation principles for Gaussian measures, which can be used to obtain the logarithmic tail asymptotics for the aforementioned quantities. However, different techniques are required for exact asymptotics, such as sharp inequalities. Therefore, we give two fundamental inequalities related to extremes of Gaussian processes. Finally, we sketch in a classical example how the *double sum method* can be applied to find the tail asymptotics for the maximum of a Gaussian process.

A Gaussian (or normal) random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ is a real-valued random variable such that for each $\beta \in \mathbb{R}$,

$$\mathbb{E}e^{i\beta X} = e^{i\mu\beta - \frac{1}{2}\sigma^2\beta^2},$$

or, equivalently, the distribution of X has density

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{\sigma^2}}.$$

We call X *centered* if $\mu = 0$, and *standard normal* if in addition $\sigma = 1$.

A (centered) *Gaussian process* is a family $\{X_t : t \in \mathbf{T}\}$ of random variables on some probability space, indexed by a parameter set $\mathbf{T}$, such that each finite linear combination $\sum_i \alpha_i X_{t_i}$ for $t_i \in \mathbf{T}$ is (centered) Gaussian. A Gaussian process is *separable* if there exists a countable set $\mathbf{S} \subset \mathbf{T}$ such that for any open set $U \subset \mathbf{T}$, almost surely

$$\sup_{t\in U} X_t = \sup_{t\in U\cap\mathbf{S}} X_t, \quad \inf_{t\in U} X_t = \inf_{t\in U\cap\mathbf{S}} X_t.$$

Note that this condition ensures that the supremum functional be measurable, and that separability is redundant if X is continuous.

An important role in Part A of this thesis is played by the tail of a standard normal random variable. This tail is denoted by Ψ:

$$\Psi(x) := \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{1}{2}w^2} dw, \tag{3.1}$$

and it is standard that, for $x \to \infty$,

$$\Psi(x) \sim \frac{1}{\sqrt{2\pi}x} e^{-x^2/2}. \tag{3.2}$$

In particular, we have the logarithmic tail asymptotics $\log \Psi(x) \sim -x^2/2$.

Gaussian (fluid) queues arise from the situation that the amount of traffic arriving at a buffer in the interval $[0, t]$ is given by $A_t = Y_t + \mu t$, where Y is a centered Gaussian process. A more detailed description of this fluid model can be found in Section 1.1. To ensure stability, the buffer is drained at some rate $r > \mu$. The case where Y has stationary increments is of special interest. All its finite-dimensional distributions are then specified by a single function, the *variance function* $\sigma^2 : t \mapsto \mathbb{V}\mathrm{ar} Y_t$, since we have for $s, t \geq 0$,

$$\mathbb{C}\mathrm{ov}(Y_s, Y_t) = \frac{1}{2}\left[\sigma^2(s) + \sigma^2(t) - \sigma^2(|s-t|)\right]. \tag{3.3}$$

An inherent conceptual problem of Gaussian queues is that the input process can be negative. However, in view of the discussion in Section 1.3.2, the steady-state buffer-content distribution for the Gaussian system can still be a good approximation for the corresponding distribution in a 'more realistic' model. This is comparable to the situation where the number of successes in n Bernoulli trials is approximated by a Gaussian random variable for large n; the distributions are somehow 'close' to each other, even though their respective supports do not agree.

In Part A of the thesis, we are interested in two aspects of Gaussian queues: the steady-state buffer-content distribution and the steady-state total length of the busy period. Note that these quantities can be analyzed by considering the free process, as explained in Section 1.1. To find their logarithmic tail asymptotics, it is useful to have large-deviation principles (LDPs) at our disposal. Therefore, we show in Section 3.1 how such LDPs can be proven. Two fundamental inequalities that are useful for analyzing the maximum of a Gaussian process are given in Section 3.2. In the last section of this chapter, Section 3.3, we illustrate a technique that can be used to find the exact tail asymptotics for the maximum of a stationary Gaussian process.

3.1 Gaussian measures and large-deviation principles

In this section, we present large-deviation principles for Gaussian measures, culminating in a powerful theorem due to Chevet [70]. Recall from Section 2.3 that these principles play a pivotal role in large-deviation theory, which can be used to study logarithmic tail asymptotics.

Before giving the LDPs, we first discuss a space that plays a key role in the definition of the rate function, the *Cameron-Martin space*, which is a special *reproducing kernel Hilbert space*. Let ν be a measure on some separable Banach space $\mathcal{X}$ equipped with its Borel σ-field. The measure ν is called a *Gaussian measure* if the image of ν under any continuous linear mapping $\xi : \mathcal{X} \to \mathbb{R}$ is a Gaussian distribution. More details can be found in the monographs by Bogachev [52], Kuo [205], Lifshits [219], and Üstünel [299].

Gaussian measures naturally arise from Gaussian processes. For instance, consider a Gaussian process $X = \{X_t : t \in \mathbf{T}\}$. If the paths $t \mapsto X_t$ are continuous and if $\mathbf{T}$ is compact,

one can take $\mathcal{X}$ as the space of continuous functions on $\mathbf{T}$ with the topology of uniform convergence. The law ν of X on $\mathcal{X}$ is then a Gaussian measure (for existence, see the proof of Lemma 4.7). Indeed, for any finite set $\{t_i\} \in \mathbf{T}$, the image of the continuous linear mapping $\xi : x \in \mathcal{X} \mapsto \sum_i \alpha_i x(t_i)$ has the same distribution as $\sum_i \alpha_i X(t_i)$, which is Gaussian by definition of a Gaussian process.

The covariance function $\Gamma : (s,t) \mapsto \mathbb{C}\mathrm{ov}(X_s, X_t)$ of X induces the space

$$S_\nu := \left\{ x : \mathbf{T} \to \mathbb{R} \text{ such that } x(\cdot) = \sum_{i=1}^{n} \alpha_i \Gamma(s_i, \cdot), \text{ for some } \alpha_i \in \mathbb{R}, s_i \in \mathbf{T}, n \in \mathbb{N} \right\}.$$

An inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_\nu}$ on S_ν is defined by

$$\left\langle \sum_{i=1}^{n} \alpha_i \Gamma(s_i, \cdot), \sum_{j=1}^{m} \beta_j \Gamma(t_j, \cdot) \right\rangle_{\mathcal{H}_\nu} := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \Gamma(s_i, t_j). \tag{3.4}$$

This inner product has the following unusual property: if $x(\cdot) = \sum_{i=1}^{n} \alpha_i \Gamma(s_i, \cdot)$, then

$$x(t) = \left\langle \sum_{i=1}^{n} \alpha_i \Gamma(s_i, \cdot), \Gamma(t, \cdot) \right\rangle_{\mathcal{H}_\nu} = \langle x, \Gamma(t, \cdot) \rangle_{\mathcal{H}_\nu}, \tag{3.5}$$

which is sometimes called the *reproducing kernel* property. It can be checked that (3.4) defines indeed an inner product, which induces a norm $\|f\|_{\mathcal{H}_\nu} = \sqrt{\langle f, f \rangle_{\mathcal{H}_\nu}}$. In particular, by (3.5) and the Cauchy-Schwarz inequality, for any $t \in \mathbf{T}$,

$$|f_n(t) - f_m(t)| \leq \|f_n - f_m\|_{\mathcal{H}_\nu} \sqrt{\mathbb{V}\mathrm{ar} X_t},$$

implying that Cauchy sequences in S_ν converge pointwise. This motivates the following definition.

Definition 3.1 (Cameron-Martin space) *The function $x \in \mathcal{X}$ belongs to the* Cameron-Martin space *of ν, denoted by $\mathcal{H}_\nu$, if there exists a Cauchy sequence in S_ν with x as its pointwise limit.*

The definition of $\langle \cdot, \cdot \rangle_{\mathcal{H}_\nu}$ can be extended from S_ν to $\mathcal{H}_\nu$ using the polarization formula

$$\langle x, y \rangle_{\mathcal{H}_\nu} = \frac{1}{2} \left[\|x\|^2_{\mathcal{H}_\nu} + \|y\|^2_{\mathcal{H}_\nu} - \|x - y\|^2_{\mathcal{H}_\nu} \right].$$

The reproducing kernel property (3.5) then continues to hold for any $x \in \mathcal{H}_\nu$. Importantly, this construction makes $\mathcal{H}_\nu$ a Hilbert space.

We will shortly see that the Cameron-Martin space plays a crucial role in large-deviation theory for Gaussian processes, but this is not the only setting where it is useful. In fact, the space is closely related to changes of measures, see for instance Bogachev [52, Cor. 2.4.3]. To gain intuition for this space, the reader may consult Adler [3], Chover and Feldman [74], or Parzen [250].

When reasoning intuitively about Cameron-Martin spaces and large deviations for Gaussian processes, one should keep the following caveat in mind. If $\mathcal{X} = \mathcal{H}_\nu = \mathbb{R}^d$, one has $\nu(\mathcal{H}_\nu) = 1$, but if $\mathcal{H}_\nu$ is infinite-dimensional, one has $\nu(\mathcal{H}_\nu) = 0$; see [52, Thm. 2.4.7].

We now turn to large deviations for Gaussian processes. The most well-known result in this framework is *(generalized) Schilder's theorem*, which states that the LDP holds for the empirical mean of i.i.d. copies of a centered Gaussian measure ν. More precisely, given a

sequence $\{a_n\}$ tending to ∞, Schilder's theorem states that the family $\{\nu(a_n\cdot)\}$ satisfies the LDP with the scale sequence $\{a_n^2\}$ and good rate function

$$I(x) = \begin{cases} \frac{1}{2}\|x\|_{\mathcal{H}_\nu}^2 & \text{if } x \in \mathcal{H}_\nu; \\ \infty & \text{otherwise.} \end{cases} \tag{3.6}$$

This result goes back to Azencott [32] and Bahadur and Zabell [34]; more details can be found in Deuschel and Stroock [104] or Lifshits [219]. It shows that, from a large-deviation point of view, the Cameron-Martin space is the analogue of the space of absolutely continuous functions in Mogul′skiĭ's theorem (Proposition 2.18).

Schilder's theorem is a special case of the following theorem, which is Theorem 2 of Chevet [70]. For the meaning of $\nu_n \Rightarrow \nu$, we refer to Section 2.2.

Theorem 3.2 *Let ν, ν_n be centered Gaussian measures on $\mathcal{X}$, and let $\{a_n\}$ be a sequence of positive real numbers tending to ∞. If $\nu_n \Rightarrow \nu$ in $\mathcal{X}$, then $\{\nu_n(a_n\cdot)\}$ satisfies the LDP with the good rate function I and scale sequence $\{a_n^2\}$, where I is the good rate function associated with Schilder's theorem for $\{\nu(a_n\cdot)\}$.*

Informally, Theorem 3.2 states that the families $\{\nu(a_n\cdot)\}$ and $\{\nu_n(a_n\cdot)\}$ have the same large-deviation behavior if ν_n converges weakly to ν.

3.2 Two fundamental inequalities

In this section, we present two inequalities related to extremes of Gaussian processes. The first intuitively entails that the maximum of a Gaussian process behaves like a single Gaussian variable with mean equal to the largest mean achieved by the entire process, and similarly for its variance. The second provides a way to compare (the distributions of) the extremes of two centered Gaussian processes with equal variances; if one process is 'more correlated' than the other, then its maximum is (stochastically) smaller.

The following lemma is Theorem D.1 of [257]. We refer to it as Borell's inequality, but it is also due to Cirel′son *et al.* [77].

Lemma 3.3 (Borell) *Let $\{X_t : t \in \mathbf{T}\}$ be a separable Gaussian process. Suppose that*

$$\sigma_{\mathbf{T}}^2 := \sup_{t\in\mathbf{T}} \mathbb{V}\mathrm{ar} X_t < \infty, \quad m_{\mathbf{T}} := \sup_{t\in\mathbf{T}} \mathbb{E}X_t < \infty,$$

and that, for some a,

$$\mathbb{P}\left(\sup_{t\in\mathbf{T}} X_t - \mathbb{E}X_t \geq a\right) \leq \frac{1}{2}.$$

Then we have for all $x \in \mathbb{R}$,

$$\mathbb{P}\left(\sup_{t\in\mathbf{T}} X_t > x\right) \leq 2\Psi\left(\frac{x - m_{\mathbf{T}} - a}{\sigma_{\mathbf{T}}}\right). \tag{3.7}$$

Note that this lemma provides no information for $x < m_{\mathbf{T}} + a$. Borell's inequality is closely related to the so-called isoperimetric inequality for the standard Gaussian measure; see Theorem 2.5 of Ledoux [213], Theorem 2.1 of Li and Shao [218], or Ledoux and Talagrand [214].

Sometimes the following variant of Lemma 3.3 is called Borell's inequality, as in Theorem 2.1.1 of Adler and Taylor [4]. If $\{X_t : t \in \mathbf{T}\}$ is a centered, almost surely bounded Gaussian process, then $\mathbb{E}\sup_{t\in\mathbf{T}} X_t < \infty$, and for all $x > 0$,

$$\mathbb{P}\left(\sup_{t\in\mathbf{T}} X_t - \mathbb{E}\sup_{t\in\mathbf{T}} X_t > x\right) \leq e^{-\frac{1}{2}x^2/\sigma_{\mathbf{T}}}.$$

This inequality is Proposition 2.18 in [213], and we refer the reader to this monograph for the exact relationship between the two inequalities.

Sometimes, we need sharper bounds than Borell's inequality, mainly because the presence of a (or $\mathbb{E}\sup_{t\in\mathbf{T}} X_t$) is undesirable in (3.7). Such inequalities (under more restrictive assumptions) are obtained by Berman [40], Fernique (see [211, Lem. 12.2.1]), and Piterbarg [257, Thm. D.4]. See also Section 4.1 in [4].

We now turn to the second inequality, which compares the extremes of two Gaussian processes with the same variances and different correlations. A proof can be found in [257, Thm. C.1] or [4, Thm. 2.2.1].

Lemma 3.4 (Slepian) *Let $\{X_t : t \in \mathbf{T}\}$ and $\{Y_t : t \in \mathbf{T}\}$ be separable Gaussian processes. If $\mathbb{E}X_t = \mathbb{E}Y_t$ and the covariance functions satisfy*

$$\mathbb{V}\mathrm{ar}X_t = \mathbb{V}\mathrm{ar}Y_t, \quad \mathbb{V}\mathrm{ar}(X_s - X_t) \le \mathbb{V}\mathrm{ar}(Y_s - Y_t), \quad s,t \in \mathbf{T},$$

then for any x,

$$\mathbb{P}\left(\sup_{t\in\mathbf{T}} X_t > x\right) \le \mathbb{P}\left(\sup_{t\in\mathbf{T}} Y_t > x\right).$$

A related inequality, the Sudakov-Fernique inequality (Theorem 2.2.3 of [4]), does not require equal variances, but only allows the comparison of means rather than probabilities.

3.3 The double sum method

In this section, we discuss one of the available tools to study tail asymptotics for Gaussian processes, the *double sum method*. It serves as a stepping stone for the analysis in Chapter 5. The presentation closely follows Piterbarg [257] (see also [259]).

The double sum method is best explained by working through an example. Suppose we are interested in determining the asymptotics (as $u \to \infty$) of the probability

$$\mathbb{P}\left(\sup_{t\in[0,S]} X_t > u\right),$$

where X is a centered standardized stationary Gaussian process with covariance function $R : t \mapsto \mathbb{C}\mathrm{ov}(X_0, X_t)$ satisfying $R(t) = 1 - |t|^\alpha + o(|t|^\alpha)$ for some $\alpha \in (0,2]$ as $t \downarrow 0$, and $R(t) < 1$ for all $t > 0$. This is Pickands' classical example [252, 253].

For a centered separable Gaussian process η with stationary increments and variance function $\sigma_\eta^2(\cdot) := \mathbb{V}\mathrm{ar}\,\eta_\cdot$, we define the so-called *Pickands' constant*

$$\mathcal{H}_\eta := \lim_{T\to\infty} \frac{1}{T}\mathcal{H}_\eta(T) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\exp\left(\sup_{t\in[0,T]}\left[\sqrt{2}\eta_t - \sigma_\eta^2(t)\right]\right), \tag{3.8}$$

provided both the mean and the limit exist. Depending on the context, we also write $\mathcal{H}_{\sigma_\eta^2}$ for $\mathcal{H}_\eta$. If η is a fractional Brownian motion B_H with Hurst parameter $H \in (0,1)$, i.e., $\sigma_\eta^2(t) = t^{2H}$, the constant is strictly positive (in particular, it exists). In the present generality, Pickands' constant has been introduced by Dębicki [87], and the field analogue shows up in the study of Gaussian fields; see Piterbarg [257].

In this section, we sketch the proof of the following theorem.

Theorem 3.5 *Consider a stationary process as described above. Then for any $S > 0$, as $u \to \infty$,*

$$\mathbb{P}\left(\sup_{t\in[0,S]} X_t > u\right) \sim S\mathcal{H}_{B_{\alpha/2}} u^{2/\alpha}\Psi(u).$$

The first step in its proof is an important lemma in the theory of asymptotics for Gaussian processes. It is often called Pickands' lemma; see Lemma D.1 of Piterbarg [257]. A proof is omitted, as we prove a more general version in Lemma 5.9.

Lemma 3.6 *For any $T > 0$, as $u \to \infty$,*

$$\mathbb{P}\left(\sup_{t\in[0,Tu^{-2/\alpha}]} X_t > u\right) \sim \mathcal{H}_{B_{\alpha/2}}(T)\Psi(u).$$

Write $I_k^T(u) := [kTu^{-2/\alpha}, (k+1)Tu^{-2/\alpha}]$. The 'upper bound' part of Theorem 3.5 uses Pickands' lemma and exploits the stationarity of X:

$$\mathbb{P}\left(\sup_{t\in[0,S]} X_t > u\right) \le \frac{S}{Tu^{-2/\alpha}}\mathbb{P}\left(\sup_{t\in I_0^T(u)} X_t > u\right) \sim \frac{S}{T}\mathcal{H}_{B_{\alpha/2}}(T)u^{2/\alpha}\Psi(u),$$

yielding the upper bound upon letting $T \to \infty$.

More work is needed to prove the 'lower bound' part of Theorem 3.5. For this, the idea is to use the following inequality:

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,S]} X_t > u\right) \quad \ge \quad & \frac{S}{Tu^{-2/\alpha}}\mathbb{P}\left(\sup_{t\in I_0^T(u)} X_t > u\right) \\ & - \sum_{0\le k\le Su^{2/\alpha}/T} \mathbb{P}\left(\sup_{t\in I_k^T(u)} X_t > u, \sup_{t\in[0,S]\setminus I_k^T(u)} X_t > u\right).\end{aligned}$$

As the first term has the 'right' asymptotic behavior as $u \to \infty$ and then $T \to \infty$, it suffices to show that the second term is $o(u^{2/\alpha}\Psi(u))$ as $u \to \infty$ and then $T \to \infty$. The k-th summand does not exceed

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in I_k^T(u)} X_t > u, \sup_{t\in[0,(kT-\sqrt{T})u^{-2/\alpha})\cup(((k+1)T+\sqrt{T})u^{-2/\alpha},S]} X_t > u\right) \\ &+ 2\mathbb{P}\left(\sup_{t\in[0,\sqrt{T}u^{-2/\alpha}]} X_t > u\right),\end{aligned}$$

and it follows again from Lemma 3.6 that the last term is indeed small enough (note that $\lim_{T\to\infty} \mathcal{H}_{B_{\alpha/2}}(\sqrt{T})/T = 0$).

Therefore, it suffices to study

$$\begin{aligned}&\sum_{0\le k<\ell\le Su^{2/\alpha}/T} 2\mathbb{P}\left(\sup_{t\in I_k^T(u)} X_t > u, \sup_{t\in I_\ell^T(u)} X_t > u\right) \\ &+ \quad 2\frac{S}{Tu^{-2/\alpha}}\mathbb{P}\left(\sup_{t\in I_0^T(u)} X_t > u, \sup_{t\in[(T+\sqrt{T})u^{-2/\alpha},(2T+\sqrt{T})u^{-2/\alpha}]} X_t > u\right). \qquad (3.9)\end{aligned}$$

The double sum in this expression (i.e., the sum over k, ℓ) explains the name of the method. The summands are estimated from above by

$$\begin{aligned}&\mathbb{P}\left(\sup_{(s,t)\in I_k^T(u)\times I_\ell^T(u)} X_s + X_t > 2u\right)\\ &\quad\leq \mathbb{P}\left(\sup_{(s,t)\in I_k^T(u)\times I_\ell^T(u)} \frac{X_s+X_t}{\sqrt{\mathbb{V}\mathrm{ar}(X_s+X_t)}} > \frac{2u}{\sqrt{4-\inf_{(s,t)\in I_k^T(u)\times I_\ell^T(u)}\mathbb{V}\mathrm{ar}(X_s-X_t)}}\right).\end{aligned}$$

We now briefly indicate how the proof of Theorem 3.5 is completed, and refer to [257] for details.

It is possible to compare the covariance structure of the standardized random field $(X_s + X_t)/\sqrt{\mathbb{V}\mathrm{ar}(X_s + X_t)}$ to the one of $\sqrt{2}[\theta_1(s) + \theta_2(t)]$, where the θ_i are two independent standardized stationary processes with a certain covariance structure. With Slepian's inequality (Lemma 3.4), this yields an upper bound for which the asymptotics can be found with a field version of Lemma 3.6. A careful analysis of the behavior of

$$\frac{2u}{\sqrt{4-\inf_{(s,t)\in I_k^T(u)\times I_\ell^T(u)}\mathbb{V}\mathrm{ar}(X_s-X_t)}}$$

for large u and small $|k - \ell|$ is then needed to derive a further upper bound on the double sum. Borell's inequality (Lemma 3.3) shows that large values of $|k - \ell|$ do not contribute to the double sum. Again, the argument is completed by sending first $u \to \infty$ and then $T \to \infty$.

The second term in (3.9) is bounded in a similar way; this finishes the 'proof' of Theorem 3.5.

Pickands' constant and alternative approaches

Pickands' constant $\mathcal{H}_\eta$ introduced in (3.8) appears in many results on tail asymptotics for Gaussian processes, sometimes in an alternative form. However, not so much is known about the values of the constant; in fact, $\mathcal{H}_\eta$ is only known if η is a Brownian motion ($\mathcal{H}_{B_{1/2}} = 1$), a degenerate Gaussian process ($\mathcal{H}_{B_1} = 1/\sqrt{\pi}$), or an integrated Gaussian process with $\sigma_\eta^2(u) \sim \mathcal{G}u$ for some $\mathcal{G} \in (0, \infty)$ (see Kobelkov [193]).

There are some qualitative properties known about the constant. Most notably in this respect is Dębicki [88], who proves that $\mathcal{H}_{B_H}$ is continuous as a function of H. The papers by Shao [287] and Dębicki *et al.* [94] show that it is even nontrivial (yet possible) to bound Pickands' constant.

Different approaches to extremes of Gaussian processes often lead to alternative descriptions of Pickands' constant. Apart from the double sum method, we mention the method of refined grids, which is based on extremes for discrete-time stochastic processes, see Leadbetter *et al.* [211, Ch. 12]. In this context, we also refer to Hüsler [162], who uses triangular arrays to interpret Pickands' constant as a clustering index.

Another approach is based on high-level crossings. Due to sample-path irregularities, some care is needed to define such crossings, see Piterbarg [257]. For Gaussian processes that are sufficiently smooth, the *Rice formula* is an indispensable tool; recent contributions include Azaïs and Wschebor [31] and Kobelkov [193]. Interestingly, high-level crossings occur approximately according to a Poisson process; see, for instance, Leadbetter *et al.* [211, Sec. 13.6] and Piterbarg and Stamatovic [261].

This Poissonian character connects high-level crossings to Berman's sojourn approach [41]. Aldous [8] explains this heuristically and also gives intuition behind other fundamental results in extreme-value theory.

Finally, we mention a Laplace-transform approach to extremes; see Lifshits [219, Sec. 13]. Although the underlying assumptions are mild, a drawback of this method is that the tail asymptotics are expressed in terms of the asymptotic behavior of an unknown Laplace transform.

3.4 Outline of Part A

In Chapter 4, we study a centered Gaussian process Y with stationary increments. Under some assumptions on the variance function $\sigma^2(t) := \mathbb{V}\mathrm{ar}Y_t$ of Y, we prove the convergence in distribution (see Definition 2.9) of $Y_{\alpha t}/\sigma(\alpha)$ in an appropriate space. With Theorem 3.2, this immediately yields a number of LDPs, which are useful to derive two conditional limit theorems for Gaussian queues. The first conditional limit theorem describes the behavior of Y if the steady-state buffer content exceeds u; the logarithmic asymptotics of the corresponding tail probability are also found. The second deals with the length of the steady-state busy period in a Gaussian queue. Again, we also establish the logarithmic tail asymptotics.

Chapter 5 further investigates the tail asymptotics for the steady-state buffer content. As opposed to Chapter 4, where the logarithmic asymptotics are found as a corollary of an LDP, Chapter 5 focuses on *exact* asymptotics. Using a variant of the double sum method of Section 3.3, we also relax the assumption that the increments of Y be stationary. The case where Y is self-similar (see Section 1.3) is worked out in detail.

In Chapter 6, we apply the results of Chapter 5 to study *reduced-load equivalence*. That is, we consider a Gaussian queue for which the input A can be written as a sum of M (independent) Gaussian processes. We present a necessary and sufficient condition for a subset S of these M Gaussian processes to dominate the tail behavior of the steady-state buffer-content distribution. This analysis relies extensively on de Haan theory, as introduced in Section 2.1.

The results of Chapter 4, 5, and 6 have been published as [106], [107], and [108] respectively.

CHAPTER 4

Conditional limit theorems

In this chapter, we study a fluid queue fed by a stationary Gaussian source. By proving conditional limit theorems, we investigate how a high buffer level is typically achieved. The underlying large-deviation analysis also enables us to establish the logarithmic tail asymptotics for the buffer content. In addition, we study how a long busy period typically occurs, and we find the corresponding logarithmic tail asymptotics.

The study relies on weak convergence in an appropriate space of $\{Y_{\alpha t}/\sigma(\alpha) : t \in \mathbb{R}\}$ to a fractional Brownian motion with Hurst parameter H as $\alpha \to \infty$. Here Y is a separable centered Gaussian process with stationary increments. We assume that its variance function $\sigma^2 : t \mapsto \mathbb{V}\mathrm{ar}Y_t$ is regularly varying with index $2H$, for some $0 < H < 1$.

We prove this weak convergence under a fairly general condition on σ^2, sharpening recent results of Kozachenko *et al.* [201]. The core of the proof is a new uniform convergence theorem for regularly varying functions with positive index.

4.1 Introduction

When studying a buffered queueing system, one is often interested in the following two questions:

Q1: How is a high buffer level achieved?

Q2: If the buffer is nonempty for a long time, how does this event occur?

This chapter considers these questions for a fluid queue with stationary Gaussian input.

There are good reasons to investigate the above questions in a Gaussian framework. First, Gaussian processes can model both short-range dependence (as in, e.g., an integrated Ornstein-Uhlenbeck process) and long-range dependence (as in, e.g., a fractional Brownian motion with Hurst parameter exceeding 1/2). This flexibility is particularly relevant in view of the above questions, since fundamentally different answers can be expected in the short-range and long-range dependent case. Moreover, as argued in Section 1.3.2, the normality assumption is motivated by a central limit-type result. A third, pragmatic reason to study Gaussian processes is that a vast body of literature facilitates their investigation.

The behavior of a queue conditioned on the occurrence of a rare event has been studied in different contexts. Hooghiemstra [160] studies waiting times if a long busy period occurs,

and obtains in the (weak) limit a Brownian excursion. A different type of limit theorem is found by Anantharam [10]. He studies how a queue must have evolved when the waiting time has become large. By underlying independence assumptions, this occurs by 'staying close' to a piecewise linear path. We also mention Bertoin and Doney [46], who focus on the initial behavior of a random walk conditioned to stay nonnegative.

In the literature, the typical behavior of a queue has already been studied in connection to the first question Q1. Due to the close relationship between queueing processes and risk processes (see Section 1.1.5), it is equivalent to ask how ruin occurs in the corresponding risk model. In a compound Poisson setting, this is addressed in Chapter IV.7 of Asmussen [18] (see also [13]). It turns out that the path is linear for this risk process (and similarly for processes with a weak dependence, see Nyrhinen [245]). However, as a result of possible correlations in the system input, this need not be the case in the Gaussian setting.

Although the second question Q2 has not been investigated explicitly in the literature, there is some related work. In queueing language, the question deals with the length of the steady-state busy period. Norros [244] considers a queue with fractional Brownian motion input, and studies the probability that the length of the busy period exceeds T as $T \to \infty$. He formulates a variational problem for which the solution determines the logarithmic asymptotics. This result is generalized by Kozachenko *et al.* [201], who also allow for other Gaussian input processes. In the present chapter, we considerably widen the class of Gaussian processes for which these logarithmic asymptotics are valid.

As answers to the above questions, we provide two conditional limit theorems. As for Q1, we identify a path x^* such that, under a certain condition, the (scaled) distribution of the Gaussian process Y given that the buffer content reaches a high level u converges (in a sense that will be made precise) to a Dirac mass δ_{x^*} at x^*. In other words, even though the probability that the buffer content exceeds u becomes increasingly rare as $u \to \infty$, it becomes easier to predict the behavior of a (scaled) sample path of Y. Formally, we prove that for every regular set of paths A, as $u \to \infty$,

$$\mathbb{P}\left(\left\{\frac{1}{u}Y_{ut} : t \in \mathbb{R}\right\} \in A \,\middle|\, \sup_{t\geq 0} Y_t - t \geq u\right) \to \begin{cases} 1 & \text{if } x^* \in A; \\ 0 & \text{otherwise.} \end{cases}$$

A similar conditional limit theorem is given for the busy-period problem.

A weak convergence approach

In order to explain the contributions of this chapter, we need to formalize our framework.

Let Y denote a centered separable Gaussian process with stationary increments. The central assumption is that the *variance function* $\sigma^2 : t \in \mathbb{R} \to \mathbb{V}\mathrm{ar}Y_t \in [0, \infty)$ is continuous and regularly varying with index $2H$ for some $0 < H < 1$, i.e.,

$$\lim_{\alpha\to\infty} \frac{\sigma^2(\alpha t)}{\sigma^2(\alpha)} = |t|^{2H}.$$

Note that the function σ^2 characterizes the finite-dimensional distributions of Y.

Interestingly, by a powerful theorem of Chevet (see Theorem 3.2), the proofs of the conditional limit theorems rely only on a (sufficiently strong) type of weak convergence of the processes Y^α as $\alpha \to \infty$, with

$$Y_t^\alpha := \frac{Y_{\alpha t}}{\sigma(\alpha)}.$$

We now precisely describe the type of weak convergence that we show; it is explained in Section 4.3.3 that other types of weak convergence are not strong enough to provide satisfactory

answers to the above questions. For $\gamma \geq 0$, set

$$\Omega^\gamma := \left\{ x : \mathbb{R} \to \mathbb{R} \text{ such that } x \text{ continuous}, x(0) = 0, \lim_{t \to \pm\infty} \frac{x(t)}{1 + |t|^\gamma} = 0 \right\},$$

and equip Ω^γ with the topology generated by the norm

$$\|x\|_{\Omega^\gamma} := \sup_{t \in \mathbb{R}} \frac{|x(t)|}{1 + |t|^\gamma},$$

under which Ω^γ is a separable Banach space. Endow Ω^γ with the Borel σ-field induced by this topology, denoted by $\mathcal{B}(\Omega^\gamma)$. As pointed out in Section 4.4.2, under the condition $\lim_{t \to 0} \sigma^2(t) |\log |t||^{1+\epsilon} < \infty$ for some $\epsilon > 0$, Y^α takes almost surely values in Ω^γ for $\gamma > H$ and the law of Y^α in $(\Omega^\gamma, \mathcal{B}(\Omega^\gamma))$ exists; it is denoted by ν_α^γ. Hence, it is legitimate to ask whether ν_α^γ has a weak limit for $\alpha \to \infty$.

By considering the finite-dimensional distributions, it is readily seen that the only candidate weak limit is the law $\mathcal{L}(B_H)$ in Ω^γ of a fractional Brownian motion B_H with Hurst parameter H. Recall that a fractional Brownian motion B_H is a continuous centered Gaussian process with stationary increments and variance function $\mathbb{V}\mathrm{ar} B_H(t) = |t|^{2H}$; for $H = 1/2$, it reduces to ordinary Brownian motion. We write $Y^\alpha \Rightarrow B_H$ and $\nu_\alpha^\gamma \Rightarrow \mathcal{L}(B_H)$ for convergence in distribution and weak convergence respectively, cf. Section 2.2; when this notation is used, we also specify the space (and topology) in which this convergence takes place.

Comparison with previous results

Conditions for the weak convergence of Y^α in Ω^1 have been derived by Kozachenko *et al.* [201]. Their conditions are based on the *majorizing variance*

$$\overline{\sigma}^2(t) = \sup_{0 < s < t} \sup_{\alpha \geq 1} \frac{\sigma^2(\alpha s)}{\sigma^2(\alpha)}. \tag{4.1}$$

Unfortunately, apart from some special cases, $\overline{\sigma}^2(t)$ is difficult to bound or compute.

By taking a different approach than Kozachenko *et al.* [201], we show that Y^α converges weakly to B_H in Ω^γ for $\gamma > H$ under the *same* condition that we use to guarantee the existence of ν_α^γ: $\lim_{t \to 0} \sigma^2(t) |\log |t||^{1+\epsilon} < \infty$ for some $\epsilon > 0$. This not only relaxes the condition in Proposition 2.9 of [201], but is also easier to check. As in [201], we rely on metric entropy techniques. However, we first exploit the regular variation of the variance function before applying these techniques. Specifically, we present a new type of uniform convergence theorem for regularly varying functions with positive index.

To illustrate the advantage of the condition developed in this chapter, consider the situation that the process Y is the superposition of a finite number m of independent Gaussian process with stationary increments. The variance functions of the m individual Gaussian processes are denoted by $\sigma_1^2, \ldots, \sigma_m^2$, and the σ_i are assumed to be regularly varying with index $H_i \in (0, 1)$. The variance function $\sigma^2 = \sum_i \sigma_i^2$ of Y is then regularly varying with index $2 \max_i H_i$, but it is in general impossible to compute the majorizing variance (4.1). In contrast, $\lim_{t \to 0} \sigma^2(t) |\log |t||^{1+\epsilon} < \infty$ for some $\epsilon > 0$ if and only if the same is true for the *individual* variance functions $\sigma_1^2, \ldots, \sigma_m^2$. This situation is further investigated in Chapter 6; see also Ganesh and Wischik [141].

The outline of the chapter is as follows. In Section 4.2, we introduce the two queueing problems in more detail, and we state the theorems that provide answers to the above questions. Section 4.3 provides preliminaries on notions that are crucial in the proofs of these theorems, including the new uniform convergence result for regularly varying functions (proven

in Section 4.6). The convergence in distribution of Y^α to a fractional Brownian motion is the subject of Section 4.4; we present both a necessary and a sufficient condition. With these weak-convergence results at our disposal, the proofs of the claims in Section 4.2 are given in Section 4.5.

4.2 Queueing results

In this section, we present the two conditional limit theorems that serve as answers to the two questions raised in the introduction. As indicated there, a key role in the proofs of the results is played by the convergence in distribution of Y^α to B_H in Ω^γ. Since this convergence is the subject of Section 4.4, we defer all proofs for the present section to Section 4.5.

4.2.1 Conditional limit theorem for high buffer level

Before presenting the announced conditional limit theorem, it is insightful to first have a closer look at the probability

$$\mathbb{P}\left(\sup_{t\geq 0} Y_t - ct^\beta \geq u\right) \tag{4.2}$$

for $\beta > H$ and $c > 0$, as $u \to \infty$. In case $\beta = 1$, this probability equals the steady-state probability that the buffer content exceeds u when the input process A is given by $A_t = Y_t + \mu t$ the drain rate r is $c + \mu$; this situation is described in Section 1.1. Since we allow $\beta > H$, we analyze the problem slightly more generally. We note that studying $u \to \infty$ is known as considering the *large-buffer* asymptotic regime.

There exists a vast body of literature dealing with the *logarithmic* asymptotics of (4.2), under different levels of generality (Duffield and O'Connell [123], Dębicki *et al.* [93], and Kozachenko *et al.* [201]). An important contribution in this setting was made by Dębicki [85], who establishes the logarithmic asymptotics for $\beta = 1$ under the technical requirement that $\lim_{u\to\infty} \mathbb{P}(\sup_{t\geq 0} Y_t - \epsilon t > u) = 0$ for $\epsilon > 0$. However, this condition is automatically satisfied in case Y has stationary increments, since $Y_t/t \to 0$ almost surely (see Lemma 4.13). In [85] it is also assumed that σ^2 increases, but this assumption can be avoided by invoking the uniform convergence theorem (UCT) for regularly varying functions (Theorem 2.3) in Lemma 3.1. Hence, only assuming continuity of the sample paths of Y and regular variation of the variance function suffice to establish the logarithmic asymptotics of (4.2).

We remark that the *exact* asymptotics of (4.2) have been studied extensively in the past few years. Recall from Section 2.3 that exact asymptotics are more powerful than logarithmic asymptotics. Since exact asymptotics are discussed in Chapter 5, we refer to that chapter for background and references.

We now return to logarithmic asymptotics. It was already noted that these are known to hold under the condition that Y has continuous sample paths. However, the proof given in this chapter relies on the weak convergence of the processes Y^α, and we therefore require the (stronger) condition $\lim_{t\to 0} \sigma^2(t)|\log|t||^{1+\epsilon} < \infty$ for some $\epsilon > 0$; see Section 4.5.

Proposition 4.1 *If* $\lim_{t\to 0} \sigma^2(t)|\log|t||^{1+\epsilon} < \infty$ *for some* $\epsilon > 0$*, then for* $\beta > H$*,*

$$\lim_{u\to\infty} \frac{\sigma^2(u^{1/\beta})}{u^2} \log \mathbb{P}\left(\sup_{t\geq 0} Y_t - ct^\beta \geq u\right) = -\frac{1}{2} c^{2H/\beta} \left(\frac{H}{\beta - H}\right)^{-2H/\beta} \left(\frac{\beta}{\beta - H}\right)^2.$$

One of the advantages of using the weak convergence approach is that one can analyze the large deviations on a path level. This large-deviation study yields a path x^* that can be

interpreted as the 'most likely' path. In the setting of this subsection, the path is given by

$$\begin{aligned} x^*(t) &= \frac{1+(t^*)^\beta}{2(t^*)^{2H}}\mathbb{C}\mathrm{ov}(B_H(t), B_H(t^*)) \\ &= \frac{\beta}{2(\beta-H)}\left(1+\left|\frac{t}{t^*}\right|^{2H} - \left|1-\frac{t}{t^*}\right|^{2H}\right), \end{aligned} \tag{4.3}$$

for $t \in \mathbb{R}$, where $t^* = (H/[\beta - H])^{1/\beta}$; see Addie *et al.* [2] for a detailed derivation, and Wischik [308] for a general approach. We now formalize the intuition that x^* (suitably scaled) is the most likely trajectory of Y when $Y_t - ct^\beta$ reaches u. For convenience, the law of the stochastic process $Z^{(u)}$ given the event A_u is said to be the law of $Z^{(u)}|A_u$.

Theorem 4.2 *If* $\lim_{t\to 0}\sigma^2(t)|\log|t||^{1+\epsilon} < \infty$ *for some* $\epsilon > 0$*, then for any* $\beta > H$*, the law of*

$$\frac{\sigma([u/c]^{1/\beta})}{u} Y^{(u/c)^{1/\beta}} \,\Big|\, \sup_{t\geq 0} Y_t - ct^\beta \geq u$$

converges weakly in Ω^β *to the Dirac measure* δ_{x^*} *at* x^* *as* $u \to \infty$*.*

Note that the weak convergence stated in the theorem implies, for instance, that for any $\eta > 0$, $\beta > H$,

$$\lim_{u\to\infty} \mathbb{P}\left(\sup_{t\in\mathbb{R}} \frac{\left|\frac{1}{u}Y_{u^{1/\beta}t} - x^*(c^{1/\beta}t)\right|}{1+c|t|^\beta} \geq \eta \,\Big|\, \sup_{t\geq 0} Y_t - ct^\beta \geq u\right) = 0.$$

For $\beta = 1$, the most likely time epoch for $Y_t - ct$ to hit u is ut^* and hence linear in u. Interestingly, according to Theorem 4.2, if $Y_t - ct^\beta$ reaches u, $\frac{\sigma([u/c]^{1/\beta})}{u} Y^{(u/c)^{1/\beta}}$ is typically 'close' to x^*, which is only a straight line when $H = 1/2$. See Addie *et al.* [2] for 'most likely paths' in the *many-sources* asymptotic regime; Chapter 10 addresses this regime in the context of simulation.

4.2.2 Conditional limit theorem for the length of a busy period

In this subsection, we gain some insight into the steady-state distribution of the length of a so-called busy period. We start by introducing some notation.

For $x \in \Omega^1$, define the function $s : \Omega^1 \to \mathbb{R}^{\mathbb{R}}$ as

$$s(x)(t) := \sup_{s\leq t} x(t) - x(s) - (t-s). \tag{4.4}$$

While $x(t)$ (or $-x(-t)$) represents the amount of work arriving in the interval $[0,t]$ (or $[-t,0]$) for $t \geq 0$, $s(x)(t)$ can be thought of as the amount of work in the buffer at time t when the system 'input' is x, cf. Section 1.1.3. We set

$$t_-(x) := \sup\{t \leq 0 : s(x)(t) = 0\}, \quad t_+(x) := \inf\{t \geq 0 : s(x)(t) = 0\},$$

i.e., $t_-(x)$ ($t_+(x)$) is the last (first) time $s(x)$ hits zero before (after) time zero. We say that zero is contained in a *busy period*, since an imaginary server is constantly draining the buffer during the time interval $[t_-(x), t_+(x)]$.

Note that $t_-(x)$ has the following alternative interpretation in terms of x. By definition, $t_-(x)$ is the largest $t \leq 0$ such that $\sup_{s\leq t} s - x(s) = t - x(t)$, i.e., it is the epoch where $s - x(s)$ attains its maximum on $(-\infty, 0]$ (assuming its uniqueness for simplicity). In Chapter 13, this is made precise in a network context with the help of Skorokhod problems.

We write K_T for the set of paths in Ω^1 for which the busy period straddling zero is strictly longer than T, i.e.,

$$K_T := \{x \in \Omega^1 : t_-(x) < 0 < t_+(x), t_+(x) - t_-(x) > T\}.$$

It is our aim to find the logarithmic asymptotics of $\mathbb{P}(Y \in K_T)$ as $T \to \infty$.

Norros [244] considers this setting in case $Y = B_H$, and his results are generalized by Kozachenko *et al.* [201] to allow for more general input processes. The next proposition generalizes their findings. A key role in the result is played by a separable Hilbert space $\mathcal{H}_H$, the Cameron-Martin space associated with the law of B_H; see Definition 3.1. The norm induced by the inner product on $\mathcal{H}_H$ is denoted by $\|\cdot\|_{\mathcal{H}_H}$.

Proposition 4.3 *If* $\lim_{t\to 0} \sigma^2(t)|\log|t||^{1+\epsilon} < \infty$ *for some* $\epsilon > 0$*, then*

$$\lim_{T\to\infty} \frac{\sigma^2(T)}{T^2} \log \mathbb{P}(Y \in K_T) = -\frac{1}{2} \inf_{x\in K_1\cap\mathcal{H}_H} \|x\|^2_{\mathcal{H}_H}. \tag{4.5}$$

Let us stress the fact that $\mathcal{H}_H$ is the Cameron-Martin space associated with the law of B_H; the right-hand side of (4.5) does not depend on the specific form of σ^2, but only on its index of variation.

Proposition 4.3 can also be used to derive the logarithmic asymptotics in case $c(t-s)$ is subtracted from $x(t) - x(s)$ in the definition of $s(x)$ in (4.4). Equation (4.4) shows that we essentially replace the distribution of Y by the distribution of $\tilde{Y} = Y/c$. Evidently, the variance function $\tilde{\sigma}^2$ of $\tilde{Y}$ then equals $\tilde{\sigma}^2 = \sigma^2/c^2$. We conclude that the following logarithmic asymptotics apply:

$$\lim_{T\to\infty} \frac{\sigma^2(T)}{T^2} \log \mathbb{P}(\tilde{Y} \in K_T) = -\frac{c^2}{2} \inf_{x\in K_1\cap\mathcal{H}_H} \|x\|^2_{\mathcal{H}_H}.$$

The constant $\inf_{x\in K_1\cap\mathcal{H}_H} \|x\|^2_{\mathcal{H}_H}$ is generally difficult to identify, except for the case $H = 1/2$; in that case, it equals one. An expression for the path $\overline{x} \in \overline{K_1} \cap \mathcal{H}_H$ with $\inf_{x\in K_1\cap\mathcal{H}_H} \|x\|^2_{\mathcal{H}_H} = \|\overline{x}\|^2_{\mathcal{H}_H}$ in the complementary case $H \neq 1/2$ has been found recently by Mandjes *et al.* [224]. Even without this knowledge, it is possible to formulate the analogue of Theorem 4.2.

Theorem 4.4 *If* $\lim_{t\to 0} \sigma^2(t)|\log|t||^{1+\epsilon} < \infty$ *for some* $\epsilon > 0$*, then the law of*

$$\frac{\sigma(T)}{T} Y^T \,\Big|\, Y \in K_T$$

converges for $\gamma > H$ *weakly in* Ω^γ *to the Dirac measure* $\delta_{\overline{x}}$ *at* $\overline{x}$ *as* $T \to \infty$*.*

4.3 Preliminaries

In this section, we discuss a property of regularly varying functions and introduce the notion of metric entropy. We also address the topological issues raised in Section 4.1.

Before we start, we introduce the notation

$$C([-T,T]) := \{x : [-T,T] \to \mathbb{R} \text{ such that } x \text{ continuous}, x(0) = 0\},$$

and equip $C([-T,T])$ with the topology of uniform convergence, i.e., the topology generated by the norm $\|x\|_T := \sup_{t\in[-T,T]} |x(t)|$. Note that $C([-T,T])$ equipped with this topology is a separable Banach space. We write $\mathcal{B}(C([-T,T]))$ for the Borel σ-field on $C([-T,T])$ generated by the topology of uniform convergence.

4.3.1 A property of regularly varying functions

The results in this chapter, particularly those in Section 4.4, rely extensively on a property of regularly varying functions (for the definition and background, see Section 2.1) that is intimately related to the UCT (Theorem 2.3). To formulate this property, first define $L_\epsilon : \mathbb{R} \to [0, \infty)$ as

$$L_\epsilon(t) := \begin{cases} |\log|t||^{1+\epsilon} & \text{if } |t| \le 1/e; \\ 1 & \text{otherwise.} \end{cases} \tag{4.6}$$

Proposition 4.5 *Let f be regularly varying with index $\rho > 0$. If fL_ϵ is bounded on each interval $(0, \cdot]$ for some $\epsilon > 0$, then we have*

$$\lim_{\alpha \to \infty} \frac{f(\alpha t)}{f(\alpha)} L_\epsilon(t) = t^\rho L_\epsilon(t),$$

uniformly in t on each $(0, \cdot]$.

Proof. The proof is given in Section 4.6. □

Note that the requirement that fL_ϵ be bounded on intervals of the form $(0, \cdot]$ is equivalent to local boundedness of f and $\limsup_{t \downarrow 0} f(t)|\log t|^{1+\epsilon} < \infty$. Alternatively, one can replace the L_ϵ by another continuous positive function with the following two properties: on compact subsets of $(0, \infty)$, it is bounded away from zero and bounded from above, and near zero it is equivalent to L_ϵ. An example of such a function is $(\log(1 + 1/t))^{1+\epsilon}$.

4.3.2 Metric entropy

Metric entropy is an important tool in studying continuity and boundedness of trajectories of Gaussian processes, see for instance [3, 4, 52]. In order to introduce the main ideas of the concept, let X be a centered Gaussian process on a set $\mathbf{T} \subset \mathbb{R}$ and define the semimetric

$$d(s, t) := \sqrt{\mathbb{E}|X_s - X_t|^2}, \quad s, t \in \mathbf{T}.$$

For simplicity, we suppose that d is continuous on $\mathbf{T} \times \mathbf{T}$; in the context of the present chapter, this is guaranteed by the fact that σ^2 is continuous. We say that $S \subset \mathbf{T}$ is a ϑ-net in $\mathbf{T}$ with respect to the semimetric d, if for any $t \in \mathbf{T}$ there exists an $s \in S$ such that $d(s, t) \le \vartheta$.

Definition 4.6 *The* metric entropy $\mathbb{H}_d(\mathbf{T}, \vartheta)$ *is defined as* $\log N_d(\mathbf{T}, \vartheta)$*, where $N_d(\mathbf{T}, \vartheta)$ denotes the minimal number of points in a ϑ-net in $\mathbf{T}$ with respect to d.*

The quantity $\int_0^\infty \sqrt{\mathbb{H}_d(\mathbf{T}, \vartheta)} d\vartheta$ is called the Dudley integral.

If $\mathbf{T}$ is completely bounded with respect to d, then $\mathbb{H}_d(\mathbf{T}, \vartheta) = 0$ for ϑ large enough, so that the convergence of the Dudley integral is equivalent to its convergence at zero.

A useful fact is that X has an almost-surely continuous modification if the Dudley integral converges, see, e.g., Lemma 1.3.1 and Theorem 1.3.5 of Adler and Taylor [4], or Corollary 4.15 of Adler [3]. A simple sufficient condition for this is given in the next lemma; it is satisfied by many processes.

Lemma 4.7 *Let $\mathbf{T} = [-T, T]$ for some $T > 0$. If there exist $\epsilon, \kappa, C > 0$ such that for any $s, t \in \mathbf{T}$ with $|s - t| < \kappa$,*

$$\mathbb{E}|X_s - X_t|^2 \le \frac{C}{L_\epsilon(s - t)},$$

then there exists a probability measure ν on $(C([-T, T]), \mathcal{B}(C([-T, T])))$ such that for any finite sequence $\{t_1, \ldots, t_n\} \subset [-T, T]$ and sets $A_i \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}(X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) = \nu\left(x \in C([-T, T]) : x(t_1) \in A_1, \ldots, x(t_n) \in A_n\right).$$

Proof. We adopt the (mostly standard) terminology on stochastic processes (see, e.g., Revuz and Yor [271]). Without loss of generality, we may suppose that X is the coordinate mapping on the canonical probability space $(\mathbb{R}^{[-T,T]}, \mathcal{B}(\mathbb{R})^{[-T,T]}, \tilde{P})$, where $\mathcal{B}(\mathbb{R})$ denotes the usual Borel σ-field on $\mathbb{R}$, and the superscripts indicate that we deal with product spaces and product σ-fields. We refer to Section I.3 of [271] for more details.

Since the metric entropy $\mathbb{H}_d([-T,T],\vartheta)$ is upper bounded by $C_0\vartheta^{-\frac{2}{1+\epsilon}}$ for some constant $C_0 > 0$ and for $\vartheta > 0$ small, the Dudley integral converges and there exists a continuous modification of X. One can now construct a probability space $(C([-T,T]), \mathcal{B}(\mathbb{R})^{[-T,T]} \cap C([-T,T]), \nu)$ with the required property. The claim follows by noting that $\mathcal{B}(\mathbb{R})^{[-T,T]} \cap C([-T,T]) = \mathcal{B}(C([-T,T]))$ (see, e.g., Theorem VII.2.1 of Parthasarathy [249]). □

In Section 4.4.1, we establish tightness of a sequence of probability measures on $C([-T,T])$ in a similar way. However, instead of using the convergence of the Dudley integral, we then carefully derive upper bounds in order to obtain the desired uniformity. A key tool in this analysis is Proposition 4.5.

4.3.3 Topological issues

In this subsection, we motivate the choice for the space Ω^γ and its topology. As pointed out in the introduction, convergence in distribution of Y^α to B_H in Ω^γ is only useful in applications if the topology on Ω^γ is strong enough. For explanatory reasons, we suppose in this subsection that Y has continuous sample paths.

The most natural path space to work with is the space $C(\mathbb{R})$ of continuous functions on $\mathbb{R}$; it is usually equipped with the topology induced by the metric

$$d_p(x,y) = \sum_{n=1}^{\infty} 2^{-n} \sup_{t\in[-n,n]} \min(|x(t)-y(t)|, 1).$$

This (product) topology is also referred to as the topology of uniform convergence on compacts. Note that convergence of a sequence in $(C(\mathbb{R}), d_p)$ is equivalent to uniform convergence in $C([-T,T])$ for any $T > 0$. A similar statement holds for weak convergence of measures on $C(\mathbb{R})$: a sequence of measures converges weakly in $(C(\mathbb{R}), d_p)$ if and only if the image measure under the projection mapping $p_T : C(\mathbb{R}) \to C([-T,T])$ converges weakly in $C([-T,T])$ for any $T > 0$.

However, for many applications the product topology is not strong enough; the weaker the topology, the less information is contained by stating that measures converge weakly. In fact, the topology of uniform convergence on compacts cannot be used in either of the settings studied in Section 4.2. To illustrate this, we introduce a set A^β that is used in the first application. For $\beta > 0$, we set

$$\begin{aligned} A^\beta &:= \{x \in C(\mathbb{R}) : \sup_{t\geq 0} x(t) - t^\beta \geq 1\}, \\ A^\beta_o &:= \{x \in C(\mathbb{R}) : \sup_{t\geq 0} x(t) - t^\beta > 1\}. \end{aligned} \tag{4.7}$$

Suppose we have an LDP in the space $(C(\mathbb{R}), d_p)$. As this provides an upper bound for closed sets, it is desirable that A^β is closed in $(C(\mathbb{R}), d_p)$. However, this is not the case; construct a sequence $\{x_n\} \subset A^\beta$ as follows:

$$x_n(t) = \begin{cases} (1+n^\beta)(t-n+1) & t \in [n-1, n] \\ -(1+n^\beta)(t-n-1) & t \in (n, n+1] \\ 0 & \text{otherwise.} \end{cases}$$

It is readily seen that x_n converges in $(C(\mathbb{R}), d_p)$ to zero, but $0 \notin A^\beta$.

This example indicates that the 'tail' of the sample paths in $(C(\mathbb{R}), d_p)$ may cause problems. However, the topology in Ω^γ is sufficiently strong to make A^β closed, as the following lemma shows. The lemma is used in the proofs of Proposition 4.1 and Theorem 4.2.

Lemma 4.8 *For $\beta \geq \gamma$, $A^\beta \cap \Omega^\gamma$ is closed in $(\Omega^\gamma, \|\cdot\|_{\Omega^\gamma})$, and $A^\beta_o \cap \Omega^\gamma$ is open in $(\Omega^\gamma, \|\cdot\|_{\Omega^\gamma})$.*

Proof. To prove the first claim, we consider an arbitrary sequence $\{x_n\} \subset A^\beta \cap \Omega^\gamma$ converging in Ω^γ to some $x \in \Omega^\gamma$; we show that $x \in A^\beta \cap \Omega^\gamma$. We derive a contradiction by supposing that this is not the case. Note that A^β can be written as

$$A^\beta = \left\{ x \in C(\mathbb{R}) : \sup_{t \geq 0} \frac{x(t)}{1+t^\beta} \geq 1 \right\},$$

and define

$$\eta := 1 - \sup_{t \geq 0} \frac{x(t)}{1+t^\beta} > 0. \tag{4.8}$$

As $x_n \to x$ in Ω^γ, we can select an n_0 such that for $n \geq n_0$,

$$\sup_{t \geq 0} \frac{|x_n(t) - x(t)|}{1+t^\beta} < \eta/2. \tag{4.9}$$

We combine (4.8) with (4.9) to see that

$$\sup_{t \geq 0} \frac{x_n(t)}{1+t^\beta} \leq \sup_{t \geq 0} \frac{x_n(t) - x(t)}{1+t^\beta} + \sup_{t \geq 0} \frac{x(t)}{1+t^\beta} < \eta/2 + 1 - \eta < 1,$$

implying $x_n \notin A^\beta$; a contradiction.

A similar argument can be given to see that $A^\beta_o \cap \Omega^\gamma$ is open in Ω^γ. □

4.4 Weak convergence results

This section is devoted to necessary and sufficient conditions for the convergence in distribution of Y^α to a fractional Brownian motion. For background and references on weak convergence, see Section 2.2. After dealing with weak convergence on compact intervals (Section 4.4.1), we extend the results to weak convergence in Ω^γ for $\gamma > H$ (Section 4.4.2).

Throughout, σ^2_α denotes the variance function of Y^α, i.e.,

$$\sigma^2_\alpha(t) := \frac{\sigma^2(\alpha t)}{\sigma^2(\alpha)}.$$

There is a specific reason why the candidate weak limit is self-similar in the sense of (1.12); see Lamperti [209].

4.4.1 Weak convergence on compacts

Fix some time horizon $T > 0$ throughout this subsection and consider the compact interval $[-T, T]$. We slightly abuse notation by restricting Y^α to $[-T, T]$ while keeping the notation Y^α. Under the condition that $\lim_{t \to 0} \sigma^2(t) |\log |t||^{1+\epsilon} < \infty$ for some $\epsilon > 0$, Lemma 4.7 implies that the distribution of Y^α is equivalent to a probability measure ν_α on the measurable space $(C([-T,T]), \mathcal{B}(C([-T,T])))$.

To get some feeling for the necessity of this condition, we note that continuity of the sample paths of Y imply that $\lim_{t \to 0} \sigma^2(t) |\log |t|| < \infty$ under an extremely weak condition on

σ. Indeed, the following theorem, based on Sudakov's inequality, provides a simple necessary condition for continuity of the sample paths. We omit a proof, since one can repeat the arguments in van der Vaart and van Zanten [300, Cor. 2.7].

Theorem 4.9 (necessity) *Suppose that σ is strictly increasing on some neighborhood of zero. If Y has continuous sample paths, then $\lim_{t\to 0}\sigma^2(t)|\log|t||<\infty$.*

We now turn to a sufficient condition for the weak convergence of Y^α. Since $C([-T,T])$ is a separable and complete metric space, as a result of Prokhorov's theorem (Lemma 2.10), weak convergence in $C([-T,T])$ to a fractional Brownian motion is equivalent to convergence of finite-dimensional distributions in conjunction with tightness of $\{\nu_\alpha\}$.

With (3.3) and the fact that σ^2 is regularly varying, it is easy to see that the finite-dimensional distributions of Y^α converge in distribution to B_H. Therefore, weak convergence of Y^α in $C([-T,T])$ is equivalent to tightness of $\{\nu_\alpha\}$ in $C([-T,T])$. By Theorem 7.3 of [50], this is in turn equivalent to

$$\lim_{\delta\to 0}\limsup_{\alpha\to\infty}\mathbb{P}\left(\sup_{\substack{|s-t|\le\delta\\ s,t\in[-T,T]}}|Y_s^\alpha-Y_t^\alpha|\ge\zeta\right)=0, \tag{4.10}$$

for any $\zeta>0$. For notational convenience, we leave out the requirement $s,t\in[-T,T]$ explicitly in the remainder.

Theorem 4.10 (sufficiency) *If $\lim_{t\to 0}\sigma^2(t)|\log|t||^{1+\epsilon}<\infty$ for some $\epsilon>0$, then $Y^\alpha\Rightarrow B_H$ in $C([-T,T])$.*

Proof. Our objective is to prove (4.10). Since σ^2 is assumed to be continuous, the condition in the theorem implies the boundedness of $\sigma^2 L_\epsilon$ on intervals of the form $(0,\cdot]$ (recall the definition of L_ϵ in (4.6)). Therefore, as a consequence of Proposition 4.5, we have for any $\delta>0$, and α large enough, uniformly in $t\in[-\delta,\delta]\backslash\{0\}$ (obviously, $\sigma_\alpha^2(0)=0$):

$$\sigma_\alpha^2(t)\le 2\delta^{2H}\frac{L_\epsilon(\delta)}{L_\epsilon(t)}. \tag{4.11}$$

Use the fact that L_ϵ is nonincreasing and (4.11) to see that, for any $\zeta>0$ and α sufficiently large,

$$\begin{aligned}\mathbb{P}\left(\sup_{|s-t|\le\delta}|Y_s^\alpha-Y_t^\alpha|\ge\zeta\right) &= \mathbb{P}\left(\sup_{\{(s,t):2\delta^{2H}L_\epsilon(\delta)/L_\epsilon(s-t)\le 2\delta^{2H}\}}|Y_s^\alpha-Y_t^\alpha|\ge\zeta\right)\\ &\le \mathbb{P}\left(\sup_{\sigma_\alpha^2(|s-t|)\le 2\delta^{2H}}|Y_s^\alpha-Y_t^\alpha|\ge\zeta\right)\\ &\le \frac{1}{\zeta}\mathbb{E}\left(\sup_{\sigma_\alpha^2(|s-t|)\le 2\delta^{2H}}|Y_s^\alpha-Y_t^\alpha|\right).\end{aligned}$$

Define $\mathbb{H}_\alpha(\mathbf{T},\cdot)$ as the metric entropy of $\mathbf{T}\subset\mathbb{R}$ under the semimetric induced by σ_α^2; see Section 4.3.2 for definitions. Motivated by the proof of Lemma 4.7, we set $\mathbb{H}(\mathbf{T},\vartheta)=C_0\vartheta^{-2/(1+\epsilon)}$ for some constant C_0 depending on the Lebesgue measure of $\mathbf{T}$; it can be regarded as the metric entropy under the semimetric induced by L_ϵ, being only valid for small $\vartheta>0$.

We use Corollary 1.3.4 of [4] to see that there exists a constant $C>0$ such that

$$\mathbb{E}\left(\sup_{\sigma_\alpha^2(|s-t|)\le 2\delta^{2H}}|Y_s^\alpha-Y_t^\alpha|\right)\le C\int_0^{2\delta^{2H}}\sqrt{\mathbb{H}_\alpha([-T,T],\vartheta)}d\vartheta.$$

Another application of (4.11) shows that for $\vartheta > 0$,

$$\mathbb{H}_\alpha([-T,T],\vartheta) \le \mathbb{H}\left([-T,T], \frac{\vartheta}{\sqrt{2\delta^{2H}L_\epsilon(\delta)}}\right),$$

so that

$$\int_0^{2\delta^{2H}} \sqrt{\mathbb{H}_\alpha([-T,T],\vartheta)}d\vartheta \le \sqrt{2\delta^{2H}L_\epsilon(\delta)} \int_0^{2\delta^{2H}/\sqrt{2\delta^{2H}L_\epsilon(\delta)}} \sqrt{\mathbb{H}([-T,T],\vartheta)}d\vartheta.$$

To summarize, we have

$$\limsup_{\alpha\to\infty} \mathbb{P}\left(\sup_{|s-t|\le\delta} |Y_s^\alpha - Y_t^\alpha| \ge \zeta\right) \le \frac{C\sqrt{2\delta^{2H}L_\epsilon(\delta)}}{\zeta} \int_0^{\delta^H/\sqrt{L_\epsilon(\delta)/2}} \sqrt{\mathbb{H}([-T,T],\vartheta)}d\vartheta.$$

As $\int_0^\infty \sqrt{\mathbb{H}([-T,T],\vartheta)}d\vartheta < \infty$, we obtain (4.10) by letting $\delta \to 0$. □

The remainder of this subsection is devoted to easy corollaries of the sufficient condition in Theorem 4.10. We first show the relation with Lemma 4.2 of [98].

Corollary 4.11 *Suppose that σ^2 is regularly varying at zero with index $\lambda \in (0,2]$, and that σ^2 is continuous. Then we have $Y^\alpha \Rightarrow B_H$ in $C([-T,T])$.*

Proof. Since σ^2 is regularly varying at zero with index λ, $t \mapsto \sigma^2(1/t)$ is regularly varying at infinity with index $-\lambda$. With (2.2), we conclude that $\sigma^2(1/t)|t|^{\lambda/2} \to 0$ as $t \to \infty$. Equivalently, $\sigma^2(t)|t|^{-\lambda/2} \to 0$ as $t \to 0$, implying the condition in Theorem 4.10. □

Similarly, one proves the following Kolmogorov-type criterion for tightness.

Corollary 4.12 *If $\lim_{t\to 0} \sigma^2(t)|t|^{-\lambda} < \infty$ for some $\lambda \in (0,2]$, then $Y^\alpha \Rightarrow B_H$ in $C([-T,T])$.*

4.4.2 Weak convergence on Ω^γ

In this subsection, we focus on the weak convergence of Y^α to B_H in Ω^γ for $\gamma > H$. Obviously, this convergence can only take place when the laws ν_α^γ of Y^α in Ω^γ exist.

Lemma 4.13 *If $\lim_{t\to 0} \sigma^2(t)|\log|t||^{1+\epsilon} < \infty$ for some $\epsilon > 0$, then the probability measures ν_α^γ on $(\Omega^\gamma, \mathcal{B}(\Omega^\gamma))$ exist for $\gamma > H$.*

Proof. We first note that, by the assumption on $\sigma^2 L_\epsilon$, Y has almost surely continuous trajectories as detailed in the proof of Lemma 4.7. Therefore, in order to show that $Y^\alpha \in \Omega^\gamma$ almost surely, it suffices to prove that $\lim_{t\to\pm\infty} Y_t^\alpha/t^\gamma = 0$ almost surely. We use the reasoning in Section 2.1 of Addie *et al.* [2], to which we add an essential argument.

Since σ^2 is supposed to be regularly varying with index $2H$, we have $\sigma^2(t)/t^{\gamma+H} \to 0$, which can be exploited to see that for $\epsilon > 0$, $\sum_k \mathbb{P}(Y_k/k^\gamma > \epsilon) < \infty$. By the Borel-Cantelli lemma, $Y_k/k^\gamma \to 0$ almost surely. Note that, for $Z_k := \sup_{s\in[k,k+1]} |Y_s - Y_k|$,

$$|Y_t| \le |Y_{\lfloor t \rfloor}| + Z_{\lfloor t \rfloor},$$

so that it suffices to show that $Z_k/k^\gamma \to 0$ almost surely. For this, we first remark that $\mathbb{E}\exp(\alpha Z_k^2) = \mathbb{E}\exp(\alpha Z_1^2) < \infty$ for $\alpha > 0$ small enough, as a consequence of Borell's inequality (Lemma 3.3). Note that we used the continuity to ensure that this inequality can be applied. By Chernoff's bound, we have for any $\epsilon > 0$,

$$\sum_k \mathbb{P}\left(Z_k/k^\gamma > \epsilon\right) \le \sum_k \mathbb{P}\left(Z_k^2 > \epsilon^2 k^{2\gamma}\right) \le \sum_k \exp(-\alpha\epsilon^2 k^{2\gamma})\mathbb{E}\exp(\alpha Z_1^2) < \infty.$$

Another application of the Borel-Cantelli lemma proves that $Z_k/k^\gamma \to 0$ and therefore $Y_t^\alpha/t^\gamma \to 0$ almost surely.

The measure ν_α^γ can now be constructed as in Lemma 4.7; it only remains to show that $\mathcal{B}(\mathbb{R})^{\mathbb{R}} \cap \Omega^\gamma = \mathcal{B}(\Omega^\gamma)$. It is easy to see that this holds for $\gamma = 0$, and Ω^γ is isometrically isomorphic to Ω^0. □

We now investigate the probabilistic meaning of weak convergence in Ω^γ. While the weak convergence in the uniform topology on compacts is obtained by applying Theorem 4.10, the convergence in Ω^γ is substantially stronger (see Section 4.3.3). Therefore, an additional condition is needed to strengthen the convergence. Such a condition is given in Lemma 3 of Buldygin and Zaiats [65, cited according to [201]]. Lemma 4.14 below is closely related to this key result; only 'sup' has been replaced by 'lim sup'. See also Majewski [222] for a related result in a large-deviation setting.

Lemma 4.14 *Let a family of probability measures $\{\mu_n\}$ on Ω^γ be given. Suppose that the image of $\{\mu_n\}$ under the projection mapping $p_T : \Omega^\gamma \to C([-T,T])$ is tight in $C([-T,T])$ for all $T > 0$. Then $\{\mu_n\}$ is tight in Ω^γ if and only if for any $\zeta > 0$,*

$$\lim_{T\to\infty} \limsup_{n\to\infty} \mu_n \left(x \in \Omega^\gamma : \sup_{|t|\geq T} \frac{|x(t)|}{1+|t|^\gamma} \geq \zeta \right) = 0. \tag{4.12}$$

Proof. To prove necessity, let $\{\mu_n\}$ be tight in Ω^γ and fix $\zeta > 0$. Given $\eta > 0$, choose an Ω^γ-compact set K such that $\mu_n(K) > 1 - \eta$ for all n. We denote an Ω^γ-ball centered at x with radius ζ by $B_\zeta(x)$, so that $\{B_{\zeta/2}(x) : x \in K\}$ is an Ω^γ-open cover of K. Since K is Ω^γ-compact, one can select $m < \infty$ and $x_1, \ldots, x_m$ such that $\{B_{\zeta/2}(x_i) : i = 1, \ldots, m\}$ also covers K. Let T_i be such that $\sup_{|t|\geq T_i} |x_i(t)|/(1+|t|^\gamma) < \zeta/2$, and set

$$A_{\zeta,T} := \left\{ x \in \Omega^\gamma : \sup_{|t|\geq T} \frac{|x(t)|}{1+|t|^\gamma} < \zeta \right\}.$$

Note that for $T := \max_i T_i$, $K \subset A_{\zeta,T}$. We have now shown that for any $\eta > 0$ one can find $T > 0$ such that

$$\sup_{n\geq 1} \mu_n(A_{\zeta,T}^c) \leq \eta. \tag{4.13}$$

Obviously, this implies (4.12).

For sufficiency, instead of supposing (4.12), we may suppose without loss of generality that for any $\eta > 0$ there exists a $T > 0$ such that (4.13) holds. Indeed, since Ω^γ is separable and complete, any probability measure on Ω^γ is tight (Lemma 2.10). In particular, the above reasoning used to prove necessity implies that for any $\eta > 0$ and $n \geq 1$, one can find $T_n > 0$ such that $\mu_n(A_{\zeta,T_n}^c) \leq \eta$. As a consequence of (4.12), there exists a $T' > 0$ and n_0 such that $\sup_{n\geq n_0} \mu_n(A_{\zeta,T'}^c) \leq \eta$. Hence we have (4.13) for $T := \max(T', \max_{n\leq n_0} T_n)$.

Suppose the image of $\{\mu_n\}$ under the projection mapping is tight in $C([-T,T])$ for all $T > 0$. We can then choose a set K that is compact in the topology of uniform convergence on compact intervals such that $\sup_n \mu_n(K^c) \leq \eta/2$. For brevity, we call K $\mathcal{U}$-compact. Using (4.13), we can select for any $m \in \mathbb{N}$ a $T_m > 0$ such that $\sup_{n\geq 1} \mu_n(A_{1/m,T_m}^c) \leq \eta/2^{m+1}$. Set $K' := K \cap \bigcap_{m\in\mathbb{N}} A_{1/m,T_m}$ and note that $\inf_{n\geq 1} \mu_n(K') \geq 1 - \eta$. Therefore, we have established the claim once we have shown that the Ω^γ-closure of K' is Ω^γ-compact. For this, let $\{x_\ell\}$ be a sequence in K', and let $\delta > 0$ be arbitrary. Since K is $\mathcal{U}$-compact, we can find a subsequence $\{x_{\ell_k}\}$ of $\{x_\ell\}$ that converges uniformly on compact intervals, say, to x. Moreover, $\{x_{\ell_k}\} \subset \bigcap_{m\in\mathbb{N}} A_{1/m,T_m}$ implies that we can find a $T > 0$ such that $\sup_{k\geq 1} \sup_{|t|\geq T} |x_{\ell_k}(t)|/(1+|t|) < \delta/2$. As $x \in \Omega^\gamma$, we can also choose $T' > 0$ such that $\sup_{|t|\geq T'} |x(t)|/(1+|t|^\gamma) \leq \delta/2$. From the convergence of

x_{ℓ_k} to x on compacts we deduce that $\sup_{|t|\le \max(T,T')} |x_{\ell_k}(t) - x(t)|/(1+|t|^\gamma) \le \delta$. We now readily infer that $\|x_{\ell_k} - x\|_\Omega \le \delta$, i.e., K' is Ω^γ-compact. □

Having a characterization of weak convergence in Ω^γ at our disposal, we now specialize to the framework of the present chapter. The main result of this section is that the family $\{\nu_\alpha^\gamma\}$ is tight in Ω^γ under the conditions of Theorem 4.10. It may seem rather surprising that it is possible to establish this tightness in Ω^γ without an additional condition on long-term behavior of sample paths. Apparently, the fact that the variance function varies regularly with an index $2H < 2\gamma$, in conjunction with the Gaussian nature, suffice to control the process over large time periods.

Theorem 4.15 (sufficiency) *If* $\lim_{t\to 0} \sigma^2(t)|\log|t||^{1+\epsilon} < \infty$ *for some* $\epsilon > 0$*, then* $Y^\alpha \Rightarrow B_H$ *in* Ω^γ *for any* $\gamma > H$*.*

Proof. Having Lemma 4.14 at our disposal, we need to establish (4.12) for the family $\{\nu_\alpha^\gamma\}$, or, equivalently, for any $\zeta > 0$,

$$\lim_{T\to\infty} \limsup_{\alpha\to\infty} \mathbb{P}\left(\sup_{t\ge T} \frac{|Y_t^\alpha|}{1+t^\gamma} \ge \zeta \right) = 0.$$

An upper bound for the probability in the preceding display is based on Markov's inequality and some elementary considerations (see [3, Lem. 3.1]): for $\zeta > 0$, $\alpha, k \ge 1$,

$$\begin{aligned} \mathbb{P}\left(\sup_{t\ge e^k} \frac{|Y_t^\alpha|}{1+t^\gamma} \ge \zeta \right) &\le \sum_{j=k}^\infty \mathbb{P}\left(\sup_{t\in[e^j,e^{j+1}]} \frac{|Y_t^\alpha|}{1+t^\gamma} \ge \zeta \right) \le \frac{1}{\zeta} \sum_{j=k}^\infty \frac{\mathbb{E}\sup_{t\in[e^j,e^{j+1}]} |Y_t^\alpha|}{1+e^{j\gamma}} \\ &\le \frac{1}{\zeta} \sum_{j=k}^\infty \frac{\mathbb{E}|Y_{e^j}^\alpha|}{1+e^{j\gamma}} + \frac{2}{\zeta} \sum_{j=k}^\infty \frac{\mathbb{E}\sup_{t\in[e^j,e^{j+1}]} Y_t^\alpha}{1+e^{j\gamma}}. \end{aligned} \tag{4.14}$$

As a consequence of the UCT in Corollary 2.4 applied to $\sigma(t)/t^{(\gamma-H)/2}$, the first term is (at least for large k) of the order

$$\sum_{j=k}^\infty \frac{\sigma(\alpha e^j)}{e^{j\gamma}\sigma(\alpha)} \le \sum_{j=k}^\infty e^{-j(H+\gamma)/2},$$

and the resulting upper bound tends to zero as $k \to \infty$.

As in the proof of Theorem 4.10, we use metric entropy techniques to find an upper bound on the second term in (4.14). Recall the notation $\mathbb{H}_\alpha(\mathbf{T},\cdot)$ and $\mathbb{H}(\mathbf{T},\cdot)$ that we used in the proof of Theorem 4.10. By Theorem 14.1 of Lifshits [219], there exists a constant $C > 0$ such that $\mathbb{E}\sup_{t\in[e^j,e^{j+1}]} Y_t^\alpha \le C \int \sqrt{\mathbb{H}_\alpha([e^j,e^{j+1}],\vartheta)}d\vartheta$.

We now derive a bound on $\int \sqrt{\mathbb{H}_\alpha([e^k,e^{k+1}],\vartheta)}d\vartheta$ for k large, uniformly in α. The first step is to bound the variance σ_α^2. As a consequence of Proposition 4.5, we have for $\alpha \to \infty$,

$$\sup_{|t|\le 1/e} \sigma_\alpha^2(t) L_\epsilon(t) \to \sup_{|t|\le 1/e} t^{2H} L_\epsilon(t) = e^{-2H}.$$

Moreover, by the UCT, for large α, we have $\sigma_\alpha^2(t) \le 2e^{\gamma-H} t^{H+\gamma}$ for all $|t| \ge 1/e$. Therefore, the function M_ϵ given by

$$M_\epsilon(t) := \begin{cases} 2e^{-2H}/L_\epsilon(t) & \text{if } |t| \le 1/e; \\ 2e^{\gamma-H} t^{H+\gamma} & \text{otherwise.} \end{cases}$$

majorizes σ_α^2 uniformly in (large) α. It is important to note that M_ϵ is continuous and strictly increasing for $t \in \mathbb{R}_+$.

Using the stationarity of the increments and the fact that the inverse of $1/\sqrt{L_\epsilon(\cdot)}$ is given by $\vartheta \mapsto \exp\left(-\vartheta^{-2/(1+\epsilon)}\right)$ for $\vartheta \in [0, \sqrt{2}e^{-H}]$, we see that for large j,

$$\begin{aligned}\int_0^{\sqrt{2}e^{-H}} \sqrt{\mathbb{H}_\alpha([e^j, e^{j+1}], \vartheta)}d\vartheta &= \int_0^{\sqrt{2}e^{-H}} \sqrt{\mathbb{H}_\alpha([0, e^j(e-1)], \vartheta)}d\vartheta \\ &\leq \sqrt{2}e^{-H}\int_0^1 \sqrt{\log\left(\frac{e^j(e-1)}{2\exp\left(-\vartheta^{-2/(1+\epsilon)}\right)}+1\right)}d\vartheta \\ &\leq e^{-H}\log[e^j(e-1)] + 1 + \frac{1}{\epsilon},\end{aligned}$$

implying that

$$\lim_{k\to\infty}\sum_{j=k}^{\infty}\frac{\int_0^{\sqrt{2}e^{-H}}\sqrt{\mathbb{H}_\alpha([e^j, e^{j+1}], \vartheta)}d\vartheta}{1+e^{j\gamma}} = 0,$$

so that it remains to show a similar statement for the integration interval $[\sqrt{2}e^{-H}, \infty)$. For this, observe that, for some constant $\mathcal{C} > 0$,

$$\begin{aligned}&\int_{\sqrt{2}e^{-H}}^{\infty}\sqrt{\mathbb{H}_\alpha([e^j, e^{j+1}], \vartheta)}d\vartheta \\ &\leq \int_{\sqrt{2}e^{-H}}^{e^{H(j+1)}}\sqrt{\log\left(\frac{e^j(e-1)}{2^{1-1/(H+\gamma)}e^{(H-\gamma)/(H+\gamma)}\vartheta^{2/(H+\gamma)}}\right)+1}d\vartheta \\ &\leq (e^{H(j+1)} - \sqrt{2}e^{-H})\sqrt{\log\left(\frac{\mathcal{C}e^j}{(2e^{-2H})^{1/(H+\gamma)}}\right)},\end{aligned}$$

from which the claim is readily obtained. □

We now relate Theorem 4.15 to Proposition 2.9 of Kozachenko *et al.* [201]. The criterion given in Proposition 2.9 of [201] states that

$$\sup_{t\in\mathbb{R}_+}\overline{\sigma}^2(t)(\log(1+1/t))^{1+\epsilon} < \infty \tag{4.15}$$

for some $\epsilon \in (0,1)$ (recall the definition of $\overline{\sigma}^2$ in (4.1)). Since we already noted that $\lim_{t\to 0}(\log(1+1/t))^{1+\epsilon}/L_\epsilon(t) = 1$, this condition implies that $\lim_{t\to 0}\sigma^2(t)|\log|t||^{1+\epsilon} < \infty$ for some $\epsilon > 0$. Although the continuity of σ is not stated explicitly in [201], it is necessary to obtain continuity of the sample paths of Y. Indeed, if Y has continuous sample paths and σ^2 is locally bounded (as implied by (4.15)), then the dominated convergence theorem implies the continuity of σ^2.

In conclusion, the condition in Theorem 4.15 improves this result of [201] in two ways: the condition is both easier to check and weaker.

We get the following important corollary by combining Theorem 4.15 with Chevet's theorem 3.2.

Corollary 4.16 *Let $\{a_\alpha\}$ be a sequence of positive real numbers tending to infinity as $\alpha \to \infty$. If $\lim_{t\to 0}\sigma^2(t)|\log|t||^{1+\epsilon} < \infty$ for some $\epsilon > 0$, then the distributions in Ω^γ ($\gamma > H$) of Y^α/a_α satisfy the LDP in Ω^γ with the scale sequence $\{a_\alpha^2\}$ and rate function I given by (3.6), where $\mathcal{H}_H$ is the Cameron-Martin space associated with fractional Brownian motion on Ω^γ.*

4.5 Proofs for Section 4.2

4.5.1 Proof of Proposition 4.1

For A^β be given by (4.7), we note

$$\begin{aligned}\mathbb{P}\left(\sup_{t\geq 0} Y_t - ct^\beta \geq u\right) &= \mathbb{P}\left(\sup_{t\geq 0} Y_{(u/c)^{1/\beta}t} - ut^\beta \geq u\right)\\ &= \mathbb{P}\left(\sup_{t\geq 0} \frac{1}{u} Y_{(u/c)^{1/\beta}t} - t^\beta \geq 1\right)\\ &= \nu^\beta_{(u/c)^{1/\beta}}\left(\frac{u}{\sigma([u/c]^{1/\beta})} A^\beta \cap \Omega^\beta\right).\end{aligned}$$

Since $A^\beta \cap \Omega^\beta$ is closed in Ω^β by Lemma 4.8, we have by Corollary 4.16,

$$\begin{aligned}&\limsup_{u\to\infty} \frac{\sigma^2(u^{1/\beta})}{u^2} \log \mathbb{P}\left(\sup_{t\geq 0} Y_t - ct^\beta \geq u\right)\\ &= \limsup_{u\to\infty} c^{2H/\beta} \frac{\sigma^2([u/c]^{1/\beta})}{u^2} \log \mathbb{P}\left(\sup_{t\geq 0} Y_t - ct^\beta \geq u\right)\\ &\leq -c^{2H/\beta} \inf_{x\in A^\beta\cap\Omega^\beta} I(x),\end{aligned}$$

where I is given by (3.6). It remains to calculate the quantity $\inf_{x\in A^\beta\cap\Omega^\beta} I(x)$, or equivalently $\inf_{t\geq 0}\inf_{\{x\in\Omega^\beta : x(t)-t^\beta\geq 1\}} I(x)$. It is left to the reader to repeat the argument in the proof of Proposition 2 in Addie *et al.* [2] to see that, also for $\beta \neq 1$, $2\inf_{\{x\in\Omega^\beta : x(t)-t^\beta\geq 1\}} I(x) = (1+t^\beta)^2/t^{2H}$. Straightforward calculus shows that the (unique) infimum over t is attained at $t = (H/[\beta - H])^{1/\beta}$. The analysis in [2] also shows that the minimizing argument x^* of $\inf_{x\in A^\beta\cap\Omega^\beta} I(x)$ is indeed given by (4.3).

For the lower bound, note that $A^\beta \supset A^\beta_o$, and that $A^\beta_o \cap \Omega^\beta$ is Ω^β-open by Lemma 4.8. Therefore,

$$\liminf_{u\to\infty} \frac{\sigma^2(u^{1/\beta})}{u^2} \log \mathbb{P}\left(\sup_{t\geq 0} Y_t - ct^\beta \geq u\right) \geq -c^{2H/\beta} \inf_{x\in A^\beta_o\cap\Omega^\beta} I(x).$$

An elementary argument shows that $\inf_{x\in A^\beta_o\cap\Omega^\beta} I(x) = \inf_{x\in A^\beta\cap\Omega^\beta} I(x)$, so that the claim is proven.

4.5.2 Proof of Theorem 4.2

By the portmanteau theorem (Theorem 2.11), it suffices to show that for all Ω^β-closed sets F,

$$\limsup_{u\to\infty} \mathbb{P}\left(\frac{1}{u} Y_{(u/c)^{1/\beta}\cdot} \in F \,\middle|\, \sup_{t\geq 0} Y_t - ct^\beta \geq u\right) \leq \delta_{x^*}(F). \tag{4.16}$$

Since this assertion is trivial if $x^* \in F$, we suppose that $x^* \notin F$. Denote the probability on the left-hand side of (4.16) by p_u, so that

$$\log p_u = \log \nu^\beta_{(u/c)^{1/\beta}}\left(\frac{u}{\sigma([u/c]^{1/\beta})}\left(A^\beta \cap F\right)\right) - \log \nu^\beta_{(u/c)^{1/\beta}}\left(\frac{u}{\sigma([u/c]^{1/\beta})} A^\beta \cap \Omega^\beta\right)$$

and by Corollary 4.16, as both F and $A^\beta \cap \Omega^\beta$ are closed,

$$\limsup_{u\to\infty} \frac{\sigma^2(u^{1/\beta})}{u^2} \log p_u \leq -c^{2H/\beta} \inf_{x\in A^\beta\cap F} I(x) + c^{2H/\beta} \inf_{x\in A^\beta_o\cap\Omega^\beta} I(x). \tag{4.17}$$

We proceed by showing that

$$\inf_{x\in A^\beta\cap F} I(x) > \inf_{x\in A^\beta\cap\Omega^\beta} I(x). \tag{4.18}$$

For this, we suppose that we have equality, so that we can find a sequence $\{x_n\}\subset A^\beta\cap F$ with $I(x_n) < I(x^*)+1/n$. Without loss of generality, we may suppose that x_n is $(1+t_n^\beta)/(2t_n^{2H})\mathbb{C}\mathrm{ov}(B_H(\cdot),B_H(t_n))$ for some $t_n\geq 0$, since the minimizer of the rate function over the set $\{x : x(t_n)-t_n^\beta\geq 1\}$ has this form, cf. (4.3). Moreover, by uniqueness of t^*, we must have $t_n\to t^*$ in order to ensure that $(1+t_n^\beta)^2/(2t_n^{2H}) = I(x_n) < I(x^*)+1/n$. An easy calculation shows that then x_n converges in Ω^β to $x^*\notin F$, which contradicts the fact that F is closed.

The claim follows by combining (4.17) with (4.18) and observing that $\inf_{x\in A_o^\beta\cap\Omega^\beta} I(x) = \inf_{x\in A^\beta\cap\Omega^\beta} I(x)$.

4.5.3 Proof of Proposition 4.3

Corollary 4.16 implies that $Y_{\cdot T}/T$ satisfies the large-deviation principle in Ω^1 with the rate function I and scale sequence $\sigma^2(T)/T^2$. Having observed this, the remainder of the proof is a combination of the arguments contained in Kozachenko *et al.* [201] and Norros [244]; we do not repeat them. The idea is to use $\mathbb{P}(Y\in K_T) = \mathbb{P}(Y_{\cdot T}/T\in K_1)$, and then justify the limit in (4.5) by showing that K_1 is open and that $\inf_{x\in K_1} I(x) = \inf_{x\in\overline{K_1}} I(x)$.

4.5.4 Proof of Theorem 4.4

We first show the existence and uniqueness of $\overline{x}$. For this, note that the large-deviation principle for $Y_{\cdot T}/T$ is governed by a strictly convex rate function. Both existence and uniqueness follow from Proposition 4.4 of [244], since $\inf_{x\in\overline{K_1}} I(x)$ can be written as an infimum of the rate function over a convex set. Note that Proposition 4.3 and Proposition 4.4 of [244] together imply that $\inf_{x\in K_1} I(x) = I(\overline{x})$.

By a similar reasoning as in the proof of Theorem 4.2 and the fact that K_1 is open [244], the claim follows after showing that for Ω^γ-closed sets F with $\overline{x}\notin F$,

$$\inf_{x\in\overline{K_1}\cap F} I(x) > \inf_{x\in K_1} I(x) = I(\overline{x}),$$

cf. (4.18). Suppose we have equality in the preceding display. For every $n\in\mathbb{N}$, one can then select an $x_n\in\overline{K_1}\cap F$ such that $I(x_n)\leq I(\overline{x})+1/n$. Now define the sets

$$M_n := \{x\in\Omega : I(x)\leq I(\overline{x})+1/n\}.$$

By the goodness of the rate function, these sets are Ω^γ-compact. Since $\{x_n\}\subset M_1$, one can select a subsequence of $\{x_n\}$ that converges (in Ω^γ) to some $\underline{x}$. As the M_n decrease, one then has that $\underline{x}\in M_n$ for every $n\in\mathbb{N}$, implying that $I(\underline{x}) = I(\overline{x})$. By construction we also have $\underline{x}\in\overline{K_1}\cap F$ as the latter set is closed in Ω^γ. Uniqueness yields $\overline{x} = \underline{x}$, contradicting $\underline{x}\in F$.

4.6 Proof of Proposition 4.5

The proof is modeled after the proof of the UCT; see Theorem 1.5.2 in Bingham *et al.* [51]. We start with some notation. Let $\eta>0$ be arbitrary and let $\mathcal{T} = \mathcal{T}(\eta) < 1$ be such that

$$\sup_{t\in[0,\mathcal{T}]} t^\rho L_\epsilon(t) < \frac{1}{9}\eta. \tag{4.19}$$

Since f is regularly varying with index $\rho > 0$, we can find A_1 so that for $\alpha \geq A_1$,

$$\frac{f(\alpha)}{f(\alpha/\mathcal{T})} \leq 2\mathcal{T}^{\rho}. \tag{4.20}$$

Define

$$M := \sup_{0<\alpha\leq A_1} f(\alpha)L_{\epsilon}(\alpha),$$

which is finite by assumption. Using (2.2), we pick A_2 so that for $\alpha \geq A_2$,

$$\frac{(\log A_1)^{1+\epsilon} + (\log \alpha)^{1+\epsilon}}{f(\alpha)} \leq \frac{\eta}{2^{1+\epsilon}M}. \tag{4.21}$$

Without loss of generality, we may suppose that $A_2 \geq eA_1$ and $A_1 \geq e$.

The outline of the proof is as follows. In the first step, we show

$$\sup_{t\in(0,\mathcal{T}]} \sup_{\alpha\geq A_1/t} \left| \frac{f(\alpha t)}{f(\alpha)} L_{\epsilon}(t) - t^{\rho}L_{\epsilon}(t) \right| < \eta, \tag{4.22}$$

and then we show that

$$\sup_{t\in(0,A_1/A_2]} \sup_{A_2\leq\alpha\leq A_1/t} \left| \frac{f(\alpha t)}{f(\alpha)} L_{\epsilon}(t) - t^{\rho}L_{\epsilon}(t) \right| < \eta. \tag{4.23}$$

In the third and last step, we use (4.22) and (4.23) to establish the claim.

Step 1: Proof of (4.22)

Apply Theorem 1.5.4 of [51] to the regularly varying function $f(\alpha)\alpha^{-\rho}$ to see that

$$f(\alpha) = c(\alpha)\alpha^{\rho/2}\phi(\alpha),$$

where $c(\alpha) \to 1$ as $\alpha \to \infty$, and ϕ is nondecreasing. Without loss of generality, we may assume that A_1 is such that $1/2 \leq c(\alpha) \leq 2$ for $\alpha \geq A_1$. For fixed $t \in (0, \mathcal{T}]$, we have $t \leq 1$ (since $\mathcal{T} < 1$), so that

$$\begin{aligned}
\sup_{\alpha\geq A_1/t} \frac{f(\alpha t)}{f(\alpha)} L_{\epsilon}(t) &= \sup_{\alpha\geq A_1} \frac{f(\alpha)}{f(\alpha/t)} L_{\epsilon}(t) \\
&= \sup_{\alpha\geq A_1} \frac{f(\alpha)}{c(\alpha/t)\alpha^{\rho/2}\phi(\alpha/t)} t^{\rho/2}L_{\epsilon}(t) \\
&\leq 2 \sup_{\alpha\geq A_1} \frac{f(\alpha)}{\alpha^{\rho/2}\phi(\alpha/t)} t^{\rho/2}L_{\epsilon}(t).
\end{aligned}$$

Since both $t^{\rho/2}L_{\epsilon}(t)$ and (for any α) $f(\alpha)\alpha^{-\rho/2}/\phi(\alpha/t)$ are nondecreasing in t on $(0, \mathcal{T}]$, we conclude with (4.20) that

$$\begin{aligned}
\sup_{t\in(0,\mathcal{T}]} \sup_{\alpha\geq A_1/t} \frac{f(\alpha t)}{f(\alpha)} L_{\epsilon}(t) &\leq 2 \sup_{\alpha\geq A_1} \frac{f(\alpha)}{\alpha^{\rho/2}\phi(\alpha/\mathcal{T})} \mathcal{T}^{\rho/2}L_{\epsilon}(\mathcal{T}) \\
&\leq 4 \sup_{\alpha\geq A_1} \frac{f(\alpha)}{c(\alpha/\mathcal{T})\alpha^{\rho/2}\phi(\alpha/\mathcal{T})} \mathcal{T}^{\rho/2}L_{\epsilon}(\mathcal{T}) \\
&= 4 \sup_{\alpha\geq A_1} \frac{f(\alpha)}{f(\alpha/\mathcal{T})} L_{\epsilon}(\mathcal{T}) \\
&\leq 8\mathcal{T}^{\rho}L_{\epsilon}(\mathcal{T}) \\
&\leq 8 \sup_{t\in[0,\mathcal{T}]} t^{\rho}L_{\epsilon}(t).
\end{aligned}$$

Inequality (4.22) is an easy consequence of combining this with (4.19).

Step 2: Proof of (4.23)

Note that, since $t \in [0,\infty) \mapsto |t|^{1+\epsilon}$ is convex, we have for $t/\alpha \le 1/e$

$$\begin{aligned} L_\epsilon(t/\alpha) &= |\log t/\alpha|^{1+\epsilon} \le (|\log t| + \log\alpha)^{1+\epsilon} \le 2^{1+\epsilon}\left(\frac{1}{2}|\log t|^{1+\epsilon} + \frac{1}{2}(\log\alpha)^{1+\epsilon}\right) \\ &= 2^\epsilon(|\log t|^{1+\epsilon} + (\log\alpha)^{1+\epsilon}). \end{aligned}$$

Observe that $t/\alpha \le 1/e$ for $t \in [0, A_1]$ and $\alpha \ge A_2$, so that

$$\begin{aligned} \sup_{t\in(0,A_1/A_2]} \sup_{A_2\le\alpha\le A_1/t} \frac{f(\alpha t)}{f(\alpha)} L_\epsilon(t) &= \sup_{\alpha\ge A_2} \sup_{t\in(0,A_1/\alpha]} \frac{f(\alpha t)}{f(\alpha)} L_\epsilon(t) \\ &= \sup_{\alpha\ge A_2} \sup_{t\in(0,A_1]} \frac{f(t)}{f(\alpha)} L_\epsilon(t/\alpha) \\ &\le 2^\epsilon \sup_{\alpha\ge A_2} \sup_{t\in(0,A_1]} \frac{f(t)}{f(\alpha)} \left(|\log t|^{1+\epsilon} + (\log\alpha)^{1+\epsilon}\right). \end{aligned}$$

Since $A_1 \ge e$, we can bound $f(t)|\log t|^{1+\epsilon}$ as follows on $(0, A_1]$:

$$\begin{aligned} \sup_{t\in(0,A_1]} f(t)|\log t|^{1+\epsilon} &\le \max\left(\sup_{t\in(0,1/e]} f(t)L_\epsilon(t), \sup_{t\in(1/e,A_1]} f(t)|\log A_1|^{1+\epsilon}\right) \\ &\le \max(M, M(\log A_1)^{1+\epsilon}) \\ &= M(\log A_1)^{1+\epsilon}. \end{aligned}$$

By combining the two preceding displays, we obtain

$$\sup_{t\in(0,A_1/A_2]} \sup_{A_2\le\alpha\le A_1/t} \frac{f(\alpha t)}{f(\alpha)} L_\epsilon(t) \le 2^\epsilon M \sup_{\alpha\ge A_2} \frac{(\log A_1)^{1+\epsilon} + (\log\alpha)^{1+\epsilon}}{f(\alpha)} \le \frac{1}{2}\eta,$$

where the last inequality is (4.21). Inequality (4.23) readily follows by using (4.19).

Step 3: Proof of the claim

It is readily checked that the first two steps imply that for any (small) $\eta > 0$ we can find a (small) κ and (large) A_2 such that

$$\sup_{\alpha\ge A_2} \sup_{t\in(0,\kappa]} \left|\frac{f(\alpha t)}{f(\alpha)} L_\epsilon(t) - t^\rho L_\epsilon(t)\right| < \eta. \tag{4.24}$$

In this last step, we establish the uniform convergence on each interval $(0, \cdot]$.

Let $T > 0$ be arbitrary, and set $M' := \sup_{t\in(0,T]} t^\rho L_\epsilon(t) < \infty$. By the UCT (Theorem 2.3), it is possible to select A_3 so that for $\alpha \ge A_3$, uniformly in $t \in [\kappa, T]$,

$$\left|\frac{f(\alpha t)}{f(\alpha)t^\rho} - 1\right| < \frac{\eta}{M'}.$$

Now, for $\alpha \ge A_3$, we have

$$\sup_{t\in[\kappa,T]} \left|\frac{f(\alpha t)}{f(\alpha)} L_\epsilon(t) - t^\rho L_\epsilon(t)\right| = \sup_{t\in[\kappa,T]} t^\rho L_\epsilon(t) \left|\frac{f(\alpha t)}{f(\alpha)t^\rho} - 1\right| < \eta.$$

Combining this with (4.24) yields the claim.

CHAPTER 5

Extremes of Gaussian processes

In this chapter, we continue the investigation of the tail asymptotics for the buffer content as started in Chapter 4. Again, we rely on the representation of the buffer-content distribution as the maximum of the free process.

This naturally leads to the following setup. Let Y be a centered separable Gaussian process with a variance function that is regularly varying at infinity with index $2H \in (0, 2)$. Moreover, let ϕ be a 'drift' function that is strictly increasing, regularly varying at infinity with index $\beta > H$, and vanishing at the origin. We examine the exact asymptotics of the probability $\mathbb{P}(\sup_{t\geq 0} Y_t - \phi(t) > u)$ as $u \to \infty$.

To obtain these asymptotics, we tailor the celebrated double sum method (see Section 3.3) to our general framework. Two different families of correlation structures are studied, leading to four qualitatively different types of asymptotic behavior. Our results cover both processes with stationary increments (including Gaussian integrated processes) and self-similar processes.

5.1 Introduction

Let Y be a centered separable Gaussian process, and let ϕ be a strictly increasing 'drift' function with $\phi(0) = 0$. Motivated by applications in telecommunications engineering and insurance mathematics, the probability

$$\mathbb{P}\left(\sup_{t\geq 0} Y_t - \phi(t) > u\right) \tag{5.1}$$

has been analyzed under different levels of generality as $u \to \infty$. In these applications, Y_0 is supposed to be degenerate, i.e., $Y_0 = 0$. Letting u tend to infinity is known as investigating the *large-buffer regime*, since u can be interpreted as a buffer level of a fluid queue, see Section 1.1. Note that (5.1) can be rewritten as

$$\mathbb{P}\left(\sup_{t\geq 0} \frac{Y_{\mu(ut)}}{1+t} > u\right), \tag{5.2}$$

where μ is the inverse of ϕ. Special attention has been paid to the case where Y has stationary increments (e.g., [71, 72, 85, 86, 87, 93, 96, 123, 164, 201, 225, 241]), and to the case where Y is self-similar or 'almost' self-similar [163].

The main contribution of the present chapter is that we extend the known results on the exact asymptotics of (5.1). For this, we introduce a wide class of local correlation structures,

covering both processes with stationary increments and 'almost' self-similar processes. A motivation for studying the problem in this generality is to gain insight into the case where Y is the sum of a number of independent Gaussian processes, e.g., the sum of a Gaussian integrated process and a number of fractional Brownian motions with different Hurst parameters. We study this situation in detail in the next chapter.

Some words for the technical aspects of this chapter. We use the *double sum method* to establish the asymptotics of (5.2). In Section 3.3, it has been sketched how this method can be applied to find the asymptotics of $\mathbb{P}(\sup_{t\in[0,S]} X(t) > u)$, where X is a stationary Gaussian process. These asymptotics have also been investigated if X is a Gaussian process with a unique point of maximum variance [260], and there also exist analogues for fields [257, Sec. 8]. However, these results cannot be applied to establish the asymptotics of (5.1).

In this chapter, we approach the double sum method differently. The idea in [260] is to first establish the tail asymptotics for the maximum of some stationary Gaussian process on a subinterval of $[0, S]$. Then a comparison inequality is applied to see that the asymptotics of $\mathbb{P}(\sup_{t\in[0,S]} X(t) > u)$ equal the asymptotics of this stationary field. Here, we do not make a comparison to stationary processes, but we apply the ideas underlying the double sum method directly to the processes $Y_{\mu(ut)}/(1+t)$. Given our results, it can be seen immediately that the comparison approach cannot work in the generality of this chapter: a so-called *generalized* Pickands' constant appears, which is not present in the stationary case. It is also obtained in the analysis of extremes of Gaussian integrated processes, see Dębicki [87]. The appearance of this constant in the present study is not surprising, since our results also cover Gaussian integrated processes. We refer to Kobelkov [193] for recent results on generalized Pickands' constants.

Several related problems appear in the vast body of literature on asymptotics for Gaussian processes. For instance, Dębicki and Rolski [97] and Duncan *et al.* [125] study the asymptotics of (5.1) over a finite horizon, i.e., the supremum is taken over $[0, S]$ for some $S > 0$. These asymptotics differ qualitatively from the asymptotics established in the present chapter. We also mention the work of Zeevi and Glynn [312] and Piterbarg [258]. In queueing terminology, they examine the maximum of the stationary buffer-content process for a Gaussian queue (see Section 1.1.3), while we study its marginal distribution (i.e., the steady-state buffer-content distribution).

Another problem closely related to the present setting is where Y has the form $Z/\sqrt{n}$ for some Gaussian process Z independent of n. One then fixes u and studies the probability (5.1) as $n \to \infty$. The resulting asymptotics were studied by Dębicki and Mandjes [90]; these asymptotics are often called *many-sources asymptotics*, since convolving identical Gaussian measures amounts to scaling a single measure. The many-sources asymptotic regime is studied in Chapter 10.

It is also worthwhile to compare our results with those of Berman [39] on extremes of Gaussian processes with stationary increments. Berman studies the probability $\mathbb{P}(\sup_{t\in B} \overline{Y}_t > u)$ for $u \to \infty$, where $\overline{Y}$ is constructed from Y by standardization (so that its variance is constant) and B is some fixed compact interval. The problem of finding the asymptotics of (5.2) does not fit into Berman's framework: our assumptions will imply that $Y_{\mu(ut)}/(1+t)$ has a point of maximum variance, which is asymptotically unique. Another difference is that this point depends (asymptotically) linearly on u, so that it cannot belong to B for large u.

The chapter is organized as follows. The main result and its assumptions are described in Section 5.2. In Section 5.3, we work out two cases of special interest: processes with stationary increments and self-similar processes. Furthermore, we relate our formulas with the literature by giving some examples. Interestingly, these examples show that the tail asymptotics for the buffer content in a Gaussian fluid queue are qualitatively different if the input is short-range dependent or long-range dependent.

Sections 5.4–5.7 are devoted to proofs. In Section 5.4, the classical Pickands' lemma

(Lemma 3.6) is generalized. Section 5.5 distinguishes four instances of this lemma. The resulting observations are key to the derivation of the upper bounds, which is the topic of Section 5.6. Lower bounds are given in Section 5.7, where we use a double sum-type argument to see that the upper and lower bounds coincide asymptotically.

To slightly reduce the length of the proofs and make them more readable, details are often omitted when a similar argument has already been given, or when the argument is standard. We then use curly brackets (e.g., {**T1**}) to indicate which assumptions are needed to make the claim precise.

The results in this chapter rely on standard results for regularly varying functions, see Section 2.1. Specifically, the uniform convergence theorem (Theorem 2.3) and its corollary (Corollary 2.4) are used extensively. For convenience, we call both results the UCT and we apply these two results without reference to the specific version that is used.

5.2 Description of the results and assumptions

This section presents our main theorem. Since many (yet natural and weak) assumptions underly our result, we defer a detailed description of these assumptions to Section 5.2.2.

5.2.1 Main theorem

The supremum in (5.2) is asymptotically 'most likely' attained at a point where the variance is close to its maximum value. Let t_u^* denote a point that maximizes the variance $\sigma^2(\mu(ut))/(1+t)^2$ (existence will be ensured by continuity conditions). Our main assumptions are that σ^2 (defined as $\sigma^2(t) := \mathbb{V}\mathrm{ar}Y_t$) and μ (defined as the inverse of ϕ in (5.1)) are regularly varying at infinity with indices $2H \in (0,2)$ and $1/\beta < 1/H$ respectively. Note that the UCT implies that t_u^* converges to $t^* := H/(\beta - H)$. In this sense, t_u^* is asymptotically unique.

For an appropriately chosen δ with $\delta(u)/u \to 0$ and $\sigma(\mu(u))/\delta(u) \to 0$, (5.1) and (5.2) are asymptotically equivalent to

$$\mathbb{P}\left(\sup_{t \in [t_u^* \pm \delta(u)/u]} \frac{Y_{\mu(ut)}}{1+t} > u \right),$$

see Lemma 5.14. Hence, in some sense, the variance $\sigma^2(\mu(ut))$ of $Y_{\mu(ut)}$ determines the *length* of the 'most likely' hitting interval by the requirement that $\sigma(\mu(u))/\delta(u) \to 0$.

Not only the length of this interval plays a role in the asymptotics of (5.2). There is one other important element: the *local correlation structure* of the process on $[t_u^* \pm \delta(u)/u]$. Traditionally, it was assumed that $\mathbb{V}\mathrm{ar}\left(Y_{\mu(us)}/\sigma(\mu(us)) - Y_{\mu(ut)}/\sigma(\mu(ut))\right)$ behaves locally like $|s-t|^\alpha$ for some $\alpha \in (0,2]$, as in Section 3.3. It was soon realized that $|s-t|^\alpha$ can be replaced by a regularly varying function (at zero) with minimal additional effort [267]; see also [41, 87, 163], to mention a few recent contributions.

However, by imposing such a correlation structure, it is impossible to find the asymptotics of (5.1) for a general Gaussian process with stationary increments, for instance. We solve this problem by introducing two wide classes of correlation structures, resulting in qualitatively different asymptotics in four cases. These specific structures must be imposed to be able to perform explicit calculations. The main novelty of this chapter is that the local behavior may depend on u. Our framework is specific enough to derive generalities, yet general enough to include many interesting processes as special cases (to our best knowledge, all processes are covered for which the asymptotics of (5.1) appear in the literature; see the examples in Section 5.3.3).

Often there is a third element that plays a role in the asymptotics: the *local variance structure* of $Y_{\mu(ut)}/(1+t)$ near $t = t_u^*$. By the structure of the problem and the differentiability

assumptions that we will impose on σ and μ, this third element is only implicitly present in our analysis. However, if one is interested in the asymptotics of some probability different from (5.1), it may play a role. In that case, the reasoning of the present chapter is readily adapted.

We now introduce the first family of correlation structures, leading to three different types of asymptotics. Suppose that the following holds:

$$\sup_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ s\neq t}} \left| \frac{\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))} - \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right)}{\mathcal{D}\tau^2(|\nu(us)-\nu(ut)|)/\tau^2(\nu(u))} - 1 \right| \to 0, \tag{5.3}$$

as $u \to \infty$, where $\mathcal{D}$ is some constant and τ and ν are suitable functions. It is assumed that τ and ν are regularly varying at infinity with indices $\iota_\tau \in (0,1)$ and $\iota_\nu > 0$ respectively. To gain some intuition, suppose that ν is the identity, and write $\tau(t) = \ell(t)t^{\iota_\tau}$ for some slowly varying function at infinity ℓ. The denominator in (5.3) then equals $\mathcal{D}|s-t|^{2\iota_\tau}\ell^2(u|s-t|)/\ell^2(u)$. From the analysis of the problem it follows that one must consider $|s-t| \leq \Delta(u)/u$, where Δ is some function satisfying $\Delta(u) = o(\delta(u))$. As a result, the denominator is of the order $[\Delta(u)/u]^{2\iota_\tau}\ell^2(\Delta(u))/\ell^2(u)$; due to the term $\ell^2(\Delta(u))$, three cases can now be distinguished: Δ tends to infinity, to a constant, or to zero. Interestingly, the Pickands' constant appearing in the asymptotics is determined by the behavior of τ at infinity in the first case, and at zero in the last case (one needs an additional assumption on the behavior of τ at zero). The second 'intermediate' case is special, resulting in the appearance of a so-called generalized Pickands' constant.

The second family of correlation structures, resulting in the fourth type of asymptotics, is given by

$$\sup_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ s\neq t}} \left| \frac{\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))} - \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right)}{\tau^2(|\nu(us)-\nu(ut)|/\nu(u))} - 1 \right| \to 0, \tag{5.4}$$

where ν is regularly varying at infinity with index $\iota_\nu > 0$ and τ is regularly varying at zero with index $\tilde{\iota}_\tau \in (0,1)$ (the tilde emphasizes that we consider regular variation at zero). A detailed description of the assumptions on each of the functions are given in Section 5.2.2. Here, if ν is the identity, the denominator equals $\ell^2(|s-t|)|s-t|^{2\tilde{\iota}_\tau}$ for some slowly varying function at zero ℓ. Therefore, it cannot be written in the form (5.3) unless ℓ is constant.

Having introduced the four cases informally, we now present them in more detail. The cases are referred to as case A, B, C, and D; the precise underlying assumptions are given in Section 5.2.2. We set

$$\mathcal{G} := \lim_{u\to\infty} \frac{\sigma(\mu(u))\tau(\nu(u))}{u}, \tag{5.5}$$

assuming the limit exists.

A. Case A applies when (5.3) holds and $\mathcal{G} = \infty$.

B. Case B applies when (5.3) holds and $\mathcal{G} \in (0,\infty)$.

C. Case C applies when (5.3) holds and $\mathcal{G} = 0$. We then also suppose that τ be regularly varying at zero with index $\tilde{\iota}_\tau \in (0,1)$.

D. Case D applies when (5.4) holds.

In order to formulate the main result of this chapter, it is convenient to introduce the notation

$$C_{H,\beta,\iota_\nu,\iota_\tau} := \sqrt{2^{1-1/\iota_\tau}\pi\iota_\nu}\left(\frac{\beta}{H}\right)^{1/\iota_\tau}\left(\frac{H}{\beta-H}\right)^{\iota_\nu+\frac{H}{\beta}-\frac{1}{2}+\frac{1}{\iota_\tau}\left(1-\frac{H}{\beta}\right)}$$

and, for case B,

$$\mathcal{M} := \frac{\beta^2}{2\mathcal{G}^2 H^{2H/\beta}(\beta - H)^{2-2H/\beta}},$$

where $\mathcal{G} \in (0, \infty)$ is defined as in (5.5). Recall the definitions of Ψ and $\mathcal{H}_\eta$ in (3.1) and (3.8). Here is our main result; the assumptions are detailed in Section 5.2.2.

Theorem 5.1 *Let μ and σ satisfy assumptions* **M1**–**M4** *and* **S1**–**S4** *below for some $\beta > H$. In case A, i.e., when* **A1**, **A2**, **T1**, **T2**, **N1**, **N2** *below hold, we have*

$$\mathbb{P}\left(\sup_{t\geq 0} Y_{\mu(t)} - t > u\right) \sim \mathcal{H}_{B_{\iota_\tau}} C_{H,\beta,\iota_\nu,\iota_\tau} \sqrt{\mathcal{D}^{1/\iota_\tau}} \frac{\sigma(\mu(u))\nu(u)}{u^{\overleftarrow{\tau}}\left(\frac{\sigma(\mu(u))\tau(\nu(u))}{u}\right)} \Psi\left(\inf_{t\geq 0} \frac{u(1+t)}{\sigma(\mu(ut))}\right).$$

In case B, i.e., when **B1**, **B2**, **T1**, **T2**, **N1**, **N2** *below hold, then $\mathcal{H}_{\mathcal{D}\mathcal{M}^{\tau^2}}$ exists and we have*

$$\mathbb{P}\left(\sup_{t\geq 0} Y_{\mu(t)} - t > u\right) \sim \mathcal{H}_{\mathcal{D}\mathcal{M}^{\tau^2}} \sqrt{2\pi}\iota_\nu \left(\frac{H}{\beta - H}\right)^{\iota_\nu + \frac{H}{\beta} - \frac{1}{2}} \frac{\sigma(\mu(u))\nu(u)}{u} \Psi\left(\inf_{t\geq 0} \frac{u(1+t)}{\sigma(\mu(ut))}\right).$$

In case C, i.e., when **C1**–**C3**, **T1**, **N1**, **N2** *below hold, we have*

$$\mathbb{P}\left(\sup_{t\geq 0} Y_{\mu(t)} - t > u\right) \sim \mathcal{H}_{B_{\tilde{\iota}_\tau}} C_{H,\beta,\iota_\nu,\tilde{\iota}_\tau} \sqrt{\mathcal{D}^{1/\tilde{\iota}_\tau}} \frac{\sigma(\mu(u))\nu(u)}{u^{\overleftarrow{\tau}}\left(\frac{\sigma(\mu(u))\tau(\nu(u))}{u}\right)} \Psi\left(\inf_{t\geq 0} \frac{u(1+t)}{\sigma(\mu(ut))}\right).$$

In case D, i.e., when **D1**, **D2**, **N1**, **N2** *below hold, we have*

$$\mathbb{P}\left(\sup_{t\geq 0} Y_{\mu(t)} - t > u\right) \sim \mathcal{H}_{B_{\tilde{\iota}_\tau}} C_{H,\beta,\iota_\nu,\tilde{\iota}_\tau} \frac{\sigma(\mu(u))}{u^{\overleftarrow{\tau}}\left(\frac{\sigma(\mu(u))}{u}\right)} \Psi\left(\inf_{t\geq 0} \frac{u(1+t)}{\sigma(\mu(ut))}\right).$$

Observe that $\overleftarrow{\tau}$ is an asymptotic inverse of τ at infinity in case A, and at zero in case C and D (see Section 2.1 for definitions). Hence, the factors preceding the function Ψ are regularly varying with index $(H/\beta + \iota_\nu\iota_\tau - 1)(1 - 1/\iota_\tau) + (1 - \iota_\tau)\iota_\nu$ in case A, with index $H/\beta + \iota_\nu - 1$ in case B, with index $H/\beta + \iota_\nu - 1 - (H/\beta + \iota_\tau\iota_\nu - 1)/\tilde{\iota}_\tau$ in case C, and with index $(H/\beta - 1)(1 - 1/\tilde{\iota}_\tau)$ in case D. Note that case B is special in a number of ways: a nonclassical Pickands' constant is present and no inverse appears in the formula.

We now formally state the underlying assumptions.

5.2.2 Assumptions

Two types of assumptions are distinguished: general assumptions and case-specific assumptions. The general assumptions involve the variance σ^2 of Y, the time-change μ, and the functions ν and τ appearing in (5.3) and (5.4). The case-specific assumptions formalize the four regimes introduced in the previous subsection.

General assumptions

We start by stating the assumptions on μ.

M1 μ is regularly varying at infinity with index $1/\beta$,

M2 μ is strictly increasing, $\mu(0) = 0$,

M3 μ is ultimately continuously differentiable and its derivative $\dot{\mu}$ is ultimately monotone,

M4 μ is twice continuously differentiable and its second derivative $\ddot{\mu}$ is ultimately monotone.

Assumption **M2** is needed to ensure that the probabilities (5.1) and (5.2) be equal. The remaining conditions imply that $\beta u\dot{\mu}(u) \sim \mu(u)$ and $\beta^2 u^2 \ddot{\mu}(u) \sim (1-\beta)\mu(u)$, see Exercise 1.11.13 of Bingham *et al.* [51]. In particular, $\dot{\mu}$ and $\ddot{\mu}$ are regularly varying with index $1/\beta - 1$ and $1/\beta - 2$ respectively.

Now we formulate the assumptions on σ and one assumption on both μ and σ.

S1 σ is continuous and regularly varying at infinity with index H for some $H \in (0,1)$,

S2 σ^2 is ultimately continuously differentiable and its first derivative $\dot{\sigma}^2$ is ultimately monotone,

S3 σ^2 is ultimately twice continuously differentiable and its second derivative $\ddot{\sigma}^2$ is ultimately monotone,

S4 there exist some $T, \epsilon > 0$, $\gamma \in (0,2]$ such that

1. $\limsup_{u\to\infty} \sup_{s,t\in(0,(1+\epsilon)T^{1/\beta}]} \mathbb{V}\mathrm{ar}(Y_{us} - Y_{ut})\sigma^{-2}(u)|s-t|^{-\gamma} < \infty$ and
2. $\limsup_{u\to\infty} \frac{\sigma^2(\mu(u))}{u^2} \log \mathbb{P}\left(\sup_{t\geq T} \frac{Y_{\mu(ut)}}{1+t} > u\right) < -\frac{1}{2}\frac{(1+t^*)^2}{(t^*)^{H/\beta}}$.

We emphasize that $\dot{\sigma}^2$ denotes the derivative of σ^2, and not the square derivative of σ. As before, conditions **S1**–**S3** imply that $u\dot{\sigma}^2(u) \sim 2H\sigma^2(u)$ and $u^2\ddot{\sigma}^2(u) \sim 2H(2H-1)\sigma^2(u)$. The first point of **S4**, which is Kolmogorov's weak convergence criterion, ensures the existence of a modification with continuous sample paths, as in Lemma 4.7; we always assume to work with this modification. The second point of **S4** ensures that the probability $\mathbb{P}(\sup_{t\geq uT} Y_{\mu(t)} - t > u)$ cannot dominate the asymptotics. We choose to formulate this as an assumption, although it is possible to give sharp conditions for **S4**.2 to hold. However, these conditions look relatively complicated, while the second point is in general easier to verify on a case by case basis. In the next section, we show that it holds for processes with stationary increments and self-similar processes.

Note that if **M1**–**M4** and **S1**–**S4** hold, the first and second derivative of $\sigma^2(\mu(\cdot))$ are also regularly varying, with indices $2H/\beta-1$ and $2H/\beta-2$ respectively. It is this fact that guarantees the existence of the limits that are implicitly present in the notation '$\sim$' in Theorem 5.1.

The function ν appearing in (5.3) and (5.4) also has to satisfy certain assumptions, which are similar to those imposed on μ.

N1 ν is regularly varying at infinity with index $\iota_\nu > 0$,

N2 ν is ultimately continuously differentiable and its derivative $\dot{\nu}$ is ultimately monotone.

Finally, we formulate the assumptions on τ in (5.3) or (5.4).

T1 τ is continuous and regularly varying at infinity with index ι_τ for some $\iota_\tau \in (0,1)$,

T2 $\tau(t) \leq Ct^{\gamma'}$ on a neighborhood of zero for some $C, \gamma' > 0$.

Assumption **T2** is essential to prove uniform tightness at some point in the proof, which yields the existence of the Pickands' constants.

Case-specific assumptions

We now formulate the case-specific assumptions in each of the cases A, B, C, and D. These assumptions are also mentioned in the introduction, but it is convenient to label them for reference purposes. If we write that the correlation structure is determined by (5.3) or (5.4), the function δ is supposed to satisfy $\delta(u) = o(u)$ and $\sigma(\mu(u)) = o(\delta(u))$ as $u \to \infty$.

After recalling the definition of $\mathcal{G}$ in (5.5), we start with case A.

A1 The correlation structure is determined by (5.3),

A2 $\mathcal{G} = \infty$.

Similar conditions are imposed in case B.

B1 The correlation structure is determined by (5.3),

B2 $\mathcal{G} \in (0, \infty)$.

In case C, we need an additional condition (**C3**). Note that the index of variation in **C3** appears at several places in the asymptotics, cf. Theorem 5.1. It also implies the existence of an asymptotic inverse $\overleftarrow{\tau}$ at zero, cf. Lemma 2.6.

C1 The correlation structure is determined by (5.3),

C2 $\mathcal{G} = 0$,

C3 τ is regularly varying at zero with index $\tilde{\iota}_\tau \in (0, 1)$.

Case D is slightly different from the previous three cases. As in case C, regular variation of τ at zero plays a role.

D1 The correlation structure is determined by (5.4),

D2 τ is regularly varying at zero with index $\tilde{\iota}_\tau \in (0, 1)$.

5.3 Special cases: processes with stationary increments and self-similar processes

In this section, we apply Theorem 5.1 to calculate the asymptotics of (5.2) for two specific cases: (i) Y has stationary increments and (ii) Y is self-similar. In both examples, the imposed assumptions imply that $\sigma^2(0) = 0$, so that $Y_0 = 0$ almost surely.

In case Y has stationary increments, the finite-dimensional distributions are completely determined by the variance function σ^2. For self-similar processes, (5.2) has been studied by Hüsler and Piterbarg [163]. We show that their results are reproduced and even slightly generalized by Theorem 5.1.

We conclude this section with some examples that have been studied in the literature.

5.3.1 Stationary increments

Since σ determines the finite-dimensional distributions of Y, it also fixes the local correlation structure; we record this in the next proposition. To get some feeling for the result, observe that for $s, t \in [t_u^* \pm \delta(u)/u]$,

$$\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))} - \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right) \approx \frac{\mathbb{V}\mathrm{ar}\left(Y_{\mu(us)} - Y_{\mu(ut)}\right)}{\sigma^2(\mu(ut^*))} = \frac{\sigma^2(|\mu(us) - \mu(ut)|)}{\sigma^2(\mu(ut^*))}.$$

This intuitive reasoning is now made precise. Note the proposition also entails that case D does *not* occur in this setting.

Proposition 5.2 *Let* **S1**–**S2**, **M1**–**M3** *hold for some* $\beta > H$. *Let* δ *be regularly varying with index* $\iota_\delta \in (1-1/\beta, 1)$. *Then (5.3) holds with* $\tau = \sigma$, $\nu = \mu$ *and* $\mathcal{D} = (t^*)^{-2H/\beta}$.

Proof. Since $s, t \in [t_u^* \pm \delta(u)/u]$, we have by the UCT {**S1**, **M1**},

$$\lim_{u\to\infty} \sup_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ s\neq t}} \left| \frac{\sigma^2(\mu(u))}{\mathcal{D}\sigma(\mu(us))\sigma(\mu(ut))} - 1 \right| = 0.$$

Moreover, the stationarity of the increments implies that

$$2\left[\sigma(\mu(us))\sigma(\mu(ut)) - \mathbb{C}\mathrm{ov}\left(Y_{\mu(us)}, Y_{\mu(ut)}\right)\right] = \sigma^2(|\mu(us)-\mu(ut)|) - [\sigma(\mu(us)) - \sigma(\mu(ut))]^2.$$

Hence, it suffices to prove that

$$\lim_{u\to\infty} \sup_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ s\neq t}} \frac{[\sigma(\mu(us)) - \sigma(\mu(ut))]^2}{\sigma^2(|\mu(us)-\mu(ut)|)} = 0. \tag{5.6}$$

For this, observe that the left-hand side of (5.6) is majorized by $t_1(u)t_2(u)$, where

$$\begin{aligned} t_1(u) &:= \sup_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ s\neq t}} \frac{[\sigma(\mu(us)) - \sigma(\mu(ut))]^2}{[\mu(us)-\mu(ut)]^2}, \\ t_2(u) &:= \sup_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ s\neq t}} \frac{[\mu(us)-\mu(ut)]^2}{\sigma^2(|\mu(us)-\mu(ut)|)}. \end{aligned}$$

As for $t_1(u)$, by the mean value theorem {**S2**, **M3**} there exist $t^\wedge(u,s,t), t^\vee(u,s,t)$ such that, for u large enough,

$$t_1(u) = \sup_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ s\neq t}} \frac{[\tilde\sigma_\mu(ut^\wedge(u,s,t))]^2}{[\dot\mu(ut^\vee(u,s,t))]^2} \le \left(\frac{\sup_{t\in[t_u^*\pm\delta(u)/u]} \tilde\sigma_\mu(ut)}{\inf_{t\in[t_u^*\pm\delta(u)/u]} \dot\mu(ut)} \right)^2,$$

where $\tilde\sigma_\mu(\cdot)$ denotes the derivative of $\sigma(\mu(\cdot))$. As a consequence of the UCT {**M1**, **M3**, **S1**, **S2**}, $t_1(u)$ can therefore be upper bounded by $C'\sigma^2(\mu(u))/\mu^2(u)$ for some constant $C' < \infty$.

We now turn to $t_2(u)$. A substitution {**M2**} shows that

$$t_2(u) = \sup_{\substack{s,t\in[\mu(ut_u^*-\delta(u)),\mu(ut_u^*+\delta(u))]\\ s>t}} \frac{(s-t)^2}{\sigma^2(s-t)} = \sup_{0<t\le\mu(ut_u^*+\delta(u))-\mu(ut_u^*-\delta(u))} \frac{t^2}{\sigma^2(t)}.$$

Observe that, again by the mean value theorem and the UCT {**M1**, **M3**},

$$\mu(ut_u^*+\delta(u)) - \mu(ut_u^*-\delta(u)) \le 2 \sup_{t\in[t_u^*\pm\delta(u)/u]} \dot\mu(ut)\delta(u) \sim \frac{2}{\beta}(t^*)^{1/\beta-1}\mu(u)\delta(u)/u,$$

which tends to infinity by the assumption on ι_δ.

Suppose for the moment that the mapping $x \mapsto x^2/\sigma^2(x)$ is bounded on sets of the form $(0,\cdot]$. Since it is regularly varying with index $2-2H > 0$ {**S1**}, we have by the UCT and the assumption that $\iota_\delta > 1-1/\beta$, for u large enough,

$$t_2(u) \le \sup_{0<t\le 3/\beta(t^*)^{1/\beta-1}} \frac{[\mu(u)\delta(u)/u]^2t^2}{\sigma^2(\mu(u)\delta(u)/ut)} \sim \left(\frac{3}{\beta}(t^*)^{1/\beta-1}\right)^{2-2H} \frac{[\mu(u)\delta(u)/u]^2}{\sigma^2(\mu(u)\delta(u)/u)}.$$

In conclusion, there exists a constant $\mathcal{K} < \infty$ such that

$$\sup_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ s\neq t}} \frac{[\sigma(\mu(us))-\sigma(\mu(ut))]^2}{\sigma^2(|\mu(us)-\mu(ut)|)} \leq \mathcal{K}\frac{\sigma^2(\mu(u))\delta^2(u)/u^2}{\sigma^2(\mu(u)\delta(u)/u)},$$

which is regularly varying with index $2(1-H)(\iota_\delta - 1) < 0$, so that (5.6) follows.

It remains to show that $x \mapsto x^2/\sigma^2(x)$ is locally bounded. To see this, we use an argument introduced by Dębicki [87, Lem. 2.1]. By **S2**, one can select some (large) $s \geq 0$ such that σ^2 is continuously differentiable at s. Then, for some small $x > 0$,

$$\sigma^2(s) - \sigma^2(s-x) \leq \sigma^2(s) + \sigma^2(x) - \sigma^2(s-x) = 2\mathbb{C}\mathrm{ov}(Y_s, Y_x) \leq 2\sigma(s)\sigma(x),$$

and by the mean value theorem there exists some $\rho_x \in [s-x, s]$ such that $\sigma^2(s) - \sigma^2(s-x) = \dot{\sigma}^2(\rho_x)x$. By continuity of $\dot{\sigma}^2$ at s, we have

$$\limsup_{x\downarrow 0} \frac{x}{\sigma(x)} \leq \limsup_{x\downarrow 0} 2\frac{\sigma(s)}{\dot{\sigma}^2(\rho_x)} = 2\frac{\sigma(s)}{\dot{\sigma}^2(s)} < \infty.$$

The claim follows upon combining this observation with **S1**. □

Lemma 5.3 *Let Y have stationary increments, and suppose that* **S1** *and* **M1** *hold. If $\sigma^2(t) \leq Ct^\gamma$ on a neighborhood of zero for some $C, \gamma > 0$, then* **S4** *holds.*

Proof. By the stationarity of the increments, the first point of **S4** follows immediately from the UCT for $t \mapsto \sigma^2(t)t^{-\gamma}$ (this mapping is locally bounded by the condition in the lemma). In fact, it holds for all $T, \epsilon > 0$.

To check the second requirement of **S4**, select some ω such that $H/\beta < \omega < 1$. By the UCT {**M1**},

$$\lim_{T\to\infty}\lim_{u\to\infty}\sup_{t\geq T}\frac{ut^\omega}{\overleftarrow{\mu}(\mu(u)t^{1/\beta})} = \lim_{T\to\infty} T^{\omega-1} = 0.$$

Hence, we may suppose without loss of generality that T is such that $\overleftarrow{\mu}(\mu(u)t^{1/\beta})/u \geq t^\omega$ for every $t \geq T$ and large u. This implies that

$$\mathbb{P}\left(\sup_{t\geq T}\frac{Y_{\mu(ut)}}{1+t} > u\right) \leq \mathbb{P}\left(\sup_{t\geq[\mu(uT)/\mu(u)]^\beta}\frac{Y_{\mu(u)t^{1/\beta}}}{1+t^\omega} > u\right) \leq \mathbb{P}\left(\sup_{t\geq T/2}\frac{Y_{\mu(u)t^{1/\beta}}}{1+t^\omega} > u\right).$$

We now apply some results from Chapter 4. By Corollary 4.16 and the arguments in the proof of Proposition 4.1, we have

$$\limsup_{u\to\infty}\frac{\sigma^2(\mu(u))}{u^2}\log\mathbb{P}\left(\sup_{t\geq T}\frac{Y_{\mu(ut)}}{1+t} > u\right) \leq -\frac{1}{2}\inf_{t\geq T/2}\frac{(1+t^\omega)^2}{t^{2H/\beta}}.$$

Note that we have used the continuity of the functional $x \mapsto \sup_{t\geq(T/2)^{1/\beta}} x(t)/(1+t^{\omega\beta})$ in the topology on $\Omega^{\omega\beta}$, cf. Lemma 4.8 and Corollary 4.16. The claim is obtained by choosing T large enough, which is possible since $t^{2\omega}/t^{2H/\beta} \to \infty$ as $t \to \infty$. □

With Proposition 5.2 and Lemma 5.3 at our disposal, we readily find the asymptotics of (5.1) when Y has stationary increments.

Proposition 5.4 *Let Y have stationary increments. Suppose that* **S1**–**S3** *hold, and that $\sigma^2(t) \leq Ct^\gamma$ on a neighborhood of zero for some $C, \gamma > 0$. Moreover, suppose that* **M1**–**M4** *hold for some $\beta > H$.*

If $\sigma^2(\mu(u))/u \to \infty$, *then*

$$\mathbb{P}\left(\sup_{t\geq 0} Y_{\mu(t)} - t > u\right) \sim \mathcal{H}_{B_H} C_{H,\beta,1/\beta,H} \left(\frac{\beta - H}{H}\right)^{\frac{1}{\beta}} \frac{\sigma(\mu(u))\mu(u)}{u \overleftarrow{\sigma}\left(\frac{\sigma^2(\mu(u))}{u}\right)} \Psi\left(\inf_{t\geq 0} \frac{u(1+t)}{\sigma(\mu(ut))}\right).$$

If $\sigma^2(\mu(u))/u \to \mathcal{G} \in (0,\infty)$, *then*

$$\mathbb{P}\left(\sup_{t\geq 0} Y_{\mu(t)} - t > u\right) \sim \mathcal{H}_{(2/\mathcal{G}^2)\sigma^2} \frac{\sqrt{\pi/2}}{H} \frac{\sigma(\mu(u))\mu(u)}{u} \Psi\left(\inf_{t\geq 0} \frac{u(1+t)}{\sigma(\mu(ut))}\right).$$

If $\sigma^2(\mu(u))/u \to 0$ *and* σ *is regularly varying at zero with index* $\lambda \in (0,1)$, *then*

$$\mathbb{P}\left(\sup_{t\geq 0} Y_{\mu(t)} - t > u\right) \sim \mathcal{H}_{B_\lambda} C_{H,\beta,1/\beta,\lambda} \left(\frac{\beta - H}{H}\right)^{\frac{H}{\beta\lambda}} \frac{\sigma(\mu(u))\mu(u)}{u \overleftarrow{\sigma}\left(\frac{\sigma^2(\mu(u))}{u}\right)} \Psi\left(\inf_{t\geq 0} \frac{u(1+t)}{\sigma(\mu(ut))}\right).$$

Proof. Directly from Theorem 5.1. For the case $\sigma^2(\mu(u))/u \to \mathcal{G} \in (0,\infty)$, observe that necessarily $2H = \beta$. □

5.3.2 Self-similarity

We now suppose that Y is a self-similar process with Hurst parameter H, i.e., $\mathbb{V}\mathrm{ar}(Y_t) = t^{2H}$ and for any $\alpha > 0$ and $s, t \geq 0$,

$$\mathbb{C}\mathrm{ov}\left(Y_{\alpha t}, Y_{\alpha s}\right) = \alpha^{2H} \mathbb{C}\mathrm{ov}\left(Y_t, Y_s\right), \tag{5.7}$$

cf. (1.12). The self-similarity property has been observed statistically in several types of data traffic, see Section 1.3 for more details. Two examples of self-similar Gaussian processes are the fractional Brownian motion and the Riemann-Liouville process.

Another (undoubtedly related) reason why self-similar processes are interesting is that the weak limit obtained by scaling a process both in time and space must be self-similar (if it exists); see Lamperti [209]. In the setting of Gaussian processes with stationary increments, a strong type of weak convergence is studied in Chapter 4. We also mention the interesting fact that self-similar processes are closely related to stationary processes by the so-called Lamperti-transformation; see [7] for more details.

We make the following assumption about the behavior of the (standardized) variance of Y near $t = t^*$: for some function τ which is regularly varying at zero with index $\tilde{\iota}_\tau \in (0,1)$,

$$\lim_{s,t\to t^*} \frac{\mathbb{V}\mathrm{ar}\left(\frac{Y_s}{s^{H/\beta}} - \frac{Y_t}{t^{H/\beta}}\right)}{\tau^2(|s-t|)} = 1. \tag{5.8}$$

By the self-similarity, one may equivalently require that a similar condition holds for s, t tending to an arbitrary strictly positive number; see [163]. In the proof of Proposition 5.5 below we show that (5.8) implies that self-similar processes are covered by case D.

We also need the following assumption on the variance structure of Y: for some $\gamma > 0$,

$$\sup_{s,t\in(0,1]} \mathbb{V}\mathrm{ar}(Y_s - Y_t)|s-t|^{-\gamma} < \infty. \tag{5.9}$$

This Kolmogorov criterion ensures that there exists a continuous modification of Y. Notice that without loss of generality it suffices to take the supremum over any interval $(0,\cdot]$ by the self-similarity.

The following proposition generalizes Theorem 1 of Hüsler and Piterbarg [163]; it is left to the reader to check that the formulas indeed coincide when $\phi(t) = ct^\beta$ for some $c > 0$. Although no condition of the type (5.9) appears in [163], it is implicitly present; the process $\tilde{Z}$ in [163] is claimed to satisfy condition (E3) on page 19 of [257].

Proposition 5.5 *Let Y be self-similar with Hurst parameter H, and let μ satisfy* **M1**–**M4** *for some $\beta > H$. If (5.8) and (5.9) hold, then*

$$\mathbb{P}\left(\sup_{t\geq 0} Y_{\mu(t)} - t > u\right) \sim \mathcal{H}_{B_{\tilde{\iota}_\tau}} C_{H,\beta,1,\tilde{\iota}_\tau} \frac{\mu(u)^H}{u \overleftarrow{\tau}\left(\frac{\mu(u)^H}{u}\right)} \Psi\left(\inf_{t\geq 0} \frac{u(1+t)}{\mu(ut)^H}\right).$$

Proof. Note that by (5.8), for δ with $\delta(u) = o(u)$,

$$\lim_{u\to\infty} \sup_{s,t\in[t_u^*\pm\delta(u)/u]} \left| \frac{\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)/\mu(u)}}{(\mu(us)/\mu(u))^H} - \frac{Y_{\mu(ut)/\mu(u)}}{(\mu(ut)/\mu(u))^H}\right)}{\tau^2(|\mu(us)^\beta - \mu(ut)^\beta|/\mu(u)^\beta)} - 1 \right| = 0.$$

The self-similarity implies

$$\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)/\mu(u)}}{(\mu(us)/\mu(u))^H} - \frac{Y_{\mu(ut)/\mu(u)}}{(\mu(ut)/\mu(u))^H}\right) = \mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\mu(us)^H} - \frac{Y_{\mu(ut)}}{\mu(ut)^H}\right),$$

so that (5.4) holds for $\nu(t) = \mu(t)^\beta$ and the τ of (5.8); then we have **N1** and **N2** as a consequence of the assumption that **M1**–**M3** hold. Moreover, it is trivial that $\sigma^2(t) = t^{2H}$ satisfies **S1**–**S3**. We now show that **S4** holds. By the self-similarity, for any $T > 0$,

$$\sup_{s,t\in(0,T]} \frac{\mathbb{V}\mathrm{ar}(Y_{us} - Y_{ut})}{u^{2H}|s-t|^\gamma} = T^{2H-\gamma} \sup_{s,t\in(0,1]} \frac{\mathbb{V}\mathrm{ar}(Y_s - Y_t)}{|s-t|^\gamma},$$

so that the first condition of **S4** is satisfied due to (5.9). As for the second point, by the self-similarity and the reasoning in the proof of Lemma 5.3, it suffices to show that for large T

$$\limsup_{u\to\infty} \frac{\mu(u)^{2H}}{u^2} \log \mathbb{P}\left(\sup_{t\geq T/2} \frac{Y_{t^{1/\beta}}}{1+t^\omega} > \frac{u}{\mu(u)^H}\right) < -\frac{1}{2}\frac{(1+t^*)^2}{(t^*)^{H/\beta}},$$

for some ω satisfying $H/\beta < \omega < 1$. This follows from Borell's inequality (Lemma 3.3) once it has been shown that $Y_t/t^{\omega\beta} \to 0$ as $t \to \infty$. We use a reasoning as in Lemma 4.13 to see that this is the case. First, one can exploit the fact that $\omega\beta > H$ to establish $\lim_{k\to\infty} Y_{2^k}/2^{k\omega\beta} = 0$ by the Borel-Cantelli lemma. It then suffices to show that also $Z_k/2^{k\omega\beta} \to 0$, where $Z_k := \sup_{s\in[2^k,2^{k+1}]} |Y_s - Y_{2^k}|$. Note that Z_k has the same distribution as $2^{kH}Z_0$ by the self-similarity of Y. The almost sure convergence follows again from the Borel-Cantelli lemma: for $\alpha, \epsilon > 0$,

$$\sum_k \mathbb{P}(Z_k/2^{k\omega\beta} > \epsilon) \leq \sum_k \mathbb{P}(Z_0 > \epsilon 2^{k(\omega\beta-H)}) \leq \sum_k \exp\left(-\alpha\epsilon^2 2^{2k(\omega\beta-H)}\right) \mathbb{E}\exp\left(\alpha Z_0^2\right).$$

If one chooses $\alpha > 0$ appropriately small, $\mathbb{E}\exp\left(\alpha Z_0^2\right)$ is finite as a consequence of Borell's inequality (Lemma 3.3; it can be applied since Y is continuous).

In conclusion, case D applies and the asymptotics are given by Theorem 5.1. □

Hüsler and Piterbarg [163, Sec. 3] also consider a class of Gaussian processes that behave somewhat like self-similar processes. Although we do not work this out, this class is also covered by (case D of) Theorem 5.1; note that their condition (18) is a special case of (5.4), for $\nu(t) = t$.

5.3.3 Examples

We now work out some examples that appear in the literature. In all examples, we obtain extensions of what is known already. For Gaussian integrated processes (Section 5.3.3), we also remove some technical conditions.

Fractional Brownian motion

In some sense, fractional Brownian motion (fBm) is the easiest instance of a process Y that fits into the framework of Proposition 5.4. Indeed, the variance function σ^2 of fBm is the canonical regularly varying function, $\sigma^2(t) = t^{2H}$ for some $H \in (0,1)$.

A fractional Brownian motion B_H is self-similar in the sense of (5.7). Therefore, it can appear as a weak limit of a time- and space-scaled process; for examples, see Chapter 4 and [306]. The increments of a fractional Brownian motion are *long-range dependent* if and only if $H > 1/2$, i.e., the covariance function of the increments on an equispaced grid is then nonsummable. For more details on long-range dependence and an extensive list of references, see Doukhan *et al.* [122].

As fBm is both self-similar and has stationary increments, the asymptotics can be obtained by applying either Proposition 5.4 or Proposition 5.5. Interestingly, this implies that it should be possible to write the formulas in the three cases of Proposition 5.4 as a single formula for fBm. The proof given below is based on Proposition 5.4, but the reader easily verifies that Proposition 5.5 yields the same formula; one then uses the fact that

$$\left(\frac{\beta - H}{\beta}\right)^{1/\beta} C_{H,\beta,1/\beta,H} = \frac{\beta - H}{\beta H} C_{H,\beta,1,H}.$$

Note that fBm is the only process for which both Proposition 5.4 and 5.5 can be applied: it is the only Gaussian self-similar process with stationary increments.

Corollary 5.6 *Let B_H be a fractional Brownian motion with Hurst parameter $H \in (0,1)$. If μ satisfies conditions* **M1**–**M4** *for some $\beta > H$, then*

$$\mathbb{P}\left(\sup_{t\geq 0} B_H(\mu(t)) - t > u\right) \sim \mathcal{H}_{B_H} C_{H,\beta,1/\beta,H} \left(\frac{\beta - H}{H}\right)^{1/\beta} \frac{u^{1/H-1}}{\mu(u)^{1-H}} \Psi\left(\inf_{t\geq 0} \frac{u(1+t)}{\mu(ut)^H}\right).$$

Proof. First note that $\mu(u)^{2H}/u$ has a limit in $[0,\infty]$ as a consequence of **M2**. If $\mu(u)^{2H}/u$ tends to either zero or infinity, the formula follows readily from Proposition 5.4 by setting $\sigma^2(t) = t^{2H}$ (so that $\lambda = H$ in case C). In case $\mu(u)^{2H}/u \to \mathcal{G} \in (0,\infty)$, the generalized Pickands' constant can be expressed in a classical one by exploiting the self-similarity of B_H; one easily checks that $\mathcal{H}_{(\sqrt{2}/\mathcal{G})B_H} = (\sqrt{2}/\mathcal{G})^{1/H} \mathcal{H}_{B_H}$. Now note that $\beta = 2H$ and that

$$\frac{\mu(u)^{H+1}}{u} \sim \mathcal{G}^{1/H} \frac{u^{1/H-1}}{\mu(u)^{1-H}},$$

and the assertion follows. □

For a standard Brownian motion ($H = 1/2$), Pickands' constant equals $\mathcal{H}_{B_{1/2}} = 1$, so that the formula reduces to

$$\mathbb{P}\left(\sup_{t\geq 0} B_{\mu(t)} - t > u\right) \sim 2\sqrt{2\pi}\beta(2\beta-1)^{\frac{1}{2}(1/\beta-3)} \frac{u}{\sqrt{\mu(u)}} \Psi\left(\inf_{t\geq 0} \frac{u(1+t)}{\sqrt{\mu(ut)}}\right). \tag{5.10}$$

This probability has been extensively studied in the literature; the whole distribution of $\sup_{t\geq 0} B_{\mu(t)} - t$ is known in a number of cases. We refer to some recent contributions [82, 128, 158] for background and references.

The tail asymptotics for $\sup_{t\geq 0} B_{\mu(t)} - t$ are studied in Dębicki [86], but we believe that formula (5.10) does not appear elsewhere in the literature.

Gaussian integrated process

A Gaussian integrated process Y has the form

$$Y_t = \int_0^t Z_s ds, \tag{5.11}$$

where Z is a centered stationary Gaussian process with covariance function R, see Section 3.3. We suppose that R be ultimately continuous and that $R(0) > 0$. It is easy to see that

$$\sigma^2(t) = 2\int_0^t \int_0^s R(v)dvds.$$

In the literature, μ is assumed to be of the form $\mu(t) = t/c$ for some $c > 0$, so that **M1**–**M4** obviously hold. For an easy comparison, we also adopt this particular choice for μ here (elementary scaling arguments show that we may have assumed $c = 1$ without loss of generality). Evidently, the results of this chapter allow for much more general drift functions, and the reader has no difficulties to write out the corresponding formula.

The structure of the problem ensures that **S2** and **S3** hold, and that $\sigma(t) \leq Ct^\gamma$ for some $C, \gamma > 0$ since $\sigma^2(t)/t^2 = 2\int_0^1 \int_0^s R(tv)dvds$ tends to $R(0)$ as $t \downarrow 0$.

Short-range dependent case

A number of important Gaussian integrated processes have short-range dependent characteristics. Perhaps the most well-known example is an Ornstein-Uhlenbeck process, for which $R(t) = \exp(-\alpha t)$, where $\alpha > 0$ is a constant. Dębicki and Rolski [96] study the more general case where $Z = r'X$ for some k-vector r and X is the stationary solution of the stochastic differential equation

$$dX_t = AX_t dt + \sigma dW_t,$$

for $k \times k$ matrices A, σ (satisfying certain conditions) and a standard k-dimensional Brownian motion W. Then $R(t) = r'\Sigma e^{tA'} r$ for some covariance Σ.

By stating that a Gaussian integrated process is short-range dependent, we mean that $\mathcal{R} := \lim_{t\to\infty} \int_0^t R(s)ds$ exists as a strictly positive real number and that R is integrable, i.e., $\int_0^\infty |R(s)|ds < \infty$. We can now specialize Proposition 5.4 to this case.

Corollary 5.7 *Let Y be a Gaussian integrated process with short-range dependence. Then*

$$\mathbb{P}\left(\sup_{t\geq 0} Y_t - ct > u\right) \sim \mathcal{H}_{\frac{c}{\sqrt{2\mathcal{R}}}Y} \sqrt{\pi} \frac{2\sqrt{\mathcal{R}}}{c^{3/2}} \sqrt{u}\Psi\left(\inf_{t\geq 0} \frac{u(1+ct)}{\sqrt{2\int_0^{ut}\int_0^s R(v)dvds}}\right). \tag{5.12}$$

Proof. By the existence of $\mathcal{R}$, continuity of $t \mapsto \int_0^t R(s)ds$, and bounded convergence, we have

$$\lim_{t\to\infty} \frac{\sigma^2(t/c)}{t} = \frac{2}{c} \lim_{t\to\infty} \int_0^1 \int_0^{st} R(v)dvds = \frac{2\mathcal{R}}{c} < \infty,$$

so that **S1** holds with $H = 1/2$ and we are in the second case of Proposition 5.4 with $\mathcal{G} = 2\mathcal{R}/c$. □

Notice that Corollary 5.7 is a modest generalization of the results of Dębicki [87]. To see this, note that (5.12) is asymptotically equivalent with

$$\mathcal{H}_{\frac{c}{\sqrt{2\mathcal{R}}}Y}\frac{\mathcal{R}}{c^2}\exp\left(-\frac{1}{4}\inf_{t\geq 0}\frac{u^2(1+ct)^2}{\int_0^{ut}\int_0^s R(v)dvds}\right),$$

since $t^* = H/(\beta - H) = 1$ and $\sqrt{u}\sigma(u) \sim \sqrt{2\mathcal{R}u}$. Proposition 6.1 of [87] shows that this expression is in agreement with the findings of [87]. In the present context, we also mention the recent contribution of Kobelkov [193].

Long-range dependent case

Consider a Gaussian integrated process as in (5.11), but now with a covariance function R that is regularly varying at infinity with index $2H-2$ for some $H \in (1/2, 1)$ (in addition to the regularity assumptions above). Since there is so much long-term correlation that $\int_0^\infty |R(t)|dt = \infty$, the process is long-range dependent. The motivation for studying this long-range dependent case stems from the fact that it arises as a limit in heavy traffic of ON/OFF fluid models with heavy-tailed ON-periods or OFF-periods, see Section 1.3.2.

By the direct half of Karamata's theorem (Lemma 2.5), we have for $t \to \infty$,

$$\sigma^2(t) = 2\int_0^t\int_0^s R(v)dvds \sim \frac{t\int_0^t R(v)dv}{H} \sim \frac{t^2R(t)}{H(2H-1)}.$$

Therefore, since $H > 1/2$, we have $\sigma^2(t)/t \to \infty$ and we are in the first case of Proposition 5.4.

Corollary 5.8 *Let Y be a Gaussian integrated process with long-range dependence. Then*

$$\mathbb{P}\left(\sup_{t\geq 0} Y_t - ct > u\right) \sim \mathcal{H}_{B_H}\widehat{C}_H\frac{u\sqrt{R(u)}}{\overleftarrow{\tau}(uR(u))}\Psi\left(\inf_{t\geq 0}\frac{u(1+ct)}{\sqrt{2\int_0^{ut}\int_0^s R(v)dvds}}\right),$$

where $\overleftarrow{\tau}$ denotes an asymptotic inverse of $t \mapsto t\sqrt{R(t)}$ (at infinity) and the constant $\widehat{C}_H$ equals $C_{H,1,1,H}c^{1-H}\frac{1-H}{H}[H(2H-1)]^{\frac{1}{2H}-\frac{1}{2}}$.

The case of a Gaussian integrated process with long-range dependence is also studied by Hüsler and Piterbarg [164]. The reasoning following Equation (7) of [164] shows that the formulas are the same (up to the constants; we leave it to the reader to check that these coincide).

5.4 A variant of Pickands' lemma

In this section, we present a generalization of Pickands' lemma (Lemma 3.6). As we need a field version of this lemma, we let time be indexed by $\mathbb{R}^n$ for some $n \geq 1$, and we write $\mathbf{t} = (t_1, \ldots, t_n)$.

Given an even functional $\xi_\eta : \mathbb{R}^n \to \mathbb{R}$ (i.e., $\xi_\eta(\mathbf{t}) = \xi_\eta(-\mathbf{t})$ for $\mathbf{t} \in \mathbb{R}^n$), we define the centered Gaussian field η by its covariance

$$\mathbb{C}\mathrm{ov}(\eta_{\mathbf{s}}, \eta_{\mathbf{t}}) = \xi_\eta(\mathbf{s}) + \xi_\eta(\mathbf{t}) - \xi_\eta(\mathbf{s}-\mathbf{t}), \tag{5.13}$$

provided it is a proper covariance in the sense that the field η exists.

A central role in the lemma is played by functions g_k, ξ_η, and θ_k. These functions are in principle arbitrary, but they are assumed to satisfy certain conditions, which we now formulate. To get some feeling for these conditions, the reader may want to look in the proof of Lemma 5.10, for instance, to see how the functions are chosen in a particular situation.

Throughout, $\{K_u\}$ denotes a nondecreasing family of countable sets (say $K_u \subset \mathbb{Z}$), and $\{X_{\mathbf{t}}^{(u,k)} : \mathbf{t} \in [0,T]^n\}$, $u \in \mathbb{N}$, $k \in K_u$ denotes a collection of centered continuous Gaussian fields on $[0,T]^n$ for some fixed $T > 0$. We suppose that $X^{(u,k)}$ has unit variance. It is important to note that we do not assume stationarity of the $X^{(u,k)}$. The following conditions are imposed.

P1 $\inf_{k\in K_u} g_k(u) \to \infty$ as $u \to \infty$,

P2 for some even functional ξ_η, $\sup_{k\in K_u} |\theta_k(u,\mathbf{s},\mathbf{t}) - 2\xi_\eta(\mathbf{s}-\mathbf{t})| \to 0$ for any $\mathbf{s},\mathbf{t} \in [0,T]^n$,

P3 for some $\gamma_1,\dots,\gamma_n > 0$,

$$\limsup_{u\to\infty} \sup_{k\in K_u} \sup_{\mathbf{s},\mathbf{t}\in[0,T]^n} \frac{\theta_k(u,\mathbf{s},\mathbf{t})}{\sum_{i=1}^n |s_i - t_i|^{\gamma_i}} < \infty,$$

P4 $\mathbf{t} \mapsto g_k^2(u)\mathbb{C}\mathrm{ov}\left(X_{\mathbf{t}}^{(u,k)}, X_{\mathbf{0}}^{(u,k)}\right)$ is uniformly continuous in the sense that

$$\lim_{\varepsilon\to 0} \limsup_{u\to\infty} \sup_{k\in K_u} \sup_{\substack{|\mathbf{s}-\mathbf{t}|<\varepsilon \\ \mathbf{s},\mathbf{t}\in[0,T]^n}} g_k^2(u)\mathbb{C}\mathrm{ov}\left(X_{\mathbf{s}}^{(u,k)} - X_{\mathbf{t}}^{(u,k)}, X_{\mathbf{0}}^{(u,k)}\right) = 0.$$

We use the following lemma in Section 5.6 for $n = 1$ to establish the upper bound, and in Section 5.7 for $n = 2$ to establish the lower bound. The main assumption of the lemma is that $\mathbb{C}\mathrm{ov}\left(X_{\mathbf{s}}^{(u,k)}, X_{\mathbf{t}}^{(u,k)}\right)$ tends uniformly to 1 at rate $2\theta_k(u,\mathbf{s},\mathbf{t})/g_k^2(u)$ as $u \to \infty$.

Lemma 5.9 *Suppose there exist functions g_k, ξ_η, and θ_k satisfying* **P1**–**P4**. *If*

$$\lim_{u\to\infty} \sup_{k\in K_u} \sup_{\substack{\mathbf{s},\mathbf{t}\in[0,T]^n \\ \mathbf{s}\neq\mathbf{t}}} \left| g_k^2(u) \frac{\mathbb{V}\mathrm{ar}\left(X_{\mathbf{s}}^{(u,k)} - X_{\mathbf{t}}^{(u,k)}\right)}{\theta_k(u,\mathbf{s},\mathbf{t})} - 1 \right| = 0, \tag{5.14}$$

then for any $k \in \bigcup_u K_u$, as $u \to \infty$,

$$\mathbb{P}\left(\sup_{\mathbf{t}\in[0,T]^n} X_{\mathbf{t}}^{(u,k)} > g_k(u)\right) \sim \mathcal{H}_\eta([0,T]^n)\Psi(g_k(u)), \tag{5.15}$$

where

$$\mathcal{H}_\eta([0,T]^n) = \mathbb{E}\exp\left(\sup_{\mathbf{t}\in[0,T]^n} \sqrt{2}\eta_{\mathbf{t}} - \xi_\eta(\mathbf{t})\right).$$

Moreover, we have

$$\limsup_{u\to\infty} \sup_{k\in K_u} \frac{\mathbb{P}\left(\sup_{\mathbf{t}\in[0,T]^n} X_{\mathbf{t}}^{(u,k)} > g_k(u)\right)}{\Psi(g_k(u))} < \infty. \tag{5.16}$$

Proof. The proof is based on a standard approach in the theory of Gaussian processes; see for instance (the proof of) Lemma D.1 of Piterbarg [257].

First note that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{t}\in[0,T]^n} X_{\mathbf{t}}^{(u,k)} > g_k(u)\right) \\
&= \frac{1}{\sqrt{2\pi}g_k(u)} \exp\left(-\frac{1}{2}g_k^2(u)\right) \int_{\mathbb{R}} \exp(w) \exp\left(-\frac{1}{2}\frac{w^2}{g_k^2(u)}\right) \\
&\quad \times \mathbb{P}\left(\sup_{\mathbf{t}\in[0,T]^n} X_{\mathbf{t}}^{(u,k)} > g_k(u) \,\middle|\, X_{\mathbf{0}}^{(u,k)} = g_k(u) - \frac{w}{g_k(u)}\right) dw.
\end{aligned}$$

For fixed w, we set $\chi_{u,k}(\mathbf{t}) := g_k(u)[X_{\mathbf{t}}^{(u,k)} - g_k(u)] + w$, so that the conditional probability that appears in the integrand equals $\mathbb{P}(\sup_{\mathbf{t}\in[0,T]^n} \chi_{u,k}(\mathbf{t}) > w | \chi_{u,k}(\mathbf{0}) = 0)$.

We first study the field $\chi_{u,k}|\chi_{u,k}(\mathbf{0}) = 0$ as $u \to \infty$, starting with the finite-dimensional (cylinder) distributions. These converge uniformly in $k \in K_u$ to the corresponding distributions of $\sqrt{2}\eta - \xi_\eta$. To see this, we set $v_{u,k}(\mathbf{s},\mathbf{t}) := \mathbb{V}\mathrm{ar}(X_{\mathbf{s}}^{(u,k)} - X_{\mathbf{t}}^{(u,k)})$, so that by **P1**, **P2**, and (5.14), uniformly in $k \in K_u$,

$$\begin{aligned}\mathbb{E}[\chi_{u,k}(\mathbf{t})|\chi_{u,k}(\mathbf{0}) = 0] &= -\frac{1}{2}g_k^2(u)v_{u,k}(\mathbf{0},\mathbf{t}) + \frac{1}{2}wv_{u,k}(\mathbf{0},\mathbf{t}) \\ &= -\frac{1}{2}\theta_k(u,\mathbf{0},\mathbf{t})(1+o(1)) + o(1) \to -\xi_\eta(\mathbf{t}),\end{aligned}$$

and similarly, also uniformly in $k \in K_u$,

$$\begin{aligned}&\mathbb{V}\mathrm{ar}(\chi_{u,k}(\mathbf{s}) - \chi_{u,k}(\mathbf{t})|\chi_{u,k}(\mathbf{0}) = 0) \\ &= g_k^2(u)v_{u,k}(\mathbf{s},\mathbf{t}) - \frac{1}{4}g_k^2(u)\left[v_{u,k}(\mathbf{0},\mathbf{t}) - v_{u,k}(\mathbf{0},\mathbf{s})\right]^2 \\ &= \theta_k(u,\mathbf{s},\mathbf{t})(1+o(1)) + o(1) \to 2\xi_\eta(\mathbf{s}-\mathbf{t}).\end{aligned}$$

Denoting the law of a field X by $\mathcal{L}(X)$, we next show that the family $\{\mathcal{L}(\chi_{u,k}|\chi_{u,k}(\mathbf{0}) = 0) : u \in \mathbb{N}, k \in K_u\}$ is tight. Since $\mathbf{t} \mapsto \mathbb{E}(\chi_{u,k}(\mathbf{t})|\chi_{u,k}(\mathbf{0}) = 0)$ is uniformly continuous in the sense that **P4** holds, it suffices to show that the family of centered distributions is tight. We denote the centered $\chi_{u,k}$ by $\tilde{\chi}_{u,k}$, i.e., $\tilde{\chi}_{u,k}(\mathbf{t}) := \chi_{u,k}(\mathbf{t}) - \mathbb{E}[\chi_{u,k}(\mathbf{t})|\chi_{u,k}(\mathbf{0}) = 0]$. It is important to notice that $\mathcal{L}(\tilde{\chi}_{u,k}|\tilde{\chi}_{u,k}(\mathbf{0}) = 0)$ does not depend on w.

To see that $\{\mathcal{L}(\tilde{\chi}_{u,k}|\tilde{\chi}_{u,k}(\mathbf{0}) = 0) : u \in \mathbb{N}, k \in K_u\}$ is tight, observe that for u large enough, uniformly in $\mathbf{s},\mathbf{t} \in [0,T]^n$ and $k \in K_u$, we have

$$\mathbb{V}\mathrm{ar}(\tilde{\chi}_{u,k}(\mathbf{s}) - \tilde{\chi}_{u,k}(\mathbf{t})|\tilde{\chi}_{u,k}(\mathbf{0}) = 0) \le g_k^2(u)v_{u,k}(\mathbf{s},\mathbf{t}) \le 2\theta_k(u,\mathbf{s},\mathbf{t}).$$

By **P3**, there exist constants $\gamma_1,\dots,\gamma_n, C' > 0$ such that, uniformly in $\mathbf{s},\mathbf{t} \in [0,T]^n$ and $k \in K_u$,

$$\mathbb{V}\mathrm{ar}(\tilde{\chi}_{u,k}(\mathbf{s}) - \tilde{\chi}_{u,k}(\mathbf{t})|\tilde{\chi}_{u,k}(\mathbf{0}) = 0) \le C'\sum_{i=1}^n |s_i - t_i|^{\gamma_i},$$

provided u is large enough. As a corollary of Theorem 1.4.7 in Kunita [204], we have the claimed tightness.

Since the functional $x \in C([0,T]^n) \mapsto \sup_{\mathbf{t}\in[0,T]^n} x(\mathbf{t})$ is continuous in the topology of uniform convergence, the continuous mapping theorem yields for $w \in \mathbb{R}$,

$$\lim_{u\to\infty} \mathbb{P}\left(\sup_{\mathbf{t}\in[0,T]^n} \chi_{u,k}(\mathbf{t}) > w \,\middle|\, \chi_{u,k}(\mathbf{0}) = 0\right) = \mathbb{P}\left(\sup_{\mathbf{t}\in[0,T]^n} \sqrt{2}\eta_{\mathbf{t}} - \xi_\eta(\mathbf{t}) > w\right).$$

Using $\int_{\mathbb{R}} e^w \mathbb{P}(\sup_{\mathbf{t}\in[0,T]^n} \sqrt{2}\eta_{\mathbf{t}} - \xi_\eta(\mathbf{t}) > w)dw = \mathcal{H}_\eta([0,T]^n)$ and (3.2), this proves (5.15) once it has been shown that the integral and limit can be interchanged.

The dominated convergence theorem and Borell's inequality (Lemma 3.3) are used to see that this can indeed be done. For arbitrary $\delta > 0$ and u large enough,

$$\begin{aligned}\sup_{k\in K_u}\sup_{\mathbf{t}\in[0,T]^n} \mathbb{E}[\chi_{u,k}(\mathbf{t})|\chi_{u,k}(\mathbf{0}) = 0] &\le \delta|w|, \\ \sup_{k\in K_u}\sup_{\mathbf{t}\in[0,T]^n} \mathbb{V}\mathrm{ar}[\chi_{u,k}(\mathbf{t})|\chi_{u,k}(\mathbf{0}) = 0] &\le 2\sup_{k\in K_u}\sup_{\mathbf{t}\in[0,T]^n} \theta_k(u,\mathbf{t},\mathbf{0}),\end{aligned}$$

and the latter quantity remains bounded as $u \to \infty$ as a consequence of **P3**; let $\overline{\xi_\eta}$ denote an upper bound. Observe that for $a \in \mathbb{R}$, again by the continuous mapping theorem, we have

$$\lim_{u\to\infty} \sup_{k\in K_u} \mathbb{P}\left(\sup_{\mathbf{t}\in[0,T]^n} \tilde{\chi}_{u,k}(\mathbf{t}) > a \,\middle|\, \chi_{u,k}(\mathbf{0}) = 0 \right) = \mathbb{P}\left(\sup_{\mathbf{t}\in[0,T]^n} \sqrt{2}\eta_{\mathbf{t}} > a \right).$$

Since η is continuous (as remarked below), one can select an a independent of w, u, k such that the conditions for applying Borell's inequality (Lemma 3.3) are fulfilled. Hence, for every u, k, w,

$$\mathbb{P}\left(\sup_{\mathbf{t}\in[0,T]^n} \chi_{u,k}(\mathbf{t}) > w \,\middle|\, \chi_{u,k}(\mathbf{0}) = 0 \right) \le 2\Psi\left(\frac{w - \delta|w| - a}{3\overline{\xi_\eta}} \right).$$

When multiplied by $\exp(w)\exp(-\frac{1}{2}w^2/g_k^2(u))$, this upper bound is integrable with respect to w for large u. This not only shows that the dominated convergence theorem can be applied, it also implies (5.16). Indeed, using **P1**, we have

$$\lim_{u\to\infty} \sup_{k\in K_u} \frac{e^{-\frac{1}{2}g_k^2(u)}}{g_k(u)\Psi(g_k(u))} = \sqrt{2\pi}$$

by standard bounds on Ψ. □

One observation in the proof deserves to be emphasized, namely the existence and continuity of η. If θ_k satisfies (5.14) and converges uniformly in k to some $2\xi_\eta$ as in **P2**, the analysis of the finite-dimensional distributions shows that there automatically exists a field η with covariance (5.13). Moreover, η has continuous sample paths as a consequence of **P3** and **P4** (i.e., the tightness).

A number of special cases of Lemma 5.9 appear elsewhere in the literature. The best known example is Lemma 3.6; it is obtained by letting K_u consist of only a single element for every u, and by setting $g(u) = u$, $X_t^{(u)} = X_{u^{-2/\alpha}t}$, $\eta = B_{\alpha/2}$ and $\xi_\eta(t) = |t|^\alpha$.

A generalization of Lemma D.1 in [257] to a stationary field $\{X(\mathbf{t}) : \mathbf{t} \in \mathbb{R}^n\}$ is given in Lemma 6.1 of Piterbarg [257], and we now compare this generalization to Lemma 5.9. We use the notation of [257]. Lemma 5.9 deals with the case $A = 0$ and T (in the notation of [257]) equal to $[0,T]^n$ (in our notation). Again, let K_u consist of only a single element for every u, and set $g(u) = u$, $X_{\mathbf{t}}^{(u)} = X_{g_u^{-1}\mathbf{t}}$, and $\xi_\eta(\mathbf{t}) = |\mathbf{t}|_{E,\alpha}$. As the ideas of the proof are the same, Lemma 5.9 can readily be extended to also generalize Lemma 6.1 of [257]. However, we do not need this to derive the results of the present chapter.

Theorem 2.1 of Dębicki [87] can also be considered to be a special case of Lemma 5.9. There, again, K_u consists of a single element, and $X_{(t_1,\dots,t_n)}^{(u)} = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i^{(u)}(t_i)$ for independent processes $X_i^{(u)}$ satisfying a condition of the type (5.14), but where θ does not depend on u.

Lemma 5.9 has some interesting consequences for the properties of Pickands' constant. For instance, Pickands' constant is readily seen to be subadditive, i.e., for $T_1, T_2 > 0$ and $n = 1$,

$$\mathcal{H}_\eta([0, T_1 + T_2]) \le \mathcal{H}_\eta([0, T_1]) + \mathcal{H}_\eta([0, T_2]),$$

with appropriate generalizations to the multidimensional case. This property guarantees that the limit in (3.8) exists. For further properties of Pickands' constant and references, we refer to Section 3.3.

5.5 Four cases

We now specialize Lemma 5.9 to the four types of correlation structures introduced in Section 5.2. Throughout this section, we suppose that **S1** and **M1** hold.

Let $T > 0$ be fixed, and write $I_k^T(u)$ for the intervals $[t_u^* + kT\Delta(u)/u, t_u^* + (k+1)T\Delta(u)/u]$, where Δ is some function that depends on the correlation structure, and $\Delta(u) = o(\delta(u))$.

5.5.1 Case A

We say that case A applies if **A1**, **A2**, **T1**, **T2**, **N1**, and **N2** hold and Δ is given by

$$\Delta(u) := \frac{1}{\dot\nu(ut^*)} \overleftarrow{\tau}\left(\frac{\sqrt{2}\tau(\nu(u))}{\sqrt{\mathcal{D}}}\frac{\sigma(\mu(ut^*))}{u(1+t^*)}\right), \tag{5.17}$$

where $\overleftarrow{\tau}$ denotes an asymptotic inverse of τ at infinity (this exists when **T1** holds, see Lemma 2.6). Note that the argument of $\overleftarrow{\tau}$ tends to infinity as a consequence of **A2**, and that therefore $\nu(u)\Delta(u)/u \to \infty$. It is easy to check that Δ is regularly varying with index $(H/\beta - 1)/\iota_\tau + 1 < 1$.

The next lemma shows that this particular choice of Δ 'balances' the correlation structure on the intervals $I_k^T(u)$ (note that the interval $I_k^T(u)$ depends on Δ).

Lemma 5.10 *Let* **S1** *and* **M1** *hold and suppose that case A applies. Let δ be such that $\delta(u) = o(u)$ and $\Delta(u) = o(\delta(u))$. For any u and $-\frac{\delta(u)}{T\Delta(u)} \le k \le \frac{\delta(u)}{T\Delta(u)}$, pick some $t_k^\circ(u) \in I_k^T(u)$. Then we have for $u \to \infty$,*

$$\mathbb{P}\left(\sup_{t\in I_k^T(u)} \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))} > \frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right) \sim \mathcal{H}_{B_{\iota_\tau}}(T)\Psi\left(\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right),$$

where $\mathcal{H}_{B_{\iota_\tau}}(T)$ is defined as in (3.8). Moreover,

$$\limsup_{u\to\infty} \sup_{-\frac{\delta(u)}{T\Delta(u)}\le k\le \frac{\delta(u)}{T\Delta(u)}} \frac{\mathbb{P}\left(\sup_{t\in I_k^T(u)} \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))} > \frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right)}{\Psi\left(\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right)} < \infty. \tag{5.18}$$

Proof. The main argument in the proof is, of course, Lemma 5.9. Set

$$\kappa_k(u) := \sqrt{\frac{\mathcal{D}}{2}}\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\frac{\tau(\dot\nu(ut^*)\Delta(u))}{\tau(\nu(u))}$$

and note that by the UCT and (5.17),

$$\sup_{-\frac{\delta(u)}{T\Delta(u)}\le k\le \frac{\delta(u)}{T\Delta(u)}} \sup_{s,t\in I_k^T(u)} \left|\kappa_k^2(u) - 1\right| \to 0.$$

Equation (5.3) implies that {**A1**}

$$\sup_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ |s-t|\le T\Delta(u)/u}} \left| \frac{2\kappa_k^2(u)\tau^2(\nu(u))}{\mathcal{D}\tau^2(\dot\nu(ut^*)\Delta(u))} \frac{\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))} - \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right)}{2\tau^2(|\nu(us)-\nu(ut)|)/\tau^2(\dot\nu(ut^*)\Delta(u))} - 1\right| \to 0. \tag{5.19}$$

The preceding display suggests certain choices for the functions g_k and θ_k of Lemma 5.9, cf. (5.14); we now show that **P1**–**P4** are indeed satisfied.

As for **P1**, one readily checks that

$$g_k(u) := \sqrt{\frac{2}{\mathcal{D}}}\frac{\kappa_k(u)\tau(\nu(u))}{\tau(\dot\nu(ut^*)\Delta(u))} = \frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}$$

tends to infinity uniformly in k. We set for $s,t\in[0,T]$ and $-\frac{\delta(u)}{T\Delta(u)}\le k\le\frac{\delta(u)}{T\Delta(u)}$

$$\theta_k(u,s,t):=2\frac{\tau^2(|\nu(ut_u^*+(kT+s)\Delta(u))-\nu(ut_u^*+(kT+t)\Delta(u))|)}{\tau^2(\dot\nu(ut^*)\Delta(u))}.$$

To check that $\theta_k(u,s,t)$ converges uniformly in k as $u\to\infty$, we note that by the mean value theorem {**N2**} there exists some $t_k^\wedge(u,s,t)\in[0,T]$ such that

$$\nu(ut_u^*+(kT+s)\Delta(u))-\nu(ut_u^*+(kT+t)\Delta(u))=\Delta(u)\dot\nu(ut_u^*+[kT+t_k^\wedge(u,s,t)])(s-t).$$

Now observe that we have for $s,t\in[0,T]$,

$$\begin{aligned}
&\sup_k|\theta_k(u,s,t)-2|s-t|^{2\iota_\tau}|\\
&\quad\le\quad\sup_k\left|\theta_k(u,s,t)-2\left(\frac{\dot\nu(ut_u^*+[kT+t_k^\wedge(u,s,t)])}{\dot\nu(ut^*)}\right)^{2\iota_\tau}|s-t|^{2\iota_\tau}\right|\\
&\qquad+2\sup_k\left|\left(\frac{\dot\nu(ut_u^*+[kT+t_k^\wedge(u,s,t)])}{\dot\nu(ut^*)}\right)^{2\iota_\tau}-1\right||s-t|^{2\iota_\tau}\\
&\quad=:\quad I(u)+II(u).
\end{aligned}$$

As a consequence of the UCT {**N1**, **N2**}, we have

$$\lim_{u\to\infty}\sup_{s,t\in[0,T]}\sup_{-\frac{\delta(u)}{T\Delta(u)}\le k\le\frac{\delta(u)}{T\Delta(u)}}\frac{\dot\nu(ut_u^*+[kT+t_k^\wedge(u,s,t)])}{\dot\nu(ut^*)}(s-t)=T. \tag{5.20}$$

Since $\dot\nu(u)\Delta(u)$ tends to infinity {**A2**}, this shows that $I(u)$ is majorized by

$$\sup_{t\in[0,2T]}\left|\frac{\tau^2(\dot\nu(ut^*)\Delta(u)t)}{\tau^2(\dot\nu(ut^*)\Delta(u))}-t^{2\iota_\tau}\right|\to0.$$

The term $II(u)$ also tends to zero by the UCT. Hence, **P2** holds with $\xi_\eta(t)=|t|^{2\iota_\tau}$, so that η is a fractional Brownian motion with Hurst parameter ι_τ.

A similar reasoning is used to check **P3**. Note that $\tau^2(t)t^{-2\gamma'}$ is bounded on intervals of the form $(0,\cdot]$ {**T2**}, and that we may suppose that $\gamma'<\iota_\tau$ without loss of generality. Again using (5.20) and the UCT, we observe that for large u,

$$\begin{aligned}
&\sup_k\sup_{\substack{s,t\in(0,T]\\ s>t}}\theta_k(u,s,t)(s-t)^{-2\gamma'}\\
&\quad=\quad\sup_k\sup_{\substack{s,t\in(0,T]\\ s>t}}2\frac{\tau^2\left(\Delta(u)\dot\nu(ut_u^*+[kT+t_k^\wedge(u,s,t)])(s-t)\right)}{\tau^2(\dot\nu(ut^*)\Delta(u))}(s-t)^{-2\gamma'}\\
&\quad\le\quad2\sup_{t\in\left[0,\left(\frac{3}{2}\right)^{1/(2\iota_\tau-2\gamma')}T\right]}\frac{\tau^2(\dot\nu(ut^*)\Delta(u)t)}{\tau^2(\dot\nu(ut^*)\Delta(u))}t^{-2\gamma'}\\
&\quad\le\quad4T^{2\iota_\tau-2\gamma'},
\end{aligned}$$

which is clearly finite (the factor 4 appears again in the proof of Lemma 5.18 below).

It remains to check **P4**. For this, it suffices to show that

$$\limsup_{u\to\infty}\sup_{s,s',t\in[t_u^*\pm\delta(u)/u]}\sup_{\substack{|s-t|<\varepsilon T\Delta(u)/u\\ |s'-t|<\varepsilon T\Delta(u)/u}}g_k^2(u)\mathbb{C}\mathrm{ov}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))}-\frac{Y_{\mu(us')}}{\sigma(\mu(us'))},\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right)$$

vanishes as $\varepsilon \to 0$, and hence that

$$\lim_{\varepsilon\to 0}\limsup_{u\to\infty}\sup_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ |s-t|<\varepsilon\Delta(u)/u}} g_k^2(u)\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))}-\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right)=0. \tag{5.21}$$

For large u, by (5.19) and the mean value theorem, uniformly in $s,t\in[t_u^*\pm\delta(u)/u]$, we have

$$\begin{aligned}\sup_{|s-t|<\varepsilon\Delta(u)/u} g_k^2(u)\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))}-\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right) &\le 4\sup_{|s-t|<\varepsilon\Delta(u)/u}\frac{\tau^2(|\nu(us)-\nu(ut)|)}{\tau^2(\dot\nu(ut^*)\Delta(u))}\\ &\le 4\sup_{t<2\varepsilon\Delta(u)/u}\frac{\tau^2(u\dot\nu(ut^*)t)}{\tau^2(\dot\nu(ut^*)\Delta(u))}\\ &\le 8(2\varepsilon)^{2\iota_\tau}\to 0,\end{aligned}$$

as $\varepsilon\to 0$.

Having checked that Lemma 5.9 can be applied, we use the definition of $\Delta(u)$ to see that

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}>\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right) &= \mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}>\sqrt{\frac{2}{\mathcal{D}}}\frac{\kappa_k(u)\tau(\nu(u))}{\tau(\dot\nu(ut^*)\Delta(u))}\right)\\ &\sim \mathcal{H}_{B_{\iota_\tau}}(T)\Psi\left(\sqrt{\frac{2}{\mathcal{D}}}\frac{\kappa_k(u)\tau(\nu(u))}{\tau(\dot\nu(ut^*)\Delta(u))}\right)\\ &= \mathcal{H}_{B_{\iota_\tau}}(T)\Psi\left(\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right),\end{aligned}$$

as claimed. □

5.5.2 Case B

Case B differs from the other cases in the sense that no (asymptotic) inverse is involved in the definition of Δ. As a consequence, a nonclassical Pickands' constant appears in the asymptotics.

We say that case B applies when **B1**, **B2**, **T1**, **T2**, **N1**, and **N2** hold and Δ is given by

$$\Delta(u):=\frac{1}{\dot\nu(ut^*)}. \tag{5.22}$$

Moreover, we set

$$\mathcal{F}:=\frac{\mathcal{D}(1+t^*)^2}{2\mathcal{G}^2(t^*)^{2H/\beta}}.$$

Under these assumptions, $\lim_{u\to\infty}\nu(u)\Delta(u)/u$ exists in $(0,\infty)$.

Lemma 5.11 *Let* **S1** *and* **M1** *hold and suppose that case B applies. Let δ be such that $\delta(u)=o(u)$ and $\Delta(u)=o(\delta(u))$. For any u and $-\frac{\delta(u)}{T\Delta(u)}\le k\le\frac{\delta(u)}{T\Delta(u)}$, pick some $t_k^\circ(u)\in I_k^T(u)$. For T large enough, we have for $u\to\infty$,*

$$\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}>\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right)\sim\mathcal{H}_{\mathcal{F}\tau^2}(T)\Psi\left(\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right),$$

where $\mathcal{H}_{\mathcal{F}\tau^2}(T)$ is defined as in (3.8). Moreover, (5.18) holds.

Proof. Define

$$\kappa_k(u) := \sqrt{\frac{\mathcal{D}}{2\mathcal{F}}}\frac{u(1+t_k^\circ(u))}{\tau(\nu(u))\sigma(\mu(ut_k^\circ(u)))},$$

which converges uniformly in k to 1 as a consequence of the fact that by **B2**,

$$\frac{2\mathcal{F}\tau^2(\nu(u))}{\mathcal{D}} = \frac{(1+t^*)^2}{\mathcal{G}^2(t^*)^{2H/\beta}}\tau^2(\nu(u)) \sim \frac{u^2(1+t^*)^2}{\sigma^2(\mu(ut^*))}.$$

Therefore, as in Lemma 5.10, we have by (5.3),

$$\sup_{\substack{-\frac{\delta(u)}{T\Delta(u)}\le k\le \frac{\delta(u)}{T\Delta(u)}\\ k\in\mathbb{Z}}}\sup_{s,t\in I_k^T(u)}\left|\frac{2\mathcal{F}\kappa_k^2(u)\tau^2(\nu(u))}{\mathcal{D}}\frac{\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))}-\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right)}{2\mathcal{F}\tau^2(|\nu(us)-\nu(ut)|)}-1\right|\to 0.$$

Again, this should be compared to (5.14). Set $g_k(u) := \sqrt{2\mathcal{F}/\mathcal{D}}\kappa_k(u)\tau(\nu(u))$, and

$$\theta_k(u,s,t) := 2\mathcal{F}\tau^2(|\nu(ut_u^* + (kT+s)\Delta(u)) - \nu(ut_u^* + (kT+t)\Delta(u))|).$$

Obviously, **P1** holds. We now check that **P2** holds with $\xi_\eta(t) = \mathcal{F}\tau^2(|t|)$. Let $s,t\in[0,T]$, and observe that by the mean value theorem there exist $t_k^\wedge(u,s,t)\in[0,T]$ such that for every $\epsilon>0$,

$$\begin{aligned}
&\sup_k|\theta_k(u,s,t)-2\mathcal{F}\tau^2(s-t)|\\
&\quad= 2\sup_k|\mathcal{F}\tau^2(\Delta(u)\dot\nu(ut_u^*+[kT+t_k^\wedge(u,s,t)])|s-t|)-\mathcal{F}\tau^2(|s-t|)|\\
&\quad\le 2\mathcal{F}\sup_{s\in[1-\epsilon,1+\epsilon]}\sup_{t\in[0,T]}|\tau^2(st)-\tau^2(t)|\\
&\quad\le 2\mathcal{F}\sup_{\substack{s,t\in[0,2T]\\|s-t|\le\epsilon T}}|\tau^2(s)-\tau^2(t)|,
\end{aligned}$$

where we used the definition of Δ and the UCT. By continuity of τ {**T1**}, this upper bound (which is a modulus of continuity) tends to zero as $\epsilon\to 0$. As for **P3**, the same arguments show that for large T (by the UCT) {**T1**, **T2**}

$$\sup_k\sup_{s,t\in[0,T]}\frac{\theta_k(u,s,t)}{|s-t|^{2\gamma'}}\le 2\mathcal{F}\sup_{t\in\left[0,\left(\frac{3}{2}\right)^{1/(2\iota_\tau-2\gamma')}T\right]}\frac{\tau^2(t)}{t^{2\gamma'}}\le 4\mathcal{F}T^{2(\iota_\tau-\gamma')}.$$

It remains to verify **P4**. As in the proof of Lemma 5.10, it suffices to show that (5.21) holds. By again applying the UCT, one can check that for $s,t\in[t_u^*\pm\delta(u)/u]$,

$$\sup_k\sup_{|s-t|<\varepsilon\Delta(u)/u}g_k^2(u)\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))}-\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right)\le 2\mathcal{F}\sup_{t\in[0,2\varepsilon]}\tau^2(t),$$

showing **P4** since τ^2 is continuous at zero.

In conclusion, Lemma 5.9 can be applied and therefore

$$\begin{aligned}
\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}>\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right) &= \mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}>\sqrt{\frac{2\mathcal{F}}{\mathcal{D}}}\kappa_k(u)\tau(\nu(u))\right)\\
&\sim \mathcal{H}_{\mathcal{F}\tau^2}(T)\Psi\left(\sqrt{\frac{2\mathcal{F}}{\mathcal{D}}}\kappa_k(u)\tau(\nu(u))\right)\\
&= \mathcal{H}_{\mathcal{F}\tau^2}(T)\Psi\left(\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right),
\end{aligned}$$

as claimed. □

5.5.3 Case C

We say that case C applies when **C1**–**C3**, **T1**, **N1**, and **N2** hold and Δ is given by

$$\Delta(u) := \frac{1}{\dot{\nu}(ut^*)} \overleftarrow{\tau} \left(\frac{\sqrt{2}\tau(\nu(u))}{\sqrt{\mathcal{D}}} \frac{\sigma(\mu(ut^*))}{u(1+t^*)} \right), \tag{5.23}$$

where $\overleftarrow{\tau}$ denotes an asymptotic inverse of τ at zero (which exists due to **T1**, see Lemma 2.6). Here, the argument of $\overleftarrow{\tau}$ tends to zero as a consequence of **C2**, and therefore $\nu(u)\Delta(u)/u \to 0$. Note that we do not impose **T2**, since it is automatically satisfied once **C3** holds.

The following lemma is the analogue of Lemma 5.10 and Lemma 5.11 for case C.

Lemma 5.12 *Let* **S1** *and* **M1** *hold and suppose that case C applies. Let δ be such that $\delta(u) = o(u)$ and $\Delta(u) = o(\delta(u))$. For any u and $-\frac{\delta(u)}{T\Delta(u)} \le k \le \frac{\delta(u)}{T\Delta(u)}$, pick some $t_k^\circ(u) \in I_k^T(u)$. Then we have for $u \to \infty$,*

$$\mathbb{P}\left(\sup_{t \in I_k^T(u)} \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))} > \frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))} \right) \sim \mathcal{H}_{B_{\tilde{\iota}_\tau}}(T)\Psi\left(\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))} \right),$$

where $\mathcal{H}_{B_{\tilde{\iota}_\tau}}(T)$ is defined as in (3.8). Moreover, (5.18) holds.

Proof. The proof is exactly the same as the proof of Lemma 5.10, except that now ι_τ is replaced by $\tilde{\iota}_\tau$. □

5.5.4 Case D

We say that case D applies when **D1**, **D2**, **N1**, **N2** hold and Δ is given by

$$\Delta(u) := \frac{u}{\iota_\nu (t^*)^{\iota_\nu - 1}} \overleftarrow{\tau} \left(\frac{\sqrt{2}\sigma(\mu(ut^*))}{u(1+t^*)} \right). \tag{5.24}$$

The local behavior is described by the following lemma.

Lemma 5.13 *Let* **S1** *and* **M1** *hold and suppose that case D applies. Let δ be such that $\delta(u) = o(u)$ and $\Delta(u) = o(\delta(u))$. For any u and $-\frac{\delta(u)}{T\Delta(u)} \le k \le \frac{\delta(u)}{T\Delta(u)}$, pick some $t_k^\circ(u) \in I_k^T(u)$. Then we have for $u \to \infty$,*

$$\mathbb{P}\left(\sup_{t \in I_k^T(u)} \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))} > \frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))} \right) \sim \mathcal{H}_{B_{\tilde{\iota}_\tau}}(T)\Psi\left(\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))} \right),$$

where $\mathcal{H}_{B_{\tilde{\iota}_\tau}}(T)$ is defined as in (3.8). Moreover, (5.18) holds.

Proof. The arguments are similar to those in the proof of Lemma 5.10. Therefore, we only show how the functions in Lemma 5.9 should be chosen in order to match (5.4) with (5.14).

Define for $-\frac{\delta(u)}{T\Delta(u)} \le k \le \frac{\delta(u)}{T\Delta(u)}$,

$$\kappa_k(u) := \frac{u\tau(\iota_\nu (t^*)^{\iota_\nu - 1}\Delta(u)/u)(1+t_k^\circ(u))}{\sqrt{2}\sigma(\mu(ut_k^\circ))}, \quad g_k(u) := \frac{\sqrt{2}\kappa_k(u)}{\tau(\iota_\nu (t^*)^{\iota_\nu - 1}\Delta(u)/u)},$$

and

$$\theta_k(u,s,t) := 2\frac{\tau^2(\nu(|ut_u^* + (kT+s)\Delta(u)) - \nu(ut_u^* + (kT+t)\Delta(u))|/\nu(u))}{\tau^2(\iota_\nu (t^*)^{\iota_\nu - 1}\Delta(u)/u)}.$$

It follows from Lemma 5.9 with $\eta = B_{\check{\iota}_\tau}$ that

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))} > \frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right) &= \mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))} > \frac{\sqrt{2}\kappa_k(u)}{\tau(\iota_\nu(t^*)^{\iota_\nu-1}\Delta(u)/u)}\right)\\ &\sim \mathcal{H}_{B_{\check{\iota}_\tau}}(T)\Psi\left(\frac{\sqrt{2}\kappa_k(u)}{\tau(\iota_\nu(t^*)^{\iota_\nu-1}\Delta(u)/u)}\right)\\ &= \mathcal{H}_{B_{\check{\iota}_\tau}}(T)\Psi\left(\frac{u(1+t_k^\circ(u))}{\sigma(\mu(ut_k^\circ(u)))}\right),\end{aligned}$$

as claimed. □

5.6 Upper bounds

In this section, we prove the 'upper bound' part of Theorem 5.1 in each of the four cases. Since the proof is almost exactly the same for each of the regimes, we only give it once by using the following notation in both the present and the next section.

We denote the Pickands' constants $\mathcal{H}_{B_{\iota_\tau}}(T)$, $\mathcal{H}_{\mathcal{DM}^{\tau^2}}(T)$, and $\mathcal{H}_{B_{\check{\iota}_\tau}}(T)$ by $\mathcal{H}(T)$. The abbreviation $\mathcal{H} := \lim_{T\to\infty}\mathcal{H}(T)/T$ is used for the corresponding limits. The definition of Δ also depends on the regime; it is defined in (5.17), (5.22), (5.23), and (5.24) for the cases A, B, C, and D, respectively. Note that the dependence on Δ is suppressed in the notation $I_k^T(u) = [t_u^* + kT\Delta(u)/u, t_u^* + (k+1)T\Delta(u)/u]$. It is convenient to define $\underline{t}_k^T(u)$ and $\overline{t}_k^T(u)$ as the left and right end of $I_k^T(u)$ respectively. In the proofs of the upper and lower bounds, we write

$$\mathcal{C} := \frac{1}{2}\frac{d^2}{dt^2}\frac{(1+t)^2}{t^{2H/\beta}}\bigg|_{t=t^*} = (t^*)^{-2H/\beta-1}. \tag{5.25}$$

We start with an auxiliary lemma, which shows that it suffices to focus on *local* behavior near t_u^*. This observation is important since the lemmas of the previous section only yield local uniformity (note that $I_k^T(u) \subset [t_u^* \pm \delta(u)/u]$ and $\delta(u) = o(u)$).

Lemma 5.14 *Suppose that* **S1**–**S4**, *and* **M1**–**M4** *hold for some* $\beta > H$. *Let* δ *be such that* $\delta(u) = o(u)$ *and* $\sigma(\mu(u)) = o(\delta(u))$. *Then we have*

$$\mathbb{P}\left(\sup_{t\notin[t_u^*\pm\delta(u)/u]}\frac{Y_{\mu(ut)}}{1+t} > u\right) = o\left(\frac{\sigma(\mu(u))}{\Delta(u)}\Psi\left(\inf_{t\geq 0}\frac{u(1+t)}{\sigma(\mu(ut))}\right)\right). \tag{5.26}$$

Proof. The proof consists of three parts: we show that the intervals $[0,\omega]$, $[\omega,T]\backslash[t_u^*\pm\delta(u)/u]$ and $[T,\infty]$ play no role in the asymptotics, where $\omega, T > 0$ are chosen appropriately.

We start with the interval $[T,\infty)$. If T is chosen as in **S4**, this interval is asymptotically negligible by assumption.

As for the remaining intervals, by **S4** we can find some $\epsilon, C \in (0,\infty)$, $\gamma \in (0,2]$ such that for each $s,t \in [0,(1+\epsilon)T^{1/\beta}]$

$$\mathbb{V}\mathrm{ar}\,(Y_{us} - Y_{ut}) \leq C\sigma^2(u)|s-t|^\gamma, \tag{5.27}$$

where u is large. Starting with $[0,\omega]$, we select ω so that for large u,

$$\sup_{t\in[0,\omega]}\frac{\sigma(\mu(ut))}{1+t} \leq \frac{1}{2}\frac{\sigma(\mu(ut_u^*))}{1+t_u^*}. \tag{5.28}$$

The main argument is Borell's inequality, but we first have to make sure that it can be applied. For $a > 0$, there exists constants c_γ, C independent of u and a such that for large u, {**M2**}

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[0,\omega]}\frac{Y_{\mu(ut)}}{\sigma(\mu(u))(1+t)}>a\right) &\le \mathbb{P}\left(\sup_{t\in\left[0,\left(\frac{\mu(u\omega)}{\mu(u)}\right)^\beta\right]}\frac{Y_{\mu(u)t^{1/\beta}}}{\sigma(\mu(u))}>a\right)\\ &\le \mathbb{P}\left(\sup_{t\in[0,2\omega]}\frac{Y_{\mu(u)t^{1/\beta}}}{\sigma(\mu(u))}>a\right)\\ &\le 4\exp\left(-\frac{c_\gamma a^2}{C}\right),\end{aligned}$$

where the last inequality follows from (5.27) and Fernique's lemma [211, p. 219] as $\gamma \in (0, 2]$. By choosing a sufficiently large, we have by Borell's inequality (Lemma 3.3)

$$\mathbb{P}\left(\sup_{t\in[0,\omega]}\frac{Y_{\mu(ut)}}{1+t}>u\right)\le 2\Psi\left(\frac{1-a\sigma(\mu(u))/u}{\sup_{t\in[0,\omega]}\frac{\sigma(\mu(ut))}{u(1+t)}}\right).$$

Since (5.28) holds, there exist constants $\mathcal{K}_1, \mathcal{K}_2 < \infty$ such that

$$\mathbb{P}\left(\sup_{t\in[0,\omega]}\frac{Y_{\mu(ut)}}{1+t}>u\right)\le \mathcal{K}_1\exp\left(-2\frac{u^2(1+t_u^*)^2}{\sigma^2(\mu(ut_u^*))}+\mathcal{K}_2\frac{u(1+t_u^*)}{\sigma(\mu(ut_u^*))}\right).$$

This shows that the interval $[0, \omega]$ is asymptotically negligible in the sense of (5.26).

We next consider the contribution of the set $[\omega, T]\backslash[t_u^* \pm \delta(u)/u]$ to the asymptotics. Define

$$\overline{\sigma}(u)=\sup_{t\in[\omega,T]\backslash[t_u^*\pm\delta(u)/u]}\frac{\sigma(\mu(ut))}{1+t}=\max\left(\frac{\sigma(\mu(ut_u^*-\delta(u)))}{1+t_u^*-\delta(u)/u},\frac{\sigma(\mu(ut_u^*+\delta(u)))}{1+t_u^*+\delta(u)/u}\right),$$

where the last equality holds for large u. Now observe that by the UCT {**M1**}, for large u,

$$\begin{aligned}\mathbb{P}\left(\sup_{t\in[\omega,T]\backslash[t_u^*\pm\delta(u)/u]}\frac{Y_{\mu(ut)}}{1+t}>u\right) &\le \mathbb{P}\left(\sup_{t\in[\omega,T]\backslash[t_u^*\pm\delta(u)/u]}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}>\frac{u}{\overline{\sigma}(u)}\right)\\ &\le \mathbb{P}\left(\sup_{t\in\left[\omega^{1/\beta}/2,2T^{1/\beta}\right]}\frac{Y_{\mu(u)t}}{\sigma(\mu(u)t)}>\frac{u}{\overline{\sigma}(u)}\right).\end{aligned}$$

In order to further bound this quantity, we use (5.27) and the inequality $2ab \le a^2 + b^2$: for $s, t \in \left[\omega^{1/\beta}/2, 2T^{1/\beta}\right]$, {**M2**}

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(u)s}}{\sigma(\mu(u)s)}-\frac{Y_{\mu(u)t}}{\sigma(\mu(u)t)}\right) &\le \frac{\mathbb{V}\mathrm{ar}\left(Y_{\mu(u)s}-Y_{\mu(u)t}\right)}{\sigma(\mu(u)s)\sigma(\mu(u)t)}\\ &\le \sup_{v\in[\omega^{1/\beta}/2,2T^{1/\beta}]}\frac{\mathbb{V}\mathrm{ar}\left(Y_{\mu(u)s}-Y_{\mu(u)t}\right)}{\sigma^2(\mu(u)v)}\\ &\le \frac{2^{1+2H}\omega^{-2H/\beta}}{\sigma^2(\mu(u))}\mathbb{V}\mathrm{ar}\left(Y_{\mu(u)s}-Y_{\mu(u)t}\right)\\ &\le \mathcal{K}'|s-t|^\gamma,\end{aligned}$$

where $\mathcal{K}' < \infty$ is some constant (depending on ω and T). Hence, by Theorem D.4 of Piterbarg [257] there exists a constant $\mathcal{K}''$ depending only on $\mathcal{K}'$ and γ such that

$$\mathbb{P}\left(\sup_{t\in[\omega,T]\setminus[t_u^*\pm\delta(u)/u]}\frac{Y_{\mu(ut)}}{1+t}>u\right)\leq T\mathcal{K}''\left(\frac{u}{\overline{\sigma}(u)}\right)^{2/\gamma}\Psi\left(\frac{u}{\overline{\sigma}(u)}\right).$$

Consider the expression

$$\frac{u^2}{\mathcal{C}}\left(\frac{(1+t_u^*+\delta(u)/u)^2}{\sigma^2(\mu(ut_u^*+\delta(u)))}-\frac{(1+t_u^*)^2}{\sigma^2(\mu(ut_u^*))}\right)\Bigg/\left[\frac{\delta(u)}{\sigma(\mu(u))}\right]^2, \tag{5.29}$$

where $\mathcal{C}$ is given by (5.25). By Taylor's mean value theorem {**S3**, **M4**}, there exists some $t_\# = t_\#(u) \in [t_u^*, t_u^* + \delta(u)/u]$ such that this expression equals

$$\frac{\delta^2(u)}{2\mathcal{C}}\frac{d^2}{dt^2}\frac{(1+t)^2}{\sigma^2(\mu(ut))}\bigg|_{t=t_\#}\Bigg/\left[\frac{\delta(u)}{\sigma(\mu(u))}\right]^2.$$

Recall that $\sigma^2(\mu(\cdot))$ is regularly varying with index $2H/\beta > 0$, and that (under the present conditions) both its first and second derivative are regularly varying with respective indices $2H/\beta - 1$ and $2H/\beta - 2$. The UCT now yields

$$\lim_{u\to\infty}\frac{\sigma^2(\mu(u))}{2}\frac{d^2}{dt^2}\frac{(1+t)^2}{\sigma^2(\mu(ut))}\bigg|_{t=t_\#}=\mathcal{C}.$$

Since $\sigma(\mu(u)) = o(\delta(u))$, the expression in (5.29) converges to 1 as $u \to \infty$. Hence, we have

$$\frac{\Psi\left(\frac{u}{\overline{\sigma}(u)}\right)}{\Psi\left(\frac{u(1+t_u^*)}{\sigma(\mu(ut_u^*))}\right)}=\exp\left(-\frac{1}{2}\mathcal{C}\frac{\delta^2(u)}{\sigma^2(\mu(u))}(1+o(1))\right)(1+o(1)),$$

showing that the interval $[\omega, T]\setminus[t_u^* \pm \delta(u)/u]$ plays no role in the asymptotics. □

We can now prove the upper bounds. In the proof, it is essential that $\sigma(\mu(u))/\Delta(u) \to \infty$ in all four cases. To see that this holds, note that this function is regularly varying with index $(1 - H/\beta)(1/\iota_\tau - 1) > 0$ in case A and B (use $\iota_\nu = (1 - H/\beta)/\iota_\tau$ in the latter case). In case C, the index of variation is

$$\frac{H}{\beta}+\iota_\nu-1+\frac{1-\iota_\tau\iota_\nu-H/\beta}{\tilde{\iota}_\tau}>\left(1-\iota_\tau\iota_\nu-\frac{H}{\beta}\right)\left(\frac{1}{\tilde{\iota}_\tau}-1\right)>0.$$

Finally, it is regularly varying with index $(1 - H/\beta)(1/\tilde{\iota}_\tau - 1) > 0$ in case D.

The upper bounds are stated in the following proposition.

Proposition 5.15 *Let μ and σ satisfy assumptions* **M1**–**M4** *and* **S1**–**S4** *for some $\beta > H$. Moreover, let case A, B, C, or D apply. We then have*

$$\limsup_{u\to\infty}\frac{\mathbb{P}\left(\sup_{t\geq0}Y_{\mu(t)}-t>u\right)}{\frac{\sigma(\mu(u))}{\Delta(u)}\Psi\left(\inf_{t\geq0}\frac{u(1+t)}{\sigma(\mu(ut))}\right)}\leq\mathcal{H}\sqrt{\frac{2\pi}{\mathcal{C}}}.$$

Proof. Select some δ such that $\delta(u) = o(u)$, $\sigma(\mu(u)) = o(\delta(u))$, $\Delta(u) = o(\delta(u))$, and $u = o(\delta(u)\nu(u))$. While the specific choice is irrelevant, it is left to the reader that such δ exists in each of the four cases. In view of Lemma 5.14, we need to show that

$$\limsup_{u\to\infty}\frac{\mathbb{P}\left(\sup_{t\in[t_u^*\pm\delta(u)/u]}\frac{Y_{\mu(ut)}}{1+t}>u\right)}{\frac{\sigma(\mu(u))}{\Delta(u)}\Psi\left(\frac{u(1+t_u^*)}{\sigma(\mu(ut_u^*))}\right)}\leq\mathcal{H}\sqrt{\frac{2\pi}{\mathcal{C}}}.$$

For this, note that by definition of t_u^* and continuity of σ and μ, for large u,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[t_u^*\pm\delta(u)/u]}\frac{Y_{\mu(ut)}}{1+t}>u\right)\\
&\le \sum_{-\frac{\delta(u)}{T\Delta(u)}\le k\le\frac{\delta(u)}{T\Delta(u)}}\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{1+t}>u\right)\\
&\le \sum_{0\le k\le\frac{\delta(u)}{T\Delta(u)}}\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}>\frac{u(1+\underline{t}_k^T(u))}{\sigma(\mu(u\underline{t}_k^T(u)))}\right)\\
&\quad+\sum_{-\frac{\delta(u)}{T\Delta(u)}\le k<0}\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}>\frac{u(1+\overline{t}_k^T(u))}{\sigma(\mu(u\overline{t}_k^T(u)))}\right).
\end{aligned}\tag{5.30}$$

By Lemmas 5.10–5.13, the UCT, and (3.2), as $u\to\infty$,

$$\begin{aligned}
&\frac{\Delta(u)}{\sigma(\mu(u))}\sum_{0\le k\le\frac{\delta(u)}{T\Delta(u)}}\frac{\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}>\frac{u(1+\overline{t}_k^T(u))}{\sigma(\mu(u\overline{t}_k^T(u)))}\right)}{\Psi\left(\frac{u(1+t_u^*)}{\sigma(\mu(ut_u^*))}\right)}\\
&=\mathcal{H}(T)\frac{\Delta(u)}{\sigma(\mu(u))}\sum_{0\le k\le\frac{\delta(u)}{T\Delta(u)}}\left[\frac{\Psi\left(\frac{u(1+\overline{t}_k^T(u))}{\sigma(\mu(u\overline{t}_k^T(u)))}\right)}{\Psi\left(\frac{u(1+t_u^*)}{\sigma(\mu(ut_u^*))}\right)}(1+o(1))\right]\\
&=\mathcal{H}(T)\frac{\Delta(u)}{\sigma(\mu(u))}\sum_{0\le k\le\frac{\delta(u)}{T\Delta(u)}}\left[\frac{\exp\left(-\frac12\frac{u^2(1+\overline{t}_k^T(u))^2}{\sigma^2(\mu(u\overline{t}_k^T(u)))}\right)}{\exp\left(-\frac12\frac{u^2(1+t_u^*)^2}{\sigma^2(\mu(ut_u^*))}\right)}(1+o(1))\right].
\end{aligned}\tag{5.31}$$

As in the proof of Lemma 5.14, one can show that, uniformly in k by the UCT,

$$u^2\left(\frac{(1+\overline{t}_k^T(u))^2}{\sigma^2(\mu(u\overline{t}_k^T(u)))}-\frac{(1+t_u^*)^2}{\sigma^2(\mu(ut_u^*))}\right)\Big/\,\mathcal{C}\left[\frac{(k+1)T\Delta(u)}{\sigma(\mu(u))}\right]^2\to 0,$$

where $\mathcal{C}$ is given in (5.25). Hence, (5.31) can be written as

$$\frac{\mathcal{H}(T)}{T}\frac{T\Delta(u)}{\sigma(\mu(u))}\sum_{0\le k\le\frac{\delta(u)}{T\Delta(u)}}\left[\exp\left(-\frac12\mathcal{C}\frac{[(k+1)T\Delta(u)]^2}{\sigma^2(\mu(u))}(1+o(1))\right)(1+o(1))\right].$$

By Lemmas 5.10–5.13, the fact that $\sigma(\mu(u))=o(u)$, and the dominated convergence theorem, this tends to

$$\frac{\mathcal{H}(T)}{T}\int_0^\infty\exp\left(-\frac12\mathcal{C}x^2\right)dx=\frac{\mathcal{H}(T)}{T}\sqrt{\frac{\pi/2}{\mathcal{C}}}.$$

The second term in (5.30) is bounded from above similarly. Hence, we have shown that for any $T>0$,

$$\limsup_{u\to\infty}\frac{\Delta(u)}{\sigma(\mu(u))}\frac{\mathbb{P}\left(\sup_{t\in[t_u^*+\delta(u)/u]}\frac{Y_{\mu(ut)}}{1+t}>u\right)}{\Psi\left(\frac{u(1+t_u^*)}{\sigma(\mu(ut_u^*))}\right)}\le\frac{\mathcal{H}(T)}{T}\sqrt{\frac{2\pi}{\mathcal{C}}}.$$

The claim is obtained by letting $T\to\infty$. □

5.7 Lower bounds

In this section, we prove the 'lower bound' part of Theorem 5.1 using an appropriate modification of the corresponding argument in the double sum method. For notational conventions, see Section 5.6.

Proposition 5.16 *Let μ and σ satisfy assumptions* **M1**–**M4** *and* **S1**–**S4** *for some $\beta > H$. Moreover, let case A, B, C, or D apply. We then have*

$$\liminf_{u\to\infty} \frac{\mathbb{P}\left(\sup_{t\geq 0} Y_{\mu(t)} - t > u\right)}{\frac{\sigma(\mu(u))}{\Delta(u)}\Psi\left(\inf_{t\geq 0}\frac{u(1+t)}{\sigma(\mu(ut))}\right)} \geq \mathcal{H}\sqrt{\frac{2\pi}{\mathcal{C}}}.$$

The proof of this proposition requires some auxiliary observations, resulting in a bound on probabilities involving the maximum on a two-dimensional field. The first step in establishing those bounds is to study the variances; for this, it is convenient to introduce the notation

$$\underline{\sigma}^2_{k,\ell}(u) := \inf_{(s,t)\in I_k^T(u)\times I_\ell^T(u)} \mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))} - \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right)$$

and

$$\overline{\sigma}^2_{k,\ell}(u) := \sup_{(s,t)\in I_k^T(u)\times I_\ell^T(u)} \mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))} - \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right).$$

Lemma 5.17 *Suppose that one of the cases A, B, C, or D applies, and that both $\delta(u) = o(u)$ and $\Delta(u) = o(\delta(u))$. Then there exist constants $\zeta \in (0,2)$ and $\mathcal{K} \in (0,\infty)$, independent of T, such that for large T the following holds. Given $\epsilon > 0$, there exists some u_0 such that for all $u \geq u_0$ and all $-\frac{\delta(u)}{T\Delta(u)} \leq k,\ell \leq \frac{\delta(u)}{T\Delta(u)}$ with $|\ell - k| > 1$,*

$$\underline{\sigma}^2_{k,\ell}(u) \geq (1-\epsilon)^3\mathcal{K}\left[\left(\frac{T(|k-\ell|-1)}{2}\right)^\zeta - \epsilon\right]\frac{\sigma^2(\mu(u))}{u^2}.$$

Moreover,

$$\sup_{\substack{-\frac{\delta(u)}{T\Delta(u)}\leq k,\ell\leq\frac{\delta(u)}{T\Delta(u)}\\ |k-\ell|>1}} \overline{\sigma}^2_{k,\ell}(u) \to 0.$$

Proof. Let $\epsilon > 0$ be given. By (5.3), the first claim is proven for case A, B, and C once it has been shown that for large u, uniformly in $\alpha \in \left[1, \frac{\delta(u)}{T\Delta(u)}\right]$,

$$\inf_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ |s-t|\geq\alpha T\Delta(u)/u}} \frac{\tau^2(|\nu(us)-\nu(ut)|)}{\tau^2(\nu(u))} \geq (1-\epsilon)^2\frac{\mathcal{K}}{\mathcal{D}}\left[\left(\frac{\alpha T}{2}\right)^\zeta - \epsilon\right]\frac{\sigma^2(\mu(u))}{u^2},$$

since one can then set $\alpha = |k-\ell| - 1$. By the mean value theorem {**N2**} we have, for certain $t^\wedge(u,s,t) \in [t_u^* \pm \delta(u)/u]$,

$$\begin{aligned}
\inf_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ |s-t|\geq\alpha T\Delta(u)/u}} \frac{\tau^2(|\nu(us)-\nu(ut)|)}{\tau^2(\nu(u))} &= \inf_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ |s-t|\geq\alpha T\Delta(u)/u}} \frac{\tau^2(u\dot\nu(ut^\wedge(u,s,t))|s-t|)}{\tau^2(\nu(u))}\\
&\geq \inf_{\substack{s,t\in[t_u^*\pm\delta(u)/u]\\ |s-t|\geq\frac{1}{2}\alpha T\Delta(u)/u}} \frac{\tau^2(u\dot\nu(ut^*)|s-t|)}{\tau^2(\nu(u))}\\
&\geq \inf_{t\geq\alpha T/2} \frac{\tau^2(\dot\nu(ut^*)\Delta(u)t)}{\tau^2(\nu(u))},
\end{aligned}$$

where the first inequality follows from the UCT {**N1**}; the details are left to the reader.

We investigate the lower bound in each of the three cases. In case A, $\dot\nu(ut^*)\Delta(u)$ tends to infinity. By the UCT and the definition of Δ, we have for any $\alpha \geq 1$,

$$\begin{aligned}\inf_{t\geq \alpha T/2} \frac{\tau^2(\dot\nu(ut^*)\Delta(u)t)}{\tau^2(\nu(u))} &\geq (1-\epsilon)\frac{\tau^2(\dot\nu(ut^*)\Delta(u))}{\tau^2(\nu(u))}\left[\left(\frac{\alpha T}{2}\right)^{2\iota_\tau} - \epsilon\right] \\ &\geq (1-\epsilon)^2 \frac{2}{\mathcal{D}(1+t^*)^2}\frac{\sigma^2(\mu(ut^*))}{u^2}\left[\left(\frac{\alpha T}{2}\right)^{2\iota_\tau} - \epsilon\right].\end{aligned}$$

Case C is similar, except that now $\dot\nu(ut^*)\Delta(u)$ tends to zero (so that one can apply the UCT as τ is continuous and regularly varying at zero):

$$\inf_{t\geq \alpha T/2} \frac{\tau^2(\dot\nu(ut^*)\Delta(u)t)}{\tau^2(\nu(u))} \geq (1-\epsilon)^2 \frac{2}{\mathcal{D}(1+t^*)^2}\frac{\sigma^2(\mu(ut^*))}{u^2}\left[\left(\frac{\alpha T}{2}\right)^{2\bar\iota_\tau} - \epsilon\right].$$

In case B, we note that $\sigma(\mu(u))\tau(\nu(u)) \sim \mathcal{G}u$ implies that for small $\zeta > 0$, there exists some t_0 such that for $t \geq t_0$, $\tau^2(t) \geq t^\zeta$. Therefore, for T large enough, since $\dot\nu(ut^*)\Delta(u) = 1$, uniformly in $\alpha \geq 1$,

$$\inf_{t\geq \alpha T/2} \frac{\tau^2(\dot\nu(ut^*)\Delta(u)t)}{\tau^2(\nu(u))} \geq \inf_{t\geq \alpha T/2} \frac{t^\zeta}{\tau^2(\nu(u))} = \frac{(\alpha T/2)^\zeta}{\tau^2(\nu(u))} \geq (1-\epsilon)^2 \left(\frac{\alpha T}{2}\right)^\zeta \frac{1}{\mathcal{G}^2}\frac{\sigma^2(\mu(u))}{u^2},$$

implying the stated.

We leave the proof of the assertion for case D to the reader; one then exploits the regular variation of τ at zero and uses the definition of Δ.

To prove the second claim of the lemma in case A, B, and C, we use the mean value theorem and the UCT: {**N1**, **N2**}

$$\begin{aligned}\sup_{s,t\in[t_u^*\pm\delta(u)/u]} \mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))} - \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right) &\sim \sup_{s,t\in[t_u^*\pm\delta(u)/u]} \frac{\mathcal{D}\tau^2(|\nu(us)-\nu(ut)|)}{\tau^2(\nu(u))} \\ &\leq \sup_{s,t\in[t_u^*\pm 2\delta(u)/u]} \frac{\mathcal{D}\tau^2(u\dot\nu(ut^*)|s-t|)}{\tau^2(\nu(u))} \\ &= \sup_{t\in[0,2]} \frac{\mathcal{D}\tau^2(\delta(u)\dot\nu(ut^*)t)}{\tau^2(\nu(u))}.\end{aligned}$$

Since $\delta(u)\dot\nu(ut^*)$ tends to infinity by assumption, **T1** implies that the latter expression is of order $\tau^2(\delta(u)\dot\nu(u))/\tau^2(\nu(u))$. In particular, it tends to zero as $u \to \infty$.

We do not prove the claim for case D, since the same arguments apply. □

The next lemma exploits the two statements of Lemma 5.17 on the correlation structure. Let $\kappa_{k,\ell}$ be arbitrary functions of u which converge uniformly in $-\frac{\delta(u)}{T\Delta(u)} \leq k, \ell \leq \frac{\delta(u)}{T\Delta(u)}$ to $2(1+t^*)$.

Lemma 5.18 *Suppose that one of the cases A, B, C, or D applies, and that $\delta(u) = o(u)$. There exist constants $\alpha, \mathcal{K}' < \infty$, independent of k, ℓ, such that for large u, uniformly in k, ℓ with $|k-\ell| > 1$,*

$$\mathbb{P}\left(\sup_{(s,t)\in I_k^T(u)\times I_\ell^T(u)} \frac{Y_{\mu(us)}}{\sigma(\mu(us))} + \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))} > \frac{u\kappa_{k,\ell}(u)}{\sigma(\mu(ut^*))}\right) \leq \mathcal{K}' T^\alpha \Psi\left(\frac{\frac{u\kappa_{k,\ell}(u)}{\sigma(\mu(ut^*))}}{\sqrt{4-\underline{\sigma}^2_{k,\ell}(u)}}\right). \quad (5.32)$$

Proof. Define

$$Y^*_{(s,t)}(u) := \frac{\frac{Y_{\mu(us)}}{\sigma(\mu(us))} + \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}}{\sqrt{\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))} + \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right)}}, \quad u^*_{k,\ell} = \frac{\frac{u\kappa_{k,\ell}(u)}{\sigma(\mu(ut^*))}}{\sqrt{4 - \underline{\sigma}^2_{k,\ell}(u)}},$$

so that the left-hand side of (5.32) is majorized by

$$\mathbb{P}\left(\sup_{(s,t)\in I^T_k(u)\times I^T_\ell(u)} Y^*_{(s,t)}(u) > u^*_{k,\ell}\right). \tag{5.33}$$

As a consequence of (the second claim in) Lemma 5.17, we have for large u

$$\inf_{k,\ell}\inf_{(s,t)\in I^T_k(u)\times I^T_\ell(u)} \mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))} + \frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}\right) \geq 2.$$

The remainder of the proof closely follows the reasoning on page 102 of Piterbarg [257]. In particular, for $(s,t),(s',t') \in I^T_k(u) \times I^T_\ell(u)$, we have

$$\begin{aligned}&\mathbb{V}\mathrm{ar}\left(Y^*_{(s,t)}(u) - Y^*_{(s',t')}(u)\right)\\ &\quad\leq\quad 4\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(us)}}{\sigma(\mu(us))} - \frac{Y_{\mu(us')}}{\sigma(\mu(us'))}\right) + 4\mathbb{V}\mathrm{ar}\left(\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))} - \frac{Y_{\mu(ut')}}{\sigma(\mu(ut'))}\right).\end{aligned} \tag{5.34}$$

Define

$$\upsilon(u) := \begin{cases} \sqrt{2}\tau(\nu(u))\frac{\sigma(\mu(ut^*))}{u(1+t^*)} & \text{in case A, C and D;}\\ \frac{1}{2}\sqrt{2\mathcal{D}} & \text{in case B.}\end{cases}$$

Now we have to distinguish between case D and the other cases. First we focus on the cases A, B, and C; then one can use (5.3) to see that (5.34) is asymptotically at most

$$4\frac{\mathcal{D}\tau^2(|\nu(us) - \nu(us')|)}{\tau^2(\nu(u))} + 4\frac{\mathcal{D}\tau^2(|\nu(ut) - \nu(ut')|)}{\tau^2(\nu(u))}. \tag{5.35}$$

As shown in the proofs of Lemmas 5.10–5.12,

$$\limsup_{u\to\infty} \sup_{-\frac{\delta(u)}{T\Delta(u)}\leq k\leq \frac{\delta(u)}{T\Delta(u)}} \sup_{(s,t)\in I^T_k(u)} \frac{\mathcal{D}\tau^2(|\nu(us) - \nu(ut)|)}{\upsilon^2(u)} \left(\frac{u}{\Delta(u)}(s-t)\right)^{-2\gamma'} \leq 2T^{\alpha'},$$

where $\alpha' = 2(\iota_\tau - \gamma')$ in case A and B, and $\alpha' = 2(\tilde{\iota}_\tau - \gamma')$ in case C. Therefore, we find the following asymptotic upper bound for (5.35) and hence for (5.34):

$$8T^{\alpha'}\frac{\upsilon^2(u)}{\tau^2(\nu(u))}\left[\left(\frac{u}{\Delta(u)}(s-s')\right)^{2\gamma'} + \left(\frac{u}{\Delta(u)}(t-t')\right)^{2\gamma'}\right]. \tag{5.36}$$

We now show that (5.36) is also an asymptotic upper bound in case D. For this, we note that in this case (5.34) is asymptotically at most

$$4\tau^2\left(\frac{|\nu(us) - \nu(us')|}{\nu(u)}\right) + 4\tau^2\left(\frac{|\nu(ut) - \nu(ut')|}{\nu(u)}\right),$$

and the reader can check with the mean value theorem and the UCT that (5.36) holds for $\gamma' = \tilde{\iota}_\tau/2$ and $\alpha' = \tilde{\iota}_\tau$ (say).

For any u, we now introduce two independent centered Gaussian stationary processes $\vartheta_1^{(u)}$ and $\vartheta_2^{(u)}$. These processes have unit variance and covariance function equal to

$$r_\vartheta^{(u)}(t) := \mathbb{C}\mathrm{ov}\left(\vartheta_i^{(u)}(t), \vartheta_i^{(u)}(0)\right) = \exp\left(-32\frac{\upsilon^2(u)}{\tau^2(\nu(u))}t^{2\gamma'}\right).$$

Observe that $\upsilon^2(u)/\tau^2(\nu(u)) \to 0$ in each of the four cases, so that for $s,t,s',t' \in [0,T]$ and u large enough,

$$\begin{aligned}
&\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{2}}\left[\vartheta_1^{(u)}(s) + \vartheta_2^{(u)}(t) - \vartheta_1^{(u)}(s') - \vartheta_2^{(u)}(t')\right]\right) \\
&\quad = \quad 2 - \exp\left(-32\frac{\upsilon^2(u)}{\tau^2(\nu(u))}|s-s'|^{2\gamma'}\right) - \exp\left(-32\frac{\upsilon^2(u)}{\tau^2(\nu(u))}|t-t'|^{2\gamma'}\right) \\
&\quad \geq \quad 16\frac{\upsilon^2(u)}{\tau^2(\nu(u))}|s-s'|^{2\gamma'} + 16\frac{\upsilon^2(u)}{\tau^2(\nu(u))}|t-t'|^{2\gamma'}.
\end{aligned}$$

We now apply Slepian's inequality (Lemma 3.4) to compare the suprema of the two fields Y^* and $2^{-1/2}\left[\vartheta_1^{(u)} + \vartheta_2^{(u)}\right]$: for $-\frac{\delta(u)}{T\Delta(u)} \leq k,\ell \leq \frac{\delta(u)}{T\Delta(u)}$, (5.33) is majorized by

$$\begin{aligned}
&\mathbb{P}\left(\sup_{(s,t)\in[0,T]^2}\frac{1}{\sqrt{2}}\left[\vartheta_1^{(u)}(T^{\alpha'/(2\gamma')}s) + \vartheta_2^{(u)}(T^{\alpha'/(2\gamma')}t)\right] > u_{k,\ell}^*\right) \\
&\quad = \quad \mathbb{P}\left(\sup_{(s,t)\in[0,T^{\alpha'/(2\gamma')+1}]^2}\frac{1}{\sqrt{2}}\left[\vartheta_1^{(u)}(s) + \vartheta_2^{(u)}(t)\right] > u_{k,\ell}^*\right). \qquad (5.37)
\end{aligned}$$

Lemma 5.9 is used to investigate the asymptotics of (5.37), yielding the desired bound. For notational convenience, we set $T' = T^{\alpha'/(2\gamma')+1}$. Observe that the mapping

$$(\alpha_1,\alpha_2) \mapsto [2 - \exp(-\alpha_1) - \exp(-\alpha_2)]/[\alpha_1+\alpha_2] - 1$$

is nonpositive and that the minimum over the set $[0,\theta]^2$ is achieved at $(\alpha_1,\alpha_2) = (\theta,\theta)$. Therefore,

$$\sup_{(s,t),(s',t')\in[0,T']^2}\left|\frac{2 - r_\vartheta^{(u)}(|s-s'|) - r_\vartheta^{(u)}(|t-t'|)}{32\frac{\upsilon^2(u)}{\tau^2(\nu(u))}\left[|s-s'|^{2\gamma'} + |t-t'|^{2\gamma'}\right]} - 1\right| = 1 - \frac{2 - r_\vartheta^{(u)}(T') - r_\vartheta^{(u)}(T')}{64\frac{\upsilon^2(u)}{\tau^2(\nu(u))}(T')^{2\gamma'}},$$

which tends to zero if $u \to \infty$. Moreover, we have

$$\sup_{-\frac{\delta(u)}{T\Delta(u)}\leq k,\ell\leq\frac{\delta(u)}{T\Delta(u)}}\left|\frac{\sigma^2(\mu(ut^*))(u_{k,\ell}^*)^2}{u^2(1+t^*)^2} - 1\right| \to 0.$$

To see that Lemma 5.9 can be applied, set $g_{k,\ell}(u) = u_{k,\ell}^*$, and

$$\theta_{k,\ell}(u,s,s',t,t') := 32(1+t^*)^2\frac{u^2\upsilon^2(u)}{\sigma^2(\mu(ut^*))\tau^2(\nu(u))}\left[|s-s'|^{2\gamma'} + |t-t'|^{2\gamma'}\right].$$

Condition **P1** clearly holds, and $\theta_{k,\ell}(u,s,s',t,t')$ tends to

$$2\xi_\eta(s,s',t,t') := \begin{cases} 64\left[|s-s'|^{2\gamma'} + |t-t'|^{2\gamma'}\right] & \text{in case A, C, and D;} \\ \frac{16(1+t^*)^2}{(t^*)^{2H/\beta}\mathcal{G}^2}\left[|s-s'|^{2\gamma'} + |t-t'|^{2\gamma'}\right] & \text{in case B,} \end{cases}$$

showing that **P2** also holds. As **P3** is immediate, it remains to investigate **P4**. The reasoning in the proof of Lemma 5.10 shows that it suffices to have

$$\lim_{\varepsilon\to 0}\limsup_{u\to\infty}\sup_{k,\ell}\sup_{|s-s'|^{2\gamma'}+|t-t'|^{2\gamma'}<\varepsilon}\theta_{k,\ell}(u,s,s',t,t')<\infty,$$

which is trivial. Define for $s,t\in[0,T']$,

$$\eta(s,t):=B^1_{\gamma'}(s)+B^2_{\gamma'}(t),$$

where $B^1_{\gamma'}$ and $B^2_{\gamma'}$ are independent fractional Brownian motions with Hurst parameter γ'. Then, the probability in (5.37) is asymptotically equivalent to

$$\begin{cases}\mathbb{E}\exp\left(\sup_{(s,t)\in[0,T']^2}8\eta(s,t)-32s^{2\gamma'}-32t^{2\gamma'}\right)\Psi(u^*_{k,\ell}) & \text{in case A, C, D;}\\ \mathbb{E}\exp\left(\sup_{(s,t)\in[0,T']^2}\frac{4(1+t^*)}{(t^*)^{H/\beta}\mathcal{G}}\eta(s,t)-\frac{8(1+t^*)^2}{(t^*)^{2H/\beta}\mathcal{G}^2}\left[s^{2\gamma'}+t^{2\gamma'}\right]\right)\Psi(u^*_{k,\ell}) & \text{in case B.}\end{cases}$$

By exploiting the self-similarity of fractional Brownian motion one can see that the expectation equals $(T')^2\mathcal{K}'$ for some constant $\mathcal{K}'<\infty$. □

Proof of Proposition 5.16. Note that

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in[t^*_u\pm\delta(u)/u]}\frac{Y_{\mu(ut)}}{1+t}>u\right)\\ &\geq\sum_{-\frac{\delta(u)}{T\Delta(u)}\leq k\leq\frac{\delta(u)}{T\Delta(u)}}\mathbb{P}\left(\sup_{t\in I^T_k(u)}\frac{Y_{\mu(ut)}}{1+t}>u;\ \sup_{t\in[t^*_u\pm\delta(u)/u]\setminus I^T_k(u)}\frac{Y_{\mu(ut)}}{1+t}\leq u\right)\\ &=\sum_{-\frac{\delta(u)}{T\Delta(u)}\leq k\leq\frac{\delta(u)}{T\Delta(u)}}\mathbb{P}\left(\sup_{t\in I^T_k(u)}\frac{Y_{\mu(ut)}}{1+t}>u\right)\\ &\quad-\sum_{-\frac{\delta(u)}{T\Delta(u)}\leq k\leq\frac{\delta(u)}{T\Delta(u)}}\mathbb{P}\left(\sup_{t\in I^T_k(u)}\frac{Y_{\mu(ut)}}{1+t}>u;\ \sup_{t\in[t^*_u\pm\delta(u)/u]\setminus I^T_k(u)}\frac{Y_{\mu(ut)}}{1+t}>u\right).\end{aligned}\tag{5.38}$$

A similar reasoning as in the proof of Proposition 5.15 can be used to see that

$$\lim_{T\to\infty}\liminf_{u\to\infty}\frac{\sum_{-\frac{\delta(u)}{T\Delta(u)}\leq k\leq\frac{\delta(u)}{T\Delta(u)}}\mathbb{P}\left(\sup_{t\in I^T_k(u)}\frac{Y_{\mu(ut)}}{1+t}>u\right)}{\frac{\sigma(\mu(u))}{\Delta(u)}\Psi\left(\frac{u(1+t^*_u)}{\sigma(\mu(ut^*_u))}\right)}\geq\mathcal{H}\sqrt{\frac{2\pi}{\mathcal{C}}}.$$

It remains to find an appropriate upper bound for the second term in (5.38). For this, observe that

$$\begin{aligned}&\mathbb{P}\left(\sup_{t\in I^T_k(u)}\frac{Y_{\mu(ut)}}{1+t}>u;\ \sup_{t\in[t^*_u\pm\delta(u)/u]\setminus I^T_k(u)}\frac{Y_{\mu(ut)}}{1+t}>u\right)\\ &\leq\mathbb{P}\left(\sup_{t\in I^T_k(u)}\frac{Y_{\mu(ut)}}{1+t}>u;\ \sup_{t\in\left[t^*_u-\frac{\delta(u)}{u},\underline{t}^T_k(u)-\sqrt{T}\frac{\Delta(u)}{u}\right)\cup\left(\overline{t}^T_k(u)+\sqrt{T}\frac{\Delta(u)}{u},t^*_u+\frac{\delta(u)}{u}\right]}\frac{Y_{\mu(ut)}}{1+t}>u\right)\\ &\quad+\mathbb{P}\left(\sup_{t\in\left[\underline{t}^T_k(u)-\sqrt{T}\frac{\Delta(u)}{u},\underline{t}^T_k(u)\right)}\frac{Y_{\mu(ut)}}{1+t}>u\right)+\mathbb{P}\left(\sup_{t\in\left(\overline{t}^T_k(u),\overline{t}^T_k(u)+\sqrt{T}\frac{\Delta(u)}{u}\right]}\frac{Y_{\mu(ut)}}{1+t}>u\right)\\ &=:\ p_1(u,k,T)+p_2(u,k,T)+p_3(u,k,T).\end{aligned}$$

One can apply the arguments that are detailed in the proof of Proposition 5.15 to infer that

$$\limsup_{u\to\infty} \frac{\sum_{-\frac{\delta(u)}{T\Delta(u)}\le k\le \frac{\delta(u)}{T\Delta(u)}} p_2(u,k,T)}{\frac{\sigma(\mu(u))}{\Delta(u)}\Psi\left(\frac{u(1+t_u^*)}{\sigma(\mu(ut_u^*))}\right)} \le \frac{\mathcal{H}\left(\sqrt{T}\right)}{T}\sqrt{\frac{2\pi}{\mathcal{C}}},$$

which converges to zero as $T\to\infty$. The term $p_3(u,k,T)$ is bounded from above similarly.

We now study $\sum_k p_1(u,k,T)$ in more detail; for this we need the technical lemmas that were established earlier. Observe that it is majorized by

$$\begin{aligned}
&\sum_{-\frac{\delta(u)}{T\Delta(u)}\le k<\ell\le\frac{\delta(u)}{T\Delta(u)}} 2\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{1+t}>u;\ \sup_{t\in I_\ell^T(u)}\frac{Y_{\mu(ut)}}{1+t}>u\right)\\
&\quad+\sum_{-\frac{\delta(u)}{T\Delta(u)}\le k\le\frac{\delta(u)}{T\Delta(u)}}\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{1+t}>u;\ \sup_{t\in[\bar t_u^k+\sqrt{T}\frac{\Delta(u)}{u},\bar t_u^k+(T+\sqrt{T})\frac{\Delta(u)}{u}]}\frac{Y_{\mu(ut)}}{1+t}>u\right)\\
&\quad+\sum_{-\frac{\delta(u)}{T\Delta(u)}\le k\le\frac{\delta(u)}{T\Delta(u)}}\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{1+t}>u;\ \sup_{t\in[\underline t_u^k-(T+\sqrt{T})\frac{\Delta(u)}{u},\underline t_u^k-\sqrt{T}\frac{\Delta(u)}{u}]}\frac{Y_{\mu(ut)}}{1+t}>u\right)\\
&=:\ I(u,T)+II(u,T)+III(u,T).
\end{aligned}$$

By symmetry, $I(u,T)$ is bounded from above by

$$2\sum_{\substack{-\frac{\delta(u)}{T\Delta(u)}\le k\le\frac{\delta(u)}{T\Delta(u)}\\ |k-\ell|>1,\sup_{t\in I_k^T(u)}\frac{\sigma(\mu(ut))}{1+t}\le\sup_{t\in I_\ell^T(u)}\frac{\sigma(\mu(ut))}{1+t}}}\ \sum_{-\frac{\delta(u)}{T\Delta(u)}\le\ell\le\frac{\delta(u)}{T\Delta(u)}}\mathbb{P}\left(\sup_{t\in I_k^T(u)}\frac{Y_{\mu(ut)}}{1+t}>u;\ \sup_{t\in I_\ell^T(u)}\frac{Y_{\mu(ut)}}{1+t}>u\right).$$

Each of the summands cannot exceed

$$\mathbb{P}\left(\sup_{(s,t)\in I_k^T(u)\times I_\ell^T(u)}\frac{Y_{\mu(us)}}{\sigma(\mu(us))}+\frac{Y_{\mu(ut)}}{\sigma(\mu(ut))}>\inf_{t\in I_k^T(u)}\frac{2u(1+t)}{\sigma(\mu(ut))}\right),$$

and we are in the setting of Lemma 5.18. Hence, there exist constants $\mathcal{K}',\alpha$ such that $I(u,T)$ is majorized by

$$2\mathcal{K}'T^\alpha\sum_{-\frac{\delta(u)}{T\Delta(u)}\le k\le\frac{\delta(u)}{T\Delta(u)}}\ \sum_{\substack{-\frac{\delta(u)}{T\Delta(u)}\le\ell\le\frac{\delta(u)}{T\Delta(u)}\\ |k-\ell|>1}}\Psi\left(\frac{\inf_{t\in I_k^T(u)}\frac{2u(1+t)}{\sigma(\mu(ut))}}{\sqrt{4-\underline{\sigma}_{k,\ell}^2(u)}}\right).\tag{5.39}$$

Since

$$-\frac{\inf_{t\in I_k^T(u)}\frac{u^2(1+t)^2}{\sigma^2(\mu(ut))}}{1-\frac{1}{4}\underline{\sigma}_{k,\ell}^2(u)}\le-\frac{1}{4}\inf_{t\in I_k^T(u)}\frac{u^2(1+t)^2}{\sigma^2(\mu(ut))}\underline{\sigma}_{k,\ell}^2(u)-\inf_{t\in I_k^T(u)}\frac{u^2(1+t)^2}{\sigma^2(\mu(ut))},$$

the summand in (5.39) is bounded from above by

$$\exp\left(-\frac{1}{8}\inf_{t\in I_k^T(u)}\frac{u^2(1+t)^2}{\sigma^2(\mu(ut))}\underline{\sigma}_{k,\ell}^2(u)\right)\Psi\left(\inf_{t\in I_k^T(u)}\frac{u(1+t)}{\sigma(\mu(ut))}\right)(1+o(1)),$$

where the $o(1)$ term is uniformly in k, ℓ as a consequence of the second claim of Lemma 5.17, cf. Equation (3.2). By the first claim of Lemma 5.17 for $\epsilon = 1/2$, say, and the UCT, there exist constants $\mathcal{K}'', \mathcal{K}''', \zeta$ such that

$$\begin{aligned}
&\sum_{\substack{-\frac{\delta(u)}{T\Delta(u)} \le \ell \le \frac{\delta(u)}{T\Delta(u)} \\ |k-\ell|>1}} \exp\left(-\frac{1}{8} \inf_{t \in I_k^T(u)} \frac{u^2(1+t)^2}{\sigma^2(\mu(ut))} \underline{\sigma}^2_{k,\ell}(u)\right) \\
&\quad \le \sum_{\substack{-\frac{\delta(u)}{T\Delta(u)} \le \ell \le \frac{\delta(u)}{T\Delta(u)} \\ |k-\ell|>1}} \exp\left(-\mathcal{K}'' \left[T^\zeta (|k-\ell|-1)^\zeta - 2^{\zeta-1}\right]\right) \\
&\quad \le 2e^{\mathcal{K}'' 2^{\zeta-1}} \sum_{j=1}^{\infty} \exp\left(-\mathcal{K}'' T^\zeta j^\zeta\right) \\
&\quad \le \mathcal{K}''' \exp\left(-T^\zeta\right).
\end{aligned}$$

Therefore, (5.39) cannot be not larger than

$$\begin{aligned}
&2\mathcal{K}'\mathcal{K}'''T^\alpha \exp\left(-T^\zeta\right) \sum_{-\frac{\delta(u)}{T\Delta(u)} \le k \le \frac{\delta(u)}{T\Delta(u)}} \Psi\left(\inf_{t \in I_k^T(u)} \frac{u(1+t)}{\sigma(\mu(ut))}\right)(1+o(1)) \\
&\quad = 2\sqrt{\frac{2\pi}{\mathcal{C}}}\mathcal{H}(T)\mathcal{K}'\mathcal{K}'''T^\alpha \exp\left(-T^\zeta\right) \frac{\sigma(\mu(u))}{\Delta(u)} \Psi\left(\frac{u(1+t_u^*)}{\sigma(\mu(ut_u^*))}\right)(1+o(1)),
\end{aligned}$$

where the last equality was shown in the proof of Proposition 5.15. Now first send $u \to \infty$, and then $T \to \infty$ to see that $I(u,T)$ plays no role in the asymptotics. One can also see that $II(u,T)$ and $III(u,T)$ can be neglected, but one needs suitable analogues of Lemma 5.17 and Lemma 5.18 to see this. Except that there is no summation over ℓ, the arguments are exactly the same as for $I(u,T)$. Since it is notationally more involved, we leave this to the reader. □

CHAPTER 6

Reduced-load equivalence

In this chapter, we consider a queue fed by a number of independent heterogeneous Gaussian sources. We study under which conditions a reduced-load equivalence holds, i.e., when a subset of the sources becomes asymptotically dominant as the buffer size increases. For this, the results on extremes of Gaussian processes of Chapter 5 are combined with de Haan theory. We explain how the results of this chapter relate to square-root insensitivity and moderately heavy tails.

6.1 Introduction

Consider a fluid queue fed by the superposition of M independent stationary Gaussian sources with mean input rate $\mu > 0$. As seen in Section 1.1.3, if the buffer is drained at some constant rate $r > \mu$, the steady-state probability that the buffer content exceeds u is given by

$$\mathbb{P}\left(\sup_{t\geq 0}\sum_{i=1}^{M} Y_i(t) - (r-\mu)t > u\right), \tag{6.1}$$

where Y_i is a centered separable Gaussian process with stationary increments, and the Y_i are independent. In Chapter 5, the asymptotics of this probability for $u \to \infty$ have been found under some regularity conditions.

As explained in Section 1.1.5, the probability (6.1) plays not only a role in queueing theory. It is well-known that it can be interpreted alternatively as the ruin probability of an insurance company with total premium rate r, initial capital u, and cumulative claim process $\sum_i Y_i(t) + \mu t$. The case $M > 1$ has recently attracted attention in the insurance literature under the name 'perturbed risk models'. We mention in particular Huzak *et al.* [166], who study the influence of a perturbation with stationary independent increments on the classical Cramér-Lundberg process. We study this model in Section 12.4; see also [285, 310] for further examples and references.

In this chapter, we consider a slightly more general setting than (6.1): we replace the 'drift' $(r-\mu)t$ by ct^β for some $\beta, c > 0$. We study the question in which cases only a subset of the M input sources contribute to a high value of $\sup_{t\geq 0}\sum_i Y_i(t) - ct^\beta$, i.e., when for $u \to \infty$,

$$\mathbb{P}\left(\sup_{t\geq 0}\sum_{i=1}^{M} Y_i(t) - ct^\beta > u\right) \sim \mathbb{P}\left(\sup_{t\geq 0}\sum_{i\in S} Y_i(t) - ct^\beta > u\right), \tag{6.2}$$

where $S \subseteq \{1, \ldots, M\}$. We present necessary and sufficient conditions for (6.2), and we say that we have *reduced-load equivalence* when (6.2) holds for some $S \neq \{1, \ldots, M\}$.

To explain the term 'reduced-load equivalence', consider two Gaussian sources with mean input rates μ_1 and μ_2, and suppose the buffer is drained at rate $r > \mu_1 + \mu_2$. In the interval $[0, t]$, the input to the system is $Y_1(t) + \mu_1 t + Y_2(t) + \mu_2 t$ and the output is rt. Hence, if (6.2) holds for $c = r - \mu_1 - \mu_2$, $\beta = 1$, and $S = \{1\}$, the system behaves asymptotically as if only the first source is present and the buffer is drained at rate $r - \mu_2$. It is easy to see that the new load $\mu_1/(r - \mu_2)$ is smaller than the old load $(\mu_1 + \mu_2)/r$. The same reasoning applies if one has more than two sources.

Our main assumptions are the following. First, the variance function σ^2 of $\sum_i Y_i$ is supposed to be regularly varying at infinity with index $2H \in (0, 2)$; see Section 2.1. Moreover, we must impose that $\beta \in (H, 2H)$; otherwise (6.2) cannot hold, since the key ingredient from Chapter 5, Proposition 6.1 below, cannot be applied. As before, we write UCT for uniform convergence theorem, which refers to either Theorem 2.3 or Corollary 2.4.

The motivation for pursuing our analysis stems from recent results on certain non-Gaussian queueing models [5, 172, 175]. An interesting special case of the results in [175] is $M = 2$ and Y_1, Y_2 being ON/OFF sources with exponentially distributed ON and OFF periods, except for Y_2; the latter process is supposed to have activity periods with a general distribution G. It is then known that if the tail $P(G > x)$ of G is heavier than $\exp(-x^{-1/2})$, the buffer content behaves asymptotically as if the second source produces traffic at a constant rate equal to its load. This does not hold if the tail of G is lighter than $\exp(-x^{-1/2})$. In a slightly different setting, Zwart *et al.* [315] examine various scenarios that may then occur.

In our opinion, it is instructive to study the Gaussian case in detail for two reasons. First, there is a vast body of literature on Gaussian processes, which makes this case relatively easy to study; in the present context, we rely on the results given in Chapter 5. Moreover, one can use the Gaussian case to 'predict' reduced-load-type behavior in other models. For instance, the aforementioned critical exponent 1/2 also arises in the Gaussian framework, and a simple formula shows how this quantity changes if the underlying source characteristics are altered, cf. Theorem 6.2. We remark that this phenomenon plays no role when studying logarithmic asymptotics, as seen in Chapter 4.

Our results are closely related to those of Zwart *et al.* [314]. While we present a condition that is both necessary and sufficient, it is not clear that the sufficient condition for reduced-load equivalence of [314] coincides with their necessary condition. Another difference is that only the superposition of two Gaussian sources ($M = 2$) is studied in [314]. In conclusion, we prove a condition that is at least as good as the condition of [314], but our condition is extremely simple and easy to check.

Another related paper is Hüsler and Schmid [165]. They establish exact asymptotics of (6.1) for $M = 2$. The difference with the present chapter is that we assume stationary increments and allow general variance functions, while Hüsler and Schmid restrict the variance function to be of the form $\sum_i \alpha_i t^{2H_i}$ but do not require stationarity of the increments. Another, more fundamental, difference is the type of question we pose. Exact asymptotics for (6.1) have already been established in Chapter 5 in the present setting, but here we focus on the reduced-load equivalence (6.2).

The results in this chapter are readily adapted if ct^β is replaced by a regularly varying function ϕ; the only reason for considering ct^β is to avoid cumbersome notation, cf. Assumptions **M1**–**M4** in Chapter 5. The case $\beta \neq 1$ may also be relevant in a queueing context; see [194].

Some words for the organization of this chapter. Section 6.2 describes the results, and Section 6.3 gives some examples. The required proofs, which rely extensively on de Haan theory (see Section 2.1), can be found in Section 6.4.

6.2 Description of the result

We first introduce some notation. Let $Y_1, \dots, Y_M$ be independent centered Gaussian processes with stationary increments. As indicated in Chapter 3, the finite-dimensional distributions are then completely determined by their respective variance functions $\sigma_1^2, \dots, \sigma_M^2$: for $s, t \geq 0$,

$$\mathbb{C}\mathrm{ov}(Y_i(s), Y_i(t)) = \frac{1}{2}\left[\sigma_i^2(s) + \sigma_i^2(t) - \sigma_i^2(|t-s|)\right].$$

The sum of these processes is denoted by Y, so that Y has stationary increments and variance function $\sigma^2 = \sum_{i=1}^M \sigma_i^2$.

Given a subset S of $\{1, \dots, M\}$, we set

$$\sigma_S^2(t) := \sum_{i \in S} \sigma_i^2(t),$$

and $S^c := \{1, \dots, M\} \backslash S$.

Now we formulate the assumptions on the variance functions.

S1 For $i = 1, \dots, M$, σ_i^2 is continuous and regularly varying at infinity with index $2H_i$ for some $H_i \in (0, 1)$,

S2 for $i = 1, \dots, M$, σ_i^2 is ultimately continuously differentiable and its first derivative $\dot{\sigma}_i^2$ is ultimately monotone,

S3 for $i = 1, \dots, M$, σ_i^2 is ultimately twice continuously differentiable and its second derivative $\ddot{\sigma}_i^2$ is ultimately monotone,

S4 $\sigma(t) \leq Ct^\gamma$ on a neighborhood of zero for some $C, \gamma > 0$.

Note that **S1** implies that the variance function σ^2 of the sum Y is regularly varying with index $2H$, with $H := \max_{i=1}^M H_i$. Define

$$S_* := \{i \in \{1, \dots, M\} : \sigma_i^2 \text{ is regularly varying with index } 2H\},$$

as the family of indices with maximum index of variation.

We start with a simple consequence of Proposition 5.2. Note that for $\beta = 1$ the condition $\beta \in (H, 2H)$ is equivalent to $H \in (1/2, 1)$.

Proposition 6.1 *Let the σ_i^2 and σ^2 satisfy* **S1**–**S4**. *For $H \in (0, 1)$ and $\beta \in (H, 2H)$, we have for any $S \supseteq S_*$,*

$$\frac{\mathbb{P}\left(\sup_{t \geq 0} Y(t) - ct^\beta > u\right)}{\sup_{t \geq 0} \mathbb{P}\left(Y(t) - ct^\beta > u\right)} \sim \frac{\mathbb{P}\left(\sup_{t \geq 0} \sum_{i \in S} Y_i(t) - ct^\beta > u\right)}{\sup_{t \geq 0} \mathbb{P}\left(\sum_{i \in S} Y_i(t) - ct^\beta > u\right)}.$$

Proposition 6.1 shows that (6.2) holds if and only if

$$\sup_{t \geq 0} \mathbb{P}\left(\sum_{i=1}^M Y_i(t) - ct^\beta > u\right) \sim \sup_{t \geq 0} \mathbb{P}\left(\sum_{i \in S} Y_i(t) - ct^\beta > u\right). \tag{6.3}$$

The following theorem, which is the main result of this chapter, gives a simple necessary and sufficient condition for this to hold. We emphasize that the theorem gives a statement for *any* set $S \supseteq S_*$, but that one usually tries to find the *smallest* S for which a reduced-load equivalence holds. The theorem is proven in Section 6.4.

Theorem 6.2 *Let $S \supseteq S_*$. Under the conditions of Proposition 6.1, the reduced-load equivalence (6.2) holds if and only if*

$$\lim_{u\to\infty} \frac{u\sigma_{S^c}(u^{1/\beta})}{\sigma_S^2(u^{1/\beta})} = 0. \tag{6.4}$$

To intuitively understand this result, it is important to gain insight into the optimizers on both sides of (6.3). As for the left-hand side, we have

$$\arg\sup_{t\geq 0} \mathbb{P}\left(\sum_{i=1}^{M} Y_i(t) - ct^\beta > u\right) = \arg\inf_{t\geq 0} \frac{u+ct^\beta}{\sigma(t)} = \left(u \arg\inf_{t\geq 0} \frac{\sigma(u^{1/\beta})(1+ct)}{\sigma([ut]^{1/\beta})}\right)^{1/\beta},$$

and by the regular variation of σ, it is plausible that the optimizer in the latter expression tends to $t^* := \arg\inf_{t\geq 0}(1+ct)/t^{H/\beta}$. The same reasoning applies to the right-hand side of (6.3). Hence, suppose that both suprema in (6.3) be attained for $t_u^* = (ut^*)^{1/\beta}$, where $t^* > 0$ is fixed. We then have a reduced-load equivalence if and only if

$$\Psi\left(\frac{u(1+ct^*)}{\sigma((ut^*)^{1/\beta})}\right) \sim \Psi\left(\frac{u(1+ct^*)}{\sigma_S((ut^*)^{1/\beta})}\right),$$

where Ψ denotes the complementary distribution function of the standard normal distribution, see (3.1). Standard estimates on Ψ show that this is equivalent to

$$\lim_{u\to\infty}\left[\frac{u^2(1+ct^*)^2}{\sigma_S^2((ut^*)^{1/\beta})} - \frac{u^2(1+ct^*)^2}{\sigma^2((ut^*)^{1/\beta})}\right] = \lim_{u\to\infty} \frac{u^2(1+ct^*)^2}{\sigma_S^2((ut^*)^{1/\beta})} \frac{\sigma_{S^c}^2((ut^*)^{1/\beta})}{\sigma^2((ut^*)^{1/\beta})} = 0,$$

which is readily seen to hold if and only if (6.4) holds.

6.3 Examples

In this section, we present two examples to illustrate Theorem 6.2. While the first deals with two sources, the second shows that one may need a strictly larger set than S_* for a reduced-load equivalence to hold. In both examples we set $\beta = 1$.

6.3.1 An example with $M = 2$

The first example is related to Corollary 3.1 of [314]. Consider the sum of two processes, i.e., $M = 2$. Suppose that the variance functions satisfy $\sigma_1^2(u) \sim C_1 u^{2H_1}$ and $\sigma_2^2(u) \sim C_2 u^{2H_2}$ for some constants $C_1, C_2 > 0$ and $H_1 > H_2$. Also suppose that $H_1 > 1/2$. For instance, the first process is a fractional Brownian motion with long-range dependent characteristics, and the second process a standard Brownian motion or a short-range dependent Gaussian integrated process (see Sections 1.3.2 and 5.3.3). Theorem 6.2 implies that a reduced-load equivalence holds if and only if $2H_1 > 1 + H_2$.

Of special interest is the case $H_2 = 1/2$, in which the condition reduces to $H_1 > 3/4$. Since the tail of the probability on the right-hand side of (6.2) can be written as $\exp(-\ell(u)u^{2-2H_1})$ for some slowly varying function ℓ, a reduced-load equivalence then holds if and only if the exponent in this expression is smaller than $1/2$. This is the connection with the discussion in Section 6.1 on a model with ON/OFF sources. It illustrates that *square-root insensitivity* or *moderately heavy tails* play an important role in situations with both short-range dependent processes on the one hand and long-range dependent processes or subexponential variables on the other hand; see, e.g., [5, 25, 56, 135, 174, 173, 175].

It is interesting to note that the condition $H_1 > 3/4$ also plays a role in a different problem related to the superposition of two independent processes. Indeed, set $\sigma_1^2(u) := u^{2H_1}$ and $\sigma_2^2(u) := u$ for some $H_1 > 1/2$; the distribution in $C[0,T]$ of $Y_1 + Y_2$ is *absolutely continuous* to the distribution of Y_1 (and vice versa) if and only if $H_1 > 3/4$; otherwise they are singular. These assertions are due to Cheridito [69]. From recent results of van Zanten [302], we conclude that the appearance of 3/4 in both problems is not caused by some underlying principle.

6.3.2 The set S_* is not always the dominant set

We now illustrate the fact that one may actually need a larger set than S_* for a reduced-load equivalence to hold. For simplicity, set $M = 3$ and suppose that Y_1 is a fractional Brownian motion B_{H_1} with Hurst parameter H_1, Y_2 a fractional Brownian motion B_{H_2} with Hurst parameter H_2, and Y_3 a standard Brownian motion B (all mutually independent). Let H_1 and H_2 satisfy

$$\frac{3}{4} < H_2 < H_1 < \frac{H_2 + 1}{2}.$$

It is easy to verify that Theorem 6.2 implies that the following does *not* hold:

$$\mathbb{P}\left(\sup_{t\geq 0} B_{H_1}(t) + B_{H_2}(t) + B(t) - t > u\right) \sim \mathbb{P}\left(\sup_{t\geq 0} B_{H_1}(t) - t > u\right).$$

Hence, B_{H_1} alone does not determine the asymptotics, although $S_* = \{1\}$. However, Theorem 6.2 also shows that the following *does* hold:

$$\mathbb{P}\left(\sup_{t\geq 0} B_{H_1}(t) + B_{H_2}(t) + B(t) - t > u\right) \sim \mathbb{P}\left(\sup_{t\geq 0} B_{H_1}(t) + B_{H_2}(t) - t > u\right),$$

i.e., the first two processes are asymptotically dominant.

6.4 Proof of Theorem 6.2

The proof consists of three steps. We first present an 'intermediate' necessary and sufficient condition, which is not so explicit. In the second step, this condition is shown to be necessary for (6.4), and the last step shows sufficiency.

For notational convenience, we set $\mu(t) := t^{1/\beta}/c$. Moreover, we let t^* denote the argument of the infimum of $(1+t)^2/t^{2H/\beta}$ over $\mathbb{R}_+$, i.e., $t^* = H/(\beta - H)$.

Step 1: auxiliary necessary and sufficient condition

Proposition 6.1 implies that (6.2) holds if and only if

$$\lim_{u\to\infty}\left[\inf_{t\geq 0}\frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))} - \inf_{t\geq 0}\frac{u^2(1+t)^2}{\sigma^2(\mu(ut))}\right] \leq 0. \tag{6.5}$$

Let $\epsilon > 0$ be small. In the first step of the proof, we show that the condition in the preceding display is equivalent to

$$\lim_{u\to\infty}\left[\inf_{t\in[t^*\pm\epsilon]}\frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))} - \inf_{t\in[t^*\pm\epsilon]}\frac{u^2(1+t)^2}{\sigma^2(\mu(ut))}\right] \leq 0, \tag{6.6}$$

where we denoted the interval $[t^* - \epsilon, t^* + \epsilon]$ by $[t^* \pm \epsilon]$. We only show this for the second infimum; a similar reasoning applies to the first.

Select some (large) $T > 0$ such that

$$\frac{1}{4}T^{1-H/\beta} \geq \frac{1+t^*}{(t^*)^{H/\beta}}.$$

We start by showing that the infima in (6.5) can be taken over the interval $[0, T]$. To see this, note that for large u,

$$\inf_{t\geq T} \frac{1+t}{\sigma(\mu(ut))} \geq \inf_{t\geq T} \frac{t}{\sigma(\mu(ut))} \geq \frac{1}{2}\frac{T^{1-H/\beta}}{\sigma(\mu(u))} \geq 2\frac{1+t^*}{\sigma(\mu(u))(t^*)^{H/\beta}} \geq \frac{1+t^*}{\sigma(\mu(ut^*))},$$

where the second inequality is a consequence of the UCT of Corollary 2.4 (indeed, the mapping $t \mapsto t/\sigma(\mu(t))$ is locally bounded on $[1, \infty)$ and regularly varying with index $1 - H/\beta > 0$). The last inequality follows from the definition of regular variation.

Now we show that infima cannot be attained on $[0, T]\backslash[t^* \pm \epsilon]$ for large u. Choose some $\eta > 0$ such that

$$(1+\eta)\left[\sup_{t\in[0,T]\backslash[t^*\pm\epsilon]} \frac{t^{H/\beta}}{1+t} + \eta\right] \leq \frac{(t^*)^{H/\beta}}{1+t^*},$$

which is possible since $\epsilon > 0$. Again exploiting the local boundedness of σ^2, by the UCT, we have for large u,

$$\begin{aligned}
\inf_{t\in[0,T]\backslash[t^*\pm\epsilon]} \frac{1+t}{\sigma(\mu(ut))} &\geq \inf_{t\in[0,T]\backslash[t^*\pm\epsilon]} \frac{1+t}{\sigma(\mu(u))[t^{H/\beta}+\eta]} \\
&\geq \frac{1}{\sigma(\mu(u))\left[\sup_{t\in[0,T]\backslash[t^*\pm\epsilon]} \frac{t^{H/\beta}}{1+t} + \eta\right]} \\
&\geq (1+\eta)\frac{1+t^*}{\sigma(\mu(u))(t^*)^{H/\beta}},
\end{aligned}$$

which majorizes $\inf_{t\in[t^*\pm\epsilon]}(1+t)/\sigma(\mu(ut))$ again by the UCT.

Step 2: (6.4) implies (6.6)

Before proving that (6.6) is equivalent to (6.4) by combining Step 2 and 3, we first make an observation concerning the second-order behavior of the function $\sigma_S^2(\mu(\cdot))/\sigma^2(\mu(\cdot))$.

Observe that $\sigma^2 = \sigma_{S^c}^2 + \sigma_S^2$ and that $\sigma_{S^c}^2(u)/\sigma_S^2(u) \to 0$ since $S \supseteq S_*$, so that the function $\sigma_S^2(\mu(\cdot))/\sigma^2(\mu(\cdot))$ lies in the de Haan class of $\sigma_{S^c}^2(\mu(\cdot))/\sigma_S^2(\mu(\cdot))$ with index $2(H - \max_{i\in S^c} H_i)/\beta$, i.e., for $t > 0$,

$$\frac{\sigma_S^2(\mu(ut))}{\sigma^2(\mu(ut))} - \frac{\sigma_S^2(\mu(u))}{\sigma^2(\mu(u))} \sim \left(1 - t^{2(\max_{i\in S^c} H_i - H)/\beta}\right)\frac{\sigma_{S^c}^2(\mu(u))}{\sigma_S^2(\mu(u))}. \tag{6.7}$$

We refer to Section 2.1 for more details.

Now we prove the sufficiency of (6.4). By (6.7), we have

$$\begin{aligned}
&\inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma^2(\mu(ut))} \\
&= \inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))}\left[\frac{\sigma_S^2(\mu(u))}{\sigma^2(\mu(u))} + (1+o(1))\left(1 - t^{2(\max_{i\in S^c} H_i - H)/\beta}\right)\frac{\sigma_{S^c}^2(\mu(u))}{\sigma_S^2(\mu(u))}\right] \qquad (6.8)\\
&\geq \frac{\sigma_S^2(\mu(u))}{\sigma^2(\mu(u))} \inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))} \\
&+ \frac{\sigma_{S^c}^2(\mu(u))}{\sigma_S^2(\mu(u))} \inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))}\left(1 - t^{2(\max_{i\in S^c} H_i - H)/\beta}\right)(1+o(1)) \\
&=: I(u) + II(u),
\end{aligned}$$

and $o(1)$ is uniform in $t \in [t^* \pm \epsilon]$ by the UCT for de Haan functions, see Theorem 2.8. Observe that

$$\begin{aligned}\inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))} - I(u) &= \left(1 - \frac{\sigma_S^2(\mu(u))}{\sigma^2(\mu(u))}\right) \inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))} \\ &= \frac{\sigma_{S^c}^2(\mu(u))}{\sigma^2(\mu(u))} \inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))} \\ &= \frac{u^2\sigma_{S^c}^2(\mu(u))}{\sigma_S^4(\mu(u))} \frac{\sigma_S^2(\mu(u))}{\sigma^2(\mu(u))} \inf_{t\in[t^*\pm\epsilon]} \frac{\sigma_S^2(\mu(u))(1+t)^2}{\sigma_S^2(\mu(ut))}.\end{aligned}$$

The infimum tends to a constant in $(0, \infty)$ by the UCT, and the other terms tend to zero as a consequence of (6.4). Step 2 is complete once it has been shown that $II(u)$ tends to zero, or equivalently that

$$\lim_{u\to\infty} \frac{\sigma_{S^c}^2(\mu(u))}{\sigma_S^2(\mu(u))} \inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))} \left(1 - t^{2(\max_{i\in S^c} H_i - H)/\beta}\right) = 0.$$

For this, note that the left-hand side equals

$$\begin{aligned}&\lim_{u\to\infty} \frac{u^2\sigma_{S^c}^2(\mu(u))}{\sigma_S^4(\mu(u))} \inf_{t\in[t^*\pm\epsilon]} \frac{\sigma_S^2(\mu(u))(1+t)^2}{\sigma_S^2(\mu(ut))} \left(1 - t^{2(\max_{i\in S^c} H_i - H)/\beta}\right) \\ &= \lim_{u\to\infty} \frac{u^2\sigma_{S^c}^2(\mu(u))}{\sigma_S^4(\mu(u))} \inf_{t\in[t^*\pm\epsilon]} \frac{(1+t)^2}{t^{2H/\beta}} \left(1 - t^{2(\max_{i\in S^c} H_i - H)/\beta}\right),\end{aligned}$$

by the UCT and the fact that $1 - t^{2(\max_{i\in S^c} H_i - H)/\beta}$ is bounded away from $\pm\infty$ on $[t^* \pm \epsilon]$. Evidently, (6.4) implies that the limits in the preceding display are equal to zero.

Step 3: (6.6) implies (6.4)

Now suppose that (6.4) does not hold. Observe that

$$1 - t^{2(\max_{i\in S^c} H_i - H)/\beta} \le 1 - \frac{1}{2}(t^*)^{2(\max_{i\in S^c} H_i - H)/\beta} =: \alpha < 1$$

for $t \in [t^* \pm \epsilon]$ if $\epsilon > 0$ is small enough. Hence, for large u,

$$\inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma^2(\mu(ut))} \le \inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))} \left[\frac{\sigma_S^2(\mu(u))}{\sigma^2(\mu(u))} + (\alpha + o(1))\frac{\sigma_{S^c}^2(\mu(u))}{\sigma_S^2(\mu(u))}\right],$$

so that by (6.8), for large u,

$$\begin{aligned}&\inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma_S^2(\mu(ut))} - \inf_{t\in[t^*\pm\epsilon]} \frac{u^2(1+t)^2}{\sigma^2(\mu(ut))} \\ &\ge \frac{u^2\sigma_{S^c}^2(\mu(u))}{\sigma_S^4(\mu(u))} \left[\frac{\sigma_S^2(\mu(u))}{\sigma^2(\mu(u))} - \alpha + o(1)\right] \inf_{t\in[t^*\pm\epsilon]} \frac{\sigma_S^2(\mu(u))(1+t)^2}{\sigma_S^2(\mu(ut))} \\ &\ge \frac{1-\alpha}{2} \frac{u^2\sigma_{S^c}^2(\mu(u))}{\sigma_S^4(\mu(u))} \inf_{t\in[t^*\pm\epsilon]} \frac{(1+t)^2}{t^{2H/\beta}},\end{aligned}$$

which does not converge to zero.

Part B

Simulation

CHAPTER 7

Background on simulation

In Part B of the thesis, we are interested in the simulation of probabilities that arise from the large-deviation principle (LDP). Direct Monte Carlo simulation is unsuitable in this context, since these probabilities are typically small. *Importance sampling* is a technique that can resolve these difficulties by simulating under an alternative measure, but the inherent problem is that an appropriate change of measure should be selected. To do so, we give a criterion that 'asymptotically' minimizes the number of simulation replications, subject to a fixed, predefined required accuracy (width of the confidence interval).

In this chapter, we introduce the basic notions that we use throughout Part B. Moreover, we work out examples that illustrate the interplay between large-deviation theory and rare-event simulation.

Suppose that the family $\{\nu_\epsilon\}$ satisfies the LDP with rate function I (see Section 2.3), and let some *rare event* $A \in \mathcal{B}$ be given, where $\inf_{x \in A^o} I(x) > 0$. We are interested in the simulation of $\nu(A) := \nu_{\epsilon_0}(A)$ for some small $\epsilon_0 > 0$. Direct Monte Carlo simulation of $\nu(A)$ is based on the proportion of samples for which the rare event A occurs under ν. The general rule is that, for an estimate with a fixed relative precision, the required number of simulation runs is inversely proportional to the probability to be estimated. Therefore, since $\nu_\epsilon(A)$ decays exponentially as $\epsilon \to 0$, this approach is unsuitable if ϵ_0 is small. Indeed, much time is then spent by sampling the uninteresting part of the sample space, i.e., the complement of A.

The idea of *importance sampling* is to sample from a different distribution, under which A occurs more frequently. An unbiased estimator is obtained by weighing the simulation output by likelihood ratios. However, it is usually nontrivial to select a good importance-sampling distribution, and the challenge is to find the distribution that is in some sense 'most efficient'. A widely accepted efficiency criterion for discriminating between importance-sampling distributions is *asymptotic efficiency*.

We start in Section 7.1 by reviewing importance sampling and asymptotic efficiency. By giving two random-walk examples in Section 7.2, we explain how large-deviation theory can help to select a good importance-sampling distribution.

7.1 Importance sampling and asymptotic efficiency

7.1.1 Importance sampling

Let $\mathcal{X}$ be a topological space, equipped with some σ-field $\mathcal{B}$ containing the Borel σ-field. Given a probability measure ν on $(\mathcal{X}, \mathcal{B})$, we are interested in the simulation of the ν-probability of a given event $A \in \mathcal{B}$, where $\nu(A)$ is small. The main idea of importance sampling is to speed up the simulation by working under a new probability measure λ on $(\mathcal{X}, \mathcal{B})$, for which A occurs more frequently. This is done by specifying a measurable function $d\lambda/d\nu : \mathcal{X} \to [0, \infty]$ and by setting

$$\lambda(B) := \int_B \frac{d\lambda}{d\nu} d\nu.$$

Since λ must be a probability measure, $d\lambda/d\nu$ should integrate to one with respect to ν.

In order to obtain an unbiased estimator of the desired rare event A, the samples (drawn from λ) are weighed by the *likelihood ratio* (or *Radon-Nikodym derivative*) as follows. Assuming the equivalence of the measures ν and λ, set $d\nu/d\lambda := (d\lambda/d\nu)^{-1}$ and note that

$$\nu(A) = \int_A \frac{d\nu}{d\lambda} d\lambda = \int_{\mathcal{X}} \mathbf{1}_A \frac{d\nu}{d\lambda} d\lambda,$$

where $\mathbf{1}_A$ is the indicator function of the event A. The importance-sampling estimator $\widehat{\nu_\lambda(A)}$ of $\nu(A)$ is found by drawing N independent samples $X^{(1)}, \ldots, X^{(N)}$ from λ; then

$$\widehat{\nu_\lambda(A)} := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{X^{(i)} \in A\}} \frac{d\nu}{d\lambda}(X^{(i)}).$$

It is clear that $\widehat{\nu_\lambda(A)}$ is an unbiased estimator, i.e., $\mathbb{E}_\lambda \widehat{\nu_\lambda(A)} = \nu(A)$. However, since one has the freedom to choose the importance-sampling distribution λ, it is of interest to select the distribution that minimizes the variance of the resulting estimator, or, equivalently,

$$\int_A \left(\frac{d\nu}{d\lambda}\right)^2 d\lambda = \int_{\mathcal{X}} \mathbf{1}_A \left(\frac{d\nu}{d\lambda}\right)^2 d\lambda = \int_A \frac{d\nu}{d\lambda} d\nu.$$

A zero-variance estimator is found by letting λ be equal to ν conditioned on the rare event A (see, e.g., Heidelberger [157]), but it is intractable since $d\nu/d\lambda$ then depends on the *unknown* probability $\nu(A)$. This motivates the use of another optimality criterion, *asymptotic efficiency*, which is discussed in the next subsection.

We end this subsection by noting that importance sampling is not the only way to achieve variance reduction (or to gain 'efficiency'). For surveys on other techniques, we refer to Glynn [149] and L'Ecuyer [212]. Interesting results have also been obtained with a splitting method (sometimes called RESTART); see Glasserman *et al.* [145].

7.1.2 Asymptotic efficiency

Asymptotic efficiency is related to the so-called *relative error*. Consider an i.i.d. sample $X_{\lambda_\epsilon}^{(1)}, \ldots, X_{\lambda_\epsilon}^{(N)}$ from an importance-sampling distribution λ_ϵ. We define the relative error $\eta_N(\lambda_\epsilon, A)$ of the importance-sampling estimator

$$\widehat{\nu_{\lambda_\epsilon}(A)}_N := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\left\{X_{\lambda_\epsilon}^{(i)} \in A\right\}} \frac{d\nu_\epsilon}{d\lambda_\epsilon}\left(X_{\lambda_\epsilon}^{(i)}\right) \tag{7.1}$$

as

$$\eta_N(\lambda_\epsilon, A) := \frac{\sqrt{\mathbb{V}\mathrm{ar}_{\lambda_\epsilon}\widehat{\nu_{\lambda_\epsilon}(A)}_N}}{\nu_\epsilon(A)} = \frac{1}{\sqrt{N}}\frac{\sqrt{\mathbb{V}\mathrm{ar}_{\lambda_\epsilon}\widehat{\nu_{\lambda_\epsilon}(A)}_1}}{\nu_\epsilon(A)}.$$

The relative error is proportional to the width of a confidence interval relative to the (expected) estimate itself; hence, it measures the variability of $\widehat{\nu_{\lambda_\epsilon}(A)}_N$.

For asymptotic efficiency, the number of samples required to obtain a prespecified relative error is not allowed to grow exponentially in $1/\epsilon$ as $\epsilon \to 0$. Set $N^*_{\lambda_\epsilon} := \inf\{N \in \mathbb{N} : \eta_N(\lambda_\epsilon, A) \leq \eta_{\max}\}$, for some given maximal relative error $0 < \eta_{\max} < \infty$.

Definition 7.1 *An importance-sampling family* $\{\lambda_\epsilon\}$ *is called* asymptotically efficient *if*

$$\lim_{\epsilon\to 0} \epsilon \log N^*_{\lambda_\epsilon} = 0.$$

When an importance-sampling family is asymptotically efficient, we also say that the estimator and the simulation method are asymptotically efficient. In the literature, asymptotic efficiency is sometimes referred to as *asymptotic optimality*, *logarithmic efficiency*, or *weak efficiency*. The following lemma provides an equivalent criterion, which is more convenient to work with. Note that it also shows that the specific value of $\eta_{\max}$ is irrelevant for asymptotic efficiency.

Lemma 7.2 *An importance-sampling family* $\{\lambda_\epsilon\}$ *is asymptotically efficient if and only if*

$$\lim_{\epsilon\to 0} \epsilon \log \frac{\int_A \left(\frac{d\nu_\epsilon}{d\lambda_\epsilon}\right)^2 d\lambda_\epsilon}{\nu_\epsilon(A)^2} = 0. \tag{7.2}$$

Proof. First note that by Jensen's inequality, we have

$$\int_A \left(\frac{d\nu_\epsilon}{d\lambda_\epsilon}\right)^2 d\lambda_\epsilon \geq \nu_\epsilon(A)^2,$$

meaning that (7.2) is equivalent to $\limsup_{\epsilon\to 0} \epsilon \log \frac{\int_A \left(\frac{d\nu_\epsilon}{d\lambda_\epsilon}\right)^2 d\lambda_\epsilon}{\nu_\epsilon(A)^2} \leq 0$. Using the definition of $\eta_N(\lambda_\epsilon, A)$, we obtain

$$N^*_{\lambda_\epsilon} = \left\lceil \frac{\mathbb{V}\mathrm{ar}_{\lambda_\epsilon}\widehat{\nu_{\lambda_\epsilon}(A)}_1}{\eta^2_{\max}\nu_\epsilon(A)^2} \right\rceil.$$

Since $N^*_{\lambda_\epsilon} \geq 1$, we have

$$\limsup_{\epsilon\to 0} \epsilon \log \frac{\int_A \left(\frac{d\nu_\epsilon}{d\lambda_\epsilon}\right)^2 d\lambda_\epsilon}{\nu_\epsilon(A)^2} = \limsup_{\epsilon\to 0} \epsilon \log \left(N^*_{\lambda_\epsilon} + \frac{1}{\eta^2_{\max}} \right) = \limsup_{\epsilon\to 0} \epsilon \log N^*_{\lambda_\epsilon},$$

establishing the claim. □

Let us now discuss another characterization of asymptotic optimality, which applies under a weak additional assumption. For this, note that

$$\limsup_{\epsilon\to 0} \epsilon \log \frac{\int_A \left(\frac{d\nu_\epsilon}{d\lambda_\epsilon}\right)^2 d\lambda_\epsilon}{\nu_\epsilon(A)^2} \leq \limsup_{\epsilon\to 0} \epsilon \log \int_A \left(\frac{d\nu_\epsilon}{d\lambda_\epsilon}\right)^2 d\lambda_\epsilon - 2\liminf_{\epsilon\to 0} \epsilon \log \nu_\epsilon(A), \tag{7.3}$$

with an equality if the limit $\lim_{\epsilon\to 0} \epsilon \log \nu_\epsilon(A)$ exists. It is sufficient for the existence of this limit that A be an I-continuity set, as defined in Section 2.3. In that case, we have

$\lim_{\epsilon\to 0} \epsilon \log \nu_\epsilon(A) = -\inf_{x\in A} I(x)$. In many applications, A is indeed an I-continuity set, so that asymptotic efficiency is equivalent to

$$\limsup_{\epsilon\to 0} \epsilon \log \int_A \left(\frac{d\nu_\epsilon}{d\lambda_\epsilon}\right)^2 d\lambda_\epsilon \le -2 \inf_{x\in \overline{A}} I(x). \tag{7.4}$$

In turn, this can be reformulated as $\lim_{\epsilon\to 0} \epsilon \log \int_A \left(\frac{d\nu_\epsilon}{d\lambda_\epsilon}\right)^2 d\lambda_\epsilon = -2\inf_{x\in\overline{A}} I(x)$ by Jensen's inequality. A closely related equivalent criterion is based on the *relative efficiency*

$$E_\epsilon := \frac{\log \int_A \left(\frac{d\nu_\epsilon}{d\lambda_\epsilon}\right)^2 d\lambda_\epsilon}{\log \nu_\epsilon(A)}. \tag{7.5}$$

Whenever A is an I-continuity set, asymptotic efficiency is equivalent to $\lim_{\epsilon\to 0} E_\epsilon = 2$.

Although the definition of asymptotic efficiency used here is mathematically convenient, several other criteria for discriminating between estimators have been proposed. Notably, the amount of time (or work) required to generate one simulation replication is not taken into account in our definition of asymptotic efficiency. Glynn and Whitt [151] give a definition into which this is incorporated; see also Glynn [149].

More details on importance sampling can be found in the fundamental paper by Glynn and Iglehart [150] and in the survey papers by Asmussen and Rubinstein [30], and Heidelberger [157].

It is not always possible to study a system analytically and obtain a good importance-sampling distribution as a result. In some situations, it is easier to find a good importance-sampling distribution numerically. The *cross-entropy method* calculates such a distribution iteratively by performing a simulation experiment in each iteration. We refer to Rubinstein and Kroese [277] for an extensive account.

7.2 Examples: random walks

In this section, we discuss two examples for which the large-deviation theory of Section 2.3 is a powerful guide for choosing an importance-sampling family. This illustrates the close connection between efficient simulation and large-deviation techniques. Both examples are related to random walks; an introduction to their large-deviation behavior has been given in Sections 2.3.1 and 2.3.2, and we use the same notation.

Let $\{X_n\}$ be a sequence of i.i.d. centered random variables taking values in $\mathbb{R}$, each with distribution $\mathbb{P}_X$. We set $S_0 := 0$ and, for $n \ge 1$, $S_n := \sum_{i=1}^n X_i$. The process $S := \{S_n\}$ is called a random walk; more details can be found in Chapter 9 and Section 11.1.

We define the running-maximum process $\{\overline{S}_n\}$ of $\{S_n\}$ as $\overline{S}_n := \max_{0\le k\le n} S_k$. The global maximum $\sup_{n\ge 0} S_n$ is denoted by $\overline{S}$; note that it is almost surely finite if $\mathbb{E}X_1 < 0$.

7.2.1 Cramér probabilities

Under the 'light-tail' assumption that $0 \in (\operatorname{dom}\Lambda_X)^o$, Cramér's LDP (Proposition 2.16) gives the (exponential) decay rate of the probability $\mathbb{P}(S_n \ge \gamma n)$ for some $\gamma > \mathbb{E}X_1$: if $\gamma \in (\operatorname{dom}\Lambda_X^*)^o$, then $[\gamma,\infty)$ is a Λ_X^*-continuity set, and the LDP yields

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n \ge \gamma n) = -\Lambda_X^*(\gamma).$$

We remark that the assumption $\gamma \in (\operatorname{dom}\Lambda_X^*)^o$ is not necessary for the existence of the above limit; see Corollary 2.2.19 of Dembo and Zeitouni [100].

However, it may occur that $\exp(-n\Lambda_X^*(\gamma))$ or Lemma 2.17 give unsatisfactory approximations of $\mathbb{P}(S_n \geq \gamma n)$, for instance because n is 'moderately large' (see, e.g., Ganesh *et al.* [140, Ch. 10]). In that case, one can resort to estimating this probability with importance sampling.

The observation that S_n/n satisfies the LDP suggests a useful importance-sampling distribution, which we now describe. For any *twist* $\xi \in \mathbb{R}$, we first define the *exponentially twisted* distribution as

$$\lambda_n^\xi(dx) := \exp(n\xi x - n\Lambda_X(\xi))\,\mathbb{P}(S_n/n \in dx).$$

That is, λ_n^ξ is the distribution of S_n^ξ/n, where S_n^ξ is a random walk with step-size distribution $\lambda_1^\xi(dx) = e^{\xi x - \Lambda_X(\xi)}\mathbb{P}_X(dx)$. Since $\gamma \in (\mathrm{dom}\,\Lambda_X^*)^o$, there exists a (unique) ξ^* such that $\Lambda_X^*(\gamma) = \xi^*\gamma - \Lambda_X(\xi^*)$. The next lemma shows the importance of the twist ξ^*; note that the mean of $S_n^{\xi^*}/n$ equals γ.

Lemma 7.3 *Let $\gamma \in (\mathrm{dom}\,\Lambda_X^*)^o$. The importance-sampling family $\{\lambda_n^{\xi^*}\}$ is asymptotically efficient.*

Proof. Since γ is a Λ_X^*-continuity set, we have

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n}\log \int_{[\gamma,\infty)} \exp(n\Lambda_X(\xi^*) - n\xi^* x)\,\mathbb{P}(S_n/n \in dx)\\ \leq \limsup_{n\to\infty} \frac{1}{n}\log \exp(n\Lambda_X(\xi^*) - n\xi^*\gamma)\,\mathbb{P}(S_n \geq n\gamma) = -2n\Lambda_X^*(\gamma),\end{aligned}$$

which is equivalent to asymptotic efficiency of $\{\lambda_n^{\xi^*}\}$, as observed in (7.4). □

For an in-depth study of stability issues related to the family $\{\lambda_n^{\xi^*}\}$, we refer to Sadowsky [278]. This lemma has been generalized into several directions: if $X_1, X_2, \ldots$ have a special Markovian structure, the result is established by Bucklew *et al.* [64]. For generally distributed S_n, Sadowsky and Bucklew [280] show that there exists an asymptotically efficient exponential twist if the so-called Gärtner-Ellis theorem applies to $\{S_n/n\}$; this is also observed by Szechtman and Glynn [292].

We now turn to a conditional limit theorem in the spirit of Section 4.2, which explains why the exponentially twisted distribution $\lambda_n^{\xi^*}$ is a good importance-sampling distribution. Recall that δ_y denotes the Dirac measure at y.

Lemma 7.4 *Let $\gamma \in (\mathrm{dom}\,\Lambda_X^*)^o$. The law $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ given $S_n \geq n\gamma$ converges weakly to $\lambda_1^{\xi^*}$ as $n \to \infty$.*

Proof. Similar to the proof of Theorem 4.2 in Section 4.5.2. It relies on Sanov's theorem, see Theorem 6.2.10 in Dembo and Zeitouni [100]. Sanov's theorem is an extension of Cramér's theorem (Proposition 2.16). A full proof does not fall within the scope of this chapter. □

Hence, intuitively, the step-size distribution is approximately $\lambda_1^{\xi^*}$ given that the rare event $\{S_n \geq n\gamma\}$ occurs. In other words, it is highly likely that the sample path $\{S_k/n : k \leq n\}$ is 'close' to $\{x^*(k/n) : k \leq n\}$, where the deterministic path x^* is given by $x^*(t) = \gamma t$ for $t \in [0,1]$. This is called the *most likely path*. As observed in Chapter 4, the most likely path is not necessarily a straight line.

Interestingly, if Cramér's condition $0 \in (\mathrm{dom}\,\Lambda_X)^o$ fails, the simulation of $\mathbb{P}(S_n \geq x)$ is significantly harder, even for fixed n and $X_1 \geq 0$. We emphasize that there is a wide class of distributions for which Cramér's condition does not hold, such as the subexponential distributions of Section 2.4. One of the earliest investigations of the problem in this heavy-tailed context is the paper by Asmussen *et al.* [21]. A technique somewhat akin to exponential twisting, *hazard rate twisting*, has also been investigated [54, 177]. It is still an active research area; see Asmussen and Kroese [26] for a recent contribution.

7.2.2 Level-crossing probabilities

In this subsection, we consider the simulation of the *level-crossing* probability $\mathbb{P}(\overline{S} \geq \gamma n)$ for $\gamma > 0$, where again $n \to \infty$. Suppose that $\mathbb{E}X_1 < 0$ and $\Lambda_X(\xi) < \infty$ for every $\xi \in \mathbb{R}$. Note that this requirement is stronger than Cramér's condition $0 \in (\mathrm{dom}\, \Lambda_X)^o$, which becomes clear by taking exponentially distributed step sizes. However, as argued in Section 2.3.2, this stronger assumption can often be replaced by Cramér's condition. To avoid trivialities, we also assume that $\mathbb{P}_X([0,\infty)) > 0$.

It can be seen that $\overline{S}/n$ satisfies the LDP in $\mathbb{R}_+$ with the good rate function $I(x) := \inf_{\tau>0} \tau \Lambda_X^*(x/\tau)$ and scale sequence $\{n\}$. This statement is a consequence of Mogul′skiĭ's theorem (Proposition 2.18), the contraction principle (Lemma 2.15), and Theorem 1 of Ganesh and O'Connell [139]; for more details we refer to [140, Ch. 6] and Section 9.2. An interesting alternative characterization of the rate function I is $I(x) = \xi^0 x$, where

$$\xi^0 := \sup\{\xi > 0 : \Lambda_X(\xi) \leq 0\},$$

see Rockafellar [273, Thm. 13.5] or Ganesh *et al.* [140, Lem. 1.7].

Motivated by the asymptotic efficiency of an exponentially twisted distribution in the previous subsection, we further assume that $\Lambda_X(\xi^0) = 0$, and consider a random walk $\{S_n^0\}$ with step-size distribution $\lambda_1^{\xi^0}$. Write $\tau^0(x) := \inf\{n : S_n^0 \geq x\}$. The convexity of Λ_X implies that this random walk has a positive drift, meaning that $\tau^0(\gamma n) < \infty$ almost surely. A natural importance-sampling estimator arises by simulating the random walk $\{S_n^0\}$ until the level γn is hit. This estimator has likelihood ratio

$$\exp\left(-\xi^0 S^0_{\tau^0(\gamma n)}\right). \tag{7.6}$$

Although it seems that this choice is not covered by the framework of Section 7.1.2, this problem can be circumvented by considering sample-path large deviations; Chapter 9 contains a detailed discussion of this issue.

Lemma 7.5 *Let $\Lambda_X(\xi^0) = 0$. The estimator with likelihood ratio (7.6) is asymptotically efficient.*

Proof. Observe that

$$\frac{1}{n} \log \mathbb{E} \exp\left(-2\xi^0 S^0_{\tau^0(\gamma n)}\right) \leq -2\xi^0 \gamma = -2I(\gamma),$$

and use again (7.4). □

This lemma has first been proven (with different techniques) by Lehtonen and Nyrhinen [216]; they also study the simulation of $\mathbb{P}(\overline{S}_n \geq \gamma n)$. Related results in a more general Markovian setting are obtained by Asmussen [14] and Lehtonen and Nyrhinen [215]. Asmussen and Rubinstein [30] have shown that ξ^0 is the only choice that yields asymptotic efficiency. Finally, for the simulation of multidimensional level-crossing probabilities, we refer to Collamore [80].

We next present an analogue of the conditional limit theorem in Lemma 7.4.

Lemma 7.6 *Let $\Lambda_X(\xi^0) = 0$. The law $\frac{1}{\tau(x)} \sum_{i=1}^{\tau(x)} \delta_{X_i}$ given $\tau(x) < \infty$ converges weakly to $\lambda_1^{\xi^0}$ as $x \to \infty$.*

Proof. This is a consequence of Proposition 17.9 in [30] (see also [13]). □

For further conditional limit theorems, we refer to Section IV.7 of Asmussen [18] or Section 3.6 of Robert [272].

Simulation of level-crossing probabilities in the context of heavy-tailed random variables has been much less investigated; we refer to Boots and Shahabuddin [54] and Asmussen *et al.* [27] for more details.

7.3 Outline of Part B

In each of the three chapters of Part B, we study the simulation of large-deviation probabilities. Chapter 8 contains general results for the simulation of these probabilities, and two specific problems are considered in Chapter 9 and 10.

We first describe the results in Chapter 8. Departing from the statement in Lemma 7.2, we give further necessary and sufficient conditions for asymptotic efficiency. These conditions are important, since it is not always possible to find an asymptotically efficient twist. The necessary and sufficient conditions are based on Varadhan's integral lemma (Lemma 2.14), and we therefore call them the *Varadhan conditions*. Interestingly, these elementary conditions are shown to improve upon the conditions established by Sadowsky [279].

The Varadhan conditions are used in Chapter 9 to further examine the level-crossing probabilities of Section 7.2.2. The probability $\mathbb{P}(S_k \geq e(k/n)n \text{ for some } k \leq n)$ is studied, where $e : [0,1] \to (0,\infty]$ is some 'smooth' function. Note that the choice $e(t) := \gamma$ leads to the probability $\mathbb{P}(\overline{S}_n \geq \gamma n)$, which is studied in [216]. We investigate two estimators for simulating the aforementioned probability with general e: the first relies on a change of measure on the sample-path level, and the second on the step level. For both estimators, we obtain necessary and sufficient conditions for asymptotic efficiency, and subsequently compare these conditions.

In Chapter 10, our aim is to simulate a Cramér-type probability as in Section 7.2.1. The step-size distribution $\mathbb{P}_X$ of the random walk is a multidimensional Gaussian distribution, with a possible strong correlation between the components. This situation arises in the study of a Gaussian fluid queue (see Part A) fed by a large number of Gaussian sources; for a discrete-time system, we estimate the steady-state probability that a buffer threshold is exceeded. Using the results of Chapter 8, we prove that, even if the cumulative input of the sources constitutes a (Gaussian) random walk, the natural exponentially twisted importance-sampling distribution does *not* yield asymptotic efficiency. Three alternative efficient methods are discussed and evaluated, and a detailed numerical comparison is made.

Chapters 8, 9, and 10 are based on [112], [113], and [111] respectively. All chapters in Part B are joint work with Michel Mandjes.

CHAPTER 8

Conditions for asymptotic efficiency

In this chapter, we study the simulation of $\nu_{\epsilon_0}(A)$ for some fixed event A and some $\epsilon_0 > 0$, under the assumption that $\{\nu_\epsilon : \epsilon > 0\}$ satisfies a large-deviation principle. We investigate the circumstances under which an exponentially twisted importance-sampling distribution yields an asymptotically efficient estimator. Varadhan's lemma yields necessary and sufficient conditions, and these are shown to improve upon the conditions of Sadowsky [279].

8.1 Introduction

In the previous chapter, we have argued that direct Monte Carlo methods are unsuitable for simulating probabilities that arise from a large-deviation principle. A family of probability measures $\{\nu_\epsilon : \epsilon > 0\}$ is said to satisfy a *large-deviation principle* (LDP) if $\nu_\epsilon(A)$ decays exponentially as $\epsilon \to 0$ for a wide class of sets A, cf. Definition 2.13. Given such a family, we refer to a probability of the form $\nu_{\epsilon_0}(A)$ for some $\epsilon_0 > 0$ and some event A as a *large-deviation probability*. Probabilities of this type are encountered in many fields, e.g., statistics, risk theory, operations research, information theory, and financial mathematics.

As explained in Section 7.1.1, a technique that is widely used to estimate rare-event probabilities is *importance sampling*. In importance sampling, one samples from a probability measure λ different from ν_{ϵ_0}, such that the ν_{ϵ_0}-rare event becomes λ-likely. Often, one chooses a so-called *exponentially twisted* distribution for λ, but within this class there is still freedom to select a specific twisted distribution. To evaluate the changes of measure, efficiency criteria have been developed. In this chapter, we use the *asymptotic efficiency* criterion.

Research initiated in the seminal paper of Siegmund [289] has shown that exponentially twisting is asymptotically efficient in specific cases. We mention, in particular, the Cramér and level-crossing probabilities of Section 7.2. In these two examples, large-deviation theory suggests a specific exponentially twisted importance-sampling distribution.

However, it has been noted that a successful application of an importance-sampling estimator based on large-deviation theory critically depends on the specific problem at hand. Glasserman and Wang [148] give variations on both the Cramér problem and the level-crossing problem, and show that exponential twists can be inefficient if the rare event A is irregular. In fact, they obtain the stronger result that the relative error (see Section 7.1.2) can even

become unbounded in these examples. Similar observations have been made by Glasserman and Kou [146] in a queueing context.

Given the examples of efficient and inefficient simulation with exponentially twisted importance-sampling distributions, it is natural to ask whether there exist necessary and sufficient conditions for asymptotic efficiency. This question is addressed here. In case the Gärtner-Ellis theorem applies, this question is studied by Sadowsky and Bucklew [280], while Sadowsky [279] extends their findings to a general abstract large-deviation setting.

The necessary and sufficient conditions presented in this chapter have two advantages over those in [279]. The first is that the proof is elementary; the conditions follow straightforwardly from an application of Varadhan's integral lemma. Therefore, we refer to these conditions as *Varadhan conditions*. Note that this elementary lemma has previously been applied to derive efficiency properties of certain rare event estimators, see Glasserman *et al.* [144], Dupuis and Wang [127], and Glasserman and Li [147].

A second advantage the Varadhan conditions is that they improve upon the conditions of Sadowsky [279]. This is the main result of the present chapter. To explain the improvements, it is important to realize that each set of *conditions* (the Varadhan conditions and Sadowsky's conditions) applies only under certain *assumptions*. The assumptions underlying the Varadhan conditions are less restrictive than those underlying Sadowsky's, meaning that the Varadhan conditions apply in more situations. Notably, convexity of the large-deviation rate function is not required. Furthermore, the Varadhan conditions themselves are 'better' than Sadowsky's conditions, i.e., the Varadhan sufficiency condition is implied by Sadowsky's sufficiency condition, and vice versa for the necessary condition.

The chapter is organized as follows. We state the Varadhan conditions in Section 8.2, and show in Section 8.3 that these conditions improve upon those of Sadowsky. Section 8.4 addresses the uniqueness of efficient twists. In Section 8.5, we illustrate the use of the Varadhan conditions with an example to which Sadowsky's results cannot be applied. Still, asymptotically efficient simulation is possible.

8.2 The Varadhan conditions for efficiency of exponential twisting

In this section, we investigate the asymptotic efficiency exponentially twisted importance-sampling families. After formalizing the imposed assumptions, we state necessary and sufficient conditions based on Varadhan's lemma (Lemma 2.14).

Let $\mathcal{X}$ be a topological space and $\mathcal{B}$ be a σ-field on $\mathcal{X}$ containing the Borel σ-field. We also assume that $\mathcal{X}$ is a vector space, but not necessarily a topological vector space. Throughout this section, we fix a rare event $A \in \mathcal{B}$ and a continuous linear functional $\xi : \mathcal{X} \to \mathbb{R}$. Having a topological vector space in mind, we write $\langle \xi, \cdot \rangle$ for $\xi(\cdot)$. We are given a family $\{\nu_\epsilon\}$ of probability measures on $(\mathcal{X}, \mathcal{B})$.

Assumption 8.1 (Varadhan assumptions) *Assume that*

(i) $\mathcal{X}$ *is a vector space endowed with some regular Hausdorff topology,*

(ii) $\{\nu_\epsilon\}$ *satisfies the LDP with a good rate function* I*, and*

(iii) it holds that

$$\lim_{M\to\infty} \limsup_{\epsilon\to 0} \epsilon \log \int_{\{x\in\mathcal{X}:\langle\xi,x\rangle\geq M\}} \exp[\langle \xi, x\rangle/\epsilon]\nu_\epsilon(dx) = -\infty, \tag{8.1}$$

and similarly with ξ *replaced by* $-\xi$*.*

A simple sufficient condition for (8.1) to hold is $\limsup_{\epsilon\to 0}\epsilon\log\int\exp[\gamma\langle\xi,x\rangle/\epsilon]\nu_\epsilon(dx)<\infty$ for some $\gamma>1$, see also (2.5). A similar condition can be given for $-\xi$.

A new family of probability measures $\{\lambda_\epsilon^\xi\}$ is defined as

$$\begin{aligned}\frac{d\lambda_\epsilon^\xi}{d\nu_\epsilon}(x) &:= \exp\left(\langle\xi,x\rangle/\epsilon-\log\int_{\mathcal{X}}\exp[\langle\xi,y\rangle/\epsilon]\nu_\epsilon(dy)\right) \\ &= \frac{\exp[\langle\xi,x\rangle/\epsilon]}{\int_{\mathcal{X}}\exp[\langle\xi,y\rangle/\epsilon]\nu_\epsilon(dy)}.\end{aligned} \tag{8.2}$$

The measures $\{\lambda_\epsilon^\xi\}$ are called *exponentially twisted with twist* ξ; cf. Section 7.2.1. If the family $\{\lambda_\epsilon^\xi\}$ is asymptotically efficient, we simply call the twist ξ asymptotically efficient.

The following proposition plays a key role in the proofs of this section.

Proposition 8.2 *Let $d\lambda_\epsilon^\xi/d\nu_\epsilon$ be given by (8.2), and let $B\in\mathcal{B}$. Under Assumption 8.1, we have*

$$\begin{aligned}\liminf_{\epsilon\to 0}\epsilon\log\int_B\left(\frac{d\nu_\epsilon}{d\lambda_\epsilon^\xi}\right)^2 d\lambda_\epsilon^\xi &\geq -\inf_{x\in\mathcal{X}}[I(x)-\langle\xi,x\rangle]-\inf_{x\in B^o}[I(x)+\langle\xi,x\rangle], \\ \limsup_{\epsilon\to 0}\epsilon\log\int_B\left(\frac{d\nu_\epsilon}{d\lambda_\epsilon^\xi}\right)^2 d\lambda_\epsilon^\xi &\leq -\inf_{x\in\mathcal{X}}[I(x)-\langle\xi,x\rangle]-\inf_{x\in\overline{B}}[I(x)+\langle\xi,x\rangle].\end{aligned}$$

Proof. Fix $B\in\mathcal{B}$ and note that

$$\begin{aligned}\epsilon\log\int_B\left(\frac{d\nu_\epsilon}{d\lambda_\epsilon^\xi}\right)^2 d\lambda_\epsilon^\xi &= \epsilon\log\int_B\frac{d\nu_\epsilon}{d\lambda_\epsilon^\xi}d\nu_\epsilon \\ &= \epsilon\log\int_{\mathcal{X}}\exp\left(\langle\xi,x\rangle/\epsilon\right)\nu_\epsilon(dx)+\epsilon\log\int_B\exp\left(-\langle\xi,x\rangle/\epsilon\right)\nu_\epsilon(dx).\end{aligned} \tag{8.3}$$

By Assumption 8.1 and the continuity of the functional ξ, Varadhan's integral lemma applies (see Lemma 2.14 and the remarks that follow it). Thus, the limit of the first term exists and equals

$$\lim_{\epsilon\to 0}\epsilon\log\int_{\mathcal{X}}\exp\left(\langle\xi,x\rangle/\epsilon\right)\nu_\epsilon(dx)=\sup_{x\in\mathcal{X}}[\langle\xi,x\rangle-I(x)].$$

A similar argument can be applied to the second term in (8.3). Indeed, we apply a variant of Varadhan's integral lemma (see, e.g., Exercise 4.3.11 of Dembo and Zeitouni [100]) to the continuous functional $-\xi$: for any open set G and any closed set F

$$\begin{aligned}\liminf_{\epsilon\to 0}\epsilon\log\int_G\exp\left(-\langle\xi,x\rangle/\epsilon\right)\nu_\epsilon(dx) &\geq -\inf_{x\in G}[I(x)+\langle\xi,x\rangle], \\ \limsup_{\epsilon\to 0}\epsilon\log\int_F\exp\left(-\langle\xi,x\rangle/\epsilon\right)\nu_\epsilon(dx) &\leq -\inf_{x\in F}[I(x)+\langle\xi,x\rangle].\end{aligned}$$

In particular, these inequalities hold for B^o and $\overline{B}$. The claim follows by letting $\epsilon\to 0$ in (8.3) (using the fact that the limit of the first term exists). □

The necessary and sufficient conditions, stated in the next theorem, follow almost immediately from Proposition 8.2.

Theorem 8.3 (Varadhan conditions) *Let Assumption 8.1 hold. The exponential twist ξ is asymptotically efficient if*

$$\inf_{x\in\mathcal{X}}[I(x)-\langle\xi,x\rangle]+\inf_{x\in\overline{A}}[I(x)+\langle\xi,x\rangle]\geq 2\inf_{x\in A^o}I(x). \tag{8.4}$$

Let Assumption 8.1 hold and let A be an I-continuity set. If the exponential twist ξ is asymptotically efficient, then

$$\inf_{x\in\mathcal{X}}[I(x)-\langle\xi,x\rangle]+\inf_{x\in A^o}[I(x)+\langle\xi,x\rangle]\geq 2\inf_{x\in\overline{A}}I(x). \tag{8.5}$$

Proof. Sufficiency follows from Lemma 7.2, the upper bound of Proposition 8.2, and the LDP of Assumption 8.1(ii):

$$\begin{aligned}\limsup_{\epsilon\to 0}\epsilon\frac{\int_A\left(\frac{d\nu_\epsilon}{d\lambda_\epsilon^\xi}\right)^2 d\lambda_\epsilon^\xi}{\nu_\epsilon(A)^2} &\leq \limsup_{\epsilon\to 0}\epsilon\log\int_A\left(\frac{d\nu_\epsilon}{d\lambda_\epsilon^\xi}\right)^2 d\lambda_\epsilon^\xi-2\liminf_{\epsilon\to 0}\epsilon\log\nu_\epsilon(A)\\ &\leq -\inf_{x\in\mathcal{X}}[I(x)-\langle\xi,x\rangle]-\inf_{x\in\overline{A}}[I(x)+\langle\xi,x\rangle]+2\inf_{x\in A^o}I(x).\end{aligned}$$

For necessity the argument is similar. First, the lower bound of Proposition 8.2 implies that

$$\limsup_{\epsilon\to 0}\epsilon\log\int_A\left(\frac{d\nu_\epsilon}{d\lambda_\epsilon^\xi}\right)^2 d\lambda_\epsilon^\xi\geq-\inf_{x\in\mathcal{X}}[I(x)-\langle\xi,x\rangle]-\inf_{x\in A^o}[I(x)+\langle\xi,x\rangle].$$

Furthermore, by the large-deviation upper bound,

$$\liminf_{\epsilon\to 0}\epsilon\log\nu_\epsilon(A)\leq-\inf_{x\in\overline{A}}I(x).$$

Combining these observations with the assumption that A is an I-continuity set, we have equality in (7.3), so that by Lemma 7.2,

$$\begin{aligned}0 &= \limsup_{\epsilon\to 0}\epsilon\log\int_A\left(\frac{d\nu_\epsilon}{d\lambda_\epsilon^\xi}\right)^2 d\lambda_\epsilon^\xi-2\liminf_{\epsilon\to 0}\epsilon\log\nu_\epsilon(A)\\ &\geq -\inf_{x\in\mathcal{X}}[I(x)-\langle\xi,x\rangle]-\inf_{x\in A^o}[I(x)+\langle\xi,x\rangle]+2\inf_{x\in\overline{A}}I(x),\end{aligned}$$

as desired. □

As suggested by the form of Theorem 8.3, the sufficient condition is also necessary under a weak condition on the set A. We formalize this in the following corollary, which follows straightforwardly from Theorem 8.3.

Corollary 8.4 *Let Assumption 8.1 hold, and assume that A is both an I-continuity set and an $(I+\xi)$-continuity set. Exponentially twisting with ξ is asymptotically efficient if and only if*

$$\inf_{x\in\mathcal{X}}[I(x)-\langle\xi,x\rangle]+\inf_{x\in\overline{A}}[I(x)+\langle\xi,x\rangle]=2\inf_{x\in\overline{A}}I(x).$$

We remark that Sadowsky [279] uses a more general notion than asymptotic efficiency, namely ν-efficiency. Given an I-continuity set A, the importance-sampling distribution λ_ϵ^ξ is said to be ν-efficient if

$$\limsup_{\epsilon\to 0}\epsilon\log\int_A\left(\frac{d\nu_\epsilon}{d\lambda_\epsilon^\xi}\right)^\nu d\lambda_\epsilon\leq-\nu\inf_{x\in\overline{A}}I(x).$$

In this terminology, we have established conditions for 2-efficiency (see the remarks after Definition 7.1). To obtain conditions for ν-efficiency with general $\nu\geq 2$, the statements in this section are readily modified. As an example, when A is an $(I+(\nu-1)\xi)$-continuity set and

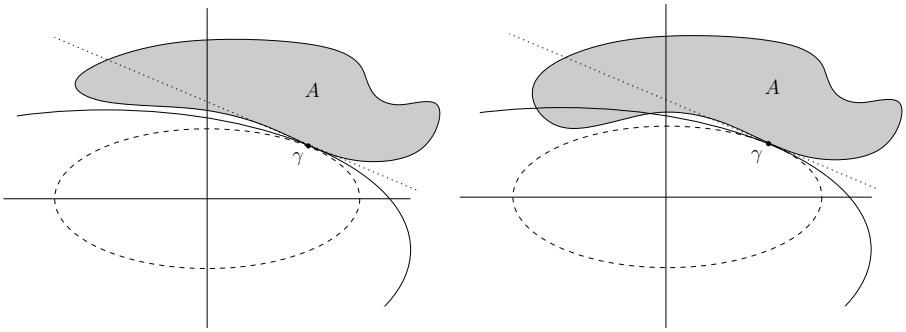

Figure 8.1: Efficient simulation with twist ξ_γ (left) and inefficient simulation with twist ξ_γ (right).

when Assumption 8.1(iii) holds with ξ replaced by both $(\nu-1)\xi$ and $-(\nu-1)\xi$, the exponential twist ξ is ν-efficient if and only if

$$\inf_{x\in\mathcal{X}}[I(x)-(\nu-1)\langle\xi,x\rangle]+\inf_{x\in\overline{A}}[I(x)+(\nu-1)\langle\xi,x\rangle]=\nu\inf_{x\in\overline{A}}I(x).$$

We now illustrate the Varadhan conditions in a simple example. Let ν be the distribution of a random variable X on $\mathbb{R}^d$, and denote the distribution of the sample mean of n i.i.d. copies of X by ν_n. Let ν be such that Cramér's theorem holds (note that $1/n$ plays the role of ϵ in this example; see Proposition 2.16).

For instance, ν_n is a zero-mean bivariate Gaussian distribution with covariance of the form Σ/n for some diagonal matrix Σ; see Figure 8.1. We are interested in $\nu_n(A)$ for two different sets A; these are drawn in the left- and right-hand diagrams of Figure 8.1. Note that the rate function has the form $I(x_1,x_2)=C_1x_1^2+C_2x_2^2$ for some constants $C_1,C_2>0$. As indicated by the dashed level curve of I, the 'most likely point' in A is in both cases γ, i.e., $\arg\inf_{x\in A}I(x)=I(\gamma)$. We can see that there is only one exponential twist $\xi_\gamma\in\mathbb{R}^2$ interesting for simulation purposes, namely the conjugate point of γ. The level curve of $I+\xi_\gamma+\inf_{x\in\mathcal{X}}[I(x)-\xi_\gamma'x]$ that goes through γ is depicted as a solid line. Since both sets A are I- and $(I+\xi_\gamma)$-continuity sets, the twist ξ_γ is asymptotically efficient if and only if A lies entirely 'outside' the solid level curve (see Corollary 8.4). Hence, in the left-hand diagram the twist ξ_γ is asymptotically efficient twist and in the right-hand diagram it is not.

In the literature, sufficient conditions for asymptotic efficiency have been given in terms of dominating points and convexity of A in case the rate function is convex (see, e.g., Sadowsky and Bucklew [280] and references therein). Using Figure 8.1, we explain how it can be seen that the Varadhan conditions improve upon these dominating-point conditions (convexity of A implies the existence of a dominating point, so we focus on dominating points). Every I-continuity set that touches γ and that is contained in the halfspace above the dotted line has dominating point γ. Obviously, such a set lies outside the solid level curve, and one can therefore estimate $\nu_n(A)$ asymptotically efficiently by an exponential twist. However, Figure 8.1 indicates that the dominating-point condition is not necessary: neither of the sets A have a dominating point, while an efficient twist exists in the left-hand diagram.

8.3 Comparison with Sadowsky's conditions

General necessary and sufficient conditions for asymptotic efficiency have been developed by Sadowsky [279]. In this section, we compare the conditions of Theorem 8.3 with Sadowsky's conditions. We show that the assumptions underlying Varadhan's conditions are less restrictive

than Sadowsky's assumptions. Moreover, the sufficient condition in Theorem 8.3 improves upon Sadowsky's sufficiency condition, and the same holds for the necessary condition.

We first introduce some further notation. Let $\mathcal{X}^*$ denote the space of linear continuous functionals $\xi : \mathcal{X} \to \mathbb{R}$ on $\mathcal{X}$. Given a convex function $f : \mathcal{X} \to (-\infty, \infty]$, a point $x \in \mathcal{X}$ is called an *exposed point* of f if there exists a $\delta \in \mathcal{X}^*$ such that $f(y) > f(x) + \langle \delta, y - x \rangle$ for all $y \neq x$. δ is then called an *exposing hyperplane* of I at x.

To compare the Varadhan conditions to Sadowsky's, we first recall Sadowsky's assumptions; shortly, these assumptions are compared to Assumption 8.1.

Assumption 8.5 (Sadowsky's assumptions) *Assume that*

(i) $\mathcal{X}$ *is a locally convex Hausdorff topological vector space,*

(ii) $\{\nu_\epsilon\}$ *satisfies the LDP with a* convex *good rate function* I*,*

(iii) for every $\delta \in \mathcal{X}^*$*,*

$$\Lambda(\delta) := \limsup_{\epsilon \to 0} \epsilon \log \int_{\mathcal{X}} \exp[\langle \delta, x \rangle / \epsilon] \nu_\epsilon(dx) < \infty, \tag{8.6}$$

and

(iv) A *satisfies*

$$0 < \inf_{x \in A^o \cap \mathcal{F}} I(x) = \inf_{x \in A} I(x) = \inf_{x \in \overline{A}} I(x) < \infty,$$

where $\mathcal{F}$ *denotes the set of exposed points of* I*.*

Although Assumption 8.5 looks similar to Assumption 8.1, there are crucial differences.

To start with, $\mathcal{X}$ is not assumed to be a *topological* vector space in Assumption 8.1(i). To see the importance of this difference for applications, note that the space $D([0,1], \mathbb{R})$ of càdlàg functions on $[0,1]$ with values in $\mathbb{R}$ is a (regular, Hausdorff) vector space but no topological vector space when equipped with the Skorokhod topology. We stress that the regularity of $\mathcal{X}$ assumed in Assumption 8.1(i) is implicit in Assumption 8.5(i): any real Hausdorff topological vector space is regular.

Moreover, the convexity of the large-deviation rate function is not assumed in Assumption 8.1(ii). Note that this convexity is a given when an LDP is derived using an (abstract) Gärtner-Ellis-type theorem, but nonconvex rate functions also arise naturally in applications; see Section 8.5 for a discussion. Assumption 8.5(iii) implies Assumption 8.1 since $\gamma\xi$ is a continuous linear functional for any $\gamma \in \mathbb{R}$, while the fourth part of Assumption 8.5 is slightly stronger than requiring that A be an I-continuity set.

In the above comparison between Assumption 8.1 and Assumption 8.5, we have shown the following.

Proposition 8.6 *Assumption 8.5 implies that Assumption 8.1 holds and that* A *is an* I*-continuity set.*

In the remainder of this section, we compare the conditions of Theorem 8.3 to the conditions in Sadowsky [279]. Such a comparison is only possible when Sadowsky's assumption holds, i.e., we must impose the (stronger) Assumption 8.5. We start by repeating Sadowsky's conditions. Given that Assumption 8.5(iv) holds for A, we call $\gamma \in \overline{A}$ a *point of continuity* if $I(\gamma) = \inf_{x \in \overline{A}} I(x)$ and if there exists a sequence $\{\gamma_n\} \subset A^o \cap \mathcal{F}$ such that $\gamma_n \to \gamma$.

Theorem 8.7 (Sadowsky's conditions) *Let Assumption 8.5 hold. The exponential twist* ξ *is asymptotically efficient if*

(a) there is a point of continuity γ such that $I(\gamma) = \langle \xi, \gamma \rangle - \Lambda(\xi)$,

(b) $I(x) + \langle \xi, x \rangle \geq I(\gamma) + \langle \xi, \gamma \rangle$ for all $x \in \overline{A}$, and

(c) either $\langle \xi, x \rangle \geq \langle \xi, \gamma \rangle$ for all $x \in \overline{A}$, or there exists an $x \in \mathcal{F}$ such that ξ is an exposing hyperplane of I at x.

Let Assumption 8.5 hold. If the twist ξ is asymptotically efficient, then

(a) there is a point of continuity γ such that $I(\gamma) = \langle \xi, \gamma \rangle - \Lambda(\xi)$, and

($\tilde{b}$) $I(x) + \langle \xi, x \rangle \geq I(\gamma) + \langle \xi, \gamma \rangle$ for all $x \in A^o \cap \mathcal{F}$.

Proposition 8.8 *Let Assumption 8.5 hold. The sufficient condition in Theorem 8.7 implies the sufficient condition in Theorem 8.3.*

Proof. By condition (a) of Theorem 8.7, there exists a point of continuity $\gamma \in \overline{A}$ such that $I(\gamma) = \inf_{x \in \overline{A}} I(x) = \langle \xi, \gamma \rangle - \Lambda(\xi)$. Since we assume that the LDP holds with some convex I (Assumption 8.5(ii)) and that Assumption 8.7(iii) holds, by Theorem 4.5.10(b) in [100] we have $I(x) = \sup_{\delta \in \mathcal{X}^*} [\langle \delta, x \rangle - \Lambda(\delta)]$, and hence $I(x) \geq \langle \xi, x \rangle - \Lambda(\xi)$. Combining this with $I(\gamma) = \langle \xi, \gamma \rangle - \Lambda(\xi)$, we conclude that

$$\inf_{x \in \mathcal{X}} [I(x) - \langle \xi, x \rangle] \geq -\Lambda(\xi) = I(\gamma) - \langle \xi, \gamma \rangle,$$

where the inequality may obviously be replaced by an equality.

It is immediate from condition (b) of Theorem 8.7 that $\inf_{x \in \overline{A}} [I(x) + \langle \xi, x \rangle] = I(\gamma) + \langle \xi, \gamma \rangle$. Since $\inf_{x \in A^o} I(x) = I(\gamma)$, this implies the sufficient condition (8.4) in Theorem 8.3. □

The above proof shows that part (c) of Sadowsky's sufficient condition is redundant.

Proposition 8.9 *Let Assumption 8.5 hold. The necessary condition in Theorem 8.7 is implied by the necessary condition in Theorem 8.3.*

Proof. Let the twist ξ be asymptotically efficient. We start by showing that a point of continuity exists under Assumption 8.5. First note that $\inf_{x \in A^o \cap \mathcal{F}} I(x) = \inf_{x \in \overline{A}} I(x)$ (Assumption 8.5(iv)) implies that, for any $n \in \mathbb{N}$, we can find some $\gamma_n \in A^o \cap \mathcal{F} \cap K_n$, where

$$K_n := \left\{ x \in \mathcal{X} : I(x) \leq \inf_{y \in \overline{A}} I(y) + 1/n \right\}.$$

Since $\inf_{x \in \overline{A}} I(x) < \infty$ and the rate function is good (Assumption 8.5(ii)), K_n is a compact subset of $\mathcal{X}$ and, hence, also necessarily sequentially compact. Since K_n decreases in n, we obviously have $\{\gamma_n\} \subset K_1$. Hence, we can subtract a subsequence that converges to $\gamma \in K_1$, say. Since K_n is closed for every n and $\{\gamma_\cdot\}$ is eventually in K_n, we must also have that $\gamma \in K_n$ for every n. As a consequence, we have $I(\gamma) \leq \inf_{x \in \overline{A}} I(x)$. Moreover, since $\{\gamma_n\} \subset A^o \cap \mathcal{F}$, we also see that $\gamma \in \overline{A^o \cap \mathcal{F}} \subset \overline{A}$. Therefore, $I(\gamma) = \inf_{x \in A^o \cap \mathcal{F}} I(x) = \inf_{x \in \overline{A}} I(x)$, and, so, γ is a point of continuity.

The necessary condition in Theorem 8.3 implies

$$2 \inf_{x \in \overline{A}} I(x) \leq [I(\gamma) - \langle \xi, \gamma \rangle] + \lim_{n \to \infty} [I(\gamma_n) + \langle \xi, \gamma_n \rangle] = 2I(\gamma).$$

As a result, the inequalities can be replaced by equalities, and we obtain

$$\sup_{x \in \mathcal{X}} [\langle \xi, x \rangle - I(x)] = \langle \xi, \gamma \rangle - I(\gamma) \quad \text{and} \quad \inf_{x \in A^o} [I(x) + \langle \xi, x \rangle] = I(\gamma) + \langle \xi, \gamma \rangle.$$

By Theorem 4.5.10(a) of [100], we also have $\sup_{x \in \mathcal{X}} [\langle \xi, x \rangle - I(x)] = \Lambda(\xi)$ under Assumption 8.5. Hence, $I(\gamma) = \langle \xi, \gamma \rangle - \Lambda(\xi)$ and part (a) of Sadowsky's necessary condition follows. Part ($\tilde{b}$) is immediate by noting that $\inf_{x \in A^o} [I(x) + \langle \xi, x \rangle] = I(\gamma) + \langle \xi, \gamma \rangle$ implies that $I(x) + \langle \xi, x \rangle \geq I(\gamma) + \langle \xi, \gamma \rangle$ for all $x \in A^o$. □

8.4 Uniqueness of the exponential twist

One of the available tools to establish LDPs is the Gärtner-Ellis theorem, see Section 4.5.3 of Dembo and Zeitouni [100]. Apart from being lower semicontinuous, the rate function governing the LDP is then convex. In that case, we can derive some further properties of exponentially twisted importance-sampling estimators.

Throughout this section, we assume that I be a convex good rate function and that (8.6) holds for every $\delta \in \mathcal{X}^*$. Therefore, we are essentially in the framework described in the previous section. By Theorem 4.5.10 of [100], we have

$$\Lambda(\xi) = \sup_{x\in\mathcal{X}}[\langle \xi, x\rangle - I(x)] \quad \text{and} \quad I(x) = \sup_{\xi\in\mathcal{X}^*}[\langle \xi, x\rangle - \Lambda(\xi)]. \tag{8.7}$$

In the next theorem, we formalize the intuition that a typical sample (path) from an importance-sampling distribution should have a lowest possible rate.

Theorem 8.10 *Let Λ be strictly convex, and let A be an I-continuity set. There exists at most one asymptotically efficient twist.*

Proof. As seen in the proof of Proposition 8.9, there exists a $\gamma \in \overline{A}$ such that $\inf_{x\in\overline{A}} I(x) = I(\gamma)$ and $\inf_{x\in A^o}[I(x)+\langle \xi, x\rangle] \le I(\gamma)+\langle \xi, \gamma\rangle$ for any $\xi \in \mathcal{X}^*$. As a result of the strict convexity of Λ and (8.7), there exists at most one $\xi_\gamma \in \mathcal{X}^*$ such that $I(\gamma) = \langle \xi_\gamma, \gamma\rangle - \Lambda(\xi_\gamma)$. For all other twists $\xi \in \mathcal{X}^*$, we have $I(\gamma) > \langle \xi, \gamma\rangle - \Lambda(\xi)$, and therefore

$$\inf_{x\in\mathcal{X}}[I(x) - \langle \xi, x\rangle] = -\sup_{x\in\mathcal{X}}[\langle \xi, x\rangle - I(x)] = -\Lambda(\xi) < I(\gamma) - \langle \xi, \gamma\rangle.$$

Such ξ cannot be asymptotically efficient, since then

$$\inf_{x\in\mathcal{X}}[I(x) - \langle \xi, x\rangle] + \inf_{x\in A^o}[I(x) + \langle \xi, x\rangle] < 2I(\gamma),$$

and (8.5) is violated.

In other words, ξ_γ is the only candidate twist for asymptotic efficiency (if ξ_γ exists). □

Under the assumptions of Theorem 8.10 and in the notation of its proof, the exponential twist ξ_γ is asymptotically efficient if $\inf_{x\in\overline{A}}[I(x) + \langle \xi_\gamma, x\rangle] = I(\gamma) + \langle \xi_\gamma, \gamma\rangle$; condition (8.4) is thus neatly rewritten. The necessary condition (8.5) can be rephrased in a similar way.

It is interesting to apply the previous theorem if there are multiple most likely elements in A. Under additional strict-convexity assumptions, we can conclude that any exponential twist is asymptotically inefficient.

Corollary 8.11 *Let both Λ and I be strictly convex, and A be an I-continuity set. If there exist $\gamma, \eta \in \overline{A}$ with $I(\gamma) = I(\eta) = \inf_{x\in\overline{A}} I(x)$ and $\gamma \neq \eta$, there is no asymptotically efficient exponential twist.*

Proof. The argument in the proof of Theorem 8.10 applies both to γ and η. Therefore, there is nothing to prove if either $I(\gamma) > \langle \xi, \gamma\rangle - \Lambda(\xi)$ for all $\xi \in \mathcal{X}^*$, or $I(\eta) > \langle \xi, \eta\rangle - \Lambda(\xi)$ for all $\xi \in \mathcal{X}^*$. Consequently, we may assume the existence of the two candidate exponentially twists, i.e., that the supremum over ξ in (8.7) is attained both for γ and η. The two twists are called ξ_γ and ξ_η, and we have to show that $\xi_\gamma \neq \xi_\eta$.

Suppose $\xi_\gamma = \xi_\eta$. Since

$$\langle \xi_\gamma, \gamma\rangle - \Lambda(\xi_\gamma) = I(\gamma) = I(\eta) = \langle \xi_\eta, \eta\rangle - \Lambda(\xi_\eta),$$

we have for $0 \le \alpha \le 1$,

$$I(\gamma) = \alpha[\langle \xi_\gamma, \gamma\rangle - \Lambda(\xi_\gamma)] + (1-\alpha)[\langle \xi_\eta, \eta\rangle - \Lambda(\xi_\eta)] = \langle \xi_\gamma, \alpha\gamma + (1-\alpha)\eta\rangle - \Lambda(\xi_\gamma).$$

From this follows that $\Lambda(\xi_\gamma) = \langle \xi_\gamma, \alpha\gamma + (1-\alpha)\eta\rangle - I(\gamma)$. Since we already know that this equality holds for $\alpha = 1$, we obtain a contradiction with (8.7) and the strict convexity of I. □

8.5 An example

In this section, we provide an example showing how Corollary 8.4 is typically used. The conditions of Sadowsky [279] do not apply to this example, since the rate function is nonconvex. Despite this nonconvexity, we show that an exponential twist may still be asymptotically efficient.

Nonconvex rate functions arise naturally in several large-deviation settings. Notably, certain large deviations of Markov chains and Markov processes (such as diffusions) are governed by rate functions that need not be convex; see, e.g., Feng and Kurtz [133]. Intuitively, analyzing the rate of convergence of random functions to a nonlinear (deterministic) function causes the rate function to be nonconvex. Nonconvex rate functions also appear when investigating the rate of convergence to a nondegenerate measure; the example of this section is of the latter type.

For our example, it appears that the event under consideration can be cut into disjoint 'subevents' that comply with Sadowsky's conditions. However, such a *cut approach* might be impossible in other cases, or lead to a large number of subevents (that need to be estimated separately). This is especially relevant for the simulation of hitting probabilities of stochastic processes, as in Collamore [80]. Such hitting probabilities are closely related to extremes. In the next chapter, we investigate under which circumstances a single twist works in a one-dimensional random-walk setting. Chapter 10 analyzes both a single twist and the cut approach in a specific situation.

To avoid irrelevant technicalities, we do not illustrate the Varadhan conditions in a sample-path setting. Instead, we discuss a relatively simple example, which nevertheless gives a good idea in what situations the Varadhan conditions are more useful than Sadowsky's conditions.

Let us first give some background on our example. Recall that a phase-type distribution is a distribution associated to a finite Markov process, which can be characterized by three quantities $(E, \boldsymbol{\alpha}, \boldsymbol{T})$, see, e.g., Asmussen [19, Sec. III.4]. Given an arbitrary distribution ν on $(0, \infty)$, there exists a sequence $\{\nu_n\}$ of phase-type distributions that converges weakly to ν [19, Thm. III.4.2]. By Theorem 2.11, this implies that $\nu_n(A) \to 0$ for a large number of sets A; in fact, $\nu_n(A)$ then vanishes at an exponential rate. We are interested in $\nu_n(A)$ for fixed n.

We consider a particularly simple distribution, namely one with ν concentrated on $\{1, 5\}$. We write $\alpha := \nu(\{1\}) = 1 - \nu(\{5\})$. It is notationally cumbersome to describe a sequence of phase-type distributions that converges weakly to ν in the $(E, \boldsymbol{\alpha}, \boldsymbol{T})$-notation; a direct description is more appropriate here. Define ν_n as the distribution of Y_n, where Y_n has an Erlang(n, n) distribution with probability α and an Erlang$(n, n/5)$ distribution with probability $1 - \alpha$. Recall that the sum of k independent exponentially distributed random variables with parameter λ has an Erlang(k, λ) distribution, and observe that ν_n converges weakly to ν.

The sequence of approximating phase-type distributions $\{\nu_n\}$ are special cases of *mixtures*, and the phenomena that we observe in our example are typically also encountered in the mixture setting. Indeed, the large deviations of some mixtures are governed by nonconvex rate functions. This holds in particular for finite mixtures (in our case, ν_n is a mixture of two distributions), for which the large deviations are readily analyzed. The infinite case is nontrivial; see Dinwoodie and Zabell [116, 117], Chaganty [66], and Biggins [48]. Importantly, mixtures also arise naturally in connection with conditional probabilities, see, e.g., [147].

Consider the simulation of $\nu_n(A)$, where $A := [0, 1/10) \cup (10, \infty)$. It is easy to see that ν_n equals the distribution of

$$\begin{cases} \frac{1}{n} \sum_{i=1}^{n} X_i & \text{with probability } \alpha; \\ \frac{5}{n} \sum_{i=1}^{n} X_i & \text{with probability } 1 - \alpha, \end{cases}$$

where $X_1, X_2, \ldots$ have a standard exponential distribution. Using Cramér's theorem (Proposi-

tion 2.16), it is readily seen that $\{\nu_n\}$ satisfies the LDP with the nonconvex good rate function

$$I(x) := \begin{cases} x - 1 - \log x & \text{if } x \in (0, \frac{5}{4}\log 5], \\ \frac{x}{5} - 1 - \log \frac{x}{5} & \text{if } x > \frac{5}{4}\log 5, \\ \infty & \text{otherwise.} \end{cases}$$

In order to avoid simulation, one could try to compute $\nu_n(A)$ by calculating $\nu_n^1(A)$ and $\nu_n^5(A)$ numerically, where ν_n^1 is the law of $\frac{1}{n}\sum_{i=1}^n X_i$ and ν_n^5 is the law of $\frac{5}{n}\sum_{i=1}^n X_i$. Indeed, both probabilities are readily expressed in terms of (incomplete) gamma functions. However, numerical problems arise already for moderate values of n, since it is necessary to divide an incomplete gamma function by $(n-1)!$.

Therefore, it is natural to attempt the estimation of $\nu_n(A)$ using simulation techniques, for which there are several possibilities. Since α is known, it suffices to simulate $\nu_n^1(A)$ and $\nu_n^5(A)$ separately. However, $\nu_n^1(A)$ cannot be simulated efficiently by twisting exponentially; see Glasserman and Wang [148] for related examples. To see that problems arise, we apply Corollary 8.4 for the simulation of $\nu_n^1(A)$, i.e., with the rate function $I^1(x) := x - 1 - \log x$. First note that $\inf_{x\in A} I^1(x) = I^1(1/10)$; one can readily check that the twist $\xi = -9$ is the only candidate twist for asymptotic efficiency. However, this twist cannot be used for simulation, because the condition in Corollary 8.4 is violated: although $\inf_{x\in\mathbb{R}} I^1(x) + 9x = I^1(1/10) + 9/10$, one also has $\inf_{x\in A} I^1(x) - 9x < I^1(1/10) - 9/10$. In this example, this can readily be overcome at some additional computational costs: one can simulate $\nu_n^1((10,\infty))$ and $\nu_n^1([0,1/10))$ separately. This is the aforementioned cut approach. These two probabilities and $\nu_n^5(A)$ can be simulated efficiently with an exponential twist, so that a reliable estimate of $\nu_n(A)$ is found by simulating *three* different probabilities.

However, it is more efficient to take a direct approach in this case: the reader immediately checks that the twist $\xi = 1/10$ is asymptotically efficient as a direct consequence of Corollary 8.4 (applied to the rate function I). For the direct approach to be more efficient than cutting, it is essential that one can easily sample from the ξ-twisted distribution.

The ξ-twisted measure is λ_n, where for Borel sets A,

$$\lambda_n(A) = \frac{\alpha \int_A \exp(n\xi x)\,\nu_n^1(dx) + (1-\alpha)\int_A \exp(n\xi x)\,\nu_n^5(dx)}{\alpha \int \exp(n\xi y)\,\nu_n^1(dy) + (1-\alpha)\int \exp(n\xi y)\,\nu_n^5(dy)} \tag{8.8}$$

$$= \overline{\alpha}_n \frac{\int_A \exp(nx/10)\,\nu_n^1(dx)}{\left(\frac{10}{9}\right)^n} + (1-\overline{\alpha}_n)\frac{\int_A \exp(nx/10)\,\nu_n^5(dx)}{2^n}, \tag{8.9}$$

with

$$\overline{\alpha}_n = \frac{\alpha}{\alpha + (1-\alpha)\left(\frac{9}{5}\right)^n}.$$

Representation (8.9) shows how we can sample from λ_n: with probability $\overline{\alpha}_n$, we draw from an Erlang$(n, \frac{10}{9}n)$ distribution, and with probability $1-\overline{\alpha}_n$, we draw from an Erlang$(n, 10n)$ distribution. Observe that $\overline{\alpha}_n \to 0$; this is quite natural as the mean of the twisted distribution then tends to $10 = \arg\inf_{x\in A} I(x)$. The likelihood ratio can be written as follows, cf. (8.8):

$$\frac{d\lambda_n}{d\nu_n}(x) = \frac{\exp(nx/10)}{\alpha\left(\frac{10}{9}\right)^n + (1-\alpha)2^n}.$$

Therefore, in the direct approach only one probability is simulated instead of three as in the cut approach. A drawback is that one has to sample from the distribution $(\overline{\alpha}_n, 1-\overline{\alpha}_n)$.

CHAPTER 9

Random walks exceeding a nonlinear boundary

In this chapter, we study a boundary-crossing probability for random walks. Let $S_n : [0,1] \to \mathbb{R}$ denote the polygonal approximation of a random walk with zero-mean increments, where both time and space are scaled by n. We are interested in the estimation of the probability that, for fixed $n \in \mathbb{N}$, S_n exceeds some positive function e.

As a result of the scaling, this probability decays exponentially in n, and importance sampling can be used to achieve variance reduction. Two simulation methods are considered: path-level twisting and step-level twisting. We give necessary and sufficient conditions for both methods to be asymptotically efficient as $n \to \infty$, partly relying on the results in Chapter 8. We subsequently compare the conditions.

9.1 Introduction

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. zero-mean random variables taking values in $\mathbb{R}$, with distribution $\mathbb{P}_X$. For $0 \le t \le 1$, let the scaled polygonal approximation for the partial sums of X_i be given by

$$S_n(t) := \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} X_i + \left(t - \frac{\lfloor nt \rfloor}{n} \right) X_{\lfloor nt \rfloor + 1}, \tag{9.1}$$

where $\lfloor t \rfloor$ denotes the largest integer smaller than or equal to t.

We consider the estimation of a 'time-varying level-crossing probability'. That is, we are interested in estimating $\mathbb{P}(S_n(\cdot) \in A)$ efficiently, where

$$A := \{x \in C([0,1]) : x(t) \ge e(t) \text{ for some } t \in [0,1]\}, \tag{9.2}$$

for some lower semicontinuous function $e : [0,1] \to (0, \infty]$. As $\mathbb{E}S_n(t) = 0$ for any $n \in \mathbb{N}$ and $t \in [0,1]$, the probability $\mathbb{P}(S_n(\cdot) \in A)$ clearly corresponds to a *rare event*, i.e., it vanishes as $n \to \infty$. In fact, since we know from Section 2.3.2 that the large deviations of $S_n(\cdot)$ are governed by Mogul′skiĭ's large-deviation principle, the theory of Chapter 8 becomes available.

We remark that it is possible to consider the (more general) problem with noncentered random variables X and with S_n defined on $[0,T]$ for some $T > 0$. If one imposes that

$e : [0,T] \to (-\infty,\infty]$ satisfies $e(t)/t > \mathbb{E}X_1$ for all $t \in [0,T]$, it is easy to see that we may restrict ourselves without loss of generality to the above simpler setup.

Only in special cases, explicit expressions are available for $\mathbb{P}(S_n(\cdot) \in A)$. When these are not known, one may resort to simulation. As the probability of interest is small, direct simulation can be extremely time-consuming. Unfortunately, the development of efficient simulation methods is usually nontrivial. For the special case that e is affine, i.e., $e(t) = a + bt$ for some $a, b \geq 0$, $\mathbb{P}(S_n(\cdot) \in A)$ corresponds to a ruin probability for the finite-horizon case. Simulation of this probability is studied by Lehtonen and Nyrhinen [216], while Asmussen [18, Sec. X.4] and Asmussen *et al.* [23] consider the analogue in continuous time; then, a Lévy process replaces the random walk. Lévy processes are discussed in more detail in Part C.

This chapter investigates the asymptotic efficiency of two approaches for simulating of $\mathbb{P}(S_n(\cdot) \in A)$, both based on importance sampling. Recall from Section 7.1.1 that samples are thus drawn from a probability measure under which A is not rare anymore, and that the simulation output is corrected with likelihood ratios to retrieve an unbiased estimate. For both methods, the importance sampling is based on sampling steps from an exponentially *twisted* distribution $\mathbb{P}_X^\theta(dz) = e^{\theta z}\mathbb{E}(e^{\theta X_1})^{-1}\mathbb{P}_X(dz)$ for some $\theta > 0$.

In the first approach, which we call *path-level twisting*, typical sample paths under the importance-sampling measure have the form of the 'most likely' path to exceedance under the original measure; let $\tilde{\tau}$ be the epoch where this most likely path exceeds e. This results in the following procedure: (i) Sample the X_i from $\mathbb{P}_X^\theta$ for $i = 1, \ldots, n\tilde{\tau}$ (with θ chosen such that these X_i get a positive mean). Then (ii) sample X_i from the original distribution $\mathbb{P}_X$ for $i = n\tilde{\tau} + 1, \ldots, n$. It is important to note that in this approach the importance sampling is 'turned off' at an *a priori* determined epoch $n\tilde{\tau}$.

In the second approach, which we call *step-level twisting*, we always draw the X_i from $\mathbb{P}_X^\theta$, and each simulation run is stopped at the *random* moment that S_n exceeds e for the first time. This, quite natural, method has been considered in [18, 216]. However, it does *not* correspond to exponentially twisting in a path space, which makes it somewhat more difficult to handle from a theoretical point of view.

Sadowsky [279] was the first to consider the above two simulation methods for estimating $\mathbb{P}(S_n(\cdot) \in A)$. The results of Chapter 8 and the ideas used in the proofs make it possible to improve upon the results of [279] for both simulation methods. Specifically, for the first method (path-level twisting), we correct Sadowsky's claim that it is *never* asymptotically optimal. For the second method (step-level twisting), we give a sufficient condition for asymptotic efficiency that is sharper than Sadowsky's. We exemplify this by establishing a closely related necessary condition.

The chapter is organized as follows. We start with some preliminaries on sample-path large deviations in Section 9.2. Section 9.3 discusses path-level twisting, and finds necessary and sufficient conditions for its asymptotic efficiency. Step-level twisting is studied in Section 9.4; also here conditions for asymptotic efficiency are derived. We compare the efficiency conditions of the two methods in Section 9.5.

9.2 Sample-path large deviations and exceedance probabilities

In Section 2.3.2, we have described the large deviations of $S_n(\cdot)$ with Mogul′skiĭ's large-deviation principle (LDP). Since this result is crucial for the present chapter, we first provide some further background on this LDP in the context of exceedance probabilities.

Let us first recall some notation from Section 2.3. The cumulant-generating function of X_1 is defined as $\Lambda_X(\theta) := \log \mathbb{E}(e^{\theta X_1})$ for $\theta \in \mathbb{R}$, and its Fenchel-Legendre transform is given

by $\Lambda_X^*(z) = \sup_{\xi\in\mathbb{R}}[\xi z - \Lambda_X(\xi)]$. Throughout, we suppose that Λ_X is everywhere finite; in Section 9.3, we discuss this assumption in detail.

Mogul′skiĭ's LDP (Proposition 2.18) states that $\{S_n(\cdot)\}$ satisfies the LDP in $C([0,1])$ with scale sequence $\{n\}$ and the good rate function

$$I(x) := \begin{cases} \int_0^1 \Lambda_X^*(\dot{x}(t))dt & \text{if } x \in \mathcal{AC}; \\ \infty & \text{otherwise,} \end{cases}$$

if $C([0,1])$ is equipped with the topology of uniform convergence. Recall that $\mathcal{AC}$ is the space of absolutely continuous functions, as defined in (2.6). It is important to remark that the level sets of I are compact, which is used in the sequel to apply some results from Chapter 8.

Since we are interested in exceedance probabilities, it is of interest to know how to minimize the rate function I over 'exceedance sets'. To this end, fix some $\tau \in [0,1]$ and set $A_\tau := \{x \in C([0,1]) : x(\tau) \geq e(\tau)\}$. It follows from Jensen's inequality that

$$\begin{aligned} \inf_{x\in A_\tau} \int_0^1 \Lambda_X^*(\dot{x}(t))dt &\geq \inf_{x\in A_\tau} \tau\frac{1}{\tau}\int_0^\tau \Lambda_X^*(\dot{x}(t))dt \geq \inf_{x\in A_\tau} \tau\Lambda_X^*\left(\int_0^\tau \dot{x}(s)ds/\tau\right) \\ &= \tau\Lambda_X^*(e(\tau)/\tau). \end{aligned}$$

Consequently, the unique minimizing argument $\tilde{x}_\tau$ of $\inf_{x\in A_\tau} I(x)$ is a piecewise straight line given by

$$\gamma_\tau(t) := \begin{cases} t(e(\tau)/\tau) & \text{if } 0 \leq t \leq \tau; \\ e(\tau) & \text{if } \tau < t \leq 1. \end{cases} \tag{9.3}$$

Let us now consider the minimizer of I on A as defined in (9.2). Note that $A = \bigcup_{\tau\in[0,1]} A_\tau$, and define

$$\tilde{\tau} := \arg\inf_{\tau\in(0,1]} \tau\Lambda_X^*(e(\tau)/\tau), \tag{9.4}$$

which exists by lower semicontinuity of $t \mapsto e(t)/t$, but is not necessarily unique. A minimizer over the set A is then given by $\arg\inf_{x\in A} I(x) = \arg\inf_{x\in A_{\tilde{\tau}}} I(x) = \gamma_{\tilde{\tau}}$.

It is also useful to define an 'ϵ-perturbed' version of γ_τ. For $\tau \in [0,1]$ and $\epsilon > 0$, define γ_τ^ϵ by

$$\gamma_\tau^\epsilon(t) := \begin{cases} (e(\tau)+\epsilon)t/\tau & \text{if } 0 \leq t \leq \tau; \\ e(\tau)+\epsilon & \text{if } \tau < t \leq 1. \end{cases} \tag{9.5}$$

9.3 Path-level twisting

In this section, we study the simulation of $\mathbb{P}(S_n(\cdot) \in A)$ by path-level twisting, where A is defined in (9.2):

$$A = \{x \in C([0,1]) : x(t) \geq e(t) \text{ for some } t \in [0,1]\}.$$

This analysis culminates in a necessary and sufficient condition for efficiency of this simulation method.

We start with the formulation of the underlying assumptions:

Assumption 9.1 *We assume that*

(i) $\Lambda_X(\theta) < \infty$ *for all* $\theta \in \mathbb{R}$,

(ii) $\mathbb{P}_X$ *is nondegenerate, and*

(iii) $0 < \inf_{t\in[0,1]} e(t) < \infty$.

Note that Assumption 9.1(i) implies that Mogul′skiĭ's LDP (Proposition 2.18) applies. This assumption can be considerably relaxed for a Mogul′skiĭ-type large-deviation principle to hold; one then uses other spaces, other topologies, and slightly modified rate functions. For instance, Mogul′skiĭ [234] allows the cumulant-generating function to be finite only in a neighborhood of zero and uses the space of càdlàg functions D endowed with the (completed) Skorokhod topology; see also Mogul′skiĭ [235]. Although Mogul′skiĭ's rate function is slightly different from I, the infima over exceedance sets are attained by straight lines as in Section 9.2, which is its only essential property for this chapter. Dembo and Zajic [99] and de Acosta [84] work under the hypothesis of a finite cumulant-generating function of $|X|$. However, in the simulation framework of the present chapter, it is more convenient to impose Assumption 9.1(i).

Let us first cast the definition of an exponentially twisted distribution into the present 'path' framework, cf. Section 8.2. For a twist ξ, we introduce the measure λ_n^ξ by setting for any Borel set B in $C([0,1])$,

$$\lambda_n^\xi(B) = \int_B \exp\left(n\xi(x) - \log \int_{C([0,1])} \exp[n\xi(y)]\mathbb{P}(S_n \in dy)\right) \mathbb{P}(S_n \in dx).$$

The measure λ_n^ξ is said to be *exponentially twisted* with twist ξ. Since λ_n^ξ is a measure on the path-space $C([0,1])$, ξ is a *path-level twist*.

Before formulating a condition that is equivalent to asymptotic efficiency, we show in the next lemma that there is at most one asymptotically efficient twist. Furthermore, the uniqueness of $\tilde\tau$, as defined in (9.4), is necessary for a path-level twist to be asymptotically efficient.

For these assertions to hold, A has to be an I-continuity set; recall from Section 2.3 that this means $\inf_{x\in\overline{A}} I(x) = \inf_{x\in A^o} I(x)$. At the end of this section we give sharp conditions for A to satisfy this and related conditions. We write $\operatorname{dom}\Lambda_X^* := \{z \in \mathbb{R} : \Lambda_X^*(z) < \infty\}$.

Lemma 9.2 *Let Assumption 9.1 hold, and let A be an I-continuity set.*

(i) There is at most one asymptotically efficient path-level twist.

(ii) If $\tilde\tau$ is unique, the only path-level exponential twist that can achieve asymptotic efficiency is $\xi_{\tilde\tau}(x) := \alpha x(\tilde\tau)$, where $\alpha = \arg\sup_{\theta\in\mathbb{R}}[\theta(e(\tilde\tau)/\tilde\tau) - \Lambda_X(\theta)]$.

(iii) If there are two points $\tilde\tau_1, \tilde\tau_2 \in \arg\inf_{\tau\in(0,1]} \tau\Lambda_X^*(e(\tau)/\tau)$ satisfying $e(\tilde\tau_i)/\tilde\tau_i \in (\operatorname{dom}\Lambda_X^*)^o$ and $\tilde\tau_1 \neq \tilde\tau_2$, there is no asymptotically efficient path-level twist.

Proof. We start with some elementary observations. Define

$$\Lambda(\xi) := \lim_{n\to\infty} \frac{1}{n} \log \int_{C([0,1])} \exp[n\xi(x)]\mathbb{P}(S_n \in dx),$$

and note that this function is finite (in particular, the limit exists) for every functional ξ on $C([0,1])$ as shown in Sadowsky [279, p. 407]. Moreover, Λ is strictly convex whenever Λ_X is strictly convex; this follows from Assumption 9.1(ii) as one easily deduces from the proof of Hölder's inequality (see, e.g., Royden [276]). Therefore, the first claim immediately follows from Theorem 8.10.

For claim (ii) we have to show that $I(\gamma_{\tilde\tau}) = \xi_{\tilde\tau}(\gamma_{\tilde\tau}) - \Lambda(\xi_{\tilde\tau})$, where $\tilde\tau$ is now unique. Observe that the minimizer in $\inf_{x\in A} I(x)$ is $\gamma_{\tilde\tau}$, and that $\xi_{\tilde\tau}$ is a continuous linear functional on $C([0,1])$. We first calculate $\Lambda(\xi_{\tilde\tau})$.

Let $\underline{\tilde\tau}_n := \lfloor n\tilde\tau\rfloor/n$, so that $\underline{\tilde\tau}_n \to \tilde\tau$ as $n\to\infty$. We then have by independence,

$$\int_{C([0,1])} \exp(n\alpha x(\tilde\tau))\mathbb{P}(S_n \in dx)$$

$$= \int_{\mathbb{R}^n} \exp\left(\alpha \sum_{i=1}^{n\underline{\tilde{\tau}}_n} z_i + \alpha(\tilde{\tau} - \underline{\tilde{\tau}}_n) z_{n\underline{\tilde{\tau}}_n+1}\right) \mathbb{P}_X(dz_1)\cdots\mathbb{P}_X(dz_n)$$

$$= \int_{\mathbb{R}} \exp(\alpha(\tilde{\tau} - \underline{\tilde{\tau}}_n)z)\mathbb{P}_X(dz) \left(\int_{\mathbb{R}} \exp(\alpha z)\mathbb{P}_X(dz)\right)^{n\underline{\tilde{\tau}}_n}.$$

Observe that $\Lambda_X(\theta) < \infty$ for all $\theta \in \mathbb{R}$ by Assumption 9.1(i), and that Λ_X is continuous due to its convexity. Consequently, the first integral of the last expression converges to one. Using the definition of $\xi_{\tilde{\tau}}$, we conclude that

$$\Lambda(\xi_{\tilde{\tau}}) = \lim_{n\to\infty} \frac{1}{n} \log \int_{C([0,1])} \exp(n\alpha x(\tilde{\tau}))\mathbb{P}(S_n \in dx) = \tilde{\tau}\Lambda_X(\alpha),$$

implying $\xi_{\tilde{\tau}}(\gamma_{\tilde{\tau}}) - \Lambda(\xi_{\tilde{\tau}}) = \alpha e(\tilde{\tau}) - \tilde{\tau}\Lambda_X(\alpha)$. By definition of α, this equals $\tilde{\tau}\Lambda_X^*(e(\tilde{\tau})/\tilde{\tau}) = I(\gamma_{\tilde{\tau}})$.

We now proceed with the proof of (iii). Recall the definition of γ_τ^ϵ in (9.5). Observe that $\{\gamma_{\tilde{\tau}_i}^{1/n}\} \subset A^o$, and that both $\gamma_{\tilde{\tau}_i}^{1/n} \to \gamma_{\tilde{\tau}_i}$ and, as a consequence of the assumption imposed on the $\tilde{\tau}_i$, $I(\gamma_{\tilde{\tau}_i}^{1/n}) \to I(\gamma_{\tilde{\tau}_i})$. Therefore, the reasoning that established the first claim shows that the only candidate path-level twists are $\xi_{\tilde{\tau}_1}$ and $\xi_{\tilde{\tau}_2}$. It suffices to observe that these twists are unequal since $\tilde{\tau}_1 \neq \tilde{\tau}_2$. □

Motivated by Lemma 9.2, we often assume the uniqueness of the minimizer $\tilde{\tau}$ of $\tau \mapsto \tau\Lambda_X^*(e(\tau)/\tau)$ in the remainder of this chapter. We now state the main theorem of this section.

Theorem 9.3 *Let Assumption 9.1 hold, and suppose that $\tau\Lambda_X^*(e(\tau)/\tau)$ has a unique minimizer $\tilde{\tau}$. Moreover, let A be both an I-continuity set and an $(I + \xi_{\tilde{\tau}})$-continuity set.*

The path-level twist $\xi_{\tilde{\tau}}$ defined as $\xi_{\tilde{\tau}}(x) = \alpha x(\tilde{\tau})$ is asymptotically efficient if and only if

$$\tilde{\tau}\Lambda_X^*\left(\frac{e(\tilde{\tau})}{\tilde{\tau}}\right) + \alpha e(\tilde{\tau}) \tag{9.6}$$

$$\leq \min\left\{\inf_{\tau\in(0,\tilde{\tau})}\left(\tau\Lambda_X^*\left(\frac{e(\tau)}{\tau}\right) + \inf_{\beta\in\mathbb{R}}\left[(\tilde{\tau}-\tau)\Lambda_X^*\left(\frac{\beta - e(\tau)}{\tilde{\tau}-\tau}\right) + \alpha\beta\right]\right),\right.$$

$$\left.\inf_{\tau\in(\tilde{\tau},1]}\inf_{\beta\in\mathbb{R}}\left[\tilde{\tau}\Lambda_X^*\left(\frac{\beta}{\tilde{\tau}}\right) + (\tau-\tilde{\tau})\Lambda_X^*\left(\frac{e(\tau)-\beta}{\tau-\tilde{\tau}}\right) + \alpha\beta\right]\right\}.$$

Proof. We prove the claim by invoking Corollary 8.4. Note that the underlying assumptions hold. Hence, the path-level twist $\xi_{\tilde{\tau}}$ is asymptotically efficient if and only if

$$\inf_{x\in C([0,1])}[I(x) - \xi_{\tilde{\tau}}(x)] + \inf_{x\in A}[I(x) + \xi_{\tilde{\tau}}(x)] = 2\inf_{x\in A} I(x). \tag{9.7}$$

The rest of the proof consists of rewriting condition (9.7).

The first term on the left-hand side of (9.7) is

$$\inf_{x\in C([0,1])}[I(x) - \xi_{\tilde{\tau}}(x)] = -\Lambda(\xi_{\tilde{\tau}}) = -\tilde{\tau}\Lambda_X(\alpha) = \tilde{\tau}\Lambda_X^*(e(\tilde{\tau})/\tilde{\tau}) - \alpha e(\tilde{\tau}),$$

so the left-hand side of (9.6) equals $2\inf_{x\in A} I(x) - \inf_{x\in C([0,1])}[I(x) - \xi_{\tilde{\tau}}(x)]$. Since clearly $\tilde{\tau}\Lambda_X^*(e(\tilde{\tau})/\tilde{\tau}) + \alpha e(\tilde{\tau}) \geq \inf_{x\in A}[I(x) + \alpha x(\tilde{\tau})]$, the condition

$$\tilde{\tau}\Lambda_X^*(e(\tilde{\tau})/\tilde{\tau}) + \alpha e(\tilde{\tau}) \leq \inf_{x\in A}[I(x) + \alpha x(\tilde{\tau})] \tag{9.8}$$

is necessary and sufficient for asymptotic efficiency. It remains to investigate the right-hand side of this inequality.

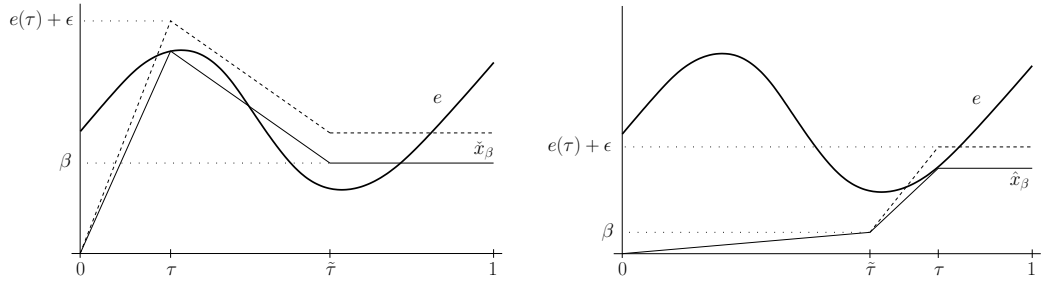

Figure 9.1: Two possibilities for the minimizing argument of $\inf_{x\in A}[I(x)+\alpha x(\tilde{\tau})]$.

Jensen's inequality shows that a minimizing argument of $\inf_{x\in A}[I(x)+\alpha x(\tilde{\tau})]$ is a piecewise straight line, which must exceed e in $[0,1]$, say at time τ, and has some value $\beta\in\mathbb{R}$ at time $\tilde{\tau}$. The right-hand side of (9.8) is the infimum over β and τ when these paths are substituted in the expression $I(x)+\alpha x(\tilde{\tau})$. Since $\tilde{\tau}\Lambda_X^*(e(\tilde{\tau})/\tilde{\tau})+\alpha e(\tilde{\tau}) = I(\gamma_{\tilde{\tau}})+\alpha e(\tilde{\tau})$, we may assume that $\tau\neq\tilde{\tau}$ in order to derive a condition that is equivalent with (9.8).

Two possibilities arise. First, x can exceed e for the first time at some $\tau<\tilde{\tau}$, then assumes some value $\beta\in\mathbb{R}$ at $\tilde{\tau}$, and is constant on $[\tilde{\tau},1]$. This path is denoted by $\check{x}_{\beta,\tau}$. Another possibility is that x has some value β at $\tilde{\tau}$, exceeds e for some $\tau>\tilde{\tau}$, and then becomes constant. This path is denoted by $\hat{x}_{\beta,\tau}$. These two possible cases are illustrated by the solid lines in Figure 9.1.

It is immediate that $\check{x}_{\beta,\tau}$ satisfies for $\tau<\tilde{\tau}$

$$I(\check{x}_{\beta,\tau})+\alpha\check{x}_{\beta,\tau}(\tilde{\tau}) = \tau\Lambda_X^*\left(\frac{e(\tau)}{\tau}\right)+(\tilde{\tau}-\tau)\Lambda_X^*\left(\frac{\beta-e(\tau)}{\tilde{\tau}-\tau}\right)+\alpha\beta.$$

This corresponds to the left-hand diagram in Figure 9.1. The same argument shows that for $\tau>\tilde{\tau}$,

$$I(\hat{x}_{\beta,\tau})+\alpha\hat{x}_{\beta,\tau}(\tilde{\tau}) = \tilde{\tau}\Lambda_X^*\left(\frac{\beta}{\tilde{\tau}}\right)+(\tau-\tilde{\tau})\Lambda_X^*\left(\frac{e(\tau)-\beta}{\tau-\tilde{\tau}}\right)+\alpha\beta,$$

which finishes the proof. □

We remark that Equation (9.6) can be slightly simplified using Λ_X. Note that

$$\begin{aligned}&\inf_{\beta\in\mathbb{R}}\left[(\tilde{\tau}-\tau)\Lambda_X^*\left(\frac{\beta-e(\tau)}{\tilde{\tau}-\tau}\right)+\alpha\beta\right]\\ &\quad= -(\tilde{\tau}-\tau)\sup_{\beta\in\mathbb{R}}\left[-\alpha\frac{\beta-e(\tau)}{\tilde{\tau}-\tau}-\Lambda_X^*\left(\frac{\beta-e(\tau)}{\tilde{\tau}-\tau}\right)\right]+\alpha e(\tau),\end{aligned}$$

and that the sup-term in this expression equals $\Lambda_X(-\alpha)$ by the duality lemma (Lemma 4.5.8 of [100]). Thus, (9.6) is equivalent to

$$\begin{aligned}&\tilde{\tau}\Lambda_X^*\left(\frac{e(\tilde{\tau})}{\tilde{\tau}}\right)+\alpha e(\tilde{\tau})\\ &\quad\leq \min\left\{\inf_{\tau\in(0,\tilde{\tau})}\left(\tau\Lambda_X^*\left(\frac{e(\tau)}{\tau}\right)-(\tilde{\tau}-\tau)\Lambda_X(-\alpha)+\alpha e(\tau)\right),\right.\\ &\qquad\left.\inf_{\tau\in(\tilde{\tau},1]}\inf_{\beta\in\mathbb{R}}\left[\tilde{\tau}\Lambda_X^*\left(\frac{\beta}{\tilde{\tau}}\right)+(\tau-\tilde{\tau})\Lambda_X^*\left(\frac{e(\tau)-\beta}{\tau-\tilde{\tau}}\right)+\alpha\beta\right]\right\}.\end{aligned}$$

To illustrate Theorem 9.3, we now work out an example. Let the X_i have a standard normal distribution, i.e., $\Lambda_X(\xi) = \Lambda_X^*(\xi) = \frac{1}{2}\xi^2$. Set $e(\tau) = 1 + |2\tau - 1|$. It is easy to check that $\tau\Lambda_X^*(e(\tau)/\tau) = e(\tau)^2/(2\tau)$ is minimized for $\tilde{\tau} = 1/2$, showing that $\alpha = 2$. It is also immediate that $e(\tau)^2/(2\tau) + 2\tau - 1 + 2e(\tau)$ 'attains' its minimum value over $(0, 1/2)$ as $\tau \uparrow 1/2$. The second minimizing β in (9.6) is then $1/(2\tau)$, and the minimum value over $(1/2, 1]$ of the resulting function is attained for $\tau \downarrow 1/2$. Consequently, we can estimate the desired probability efficiently by path-level twisting. Therefore, this example corrects the unproven claim of Sadowsky [279, p. 408] that no path-level twist is asymptotically efficient.

Different behavior is observed if $e(\tau) = 1 + |\tau - 1/2|$. Again, $\tilde{\tau} = 1/2$ and $\alpha = 2$, but now it turns out that the infimum in (9.6) is attained for $\tau = 1$. Therefore, the *same* twist as before is now asymptotically *in*efficient.

To implement the simulation procedure, the path-level twist $\xi_{\tilde{\tau}}$ should be translated into an importance sampling distribution for $(X_1, \ldots, X_n)$. Sadowsky [279] shows that the exponentially step-level θ-twisted distribution of X,

$$\mathbb{P}_X^\theta(dz) := \exp(\theta z - \Lambda_X(\theta))\mathbb{P}_X(dz),$$

are the 'building blocks' for the required path-level exponential twist. Indeed, the step sizes $X_1, \ldots, X_{\lfloor n\tilde{\tau} \rfloor}$ should be sampled from $\mathbb{P}_X^\alpha$, $X_{\lceil n\tilde{\tau} \rceil}$ from $\mathbb{P}_X^{\alpha(n\tilde{\tau} - \lfloor n\tilde{\tau} \rfloor)}$, and $X_{\lceil n\tilde{\tau} \rceil + 1}, \ldots, X_n$ from $\mathbb{P}_X$; the X_i should also be mutually independent. Using the realizations of the X_i, one can construct a sample path with (9.1). The resulting paths are samples from the path-level twisted distribution $\lambda_n^{\xi_{\tilde{\tau}}}$.

Both Lemma 9.2 and Theorem 9.3 require certain continuity properties of A. The remainder of this section is devoted to sharp conditions for these to hold.

Continuity properties of A

We start by showing that A is closed. In fact, for later use, we prove this in slightly more generality. Consider the set

$$A_m^M := \{x \in C([0,1]) : x(t) \geq M(t) \text{ for some } t \in [0,1] \text{ or } x(t) \leq m(t) \text{ for some } t \in [0,1]\},$$

where $M : [0,1] \to (-\infty, \infty]$ is lower semicontinuous and $m : [0,1] \to [-\infty, \infty)$ is upper semicontinuous with $m \leq M$ on $[0,1]$. We prove that A_m^M is closed, which implies that A is closed by choosing $m = e$ and $M \equiv -\infty$.

Lemma 9.4 *A_m^M is closed in $C([0,1])$.*

Proof. Let $\{x_n\}$ be a sequence in A_m^M converging in sup-norm to some $x \in C([0,1])$. Suppose that $x \notin A_m^M$, and set $\epsilon := \min(\inf_{t\in[0,1]}[M(t) - x(t)], \inf_{t\in[0,1]}[x(t) - m(t)])/2$. Since $[0,1]$ is compact, the infima in this expression are attained, so that $\epsilon > 0$. From the convergence in sup-norm it follows that $|x_n(t) - x(t)| \leq \epsilon$ for all $t \in [0,1]$ and n large enough. By construction of ϵ, a contradiction is obtained by noting that this would imply $x_n \notin A_m^M$. □

We now give a sharp condition for A to be an I-continuity set, for which we do not require uniqueness of $\tilde{\tau}$.

Lemma 9.5 *If $e(\tilde{\tau})/\tilde{\tau} \in (\operatorname{dom} \Lambda_X^*)^o$ for some $\tilde{\tau}$ with $\tilde{\tau}\Lambda_X^*(e(\tilde{\tau})/\tilde{\tau}) = \inf_{x\in A} I(x)$, then A is an I-continuity set.*

Proof. Similar arguments as in the proof of Lemma 9.4 show that $A^o = \{x \in C([0,1]) : x(t) > e(t) \text{ for some } t \in [0,1]\}$. As A is closed, it suffices to prove that $\inf_{x\in A} I(x) = \inf_{x\in A^o} I(x)$. Let

$\tilde{\tau}$ be such that $\tilde{\tau}\Lambda_X^*(e(\tilde{\tau})/\tilde{\tau}) = \inf_{x\in A} I(x)$. With $\gamma_{\tilde{\tau}}^\epsilon$ as in (9.5), we have $\gamma_{\tilde{\tau}}^\epsilon \in A^o$ and $I(\gamma_{\tilde{\tau}}^\epsilon) = \tilde{\tau}\Lambda_X^*([e(\tilde{\tau})+\epsilon]/\tilde{\tau})$. By convexity of Λ_X^* and the fact that there is a neighborhood of $e(\tilde{\tau})/\tilde{\tau}$ on which Λ_X^* is finite, Λ_X^* is continuous on this neighborhood, and therefore $\Lambda_X^*([e(\tilde{\tau})+\epsilon]/\tilde{\tau}) \downarrow \Lambda_X^*(e(\tilde{\tau})/\tilde{\tau})$ as $\epsilon \downarrow 0$; note that $\inf_{t\in(0,1]} e(t)/t > 0$ as a consequence of Assumption 9.1(iii). By the monotone convergence theorem, $I(\gamma_{\tilde{\tau}}^\epsilon)$ converges to $I(\gamma_{\tilde{\tau}})$. □

It is also of interest to give a condition for A to be an $(I+\xi_{\tilde{\tau}})$-continuity set. This is the content of the next lemma, but we omit the proof since it is a variation on the ϵ-argument given in the proof of Lemma 9.5. The perturbed paths are drawn as the dashed lines in Figure 9.1. Recall the definitions of $\check{x}_{\beta,\tau}$ and $\hat{x}_{\beta,\tau}$ in the proof of Theorem 9.3.

Lemma 9.6 *Let $\tilde{\tau}$ be given. If one of the following two conditions holds, then A is an $(I+\xi_{\tilde{\tau}})$-continuity set:*

(i) There exists an $\overline{x} \in \arg\inf_{x\in A}[I(x)+\alpha x(\tilde{\tau})]$ of the form $\check{x}_{\beta,\overline{\tau}}$ for some $\beta \in \mathbb{R}$ and $\overline{\tau} \le \tilde{\tau}$, for which $e(\overline{\tau})/\overline{\tau} \in (\operatorname{dom}\Lambda_X^*)^o$,

(ii) There exists an $\overline{x} \in \arg\inf_{x\in A}[I(x)+\alpha x(\tilde{\tau})]$ of the form $\hat{x}_{\beta,\overline{\tau}}$ for some $\beta \in \mathbb{R}$ and $\overline{\tau} > \tilde{\tau}$, for which $(e(\overline{\tau}) - \overline{x}(\tilde{\tau}))/(\overline{\tau}-\tilde{\tau}) \in (\operatorname{dom}\Lambda_X^*)^o$.

9.4 Step-level twisting

This section is devoted to a simplification of the simulation scheme (i.e., the measure $\lambda_n^{\xi_{\tilde{\tau}}}$) studied in Section 9.3. The new scheme overcomes an intuitive difficulty with a path-level twisted change of measure. If a path sampled from $\lambda_n^{\xi_{\tilde{\tau}}}$ remains below e on $[0, \lceil n\tilde{\tau}\rceil/n]$, it has little chance of exceeding e after $\lceil n\tilde{\tau}\rceil/n$. Indeed, since the original measure $\mathbb{P}_X$ is then used for sampling, e is rarely exceeded after $\lceil n\tilde{\tau}\rceil/n$. By the form of the estimator (7.1), a sample path that does not exceed e does not contribute to the resulting estimate, so that it is undesirable to have too many of such paths in the simulation.

The idea of the simplified simulation scheme is to sample every random variable X_i from $\mathbb{P}_X^\alpha$, *until e has been exceeded*. The simulation is then stopped and the likelihood is calculated. We refer to this setup, which has first been studied in [279], as *step-level twisting*. Note that this contrasts with path-level twisting as described in the preceding section, since there the step-size distribution is twisted up to a *fixed* twist-horizon $n\tilde{\tau}$. In the setting of this section, this horizon is sample-dependent.

Since both path-level twisting and step-level twisting are algorithms for estimating the same probability, it is legitimate to ask which procedure is better. To answer this question rigorously, it is our aim to develop necessary and sufficient conditions for asymptotic efficiency of step-level twisting. These conditions are the 'step analogue' of Theorem 9.3. A comparison of the two sets of conditions is the subject of Section 9.5.

Intuitively, it depends on the specific form of e if the probability of exceeding e on $[\lceil n\tilde{\tau}\rceil/n, 1]$ is small enough for the simplification to work. Sadowsky [279, Prop. 2] finds a sufficient condition in terms of a saddle-point inequality. The sufficient condition of Theorem 9.8 below improves upon this result significantly: our necessary condition is extremely 'close' to the sufficiency condition.

Throughout this section, we adopt the setup and notation of the previous section. It is worthwhile to specify the exact assumptions that we impose.

Assumption 9.7 *We assume that*

(i) Assumption 9.1 holds,

(ii) $\tilde{\tau} = \arg\inf \tau\Lambda_X^*(e(\tau)/\tau)$ is unique, and

(iii) $e(\tilde{\tau})/\tilde{\tau} \in (\operatorname{dom} \Lambda_X^*)^o$.

In the previous section, we have seen that this set of assumptions guarantees that Mogul′skiĭ's large-deviation principle holds, and that A is an I-continuity set, see Lemma 9.5. Lemma 9.2 shows that the uniqueness of $\tilde{\tau}$ in Assumption 9.7(ii) is required to have a unique twist α for the distribution of the X_i.

The next theorem generalizes the findings of Lehtonen and Nyrhinen [216] to nonlinear boundaries e.

Theorem 9.8 *Let Assumption 9.7 hold, and let e be lower semicontinuous. Step-level twisting is asymptotically efficient if*

$$\inf_{\tau\in(0,1]} \left[\tau\Lambda_X^*\left(\frac{e(\tau)}{\tau}\right) + \alpha e(\tau) - \tau\Lambda_X(\alpha)\right] \geq 2\tilde{\tau}\Lambda_X^*\left(\frac{e(\tilde{\tau})}{\tilde{\tau}}\right). \tag{9.9}$$

Let Assumption 9.7 hold, and let e be upper semicontinuous. If step-level twisting is asymptotically efficient, then

$$\inf_{\{\tau\in(0,1]:\, e(\tau)/\tau\in(\operatorname{dom}\Lambda_X^*)^o\}} \left[\tau\Lambda_X^*\left(\frac{e(\tau)}{\tau}\right) + \alpha e(\tau) - \tau\Lambda_X(\alpha)\right] \geq 2\tilde{\tau}\Lambda_X^*\left(\frac{e(\tilde{\tau})}{\tilde{\tau}}\right). \tag{9.10}$$

Proof. As seen in Section 7.1.2, since A is an I-continuity set, asymptotic efficiency is equivalent to

$$\limsup_{n\to\infty} \frac{1}{n} \log E_n^{(2)} \leq -2 \inf_{x\in A} I(x),$$

where $E_n^{(2)}$ denotes the second moment of the estimator.

We introduce some notation used throughout the proof.

Notation. Let $g : [0,1] \to [0,\infty]$ be given by $g(t) := t\Lambda_X^*(e(t)/t)$ for $t > 0$ and $g(0) := 0$, and define $f : [0,1] \to [-\infty,\infty)$ as

$$f(t) := -\alpha e(t) + t\Lambda_X(\alpha).$$

We also set for $\tau \in (0,1]$,

$$\tilde{A}_\tau := \{x \in C([0,1]) : x(t) < e(t) \text{ for } t \in [0,\tau), x(\tau) \geq e(\tau)\},$$

i.e., $\tilde{A}_\tau$ are the paths that exceed e for the first time at τ. Note that the $\tilde{A}_\tau$ are disjoint and that $\bigcup_{\tau\in[0,1]} \tilde{A}_\tau = A$.

Paths generated by the step-level-twisting procedure are in general no elements of $C([0,1])$, since the simulation is stopped at some random time, not necessarily at time 1. To overcome this, note that stopping a simulation run amounts to continuing the simulation by drawing from $\mathbb{P}_X$. In other words, importance sampling is 'turned off' in the sense that the sampling distribution becomes $\mathbb{P}_X$ after exceeding e. Therefore, the distribution in $C([0,1])$ of sample paths generated by step-level twisting is well-defined; we denote it by μ_n. The 'original' distribution of S_n in $C([0,1])$ is denoted by ν_n. One can readily check that on $\tilde{A}_\tau$, we have (for x in the support of ν_n)

$$\frac{d\nu_n}{d\mu_n}(x) = \exp(-n\alpha x(\tau) + n\tau\Lambda_X(\alpha)).$$

In the proof of the sufficient condition, we use the function $\zeta : C([0,1]) \to [-\infty,\infty)$ given by

$$\zeta(x) := \begin{cases} f(\tau) & \text{if } x \in \tilde{A}_\tau; \\ -\infty & \text{otherwise.} \end{cases}$$

Since ζ is in general not upper semicontinuous, we cannot apply Varadhan's integral lemma to prove the sufficient condition. However, it is quite fruitful to use some ideas of its proof (see Theorem 4.3.1 in Dembo and Zeitouni [100]).

The sufficient condition. In the proof it is essential that the functions involved have specific continuity properties. Obviously, f is upper semicontinuous under the assumption that e is lower semicontinuous. We now prove that g is lower semicontinuous. For this, let $\{t_n\}$ be a sequence in $[0,1]$ converging to some $t \in [0,1]$. For $t = 0$, it certainly holds that $\liminf_{n\to\infty} g(t_n) \geq 0 = g(0)$. Therefore, we assume $t > 0$. Since $\inf_{t\in(0,1]} e(t)/t > 0$ and Λ_X^* is nondecreasing on $[0,\infty)$ (X_1 is centered), we observe that

$$\begin{aligned}\liminf_n t_n \Lambda_X^*(e(t_n)/t_n) &= t \liminf_n \Lambda_X^*(e(t_n)/t_n) \geq t\Lambda_X^*(\liminf_n e(t_n)/t_n) \\ &= t\Lambda_X^*(\liminf_n e(t_n)/t) \geq t\Lambda_X^*(e(t)/t),\end{aligned}$$

where the last inequality uses the lower semicontinuity of e. Hence, g is lower semicontinuous.

Let $\epsilon > 0$. For any $t \in [0,1]$, by semicontinuity we know that there exists an open neighborhood T_t of t with

$$\inf_{\tau\in\overline{T_t}} g(\tau) \geq g(t) - \epsilon \quad \text{and} \quad \sup_{\tau\in\overline{T_t}} f(\tau) \leq f(t) + \epsilon. \tag{9.11}$$

Since $\bigcup_{t\in[0,1]} T_t$ is an open cover of the compact space $[0,1]$, one can find N and $t_1, \ldots, t_N \in [0,1]$ such that $\bigcup_{i=1}^N T_{t_i} = [0,1]$.

As $d\nu_n/d\mu_n \leq \exp(n\zeta)$ on each of the sets $\tilde{A}_\tau$, the cover-argument implies that (see Lemma 1.2.15 of [100])

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \int_A \frac{d\nu_n}{d\mu_n} d\nu_n &\leq \limsup_{n\to\infty} \frac{1}{n} \log \int_A \exp(n\zeta(x))\nu_n(dx) \\ &= \max_{i=1}^N \limsup_{n\to\infty} \frac{1}{n} \log \int_{\bigcup_{\tau\in T_{t_i}} \tilde{A}_\tau} \exp(n\zeta(x))\nu_n(dx).\end{aligned}$$

The integral in this expression can be bounded by noting that ζ is majorized on $\bigcup_{\tau\in T_{t_i}} \tilde{A}_\tau$ using (9.11):

$$\int_{\bigcup_{\tau\in T_{t_i}} \tilde{A}_\tau} \exp(n\zeta(x))\nu_n(dx) \leq \exp[f(t_i) + \epsilon]\nu_n\left(\bigcup_{\tau\in T_{t_i}} \tilde{A}_\tau\right).$$

Although $\bigcup_{\tau\in T_{t_i}} \tilde{A}_\tau$ is in general not closed, it is a subset of $\{x : x(t) \geq e(t) \text{ for some } t \in \overline{T_{t_i}}\}$. This set is closed by Lemma 9.4 for $M = e$ on $\overline{T_{t_i}}$ and $M = \infty$ on $[0,1]\backslash\overline{T_{t_i}}$. Therefore, by the large-deviation upper bound, Jensen's inequality, and (9.11),

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \nu_n\left(\bigcup_{\tau\in T_{t_i}} \tilde{A}_\tau\right) &\leq -\inf_{\{x:x(t)\geq e(t) \text{ for some } t\in\overline{T_{t_i}}\}} I(x) = -\inf_{t\in\overline{T_{t_i}}} g(t) \\ &\leq -g(t_i) + \epsilon.\end{aligned}$$

Combining the preceding three displays, we obtain

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \int_A \frac{d\nu_n}{d\mu_n} d\nu_n &\leq \max_{i=1}^N [f(t_i) - g(t_i)] + 2\epsilon \\ &\leq \sup_{t\in[0,1]} [f(t) - g(t)] + 2\epsilon.\end{aligned}$$

The sufficient condition follows by letting $\epsilon \to 0$.

The necessary condition. We now turn to the necessary condition. Since A is an I-continuity set and we suppose that step-level twisting is asymptotically efficient, we have

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_A \frac{d\nu_n}{d\mu_n} d\nu_n \leq -2\tilde{\tau}\Lambda_X^* \left(\frac{e(\tilde{\tau})}{\tilde{\tau}} \right). \tag{9.12}$$

Let $\epsilon > 0$. The upper semicontinuity of e implies that for all $t \in (0,1]$ there exists some $\delta \in (0,t)$ such that

$$\sup_{\tau\in(t-\delta,t]} e(\tau) \leq e(t) + \epsilon. \tag{9.13}$$

Fix $t \in (0,1]$, and define

$$A_t^{\delta,\epsilon} := \left\{ \begin{array}{ll} x: & x(\tau) < e(\tau) \text{ for } \tau \in [0, t-\delta]; x(t) > e(t); \\ & x(\tau) < \sup_{s\in(t-\delta,t]} e(s) + \epsilon \text{ for } \tau \in (t-\delta, t] \end{array} \right\}.$$

Note that $A_t^{\delta,\epsilon} \subset A$ and that it is open by the fact that A_m^M in Lemma 9.4 is closed. Indeed, set $m(t) = e(t)$ and $m = -\infty$ on $[0,1]\backslash\{t\}$; $M = e$ on $[0, t-\delta]$ and $M = \sup_{s\in(t-\delta,t]} e(s) + \epsilon$ on $(t-\delta, t]$.

We deduce that by definition of $A_t^{\delta,\epsilon}$,

$$\begin{aligned} \frac{1}{n} \log \int_A \frac{d\nu_n}{d\mu_n} d\nu_n &\geq \frac{1}{n} \log \int_{A_t^{\delta,\epsilon}} \frac{d\nu_n}{d\mu_n} d\nu_n \\ &\geq \frac{1}{n} \log \int_{A_t^{\delta,\epsilon}} \exp\left(-n\alpha \left[\sup_{\tau\in(t-\delta,t]} e(\tau) + \epsilon \right] + nt\Lambda_X(\alpha) \right) \nu_n(dx) \\ &\geq -\alpha[e(t) + 2\epsilon] + t\Lambda_X(\alpha) + \frac{1}{n} \log \nu_n(A_t^{\delta,\epsilon}), \end{aligned}$$

where we used (9.13) for the last inequality.

Recall the definition of γ_τ and γ_τ^ϵ in (9.3) and (9.5). Now two cases are distinguished.

Case 1: γ_t and e do not intersect before t.

Let t be such that γ_t and e do not intersect before t. Choose δ such that (9.13) is met, and set

$$\eta := \frac{1}{2} \min \left(\inf_{\tau\in[0,t-\delta]} \left[e(\tau) - \gamma_t(\tau) \right], \epsilon \right).$$

By the usual arguments, it is readily seen that $\eta > 0$ and $\gamma_t^\eta \in A_t^{\delta,\epsilon}$. Since $I(\gamma_t^\eta) = t\Lambda_X^*([e(t) + \eta]/t)$, we have by monotonicity of Λ_X^* on $[0,\infty)$ and the large-deviation lower bound,

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \log \int_A \frac{d\nu_n}{d\mu_n} d\nu_n &\geq f(t) - 2\alpha\epsilon - \inf_{x\in A_t^{\delta,\epsilon}} I(x) \\ &\geq f(t) - 2\alpha\epsilon - t\Lambda_X^*([e(t) + \eta]/t) \\ &\geq f(t) - 2\alpha\epsilon - t\Lambda_X^*([e(t) + \epsilon/2]/t). \end{aligned}$$

Since ϵ was arbitrary, we obtain a nontrivial lower bound if $e(t)/t \in (\operatorname{dom} \Lambda_X^*)^o$.

An auxiliary result. Before proceeding with the complementary case, we first prove an auxiliary result: asymptotic efficiency implies that for any $t \in (0,1]$ with $e(t)/t \in (\operatorname{dom} \Lambda_X^*)^o$,

$$\alpha \frac{e(t)}{t} - \Lambda_X(\alpha) + \Lambda_X^* \left(\frac{e(t)}{t} \right) \geq 0. \tag{9.14}$$

We work towards a contradiction by supposing that (9.14) is not satisfied for some $\hat{t}$ with $e(\hat{t})/\hat{t} \in (\operatorname{dom} \Lambda_X^*)^o$. Without loss of generality, we may suppose that $\gamma_{\hat{t}}$ does not intersect with e before $\hat{t}$. By the above derived lower bound for 'Case 1',

$$\begin{aligned} \limsup_{n\to\infty} \frac{1}{n} \log \int_A \frac{d\nu_n}{d\mu_n} d\nu_n &\geq \liminf_{n\to\infty} \frac{1}{n} \log \int_A \frac{d\nu_n}{d\mu_n} d\nu_n \\ &\geq f(\hat{t}) - \hat{t}\Lambda_X^* \left(\frac{e(\hat{t})}{\hat{t}} \right) \\ &> 0. \end{aligned}$$

Since $-2\tilde{\tau}\Lambda_X^*(e(\tilde{\tau})/\tilde{\tau}) \leq 0$, this contradicts the assumption that step-level twisting is asymptotically efficient.

Case 2: γ_t intersects e before t. We now suppose that γ_t intersects e before t, and the first time that this occurs is denoted by $\bar{t} < t$. Use $e(t)/t = e(\bar{t})/\bar{t}$ and the 'auxiliary result' to see that

$$\begin{aligned} -f(t) + t\Lambda_X^* \left(\frac{e(t)}{t} \right) &= t \left[\alpha \frac{e(t)}{t} - \Lambda_X(\alpha) + \Lambda_X^* \left(\frac{e(t)}{t} \right) \right] \\ &\geq \bar{t} \left[\alpha \frac{e(\bar{t})}{\bar{t}} - \Lambda_X(\alpha) + \Lambda_X^* \left(\frac{e(\bar{t})}{\bar{t}} \right) \right] \\ &= -f(\bar{t}) + \bar{t}\Lambda_X^* \left(\frac{e(\bar{t})}{\bar{t}} \right). \end{aligned}$$

Hence, the infimum in (9.10) is not attained by t for which γ_t intersects with e before t.

Therefore, if step-level twisting is asymptotically efficient, we must have by (9.12)

$$\begin{aligned} \inf_{\{t\in(0,1]:e(t)/t\in(\operatorname{dom}\Lambda_X^*)^o\}} \left[t\Lambda_X^* \left(\frac{e(t)}{t} \right) - f(t) \right] &\geq -\liminf_{n\to\infty} \frac{1}{n} \log \int_A \frac{d\nu_n}{d\mu_n} d\nu_n \\ &\geq -\limsup_{n\to\infty} \frac{1}{n} \log \int_A \frac{d\nu_n}{d\mu_n} d\nu_n \\ &\geq 2\tilde{\tau}\Lambda_X^* \left(\frac{e(\tilde{\tau})}{\tilde{\tau}} \right), \end{aligned}$$

which proves the claim. □

As a result of the sufficient condition in Theorem 9.8, step-level twisting is asymptotically efficient if the saddle-point inequality

$$\alpha e(t) - t\Lambda_X(\alpha) \geq \tilde{\tau}\Lambda_X^*(e(\tilde{\tau})/\tilde{\tau})$$

holds for all $t \in [0, 1]$. This is Sadowsky's sufficient condition [279].

9.5 A comparison

In Theorems 9.3 and 9.8, we have provided necessary and sufficient conditions for asymptotic efficiency of path-level twisting and step-level twisting respectively. It is our present aim to compare these conditions, and we start by showing that the conditions must be different.

Consider the example given on page 119, in which $e(\tau) = 1 + |\tau - 1/2|$. We saw already that $\tilde{\tau} = 1/2$ and $\alpha = 2$. The infimum on the left-hand side of (9.9) is attained at $\tau = 1/2$, implying that step-level twisting is asymptotically efficient. Note that path-level twisting was not asymptotically efficient.

This raises the question how the conditions for the two methods are related.

Corollary 9.9 *Condition (9.6) for path-level efficiency implies both the sufficient condition (9.9) and the necessary condition (9.10) for step-level efficiency.*

Proof. Since the sufficient condition (9.9) implies the necessary condition (9.10), it suffices to compare (9.6) and (9.9). The first step is to note that $\tilde{\tau}\Lambda_X^*\left(\frac{e(\tilde{\tau})}{\tilde{\tau}}\right) > 0$ and that

$$\alpha e(\tilde{\tau}) - \tilde{\tau}\Lambda_X(\alpha) = \tilde{\tau}\Lambda_X^*\left(\frac{e(\tilde{\tau})}{\tilde{\tau}}\right),$$

so that (9.9) is equivalent to

$$\begin{aligned}
&\tilde{\tau}\Lambda_X^*\left(\frac{e(\tilde{\tau})}{\tilde{\tau}}\right) + \alpha e(\tilde{\tau}) \\
&\quad\leq \quad \min\left\{\inf_{\tau\in(0,\tilde{\tau})}\left(\tau\Lambda_X^*\left(\frac{e(\tau)}{\tau}\right) + \alpha e(\tau) + (\tilde{\tau}-\tau)\Lambda_X(\alpha)\right),\right. \\
&\qquad\qquad \left.\inf_{\{\tau\in(\tilde{\tau},1]:\,\alpha e(\tau)-\tau\Lambda_X(\alpha)\geq 0\}}\left(\tau\Lambda_X^*\left(\frac{e(\tau)}{\tau}\right) + \alpha e(\tau) - (\tau-\tilde{\tau})\Lambda_X(\alpha)\right)\right\}.
\end{aligned} \tag{9.15}$$

We now prove that the right-hand side of (9.6) cannot exceed the right-hand side of (9.15).

Clearly, the last term in the minimum of (9.15) cannot be smaller than

$$\inf_{\tau\in(\tilde{\tau},1]}\left(\tau\Lambda_X^*\left(\frac{e(\tau)}{\tau}\right) + \frac{\tilde{\tau}}{\tau}\big[\alpha e(\tau) - \tau\Lambda_X(\alpha)\big]\right) + \tilde{\tau}\Lambda_X(\alpha).$$

Since $\Lambda_X(\alpha) \geq 0$ as a result of the fact that X_1 has zero mean, this immediately yields that the right-hand side of (9.15) cannot be smaller than

$$\min\left\{\inf_{\tau\in(0,\tilde{\tau}]}\left(\tau\Lambda_X^*\left(\frac{e(\tau)}{\tau}\right) + \alpha e(\tau)\right), \inf_{\tau\in(\tilde{\tau},1]}\left(\tau\Lambda_X^*\left(\frac{e(\tau)}{\tau}\right) + \alpha\tilde{\tau}\frac{e(\tau)}{\tau}\right)\right\}.$$

To see that the right-hand side of (9.6) does not exceed this quantity, choose $e(\tau)$ and $\tilde{\tau}e(\tau)/\tau$ for the first and second β in (9.6) respectively. □

CHAPTER 10

Simulation of a Gaussian queue

In this chapter, we study a queue fed by a large number n of independent discrete-time Gaussian processes with stationary increments. We consider the *many-sources* asymptotic regime, in which the drain rate is also scaled with n.

We discuss four methods for simulating the steady-state probability that the buffer content exceeds $nb > 0$: the single-twist method (suggested by large-deviation theory), the cut-and-twist method (simulating timeslot by timeslot), the random-twist method (the twist is sampled from a discrete distribution), and the sequential-twist method (simulating source by source).

The asymptotic efficiency of these four methods is analytically investigated for $n \to \infty$. A necessary and sufficient condition is derived for the efficiency of the single-twist method, indicating that it is nearly always asymptotically inefficient. The other three methods are asymptotically efficient. We numerically evaluate the four methods by performing a detailed simulation study, where it is our main objective to compare the three efficient methods in practical situations.

10.1 Introduction

In Chapter 8, we have studied conditions for asymptotic efficiency on a rather abstract level. It is the main objective of the present chapter to apply these conditions to a concrete setting, which is closely related to Part A of this thesis. That is, we consider a fluid queue with stationary Gaussian input, meaning that the input process is Gaussian and that it has stationary increments. In Section 1.3.2, we have argued why Gaussian fluid queues are particularly interesting. The present chapter focuses on fast simulation techniques to estimate (the tail of) the steady-state buffer-content distribution. A key element in our analysis is the equality in distribution of the steady-state buffer content and the maximum of the free process.

It is notoriously hard to calculate the full buffer-content distribution of a queue with Gaussian input, but some limiting regimes allow explicit analysis; we refer to Part A for a detailed investigation of Gaussian queues in the *large-buffer regime*. The present chapter focuses on the so-called *many-sources regime*, which has been generally accepted as a framework that is particularly suitable for studying large multiplexing systems. In the many-sources regime, we suppose that there are n i.i.d. Gaussian input sources, and that the drain rate is also scaled with n, i.e., it equals nr for some $r > 0$. The probability that the steady-state buffer content exceeds level nb becomes small when n grows large. For fixed but large n, we study this *buffer-content probability* p_n in a discrete-time model.

Likhanov and Mazumdar [220] find the asymptotics of p_n, i.e., they identify a sequence $\{g_n\}$ such that $p_n g_n \to 1$ as $n \to \infty$; notably, they find that p_n decays roughly exponentially in n. Based on these asymptotics, one could estimate p_n by $1/g_n$. However, due to the lack of error bounds one does not know *a priori* whether these estimates are any good. In fact, the derivation of $p_n g_n \to 1$ indicates that $1/g_n$ has the undesirable property that it tends to underestimate p_n, cf. Equations (2.1) and (3.4) in [220].

In the absence of analytical results (or asymptotic results that are backed up by error bounds), one could resort to simulation. When simulating loss probabilities in queues with Gaussian input, essentially two problems arise. The first is that it is not straightforward to quickly simulate Gaussian processes, see for instance [110]. Although 'exact' methods for generating (discrete versions of) Gaussian processes are in general quite slow, a sophisticated simulation technique becomes available by exploiting the stationarity of the input [83]. In the important case of fBm, this leads to a fast algorithm (order of $T \log T$ for a trace of length T) for generating fBm traces. An inherent difficulty with this algorithm is that the trace length should be specified before the simulation is started.

The second problem of simulation is that it is typically hard to estimate small probabilities; we mainly focus on this issue in this chapter, since in our setting $p_n \to 0$ as $n \to \infty$. We use importance sampling to resolve this problem; however, as explained in Section 7.1, the main difficulty with this technique is that a good importance-sampling family needs to be selected. It is desirable that the importance-sampling estimator be *asymptotically efficient*.

Estimators based on large-deviation theory are natural candidates for efficient simulation, but we have seen in Chapter 8 that the resulting estimators are not always asymptotically efficient. In this chapter, we state the conditions for asymptotic efficiency (as $n \to \infty$) of the large-deviation estimator that would apply to our buffer-content probability. It turns out that this estimator is predominantly asymptotically *inefficient* for a wide range of Gaussian input processes, including fBm and (perhaps surprisingly) even standard Brownian motion. Since we essentially work under path-level twisting in the terminology of Chapter 9, it is interesting to compare this assertion to Corollary 9.9.

As the large-deviation estimator is inefficient in practice, a different approach needs to be taken. We present three other methods that can be proven to be asymptotically efficient. The first uses ideas of Boots and Mandjes [53], and simulates timeslot by timeslot. The second method is a randomized version of the large-deviation estimator; it is based on the work of Sadowsky and Bucklew [280]. A third method relies on a recent paper by Dupuis and Wang [127], and simulates source by source. In the latter approach, the change of measure of the source under consideration depends on the input generated by the sources that have already been simulated. We present a detailed performance evaluation of the four resulting approaches, both analytically and empirically.

Some related results on fast simulation of queues with Gaussian input have been reported by Michna [227] and by Huang *et al.* [161]. Michna focuses on fBm input under the *large-buffer scaling* of Part A, but does not consider asymptotic efficiency of his simulation scheme. The study of Huang *et al.* also relates to the large-buffer asymptotic regime for fBm input. They empirically assess the variance reduction of their proposed change of measure, but do not formally derive properties of their estimator (such as asymptotic efficiency); in fact, Lemma 10.5 below entails that their estimator is unnatural from the point of view of asymptotic efficiency. We would like to stress that the present chapter *only* focuses on the simulation of (the tail of) the buffer-content distribution in the many-sources regime (with general Gaussian input, not necessarily fBm).

This chapter is organized as follows. Section 10.2 formalizes the queueing framework used in this chapter. It also discusses how the simulation horizon can be truncated, so that we

can work with traces of prespecified length. Section 10.3 studies the asymptotic efficiency of the four simulation methods mentioned above from an analytical perspective. Section 10.4 contains a numerical evaluation of these methods, to assess their performance under practical circumstances. The chapter concludes with a discussion on continuous-time Gaussian fluid queues in Section 10.5.

10.2 The buffer-content probability

The present section contains the description of our queueing model. In particular, we argue in Section 10.2.1 that the simulation context is particularly suitable for studying fluid queues through extremes (see Section 1.1). Recall that this approach translates the buffer-content probability into an exceedance probability of the free process on an infinite time interval. To simulate this exceedance probability, this infinite time interval needs to be truncated, where the neglected probability mass is below a tolerable level. Under the truncation, we can work with Gaussian traces of prespecified length. This truncation issue is addressed in Sections 10.2.2 and 10.2.3.

10.2.1 Description of the model — many-sources framework

We focus on a discrete-time fluid queue in the spirit of Section 1.1.2; this framework is particularly suitable in a simulation context.

First we describe the input process. Consider n i.i.d. sources feeding into a buffered resource. The sources are assumed to be *stationary*, so that the distribution of the input ('traffic') generated in an interval $[s, s+t)$ only depends on the interval length t (and not on the 'position' s). Define A_n as the aggregate input process, i.e., $A_n(t)$ denotes the traffic generated by the superposition of the n sources in the interval $\{1, \ldots, t\}$. For notational convenience, we set $A_n(0) := 0$.

In this chapter, we assume that the sources are *Gaussian*, so that the distribution of $A_n(\cdot)$ is completely determined by the mean input rate and the covariance structure. Let μ denote the mean input rate of a single source, i.e., $\mathbb{E}A_n(t) =: n\mu t$. Because the stationarity of the sources results in stationary increments of the process A_n, the covariance structure is determined by the variance function $\sigma^2(t) := \mathbb{V}\mathrm{ar}A_1(t)$, see (3.3). We suppose that $\sigma^2(t)t^{-\alpha} \to 0$ as $t \to \infty$ for some $\alpha \in (0, 2)$; the Borel-Cantelli lemma then shows that $A_1(t)/t \to \mu$ almost surely, as seen in the proof of Lemma 4.13.

It is readily deduced that the covariance of $A_n(\cdot)$ is given by $\Gamma_n(s,t) = n\Gamma(s,t)$, where for $s, t \in \mathbb{N}$,

$$\Gamma(s,t) := \mathbb{C}\mathrm{ov}(A_1(s), A_1(t)) = \frac{\sigma^2(s) + \sigma^2(t) - \sigma^2(|s-t|)}{2}.$$

An important special case of Gaussian input is *fractional Brownian motion* (fBm), for which $\sigma^2(t)$ is (proportional to) t^{2H}.

Now consider a fluid queue with input process A_n. In this chapter, we scale the queue's (deterministic) drain rate with the number of sources: the queue releases fluid at rate nr. To ensure stability, we assume that $\mu < c$. We are interested in the steady-state probability p_n of the buffer content exceeding some prescribed level $nb > 0$, which also scales with the number of sources. As seen in Section 1.1, due to the assumed stationarity of the input, the steady-state probability p_n of the buffer content exceeding nb equals

$$p_n = \mathbb{P}\left(\sup_{t \in \{1,2,\ldots,\infty\}} A_n(t) - nrt > nb \right). \tag{10.1}$$

We conclude that there are essentially two ways of simulating the buffer-content probability:

- In the first place, one could estimate the buffer-content probability from the evolution of the *reflected process* (i.e., the buffer-content process). However, standard simulation approaches to do this have their intrinsic difficulties: a regenerative approach fails in the context of Gaussian inputs (note that busy periods are dependent!), whereas in a 'batch-means' approach the estimator could suffer from the relatively strong dependencies between the batches (particularly when the Gaussian process is long-range dependent).
- As an alternative, one could estimate the buffer-content probability from sample paths of the *free process* $A_n(t) - nrt$. Every run is an independent sample of this free process, and the corresponding estimator is the fraction of runs in which nb is exceeded (for some $t \in \mathbb{N}$). This approach clearly overcomes the aforementioned problems arising when estimating p_n from the reflected process.

A practical difficulty of the latter 'extreme-based' approach, however, relates to the infinite 'simulation horizon' involved (it needs to be verified whether the free process exceeds nb for some $t \in \mathbb{N}$); this issue is addressed in the next subsection. Motivated by these arguments, we use the representation of the buffer-content probability as an exceedance probability of the free process; in other words: we estimate p_n relying on the right-hand side of (10.1).

We remark that the probability p_n of the steady-state buffer content being larger than nb in a system with *infinite* buffer is often used as an approximation for the loss probability in a system with *finite* buffer nb. In some research papers, our buffer-content probability is therefore called the *overflow probability*.

The behavior of the probability p_n in discrete time is essentially different from continuous time. The buffer-content probability in continuous time is obtained by replacing $\mathbb{N}$ by $\mathbb{R}_+$ in Equation (10.1). Notably, the tail asymptotics for the buffer content in continuous time differ qualitatively from those of p_n, see Dębicki and Mandjes [90]. A further discussion of this issue is relegated to Section 10.5.

10.2.2 The simulation horizon

Representation (10.1) shows that the buffer-content probability equals an exceedance probability on an *infinite* time horizon. Hence, to estimate p_n through simulation, we first have to truncate $\mathbb{N}$ to $\{1, \ldots, T\}$, for some finite T, while still controlling the error made. That can be done as follows.

Suppose we approximate p_n by

$$p_n^T := \mathbb{P}\left(\sup_{t \in \{1,\ldots,T\}} A_n(t) - nrt > nb \right). \tag{10.2}$$

This is evidently a probability smaller than p_n, but the larger T the smaller the error. We now analyze how large T should be. Define $\tau_n := \inf\{t \in \mathbb{N} : A_n(t) - nrt > nb\}$, so that $p_n = \mathbb{P}(\tau_n < \infty)$. As we propose to approximate p_n by $\mathbb{P}(\tau_n \leq T)$, we discard the contribution of $\mathbb{P}(T < \tau_n < \infty)$. As in Boots and Mandjes [53], we choose T such that

$$\frac{\mathbb{P}(T < \tau_n < \infty)}{p_n} < \epsilon, \tag{10.3}$$

for some predefined, typically small, $\epsilon > 0$. When choosing ϵ small enough, the truncation is of minor impact. Clearly, the smaller ϵ, the larger the T required.

The requirement in (10.3) does not directly translate into an explicit expression for the simulation horizon T as a function of ϵ and n. Following [53], this problem is tackled by

establishing tractable bounds on $\mathbb{P}(T < \tau_n < \infty)$ and p_n: with a lower bound on p_n and an upper bound on $\mathbb{P}(T < \tau_n < \infty)$, we can choose T so large that $\mathbb{P}(T < \tau_n < \infty)/p_n < \epsilon$. We write

$$I_t := \frac{(b + (r-\mu)t)^2}{2\sigma^2(t)}.$$

A lower bound on p_n

For any $t \in \mathbb{N}$, application of (10.1) entails

$$\begin{aligned} p_n &\geq \mathbb{P}(A_n(t) > nb + nrt) \\ &= \int_{\sqrt{n}\frac{b+(r-\mu)t}{\sigma(t)}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx \\ &\geq \frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{nI_t} + \sqrt{nI_t + 2}} e^{-nI_t}, \end{aligned} \tag{10.4}$$

where the last inequality is a standard bound for the standard normal cumulative density function (see [231, p. 177–181] for related inequalities and references).

In order to find the best possible lower bound, we compute $t^* := \arg\inf_{t\in\mathbb{N}} I_t$ and use the lower bound (10.4) for $t = t^*$. The existence of t^* is guaranteed by the assumption that $\sigma^2(t)t^{-\alpha} \to 0$ as $t \to \infty$ for some $\alpha \in (0, 2)$. In case t^* is unique, it is usually referred to as the 'most likely' exceedance epoch: given that the free process $A_n(t) - nrt$ exceeds nb, it is most likely that it happens at epoch t^*; see for instance Wischik [308].

An upper bound on $\mathbb{P}(T < \tau_n < \infty)$

By a Chernoff-bound argument, we have

$$\mathbb{P}(T < \tau_n < \infty) = \sum_{t=T+1}^{\infty} \mathbb{P}(\tau_n = t) \leq \sum_{t=T+1}^{\infty} \mathbb{P}(A_n(t) - nrt > nb) \leq \sum_{t=T+1}^{\infty} e^{-nI_t}. \tag{10.5}$$

In the present generality, it is difficult to further bound this sum. We could proceed by focusing on a specific correlation structure, such as fBm, for which $\sigma^2(t) = t^{2H}$, for $H \in (0, 1)$. Instead, we focus on the somewhat more general situation that the variance function can be bounded (from above) by a polynomial: $\sigma^2(t) \leq Ct^{2H}$, for some $H \in (0, 1)$ and $C \in (0, \infty)$. For instance, if $\sigma^2(\cdot)$ is regularly varying (Definition 2.1) with index α, then $\sigma^2(t)$ can be bounded from above by $Ct^{\alpha+\delta}$, for some $C, \delta > 0$; see (2.2). Obviously, it is desirable to choose the horizon as small as possible under the restriction that (10.3) holds; for this, C and H should be chosen as small as possible.

Under $\sigma^2(t) \leq Ct^{2H}$ we can bound (10.5) as follows:

$$\sum_{t=T+1}^{\infty} e^{-nI_t} \leq \sum_{t=T+1}^{\infty} \exp\left(-n\frac{(r-\mu)^2}{2C}t^{2-2H}\right) \leq \int_T^{\infty} \exp\left(-n\frac{(r-\mu)^2}{2C}t^{2-2H}\right) dt.$$

It turns out that we have to consider the cases $H \leq 1/2$ and $H > 1/2$ separately. For $H \leq 1/2$, the following bound is readily found (its proof is deferred to Appendix 10.A.1). Set $C_0 := (r-\mu)^2/(2C)$ and $q := 1/(2-2H)$ for notational convenience.

Lemma 10.1 *In case $H \leq 1/2$, we have*

$$\int_T^{\infty} \exp\left(-nC_0t^{1/q}\right) dt \leq \frac{q}{C_0 n} \exp\left(-nC_0T^{1/q}\right). \tag{10.6}$$

We now focus on $H > 1/2$ (and hence $q > 1$). Let m be the largest natural number such that $q - 1 - m \in (0, 1]$. Moreover, we define

$$\gamma_q := q - 1 - m, \text{ and } \beta_q := \frac{(q-1)\cdots(q-m)}{\gamma_q^m e^{\gamma_q}}. \tag{10.7}$$

These quantities play a central role in the following lemma, which is also proven in Appendix 10.A.1.

Lemma 10.2 *In case $H > 1/2$, we have*

$$\int_T^\infty \exp\left(-nC_0 t^{1/q}\right) dt \le \frac{q\beta_q}{C_0^q(n-\gamma_q)} \exp\left(-(n-\gamma_q)C_0 T^{1/q}\right).$$

By combining the upper bounds and the lower bound, we derive the following corollary:

Corollary 10.3 *For $H \le 1/2$, let $T(n)$ be the smallest integer larger than*

$$\left(-\frac{1}{nC_0}\log\left[\frac{1}{q\sqrt{\pi}}\frac{nC_0\epsilon}{\sqrt{nI_{t^*}}+\sqrt{nI_{t^*}+2}}e^{-nI_{t^*}}\right]\right)^q,$$

and for $H > 1/2$ let $T(n)$ be the smallest integer larger than

$$\left(-\frac{1}{nC_0}\log\left[\frac{1}{q\beta_q\sqrt{\pi}}\frac{(n-\gamma_q)C_0^q\epsilon}{\sqrt{nI_{t^*}}+\sqrt{nI_{t^*}+2}}e^{-nI_{t^*}}\right]\right)^q.$$

Then the error as defined in (10.3) does not exceed ϵ.
Moreover, $\overline{T} := \lim_{n\to\infty} T(n) = (I_{t^*}/C_0)^{1/(2-2H)}$.

Recall that t^* could be interpreted as the most likely epoch at which the supremum in (10.1) is attained. Hence, it is not surprising that $\overline{T} > t^*$:

$$\frac{I_{t^*}}{C_0} = \frac{(b+(r-\mu)t^*)^2}{2\sigma^2(t^*)} \Big/ \frac{(r-\mu)^2}{2C} > (t^*)^{2-2H} = (t^*)^{1/q}. \tag{10.8}$$

10.2.3 Hurst parameter

In this subsection, we investigate the influence of the Hurst parameter on the simulation horizon. This is of special interest since the computational effort to obtain estimates with the cut-and-twist method (see Section 10.3.2) is extremely sensitive to this horizon.

As already observed, the limiting value (as $n \to \infty$) of the simulation horizon is given by $(I_{t^*}/C_0)^{1/(2-2H)}$, which equals by definition

$$T = T(H) = \left(\inf_{t\in\{1,2,\ldots\}} \frac{b+(r-\mu)t}{(r-\mu)t^H}\right)^{1/(1-H)}.$$

Assuming that the infimum is taken over the whole real halfline, we see that $T(H)$ can be approximated by

$$\tilde{T}(H) := \frac{b}{r-\mu}\frac{H^{H/(H-1)}}{1-H}.$$

Clearly, $\tilde{T}(H)$ has a pole at $H = 1$, but it is insightful to plot $\tilde{T}$ as a function of H and see how quickly it tends to infinity. Set $b/(r-\mu) = 1$. In Figure 10.1, we have plotted this function and its derivative.

It is intuitively clear that $\tilde{T}(H)$ increases in H. The higher H is, the more long-term correlations are present, and more time is needed until unusual behavior is diminished. In practice, it will hardly be possible to simulate the probability with relative error at most ϵ if $H > 0.95$, cf. (10.3).

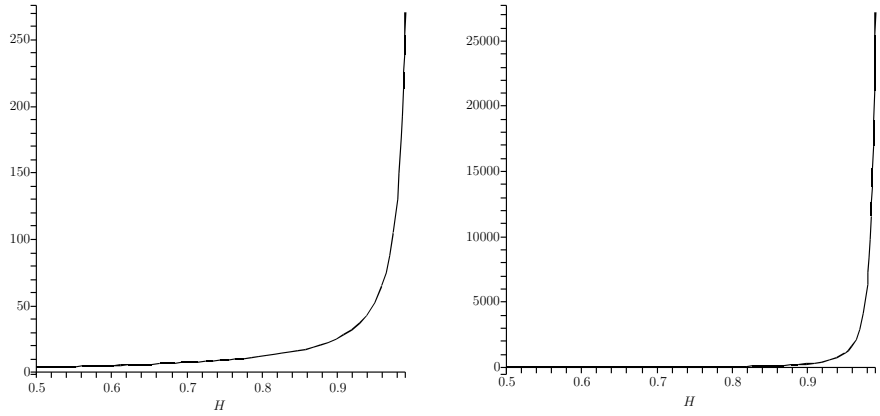

Figure 10.1: $\tilde{T}(H)$ as a function of H (left-hand diagram) and its derivative (right-hand diagram).

10.3 Simulation methods

Using the bounds of Section 10.2.2, the simulation horizon can be truncated. We therefore focus in the sequel of this chapter on the simulation of this 'truncated' buffer-content probability $p_n^{T(n)}$ defined in (10.2).

Recall from Section 7.1.2 that asymptotic efficiency corresponds to the performance of a simulation method for large n. Therefore, by virtue of Corollary 10.3, we can safely set $T(n) = \lceil \overline{T} \rceil$ for n large enough; for ease denote $T := \lceil \overline{T} \rceil$. Conclude that we can restrict ourselves to assess asymptotic efficiency of methods for estimating p_n^T.

In this chapter, we concentrate on four methods for simulating p_n^T. The first, which we refer to as the *single (exponential) twist* method, is the simplest of the four. This method corresponds to path-level twisting in the terminology of Chapter 9. We present explicit conditions on the covariance structure of the Gaussian sources under which the method is asymptotically efficient in the sense of Section 7.1.2. It appears that for important cases the method does *not* yield asymptotic efficiency. Therefore, we also discuss three asymptotically efficient alternatives: the first solves the theoretical difficulties by simulating timeslot by timeslot (which we therefore call *cut-and-twist*), the second by randomization of the twist (*random twist*), and the third by simulating source by source (*sequential twist*).

10.3.1 The single-twist method

Large-deviation theory suggests an importance-sampling distribution based on an exponential change of measure ('twist'). In a considerable number of simulation settings this alternative distribution has shown to perform well — in some cases it is asymptotically efficient, see Section 7.2 for references. However, one has to be careful, as a successful application of such an exponential twist critically depends on the specific problem at hand, see for instance Chapter 8 or [127, 148]. Before giving conditions for asymptotic optimality of the exponential twist in the setup of the present chapter, we first provide more background.

We denote

$$\mathcal{O}_T \quad := \quad \{x \in \mathbb{R}^T : x_t + \mu t \geq b + rt \text{ for some } t \in \{1, \ldots, T\}\} \tag{10.9}$$

$$= \bigcup_{t\in\{1,\ldots,T\}} \bigcup_{\{y:y+\mu t\geq b+rt\}} \{x\in\mathbb{R}^T : x_t = y\},$$

so that

$$p_n^T = \nu_n^{(T)}(\mathcal{O}_T),$$

with $\nu_n^{(T)}$ denoting the distribution of the centered (i.e., zero-mean) process $\{A_n(t)/n - \mu t : t = 1,\ldots,T\}$. The following lemma states that $\nu_n^{(T)}(\mathcal{O}_T)$ decays exponentially in n. We let $\Gamma^{(T)}$ denote the covariance matrix of $\{A_1(t) - \mu t : t = 1,\ldots,T\}$, i.e., $\Gamma^{(T)} := \{\Gamma(s,t) : s,t = 1,\ldots,T\}$. All proofs for this subsection are given in Appendix 10.A.2.

Lemma 10.4 *We have*

$$\lim_{n\to\infty} \frac{1}{n} \log \nu_n^{(T)}(\mathcal{O}_T) = -\frac{1}{2} x^{*\prime} \left(\Gamma^{(T)}\right)^{-1} x^* = -I_{t^*}, \tag{10.10}$$

where $t^* := \arg\inf_{t\in\mathbb{N}} I_t$, *and the vector* $x^* \in \mathbb{R}^T$ *is given by*

$$x_t^* := \frac{b + (r-\mu)t^*}{\sigma^2(t^*)} \Gamma(t^*, t). \tag{10.11}$$

The formula for x_t^* should be compared to (4.3) in the case of fractional Brownian motion. Time epoch t^* can be thought of as the most likely epoch that the free process $A_n(t) - nrt$ exceeds nb: as n grows the probability of exceeding nb vanishes, but given that it occurs, with overwhelming probability it occurs at t^*. Likewise, x^* can informally be interpreted as the *most likely path to exceedance*; note that indeed $x_{t^*}^* = b + (r-\mu)t^*$. Theorem 4.2 formalizes this intuition in the large-buffer asymptotic regime. It is important to realize that x^* is piecewise linear only in the case of (scaled) Brownian input (i.e., $\sigma^2(t) = Ct$ for some $C > 0$); in general x^* is a 'curved' path.

We can now introduce the family $\{\lambda_n^{(T)}\}$ of exponentially-twisted probability measures. The probability mass assigned to a Borel set $A \subset \mathbb{R}^T$ under this new distribution is

$$\lambda_n^{(T)}(A) = \int_A \exp\left(n \frac{b + (r-\mu)t^*}{\sigma^2(t^*)} x_{t^*} - nI_{t^*}\right) \nu_n^{(T)}(dx). \tag{10.12}$$

In the terminology of Chapter 9, this is a path-level twisted distribution. Note that, due to the possible presence of correlations, it is not clear how a step-level simulation procedure could be devised in the present context.

We next calculate the mean and covariance structure of vectors ('sample paths') drawn from $\lambda_n^{(T)}$. Observe that by definition of x^*, we have

$$(\Gamma^{(T)})^{-1} x^* = \frac{b + (r-\mu)t^*}{\sigma^2(t^*)} e_{t^*}, \tag{10.13}$$

where e_i is a T-dimensional vector of zeros, except for a one on the i-th position. As a result, cf. (10.10), we have $x^{*\prime} \left(\Gamma^{(T)}\right)^{-1} x^* = 2I_{t^*}$. Also, the density on $\mathbb{R}^T$ corresponding to $\lambda_n^{(T)}$ reduces to

$$\begin{aligned} &\exp\left(n \frac{b + (r-\mu)t^*}{\sigma^2(t^*)} x_{t^*} - nI_{t^*}\right) \frac{1}{(\sqrt{2\pi})^T \left|\Gamma^{(T)}/n\right|^{1/2}} \exp\left(-\frac{n}{2} x' \left(\Gamma^{(T)}\right)^{-1} x\right) \\ &= \frac{1}{(\sqrt{2\pi})^T \left|\Gamma^{(T)}/n\right|^{1/2}} \exp\left(-\frac{n}{2}(x-x^*)' \left(\Gamma^{(T)}\right)^{-1} (x-x^*)\right), \end{aligned}$$

which can immediately be seen by using (10.13) for calculating $x' \left(\Gamma^{(T)}\right)^{-1} x^*$. In other words: the new measure $\lambda_n^{(T)}$ corresponds to the distribution of a Gaussian process with mean vector

$\{x_t^* : t = 1, \ldots, T\}$ and covariance matrix $\Gamma^{(T)}/n$. Remark that the mean vector of the new measure is different from the old mean (in fact, the new Gaussian process does *not* correspond to stationary sources anymore), while the covariances under the old and new measure coincide. Since samples from $\lambda_n^{(T)}$ tend to follow the most likely path x^* for large n, we say that this exponential twist is in accordance with the large-deviation behavior of Lemma 10.4.

The above calculations show that, in the Gaussian setting, exponential twisting amounts to changing the mean vector, but not the covariance structure. Huang *et al.* [161] (see their Equation (16)) and Michna [227] propose to take a straight path as the mean vector, as opposed to the 'curved' most-likely path x^*. The following lemma shows, however, that x^* is in fact the 'best' way to change the mean (i.e., the only candidate that possibly yields asymptotic efficiency).

Lemma 10.5 *Any mean vector different from x^* does not yield asymptotic efficiency.*

Lemma 10.5 further motivates the verification of the asymptotic efficiency of the twisted distribution $\lambda_n^{(T)}$, and the following theorem is therefore the main result of this subsection. It presents sufficient and necessary conditions for asymptotic efficiency of the estimator determined by (7.1), where ϵ is replaced by n and $\lambda_n^{(T)}$ is given by (10.12).

We recently came across a related theorem by Baldi and Pacchiarotti [35]. An important difference is that these authors study the continuous-time buffer-content probability. We wish to remark, however, that our method can be extended to cover continuous time by applying standard theorems for large deviations of Gaussian measures on Banach spaces, see Section 3.1. However, we believe that discrete time is more natural in a simulation setting, and refer to Section 10.5 for a discussion. Another difference with [35] is the proof technique; Baldi and Pacchiarotti use recent insights into certain Gaussian martingales, while we take a direct approach based on Chapter 8.

Theorem 10.6 *Importance sampling under a 'single exponential twist' is asymptotically efficient for simulating p_n^T if and only if*

$$\inf_{t\in\{1,\ldots,T\}} \frac{b+(r-\mu)t+x_t^*}{\sigma(t)} = 2\,\frac{b+(r-\mu)t^*}{\sigma(t^*)}. \tag{10.14}$$

Clearly, if we set $h_t := [b+(r-\mu)t+x_t^*]/\sigma(t)$, then Theorem 10.6 states that the change of measure is asymptotically efficient if and only if $h_t \geq h_{t^*}$ for all $t = 1, \ldots, T$.

In the above we represented time by the natural numbers $\mathbb{N}$, i.e., we used a grid with a unit mesh. The same techniques can be used to prove a similar statement for any arbitrary simulation grid. In the following intermezzo, we analyze the impact of making the grid more fine-meshed.

Intermezzo: refining the simulation grid

Consider simulation on the grid $m\mathbb{N} \cap [0,T]$ for some grid mesh $m > 0$. One can repeat the proof of Theorem 10.6 to see that the infimum in (10.14) should then be taken over $m\mathbb{N} \cap [0,T]$. Thus, by refining the grid, the left-hand side of (10.14) can be made arbitrarily close to the infimum over $[0,T]$. This motivates an analysis of the function $g : \mathbb{R}_+ \to \mathbb{R}_+$ given by $g(t) := [b+(r-\mu)t+\bar{x}^*(t)]/\sigma(t)$, where $\bar{x}^*$ denotes the continuous-time analogue of (10.11):

$$\bar{x}^*(t) = \frac{b+(r-\mu)t^*}{2\sigma^2(t^*)}\left[\sigma^2(t^*)+\sigma^2(t)-\sigma^2(|t-t^*|)\right].$$

Hence, there is asymptotic optimality for any grid on $[0,T]$ if and only if $g(t) \geq g(t^*)$ for all $t \in [0,T]$. Suppose that σ^2 is twice continuously differentiable with first and second derivative

denoted by $\dot{\sigma}^2$ and $\ddot{\sigma}^2$ respectively. Necessary conditions for $\inf_{t\in[0,T]} g(t) \geq g(t^*)$ are then $\dot{g}(t^*) = 0$ and $\ddot{g}(t^*) > 0$. Therefore we compute

$$\lim_{t\uparrow t^*} \dot{g}(t) = \frac{1}{2}\frac{b+(r-\mu)t^*}{\sigma^3(t^*)}\dot{\sigma}^2(0),$$

so that $\dot{\sigma}^2(0) > 0$ implies that exponential twisting becomes asymptotically *inefficient* as the grid mesh m tends to zero. For the complementary case $\dot{\sigma}^2(0) = 0$, we can certainly find an 'inefficient' grid mesh if $\lim_{t\uparrow t^*} \ddot{g}(t) < 0$. After some calculations, one obtains

$$\lim_{t\uparrow t^*} \ddot{g}(t) = \frac{1}{4}\frac{b+(r-\mu)t^*}{\sigma^3(t^*)}\left[\frac{[\dot{\sigma}^2(t^*)]^2}{\sigma^2(t^*)} - \ddot{\sigma}^2(t^*) - \ddot{\sigma}^2(0)\right], \tag{10.15}$$

which is negative if $[\dot{\sigma}^2(t^*)]^2 < \sigma^2(t^*)[\ddot{\sigma}^2(t^*) + \ddot{\sigma}^2(0)]$.

Having these conditions at our disposal, we can study some specific cases and ask whether the single exponential twist becomes inefficient as the mesh tends to zero. For instance, suppose that the input traffic $A_1(t)$ is a fractional Brownian motion (fBm) with Hurst parameter $H \in (0,1)$, i.e., $\sigma^2(t) = t^{2H}$. Note that a special case is Brownian motion, which corresponds to $H = 1/2$. If $H \leq 1/2$, one has $\dot{\sigma}^2(0) > 0$ and a single exponential twist is therefore asymptotically inefficient for grid meshes m small enough, in line with the results of [35]. Moreover, if $H > 1/2$, it follows from (10.15) and $\ddot{\sigma}^2(0) = \infty$ that $\lim_{t\uparrow t^*} \ddot{g}(t) < 0$, so that we here have inefficiency as well.

From the above we also see that it could be that the exponential twist is asymptotically optimal for some grid mesh m, but loses the optimality at some finer threshold grid mesh m^*.

Intuition behind (in-)efficiency of exponential twist

Having seen that a single exponential twist can be asymptotically inefficient, one may wonder *why* this occurs. To this end, consider the likelihood term $d\nu_n^{(T)}/d\lambda_n^{(T)}$ following from (10.12):

$$\exp\left(-n\frac{b+(r-\mu)t^*}{\sigma^2(t^*)}\,x_{t^*} + nI_{t^*}\right) = \exp\left(-n\frac{b+(r-\mu)t^*}{\sigma^2(t^*)}\,(x_{t^*} - x^*_{t^*}) - nI_{t^*}\right),$$

where x_{t^*} corresponds to the value of $A_n(t^*)/n - \mu t^*$, with mean $x^*_{t^*} = b + (r-\mu)t^*$ under $\lambda_n^{(T)}$. For asymptotic optimality, this likelihood ratio should be 'small' for realizations in the set $\mathcal{O}_T$. If there is exceedance at time t^*, then clearly

$$\frac{d\nu_n^{(T)}}{d\lambda_n^{(T)}} \leq e^{-nI_{t^*}} \tag{10.16}$$

(use $x_{t^*} \geq b + (r-\mu)t^*$). However, if exceedance occurs at any other time epoch, the likelihood ratio can take any (positive) value. Obviously, an extremely high value has a dramatic effect on the variance of the estimator, but the probability of such an extreme value might be low. Summarizing, condition (10.14) gives a criterion to check whether high values for the likelihood are probable enough to affect (the exponential decay of) the variance of the estimator.

10.3.2 The cut-and-twist method

We have seen that the likelihood may explode while simulating p_n^T with a single exponential twist. This can be overcome by partitioning the event $\mathcal{O}_T$ into disjoint sub-events, and simulating these individually. To this end, write

$$p_n^T = \nu_n^{(T)}\left(\bigcup_{t\in\{1,\ldots,T\}} \mathcal{O}_T(t)\right) = \sum_{t\in\{1,\ldots,T\}} \nu_n^{(T)}(\mathcal{O}_T(t)),$$

where $\mathcal{O}_T(t)$ corresponds to the event that exceedance occurs *for the first time* at time t:

$$\mathcal{O}_T(t) := \{x \in \mathbb{R}^T : x_t + \mu t \geq b + rt; x_s + \mu s < b + rs \text{ for all } s \in \{1, \ldots, t-1\}\}.$$

Hence, the problem reduces to simulating T probabilities of the type $\nu_n^{(T)}(\mathcal{O}_T(t))$. This partitioning approach is also taken by Boots and Mandjes [53], where this idea was exploited for a queue fed by (discrete-time) ON/OFF sources.

The resulting simulation algorithm, to be called cut-and-twist method, works as follows. Define the exponentially twisted measure ${}^t\lambda_n^{(T)}$ as in (10.12), but with t instead of t^*, and estimate the probability $\nu_n^{(T)}(\mathcal{O}_T(t))$ with the importance-sampling distribution ${}^t\lambda_n^{(T)}$. An estimate of p_n^T is found by summing the estimates over $t \in \{1, \ldots, T\}$.

Before discussing the efficiency of this method, we note for the sake of clarity that the estimator equals

$$\frac{1}{N} \sum_{k=1}^{N} \sum_{t=1,\ldots,T} \mathbf{1}_{\left\{X_t^{(k)} \in \mathcal{O}_T(t)\right\}} \frac{d\nu_n^{(T)}}{d\lambda_n^{(T)}} \left(X_t^{(k)}\right), \tag{10.17}$$

where $X_t^{(1)}, \ldots, X_t^{(N)}$ is an i.i.d. sample from ${}^t\lambda_n^{(T)}$, and the samples $X_t^{(\cdot)}$, $t = 1, \ldots, T$ are also independent.

The following theorem is proven in Appendix 10.A.3. Its proof is based on the property that the method is such that, when estimating $\nu_n^{(T)}(\mathcal{O}_T(t))$, for any $x \in \mathcal{O}_T(t)$ the corresponding likelihood is uniformly bounded by $\exp(-nI_t)$, cf. (10.16).

Theorem 10.7 *The cut-and-twist method is asymptotically efficient for estimating p_n^T.*

This method is asymptotically optimal, but it has the obvious drawback that it may take a substantial amount of time to simulate the T probabilities individually.

Summarizing, in this approach the exceedance event is split into disjoint events that correspond to exceedance (for the first time) at time t. The main advantage of this splitting is that every of these individual events can be 'controlled' now (the corresponding likelihoods are even bounded, see the proof of Theorem 10.7), whereas the single-twist method suffers from the (potentially large) likelihoods that correspond to exceedance at time epochs different from the most likely time t^*, as was noted in Section 10.3.1. As a result, the single-twist method is *not* necessarily asymptotically efficient, while cut-and-twist *is*.

10.3.3 The random-twist method

An approach closely related to the cut-and-twist method was proposed by Sadowsky and Bucklew [280]. In their method, a *random* $\mathcal{T}$ is drawn in each simulation run according to some (arbitrary) distribution $Q = \{q_t : t = 1, \ldots, T\}$ with q_t *strictly positive* for any $t = 1, \ldots, T$; subsequently, one does a simulation run under the measure ${}^{\mathcal{T}}\lambda_n^{(T)}$ (as defined in Section 10.3.2).

The likelihood ratio becomes

$$\left[\sum_{t=1}^{T} q_t \exp\left(n \frac{b + (r-\mu)t}{\sigma^2(t)} x_t - nI_t\right)\right]^{-1}.$$

Note that this likelihood ratio depends on the whole path $\{x_t : t = 1, \ldots, T\}$, as opposed to the previous two methods.

The following result follows from Theorem 2(a) of Sadowsky and Bucklew [280].

Theorem 10.8 (Sadowsky-Bucklew) *The random-twist method is asymptotically efficient for estimating p_n^T.*

Remarkably, the asymptotic efficiency does not depend on the specific choice of the q_t, as long as they are strictly positive. Hence, a drawback of the method is that it is unclear how the distribution Q is best chosen. For instance, if Q is 'almost' degenerate in t^*, then the method is similar to the single-twist method; therefore, it may suffer in practice from the same problems as discussed in Section 10.3.2. The theorem indicates that this effect eventually vanishes (when n grows large), but this could require extremely large n.

It is not the aim of this chapter to investigate the impact of the choice of Q on the quality of the estimates; in the sequel, we suppose that the q_t correspond to a truncated Poisson distribution with mean t^*, i.e.,

$$q_t = \frac{(t^*)^t/t!}{\sum_{k=1}^{T}(t^*)^k/k!}, \quad t = 1, \ldots, T.$$

The reason for this choice is that the Poisson distribution is nicely spread around its mean value. In addition, it is straightforward to sample from a Poisson distribution, so that one can sample from Q with a simple acceptance-rejection procedure.

10.3.4 The sequential-twist method

Recently, Dupuis and Wang [127] introduced an intuitively appealing approach to rare-event simulation. We now give a brief description of their method in the setting of the present paper, although the method is known to work in a considerably more general setting. Consider a sequence $\bar{A}_1, \bar{A}_2, \ldots$ of centered i.i.d. random vectors in $\mathbb{R}^T$, where the $\bar{A}_j$ are distributed as $\{A_1(t) - \mu t : t = 1, \ldots, T\}$; as a consequence, the vectors $\bar{A}_j$ have distribution $\nu_1^{(T)}$. Note that p_n^T can be written as $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{A}_i \in \mathcal{O}_T\right)$, with $\mathcal{O}_T$ defined in (10.9), and hence

$$p_n^T = \int_{\{(x^{(1)},\ldots,x^{(n)}):\frac{1}{n}\sum_{i=1}^{n}x^{(i)}\in\mathcal{O}_T\}} \nu_1^{(T)}(dx^{(1)}) \cdots \nu_1^{(T)}(dx^{(n)}). \tag{10.18}$$

Instead of twisting ν_n^T as in the previous methods, the sequential-twist method twists *each copy* of $\nu_1^{(T)}$ (i.e., each source) in Equation (10.18) differently, exploiting the fact that the sources behave stochastically independently. Recall that exponential twisting for Gaussian random variables corresponds to shifts in the mean (and no change in the covariance structure).

This gives rise to the following sequential approach. Suppose $\bar{A}_1, \ldots, \bar{A}_j$ (i.e., source 1 up to j) are already generated, and we are about to twist the traffic generated by source $j+1$ (for $j = 0, \ldots, n-1$). We aim to find the 'cheapest' way to reach the exceedance set $\mathcal{O}_T$ *given* $\bar{A}_1, \ldots, \bar{A}_j$. Hence, we do not change the measure if already $\frac{1}{n}\sum_{i=1}^{j}\bar{A}_i \in \mathcal{O}_T$ (under this condition reaching $\mathcal{O}_T$ is not 'hard' anymore, as $\mathbb{E}\bar{A}_j(t) = 0$); otherwise we change the mean of the distribution of $\bar{A}_{j+1}$ to μ_{j+1} (recall that this is a vector in $\mathbb{R}^T$), where

$$\mu_{j+1} := \operatorname*{arg\,inf}_{\left\{y\in\mathbb{R}^T:\frac{1}{n}\sum_{i=1}^{j}\bar{A}_i+\frac{1}{n}\sum_{i=j+1}^{n}y\in\mathcal{O}_T\right\}} y'\left(\Gamma^{(T)}\right)^{-1}y;$$

an empty sum should be interpreted as zero. The following lemma gives a useful explicit expression for μ_{j+1}. The proof is given in Appendix 10.A.4.

Lemma 10.9 *Define for* $j = 0, \ldots, n-1$,

$$t_{j+1}^* := \operatorname*{arg\,inf}_{t\in\{1,\ldots,T\}} \frac{nb + n(r-\mu)t - \sum_{i=1}^{j}\bar{A}_i(t)}{(n-j)\sigma(t)}, \tag{10.19}$$

and denote the corresponding infimum by J_{j+1}. *Then we have*

$$\mu_{j+1} = \frac{J_{j+1}}{\sigma(t_{j+1}^*)}\Gamma(\cdot, t_{j+1}^*).$$

Observe that for $j = 0$ the formula reduces to the large-deviation most likely path, which is to be expected since then no information is available on the previously generated sources. The reader may check that the resulting likelihood ratio is

$$\prod_{j=1}^{n} \exp\left(-\frac{J_j}{\sigma(t_j^*)}\bar{A}_j(t_j^*) + \frac{1}{2}J_j^2\right).$$

An estimator is obtained by performing N independent runs, and computing the estimate using (7.1).

The conditions for the following theorem of Dupuis and Wang [127] are checked in Appendix 10.A.4.

Theorem 10.10 (Dupuis-Wang) *The sequential-twist method is asymptotically efficient for estimating* p_n^T.

Informally speaking, the idea behind the sequential-twist approach is that, by adapting the mean μ_j of every next source j in the way described above, the set $\mathcal{O}_T$ is reached close to its most likely point, thus avoiding large likelihood ratios. Apparently, as claimed in the above theorem, the resulting estimator is asymptotically optimal.

A drawback of this approach is that all sources should be generated individually; one does not simulate the aggregate input process, as in the previous methods. However, the sequential approach can also be used with less than n Gaussian vectors while retaining the property of asymptotic efficiency. This is done by twisting source *batches* instead of individual sources. Let M be a *batch size* such that $n/M \in \mathbb{N}$, where M does not depend on n. Define $\bar{A}_i^{(M)} := \frac{1}{M}\sum_{j=1}^{M} \bar{A}_{j+(i-1)M}$. It is important that M does not depend on n. We refer to this approach as the *batch sequential-twist method*; since

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\bar{A}_i \in \mathcal{O}_T\right) = \mathbb{P}\left(\frac{1}{n/M}\sum_{i=1}^{n/M}\bar{A}_i^{(M)} \in \mathcal{O}_T\right),$$

Theorem 10.10 also yields the asymptotic efficiency of the batch sequential-twist estimator for any fixed M.

Although the sequential-twist method and its batch counterpart are both asymptotically efficient, some practical issues arise when M is made (too) large. The relative efficiency then converges much slower to 2, so that we might not even be close to efficiency for reasonable n. This issue is addressed empirically in Section 10.4.4.

10.4 Evaluation

In this section, we evaluate the four methods of Section 10.3 as follows. First, we discuss some issues related to our implementation of the methods. Based on this, we come to preliminary conclusions on the time complexity of each of the methods. In Section 10.4.2, we check empirically that our simulations support the claims of Theorems 10.6, 10.7, 10.8, and 10.10. After this analysis, the reliability of the methods is studied by refining the simulation grid; for this, we also take the computational effort into account. Further empirical insight into the batch sequential-twist method is gained by studying the influence of the batch size on the relative efficiency and the relative error.

While the preceding sections are applicable to general Gaussian processes with stationary increments (satisfying certain conditions), in this section we focus on the important case of fractional Brownian motion.

10.4.1 Implementation and time complexity

Simulation of fractional Brownian motion is highly nontrivial. As the simulation grid is equispaced, it is best to simulate the (stationary!) incremental process, often called fractional Gaussian noise. When T is a power of two, the fastest available algorithm for simulating T points of a fractional Gaussian noise is the method of Davies and Harte [83]. In this approach, the covariance matrix is embedded in a so-called circulant matrix, for which the eigenvalues can easily be computed. The Fast Fourier Transform (FFT) is then used for maximum efficiency; the computational effort is of order $T \log T$ for a sample size of length T. For more details on this method, we refer to Dietrich and Newsam [115] and Wood and Chan [309].

Although we use this fast algorithm, the required number of traces per simulation run still highly influences the speed of the methods. The single-twist method and the random-twist method only need one trace (of length T) for each simulation run, while n such traces are needed for the sequential-twist method. For the cut-and-twist method, traces of length $t = 1, \ldots, T$ are needed. These considerations indicate that it depends on the parameter values which method performs best.

We first address the impact of the number of sources n. A clear advantage of the cut-and-twist method and the random-twist method is that the required simulation time depends just mildly on n (due to the fact that $T(n) \to T$), as opposed to sequential-twist (where the simulation time is roughly proportional to n).

As for the influence on the simulation horizon, we have already observed that T is large when either $b/(r-\mu)$ or H is large, see Section 10.2.3. This badly affects the cut-and-twist method, since such a sample is needed for each time epoch (of which there are T). The random-twist method only needs a single fBm trace, but the computation of the likelihood ratio is of the order T. Sequential-twist calculates the best twist n times, which amounts to computing the infimum in (10.19); this computation is of order T.

We raise one further implementation issue, which plays a role for all of the methods. Once we have calculated the simulation horizon T, we round it off to the smallest power of two T' with $T' \geq T$, *and we use this new horizon* T'. Similarly, since traces of length $t = 1, \ldots, T'$ are needed for the cut-and-twist method, t is rounded off, for every t.

10.4.2 Empirical validation of the theory

In Section 10.3, we studied whether the four discussed simulation methods are asymptotically efficient. In the present subsection, our aim is to validate these theoretical results by performing simulation experiments. By doing so, we gain insight into the quality of the methods.

The parameters are chosen as follows: $b = 0.3$, $r - \mu = 0.1$, $H = 0.8$, $M = 1$, $\epsilon = 0.05$, and $\eta_{\max} = 0.1/1.96$. Recall from Section 7.1.2 that the simulation is stopped when the relative error drops below $\eta_{\max}$. It is left to the reader to check that condition (10.14) does not hold, i.e., that the single-twist estimator is not asymptotically efficient. The choice $H = 0.8$ is supported by several measurement studies, see for instance Leland *et al.* [217], and $\eta_{\max}$ is chosen such that the width of the confidence interval is 20% of the estimated probability. Reduction of this value has a significant impact on the simulation time, and the present value yields typical results within a reasonable time frame.

We first study the asymptotic efficiency of the simulation methods by varying n and analyzing the number of simulation runs N_n^* needed to achieve the required relative error. In the left-hand diagram of Figure 10.2, we have plotted $\log N_n^*$ for $n = 100, 150, \ldots, 500$ and all four simulation methods, and in addition the 'naive' direct Monte Carlo estimator. The confidence intervals are not plotted, since they are completely determined by the estimates themselves and the value of $\eta_{\max}$. Note that under asymptotic efficiency, $\log N_n^*$ should be (ultimately) sublinear. Therefore, the plot supports Theorem 10.7, Theorem 10.10, and the fact

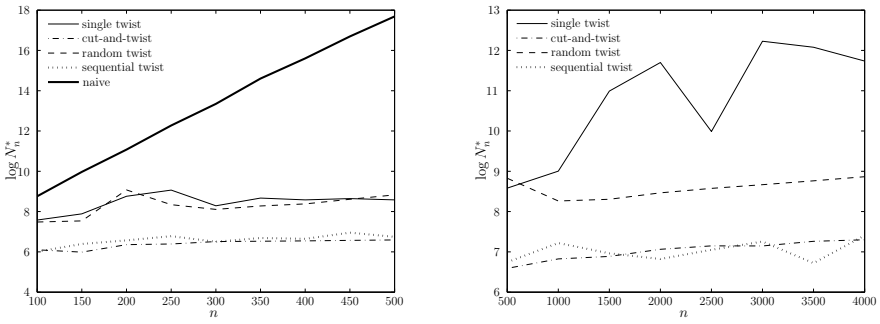

Figure 10.2: Empirical verification of the asymptotic efficiency of the simulation methods.

	$n = 300$	rel. eff.	$n = 1000$	rel. eff.
naive	6.12×10^{-4}	—	—	—
single twist	4.84×10^{-4}	1.68	1.03×10^{-10}	1.87
cut-and-twist	5.95×10^{-4}	1.86	1.32×10^{-10}	1.94
random twist	5.50×10^{-4}	1.70	1.38×10^{-10}	1.89
sequential twist	6.39×10^{-4}	1.86	1.41×10^{-10}	1.93
'exact'	5.8×10^{-4}		1.38×10^{-10}	

Table 10.1: Two of the estimates corresponding to Figure 10.2.

that the naive estimator is inefficient (in fact, the number of runs grows exponentially, in line with p_n decaying exponentially). However, it is not immediate from the left-hand diagram of Figure 10.2 that single twist is asymptotically inefficient (cf. Theorem 10.6), and that random twist is asymptotically efficient (cf. Theorem 10.8). Although the irregular behavior indicates that this might indeed be the case, we find more convincing evidence by increasing n further. This is done in the right-hand diagram of Figure 10.2.

It is interesting to see some of the estimated probabilities that correspond to Figure 10.2. We give these for two different values of n in Table 10.1. To obtain a benchmark, we also performed a very long simulation, see the row labeled 'exact'. It was obtained with the cut-and-twist method, where the simulation is stopped as soon as both ends of the confidence interval give the same value when rounded off to one digit for $n = 300$, and to two digits for $n = 1000$.

The unstable behavior of the single-twist method (also reflected in a low value of the relative efficiency in Table 10.1) has been explained theoretically through the interpretation of a possible failure of the exponential twist, see Section 10.3. As noted there, the supremum is attained at time epoch t^* in a 'typical' simulation run, but it might also happen at some other epoch $t \neq t^*$. Although such a realization is (relatively) rare, it has an impact on both the estimate and the estimated variance. Since these two estimated quantities determine whether the simulation is stopped, it may occur that the number of these 'rare' realizations is too low, so that the simulation is stopped too early and the buffer-content probability is underestimated. Likewise, random twist can lead to underestimation, but this effect vanishes when n grows large, in line with the theory of Section 10.3.3. The table shows that random twist has a low relative efficiency for small n.

It is interesting to see that the sequential-twist method and cut-and-twist method have comparable performance, both in terms of relative efficiency (Table 10.1) and number of simulation

runs (Figure 10.2).

Hence, cut-and-twist and sequential twist seem to perform best, although the random-twist method improves considerably as n grows. However, these methods are also the slowest (i.e., the effort per experiment is highest). To obtain a more realistic comparison, one should consider CPU time, rather than the number of experiments. This is done in the next subsection.

10.4.3 Simulation grid

While the observations in the previous subsection were predicted by theory, we now perform experiments that relate to the CPU time needed, for which no theory is available. This analysis provides further insight into the performance of the methods in practice, both in terms of reliability and speed. As a first step, we investigate the influence of the grid mesh on the estimated probability.

We evaluate, with again $\bar{A}_n \in \mathbb{R}^T$ denoting the centered version of A_n,

$$ {}^{\alpha}p_n := \mathbb{P}\left(\sup_{t\in\{\alpha,2\alpha,\ldots\}} \bar{A}_n(t) - n(r-\mu)t > nb \right) \tag{10.20}$$

for a range of $\alpha \geq 0$, in such a way that the simulation grid becomes finer. For instance, one can take $\alpha = 1, 1/2, 1/4, 1/8$; ${}^{\alpha}p_n$ then increases as α is made smaller, as we only add grid points, and hence the supremum of the free process becomes larger. Therefore, as a sanity check, we can test the reliability of the simulation methods by checking whether the estimates indeed increase when making the grid finer.

Before we can compare the estimated probabilities for different α, we first study the impact of α on the simulation horizon. We denote this simulation horizon, as a function of α, by T^{α}. We now verify whether also ${}^{\alpha}p_n^{T^{\alpha}}$ (defined in a self-evident manner) should increase when decreasing α. Since $\bar{A}_n(t)$ is a centered fractional Brownian motion by assumption, $\bar{A}_n(\alpha t)$ has the same distribution as $\alpha^H \bar{A}_n(t)$, see (1.12). This self-similarity property yields that (10.20) equals V

$$ \mathbb{P}\left(\sup_{t\in\mathbb{N}} \alpha^H \bar{A}_n(t) - n\alpha(r-\mu)t > nb \right) = \mathbb{P}\left(\sup_{t\in\mathbb{N}} \bar{A}_n(t) - n\alpha^{1-H}(r-\mu)t > n\alpha^{-H}b \right). $$

The above equation entails that a grid mesh α is equivalent to a unit grid mesh if b and $r-\mu$ are replaced by $b_\alpha := \alpha^{-H}b$ and $c_\alpha := \alpha^{1-H}(r-\mu)$. Note that then

$$ I^{\alpha}_{t^*} := \inf_{t\in\{\alpha,2\alpha,\ldots\}} \frac{(b+(r-\mu)t)^2}{2t^{2H}} = \inf_{t\in\mathbb{N}} \frac{(b+\alpha(r-\mu)t)^2}{2\alpha^{2H}t^{2H}}, $$

so that the (limiting) simulation horizon then becomes, see (10.8),

$$ T^{\alpha} = \frac{I^{\alpha}_{t^*}}{c_\alpha^2/2} = \inf_{t\in\mathbb{N}} \frac{(b/\alpha+(r-\mu)t)^2}{c^2t^{2H}}, $$

which is monotonic in α and tends to infinity as $\alpha \downarrow 0$. We conclude that the monotonicity is preserved: ${}^{\alpha}p_n^{T^{\alpha}}$ increases when $\alpha \downarrow 0$, just like ${}^{\alpha}p_n$ does.

In order to investigate whether the estimates indeed decrease in α, we perform some simulations with parameters $n = 150$, $b = 0.9$, $r-\mu = 0.3$, $H = 0.8$, $M = 1$, and $\epsilon = 0.05$. To obtain each of the estimates, we stop the simulation after exactly five minutes of CPU time. It would be desirable to do the simulations for grid sizes $2^0, 2^1, 2^2, 2^3, 2^4, \ldots$, but this quickly becomes computationally too intensive. Therefore, we focus on four sets of grids; $1/\alpha = 1, 2, 4, 8$, $1/\alpha = 3, 6, 12$, $1/\alpha = 3, 9$, and $1/\alpha = 5, 10$.

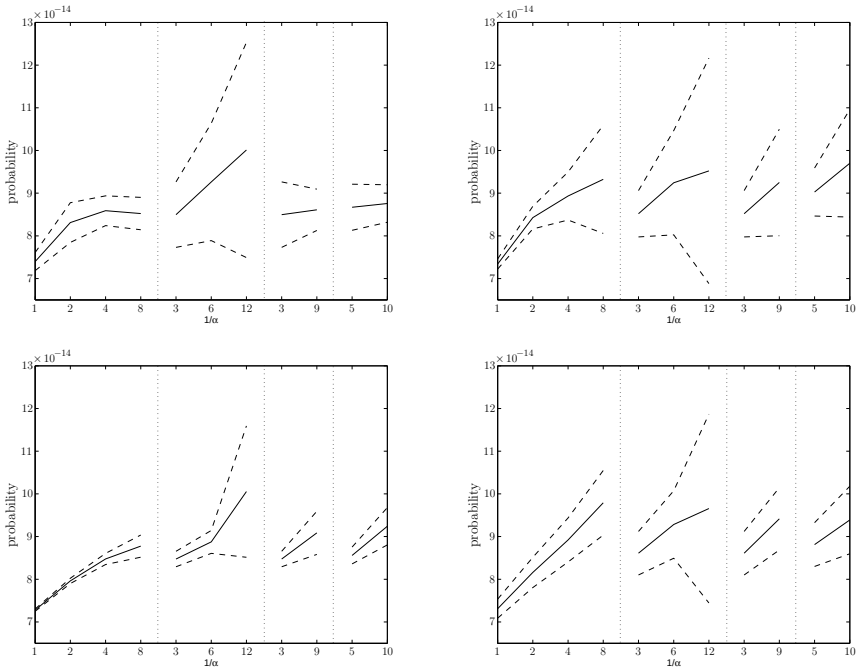

Figure 10.3: The influence of the grid mesh on the probability for single twist, cut-and-twist, random twist, and sequential twist. The solid lines represent the estimates, while the dashed lines correspond to confidence intervals.

In Figure 10.3, we have plotted these four sets using the four different methods. The dotted lines correspond to the boundaries of the confidence intervals. Roughly speaking, each of the plots shows the expected monotonicity, with the only exception of the single-twist method. This is in line with Theorem 10.6. The behavior of the single-twist confidence intervals also differs from the other methods, but it is interesting to compare these intervals for the other three (efficient) methods.

The widths of the confidence intervals do not seem to grow proportionally to the estimates, most prominently for $1/\alpha = 12$. Although the stopping criterion (CPU time) is proportional to the number of runs, the CPU time per run varies, since the time horizon T depends on α. For instance, fBm traces of length 1024 are generated if $\alpha = 1/10$, while this increases to 2048 for $\alpha = 1/12$. This is reflected in the plots (especially in the cut-and-twist plot, as anticipated in Section 10.4.1). Clearly, the random-twist confidence intervals are the smallest, but we have to keep in mind that the probabilities may be underestimated in view of the previous subsection.

10.4.4 Batch size for the sequential-twist method

The aim of the present subsection is to investigate the influence of the parameter M in the batch sequential-twist method. That is, the n sources are divided into batches of size M and each of the batches is considered a single source (using the fact that the sum of independent Gaussian vectors is again Gaussian). Then one only twists n/M times in one simulation run, which limits the flexibility of the method (i.e., the probability measure is adapted less often) and therefore makes it less efficient as M grows. Note that Theorem 10.10 states that sequential twist is still *asymptotically* efficient as $n \to \infty$, regardless the value of M. In practice, however, there is a trade-off between the batch size M and the efficiency.

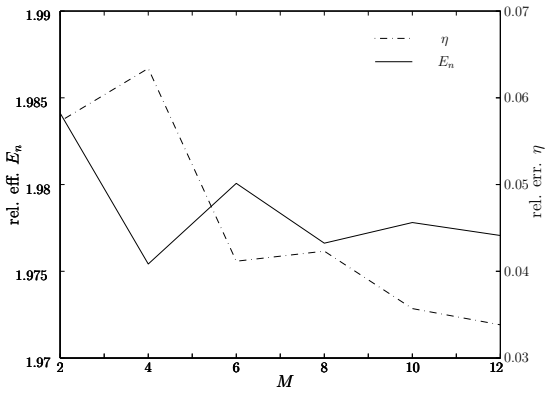

Figure 10.4: The relative efficiency as a function of M for small M.

If we let the simulation run for a specified time period (here five minutes), there are two effects as M increases: the efficiency decreases, but the relative error also decreases. It is the aim of this subsection to study these two opposite effects. The values of the parameters are the same as in Section 10.4.2, except for the value of M, which now varies.

We measure efficiency by means of the (estimated) relative efficiency. We set $n = 3840$ and estimate the relative efficiency for $M = 2, 4, 6, 8, 10, 12$. The resulting plot is given in Figure 10.4. From the plot, it is not so clear that an increase in M makes the simulation less efficient, although the relative error seems to decrease. Therefore, we also investigate what happens if $M = 80, 160, 240, 320, 480, 640, 960$; the relative efficiency (relative error) is then estimated as 1.972 (0.0164), 1.969 (0.0135), 1.966 (0.0125), 1.960 (0.0140), 1.956 (0.0139), 1.953 (0.0137), and 1.950 (0.0127) respectively. These values indeed suggest that the simulation becomes less efficient as M increases, while the relative error decreases.

Although the differences in the relative efficiency look small, one must keep in mind that this quantity relates to the *exponential* decay rate of the variance of the estimator. Therefore, small differences blow up exponentially, and we propose to always choose M as small as possible.

10.5 Concluding remarks

In this section, we explain why the buffer-content probability in discrete time does not necessarily yield a good approximation for its continuous-time counterpart. We illustrate this by recalling the asymptotics of (10.1) in both discrete and continuous time. Denote the probability in continuous time by $p_n^{\mathbb{R}_+}$.

In discrete time, there exists a constant $\mathcal{K}$ such that [220]

$$p_n \sim \frac{\mathcal{K}}{\sqrt{n}} \exp\left(-\frac{1}{2} n \frac{(b + (r-\mu)t^*)^2}{\sigma^2(t^*)}\right),$$

where t^* minimizes I_t over $\mathbb{N}$. However, in continuous time the asymptotics depend on the behavior of σ near zero. If $\sigma(t) \sim Ct^\gamma$ as $t \to 0$ for constants $C \in (0, \infty)$ and $\gamma \in (0, 2)$, then, under suitable regularity assumptions, [90]

$$p_n^{\mathbb{R}_+} \sim \mathcal{K}' n^{\frac{1}{\gamma}-1} \exp\left(-\frac{1}{2} n \frac{(b + (r-\mu)t^*)^2}{\sigma^2(t^*)}\right),$$

with t^* minimizer of I_t over $\mathbb{R}_+$, and for some constant $\mathcal{K}'$ (which involves the so-called *Pickands' constant* for which no explicit representation is available; see Section 3.3). Conclude that the polynomial term in the above asymptotic expansions is different. To our knowledge, reliable simulation methods for the continuous-time probability $p_n^{\mathbb{R}_+}$ do not exist.

10.A Appendix: proofs

In this appendix, we provide proofs of the assertions in this chapter. We start in Appendix 10.A.1 with the proofs related to the simulation horizon T, which apply to all methods discussed in Section 10.3. Appendices 10.A.2 and 10.A.3 deal with the single-twist method and cut-and-twist method respectively. The proof of Lemma 10.9 is given in Appendix 10.A.4.

10.A.1 Upper bounds on $\int_T^\infty e^{-nC_0t^{1/q}}dt$

We distinguish the cases $q \leq 1$ (Lemma 10.1) and $q > 1$ (Lemma 10.2).

Proof of Lemma 10.1. Since $q \leq 1$ and $T \in \mathbb{N}$, we can bound the left-hand side of (10.6) as follows:

$$\begin{aligned}\int_T^\infty \exp\left(-nC_0t^{1/q}\right)dt &= \frac{q}{C_0^q}\int_{C_0T^{1/q}}^\infty \exp(-ny)y^{q-1}dy \\ &\leq \frac{q}{C_0^q}\left(C_0T^{1/q}\right)^{q-1}\int_{C_0T^{1/q}}^\infty \exp(-ny)dy \\ &= \frac{q}{C_0^q n}\left(C_0T^{1/q}\right)^{q-1}\exp\left(-nC_0T^{1/q}\right) \\ &\leq \frac{q}{C_0 n}\exp\left(-nC_0T^{1/q}\right),\end{aligned}$$

as claimed. □

Proof of Lemma 10.2. First note that $q > 1$, which is crucial throughout the proof. Recall that $m \geq 0$ denotes the largest integer such that $q - 1 - m \in (0, 1]$. As before, we have by a simple substitution,

$$\int_T^\infty \exp\left(-nC_0t^{1/q}\right)dt = \frac{q}{C_0^q}\int_{C_0T^{1/q}}^\infty \exp(-ny)y^{q-1}dy. \tag{10.21}$$

The idea is to select $\beta, \gamma \in (0, \infty)$ such that

$$y^{q-1} \leq \beta e^{\gamma y} \tag{10.22}$$

for all $y \in \mathbb{R}_+$. We now discuss how these parameters can be chosen.

If $q \in (1, 2]$ (i.e., $m = 0$), then $p_q : y \mapsto y^{q-1}$ is concave. Since p_q is differentiable at 1 with derivative $q - 1$, by Theorem 25.1 of Rockafellar [273] we have for all $y \in \mathbb{R}_+$,

$$y^{q-1} \leq 1 + (q - 1)(y - 1). \tag{10.23}$$

Similarly, since $y \mapsto \beta e^{\gamma y}$ is convex and differentiable at 1 with derivative $\beta\gamma e^\gamma$, we have for all $y \in \mathbb{R}_+$,

$$\beta e^{\gamma y} \geq \beta e^\gamma + \beta\gamma e^\gamma(y - 1). \tag{10.24}$$

By comparing (10.23) to (10.24), we see that $y^{q-1} \leq \beta e^{\gamma y}$ upon choosing $\gamma = q - 1$ and $\beta = e^{-\gamma}$.

To find β, γ such that (10.22) holds for $q \in (m+1, m+2]$ where $m > 0$, the key observation is that this inequality is always satisfied for $y = 0$. Therefore, it suffices to choose β, γ such that the derivative of the left-hand side of (10.22) does not exceed the right-hand side. By applying this idea m times, one readily observes that it suffices to require that β, γ satisfy

$$\beta\gamma^m e^{\gamma y} \geq (q-1)\cdots(q-m)y^{q-m-1}.$$

Note that the right-hand side of this expression is concave as a function of y since $q-m-1 \in (0,1]$, and that the left-hand side is convex as a function of y. Therefore, we are in a similar situation as we were for $m = 0$. In this case, we choose β and γ such that

$$\begin{aligned} \beta\gamma^m e^{\gamma} &= (q-1)\cdots(q-m) \\ \beta\gamma^{m+1} e^{\gamma} &= (q-1)\cdots(q-m)(q-m-1). \end{aligned}$$

Note that β and γ as defined in (10.7) solve this system of equations uniquely. Again, Theorem 25.1 of Rockafellar [273] is applied twice to see that for $y \in \mathbb{R}_+$,

$$\begin{aligned} &(q-1)\cdots(q-m)y^{q-m-1} \\ &\quad\leq (q-1)\cdots(q-m) + (q-1)\cdots(q-m)(q-m-1)(y-1) \\ &\quad= \beta\gamma^m e^{\gamma} + \beta\gamma^{m+1}e^{\gamma}(y-1) \leq \beta\gamma^m e^{\gamma y}. \end{aligned}$$

Now that we have found simple bounds on y^{q-1}, we combine these bounds with (10.21) to see that

$$\begin{aligned} \int_T^\infty \exp\left(-nC_0 t^{1/q}\right) dt &\leq \frac{q\beta}{C_0^q} \int_{C_0 T^{1/q}}^\infty \exp(-(n-\gamma)y) dy \\ &= \frac{q\beta}{C_0^q (n-\gamma)} \exp\left(-(n-\gamma)C_0 T^{1/q}\right), \end{aligned}$$

as claimed. □

10.A.2 Proofs for the single-twist method

The key ingredient in the proofs of this subsection is Cramér's large-deviation principle (LDP), introduced in Section 2.3.1. In the present finite-dimensional Gaussian setting, Cramér's theorem coincides with Schilder's theorem, see Section 3.1. We first discuss how these theorems compile to this framework.

Recall that given some $T \in \mathbb{N}$, $\nu_n^{(T)}$ denotes the distribution of the centered process $\{A_n(t)/n - \mu t : t = 1, \ldots, T\}$. The covariance of $\nu_n^{(T)}$ is given by $\Gamma^{(T)}/n$, and this covariance defines an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\| \cdot \|_{\mathcal{H}}$ on $\mathbb{R}^T$ as follows:

$$\langle x, y \rangle_{\mathcal{H}} := x' \left(\Gamma^{(T)}\right)^{-1} y, \quad \|x\|_{\mathcal{H}} := \sqrt{\langle x, x \rangle_{\mathcal{H}}}.$$

This inner product sometimes referred to as *reproducing kernel Hilbert space* inner product or *Cameron-Martin space* inner product. It is the finite-dimensional version of the inner product discussed in Section 3.1.

Theorem 10.11 (Cramér) *$\{\nu_n^{(T)}\}$ satisfies the LDP in $\mathbb{R}^T$ with rate function $I : x \to \frac{1}{2}\|x\|_{\mathcal{H}}^2$ and scale sequence $\{n\}$.*

Proof. The claim is a consequence of Proposition 2.16, after noting that

$$\sup_{\theta \in \mathbb{R}^T} \left(\langle \theta, x \rangle - \log \int e^{\langle \theta, y \rangle} \nu^{(T)}(dy) \right) = \sup_{\theta \in \mathbb{R}^T} \left(\langle \theta, x \rangle - \frac{1}{2} \theta' \Gamma^{(T)} \theta \right),$$

which equals $\frac{1}{2}x'\left(\Gamma^{(T)}\right)^{-1}x = \frac{1}{2}\|x\|_{\mathcal{H}}^2$. □

Now we can prove Lemma 10.4.

Proof of Lemma 10.4. Lemma 10.4 is an application of Theorem 10.11. We have to prove that

$$\lim_{n\to\infty}\frac{1}{n}\log\nu_n^{(T)}(\mathcal{O}_T) = -\frac{1}{2}\inf_{x\in\mathcal{O}_T}\|x\|_{\mathcal{H}}^2 = -\frac{1}{2}\|x^*\|_{\mathcal{H}}^2. \tag{10.25}$$

The second equality in (10.25) is due to Addie *et al.* [2]. We therefore turn to the first equality. It is readily seen that $\mathcal{O}_T$ is closed in $\mathbb{R}^T$. Cramér's theorem gives an upper bound on the decay rate of $\nu_n^{(T)}(\mathcal{O}_T)$, as well as a lower bound on the decay rate of $\nu_n^{(T)}(\underline{\mathcal{O}}_T)$, where $\underline{\mathcal{O}}_T$ denotes the interior of $\mathcal{O}_T$. The first equality of (10.25) now follows upon combining these upper and lower bounds with Lemma 10.12 below (applied for $y = 0$). □

Lemma 10.12 *For all $y \in \mathbb{R}^T$, we have*

$$\inf_{x\in\underline{\mathcal{O}}_T}\|x+y\|_{\mathcal{H}}^2 = \inf_{x\in\mathcal{O}_T}\|x+y\|_{\mathcal{H}}^2 = \inf_{t\in\{1,\ldots,T\}}\frac{(b+(r-\mu)t+y_t)^2}{2\sigma^2(t)}.$$

Proof. First note that the interior of the exceedance set is given by

$$\underline{\mathcal{O}}_T := \left\{x\in\mathbb{R}^T : x_t+\mu t > b+rt \text{ for some } t\in\{1,\ldots,T\}\right\}.$$

Also, evidently,

$$\inf_{x\in\underline{\mathcal{O}}_T}\|x+y\|_{\mathcal{H}}^2 = \inf_{x\in\underline{\mathcal{O}}_{T,y}}\|x\|_{\mathcal{H}}^2,$$

where

$$\underline{\mathcal{O}}_{T,y} := \left\{x\in\mathbb{R}^T : x_t+\mu t > b+rt+y_t \text{ for some } t\in\{1,\ldots,T\}\right\}.$$

A similar reasoning that led to the second equality in (10.25) now yields the desired. □

Proof of Lemma 10.5. From the arguments in the proof of Theorem 10.6 below, the claim is an immediate consequence of Theorem 8.10. □

Proof of Theorem 10.6. We want to apply Corollary 8.4. Note that the required continuity properties of $\mathcal{O}_T$ follow from Lemma 10.12. Assumption 8.1(ii) holds by Theorem 10.11, and Assumption 8.1(iii) is implied by the fact that for $\gamma > 1$,

$$\limsup_{n\to\infty}\frac{1}{n}\log\int_{\mathbb{R}^T}\exp\left(-n\gamma\frac{b+(r-\mu)t^*}{\sigma^2(t^*)}x_{t^*}\right)\nu_n^{(T)}(dx) = \gamma^2\frac{[b+(r-\mu)t^*]^2}{2\sigma^2(t^*)} < \infty,$$

see (2.5) and use that for a zero-mean normal random variable U (with variance σ^2) the moment-generating function is $\mathbb{E}\exp(\theta U) = \exp(\theta^2\sigma^2/2)$.

Since $\inf_{x\in\mathbb{R}^T}\left[\frac{1}{2}\|x\|_{\mathcal{H}}^2 - \langle x^*, x\rangle\right] = -\frac{1}{2}\|x^*\|_{\mathcal{H}}^2$, after applying Corollary 8.4, it remains to observe that

$$\begin{aligned}
&-\inf_{x\in\mathcal{O}_T}\left[\frac{1}{2}\|x\|_{\mathcal{H}}^2 + \langle x, x^*\rangle_{\mathcal{H}} - \frac{1}{2}\|x^*\|_{\mathcal{H}}^2\right]\\
&\quad= -\left[\frac{1}{2}\inf_{x\in\mathcal{O}_T}\|x+x^*\|_{\mathcal{H}}^2\right] + \|x^*\|_{\mathcal{H}}^2\\
&\quad= -\frac{1}{2}\inf_{t\in\{1,\ldots,T\}}\frac{(b+(r-\mu)t+x_t^*)^2}{\sigma^2(t)} + \frac{(b+(r-\mu)t^*)^2}{\sigma^2(t^*)},
\end{aligned}$$

where the last equality is due to Lemma 10.12. □

10.A.3 Proofs for the cut-and-twist method

We now prove Theorem 10.7. Observe that for any $j \in \mathbb{N}$, by definition of $\mathcal{O}_T(t)$,

$$\begin{aligned}
&\int_{\mathcal{O}_T(t)} \left(\frac{\nu_n^{(T)}}{{}^t\lambda_n^{(T)}} \right)^j d\ {}^t\lambda_n^{(T)} \\
&= \int_{\mathcal{O}_T(t)} \exp\left(nj \frac{(b+(r-\mu)t)^2}{2\sigma^2(t)} - nj \frac{b+(r-\mu)t}{\sigma^2(t)} x_t \right) d\ {}^t\lambda_n^{(T)} \\
&\leq \exp\left(-nj \frac{(b+(r-\mu)t)^2}{2\sigma^2(t)} \right) = e^{-njI_t}.
\end{aligned}$$

As an aside, we mention that this gives (by choosing $j = 1$), cf. Section 10.2.2,

$$p_n^T = \sum_{t=1}^{T} \nu_n^{(T)}(\mathcal{O}_T(t)) \leq \sum_{t=1}^{T} e^{-nI_t}.$$

The second moment of the cut-and-twist estimator follows from (10.17):

$$\begin{aligned}
&\frac{1}{N} \int_{\mathbb{R}^T} \left(\sum_{t\in\{1,\dots,T\}} \mathbf{1}_{\{x_t \in \mathcal{O}_T(t)\}} \frac{d\nu_n^{(T)}}{d\ {}^t\lambda_n^{(T)}}(x_t) \right)^2 d\ {}^1\lambda_n^{(T)}(x_1) \cdots d\ {}^T\lambda_n^{(T)}(x_T) \\
&= \frac{1}{N} \sum_{t\in\{1,\dots,T\}} \int_{\mathcal{O}_T(t)} \left(\frac{\nu_n^{(T)}}{{}^t\lambda_n^{(T)}} \right)^2 d\ {}^t\lambda_n^{(T)} + \frac{1}{N} \sum_{\substack{s,t\in\{1,\dots,T\} \\ s\neq t}} {}^s\lambda_n^{(T)}(\mathcal{O}_T(t)) \cdot\ {}^t\lambda_n^{(T)}(\mathcal{O}_T(t)),
\end{aligned}$$

and therefore it is bounded by

$$\frac{1}{N} \left[\sum_{t\in\{1,\dots,T\}} \exp\left(-n \frac{(b+(r-\mu)t)^2}{2\sigma^2(t)} \right) \right]^2 \leq \frac{1}{N} T^2 \exp(-2nI_{t^*}),$$

where the last inequality is due to the definition of $t^* = \arg\inf_t I_t$. Now take logarithms, divide by n, and let $n \to \infty$ to see that the relative efficiency equals 2, cf. (7.5). □

10.A.4 Proofs for the sequential-twist method

Proof of Lemma 10.9. We have to prove that

$$\operatorname*{arg\,inf}_{\{y\in\mathbb{R}^T : \frac{1}{n}\sum_{i=1}^{j} \bar{A}_i + (1-j/n)y \in \mathcal{O}_T\}} \|y\|_{\mathcal{H}}^2 = \frac{J_{j+1}}{\sigma(t_{j+1}^*)} \Gamma(\cdot, t_{j+1}^*).$$

From Lemma 10.4, we know that the infimum equals J_{j+1}^2. It is not hard to see that μ_{j+1} attains this value (by strict convexity of $\|\cdot\|_{\mathcal{H}}$, the minimizing argument is even unique). □

Proof of Theorem 10.10. The two assumptions in Condition 2.1 of Dupuis and Wang [127] hold: since we are in a multivariate Gaussian setup, we obviously have an everywhere finite moment-generating function, and Lemma 10.12 implies that

$$\inf_{x\in\overline{\mathcal{O}_T}} x' \left(\Gamma^{(T)} \right)^{-1} x = \inf_{x\in\mathcal{O}_T^o} x' \left(\Gamma^{(T)} \right)^{-1} x.$$

The claim is Theorem 2.1 of Dupuis and Wang [127]. □

Part C

Lévy-driven fluid queues and networks

CHAPTER 11

Background on Lévy processes

Lévy processes are sometimes called random walks in continuous time. Motivated by applications to risk processes and fluid networks, it is our aim in Part C to study extremes of Lévy processes; in the present context, this is known as *fluctuation theory*.

The analysis relies extensively on a technique known as *splitting at the maximum*, which is closely related to the Wiener-Hopf factorization. In this chapter, we discuss the relevant theory for random walks and Lévy processes.

A *Lévy process* $X = \{X(t) : t \geq 0\}$ has stationary and independent increments, and is defined on the probability space of càdlàg functions with the Borel σ-field generated by the usual Skorokhod topology (see, e.g., [50, Sec. 12] or [170, Ch. VI]). Two standard references on Lévy processes are the books by Bertoin [43] and Sato [283].

The characteristic function of $X(t)$ necessarily has the form $\mathbb{E}e^{i\beta X(t)} = e^{-t\Psi_X(\beta)}$, $\beta \in \mathbb{R}$, where Ψ_X is the *characteristic exponent*

$$\Psi_X(\beta) = \frac{1}{2}\sigma_X^2\beta^2 + ic_X\beta + \int_{\mathbb{R}} \left(1 - e^{i\beta x} + i\beta x \mathbf{1}_{[-1,1]}(x)\right) \Pi_X(dx), \tag{11.1}$$

for some $\sigma_X \geq 0$, $c_X \in \mathbb{R}$ and a so-called *Lévy measure* Π_X on $\mathbb{R}\backslash\{0\}$ satisfying $\int(1 \wedge |x|^2)\Pi_X(dx) < \infty$. The *Lévy-Khinchine representation* of Ψ_X in (11.1) shows in particular that $X(0) = 0$. X is called a *compound Poisson process* if $c_X = \sigma_X = 0$ and $\Pi_X(\mathbb{R}) < \infty$. If $c_X = 0$ and $\Pi_X \equiv 0$, then X reduces to a (nonstandard) *Brownian motion*. Moreover, if $\sigma_X = 0$ and $\Pi_X \equiv 0$, then X is a *drift*.

As a consequence of the stationarity and independence of the increments, Lévy processes can be regarded as the continuous-time analogues of random walks, which have been briefly discussed in Sections 2.3 and 7.2. Therefore, before investigating the extremes of a Lévy process, it is natural to start by discussing the random-walk setting (Section 11.1). In Section 11.2, we give the Lévy analogues of some of these results.

11.1 Random walks

The process $\{S_n\}$ is a *random walk* if $S_0 = 0$ and $S_n := \sum_{i=1}^n X_i$, $n \geq 1$, where the step sizes $X_1, X_2, \ldots$ are i.i.d. random variables, not necessarily centered. Their distribution is denoted by $\mathbb{P}_X$, and we write $F(x) := \mathbb{P}_X((-\infty, x])$ for their distribution function.

It is the goal of this section to discuss some aspects of fluctuation theory for random walks. Since the pioneering work of Feller [132], this theory relies extensively on ladder variables.

11.1.1 Ladder variables

The first strict ascending (descending) *ladder epoch* $\tau_{s\pm}$ is defined as

$$\tau_{s+} = \inf\{n \geq 1 : S_n > 0\}, \quad \tau_{s-} = \inf\{n \geq 1 : S_n < 0\},$$

and the first weak ascending ladder epoch $\tau_{w\pm}$ is defined similarly with a weak inequality. The quantity $H_{s+} := S_{\tau_{s+}}$ is known as the first strict ascending *ladder height*, with similar terminology for $H_{s-} := S_{\tau_{s-}}$ and the corresponding weak analogues. Note that H_{s+} is only defined on the event $\{\tau_{s+} < \infty\}$; therefore, its distribution has defect $\mathbb{P}(\tau_{s+} = \infty)$, which is strictly positive if S drifts to $-\infty$.

When integrating with respect to defective distributions, we only carry out the integration over the set where the random variables are both finite and well-defined. For instance, we write $\mathbb{E}\rho^{\tau_{s+}} e^{-\beta H_{s+}}$ for $\mathbb{E}\left[\rho^{\tau_{s+}} e^{-\beta H_{s+}}; \tau_{s+} < \infty\right] := \mathbb{E}\left[\rho^{\tau_{s+}} e^{-\beta H_{s+}} \mathbf{1}_{\{\tau_{s+}<\infty\}}\right]$ in the remainder, unless indicated otherwise.

The following theorem relates the transforms of (τ_{s+}, H_{s+}) and (τ_{w-}, H_{w-}). A short proof of this result is given by Kennedy [189]; here, we only give the intuition behind her proof. Similar arguments are used in Chapter 14.

Theorem 11.1 (Wiener-Hopf) *For $|\rho| \leq 1$, $\beta \in \mathbb{R}$, we have*

$$\begin{aligned} 1 - \rho\mathbb{E}e^{i\beta X_1} &= \left[1 - \mathbb{E}\rho^{\tau_{s+}} e^{i\beta H_{s+}}\right]\left[1 - \mathbb{E}\rho^{\tau_{w-}} e^{i\beta H_{w-}}\right] \\ &= \left[1 - \mathbb{E}\rho^{\tau_{w+}} e^{i\beta H_{w+}}\right]\left[1 - \mathbb{E}\rho^{\tau_{s-}} e^{i\beta H_{s-}}\right]. \end{aligned}$$

Sketch of the proof. An analytic-continuation argument shows that we only need to prove the claim for ρ real and $\rho \in (0, 1]$. By introducing the (defective) step-size distribution $\rho\mathbb{P}_X$, we may further suppose that $\rho = 1$. We only prove the first equality for symmetry reasons.

Write G_{s+} for the distribution function of H_{s+}, and similarly for G_{w-}. Let $x \leq 0$. On the event $\{\tau_{s+} > 1\}$, we define

$$\sigma_{w-} := \sup\left\{k \geq 1 : S_k = \inf_{1 \leq \ell < \tau_{w-}} S_\ell\right\}.$$

Now observe that

$$G_{w-}(x) = F(x) + \mathbb{P}\left(S_{\sigma_{w-}} + \left[H_{w-} - S_{\sigma_{w-}}\right] \leq x; \tau_{s+} > 1\right), \tag{11.2}$$

and that $H_{w-} - S_{\sigma_{w-}}$ is the first weak descending ladder height if one thinks of $(\sigma_{w-}, S_{\sigma_{w-}})$ as the origin and looks 'forward' in time. Similarly, $S_{\sigma_{w-}}$ is the first strict ascending ladder height if one looks 'backward'; this is formalized by time-reversal, a technique that we often encounter in Part C. This makes it plausible that the second term on the right-hand side of (11.2) equals $G_{w-} * G_{s+}(x)$, which can be made precise by conditioning appropriately. We refer to Figure 11.1, which is taken from [189], for an intuitive explanation. This results in the identity $G_{w-}(x) = F(x) + G_{w-} * G_{s+}(x)$ for $x \leq 0$, where '$*$' denotes convolution, see Section 2.4.1. A similar argument shows that $G_{s+}(x) = F(x) + G_{w-} * G_{s+}(x)$ for $x \geq 0$, and these two statements are summarized by the equality $G_{w-} + G_{s+} = F + G_{w-} * G_{s+}$. The claim is nothing else than a reformulation of this identity. □

11.1.2 The distribution of the maximum

Next we study the running-maximum process of the random walk, relying predominantly on the fact that $\{S_n\}$ is (strongly) Markov. As indicated in Section 1.1, the maximum plays a key role in queueing theory, risk theory, and finance. Throughout, let N_ρ be geometric on $\{0, 1, 2, \ldots\}$ with parameter $\rho \in (0, 1)$, independent of S.

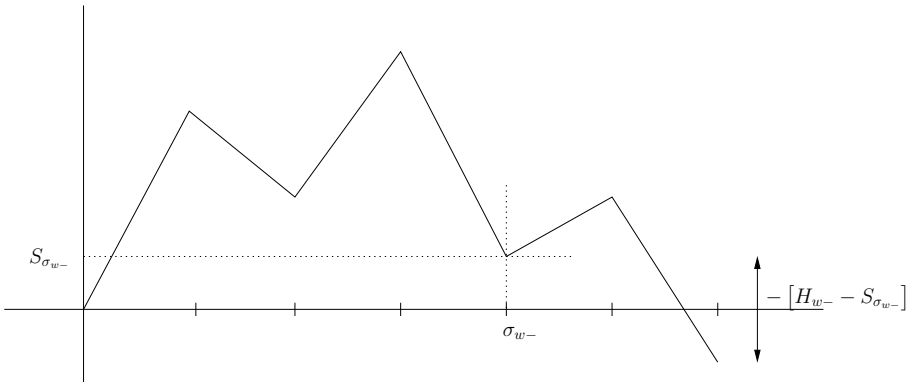

Figure 11.1: The intuition behind the Wiener-Hopf factorization.

We first introduce some further notation:

$$\begin{aligned} \overline{S}_n &:= \sup\{S_k : 0 \le k \le n\}, & \underline{S}_n &:= \inf\{S_k : 0 \le k \le n\}, \\ \overline{F}^S_n &:= \inf\{k \le n : S_k = \overline{S}_n\}, & \overline{G}^S_n &:= \sup\{k \le n : S_k = \overline{S}_n\}, \\ \underline{F}^S_n &:= \inf\{k \le n : S_k = \underline{S}_n\}, & \underline{G}^S_n &:= \sup\{k \le n : S_k = \underline{S}_n\}. \end{aligned}$$

We also define the global maximum as $\overline{S} := \overline{S}_\infty$, and note that it is almost surely finite whenever S drifts to $-\infty$, for instance if $\mathbb{E}X(1) < 0$.

The next proposition is a first result on *splitting at the maximum*, which is an important technique throughout Part C. A proof is not given, but it is important to note that the statement does not follow immediately from the Markov property, since $\overline{G}^S_{N_\rho}$ is not a stopping time.

Proposition 11.2 *The random vectors $(\overline{G}^S_{N_\rho}, \overline{S}_{N_\rho})$ and $(N_\rho - \overline{G}^S_{N_\rho}, S_{N_\rho} - \overline{S}_{N_\rho})$ are independent, and the same holds for $(\overline{F}^S_{N_\rho}, \overline{S}_{N_\rho})$ and $(N_\rho - \overline{F}^S_{N_\rho}, S_{N_\rho} - \overline{S}_{N_\rho})$.*

Moreover, $(N_\rho - \overline{G}^S_{N_\rho}, S_{N_\rho} - \overline{S}_{N_\rho})$ and $(N_\rho - \overline{F}^S_{N_\rho}, S_{N_\rho} - \overline{S}_{N_\rho})$ have the same distribution as $(\underline{F}^S_{N_\rho}, \underline{S}_{N_\rho})$ and $(\underline{G}^S_{N_\rho}, \underline{S}_{N_\rho})$ respectively.

In terms of transforms, the previous proposition implies that for $\alpha \ge 0$, $\beta \in \mathbb{R}$,

$$\mathbb{E}e^{-\alpha N_\rho + i\beta S_{N_\rho}} = \mathbb{E}e^{-\alpha \overline{G}^S_{N_\rho} + i\beta \overline{S}_{N_\rho}} \mathbb{E}e^{-\alpha \underline{F}^S_{N_\rho} + i\beta \underline{S}_{N_\rho}} = \mathbb{E}e^{-\alpha \overline{F}^S_{N_\rho} + i\beta \overline{S}_{N_\rho}} \mathbb{E}e^{-\alpha \underline{G}^S_{N_\rho} + i\beta \underline{S}_{N_\rho}}. \quad (11.3)$$

This identity can be regarded as the Wiener-Hopf factorization for extremes (as opposed to ladder variables). To see this, we first need an auxiliary lemma. The 'renewal' argument in its proof is used on several occasions in Part C.

Lemma 11.3 *For $\beta \ge 0$, $\rho \in (0,1)$, we have*

$$\mathbb{E}e^{-\beta \overline{S}_{N_\rho}} = \frac{1 - \mathbb{E}\rho^{\tau_{s+}}}{1 - \mathbb{E}\rho^{\tau_{s+}} e^{-\beta H_{s+}}}, \quad \mathbb{E}e^{\beta \underline{S}_{N_\rho}} = \frac{1 - \mathbb{E}\rho^{\tau_{w-}}}{1 - \mathbb{E}\rho^{\tau_{w-}} e^{\beta H_{w-}}}.$$

Moreover, if S drifts to $-\infty$, then $\mathbb{E}e^{-\beta \overline{S}} = \mathbb{P}(\tau_{s+} = \infty) / \left(1 - \mathbb{E}e^{-\beta H_{s+}}\right)$. That is, for $x \ge 0$,

$$\mathbb{P}\left(\overline{S} \le x\right) = \mathbb{P}(\tau_{s+} = \infty) \sum_{n=0}^{\infty} \mathbb{P}\left(H^{(1)}_{s+} + \ldots + H^{(n)}_{s+} \le x\right),$$

where $H^{(1)}_{s+}, H^{(2)}_{s+}, \ldots$ are i.i.d. with the same (defective) distribution as H_{s+}.

Proof. A random walk with geometric killing at N_ρ can be regarded as a random walk with defective step-size distribution $\rho\mathbb{P}_X$. The probability that such a random walk does not have an ascending ladder epoch is $1 - \mathbb{E}\rho^{\tau_{s+}}$. On the other hand, if there is an ascending ladder epoch, the process 'restarts' as a result of the Markov property. Therefore, we have for $\beta \geq 0$,

$$\mathbb{E}e^{-\beta\overline{S}_{N_\rho}} = 1 - \mathbb{E}\rho^{\tau_{s+}} + \mathbb{E}\rho^{\tau_{s+}}e^{-\beta H_{s+}}\mathbb{E}e^{-\beta\overline{S}_{N_\rho}},$$

and the first claim follows. The corresponding identity for the running-minimum process is proven similarly, and it is also possible to give an expression in terms of strong ladder quantities.

The assertion for $\overline{S}$ follows by letting $\rho \to 1$, and the last claim can be checked by applying the uniqueness theorem for Laplace transforms. □

Lemma 11.3 gives expressions for the transforms of $\overline{S}_{N_\rho}$ and $\underline{S}_{N_\rho}$, and the transform of S_{N_ρ} is readily calculated after conditioning on the value of N_ρ. This shows that (11.3), hence also Proposition 11.2, implies that for $\beta \geq 0$, $\rho \in (0,1)$,

$$\frac{1 - \mathbb{E}\rho^{\tau_{s+}}}{1 - \mathbb{E}\rho^{\tau_{s+}}e^{-\beta H_{s+}}} \frac{1 - \mathbb{E}\rho^{\tau_{w-}}}{1 - \mathbb{E}\rho^{\tau_{w-}}e^{\beta H_{w-}}} = \frac{1-\rho}{1 - \rho\mathbb{E}e^{i\beta S_1}}.$$

The Wiener-Hopf factorization for ladder variables (Theorem 11.1) shows why the above identity can be regarded as the Wiener-Hopf factorization for extremes. Further details can be found in most textbooks on queueing theory, see Section 1.1.2 for a selection. A particularly nice account, on which the preceding discussion is based, is given by Borovkov [55].

Exact solutions

As we have seen, determining the distribution of $\overline{S}$ amounts to studying the distribution of the first ascending ladder height H_{s+} (or, as the reader may check, its weak counterpart H_{w+}). For a given step-size distribution $\mathbb{P}_X$, this distribution follows in principle from the Wiener-Hopf factorization and standard arguments from complex analysis. However, there is a large class of distributions for which the distribution of H_{s+} is known more explicitly.

An example that is particularly tractable corresponds to the M/G/1 queue (exponentially distributed interarrival times) and the G/M/1 queue (exponentially distributed service times). For the M/G/1 queue, $\mathbb{P}_X = \mathbb{P}_B * \mathbb{P}_{-A}$, where B has a general (integrable) distribution on $[0,\infty)$ and A has an exponential distribution. Then, H_{s+} has a defect $1 - \mathbb{E}B/\mathbb{E}A$, and the distribution of H_{s+} given $H_{s+} > 0$ has the so-called integrated-tail distribution with distribution function $\int_0^x \mathbb{P}(B > y)dy/\mathbb{E}B$. For the G/M/1 queue, $\mathbb{P}_X = \mathbb{P}_B * \mathbb{P}_{-A}$, where B is exponentially distributed and A has a general distribution on $[0,\infty)$; then, H_{s+} given $H_{s+} > 0$ has the same distribution as B. Moreover, the defect of H_{s+} is $\eta\mathbb{E}B$, where $\eta > 0$ is specified as the solution to some fixed-point equation.

In these two examples, it is possible to replace the exponential distributions by distributions with a so-called rational Laplace transform, see Cohen [79, Sec. II.5.10 and II.5.11]. The class of *phase-type* distributions, which is discussed in Section 12.3, is an important subclass of these distributions.

Asymptotics

In the absence of exact results on the distribution of $\overline{S}$, one can resort to studying tail asymptotics for $\overline{S}$, i.e., $\mathbb{P}(\overline{S} > x)$ as $x \to \infty$, and local tail asymptotics, i.e., the behavior of $\mathbb{P}(\overline{S} \in (x, x+T])$ for any $T > 0$ as $x \to \infty$.

In the following theorem, we summarize results of Feller [132, Sec. XII.5.(c)] and Iglehart [167, Lem. 1] (first part), Bertoin and Doney [47, Thm. 1] (second part), and Veraverbeke [304, Thm. 2] and Asmussen *et al.* [22, Thm. 1] (third part). The first assertion in the

third part is known as *Veraverbeke's theorem*, for which a short proof has recently been found by Zachary [311]. Korshunov [197] discusses this theorem in more detail.

We impose various assumptions on the tail of the step-size distribution, see Section 2.4 for a review of the classes $\mathcal{S}(\delta)$, $\mathcal{S}$, and $\mathcal{S}^*$. For integrable X, the distribution function F_I is defined as $F_I(x) := \int_0^x \overline{F}(y)dy/\mathbb{E}X$. We restrict our attention to the nonlattice case, but similar results hold in the lattice case.

Theorem 11.4 *Let S be a random walk that drifts to $-\infty$.*

(i) Suppose that F is nonlattice, and that there exists some $\omega \in (0,\infty)$ such that $\mathbb{E}e^{\omega X} = 1$ and $\mathbb{E}Xe^{\omega X} < \infty$. Then as $x \to \infty$, we have

$$\mathbb{P}(\overline{S} > x) \sim e^{-\sum_{n=1}^{\infty} \frac{1}{n}\left\{\mathbb{P}(S_n>0)+\mathbb{E}\left[e^{\omega S_n}; S_n \le 0\right]\right\}} \frac{1}{\omega \mathbb{E}X e^{\omega X}} e^{-\omega x}.$$

Moreover, for any $T > 0$, as $x \to \infty$, we have

$$\mathbb{P}(\overline{S} \in (x, x+T]) \sim e^{-\sum_{n=1}^{\infty} \frac{1}{n}\left\{\mathbb{P}(S_n>0)+\mathbb{E}\left[e^{\omega S_n}; S_n \le 0\right]\right\}} \frac{1-e^{-\omega T}}{\omega \mathbb{E}X e^{\omega X}} e^{-\omega x}.$$

(ii) Suppose that $\delta := \sup\{\omega > 0 : \mathbb{E}e^{\omega X} < \infty\} > 0$ with $\mathbb{E}e^{\delta X} < 1$. If $F \in \mathcal{S}(\delta)$, then $\mathbb{E}e^{\delta \overline{S}} < \infty$, and as $x \to \infty$, we have

$$\mathbb{P}(\overline{S} > x) \sim \frac{\mathbb{E}e^{\delta \overline{S}}}{1 - \mathbb{E}e^{\delta X}} \overline{F}(x).$$

Moreover, for any $T > 0$, as $x \to \infty$, we have

$$\mathbb{P}(\overline{S} \in (x, x+T]) \sim \frac{\mathbb{E}e^{\delta \overline{S}}}{1 - \mathbb{E}e^{\delta X}} \mathbb{P}_X((x, x+T]).$$

(iii) Suppose that $\mathbb{E}|X| < \infty$ and $F_I \in \mathcal{S}$. Then, as $x \to \infty$, we have

$$\mathbb{P}(\overline{S} > x) \sim -\frac{\int_x^{\infty} \overline{F}(y)dy}{\mathbb{E}X}.$$

Moreover, if $F \in \mathcal{S}^*$ and F is nonlattice, then for any $T > 0$, as $x \to \infty$, we have

$$\mathbb{P}(\overline{S} \in (x, x+T]) \sim -\frac{\int_x^{x+T} \overline{F}(y)dy}{\mathbb{E}X}.$$

11.2 Lévy processes

Lévy processes can be thought of as continuous-time random walks, but the ladder analysis of the previous section is not straightforwardly generalized. For instance, different (more technical) tools are required for a ladder analysis of Brownian motion, as it crosses levels continuously. The so-called *ladder process*, which is a two-dimensional Lévy process, replaces ladder epochs and heights, and it involves the local time of the reflected process. The Wiener-Hopf factorization for Lévy processes involves the characteristics of this ladder process, but we do not address this here; however, the reader may look into the proof of Corollary 12.2. Instead, we content ourselves with an analogue of (11.3).

We start with some notation. Given a Lévy process X, we define

$$\begin{aligned}
\overline{X}(t) &:= \sup\{X(s) : 0 \le s \le t\},\\
\overline{F}^X(t) &:= \inf\{s < t : X(s) = \overline{X}(t) \text{ or } X(s-) = \overline{X}(t)\},\\
\overline{G}^X(t) &:= \sup\{s < t : X(s) = \overline{X}(t) \text{ or } X(s-) = \overline{X}(t)\},\\
\underline{X}(t) &:= \inf\{X(s) : 0 \le s \le t\},\\
\underline{F}^X(t) &:= \inf\{s < t : X(s) = \underline{X}(t) \text{ or } X(s-) = \underline{X}(t)\},\\
\underline{G}^X(t) &:= \sup\{s < t : X(s) = \underline{X}(t) \text{ or } X(s-) = \underline{X}(t)\}.
\end{aligned}$$

Here and throughout Part C, e_q denotes an exponentially distributed random variable with parameter q, independent of X. Knowing the distribution of $\overline{X}(e_q)$ for any $q > 0$ is equivalent to knowing the distribution of $\overline{X}(t)$ for any $t > 0$, since evaluation at time e_q essentially amounts to taking a Laplace transform. In analogy to the random-walk setting, where $\overline{S}$ is a limit of $\overline{S}_{N_\rho}$ as $\rho \to 1$, the global maximum $\overline{X}$ can be regarded as a limit of $\overline{X}(e_q)$ as $q \to 0$.

The following identity, referred to as the *Pecherskiĭ-Rogozin-Spitzer (PRS) factorization* throughout part C, is the analogue of (11.3). It is also called the *first factorization identity*. Importantly, since X is *time-reversible* in the sense that $\{X(t) - \lim_{u\uparrow(t-s)} X(u) : 0 \le s \le t\}$ and $\{X(s) : 0 \le s \le t\}$ have the same distribution, $(\underline{F}^X(e_q), \underline{X}(e_q))$ is distributed as $(e_q - \overline{G}^X(e_q), X(e_q) - \overline{X}(e_q))$. Therefore, the PRS factorization is essentially a Lévy version of Proposition 11.2, so that it can be regarded as a first illustration of the splitting technique for Lévy processes. The present transform version is extensively used in the next chapter. For an account of the history of the PRS factorization, we refer to Bertoin [43].

Proposition 11.5 (Pecherskiĭ-Rogozin-Spitzer) *We have for $\alpha \ge 0$, $\beta \in \mathbb{R}$, and $q > 0$,*

$$\begin{aligned}
\mathbb{E}e^{-\alpha e_q + i\beta X(e_q)} &= \mathbb{E}e^{-\alpha\overline{G}^X(e_q)+i\beta\overline{X}(e_q)}\mathbb{E}e^{-\alpha\underline{F}^X(e_q)+i\beta\underline{X}(e_q)}\\
&= \mathbb{E}e^{-\alpha\overline{F}^X(e_q)+i\beta\overline{X}(e_q)}\mathbb{E}e^{-\alpha\underline{G}^X(e_q)+i\beta\underline{X}(e_q)}.
\end{aligned}$$

This factorization is explicitly known in a number of important cases, for instance if X has one-sided jumps. We now discuss this class and other special classes of Lévy processes.

Important classes of Lévy processes

Monotonic Lévy processes are called *subordinators* [44], and they play an important role in the general theory (e.g., the ladder process is a subordinator). Note that the PRS factorization is trivial for subordinators.

Other well-studied Lévy processes are Lévy processes with one-sided jumps. If the Lévy measure of X is concentrated on the negative (or positive) halfline, then we call X a *spectrally negative* (or positive) Lévy process. We refer to Bertoin [43, Ch. VII] for results on Lévy processes in this totally asymmetric case.

If we suppose that X is spectrally negative (or spectrally positive by considering $-X$), it is possible to find the PRS factorization of Proposition 11.5 explicitly. Before writing down the factorization, we exclude the trivial situation that X is a subordinator. Since spectral negativity implies that $\Psi_X(\beta)$ is well-defined and analytic on $\{\Im(\beta) \le 0\}$ [43, p. 188], we introduce the *Laplace exponent*

$$\psi_X(\beta) := -\Psi_X(-i\beta),$$

so that $\mathbb{E}e^{\beta X(t)} = e^{t\psi_X(\beta)}$ for $\Re(\beta) \ge 0$. The mapping $\psi_X : [0,\infty) \mapsto \mathbb{R}$ is strictly convex, and $\psi_X(\beta) \to \infty$ as $\beta \to \infty$, since $\mathbb{P}(X(1) > 0) > 0$ (the subordinator case has been excluded). We denote by $\Phi_X(0)$ the largest solution to the equation $\psi_X(\beta) = 0$. Since the mapping

$\psi_X : [\Phi_X(0), \infty) \to [0, \infty)$ is a bijection, we can define the function $\Phi_X : [0, \infty) \to [\Phi_X(0), \infty)$ as its inverse, i.e., $\psi_X(\Phi_X(\beta)) = \beta$, $\beta \geq 0$. This function plays an important role throughout Part C.

The following proposition is Theorem VII.4.(i) of Bertoin [43]; for a martingale-based proof, see Kyprianou and Palmowski [208]. It is left to the reader to check that indeed

$$\mathbb{E}e^{-\alpha\overline{G}^X(e_q)+i\beta\overline{X}(e_q)}\mathbb{E}e^{-\alpha\underline{F}^X(e_q)+i\beta\underline{X}(e_q)} = \frac{q}{\alpha+q+\Psi_X(\beta)} = \mathbb{E}e^{-\alpha e_q+i\beta X(e_q)}.$$

Proposition 11.6 *Let X be spectrally negative, but not a subordinator. We have for any $\alpha, q > 0$, $\beta \geq 0$,*

$$\begin{aligned}\mathbb{E}e^{-\alpha\overline{G}^X(e_q)-\beta\overline{X}(e_q)} &= \frac{\Phi_X(q)}{\Phi_X(\alpha+q)+\beta},\\ \mathbb{E}e^{-\alpha\underline{F}^X(e_q)+\beta\underline{X}(e_q)} &= \frac{q\left(\Phi_X(\alpha+q)-\beta\right)}{\Phi_X(q)\left(\alpha+q-\psi_X(\beta)\right)},\end{aligned}$$

where the right-hand side should be interpreted as $q\Phi'_X(\beta)/\Phi_X(\beta)$ for $\beta = \Phi_X(\alpha+q)$.

As an important special case of this proposition, we obtain the Laplace transform of the vector $(\underline{F}^X(\infty), \underline{X}(\infty))$, in particular the *global* minimum. For this, we assume that X is integrable and drifts to $+\infty$; then, we let $q \to 0$ in Proposition 11.6 and use $q/\Phi_X(q) \to \mathbb{E}X$.

The *α-stable Lévy processes* constitute yet another important subclass of Lévy processes, but it is significantly harder to find the PRS factors, see Doney [118]. We have encountered α-stable Lévy processes already in Section 1.3.3. A Lévy process is α-stable if for every $t > 0$, the variables $X(t)$ and $t^{1/\alpha}X(1)$ have the same law. This is equivalent to saying that X is self-similar with index $1/\alpha$; see (1.12). The Lévy-Khinchine representation shows that we must have $\alpha \in (0, 2]$. A 2-stable Lévy process is necessarily proportional to Brownian motion, and a 1-stable Lévy process has (up to a drift) Cauchy distributed marginals, i.e., its characteristic exponent is $\lambda|\beta| + ci\beta$ for some $\lambda > 0$ and $c \in \mathbb{R}$. Let us now suppose that X is an α-stable Lévy process with $\alpha \in (0, 1) \cup (1, 2)$. The Lévy measure Π_X is then absolutely continuous with respect to the Lebesgue measure, with density

$$\frac{\Pi_X(dx)}{dx} = \begin{cases} c_+ x^{-\alpha-1} & \text{if } x > 0;\\ c_- |x|^{-\alpha-1} & \text{if } x < 0,\end{cases}$$

where $c_-, c_+ \geq 0$. Observe that the process of Section 1.3.3 has $\alpha \in (1, 2)$ and $c_- = 0$. Hence, apart from being α-stable, it is also spectrally positive. However, we stress that this process is not a subordinator, since the drift coefficient is positive. We also remark that an α-stable Lévy process is not integrable for $\alpha \in (0, 1]$.

More details and references on general stable processes (not necessarily Lévy) can be found in the book by Samorodnitsky and Taqqu [282]; see also Uchaikin and Zolotarev [298].

11.3 Outline of Part C

Chapter 12 gives three applications of an embedding based on Proposition 11.5. The first application studies the 'exact solutions' mentioned in Section 11.1.2 for general Lévy processes. Specifically, we impose a phase-type form for the Lévy measure on one halfline, while allowing for a general form on the other halfline. After this, we study a perturbed risk model, which is a variant of the classical risk model discussed in Section 1.1.5. As a third application, we establish the analogue of Theorem 11.4 for Lévy processes.

Chapter 13 continues on the splitting idea, but now within the framework of the fluid networks of Section 1.2. This naturally leads to splitting for multidimensional processes. We study the (steady-state) joint Laplace transform of the buffer contents and busy periods in each of the stations of the network. For a special tandem fluid queue, we also obtain the Laplace transform of the idle periods.

Chapter 14 investigates the extremes of a spectrally one-sided Lévy process in a random environment; the resulting process is known as a *Markov-additive process*. Such a process is governed by a Markovian background process (the 'environment'), and behaves like a spectrally one-sided Lévy process during the sojourn times of this background process; the characteristics of the Lévy process may differ for each of the background states. We derive analogues of the identities in Proposition 11.6 for $q \to 0$, and apply these to a fluid queue with Markov-additive input. We also analyze the networks of Chapter 13 in a random environment.

Chapters 12, 13, and 14 are based on [105], [89], and [114] respectively. Chapter 13 is joint work with Krzysztof Dębicki and Tomasz Rolski, while Chapter 14 is written with Michel Mandjes.

CHAPTER 12

Factorization embeddings

Using embeddings, we give three applications of the factorization identity in Proposition 11.5:

- Extremes of a Lévy process with phase-type upward jumps: we find the joint distribution of the maximum and the epoch at which it is 'attained' if a Lévy process has phase-type upward jumps.
- Perturbed risk models: we establish general properties, and obtain explicit fluctuation identities in case the Lévy process is spectrally positive.
- Tail asymptotics for the maximum of a Lévy process: we study the tail distribution of the maximum under different assumptions on the tail of the Lévy measure.

12.1 Introduction

It is the aim of this chapter to show how embedding and splitting can be used to study the fluctuations of a Lévy process. For this, we consider the sum Z of an arbitrary one-dimensional Lévy process X and a compound Poisson process Y with intensity λ, independent of X. Note that any discontinuous Lévy process Z can be written in this form; in this chapter, we are not interested in continuous Lévy processes (i.e., Brownian motions with drift), since their fluctuation theory is well-established. The representation $Z = X + Y$ need not be unique; for instance, there is a continuum of such representations if the Lévy measure has a nonvanishing absolutely continuous part. We remark that Y is not necessarily centered in Part C of this thesis, as opposed to Part A.

Before explaining the idea behind the embedding that we study in this chapter, we first introduce some notation. Write $T_1, T_2, \ldots$ for the jump epochs of Y, and set $T_0 = 0$. Define the quantities $\overline{G}_i$ and S_i for $i \geq 1$ as follows. $Z(T_{i-1}) + S_i$ stands for the value of the maximum within $[T_{i-1}, T_i)$, and $T_{i-1} + \overline{G}_i$ is the last epoch in this interval such that the value of Z at $T_{i-1} + \overline{G}_i$ or $(T_{i-1} + \overline{G}_i)-$ is $Z(T_{i-1}) + S_i$. Although formally incorrect, we say in the remainder that the maximum of Z over $[T_{i-1}, T_i)$ is attained at $T_{i-1} + \overline{G}_i$, with value $Z(T_{i-1}) + S_i$.

In the left-hand diagram of Figure 12.1, a realization of Z is given. The jumps of Y are dotted and those of X are dashed. The process Z is killed at an exponentially distributed random time $\mathtt{k}$ independent of Z, say with parameter $q \geq 0$ ($q = 0$ corresponds to no killing). The right-hand diagram in Figure 12.1 is obtained from the first by replacing the trajectory of Z between T_{i-1} and T_i by a piecewise straight line consisting of two pieces: one from

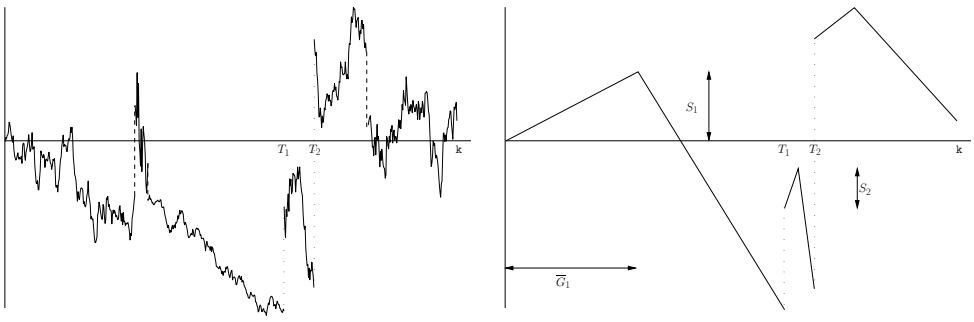

Figure 12.1: A realization of the killed Lévy process $Z = X+Y$ and the corresponding embedded (piecewise-linear) process. Jumps of Y are dotted and jumps of X are dashed.

$(T_{i-1}, Z(T_{i-1}))$ to $(T_{i-1} + \overline{G}_i, Z(T_{i-1}) + S_i)$, and one from the latter point to $(T_i, Z(T_i-))$. Obviously, by considering the embedded piecewise-linear process, no information is lost on key fluctuation quantities like the global maximum of Z and the epoch at which it is attained for the last time.

The piecewise-linear process, however, has several useful properties. First, by the Markov property, the 'hats' are mutually independent given their starting point. Moreover, obviously, the jumps of Y are independent of the 'hats'. More strikingly, the increasing and decreasing pieces of each 'hat' are also independent; indeed, $(T_i - T_{i-1}, Z(T_i-) - Z(T_{i-1})) = (\overline{G}_i, S_i) + (T_i - T_{i-1} - \overline{G}_i, Z(T_i-) - Z(T_{i-1}) - S_i)$, where the two latter vectors are independent, cf. the Pecherskiĭ-Rogozin-Spitzer factorization for Lévy processes (Proposition 11.5). This explains the name *factorization embedding*.

The right-hand diagram in Figure 12.1 can be generated without knowledge of the trajectory of Z. Indeed, since $\{T_i : i \geq 1\}$ is a Poisson point process with intensity λ and killing at rate q, it is equivalent (in law) to the first N points of a Poisson point process with intensity $\lambda + q$, where N is geometrically distributed on $\mathbb{Z}_+$ with parameter $\lambda/(\lambda + q)$ (independent of the point process).

It is not a new idea to consider an embedded process for studying fluctuations of Lévy processes. A classical example with $q = 0$ is when X is a negative drift ($X(t) = ct$ for some $c < 0$) and Y only has positive jumps. We then have that $\overline{G}_i \equiv 0$ for every i and $(\overline{G}_i, S_i) + (T_i - T_{i-1} - \overline{G}_i, Z(T_i-) - Z(T_{i-1}) - S_i)$ is distributed as (e_λ, ce_λ), where e_λ denotes an exponentially distributed random variable with parameter λ. Therefore, a random walk can be studied in order to analyze the fluctuations of Z. To the author's knowledge, nontrivial factorization embeddings have only been used to obtain results in the space domain. We mention the work of Asmussen [16] and Kennedy [188], who study a Wiener-Hopf problem with Brownian noise, and the work of Mordecki [237], who investigates the maximum of a Lévy process with phase-type upward jumps and general downward jumps. Recently, a slightly different form of this embedding has been used by Doney [119] to derive stochastic bounds on the Lévy processes Z. He defines X and Y such that the supports of Π_X and Π_Y are disjoint, and notes that $\{Z(T_{i-1}) + S_i\}$ is a random walk with a random starting point, so that it suffices to establish stochastic bounds on the starting point. Doney then uses these to analyze the asymptotic behavior of Lévy processes that converge to $+\infty$ in probability. As an aside, we remark that the factorization embedding is different from the embedding that has been used in [20, 256], where the authors work with a process in a random environment to replace jumps by drifts. In Chapter 14, we extensively study Lévy processes in a random environment.

Outline and contribution of the chapter; three applications

We now describe how this chapter is organized, thereby introducing three problems that are studied using factorization embeddings. All our results are new, with the only exception of Theorem 12.11 and the first claim in Theorem 12.13.

Section 12.2 uses the above embedding idea to express fluctuation quantities of Z in terms of those of X. In Section 12.3, we apply these results to study the case where Z has phase-type upward jumps and general downward jumps. Then, the Laplace exponent of the bivariate ladder process κ_Z can be given; this quantity lies at the heart of fluctuation theory for Lévy processes, see Chapter VI of Bertoin [43]. In particular, we give the joint law of the maximum and the epoch at which it is 'attained', generalizing Mordecki's [237] results.

Section 12.4 studies perturbed risk models, a generalization of the classical risk model of Section 1.1.5 that has drawn much attention in the literature. We prove a general Pollaczek-Khinchine formula in this framework, but more explicit results can only be obtained under further assumptions. Therefore, we impose spectral positivity of the Lévy process underlying the risk model, and extend the recent results of Huzak *et al.* [166] in the following sense. While [166] focuses on quantities related to so-called modified ladder heights, we obtain *joint* distributions related to both the modified ladder epoch and the modified ladder height. In particular, we obtain the (transform of the) distribution of the first modified ladder epoch.

Section 12.5 studies the tail of the maximum of Z under three different assumptions on the Lévy measure. We reproduce known results in the Cramér case and the subexponential case, but also give a local variant in the latter case, which is new. Our results for the intermediate case are also new, and complement recent work of Klüppelberg *et al.* [192].

After finishing this chapter, we learned about recent work of Pistorius [256], and there is some overlap between his work and Section 12.3 in the special case $K = 1$. In [256], the Laplace exponent κ_Z of the ladder process is characterized in terms of the solutions of the equation $\Psi_Z(\beta) = q$, where Ψ_Z is the characteristic exponent of Z, cf. (11.1). Our approach is different, since we express $\kappa_Z(q, \beta)$ in terms of a vector $\boldsymbol{\alpha}^q_+$, for which we give an efficient algorithm.

12.2 On factorization identities

In this section, we consider the process $Z = X + Y$, where Y is a compound Poisson process and X is a general Lévy process, independent of Y. We study the extremes of Z and the epoch at which they are attained for the first (or last) time, in terms of the corresponding distribution of X. Moreover, the characteristics of the bivariate ladder process of Z are expressed in those of X.

In order to relate the PRS factors of Z and X, we need an auxiliary random walk. We write $\lambda \in [0, \infty)$ for the intensity of Y, and ξ for its generic jump. For fixed $q > 0$, let $\{S^q_n\}$ be a random walk with step-size distribution $\xi + X(e_{\lambda+q})$, where the two summands are independent. For this random walk, we define the first strict ascending (descending) ladder epoch $\tau^q_{s\pm}$ as

$$\tau^q_{s+} := \inf\{n \geq 1 : S^q_n > 0\}, \quad \tau^q_{s-} := \inf\{n \geq 1 : S^q_n < 0\},$$

and $\tau^q_{w\pm}$ is defined similarly with a weak inequality. In analogy to the previous chapter, we write $H^q_{w\pm}$ ($H^q_{s\pm}$) for the ladder height $S^q_{\tau^q_{w\pm}}$ ($S^q_{\tau^q_{s\pm}}$). Recall that the integration is only carried out over the set where the random variables are both finite and well-defined.

The main result of this section, which we now state, relates the PRS factors of Z and X. When a specific structure is imposed on X and Y, both factors can be expressed in terms of solutions to fixed-point equations; see Section 12.3. Intuitively, a PRS factor of Z is the product of a PRS factor of X and a random-walk PRS factor. The main complication is that

the random walk is converted to a continuous-time process by 'stretching' time, but that this stretching is not done independently of the step size.

Theorem 12.1 *For every $\alpha, \beta, q > 0$, we have*

$$\mathbb{E}e^{-\alpha\overline{G}^Z(e_q)-\beta\overline{Z}(e_q)} = \mathbb{E}e^{-\alpha\overline{G}^X(e_{\lambda+q})-\beta\overline{X}(e_{\lambda+q})}\frac{1-\mathbb{E}\left(\frac{\lambda}{\lambda+q}\right)^{\tau^q_{w+}}}{1-\mathbb{E}e^{-\beta H^{q+\alpha}_{w+}}\left(\frac{\lambda}{\lambda+q+\alpha}\right)^{\tau^{q+\alpha}_{w+}}},$$

$$\mathbb{E}e^{-\alpha\overline{F}^Z(e_q)-\beta\overline{Z}(e_q)} = \mathbb{E}e^{-\alpha\overline{F}^X(e_{\lambda+q})-\beta\overline{X}(e_{\lambda+q})}\frac{1-\mathbb{E}\left(\frac{\lambda}{\lambda+q}\right)^{\tau^q_{s+}}}{1-\mathbb{E}e^{-\beta H^{q+\alpha}_{s+}}\left(\frac{\lambda}{\lambda+q+\alpha}\right)^{\tau^{q+\alpha}_{s+}}},$$

and $\mathbb{E}e^{-\alpha\underline{F}^Z(e_q)+i\beta\underline{Z}(e_q)}$, $\mathbb{E}e^{-\alpha\underline{G}^Z(e_q)+i\beta\underline{Z}(e_q)}$ follow by application of these formulas to $-Z$.

Proof. We only prove the first equality; the argument is easily adapted to obtain the second.

The first factor is a direct consequence of the independence of the first straight line in the right-hand diagram of Figure 12.1 and the other pieces; see the remarks accompanying Figure 12.1. Writing for $i \geq 1$, $w_i := T_{i-1} + \overline{G}_i - \overline{G}_1$ and $W_i := Z(T_{i-1} + \overline{G}_i) - S_1$, these arguments also yield that $\{W_i : i \geq 1\}$ is a random walk with the same distribution as $\{S^q_n : n \geq 0\}$, except for the killing in every step with probability $\lambda/(\lambda+q)$. Therefore, if we define the first (weak) ascending ladder epoch of this random walk

$$N := \inf\{i \geq 1 : W_i \geq 0\},$$

we have

$$\mathbb{P}(N < \infty) = \mathbb{E}\left(\frac{\lambda}{\lambda+q}\right)^{\tau^q_{w+}}.$$

Observe that $(\overline{G}^Z(e_q) - \overline{G}_1, \overline{Z}(e_q) - S_1)$ has the same distribution as $\sum_{j=1}^K (w^j_N, W^j_N)$, where K is geometrically distributed on $\mathbb{Z}_+$ with parameter $\mathbb{P}(N < \infty)$, and (w^j_N, W^j_N) are independent copies of (w_N, W_N), also independent of K. Note that we consider the weak ladder epoch in the definition of N, since we are interested in $\overline{G}^Z(e_q)$ (as opposed to $\overline{F}^Z(e_q)$). This shows that

$$\mathbb{E}e^{-\alpha\left(\overline{G}^Z(e_q)-\overline{G}_1\right)+i\beta\left(\overline{Z}(e_q)-S_1\right)} = \frac{1-\mathbb{E}\left(\frac{\lambda}{\lambda+q}\right)^{\tau^q_{w+}}}{1-\mathbb{E}\left(\frac{\lambda}{\lambda+q}\right)^{N}e^{-\alpha w_N+i\beta W_N}},$$

and it remains to study the denominator in more detail.

For this, we rely on Section I.1.12 of Prabhu [263]. The key observation is that $\{(w_i, W_i)\}$ is a random walk in the halfplane $\mathbb{R}_+ \times \mathbb{R}$, with step-size distribution characterized by

$$\mathbb{E}e^{-\alpha w_1+i\beta W_1} = \mathbb{E}e^{-\alpha e_{\lambda+q}+i\beta X(e_{\lambda+q})}\mathbb{E}e^{i\beta\xi}.$$

Theorem 27 of [263], which is a Wiener-Hopf factorization for random walks on the halfplane, shows that we may write for $|z| < 1$ and $\alpha \geq 0$, $\beta \in \mathbb{R}$,

$$1 - z\mathbb{E}e^{-\alpha e_{\lambda+q}+i\beta X(e_{\lambda+q})}\mathbb{E}e^{i\beta\xi} = \left[1 - \mathbb{E}z^N e^{-\alpha w_N+i\beta W_N}\right]\left[1 - \mathbb{E}z^{\bar{N}}e^{-\alpha w_{\bar{N}}+i\beta W_{\bar{N}}}\right],$$

where the bars refer to (strict) descending ladder variables. Note that Theorem 11.1 is recovered by setting $\alpha = 0$. The actual definitions of these quantities are of minor importance to us;

the crucial point is that this factorization is unique. Indeed, an alternative characterization is obtained by conditioning on the value of $e_{\lambda+q}$:

$$1 - z\mathbb{E}e^{-\alpha e_{\lambda+q}+i\beta X(e_{\lambda+q})}\mathbb{E}e^{i\beta\xi} = 1 - \frac{(\lambda+q)z}{\lambda+q+\alpha}\mathbb{E}e^{i\beta X(e_{\lambda+q+\alpha})}\mathbb{E}e^{i\beta\xi},$$

and the Wiener-Hopf factorization for random walks (Theorem 11.1) shows that this can be written as

$$\left[1 - \mathbb{E}\left(\frac{(\lambda+q)z}{\lambda+q+\alpha}\right)^{\tau_{s+}^{q+\alpha}} e^{i\beta H_{s+}^{q+\alpha}}\right]\left[1 - \mathbb{E}\left(\frac{(\lambda+q)z}{\lambda+q+\alpha}\right)^{\tau_{w-}^{q+\alpha}} e^{i\beta H_{w-}^{q+\alpha}}\right].$$

This decomposition is again unique, so that the claim follows by substituting $z = \lambda/(\lambda+q)$. □

If $\alpha = 0$, we must have $\mathbb{E}e^{-\alpha\overline{G}^Z(e_q)-\beta\overline{Z}(e_q)} = \mathbb{E}e^{-\alpha\overline{F}^Z(e_q)-\beta\overline{Z}(e_q)}$, but the formulas in Theorem 12.1 differ in the sense of weak and strict ladder variables. This is not a contradiction, as Spitzer's identity shows that the fractions are equal for both τ_{w+}^q and τ_{s+}^q.

Let us now verify that the formulas of Theorem 12.1 are in accordance with the PRS factorization of Proposition 11.5. Indeed, application of Theorems 11.1 and 12.1 (the transform $\mathbb{E}e^{-\alpha\underline{F}^Z(e_q)+i\beta\underline{Z}(e_q)}$ is obtained by considering $-Z$) yields

$$\begin{aligned}&\mathbb{E}e^{-\alpha\overline{G}^Z(e_q)+i\beta\overline{Z}(e_q)}\mathbb{E}e^{-\alpha\underline{F}^Z(e_q)+i\beta\underline{Z}(e_q)}\\ &\quad= \mathbb{E}e^{-\alpha e_{\lambda+q}+i\beta X(e_{\lambda+q})}\frac{1-\frac{\lambda}{\lambda+q}}{1-\frac{\lambda}{\lambda+q+\alpha}\mathbb{E}e^{i\beta X(e_{\lambda+q+\alpha})}\mathbb{E}e^{i\beta\xi}}.\end{aligned}$$

By conditioning on the value of $e_{\lambda+q}$ in the first factor, it is readily seen that this equals

$$\frac{q}{\frac{\lambda+q+\alpha}{\mathbb{E}e^{i\beta X(e_{\lambda+q+\alpha})}} - \lambda\mathbb{E}e^{i\beta\xi}} = \frac{q}{\lambda+q+\alpha+\Psi_X(\beta)-\lambda\mathbb{E}e^{i\beta\xi}} = \mathbb{E}e^{-\alpha e_q+i\beta Z(e_q)}.$$

Given Theorem 12.1, one can easily deduce the characteristics of the *ladder height process* of Z in terms of those of X; as the notions are standard, we refer to p. 157 of Bertoin [43] for definitions. The importance of this two-dimensional subordinator has recently been illustrated by Doney and Kyprianou [121].

The dual processes of Z and X are defined as $\hat{Z} = -Z$ and $\hat{X} = -X$ respectively.

Corollary 12.2 *For $\alpha, \beta \geq 0$, we have*

$$\kappa_Z(\alpha,\beta) = \kappa_X(\lambda+\alpha,\beta)\left(1 - \mathbb{E}e^{-\beta H_{s+}^{\alpha}}\left(\frac{\lambda}{\lambda+\alpha}\right)^{\tau_{s+}^{\alpha}}\right),$$

and

$$\begin{aligned}\hat{\kappa}_Z(\alpha,\beta) &= k\hat{\kappa}_X(\lambda+\alpha,\beta)\left(1 - \mathbb{E}e^{\beta H_{w-}^{\alpha}}\left(\frac{\lambda}{\lambda+\alpha}\right)^{\tau_{w-}^{\alpha}}\right)\\ &= k\frac{\alpha+\Psi_Z(-i\beta)}{\kappa_X(\lambda+\alpha,-\beta)\left[1-\mathbb{E}e^{\beta H_{s+}^{\alpha}}\left(\frac{\lambda}{\lambda+\alpha}\right)^{\tau_{s+}^{\alpha}}\right]},\end{aligned}$$

where k is some meaningless constant.

Proof. It suffices to note that $\kappa_Z(\alpha,-i\beta)\hat{\kappa}_Z(\alpha,i\beta) = k(\alpha+\Psi_Z(\beta))$ by the Wiener-Hopf factorization for random walks (Theorem 11.1), and to continue $\hat{\kappa}_Z$ analytically. □

12.3 Lévy processes with phase-type upward jumps

In this section, we use the results of the previous section to study Lévy processes with phase-type upward jumps and general downward jumps. A motivation for investigating this process is provided by Boucherie *et al.* [58]. According to the results in the previous section, $(\overline{G}^Z(e_q), \overline{Z}(e_q))$ can be written as the sum of $(\overline{G}^X(e_{\lambda+q}), \overline{X}(e_{\lambda+q}))$ and an (independent) random-walk term. In this section, we choose X and Y appropriately, so that the transforms of both vectors can be computed explicitly.

For this, we let X be an arbitrary spectrally negative Lévy process, and Y is a compound Poisson process (not necessarily a subordinator), independent of X, for which the upward jumps have a phase-type distribution. The exact form of the Lévy measure of Y is specified in (12.1) below.

Apart from their computational convenience, the most important property of phase-type distributions is that they are dense, in the sense of weak convergence (Definition 2.9), within the class of probability measures (although many phases may be needed to approximate a stable distribution, for instance). A phase-type distribution is the absorption time of a Markov process on a finite state space E. Its intensity matrix is determined by the $|E| \times |E|$-matrix $\boldsymbol{T}$, and its initial distribution is denoted by $\boldsymbol{\alpha}$. For more details on phase-type distributions, we refer to Asmussen [19, Sec. III.4]. We write $\boldsymbol{t} = -\boldsymbol{T}\mathbf{1}$, where $\mathbf{1}$ is the vector with ones. All vectors are column vectors.

Fluctuation theory for Lévy processes with phase-type jumps has recently been studied by Asmussen *et al.* [20] and Mordecki [237]; see also Kou and Wang [200]. As a consequence of the fact that phase-type distributions are dense within the class of probability measures, an arbitrary Lévy process can be written as the limit of a sequence of Lévy processes with phase-type jumps (in the Skorokhod topology on $D(\mathbb{R}_+)$; see, e.g., [170, Ch. VI] for definitions). In both [20] and [237], the authors obtain expressions for the Laplace transform of $\overline{Z}(e_q)$ if Y is a compound Poisson process with only positive (phase-type) jumps.

While the class of processes that we analyze here is slightly more general, the main difference with the aforementioned papers is that we calculate the Laplace transform of the *joint* distribution $(\overline{G}^Z(e_q), \overline{Z}(e_q))$; see Section 12.3. In particular, if a Lévy process Z has phase-type upward jumps, one can compute the epoch at which the maximum is attained; the latter is perhaps more surprising than the fact that it is possible to find an expression for the distribution of $\overline{Z}(e_q)$. This illustrates why Theorem 12.1 is interesting.

To the author's knowledge, the results in this section cover any Lévy process for which this joint distribution is known. The only case for which results are available but not covered here is when Z is a certain stable Lévy process; see Doney [118]. Then, only the distribution of the (marginal) law of $\overline{Z}(e_q)$ is known in a semi-explicit form.

The PRS factorization

We begin with a detailed description of the process Y. Given $K \in \mathbb{N}$, suppose that we have nonnegative random variables $\{A_j : j = 1, \ldots, K\}$ and $\{B_j : j = 1, \ldots, K\}$, where the distribution $\mathbb{P}_{B_j}$ of B_j is phase-type with representation $(E_j, \boldsymbol{\alpha}_j, \boldsymbol{T}_j)$. The distribution $\mathbb{P}_{-A_j}$ of $-A_j$ is general; the only restriction we impose is that $\mathbb{P}_{-A_j} * \mathbb{P}_{B_j}(\{0\}) = 0$ for all j, i.e., A_j and B_j do not both have an atom at zero. We assume that the process Y is a compound Poisson process with Lévy measure given by

$$\Pi_Y := \lambda \sum_{j=1}^{K} \pi_j \mathbb{P}_{B_j} * \mathbb{P}_{-A_j}, \tag{12.1}$$

where $\lambda \in (0,\infty)$, $0 \le \pi_j \le 1$ with $\sum \pi_j = 1$. In queueing theory, processes of this form arise naturally, and the B can be interpreted as the service times and the A as interarrival times. Note that Y is a subordinator if and only if Π_Y can be written as (12.1) with $K = 1$ and $A_1 \equiv 0$.

Without loss of generality, we may assume that E_j and $\boldsymbol{T}_j$ do not depend on j. Indeed, if E_j has m_j elements, one can construct an E with $\sum_{j=1}^K m_j$ elements and $\boldsymbol{T}$ can then be chosen as a block diagonal matrix with the matrices $\boldsymbol{T}_1,\dots,\boldsymbol{T}_K$ on its diagonal. The vectors $\boldsymbol{\alpha}_j$ are then padded with zeros, so that they consist of K parts of lengths $m_1,\dots,m_K$, and only the j-th part is nonzero.

Fix some $q > 0$; our first aim is to study the random walk $\{S_n^q\}$ introduced in Section 12.2, with generic step-size distribution (by the PRS factorization)

$$\mathbb{P}_{S_1^q} := \mathbb{P}_{\overline{X}(e_{\lambda+q})} * \mathbb{P}_{\underline{X}(e_{\lambda+q})} * \mathbb{P}_\xi,$$

where $\mathbb{P}_\xi = \Pi_Y/\lambda$. We exclude the case for which Z is a subordinator, so that $\mathbb{P}_{S_1^q}$ always assigns strictly positive probability to $\mathbb{R}_+$. In that case, there is no distinction between weak and strict ascending ladder heights, and we therefore write τ_+^q for $\tau_{w+}^q = \tau_{s+}^q$ throughout this section.

Since $\overline{X}(e_{\lambda+q})$ is either degenerate or exponentially distributed, the law of S_1^q can be written as $\sum \pi_j \mathbb{P}_{\widetilde{B}_j(q)} * \mathbb{P}_{\widetilde{A}_j(q)}$, where $\widetilde{B}_j(q)$ has again a phase-type distribution, say with representation $(\widetilde{E}_q, \widetilde{\boldsymbol{\alpha}}_j(q), \widetilde{\boldsymbol{T}}_q)$. It is not hard to express this triple in terms of the original triple $(E, \boldsymbol{\alpha}_j, \boldsymbol{T})$: $(\widetilde{E}_q, \widetilde{\boldsymbol{\alpha}}_j(q), \widetilde{\boldsymbol{T}}_q) = (E, \boldsymbol{\alpha}_j, \boldsymbol{T})$ if X is a nonincreasing subordinator, and otherwise $\widetilde{E}_q$ can be chosen such that $|\widetilde{E}_q| = |E| + 1$, and the dynamics of the underlying Markov chain are unchanged, except for the fact that an additional state is visited before absorption. We set $\widetilde{\boldsymbol{t}}_q = -\widetilde{\boldsymbol{T}}_q \mathbf{1}$.

Motivated by Theorem 12.1, the following lemma calculates the transform of the ladder variables (H_+^q, τ_+^q); recall that the random variables are only integrated over the subset $\{\tau_+^q < \infty\}$ of the probability space.

Lemma 12.3 *Let $\rho \in (0,1)$ and $\beta \ge 0$. Then there exists some vector $\boldsymbol{\alpha}_+^{\rho,q}$ such that*

$$\mathbb{E}\rho^{\tau_+^q} e^{-\beta H_+^q} = \left[\boldsymbol{\alpha}_+^{\rho,q}\right]' (\beta \boldsymbol{I} - \widetilde{\boldsymbol{T}}_q)^{-1} \widetilde{\boldsymbol{t}}_q.$$

Proof. The proof is similar to the proofs of Lemma VIII.5.1 and Proposition VIII.5.11 of Asmussen [19]; the details are left to the reader. □

The above lemma shows that it is of interest to be able to calculate $\boldsymbol{\alpha}_+^{\rho,q}$. Therefore, we generalize Theorem VIII.5.12 in [19] to the present setting. We omit a proof, as similar arguments apply; the only difference is that we allow for $K > 1$ and that the random walk can be killed in every step with probability ρ.

Proposition 12.4 *$\boldsymbol{\alpha}_+^{\rho,q}$ satisfies $\boldsymbol{\alpha}_+^{\rho,q} = \boldsymbol{\varphi}(\boldsymbol{\alpha}_+^{\rho,q})$, where*

$$\boldsymbol{\varphi}(\boldsymbol{\alpha}_+^{\rho,q}) = \rho \sum_{j=1}^K \pi_j \widetilde{\boldsymbol{\alpha}}_j'(q) \int_0^\infty \exp\left(\left[\widetilde{\boldsymbol{T}}_q + \widetilde{\boldsymbol{t}}_q \left[\boldsymbol{\alpha}_+^{\rho,q}\right]'\right] y\right) \widetilde{A}_j(q)(dy).$$

It can be computed as $\lim_{n\to\infty} \boldsymbol{\alpha}_+^{\rho,q}(n)$, where $\boldsymbol{\alpha}_+^{\rho,q}(0) = \mathbf{0}$ and $\boldsymbol{\alpha}_+^{\rho,q}(n) = \boldsymbol{\varphi}(\boldsymbol{\alpha}_+^{\rho,q}(n-1))$ for $n \ge 1$.

The main result of this section follows by combining Theorem 12.1 with Lemma 12.3 and the fluctuation identities of Proposition 11.6. Since Theorem 12.1 directly applies if X is a subordinator, this is excluded to focus on the most interesting case. For notational convenience, we write $\boldsymbol{\alpha}_+^q$ for $\boldsymbol{\alpha}_+^{q/(\lambda+q),q}$. Recall that Φ_X denotes the inverse of the Laplace exponent of X, see Section 11.2.

Theorem 12.5 *Suppose that both X and Z are no subordinators. Then we have for $\alpha, \beta \geq 0$,*

$$\mathbb{E}e^{-\alpha\overline{G}^Z(e_q)-\beta\overline{Z}(e_q)} = \frac{\Phi_X(\lambda+q)\left[1-\left[\boldsymbol{\alpha}_+^q\right]'\mathbf{1}\right]}{\left[\Phi_X(\lambda+q+\alpha)+\beta\right]\left[1-\left[\boldsymbol{\alpha}_+^{q+\alpha}\right]'(\beta\boldsymbol{I}-\widetilde{\boldsymbol{T}}_{q+\alpha})^{-1}\widetilde{\boldsymbol{t}}_{q+\alpha}\right]}$$

and

$$\mathbb{E}e^{-\alpha\overline{F}^Z(e_q)+\beta\underline{Z}(e_q)} = \frac{q\left[\Phi_X(\lambda+q+\alpha)-\beta\right]\left[1+\left[\boldsymbol{\alpha}_+^{q+\alpha}\right]'(\beta\boldsymbol{I}+\widetilde{\boldsymbol{T}}_{q+\alpha})^{-1}\widetilde{\boldsymbol{t}}_{q+\alpha}\right]}{\Phi_X(\lambda+q)\left[q+\alpha+\Psi_Z(-i\beta)\right]\left[1-\left[\boldsymbol{\alpha}_+^{q}\right]'\mathbf{1}\right]}.$$

While Theorem 12.5 is an immediate consequence of Theorem 12.1, we now state the corresponding analogue of Corollary 12.2. Note that the expression for $\kappa_Z(0,\beta)$ is already visible in the work of Mordecki [237]; here, we obtain a full description of κ_Z.

Corollary 12.6 *Under the assumptions of Theorem 12.5, we have for $\alpha, \beta \geq 0$,*

$$\kappa_Z(\alpha,\beta) = \left[\Phi_X(\lambda+\alpha)+\beta\right]\left[1-\left[\boldsymbol{\alpha}_+^{\alpha}\right]'(\beta\boldsymbol{I}-\widetilde{\boldsymbol{T}}_{\alpha})^{-1}\widetilde{\boldsymbol{t}}_{\alpha}\right],$$

and

$$\hat{\kappa}_Z(\alpha,\beta) = k\frac{\alpha+\Psi_Z(-i\beta)}{\left[\Phi_X(\lambda+\alpha)-\beta\right]\left[1+\left[\boldsymbol{\alpha}_+^{\alpha}\right]'(\beta\boldsymbol{I}+\widetilde{\boldsymbol{T}}_{\alpha})^{-1}\widetilde{\boldsymbol{t}}_{\alpha}\right]},$$

where k is a meaningless constant.

12.4 Perturbed risk models

Let X be an arbitrary Lévy process and Y be a compound Poisson process with intensity λ and generic positive jump ξ. In this section, we suppose that $Z = X + Y$ drifts to $-\infty$. Classical risk theory studies the maximum of Z in case X is a negative drift, i.e., $X(t) = -ct$ for some $c > \lambda\mathbb{E}\xi$, see Section 1.1.5. Then, its distribution is given by the Pollaczek-Khinchine formula. In this analysis, a key role is played by ladder epochs and heights, i.e., quantities related to the event that Z reaches a new record.

In this section, we replace the negative drift X by an arbitrary Lévy process; in the literature, this is known as a perturbed risk model; see [137, 166, 285] and references therein. We also refer to Rolski *et al.* [275, Sec. 13.2] for a textbook treatment. To analyze this model, the classical ladder epochs and heights are replaced by so-called modified ladder epochs and heights; these are related to the event that Z reaches a new record as a result of a jump of Y.

In Huzak *et al.* [166], Y is allowed to be a general subordinator, not necessarily of the compound Poisson type. Therefore, the perturbed risk models studied here are slightly less general. However, since any subordinator can be approximated by compound Poisson processes, one is led to believe that our results also hold in the general case. Since the approximation argument required for proving this is not in the spirit of this chapter, we do not address this issue here. Instead, we shall content ourselves with writing the main results (Proposition 12.7, Theorem 12.8, and Theorem 12.10) in a form that does not rely on Y being compound Poisson, although this assumption is essential for our proofs.

In Section 12.4.1, we derive a Pollaczek-Khinchine formula for perturbed risk models. Unfortunately, the formula is not so explicit. Therefore, we impose further assumptions in Section 12.4.2, where we study spectrally positive Z.

As mentioned already, a central role in perturbed risk models is played by the first time χ a new maximum is reached by a jump of Y, i.e.,

$$\chi := \inf\{t > 0 : \Delta Y_t > \overline{Z}(t-) - Z(t-)\}.$$

In Figure 12.1, we have $\chi = T_2$. On the event $\{\chi = \infty\}$, we define $(\overline{G}^Z(\chi-), \overline{Z}(\chi-))$ as $(\overline{G}^Z(\infty), \overline{Z}(\infty))$.

12.4.1 Generalities

In this subsection, we study the structure of a general perturbed risk model, i.e., we consider a *general* Lévy perturbation X. The results that we obtain are new in this generality.

The following proposition is crucial for our analysis.

Proposition 12.7 *We have*

(i) $(\overline{G}^Z(\chi-), \overline{Z}(\chi-))$ *is independent of* $\{\chi < \infty\}$,

(ii) $(\overline{G}^Z(\chi-), \overline{Z}(\chi-))$ *is distributed as* $(\overline{G}^Z(\infty), \overline{Z}(\infty))$ *given* $\{\chi = \infty\}$, *and*

(iii) $(Z(\chi) - \overline{Z}(\chi-), \overline{Z}(\chi-) - Z(\chi-), \chi - \overline{G}^Z(\chi-))$ *is conditionally independent of the vector* $(\overline{G}^Z(\chi-), \overline{Z}(\chi-))$ *given* $\{\chi < \infty\}$.

Proof. We need some definitions related to the piecewise linear (jump) process of Figure 12.1, in particular to its excursions. Let $\tilde{\mathbb{P}}$ denote the law of the piecewise linear process that is constructed by discarding the first increasing piece (which may not be present if X is a negative subordinator), and let $\tilde{\mathbb{E}}$ denote the corresponding integration operator. Under $\tilde{\mathbb{P}}$, there are two possibilities for the process to (strictly) cross the axis: it either crosses continuously or it jumps over it. The event that the first happens is denoted by $\mathcal{X}$, as it is caused by fluctuations in X. We write $\mathcal{Y}$ for the second event. The probability of no crossing (i.e., no new record) is then given by $1 - \tilde{\mathbb{P}}(\mathcal{X}) - \tilde{\mathbb{P}}(\mathcal{Y})$. Moreover, by the strong Markov property, we have

$$\mathbb{P}(\chi < \infty) = \frac{\tilde{\mathbb{P}}(\mathcal{Y})}{1 - \tilde{\mathbb{P}}(\mathcal{X})}. \tag{12.2}$$

On $\mathcal{X}$ and $\mathcal{Y}$, we also define the 'excursion lengths' L_e and 'ascending ladder heights' H_e in a self-evident manner (on $\mathcal{X}$, H_e is strictly positive and equals the maximum within the first ascending 'ladder hat'). Moreover, we let U_e be the 'undershoot' on $\mathcal{Y}$; see Figure 12.2. The dotted line is the piece that is discarded under $\tilde{\mathbb{P}}$.

For $\alpha, \beta \geq 0$, by the strong Markov property,

$$\begin{aligned}
&\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)}; \chi < \infty\right] \\
&\quad = \mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)-\beta\overline{X}(e_\lambda)}\tilde{\mathbb{P}}(\mathcal{Y}) + \tilde{\mathbb{E}}\left[e^{-\alpha L_e-\beta H_e}; \mathcal{X}\right]\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)}; \chi < \infty\right],
\end{aligned}$$

from which we obtain

$$\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)}; \chi < \infty\right] = \mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)-\beta\overline{X}(e_\lambda)}\frac{\tilde{\mathbb{P}}(\mathcal{Y})}{1 - \tilde{\mathbb{E}}\left[e^{-\alpha L_e-\beta H_e}; \mathcal{X}\right]}.$$

Along the same lines, one can deduce that

$$\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\infty)-\beta\overline{Z}(\infty)}; \chi = \infty\right] = \mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)-\beta\overline{X}(e_\lambda)}\frac{1 - \tilde{\mathbb{P}}(\mathcal{X}) - \tilde{\mathbb{P}}(\mathcal{Y})}{1 - \tilde{\mathbb{E}}\left[e^{-\alpha L_e-\beta H_e}; \mathcal{X}\right]},$$

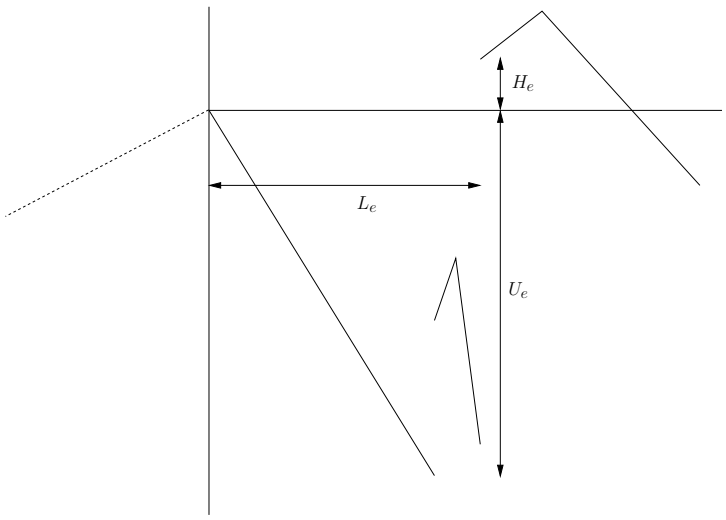

Figure 12.2: The excursion quantities in the proof of Proposition 12.7.

so that $\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)}\middle|\chi<\infty\right]$ equals

$$\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)};\chi<\infty\right]+\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\infty)-\beta\overline{Z}(\infty)};\chi=\infty\right],$$

which is $\mathbb{E}e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)}$; this is the first claim. These calculations also show that

$$\begin{aligned}&\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\infty)-\beta\overline{Z}(\infty)}\middle|\chi=\infty\right]\\ &\quad=\quad\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)};\chi<\infty\right]+\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\infty)-\beta\overline{Z}(\infty)};\chi=\infty\right],\end{aligned}$$

which is the second claim.

For the third claim, a variant of the above argument can be used to see that for $\alpha,\beta,\gamma,\delta,\varepsilon\geq 0$,

$$\begin{aligned}&\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)}e^{-\gamma\left[Z(\chi)-\overline{Z}(\chi-)\right]}e^{-\delta\left[\overline{Z}(\chi-)-Z(\chi-)\right]}e^{-\varepsilon\left[\chi-\overline{G}^Z(\chi-)\right]}\middle|\chi<\infty\right]\\ &\quad=\quad\mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)-\beta\overline{X}(e_\lambda)}\frac{1-\tilde{\mathbb{P}}(\mathcal{X})}{1-\tilde{\mathbb{E}}\left[e^{-\alpha L_e-\beta H_e};\mathcal{X}\right]}\frac{\tilde{\mathbb{E}}\left[e^{-\gamma H_e}e^{\delta U_e}e^{-\varepsilon L_e};\mathcal{Y}\right]}{\tilde{\mathbb{P}}(\mathcal{Y})},\end{aligned}$$

and the claim follows. □

The formula in the following theorem can be viewed as a generalized Pollaczek-Khinchine formula for perturbed risk models. It is a consequence of the preceding proposition and the observation that by the strong Markov property,

$$\begin{aligned}&\mathbb{E}e^{-\alpha\overline{G}^Z(\infty)-\beta\overline{Z}(\infty)}\\ &\quad=\quad\frac{\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\infty)-\beta\overline{Z}(\infty)};\chi=\infty\right]}{1-\mathbb{E}\left[e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)};\chi<\infty\right]\mathbb{E}\left[e^{-\alpha\left[\chi-\overline{G}^Z(\chi-)\right]-\beta\left[Z(\chi)-\overline{Z}(\chi-)\right]}\middle|\chi<\infty\right]}.\end{aligned}$$

Theorem 12.8 *For $\alpha, \beta \geq 0$, we have*

$$\mathbb{E}e^{-\alpha\overline{G}^Z(\infty)-\beta\overline{Z}(\infty)} = \frac{\mathbb{P}(\chi=\infty)\mathbb{E}e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)}}{1-\mathbb{P}(\chi<\infty)\mathbb{E}e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)}\mathbb{E}\left[e^{-\alpha\left[\chi-\overline{G}^Z(\chi-)\right]-\beta\left[Z(\chi)-\overline{Z}(\chi-)\right]}\middle|\chi<\infty\right]}.$$

12.4.2 Spectrally positive Z

In this subsection, we analyze the case in which Z has only positive jumps. It turns out that the transforms in the previous section can then be computed. As indicated below, this generalizes the results of Huzak *et al.* [166] (modulo the remarks at the beginning of this section). For instance, we obtain the transform of the distribution of $(\chi, Z(\chi))$. We remark that perturbed risk models with positive jumps are related to M/G/1 queueing systems with a second service; see [73].

Throughout, we exclude the case for which X is a (negative) subordinator, i.e., for which X is a negative drift; the analysis is then classical. By doing so, the fluctuation identities of Proposition 11.6 can be used for both X and Z. In this subsection, we use these identities without further reference.

Our analysis is based on Wiener-Hopf theory for Markov-additive processes. Indeed, the Lévy process can be embedded in a Markov-additive process in discrete time (e.g., [19, Ch. XI]). To see this, fix some $\alpha \geq 0$, and note that Equation (VI.1) of Bertoin [43] implies that for $q > 0$ and $\beta \in \mathbb{R}$,

$$\mathbb{E}e^{-\alpha\overline{G}^X(e_q)+i\beta\overline{X}(e_q)} = \mathbb{E}e^{-\alpha\overline{G}^X(e_q)}\mathbb{E}e^{i\beta\overline{X}(e_{q+\alpha})}, \tag{12.3}$$

and similarly for the joint distribution of $(\underline{F}^X(e_q), \underline{X}(e_q))$. In other words, since α is fixed, the joint distribution can be interpreted as a (defective) marginal distribution. Hence, a 'killing mechanism' has been introduced; a similar technique is used in Section 14.2, where we study discrete-time Markov-additive processes in detail.

Define a Markov-additive process in discrete time $\{(J_n, S_n)\}$ as the Markov process with state space $\{1,2,3\} \times \mathbb{R}$, characterized by the transform matrix

$$\boldsymbol{F}(\alpha,\beta) := \begin{pmatrix} 0 & \mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)}\mathbb{E}e^{i\beta\overline{X}(e_{\lambda+\alpha})} & 0 \\ 0 & 0 & \mathbb{E}e^{-\alpha\underline{F}^X(e_\lambda)}\mathbb{E}e^{i\beta\underline{X}(e_{\lambda+\alpha})} \\ \mathbb{E}e^{i\beta\xi} & 0 & 0 \end{pmatrix},$$

where, as before, ξ is a generic jump of Y. That is, $S_0 = 0$, and J_n is deterministic given J_0: in every time slot, it jumps from i to $i+1$, unless $i = 3$; then it jumps back to 1. If $J_{n-1} = 1$, the process is killed with probability $1 - \mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)}$, and otherwise we set $S_n = S_{n-1} + \eta_{n-1}$, where η_{n-1} is independent of S_{n-1} and distributed as $\overline{X}(e_{\lambda+\alpha})$. The cases $J_{n-1} = 2$ and $J_{n-1} = 3$ are similar, except for the absence of killing in the latter case. We also write

$$\tau_+ := \inf\{n > 0 : S_n > 0\}, \quad \tau_- := \inf\{n > 0 : S_n \leq 0\}.$$

Expressions of the type $\mathbb{P}_2(J_{\tau_+} = 2)$ should be understood as $\mathbb{P}(J_{\tau_+} = 2, \tau_+ < \infty | J_0 = 2)$, and similarly for $\mathbb{E}_2$.

In Wiener-Hopf theory for Markov-additive processes, an important role is played by the time-reversed process. To define this process, we introduce the Markov chain $\hat{J}$, for which the transitions are deterministic: it jumps from 3 to 2, from 2 to 1, and from 1 to 3. Hence, it jumps into the opposite direction of J. We set $\hat{S}_0 = 0$, and define the transition structure of the time-reversed Markov-additive process $(\hat{J}, \hat{S})$ as follows. If $\hat{J}_{n-1} = 2$, the process is

killed with probability $1-\mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)}$, and otherwise we set $\hat{S}_n=\hat{S}_{n-1}+\hat{\eta}_{n-1}$, where $\hat{\eta}_{n-1}$ is independent of $\hat{S}_{n-1}$ and distributed as $\overline{X}(e_{\lambda+\alpha})$. Similarly, if $\hat{J}_{n-1}=3$, the process is killed with probability $1-\mathbb{E}e^{-\alpha\underline{F}^X(e_\lambda)}$, and otherwise the increment is distributed as $\underline{X}(e_{\lambda+\alpha})$. If $\hat{J}_{n-1}=1$, the increment is distributed as $\xi>0$. The quantities $\hat{\tau}_+$ and $\hat{\tau}_-$ are defined as the ladder epochs for $\hat{S}$. We write $\hat{\mathbb{P}}_2$ for the conditional distribution given $\hat{J}_0=2$.

Recalling that the dependence on α is 'absorbed' in the killing mechanism, we define

$$G_+^{(k,\ell)}(\alpha,\beta):=\mathbb{E}_k\left[e^{i\beta S_{\tau_+}};J_{\tau_+}=\ell\right]$$

and

$$\hat{G}_-^{(k,\ell)}(\alpha,\beta):=\hat{\mathbb{E}}_k\left[e^{i\beta\hat{S}_{\hat{\tau}_-}};\hat{J}_{\hat{\tau}_-}=j\right].$$

Note that $G_+^{(2,2)}=\tilde{\mathbb{E}}\left[e^{-\alpha L_e+i\beta H_e};\mathcal{X}\right]$ in the notation of the proof of Proposition 12.7, and similarly for $G_+^{(2,1)}$; then $\mathcal{X}$ is replaced by $\mathcal{Y}$.

The Wiener-Hopf factorization for Markov-additive processes (see, e.g., Asmussen [19, Thm. XI.2.12] or Prabhu [263, Thm. 5.2]) states that $\boldsymbol{I}-\boldsymbol{F}(\alpha,\beta)$ (where $\boldsymbol{I}$ denotes the identity matrix) equals

$$\begin{pmatrix}1&0&0\\-\hat{G}_-^{(1,2)}&1-\hat{G}_-^{(2,2)}&-\hat{G}_-^{(3,2)}\\0&0&1\end{pmatrix}\begin{pmatrix}1&-\mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)+i\beta\overline{X}(e_\lambda)}&0\\-G_+^{(2,1)}&1-G_+^{(2,2)}&0\\-\mathbb{E}e^{i\beta\xi}&0&1\end{pmatrix},$$

where the arguments α and β of G_+ and $\hat{G}_-$ are suppressed for notational convenience.

We start by computing the first matrix. Note that $\hat{G}_-^{(3,2)}(\alpha,\beta)=\mathbb{E}e^{-\alpha\underline{F}^X(e_\lambda)+i\beta\underline{X}(e_\lambda)}$, so that two terms remain. Recall that Φ_{-X} is the inverse of the function $\beta\mapsto\psi_{-X}(\beta)=-\Psi_{-X}(-i\beta)$, and similarly for Φ_{-Z}.

Proposition 12.9 *For $\beta\in\mathbb{R}$, we have*

$$\hat{G}_-^{(1,2)}(\alpha,\beta)=\mathbb{E}e^{-\Phi_{-Z}(\alpha)\xi}\frac{\Phi_{-X}(\lambda)}{\Phi_{-X}(\lambda+\alpha)+i\beta},$$

and

$$\hat{G}_-^{(2,2)}(\alpha,\beta)=\frac{\Phi_{-X}(\lambda+\alpha)-\Phi_{-Z}(\alpha)}{\Phi_{-X}(\lambda+\alpha)+i\beta}.$$

Proof. We start with $\hat{G}_-^{(2,2)}$. By 'gluing together' the transitions $2\to1$ and $1\to3$, we see that the killing probability for going from 2 to itself now equals $\lambda/(\lambda+\alpha)$, and the distribution of a jump from 2 to itself can be written as $\xi+\overline{X}(e_{\lambda+\alpha})-e_{\Phi_{-X}(\lambda+\alpha)}$, where all three components are independent. Therefore, by standard results on random walks (e.g., Lemma I.4 of Prabhu [263]), we have

$$\hat{G}_-^{(2,2)}(\alpha,\beta)=\hat{\mathbb{E}}_2\left(\frac{\lambda}{\lambda+\alpha}\right)^{\hat{\tau}_-}e^{i\beta\hat{S}_{\hat{\tau}_-}}=\frac{\Phi_{-X}(\lambda+\alpha)}{\Phi_{-X}(\lambda+\alpha)+i\beta}\hat{\mathbb{E}}_2\left(\frac{\lambda}{\lambda+\alpha}\right)^{\hat{\tau}_-},$$

and it remains to calculate the mean in this expression, which we denote by η_α. For this, we repeat the argument that led to Theorem 12.1, but now for the minimum and in terms of η_α. We see that $\mathbb{E}e^{i\beta\underline{Z}(e_\alpha)}$ equals

$$\frac{\Phi_{-Z}(\alpha)}{\Phi_{-Z}(\alpha)+i\beta}=\frac{\Phi_{-X}(\lambda+\alpha)}{\Phi_{-X}(\lambda+\alpha)+i\beta}\frac{1-\eta_\alpha}{1-\eta_\alpha\frac{\Phi_{-X}(\lambda+\alpha)}{\Phi_{-X}(\lambda+\alpha)+i\beta}}=\frac{(1-\eta_\alpha)\Phi_{-X}(\lambda+\alpha)}{(1-\eta_\alpha)\Phi_{-X}(\lambda+\alpha)+i\beta},$$

so that $1-\eta_\alpha = \Phi_{-Z}(\alpha)/\Phi_{-X}(\lambda+\alpha)$.

Now we study $\hat{G}_-^{(1,2)}$. A descending ladder epoch occurs either at the first time that $\hat{J}$ visits 2, or in subsequent visits. The contribution to $\hat{G}_-^{(1,2)}$ of the first term is

$$\begin{aligned}
\hat{\mathbb{E}}_1\left[e^{i\beta\hat{S}_2};\hat{S}_2<0\right] &= \mathbb{E}e^{-\alpha\underline{F}^X(e_\lambda)}\mathbb{E}\left[e^{i\beta(\xi+\underline{X}(\lambda+\alpha))};\xi+\underline{X}(\lambda+\alpha)<0\right]\\
&= \int_0^\infty \Phi_{-X}(\lambda)e^{-(\Phi_{-X}(\lambda+\alpha)+i\beta)t}\mathbb{E}\left[e^{i\beta\xi};\xi<t\right]dt\\
&= \mathbb{E}e^{i\beta\xi}\int_0^\xi \Phi_{-X}(\lambda)e^{-(\Phi_{-X}(\lambda+\alpha)+i\beta)t}dt\\
&= \frac{\Phi_{-X}(\lambda)}{\Phi_{-X}(\lambda+\alpha)+i\beta}\mathbb{E}e^{-\Phi_{-X}(\lambda+\alpha)\xi}.
\end{aligned}$$

To compute the contribution to $\hat{G}_-^{(1,2)}$ of paths for which $\hat{S}_2$ is positive, we apply results of Arjas and Speed [12] on random walks with a random initial point. We also use their notation.

Using the expression for the previously computed $\hat{G}_-^{(2,2)}$ (again, the transform depends on α through the killing mechanism), we define

$$\bar{w}_{z-}(\beta) := \frac{1}{1-\hat{G}_-^{(2,2)}(\alpha,\beta)} = \frac{1}{1-\frac{\Phi_{-X}(\lambda+\alpha)-\Phi_{-Z}(\alpha)}{\Phi_{-X}(\lambda+\alpha)+i\beta}} = 1+\frac{\Phi_{-X}(\lambda+\alpha)-\Phi_{-Z}(\alpha)}{\Phi_{-Z}(\alpha)+i\beta}.$$

As in [12], define the projection operator $\mathcal{P}$ acting on a Fourier transform $f(\beta)=\int_{\mathbb{R}}e^{i\beta x}F(dx)$ as $\mathcal{P}f(\beta) := \int_{(-\infty,0]}e^{i\beta x}F(dx)$. Theorem 1(b) of [12] shows that the second contribution to $\hat{G}_-^{(1,2)}$ equals

$$\frac{1}{\bar{w}_{z-}(\beta)}\mathcal{P}\left[\mathbb{E}e^{-\alpha\underline{F}^X(e_\lambda)}\mathbb{E}\left[e^{i\beta(\xi+\underline{X}(\lambda+\alpha))};\xi+\underline{X}(\lambda+\alpha)>0\right]\bar{w}_{z-}(\beta)\right]. \tag{12.4}$$

A similar reasoning as before implies that

$$\begin{aligned}
\hat{\mathbb{E}}_1\left[e^{i\beta\hat{S}_2};\hat{S}_2>0\right] &= \mathbb{E}e^{-\alpha\underline{F}^X(e_\lambda)}\mathbb{E}\left[e^{i\beta(\xi+\underline{X}(e_{\lambda+\alpha}))};\xi+\underline{X}(e_{\lambda+\alpha})>0\right]\\
&= \Phi_{-X}(\lambda)\int_0^\infty e^{-(\Phi_{-X}(\lambda+\alpha)+i\beta)t}\mathbb{E}\left[e^{i\beta\xi};\xi>t\right]dt\\
&= \Phi_{-X}(\lambda)\frac{\mathbb{E}e^{i\beta\xi}-\mathbb{E}e^{-\Phi_{-X}(\lambda+\alpha)\xi}}{\Phi_{-X}(\lambda+\alpha)+i\beta}.
\end{aligned}$$

As this is the transform of a positive (defective) random variable, the first observation in the proof of Corollary 1 in [12] yields

$$\begin{aligned}
&\mathcal{P}\left[\frac{\Phi_{-X}(\lambda+\alpha)-\Phi_{-Z}(\alpha)}{\Phi_{-Z}(\alpha)+i\beta}\hat{\mathbb{E}}_1\left[e^{i\beta\hat{S}_2};\hat{S}_2>0\right]\right]\\
&= \frac{\Phi_{-X}(\lambda+\alpha)-\Phi_{-Z}(\alpha)}{\Phi_{-Z}(\alpha)+i\beta}\hat{\mathbb{E}}_1\left[e^{-\Phi_{-Z}(\alpha)\hat{S}_2};\hat{S}_2>0\right].
\end{aligned}$$

Therefore, (12.4) equals

$$\begin{aligned}
&\frac{\frac{\Phi_{-X}(\lambda+\alpha)-\Phi_{-Z}(\alpha)}{\Phi_{-Z}(\alpha)+i\beta}}{1+\frac{\Phi_{-X}(\lambda+\alpha)-\Phi_{-Z}(\alpha)}{\Phi_{-Z}(\alpha)+i\beta}}\Phi_{-X}(\lambda)\frac{\mathbb{E}e^{-\Phi_{-Z}(\alpha)\xi}-\mathbb{E}e^{-\Phi_{-X}(\lambda+\alpha)\xi}}{\Phi_{-X}(\lambda+\alpha)-\Phi_{-Z}(\alpha)}\\
&= \frac{\Phi_{-X}(\lambda)}{\Phi_{-X}(\lambda+\alpha)+i\beta}\left[\mathbb{E}e^{-\Phi_{-Z}(\alpha)\xi}-\mathbb{E}e^{-\Phi_{-X}(\lambda+\alpha)\xi}\right].
\end{aligned}$$

The claim follows by summing the two contributions. □

With the preceding proposition at our disposal, the Wiener-Hopf factorization yields that

$$G_+^{(2,1)}(\alpha,\beta) = \Phi_{-X}(\lambda)\frac{\mathbb{E}e^{i\beta\xi} - \mathbb{E}e^{-\Phi_{-Z}(\alpha)\xi}}{\Phi_{-Z}(\alpha)+i\beta},$$

and

$$1 - G_+^{(2,2)}(\alpha,\beta) = \frac{\Phi_{-X}(\lambda+\alpha)+i\beta - \Phi_{-X}(\lambda)\mathbb{E}e^{-\Phi_Z(\alpha)\xi}\mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)+i\beta\overline{X}(e_\lambda)}}{\Phi_{-Z}(\alpha)+i\beta},$$

where $\mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)+i\beta\overline{X}(e_\lambda)}$ is explicitly known in terms of Φ_{-X}.

From these expressions, by choosing $\alpha=\beta=0$, one obtains that $\tilde{\mathbb{P}}(\mathcal{X}) = 1+\frac{\Phi_{-X}(\lambda)}{\lambda}\mathbb{E}X(1)$ and $\tilde{\mathbb{P}}(\mathcal{Y}) = \Phi_{-X}(\lambda)\mathbb{E}\xi$ in the notation of the proof of Proposition 12.7. In particular, $1-\tilde{\mathbb{P}}(\mathcal{X})-\tilde{\mathbb{P}}(\mathcal{Y}) = -\frac{\Phi_{-X}(\lambda)}{\lambda}\mathbb{E}Z(1)$.

Our next goal is to characterize distributions related to modified ladder epochs and heights, leading to the main result of this subsection. Note that these results are closely related to the classical Pollaczek-Khinchine formula.

Theorem 12.10 *Let X be spectrally positive, but not a negative drift.*

(i) For $\alpha,\beta\geq 0$,

$$\begin{aligned}\mathbb{E}e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)} &= \mathbb{E}\left[e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)}\middle|\,\chi<\infty\right]\\ &= \mathbb{E}\left[e^{-\alpha\overline{G}^Z(\infty)-\beta\overline{Z}(\infty)}\middle|\,\chi=\infty\right]\\ &= -\mathbb{E}X(1)\frac{\Phi_{-Z}(\alpha)-\beta}{\alpha-\psi_{-X}(\beta)-\psi_{-Y}(\Phi_{-Z}(\alpha))},\end{aligned}$$

which should be interpreted as $-\mathbb{E}X(1)/\psi'_{-Z}(\beta)$ for $\beta=\Phi_{-Z}(\alpha)$.

In particular, $\overline{Z}(\chi-)$ has the same distribution as $\overline{X}(\infty)$.

(ii) For $\alpha,\beta\geq 0$,

$$\mathbb{E}\left[e^{-\alpha\left[\chi-\overline{G}^Z(\chi-)\right]-\beta\left[Z(\chi)-\overline{Z}(\chi-)\right]}\middle|\,\chi<\infty\right] = \frac{1}{\mathbb{E}Y_1}\frac{\psi_{-Y}(\beta)-\psi_{-Y}(\Phi_{-Z}(\alpha))}{\Phi_{-Z}(\alpha)-\beta},$$

which should be interpreted as $-\psi'_{-Y}(\beta)/\mathbb{E}Y_1$ for $\beta=\Phi_{-Z}(\alpha)$.

In particular, for $y,z>0$,

$$\mathbb{P}\left(Z(\chi)-\overline{Z}(\chi-)>x,\overline{Z}(\chi-)-Z(\chi-)>y\,|\chi<\infty\right) = \frac{1}{\mathbb{E}\xi}\int_{x+y}^{\infty}\mathbb{P}(\xi>u)du.$$

(iii) For $\alpha,\beta\geq 0$,

$$\mathbb{E}\left[e^{-\alpha\chi-\beta Z(\chi)};\chi<\infty\right] = \frac{\psi_{-Y}(\beta)-\psi_{-Y}(\Phi_{-Z}(\alpha))}{\alpha-\psi_{-X}(\beta)-\psi_{-Y}(\Phi_{-Z}(\alpha))},$$

which should be interpreted as $-\psi'_{-Y}(\beta)/\psi'_{-Z}(\beta)$ for $\beta=\Phi_{-Z}(\alpha)$.

In particular, $\mathbb{P}(\chi<\infty) = 1-\mathbb{E}Z(1)/\mathbb{E}X(1)$.

Proof. To compute the transform of the joint distribution of $(\overline{G}^Z(\chi-), \overline{Z}(\chi-))$, we use elements of the proof of Proposition 12.7:

$$\begin{aligned}\mathbb{E}e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)} &= \mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)-\beta\overline{X}(e_\lambda)}\frac{1-\tilde{\mathbb{P}}(\mathcal{X})}{1-G_+^{(2,2)}(\alpha,i\beta)} \\ &= -\mathbb{E}X(1)\frac{\frac{\Phi_{-X}(\lambda)}{\lambda}\left[\Phi_{-Z}(\alpha)-\beta\right]}{\frac{\Phi_{-X}(\lambda+\alpha)-\beta}{\mathbb{E}e^{-\alpha\overline{G}^X(e_\lambda)-\beta\overline{X}(e_\lambda)}}-\Phi_{-X}(\lambda)\mathbb{E}e^{-\Phi_{-Z}(\alpha)\xi}},\end{aligned}$$

from which the first claim follows.

The second claim is a consequence of the fact that the transform equals $G_+^{(2,1)}(\alpha,\beta)/\tilde{\mathbb{P}}(\mathcal{Y})$. The second statement follows by choosing $\alpha = 0$, and noting that

$$\mathbb{P}\left(\overline{Z}(\chi-) - Z(\chi-) > x \middle| Z(\chi) - \overline{Z}(\chi-) = y, \chi < \infty\right) = \mathbb{P}(\xi > x + y | \xi > y).$$

The third claim is obtained from the identity

$$\begin{aligned}&\mathbb{E}\left[e^{-\alpha\chi-\beta Z(\chi)}\middle|\chi<\infty\right] \\ &= \mathbb{E}\left[e^{-\alpha\overline{G}^Z(\chi-)-\beta\overline{Z}(\chi-)}\middle|\chi<\infty\right]\mathbb{E}\left[e^{-\alpha\left[\chi-\overline{G}^Z(\chi-)\right]-\beta\left[Z(\chi)-\overline{Z}(\chi-)\right]}\middle|\chi<\infty\right],\end{aligned}$$

and from (12.2). □

Let us now calculate the transform of $(G^Z(\infty), \overline{Z}(\infty))$ with Theorems 12.8 and 12.10: for $\alpha, \beta \geq 0$, we have

$$\mathbb{E}e^{-\alpha\overline{G}^Z(\infty)-\beta\overline{Z}(\infty)} = \frac{-\mathbb{E}Z(1)\frac{\Phi_{-Z}(\alpha)-\beta}{\alpha-\psi_{-X}(\beta)-\psi_{-Y}(\Phi_{-Z}(\alpha))}}{1-\frac{\psi_{-Y}(\beta)-\psi_{-Y}(\Phi_{-Z}(\alpha))}{\alpha-\psi_{-X}(\beta)-\psi_{-Y}(\Phi_{-Z}(\alpha))}} = -\mathbb{E}Z(1)\frac{\Phi_{-Z}(\alpha)-\beta}{\alpha-\psi_{-X}(\beta)-\psi_{-Y}(\beta)},$$

in accordance with Proposition 11.6 (for $q \to 0$).

Note that Theorem 4.7 of [166] is recovered upon combining the 'in particular'-statements of this theorem with Proposition 12.7, at least if Y is compound Poisson. There is also another way to see that $\mathbb{P}(\overline{Z}(\chi-) \leq x | \chi < \infty) = \mathbb{P}(\overline{X}(\infty) \leq x)$. Indeed, one can 'cut away' certain pieces of the path of Z to see that $\overline{X}(\infty)$ is distributed as $\overline{Z}(\infty)$ given the event $\{\chi < \infty\}$. Schmidli [285] makes this argument precise by time reversal of Z. However, this argument cannot be used to find the distribution of $\overline{G}^Z(\chi-)$.

We end this subsection by remarking that similar formulas can be derived if ξ is not necessarily positive. However, the system of Wiener-Hopf relations then becomes larger and no explicit results can be obtained, unless some structure is imposed; for instance, that Z has downward phase-type jumps.

12.5 Asymptotics of the maximum

In this section, we study the probabilities asymptotics of $\mathbb{P}(\overline{Z}(\infty) > x)$ and its local version $\mathbb{P}(\overline{Z}(\infty) \in (x, x+T])$ for fixed $T > 0$ as $x \to \infty$, where Z is a Lévy process that drifts to $-\infty$. It is our main goal to derive the Lévy analogue of Theorem 11.4 by imposing several conditions on the tail of the Lévy measure. The motivation for studying this problem stems from risk theory; the probability $\mathbb{P}(\overline{Z}(\infty) > x)$ is often called the *ruin probability*, see Section 1.1.5.

It is our aim to show that the embedding approach is a natural yet powerful method for studying tail asymptotics for the maximum. Relying on the random-walk results of Theorem 11.4, we study these asymptotics in the Cramér case, the intermediate case, and the

subexponential case. To our knowledge, this method has not been applied to this problem before, yet the asymptotics in the Cramér and subexponential case have been obtained elsewhere. Our results concerning the intermediate case and the 'local' subexponential case, however, are new. More results (and references) on asymptotics for Lévy processes can be found in [121, 192].

In order to apply the embedding approach, we write Z as a sum of two independent processes X and Y; one with small jumps ($\Pi_X((1,\infty)) = 0$), and a compound Poisson process Y with jumps exceeding 1. This decomposition has recently been used by Doney [119] and Pakes [248] in the context of asymptotics. Again, we write $\lambda = \Pi_Z([1,\infty)) \in [0,\infty)$, and ξ denotes a generic jump of Y. If $\lambda = 0$, we set $\xi = 0$. The random walk $\{S_n^q\}$ introduced in Section 12.2 plays an important role for $q = 0$. For notational convenience, we write S_n for S_n^0, i.e., S is a random walk with step-size distribution $\xi + X(e_\lambda)$.

The process X has a useful property: for any $\eta > 0$, we have $\mathbb{E}e^{\eta\overline{X}(e_\lambda)}, \mathbb{E}e^{\eta X(e_\lambda)} < \infty$. As a result, both $\mathbb{P}(\overline{X}(e_\lambda) > x)$ and $\mathbb{P}(X(e_\lambda) > x)$ decay faster than any exponential (by Chernoff's inequality). To see that the moment-generating functions are finite, first observe that for $\Re\beta = 0$, by the PRS factorization,

$$\mathbb{E}e^{\beta X(e_\lambda)} = \mathbb{E}e^{\beta\overline{X}(e_\lambda)}\mathbb{E}e^{\beta\underline{X}(e_\lambda)}.$$

This identity can be extended to $\Re\beta > 0$ by analytic continuation, since on this domain

$$\mathbb{E}e^{\beta X(e_\lambda)} = \frac{\lambda}{\lambda + \Psi_X(-i\beta)} < \infty,$$

where the finiteness follows from the fact that Π_X is supported on $(-\infty, 1]$. It is trivial that $\mathbb{E}e^{\beta\underline{X}(e_\lambda)}$ is analytic for $\Re\beta > 0$, hence the claim is obtained.

12.5.1 The Cramér case

First we deal with the Cramér case, i.e., when there exists some $\omega \in (0,\infty)$ for which we have $\mathbb{E}e^{\omega Z(1)} = 1$.

Given ω, one can define an associate probability measure $\mathbb{P}^\omega$, such that Z is a Lévy process under $\mathbb{P}^\omega$ with Lévy exponent $\Psi_Z(u - i\omega)$. This measure plays an important role in the following result, which is due to Bertoin and Doney [45]. By requiring regularity of 0 for $(0,\infty)$, we exclude the cases for which Z has a discrete ladder structure; random-walk identities then directly apply.

Theorem 12.11 *Let Z be a Lévy process for which 0 is regular for $(0,\infty)$. Moreover, suppose that there is some $\omega \in (0,\infty)$ such that $\mathbb{E}e^{\omega Z(1)} = 1$, and that $\mathbb{E}Z(1)e^{\omega Z(1)} < \infty$.*

Then, as $x \to \infty$, we have

$$\mathbb{P}(\overline{Z}(\infty) > x) \sim \frac{C_\omega}{\omega\mathbb{E}Z(1)e^{\omega Z(1)}}e^{-\omega x},$$

where

$$\log C_\omega := -\int_0^\infty t^{-1}(1 - e^{-t})\left[\mathbb{P}(Z(t) > 0) + \mathbb{P}^\omega(Z(t) \le 0)\right] dt. \tag{12.5}$$

Moreover, for any $T > 0$, we have as $x \to \infty$,

$$\mathbb{P}(\overline{Z}(\infty) \in (x, x+T]) \sim \frac{C_\omega}{\omega\mathbb{E}Z(1)e^{\omega Z(1)}}\left(1 - e^{-\omega T}\right)e^{-\omega x}.$$

Proof. As the reader readily verifies, the second claim follows immediately from the first.

Let us study the random walk S_n under the present assumptions. First note that $\mathbb{E}e^{\omega Z(1)} = 1$ is equivalent with $\mathbb{E}e^{\omega X(e_\lambda)}\mathbb{E}e^{\omega\xi} = 1$. Hence, the first part of Theorem 11.4 yields (the step-size distribution is nonlattice),

$$\mathbb{P}\left(\sup_{n\geq 1} S_n > x\right) \sim e^{-\sum_{n=1}^{\infty}\frac{1}{n}\left\{\mathbb{P}(S_n>0)+\mathbb{E}\left[e^{\omega S_n};S_n\leq 0\right]\right\}}\frac{1}{\omega\mathbb{E}S_1e^{\omega S_1}}e^{-\omega x}.$$

Since $\overline{X}(e_\lambda)$ has a finite moment-generating function, we have by Lemma 2.19 that

$$\begin{aligned}\mathbb{P}(\overline{Z}(\infty) > x) &= \mathbb{P}\left(\overline{X}(e_\lambda) + \sup_{n\geq 1} S_n > x\right)\\ &\sim e^{-\sum_{n=1}^{\infty}\frac{1}{n}\left\{\mathbb{P}(S_n>0)+\mathbb{E}\left[e^{\omega S_n};S_n\leq 0\right]\right\}}\frac{\mathbb{E}e^{\omega\overline{X}(e_\lambda)}}{\omega\mathbb{E}S_1e^{\omega S_1}}e^{-\omega x}.\end{aligned}$$

The rest of the proof consists of translating 'random-walk terminology' into 'Lévy terminology'. For this, we suppose that the ladder process of X is normalized such that for $\alpha > 0, \beta \in \mathbb{R}$,

$$\alpha + \Psi_X(\beta) = \kappa_X(\alpha, -i\beta)\hat{\kappa}_X(\alpha, i\beta),$$

and similarly for Z.

The quantity $1 - \mathbb{E}e^{i\beta S_1} = \Psi_Z(\beta)/(\lambda + \Psi_X(\beta))$ has both a 'random-walk' Wiener-Hopf decomposition and a 'Lévy' Wiener-Hopf decomposition, and their uniqueness leads to the identity

$$\exp\left(-\sum_{n=1}^{\infty}\frac{1}{n}\mathbb{E}\left[e^{i\beta S_n};S_n>0\right]\right) = \frac{\kappa_Z(0,-i\beta)}{\kappa_X(\lambda,-i\beta)}.$$

Similarly, since $1 - \mathbb{E}e^{(\omega+i\beta)S_1} = \Psi_Z(\beta - i\omega)/[\lambda + \Psi_X(\beta - i\omega)]$, we have

$$\exp\left(-\sum_{n=1}^{\infty}\frac{1}{n}\mathbb{E}\left[e^{(\omega+i\beta)S_n};S_n\leq 0\right]\right) = \frac{\hat{\kappa}_Z(0,i\beta+\omega)}{\hat{\kappa}_X(\lambda,i\beta+\omega)}.$$

Using the facts that $\mathbb{E}e^{\omega\overline{X}(e_\lambda)} = \kappa_X(\lambda,0)/\kappa_X(\lambda,-\omega)$ (cf. Equation (VI.1) of Bertoin [43]) and that (use $\mathbb{E}e^{\omega Z(1)} = 1$)

$$\mathbb{E}S_1e^{\omega S_1} = \frac{\mathbb{E}Z(1)e^{\omega Z(1)}}{\lambda\mathbb{E}e^{\omega\xi}} = \frac{\mathbb{E}Z(1)e^{\omega Z(1)}}{\lambda+\Psi_X(-i\omega)} = \frac{\mathbb{E}Z(1)e^{\omega Z(1)}}{\kappa_X(\lambda,-\omega)\hat{\kappa}_X(\lambda,\omega)},$$

the claim is obtained with $C_\omega = \kappa_Z(0,0)\hat{\kappa}_Z(0,\omega)$. Corollary VI.10 of Bertoin [43] shows that $\log C_\omega$ is given by (12.5). □

12.5.2 The intermediate case

This subsection studies the tail asymptotics for $\overline{Z}(\infty)$ under the condition

$$\delta = \sup\{\theta > 0 : \mathbb{E}e^{\theta Z(1)} < \infty\} > 0, \tag{12.6}$$

but we now suppose that we are in the intermediate case, i.e., that $\delta < \infty$ and $\mathbb{E}e^{\delta Z(1)} < 1$. These assumptions imply that $\lambda \in (0,\infty)$.

If μ is a measure with $\mu([1,\infty)) < \infty$, we write $\mu \in \mathcal{S}(\alpha)$ if $\mu([1,\cdot])/\mu([1,\infty)) \in \mathcal{S}(\alpha)$; see Section 2.4 for the definition. We remark that if (12.6) holds and $\Pi_Z \in \mathcal{S}(\alpha)$, then α necessarily equals δ, as the reader readily verifies.

The following theorem builds upon the second part of Theorem 11.4. It is closely related to Theorem 4.1 of Klüppelberg *et al.* [192], where the tail asymptotics are expressed in terms of the characteristics of the ladder process. Here, it is found directly in terms of the Lévy measure of Z and the tail asymptotics for $Z(1)$.

Theorem 12.12 *Let Z be a Lévy process that drifts to $-\infty$, for which $\delta \in (0,\infty)$ and $\mathbb{E}e^{\delta Z(1)} < 1$. If $\Pi_Z \in \mathcal{S}(\delta)$, then $\mathbb{E}e^{\delta \overline{Z}(\infty)} < \infty$ and $\mathbb{P}(\overline{Z}(\infty) \leq \cdot) \in \mathcal{S}(\delta)$; in fact, as $x \to \infty$, we have*

$$\mathbb{P}(\overline{Z}(\infty) > x) \sim -\frac{\mathbb{E}e^{\delta \overline{Z}(\infty)}}{\log \mathbb{E}e^{\delta Z(1)}} \Pi_Z((x,\infty)) \sim -\frac{\mathbb{E}e^{\delta \overline{Z}(\infty)}}{\mathbb{E}e^{\delta Z(1)} \log \mathbb{E}e^{\delta Z(1)}} \mathbb{P}(Z(1) > x).$$

Moreover, under these assumptions, we have as $x \to \infty$, for any $T > 0$,

$$\begin{aligned} \mathbb{P}(\overline{Z}(\infty) \in (x, x+T]) &\sim -\frac{\mathbb{E}e^{\delta \overline{Z}(\infty)}}{\log \mathbb{E}e^{\delta Z(1)}} \Pi_Z((x, x+T]) \\ &\sim -\frac{\mathbb{E}e^{\delta \overline{Z}(\infty)}}{\mathbb{E}e^{\delta Z(1)} \log \mathbb{E}e^{\delta Z(1)}} \mathbb{P}(Z(1) \in (x, x+T]). \end{aligned}$$

Proof. It suffices to prove the first asymptotic equivalences; for the relationship between the tail of the Lévy measures and the tail of the marginal distribution, we refer to Theorem 3.1 of Pakes [248].

With the embedding in our mind, we first note that by Lemma 2.19, we have $\mathbb{P}(\xi + X(e_\lambda) > x) \sim \mathbb{E}e^{\delta X(e_\lambda)} \mathbb{P}(\xi > x)$. Since $\mathbb{E}e^{\delta Z(1)} < 1$ is equivalent with $\mathbb{E}e^{\delta S_1} < 1$, we may apply the second part of Theorem 11.4 to see that

$$\mathbb{P}\left(\sup_{n \geq 1} S_n > x\right) \sim \frac{\mathbb{E}e^{\delta X(e_\lambda)}}{1 - \mathbb{E}e^{\delta \xi} \mathbb{E}e^{\delta X(e_\lambda)}} \mathbb{E} \exp\left(\delta \sup_{n \geq 1} S_n\right) \mathbb{P}(\xi > x).$$

Some elementary calculations show that

$$\frac{\mathbb{E}e^{\delta X(e_\lambda)}}{1 - \mathbb{E}e^{\delta \xi} \mathbb{E}e^{\delta X(e_\lambda)}} = \frac{1}{\frac{\lambda + \Psi_X(-i\delta)}{\lambda} - \mathbb{E}e^{\delta \xi}} = \frac{\lambda}{\Psi_Z(-i\delta)} = -\frac{\lambda}{\log \mathbb{E}e^{\delta Z(1)}}.$$

Using the fact that the moment-generating function of $\overline{X}(e_\lambda)$ is finite, we can again use Lemma 2.19 to see that

$$\begin{aligned} \mathbb{P}\left(\overline{X}(e_\lambda) + \sup_{n \geq 1} S_n > x\right) &\sim -\frac{\lambda}{\log \mathbb{E}e^{\delta Z(1)}} \mathbb{E}e^{\delta \overline{X}(e_\lambda)} \mathbb{E} \exp\left(\delta \sup_{n \geq 1} S_n\right) \mathbb{P}(\xi > x) \\ &= -\frac{\lambda \mathbb{E}e^{\delta \overline{Z}(\infty)}}{\log \mathbb{E}e^{\delta Z(1)}} \mathbb{P}(\xi > x), \end{aligned}$$

as claimed.

The second assertion is a consequence of the first claim and the observations $\mathbb{P}(\overline{Z}(\infty) > x + T) \sim e^{-\gamma T} \mathbb{P}(\overline{Z}(\infty) > x)$ and $\Pi_Z((x+T,\infty)) \sim e^{-\gamma T} \Pi_Z((x,\infty))$. □

It is readily checked that the statements of this theorem are equivalent to

$$\mathbb{P}(\overline{Z}(\infty) > x) \sim -\frac{\delta \mathbb{E}e^{\delta \overline{Z}(\infty)}}{\log \mathbb{E}e^{\delta Z(1)}} \int_x^\infty \Pi_Z((y,\infty)) dy.$$

In this expression, one can formally let $\delta \to 0$, so that the pre-integral factor tends to $1/\mathbb{E}Z(1)$. This naturally leads to the subexponential case.

12.5.3 The subexponential case

As in the previous subsection, given a measure μ with $\mu([1,\infty)) < \infty$, we write $\mu \in \mathcal{S}$ (or $\mathcal{S}^*$) if this property holds for the normalized measure, i.e., if $\mu([1,\cdot])/\mu([1,\infty)) \in \mathcal{S}$ ($\mathcal{S}^*$).

In this subsection, we suppose that the integrated tail of the Lévy measure

$$\Pi_I((x,\infty)) = \int_x^\infty \Pi_Z((y,\infty))dy$$

is subexponential, i.e., $\Pi_I \in \mathcal{S}$. As seen in Section 2.4, it is known that this property is implied by $\Pi \in \mathcal{S}^*$. In addition, we suppose that $Z(1)$ is integrable. The first assertion in the following theorem is due to Asmussen [17, Cor. 2.5]; see also Maulik and Zwart [226], Chan [67], and Braverman *et al.* [61]. As opposed to the Cramér and intermediate case, a local version of this theorem does not follow from the integral version, and that part of the theorem is new.

Theorem 12.13 *Let Z be a Lévy process that drifts to $-\infty$, for which $\Pi_I \in \mathcal{S}$. Then $\mathbb{P}(\overline{Z}(\infty) \le \cdot) \in \mathcal{S}$; in fact, as $x \to \infty$, we have*

$$\mathbb{P}(\overline{Z}(\infty) > x) \sim -\frac{\int_x^\infty \Pi_Z((y,\infty))dy}{\mathbb{E}Z(1)} \sim -\frac{\int_x^\infty \mathbb{P}(Z(1) > y)dy}{\mathbb{E}Z(1)}.$$

Moreover, if $\Pi_Z \in \mathcal{S}^*$ and Π_Z is (ultimately) nonlattice, then we have as $x \to \infty$, for any $T > 0$,

$$\mathbb{P}(\overline{Z}(\infty) \in (x, x+T]) \sim -\frac{\int_x^{x+T} \Pi_Z((y,\infty))dy}{\mathbb{E}Z(1)} \sim -\frac{\int_x^{x+T} \mathbb{P}(Z(1) > y)dy}{\mathbb{E}Z(1)}.$$

Proof. We have $\Pi_Z((x,\infty)) \sim \mathbb{P}(Z(1) > x)$ (see, e.g., [248]); hence, it suffices to prove only the first equivalences.

Since $\Pi_I \in \mathcal{S}$, it is in particular long-tailed, so that for $z \in \mathbb{R}$, $\int_x^\infty \mathbb{P}(\xi > y+z)dy \sim \int_x^\infty \mathbb{P}(\xi > y)dy$. Fix some $\eta > 0$. The latter observation implies that the function $x \mapsto x^\eta \int_{1\vee\log x}^\infty \mathbb{P}(\xi > y)dy$ is locally bounded and regularly varying at infinity with index η (see Definition 2.1), so that by the uniform convergence theorem (Theorem 2.3), for large x,

$$\int_x^\infty \mathbb{P}(\xi > y-z)dy \le (1+e^{\eta z}) \int_x^\infty \mathbb{P}(\xi > y)dy,$$

uniformly for $z \in [0, x-1]$. Since $X(e_\lambda) \le \overline{X}(e_\lambda)$ and $\mathbb{E}e^{\eta \overline{X}(e_\lambda)} < \infty$, this implies that

$$\begin{aligned}\int_x^\infty \mathbb{P}(\xi + X(e_\lambda) > y)dy \quad &\le \quad \int_x^\infty \mathbb{P}(\xi + \overline{X}(e_\lambda) > y)dy = O\left(\int_x^\infty \mathbb{P}(\xi > y)dy\right) \\ &\quad + \int_x^\infty \int_{(x-1,y-1]} \mathbb{P}(\xi > y-z)\mathbb{P}_{\overline{X}(e_\lambda)}(dz)dy \\ &\quad + \int_x^\infty \mathbb{P}(\overline{X}(e_\lambda) > y-1)dy,\end{aligned}$$

and the last two terms are readily seen to be $O(\mathbb{P}(\overline{X}(e_\lambda) > x))$ and $O(e^{-\eta x})$ respectively. Using Chernoff's inequality and the fact that ξ is heavy-tailed, it follows that $\int_x^\infty \mathbb{P}(\xi + X(e_\lambda) > y)dy = O\left(\int_x^\infty \mathbb{P}(\xi > y)dy\right)$.

This shows that one can apply the third part of Theorem 11.4 and the dominated convergence theorem to see that

$$\begin{aligned}\mathbb{P}\left(\sup_{n\ge1} S_n > x\right) \quad &\sim \quad -\frac{1}{\mathbb{E}\left[X(e_\lambda)+\xi\right]} \int_x^\infty \mathbb{P}(X(e_\lambda)+\xi > y)dy \\ &\sim \quad -\frac{1}{\mathbb{E}\left[X(e_\lambda)+\xi\right]} \int_x^\infty \mathbb{P}(\xi > y)dy.\end{aligned}$$

By definition of ξ, the right-hand side is equivalent to $\int_x^\infty \Pi_Z((y,\infty))dy/|\mathbb{E}Z(1)|$. Since this is the tail of a subexponential random variable, the first claim follows from the fact that $\overline{X}(e_\lambda)$ has a lighter tail, see Lemma 2.19.

The second assertion is proven similarly, using the second claim of the third part of Theorem 11.4. The rest of the argument is simpler than for the 'integral' version, since $\mathbb{P}(X(e_\lambda)+\xi > x) \sim \mathbb{P}(\xi > x)$ as $\Pi \in \mathcal{S}^* \subset \mathcal{S}$. A lattice version can also be given. □

A different proof for the first claim can be given based on recent results of Foss and Zachary [136]. Indeed, as noted in Section 12.4, a discrete-time Markov-additive process is embedded in the right-hand diagram in Figure 12.1. In order to verify the assumptions of [136] we suppose that Z is not spectrally positive, so that there exist $M_- \leq 0$ and $M_+ \geq 0$ such that $\lambda_\pm = \Pi_Z(\mathbb{R}\backslash(M_-,M_+)) < \infty$ and $\int_{\mathbb{R}\backslash(M_-,M_+)} z\Pi_Z(dz) < 0$. One can write Z as a sum of X and Y, where Y is now a compound Poisson process with Lévy measure Π_Z restricted to $\mathbb{R}\backslash(M_-,M_+)$. Further details are left to the reader.

CHAPTER 13

Quasi-product forms

This chapter investigates fluid networks driven by a multidimensional Lévy process. We are interested in (the joint distribution of) the steady-state content in each of the buffers, the busy periods, and the idle periods. To investigate these fluid networks, we relate the above three quantities to fluctuations of the free process by solving a multidimensional Skorokhod problem. This leads to the analysis of the distribution of the componentwise maxima, the corresponding epochs at which they are attained, and the beginning of the first last-passage excursion.

Using splitting techniques, we are able to find their Laplace transforms. It turns out that, if the components of the Lévy process are 'ordered', the Laplace transform has a so-called quasi-product form. The theory is illustrated by working out special cases, such as tandem networks and priority queues.

13.1 Introduction

Prompted by a series of papers by Kella and Whitt [180, 182, 186, 187], there has been a considerable interest in multidimensional generalizations of the classical fluid queue with non-decreasing Lévy input. In the resulting networks, often called *stochastic fluid networks*, the input into the buffers is governed by a multidimensional Lévy process; a special case of these networks has been briefly discussed in Section 1.2. Recently, motivated by work of Harrison and Williams on diffusion approximations [155, 156], the presence of product forms has been investigated [181, 183, 195, 254]. Recall that the stationary buffer-content vector has a product form if it has independent components, meaning that the distribution of this vector is a product of the marginal distributions.

The results in these papers show that, apart from trivial cases, the stationary buffer-content vector of stochastic fluid networks *never* has a product form. Despite this 'negative' result, we show that it may still be possible to express the joint distribution of the buffer content in terms of the marginal distributions. This fact is best visible in the Laplace domain. For certain tandem queues, for instance, the Laplace transform is a product that cannot be 'separated'; we then say that the buffer-content vector has a *quasi-product form*.

In the literature on stochastic fluid networks, there has been a focus on the stationary buffer-content vector W or one of its components. Here, we are also interested in the stationary distribution of vector of ages of the busy periods B and idle periods I. The age of a busy (or idle) period is the amount of time that the buffer content has been positive (or zero) without

being zero (positive). Knowing these, it is also possible to find the distribution of the remaining length of the busy (or idle) period and the total length of these periods.

We are interested in W, B, and I for a class of Lévy-driven fluid networks with a tree structure, which we therefore call *tree fluid networks*. Our analysis of these networks relies on a detailed study of a related multidimensional Skorokhod problem (see, e.g., Robert [272]). Using its explicit solution, we relate the triplet of vectors (W, B, I) to the fluctuations of a multidimensional Lévy process X. This is reminiscent of the reasoning in Section 1.1, where connection between the free and reflected process has been discussed in the (simpler) discrete-time one-dimensional case. We also prove that the stationary distribution of the buffer-content vector is unique.

Since our analysis of fluid tree networks is based on fluctuations of the process X, this chapter also contributes to fluctuation theory for multidimensional Lévy processes. Supposing that each of the components of X drifts to $-\infty$, we write $\overline{X}$ for the (vector of) componentwise maxima of X, $\overline{F}$ for the corresponding epochs at which they are first attained, and $\overline{H}$ for the beginning of the first last-passage excursion. Under a certain independence assumption, if the components of $\overline{F}$ are 'ordered', we express the Laplace transform of $(\overline{X}, \overline{F})$ in terms of the transforms of the marginals $(\overline{X}_j, \overline{F}_j)$. Since X_j is a real-valued Lévy process, the Laplace transform of $(\overline{X}_j, \overline{F}_j)$ is known if X_j has one-sided jumps; let $q \to 0$ in Proposition 11.6.

We also examine the distribution of $\overline{H}$ under the measure $\mathbb{P}_k^{\downarrow}$, which is the law of X given that the process X_k stays nonpositive. There exists a vast body of literature on (one-dimensional) Lévy processes conditioned to stay nonpositive (or nonnegative), see the recent paper by Chaumont and Doney [68] for references. Under the measure $\mathbb{P}_k^{\downarrow}$, we also find the transform of $(\overline{X}, \overline{F})$. As a special case, we establish the Laplace transform of the maximum of a Lévy process conditioned to stay below a subordinator, such as a (deterministic) positive-drift process.

By exploiting the solution of the aforementioned Skorokhod problem, the results that we obtain for the process X can be cast immediately into the fluid-network setting. For instance, the knowledge of $(\overline{X}, \overline{F})$ allows us to derive the Laplace transform of the stationary distribution of (W, B) in a tandem network and a priority system if there are only positive jumps, allowing Brownian input at the 'root' station. That is, we characterize the *joint* law of the buffer-content vector and the busy-period vector. With the $\mathbb{P}_k^{\downarrow}$-distribution of $\overline{H}$, we establish the transform of the idle-period vector I for a special tandem network. Our formulas generalize all explicit results for tandem fluid networks that are known to date (in the Laplace domain), such as those obtained by Kella [180] and more recently by Dębicki *et al.* [92]. Most notably, quasi-products appear in our formulas, even for idle periods.

To derive our results, we make use of *splitting times*. It has been shown in Chapter 11 that splitting at an extreme is closely related to the PRS identity. For real-valued Markov processes, splitting times have been introduced by Jacobsen [169]. A splitting time decomposes ('splits') a sample path of a Markov process into two independent pieces. A full description of the process *before* and *after* the splitting time can be given. However, since the splitting time is not necessarily a stopping time, the law of the second piece may differ from the original law of the Markov process. We refer to Millar [229, 230] for further details, and to Kersting and Memişoğlu [190] for a recent contribution.

The idea to use splitting times in the context of stochastic networks is novel. The known results to date are obtained with Itô's formula [195], a closely related martingale [187], or differential equations [254]. Intuitively, these approaches all exploit a certain harmonicity. Note that the results of Kyprianou and Palmowski [208] already indicate that there is a relation between these approaches and splitting. Splitting has the advantage that it is insightful and that proofs are short. Moreover, it can also be used for studying more complicated systems,

see Chapter 14.

This chapter is essentially divided into two parts. In the first part, consisting of Sections 13.2–13.4, we analyze the fluctuations of an n-dimensional Lévy processes X. The notion of splitting times is formalized in Section 13.2. These splitting times are first used to study the distribution of $(\overline{X}, \overline{F})$ in Section 13.3, and then to analyze the distribution of $\overline{H}$ under $\mathbb{P}_k^{\downarrow}$ in Section 13.4. The second part of this chapter deals with fluid networks. Section 13.5 ties these networks to fluctuations of X, so that the theory of the first part can be applied in Section 13.6. Finally, in Appendix 13.A, we derive some results for compound Poisson processes with negative drift. They are used in Section 13.4.

13.2 Splitting times

This chapter relies on the application of *splitting times* to a multidimensional Lévy process. After splitting times have been introduced, we examine splitting at the maximum (Section 13.2.1) and splitting at a last-passage excursion (Section 13.2.2).

Throughout, let $X = (X_1, \ldots, X_n)'$ be an n-dimensional Lévy process, that is, a càdlàg process with stationary independent increments such that $X(0) = 0 \in \mathbb{R}^n$. Without loss of generality, as in Bertoin [43], we work with the canonical measurable space $(\Omega, \mathcal{F}) = (D([0,\infty), \mathbb{R}^d \cup \{\partial\}), \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-field generated by the Skorokhod topology, and ∂ is an isolated point that serves as a cemetery state. In particular, X is the coordinate process. Unless otherwise stated, 'almost surely' refers to $\mathbb{P}$. All vectors are column vectors.

The following assumption is used extensively throughout this chapter:

D $X_k(t) \to -\infty$ almost surely, for every k.

We emphasize that a dependence between components is allowed. In the sequel, $\overline{X}_k(t)$ (or $\underline{X}_k(t)$) is shorthand for $\sup_{s\le t} X_k(s)$ (or $\inf_{s\le t} X_k(s)$). Due to **D**, $\overline{X}_k := \overline{X}_k(\infty)$ is well-defined and almost surely finite for every k. Furthermore, we write $\overline{X} = (\overline{X}_1, \ldots, \overline{X}_n)'$.

The following two definitions are key to further analysis. The second definition is closely related to the first, but somewhat more care is needed on a technical level. Intuitively, for the purposes of this chapter, there is no need to distinguish the two definitions.

Definition 13.1 *We say that a random time T is a* splitting time *for X under $\mathbb{P}$ if the two processes $\{X(t) : 0 \le t \le T\}$ and $\{X(T+t) - X(T) : t \ge 0\}$ are independent under $\mathbb{P}$. We say that T is a* splitting time from the left *for X under $\mathbb{P}$ if the two processes $\{X(t) : 0 \le t < T\}$ and $\{X(T+t) - X(T-) : t \ge 0\}$ are independent under $\mathbb{P}$.*

We need some notions related to the initial behavior of X. For $k = 1, \ldots, n$, set $\overline{R}_k = \inf\{t > 0 : X_k(t) = \overline{X}_k(t)\}$. Since $\overline{X}_k - X_k$ is a Markov process under $\mathbb{P}$ with respect to the filtration generated by X (see Proposition VI.1 of [43]), the Blumenthal zero-one law shows that either $\overline{R}_k > 0$ almost surely (0 is then called *irregular* for $\overline{X}_k - X_k$) or $\overline{R}_k = 0$ almost surely (0 is then called *regular* for $\overline{X}_k - X_k$). We also set $\underline{R}_k = \inf\{t > 0 : X_k(t) = \underline{X}_k(t)\}$, and define regularity of 0 for $X_k - \underline{X}_k$ similarly as for $\overline{X}_k - X_k$. If $\overline{R}_k = 0$ almost surely, we introduce

$$\overline{S}_k = \overline{S}_k^X := \inf\{t > 0 : X_k(t) \ne \overline{X}_k(t)\}.$$

Again, either $\overline{S}_k = 0$ almost surely (0 is then called an *instantaneous point* for $\overline{X}_k - X_k$) or $\overline{S}_k > 0$ almost surely (0 is then called a *holding point* for $\overline{X}_k - X_k$). One defines instantaneous and holding points for $X_k - \underline{X}_k$ similarly if $\underline{R}_k = 0$.

13.2.1 Splitting at the maximum under $\mathbb{P}$

Let $\overline{F}_k = \overline{F}_k^X := \inf\{t \geq 0 : X_k(t) = \overline{X}_k \text{ or } X_k(t-) = \overline{X}_k\}$ be the (first) epoch that X_k 'attains' its maximum, and write $\overline{F} = (\overline{F}_1, \ldots, \overline{F}_n)'$. Observe that $\overline{F}_k$ is well-defined and almost surely finite for every k by **D**.

Lemma 13.2 *Consider a Lévy process X that satisfies* **D**.

(i) If $\overline{R}_k > 0$ $\mathbb{P}$-almost surely or X_k is a compound Poisson process, then $\overline{F}_k$ is a splitting time for X under $\mathbb{P}$.

(ii) If $\overline{R}_k = 0$ $\mathbb{P}$-almost surely but X_k is not a compound Poisson process, then $\overline{F}_k$ is a splitting time from the left for X under $\mathbb{P}$.

Proof. We use ideas of Lemma VI.6 of Bertoin [43], who proves the one-dimensional case under exponential killing.

We start with the first case, in which the ascending ladder set is discrete. Set $\tau_0 = 0$ and define the stopping times $\tau_{n+1} = \inf\{t > \tau_n : \overline{X}_k(t) > \overline{X}_k(t-)\}$ for $n > 0$. Write $N = \sup\{n : \tau_n < \infty\}$. Note that **D** implies that $N < \infty$ almost surely.

Let F and K be bounded functionals. Apply the Markov property to see that for $n \in \mathbb{Z}_+$,

$$\begin{aligned}
&\mathbb{E}\left[F(X(t), 0 \leq t \leq \overline{F}_k) K(X(\overline{F}_k + t) - X(\overline{F}_k), t \geq 0); N = n\right] \\
&\quad = \mathbb{E}\left[F(X(t), 0 \leq t \leq \tau_n) 1_{\{N \geq n\}} K(X(\tau_n + t) - X(\tau_n), t \geq 0) 1_{\{\sup_{t \geq \tau_n} X_k(t) = X_k(\tau_n)\}}\right] \\
&\quad = \mathbb{E}\left[F(X(t), 0 \leq t \leq \tau_n) 1_{\{N \geq n\}}\right] \\
&\qquad \times \mathbb{E}\left[K(X(\tau_n + t) - X(\tau_n), t \geq 0) 1_{\{\sup_{t \geq \tau_n} X_k(t) = X_k(\tau_n)\}}\right] \\
&\quad = \mathbb{E}\left[F(X(t), 0 \leq t \leq \tau_n) 1_{\{N \geq n\}}\right] \mathbb{E}\left[K(X(t), t \geq 0) 1_{\{\sup_{t \geq 0} X_k(t) = 0\}}\right].
\end{aligned}$$

Summing over n shows that the processes $\{X(t) : 0 \leq t \leq \overline{F}_k\}$ and $\{X(\overline{F}_k + t) - X(\overline{F}_k) : t \geq 0\}$ are independent.

The argument in the case $\overline{R}_k = 0$ is more technical, but essentially the same. The idea is to discretize the ladder height structure, for which we use the local time $\overline{\ell}_k$ at zero of the process $\overline{X}_k - X_k$; see Bertoin [43, Ch. IV] for definitions. Note that $\overline{\ell}_k(\infty) < \infty$ almost surely by Assumption **D**.

Therefore, we fix some $\epsilon > 0$, and denote the integer part of $\epsilon^{-1}\overline{\ell}_k(\infty)$ by $n = \lfloor \epsilon^{-1}\overline{\ell}_k(\infty) \rfloor$. A variation of the argument for $\overline{R}_k > 0$ (using the additivity of the local time) shows that $\{X(t) : 0 \leq t \leq \overline{\ell}_k^{-1}(n\epsilon)\}$ and $\{X(\overline{\ell}_k^{-1}(n\epsilon) + t) - X(\overline{\ell}_k^{-1}(n\epsilon)) : t \geq 0\}$ are independent. According to [43, Prop. IV.7(iii)], $\overline{\ell}_k^{-1}(n\epsilon) \uparrow \overline{F}_k$ as $\epsilon \downarrow 0$, which proves the lemma. □

13.2.2 Splitting at a last-passage excursion under $\mathbb{P}_k^{\downarrow}$

Let $\overline{H}_k = \overline{H}_k^X := \inf\{t \geq 0 : \sup_{s \geq t} X_k(s) \neq X_k(t)\}$ be the beginning of the first last-passage excursion, and write $\overline{H} = (\overline{H}_1, \ldots, \overline{H}_n)'$.

In this subsection, we study the splitting properties of $\overline{H}_k$ for some fixed $k = 1, \ldots, n$. We suppose that 0 is a holding point for $X_k - \underline{X}_k$, i.e., that $\underline{R}_k = 0$ and $\underline{S}_k > 0$ $\mathbb{P}$-almost surely. Under this condition, the event $\{\overline{X}_k = 0\}$ has strictly positive probability. Therefore, one can straightforwardly define the conditional law $\mathbb{P}_k^{\downarrow}$ of X given $\overline{X}_k = 0$.

It is our aim to investigate splitting of $\overline{H}_k$ under $\mathbb{P}_k^{\downarrow}$, but we only have knowledge of X under $\mathbb{P}$. As a first step, it is therefore useful to give a *sample path construction* of the law $\mathbb{P}_k^{\downarrow}$

on the canonical measurable space $(\Omega, \mathcal{F})$. For this, we define a process $X^{k\downarrow}$ by

$$X^{k\downarrow}(t) = \begin{cases} X(t) & \text{if } t \in \left[\underline{R}_k^{(j)}, \underline{S}_k^{(j)}\right); \\ X(\underline{R}_k^{(j)}) - X((\underline{R}_k^{(j)} + \underline{S}_k^{(j)} - t)-) & \text{if } t \in \left[\underline{S}_k^{(j)}, \underline{R}_k^{(j)}\right), \end{cases} \tag{13.1}$$

where $\underline{R}_k^{(0)} = 0$, and for $j \geq 1$,

$$\underline{S}_k^{(j)} := \inf\left\{t > \underline{R}_k^{(j-1)} : \underline{X}_k(t) \neq X_k(t)\right\}, \quad \underline{R}_k^{(j)} := \inf\left\{t > \underline{S}_k^{(j)} : \underline{X}_k(t) = X_k(t)\right\}.$$

In other words, $X^{k\downarrow}$ is constructed from the coordinate process X by 'reverting' the excursions of $X_k - \underline{X}_k$.

We have the following interesting lemma, which is the key to all results related to $\mathbb{P}_k^{\downarrow}$. For the random-walk analogue, we refer to Doney [120].

Lemma 13.3 *Consider a Lévy process X that satisfies* **D**. *If $\underline{R}_k = 0$ and $\underline{S}_k > 0$ $\mathbb{P}$-almost surely, then $X^{k\downarrow}$ has law $\mathbb{P}_k^{\downarrow}$ under $\mathbb{P}$.*

Proof. Observe that $\overline{R}_k > 0$, and that the post-maximum process $\{X(\overline{F}_k+t)-X(\overline{F}_k) : t \geq 0\}$ has distribution $\mathbb{P}_k^{\downarrow}$ (a proof of this uses similar arguments as in the proof of Lemma 13.2; see Millar [229, 230] for more details).

Fix some $q > 0$, and let e_q be an exponentially distributed random variable, independent of X (obviously, one must then enlarge the probability space). The first step is to construct the law of $\{X(\overline{F}_k^q + t) - X(\overline{F}_k^q) : 0 \leq t < e_q - \overline{F}_k^q\}$, where $\overline{F}_k^q := \inf\{t < e_q : X_k(t) = \overline{X}_k(e_q)$ or $X_k(t-) = \overline{X}_k(e_q)\}$. By the time-reversibility of X (see Section 11.2), it is equivalent to construct the law of $\{X(\overline{G}_k^q) - X((\overline{G}_k^q - t)-) : 0 \leq t < \overline{G}_k^q\}$, where $\overline{G}_k^q := \sup\{t < e_q : X_k(t) = \underline{X}_k(e_q)$ or $X_k(t-) = \underline{X}_k(e_q)\}$.

To do so, we use ideas from Greenwood and Pitman [153]. Let $\underline{\ell}_k$ be the local time of $X_k - \underline{X}_k$ at zero (since $\underline{R}_k = 0$, $\overline{S}_k > 0$, we refer to Bertoin [43, Sec. IV.5] for its construction). Its right-continuous inverse is denoted by $\underline{\ell}_k^{-1}$. The X-excursion at local time s, denoted by X^s, is the càdlàg process defined by

$$X^s(u) := X\left(\left(\underline{\ell}_k^{-1}(s-) + u\right) \wedge \underline{\ell}_k^{-1}(s)\right) - X\left(\underline{\ell}_k^{-1}(s-)-\right), \quad u \geq 0.$$

If $\underline{\ell}_k^{-1}(s-) = \underline{\ell}_k^{-1}(s)$, then we let X^s be ∂, the zero function that serves as a cemetery. Since $\{X^s : s > 0\}$ is a càdlàg-valued Poisson point process as a result of **D**, one can derive (e.g., with the arguments of Lemma II.2 and Lemma VI.2 of [43]) that the process

$$W := \{W(s) = (D(s), X^s) : s > 0\}$$

is time-reversible, where $D(s) := X\left(\underline{\ell}_k^{-1}(s)\right)$. After setting $\sigma_q := \underline{\ell}_k^{-1}(e_q)$, it can be seen that this implies that $\{(D(s), X^s) : 0 < s < \sigma_q\}$ and $\{(D(\sigma_q-)-D((\sigma_q-s)-), X^{\sigma_q - s}) : 0 < s < \sigma_q\}$ have the same distribution. In other words, one can construct the law of $\{X(\overline{G}_k^q) - X((\overline{G}_k^q - t)-) : 0 \leq t < \overline{G}_k^q\}$ from the law of $\{X(t) : 0 \leq t < \overline{G}_k^q\}$ by 'reverting' excursions as in (13.1).

It remains to show that this construction is 'consistent' in the sense of Kolmogorov, so that one can let $q \to 0$ to obtain the claim. For this, note that the family $\{\sigma_q\}$ can be coupled with a single random variable through $\sigma_q = \underline{\ell}_k^{-1}(e_1/q)$. □

We now study the splitting properties of $\overline{H}_k$ using the alternative construction of $\mathbb{P}_k^{\downarrow}$ given in Lemma 13.3. Since $\underline{S}_k^{(1)}$ is a $\mathbb{P}$-stopping time with respect to the (completed) natural filtration of X, the Markov property of X under $\mathbb{P}$ with respect to this filtration [43, Prop. I.6] immediately yields the following analogue of Lemma 13.2.

Lemma 13.4 *Consider a Lévy process X that satisfies* **D**. *If $\underline{R}_k = 0$ and $\underline{S}_k > 0$ $\mathbb{P}$-almost surely, then $\overline{H}_k$ is a splitting time for X under $\mathbb{P}_k^{\downarrow}$. Moreover, it has an exponential distribution under $\mathbb{P}_k^{\downarrow}$.*

We remark that the construction and analysis of $\mathbb{P}_k^{\downarrow}$ is the easiest under the assumption that $\underline{R}_k = 0$ and $\underline{S}_k > 0$ $\mathbb{P}$-almost surely, which is exactly what we need in the remainder. A vast body of literature is devoted to the case $n = 1$, and the measure $\mathbb{P}_1^{\downarrow}$ is then studied under the assumption that $\overline{R}_1 = 0$. This is challenging from a theoretical point of view, since the condition that the process stays negative has $\mathbb{P}$-probability zero. Therefore, much more technicalities are needed to treat this case. We refer to Bertoin [42] and Doney [120] for more details. See also Chaumont and Doney [68].

13.3 The $\mathbb{P}$-distribution of $(\overline{X}, \overline{F})$

The aim of this section is to find the Laplace transform of the distribution of $(\overline{X}, \overline{F})$, assuming some additional structure on the process X. Thus, in the sequel we write $X_k \prec X_j$ if there exists some $K_{kj} > 0$ such that $X_j - K_{kj}X_k$ is nondecreasing almost surely.

Lemma 13.5 *Suppose the Lévy process X satisfies* **D**. *If $X_k \prec X_j$, then $\overline{F}_k \le \overline{F}_j$.*

Proof. First note that $\overline{F}_k, \overline{F}_j < \infty$ as a consequence of **D**. To prove the claim, let us assume instead that $\overline{F}_j < \overline{F}_k$ while $\hat{X}(t) := X_j(t) - CX_k(t)$ is nondecreasing for some arbitrary $C > 0$. Suppose that $X_k(\overline{F}_k) = \overline{X}_k$ and $X_j(\overline{F}_j) = \overline{X}_j$; the argument can be repeated if, for instance, $X_k(\overline{F}_k-) = \overline{X}_k$. The assumption $\overline{F}_j < \overline{F}_k$ implies that

$$0 \le \hat{X}(\overline{F}_k) - \hat{X}(\overline{F}_j) = X_j(\overline{F}_k) - \overline{X}_j - C\left[\overline{X}_k - X_k(\overline{F}_j)\right] \le 0,$$

meaning that $\overline{X}_k = X_k(\overline{F}_j)$. This contradicts $\overline{F}_j < \overline{F}_k$ in view of the definition of $\overline{F}_k$. □

The following proposition expresses the distribution of $(\overline{X}, \overline{F})$ in terms of the distributions of $(X(\overline{F}_k), \overline{F}_k)$ and $(X(\overline{F}_k-), \overline{F}_k)$. We denote the scalar product of x and y in $\mathbb{R}^n$ by $\langle x, y\rangle$, and we write 'cpd Ps' for 'compound Poisson'. Throughout this chapter, the expression $\prod_j \alpha_j \times \prod_j \beta_j \times \gamma$ should be read as $\left(\prod_j \alpha_j\right) \times \left(\prod_j \beta_j\right) \times \gamma$.

Proposition 13.6 *Suppose that X is an n-dimensional Lévy process satisfying* **D** *and that $X_1 \prec X_2 \prec \ldots \prec X_n$. Then for any $\alpha, \beta \in \mathbb{R}^n_+$,*

$$\begin{aligned}\mathbb{E}e^{-\langle\alpha,\overline{F}\rangle-\langle\beta,\overline{X}\rangle} = &\prod_{\substack{j=1\\ \overline{R}_j>0 \text{ or } X_j \text{ cpd Ps}}}^{n-1} \frac{\mathbb{E}e^{-\left[\sum_{\ell=j}^n \alpha_\ell\right]\overline{F}_j - \sum_{\ell=j}^n \beta_\ell X_\ell(\overline{F}_j)}}{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^n \alpha_\ell\right]\overline{F}_j - \sum_{\ell=j+1}^n \beta_\ell X_\ell(\overline{F}_j)}}\\ &\times \prod_{\substack{j=1\\ \overline{R}_j=0,\ X_j \text{ not cpd Ps}}}^{n-1} \frac{\mathbb{E}e^{-\left[\sum_{\ell=j}^n \alpha_\ell\right]\overline{F}_j - \sum_{\ell=j}^n \beta_\ell X_\ell(\overline{F}_j-)}}{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^n \alpha_\ell\right]\overline{F}_j - \sum_{\ell=j+1}^n \beta_\ell X_\ell(\overline{F}_j-)}}\\ &\times \mathbb{E}e^{-\alpha_n\overline{F}_n - \beta_n\overline{X}_n}.\end{aligned}$$

Proof. First observe that the assumptions imply that the terms $X_\ell(\overline{F}_j)$ and $X_\ell(\overline{F}_j-)$ in the formula are nonnegative for $\ell \ge j$, which legitimates the use of the Laplace transforms. Remark also that $\overline{R}_i = 0$ for $i > j$ whenever $\overline{R}_j = 0$, i.e., for some deterministic i_0 we have $\overline{R}_i > 0$ for $i \le i_0$ and $\overline{R}_i = 0$ for $i > i_0$.

Let us first suppose that $\overline{R}_j > 0$ or that X_j is a compound Poisson process. We prove that for $j = 1, \ldots, n-1$,

$$\mathbb{E}e^{-\sum_{\ell=j}^n \alpha_\ell \overline{F}_\ell - \sum_{\ell=j}^n \beta_\ell \overline{X}_\ell} = \frac{\mathbb{E}e^{-\left[\sum_{\ell=j}^n \alpha_\ell\right]\overline{F}_j - \sum_{\ell=j}^n \beta_\ell X_\ell(\overline{F}_j)}}{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^n \alpha_\ell\right]\overline{F}_j - \sum_{\ell=j+1}^n \beta_\ell X_\ell(\overline{F}_j)}} \mathbb{E}e^{-\sum_{\ell=j+1}^n \alpha_\ell \overline{F}_\ell - \sum_{\ell=j+1}^n \beta_\ell \overline{X}_\ell}.$$

The key observations are that $\overline{X}_j = X_j(\overline{F}_j)$ and that $\overline{F}_\ell \geq \overline{F}_j$ almost surely for $\ell = j, \ldots, n$ by Lemma 13.5. The fact that $\overline{F}_j$ is a splitting time by Lemma 13.2(i) then yields

$$\begin{aligned}
&\mathbb{E}e^{-\sum_{\ell=j}^n \alpha_\ell \overline{F}_\ell - \sum_{\ell=j}^n \beta_\ell \overline{X}_\ell} \\
&\quad = \mathbb{E}e^{-\left[\sum_{\ell=j}^n \alpha_\ell\right]\overline{F}_j - \sum_{\ell=j}^n \beta_\ell X_\ell(\overline{F}_j)} e^{-\sum_{\ell=j+1}^n \alpha_\ell\left[\overline{F}_\ell - \overline{F}_j\right] - \sum_{\ell=j+1}^n \beta_\ell\left[\overline{X}_\ell - X_\ell(\overline{F}_j)\right]} \\
&\quad = \mathbb{E}e^{-\left[\sum_{\ell=j}^n \alpha_\ell\right]\overline{F}_j - \sum_{\ell=j}^n \beta_\ell X_\ell(\overline{F}_j)} \mathbb{E}e^{-\sum_{\ell=j+1}^n \alpha_\ell\left[\overline{F}_\ell - \overline{F}_j\right] - \sum_{\ell=j+1}^n \beta_\ell\left[\overline{X}_\ell - X_\ell(\overline{F}_j)\right]}. \qquad (13.2)
\end{aligned}$$

The latter factor, which is rather complex to analyze directly, can be computed upon choosing $\alpha_j = \beta_j = 0$ in the above display.

Repeating this argument for the case $\overline{R}_j = 0$ yields with Lemma 13.2(i), provided that X_j is not a compound Poisson process,

$$\begin{aligned}
\mathbb{E}e^{-\sum_{\ell=j}^n \alpha_\ell \overline{F}_\ell - \sum_{\ell=j}^n \beta_\ell \overline{X}_\ell} \quad = \quad & \frac{\mathbb{E}e^{-\left[\sum_{\ell=j}^n \alpha_\ell\right]\overline{F}_j - \beta_j \overline{X}_j - \sum_{\ell=j+1}^n \beta_\ell X_\ell(\overline{F}_j-)}}{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^n \alpha_\ell\right]\overline{F}_j - \sum_{\ell=j+1}^n \beta_\ell X_\ell(\overline{F}_j-)}} \\
& \times \mathbb{E}e^{-\sum_{\ell=j+1}^n \alpha_\ell \overline{F}_\ell - \sum_{\ell=j+1}^n \beta_\ell \overline{X}_\ell}.
\end{aligned}$$

It is shown in the proof of Theorem VI.5(i) of [43] that $\overline{X}_j = X_j(\overline{F}_j-)$ almost surely, and this proves the claim. □

In the rest of this section, the following assumption is imposed.

F For $j = 1, \ldots, n-1$, we have

$$X_{j+1}(t) = K_{j+1} X_j(t) + \Upsilon_{j+1}(t), \qquad (13.3)$$

where the processes $(\Upsilon_2, \ldots, \Upsilon_n)$ are mutually independent nondecreasing subordinators and the constants $K_2, \ldots, K_n$ are strictly positive.

Note that Assumption **F** implies $X_1 \prec X_2 \prec \ldots \prec X_n$. Moreover, it entails that for $j = 1, \ldots, n-1$ and $\ell \geq j$, we have

$$X_\ell(t) = K_j^\ell X_j(t) + \sum_{i=j+1}^{\ell} K_i^\ell \Upsilon_i(t),$$

where we have set $K_j^\ell = \prod_{i=j+1}^\ell K_i$ and $K_j^j = 1$. In other words, X_ℓ can be written as the sum of X_j and $\ell - j$ independent processes, which are all mutually independent and independent of X_j.

The following reformulation of (13.3) in terms of matrices is useful in Section 13.6. Let K be the upper triangular matrix with element $(i, i+1)$ equal to K_{i+1} for $i = 1, \ldots, n-1$, and zero elsewhere. Also write $\Upsilon(t) := (\Upsilon_1(t), \ldots, \Upsilon_n(t))'$, where $\Upsilon_1(t) = X_1(t)$. Equation (13.3) is then nothing else than the identity $X(t) = (I - K')^{-1}\Upsilon(t)$. The matrix $(I - K')^{-1}$ is lower triangular, and element (i, j) equals K_j^i for $j \geq i$.

The *cumulant* of the subordinator $\Upsilon_j(t)$ is defined as

$$\theta_j^\Upsilon(\beta) := -\log \mathbb{E}e^{-\beta \Upsilon_j(1)}$$

for $\beta \geq 0$ and $j = 2, \ldots, n$.

The following theorem expresses the joint Laplace transform of $(\overline{X}, \overline{F})$ in terms of its marginal distributions and the cumulants θ^{Υ}. However, except for trivial cases, the Laplace transform is *not* the product of marginal Laplace transforms. Still, it can be expressed in terms of these marginal transforms in a product-type manner. We call this a *quasi-product* form.

Theorem 13.7 *Suppose that X is an n-dimensional Lévy process satisfying* **D** *and* **F**. *Then for any $\alpha, \beta \in \mathbb{R}^n_+$, the transform $\mathbb{E}e^{-\langle \alpha, \overline{F}\rangle - \langle \beta, \overline{X}\rangle}$ equals*

$$\prod_{j=1}^{n-1} \frac{\mathbb{E}e^{-\left[\sum_{\ell=j}^n \alpha_\ell + \sum_{\ell=j+1}^n \theta_\ell^{\Upsilon}\left(\sum_{k=\ell}^n K_\ell^k \beta_k\right)\right]\overline{F}_j - \left[\sum_{\ell=j}^n K_j^\ell \beta_\ell\right]\overline{X}_j}}{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^n \alpha_\ell + \sum_{\ell=j+1}^n \theta_\ell^{\Upsilon}\left(\sum_{k=\ell}^n K_\ell^k \beta_k\right)\right]\overline{F}_j - \left[\sum_{\ell=j+1}^n K_j^\ell \beta_\ell\right]\overline{X}_j}} \times \mathbb{E}e^{-\alpha_n \overline{F}_n - \beta_n \overline{X}_n}.$$

Proof. Let j be such that $\overline{R}_j > 0$ or X_j is compound Poisson. By Assumption **F**, we then have for $a \in \mathbb{R}_+$,

$$\begin{aligned}
&\mathbb{E}e^{-a\overline{F}_j - \sum_{\ell=j}^n \beta_\ell X_\ell(\overline{F}_j)} \\
&\quad = \mathbb{E}e^{-a\overline{F}_j - \left[\sum_{\ell=j}^n K_j^\ell \beta_\ell\right] X_j(\overline{F}_j) - \sum_{\ell=j+1}^n \left[\sum_{k=\ell}^n K_\ell^k \beta_k\right] \Upsilon_\ell(\overline{F}_j)} \\
&\quad = \mathbb{E}\left(e^{-a\overline{F}_j - \left[\sum_{\ell=j}^n K_j^\ell \beta_\ell\right] X_j(\overline{F}_j)} \mathbb{E}\left[e^{-\sum_{\ell=j+1}^n \left[\sum_{k=\ell}^n K_\ell^k \beta_k\right] \Upsilon_\ell(\overline{F}_j)} \middle| \overline{F}_j\right]\right) \\
&\quad = \mathbb{E}e^{-\left[a + \sum_{\ell=j+1}^n \theta_\ell^{\Upsilon}\left(\sum_{k=\ell}^n K_\ell^k \beta_k\right)\right]\overline{F}_j - \left[\sum_{\ell=j}^n K_j^\ell \beta_\ell\right] X_j(\overline{F}_j)}.
\end{aligned}$$

The claim now follows from Proposition 13.6 and the fact that $X_j(\overline{F}_j) = \overline{X}_j$ almost surely.

If $\overline{R}_j = 0$ but not a compound Poisson process, the same argument gives the joint transform of $\{X_\ell(\overline{F}_j-) : \ell = j, \ldots, n\}$ and $\overline{F}_j$. In the resulting formula, $X_j(\overline{F}_j-)$ can be replaced by $X_j(\overline{F}_j)$ as outlined in the proof of Theorem VI.5(i) in Bertoin [43]. □

The following corollary shows that Theorem 13.7 not only completely characterizes the law of $(\overline{X}, \overline{F})$ under $\mathbb{P}$, but also its law conditioned on one component to stay nonpositive. Indeed, let $\mathbb{P}^{\downarrow}_k$ be the law of $\{X(\overline{F}_k + t) - X(\overline{F}_k) : t \geq 0\}$ for $k = 1, \ldots, n$; it can be checked that this measure equals $\mathbb{P}^{\downarrow}_k$ as defined in Section 13.2.2 in case $\underline{R}_k = 0$ and $\underline{S}_k > 0$ $\mathbb{P}$-almost surely. Note that $\mathbb{P}^{\downarrow}_k$ can be regarded as the law of X given that X_k stays nonpositive.

Corollary 13.8 *For $\alpha, \beta \in \mathbb{R}^n_+$, we have*

$$\mathbb{E}^{\downarrow}_k e^{-\langle \alpha, \overline{X}\rangle - \langle \beta, \overline{F}\rangle} = \prod_{j=k}^{n-1} \frac{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^n \alpha_\ell + \sum_{\ell=j+2}^n \theta_\ell^{\Upsilon}\left(\sum_{i=\ell}^n K_\ell^i \beta_i\right)\right]\overline{F}_{j+1} - \left[\sum_{\ell=j+1}^n K_{j+1}^\ell \beta_\ell\right]\overline{X}_{j+1}}}{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^n \alpha_\ell + \sum_{\ell=j+1}^n \theta_\ell^{\Upsilon}\left(\sum_{i=\ell}^n K_\ell^i \beta_i\right)\right]\overline{F}_j - \left[\sum_{\ell=j+1}^n K_j^\ell \beta_\ell\right]\overline{X}_j}}.$$

Proof. Directly from Theorem 13.7 and (13.2). □

In particular, this corollary characterizes the law of the maximum of a Lévy process given that it stays below a subordinator. It provides further motivation for studying the law of the vector $\overline{H}$ under $\mathbb{P}^{\downarrow}_k$.

13.4 The $\mathbb{P}^{\downarrow}_k$-distribution of $\overline{H}$

The aim of this section is to find the Laplace transform of the distribution of $\overline{H}$ under $\mathbb{P}^{\downarrow}_k$ under the assumption that 0 is a holding point for $X_k - \underline{X}_k$ under $\mathbb{P}$.

We try to follow the same train of thoughts that led us to the results in Section 13.3. This analogy leads to Proposition 13.9, which does not yet give the Laplace transform of the distribution of $\overline{H}$ under $\mathbb{P}^{\downarrow}_k$. Therefore, we need an auxiliary result, formulated as Lemma 13.10,

which relies on Appendix 13.A. Finally, Proposition 13.11 enables us to find the Laplace transform of the distribution of $\overline{H}$ under $\mathbb{P}_k^{\downarrow}$.

As in the previous section, additional assumptions are imposed on the Lévy process X. Here, they are significantly more restrictive. The following Assumption **H** plays a similar role in the present section as Assumption **F** in Section 13.3. Note that it implies $X_1 \prec X_2 \prec \ldots \prec X_n$.

H Let $\Pi = \{\Pi(t) : t \geq 0\}$ be a compound Poisson process with positive jumps only. For each $j = 1, \ldots, n$, we have

$$X_j(t) = \Pi(t) - c_j t,$$

where c_j decreases strictly in j.

In the remainder of this section, we write $\lambda \in (0, \infty)$ for the intensity of jumps of Π. We also set $\rho_k^{(n)} := \sup\{\underline{R}_k^{(j)} : \underline{R}_k^{(j)} \leq \underline{R}_n^{(1)}\}$ and $\sigma_k^{(n)} := \sup\{\underline{S}_k^{(j)} : \underline{S}_k^{(j)} \leq \underline{R}_n^{(1)}\}$. In particular, $\rho_n^{(n)} = \underline{R}_n^{(1)}$ and $\sigma_n^{(n)} = \underline{S}_n^{(1)}$. Also, we write for $\beta \geq 0$ and $i = 1, \ldots, n$,

$$\psi_i(\beta) := \log \mathbb{E} e^{-\beta X_i(1)}$$

for the *Laplace exponent* of $-X_i$. Since we assume **D**, we can define Φ_i as the inverse of ψ_i, see Section 11.2. The function Φ_i plays an important role in this section.

Recall that we used n splitting times to arrive at Proposition 13.6. Here, we only know that $\overline{H}_k$ is a splitting time for X under $\mathbb{P}_k^{\downarrow}$ (see Lemma 13.4). In general, however, $\overline{H}_i$ $(i < k)$ is not a splitting time under $\mathbb{P}_k^{\downarrow}$, and the similarity with Proposition 13.6 is lost.

Proposition 13.9 *Suppose the Lévy process X satisfies* **D**. *For $\gamma \in \mathbb{R}_+^k$, we have*

$$\mathbb{E}_k^{\downarrow} e^{-\sum_{j=1}^k \gamma_j \overline{H}_j} = \frac{\lambda}{\lambda + \sum_{j=1}^k \gamma_j} \mathbb{E} e^{-\sum_{j=1}^{k-1} \gamma_j \left(\rho_k^{(k)} - \rho_j^{(k)}\right)}.$$

Proof. Lemma 13.4 yields

$$\mathbb{E}_k^{\downarrow} e^{-\sum_{j=1}^k \gamma_j \overline{H}_j} = \mathbb{E}_k^{\downarrow} e^{-\left(\sum_{j=1}^k \gamma_j\right)\overline{H}_k} \mathbb{E}_k^{\downarrow} e^{-\sum_{j=1}^{k-1} \gamma_j (\overline{H}_j - \overline{H}_k)}.$$

In the discussion following (13.1), we have seen that there is a simple sample-path correspondence between the laws $\mathbb{P}_k^{\downarrow}$ and $\mathbb{P}$. This yields immediately that $\overline{H}_k$ is exponentially distributed under $\mathbb{P}_k^{\downarrow}$ with parameter λ. It also gives that the $\mathbb{P}_k^{\downarrow}$-distribution of $\{\overline{H}_j - \overline{H}_k : j = 1, \ldots, k-1\}$ is the same as the $\mathbb{P}$-distribution $\{\rho_k^{(k)} - \rho_j^{(k)} : j = 1, \ldots, k-1\}$. □

Motivated by the preceding proposition, we now focus on the calculation of the distribution of the $\rho_k^{(k)} - \rho_j^{(k)}$ (that is, their joint Laplace transform). For this, we apply results from Appendix 13.A.

The following lemma is of crucial importance, as it provides a recursion for the transform of $\{\rho_{j+1}^{(i)} - \rho_j^{(i)} : j = 1, \ldots, i-1\}$ and $\{\rho_j^{(i)} - \sigma_j^{(i)} : j = 1, \ldots, i\}$ in terms of the transform of the same family with superscript $(i-1)$. The transforms of the marginals $\rho_i^{(i)} - \sigma_i^{(i)}$ and $\rho_{i-1}^{(i-1)} - \sigma_{i-1}^{(i-1)}$ also appear in the expression, but these transforms are known: for $\gamma \geq 0$, $i = 1, \ldots, n$ (cf. the proof of Proposition 13.21),

$$\lambda \mathbb{E} e^{-\gamma\left(\rho_i^{(i)} - \sigma_i^{(i)}\right)} = \lambda + \gamma - c_i \Phi_i(\gamma). \tag{13.4}$$

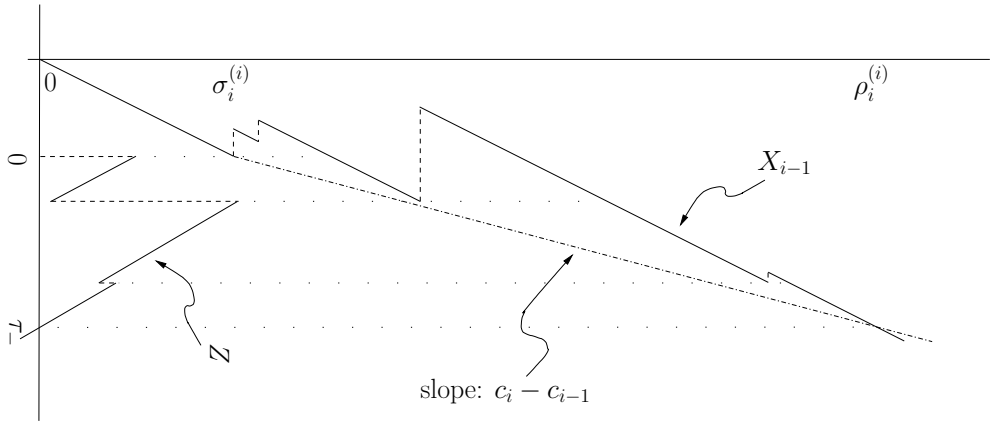

Figure 13.1: Excursions of $X_{i-1} - \underline{X}_{i-1}$ correspond to jumps of Z.

Lemma 13.10 *Suppose that X is an n-dimensional Lévy process satisfying* **D** *and* **H**. *Then for any $i = 2, \dots, n$, $\beta \in \mathbb{R}_+^{i-1}, \gamma \in \mathbb{R}_+^i$, we have the following recursion:*

$$\begin{aligned}
&\mathbb{E}e^{-\sum_{j=1}^{i-1}\beta_j\left(\rho_{j+1}^{(i)}-\rho_j^{(i)}\right)-\sum_{j=1}^{i}\gamma_j\left(\rho_j^{(i)}-\sigma_j^{(i)}\right)} \\
&\quad= \frac{\beta_{i-1}+\lambda\mathbb{E}e^{-\gamma_i\left(\rho_i^{(i)}-\sigma_i^{(i)}\right)}}{\beta_{i-1}+\lambda\mathbb{E}e^{-\delta_i(\beta_{i-1},\gamma_i)\left(\rho_{i-1}^{(i-1)}-\sigma_{i-1}^{(i-1)}\right)}} \\
&\qquad\times \mathbb{E}e^{-\sum_{j=1}^{i-2}\beta_j\left(\rho_{j+1}^{(i-1)}-\rho_j^{(i-1)}\right)-\sum_{j=1}^{i-2}\gamma_j\left(\rho_j^{(i-1)}-\sigma_j^{(i-1)}\right)-[\delta_i(\beta_{i-1},\gamma_i)+\gamma_{i-1}]\left(\rho_{i-1}^{(i-1)}-\sigma_{i-1}^{(i-1)}\right)},
\end{aligned}$$

with $\delta_i(\beta,\gamma) := \left(\frac{c_{i-1}}{c_i}-1\right)(\lambda+\beta)+\frac{c_{i-1}}{c_i}\gamma$.

Proof. Fix some $i = 2, \dots, n$, and consider the process X_{i-1} between $\sigma_i^{(i)}$ and $\rho_i^{(i)}$. There are several excursions (at least one) of the process $X_{i-1} - \underline{X}_{i-1}$ away from 0 between $\sigma_i^{(i)}$ and $\rho_i^{(i)}$, and we call these excursions the $(i-1)$-subexcursions. Each $(i-1)$-subexcursion contains excursions of the processes $X_\ell - \underline{X}_\ell$ for $\ell < i-1$; we call these the ℓ-subexcursions. To each $(i-1)$-subexcursion, we assign $2i-4$ marks, namely two for each of the $i-2$ types of further subexcursions. The first mark corresponds to the length of the last ℓ-subexcursion in the $(i-1)$-subexcursion, and the second to the difference between the end of the last ℓ-subexcursion and the end of the $(\ell+1)$-subexcursion. Observe that these marks are independent for every $(i-1)$-subexcursion between $\sigma_i^{(i)}$ and $\rho_i^{(i)}$, and that their distributions are equal to those of $\{\rho_\ell^{(i-1)} - \sigma_\ell^{(i-1)} : \ell = 1, \dots, i-2\}$ (the first marks) and $\{\rho_{\ell+1}^{(i-1)} - \rho_\ell^{(i-1)} : \ell = 1, \dots, i-2\}$ (the second marks).

The idea is to apply Proposition 13.21 to the process

$$Z(x) := \inf\left\{t \geq 0 : X_{i-1}\left(\sigma_i^{(i)}\right) - X_{i-1}\left(\sigma_i^{(i)}+t\right) = x\right\} - \frac{x}{c_{i-1}-c_i},$$

see Figure 13.1. In this diagram, excursions of $X_{i-1} - \underline{X}_{i-1}$ correspond to jumps of Z. The relevant information on the subexcursions is incorporated into Z as jump marks.

Observe that Z is a compound Poisson process with negative drift $1/c_{i-1} - 1/(c_{i-1}-c_i)$ and intensity λ/c_{i-1}, starting with a (marked) jump at zero. The jumps of Z correspond to $(i-1)$-excursions, and the above marks are assigned to the each of the jumps. In terms

of Proposition 13.21, it remains to observe that $\rho_i^{(i)} - \rho_{i-1}^{(i)}$ and $\rho_i^{(i)} - \sigma_i^{(i)}$ correspond to $(\tau_- - T_{N_-})/c_{i-1}$ and $\tau_-/(c_{i-1} - c_i)$ respectively. □

With the recursion of Lemma 13.10, we can find the joint transform of $\rho_k^{(k)} - \rho_j^{(k)}$ for $j = 1, \dots, k-1$, which is required to work out Proposition 13.9. This is done in (13.14) below. It is equivalent to find the transform of $\rho_{j+1}^{(k)} - \rho_j^{(k)}$ for $j = 1, \dots, k-1$, which is the content of the next proposition. We have also added $\rho_k^{(k)} - \sigma_k^{(k)}$ for use in subsequent work. The resulting formula has some remarkable features similar to the formula in Theorem 13.7. Most interestingly, a quasi-product form appears here as well.

For $\beta \in \mathbb{R}_+^{k-1} \geq 0$, and $j = 1, \dots, k-1$, we define

$$\mathcal{C}_j^k(\beta) := c_j \sum_{\ell=j}^{k-1} \left(\frac{1}{c_{\ell+1}} - \frac{1}{c_\ell} \right) (\lambda + \beta_\ell).$$

Proposition 13.11 *Suppose that X is an n-dimensional Lévy process satisfying* **D** *and* **H**. *Then for any $k = 2, \dots, n$, $\beta \in \mathbb{R}_+^{k-1}$, $\gamma \geq 0$, we have*

$$\begin{aligned} \mathbb{E}e^{-\sum_{j=1}^{k-1} \beta_j \left(\rho_{j+1}^{(k)} - \rho_j^{(k)}\right) - \gamma\left(\rho_k^{(k)} - \sigma_k^{(k)}\right)} &= \prod_{j=1}^{k-1} \frac{\beta_j + \lambda \mathbb{E}e^{-\left[\mathcal{C}_{j+1}^k(\beta) + \frac{c_{j+1}}{c_k}\gamma\right]\left(\rho_{j+1}^{(j+1)} - \sigma_{j+1}^{(j+1)}\right)}}{\beta_j + \lambda \mathbb{E}e^{-\left[\mathcal{C}_j^k(\beta) + \frac{c_j}{c_k}\gamma\right]\left(\rho_j^{(j)} - \sigma_j^{(j)}\right)}} \\ &\quad \times \mathbb{E}e^{-\left[\mathcal{C}_1^k(\beta) + \frac{c_1}{c_k}\gamma\right]\left(\rho_1^{(1)} - \sigma_1^{(1)}\right)}. \end{aligned}$$

Proof. Since for $\ell = 2, \dots, i$, by definition of $\mathcal{C}_\ell^k(\beta)$,

$$\left(\frac{c_{\ell-1}}{c_\ell} - 1 \right) (\lambda + \beta_{\ell-1}) + \frac{c_{\ell-1}}{c_\ell} \mathcal{C}_\ell^k(\beta) = \mathcal{C}_{\ell-1}^k(\beta),$$

it follows from Lemma 13.10 that

$$\begin{aligned} &\frac{\mathbb{E}e^{-\sum_{j=1}^{\ell-1} \beta_j \left(\rho_{j+1}^{(\ell)} - \rho_j^{(\ell)}\right) - \left[\mathcal{C}_\ell^k(\beta) + \frac{c_\ell}{c_k}\gamma\right]\left(\rho_\ell^{(\ell)} - \sigma_\ell^{(\ell)}\right)}}{\mathbb{E}e^{-\sum_{j=1}^{\ell-2} \beta_j \left(\rho_{j+1}^{(\ell-1)} - \rho_j^{(\ell-1)}\right) - \left[\mathcal{C}_{\ell-1}^k(\beta) + \frac{c_{\ell-1}}{c_k}\gamma\right]\left(\rho_{\ell-1}^{(\ell-1)} - \sigma_{\ell-1}^{(\ell-1)}\right)}} \\ &\quad = \frac{\beta_{\ell-1} + \lambda \mathbb{E}e^{-\mathcal{C}_\ell^k(\beta,\gamma)\left(\rho_\ell^{(\ell)} - \sigma_\ell^{(\ell)}\right)}}{\beta_{\ell-1} + \lambda \mathbb{E}e^{-\mathcal{C}_{\ell-1}^k(\beta,\gamma)\left(\rho_{\ell-1}^{(\ell-1)} - \sigma_{\ell-1}^{(\ell-1)}\right)}}. \end{aligned}$$

The claim follows from this recursion (start with $\ell = k$ and note that $\mathcal{C}_k^k(\beta) = 0$). □

13.5 Multidimensional Skorokhod problems

In the next sections, we apply results of the previous sections to the analysis of *fluid networks*. Such networks are closely related to (multidimensional) *Skorokhod reflection problems*, which we describe first. With the help of Skorokhod problems, the connection between the reflected process and the free process can be made precise; cf. Section 1.1 for a discrete-time framework. Subject to certain assumptions, we explicitly solve such a reflection problem in Section 13.5.1. Section 13.5.2 describes the fluid networks associated to these special Skorokhod problems.

Let P be a nonnegative matrix with spectral radius strictly smaller than 1. To a given càdlàg function Y with values in $\mathbb{R}^n$ such that $Y(0) = 0$, one can associate a càdlàg pair (W, L) with the following properties ($w \in \mathbb{R}_+^n$):

S1 $W(t) = w + Y(t) + (I - P')L(t), t \geq 0$,

S2 $W(t) \geq 0, t \geq 0$ and $W(0) = w$,

S3 $L(0) = 0$ and L is nondecreasing,

S4 $\sum_{j=1}^{n} \int_0^\infty W_j(t)\, dL_j(t) = 0$.

It is known that such a pair exists and that it is unique; see Harrison and Reiman [154] for the continuous case, and Robert [272] for the càdlàg case. It is said that (W, L) is the solution to the Skorokhod problem of Y in $\mathbb{R}_+^n$ with reflection matrix $I - P'$.

In general, the pair (W, L) cannot be expressed explicitly in terms of the driving process Y, with the notable exception of the one-dimensional case. However, if the Skorokhod problem has a special structure, this property carries over to a multidimensional setting.

13.5.1 A special Skorokhod problem

It is the aim of this subsection to solve the Skorokhod problem for the pair (W, L) under the following assumptions:

N1 P is strictly upper triangular,

N2 the j-th column of P contains exactly one strictly positive element for $j = 2, \ldots, n$,

N3 Y_j is nondecreasing for $j = 2, \ldots, n$.

In Section 13.5.2, we show that these assumptions impose a 'tree' structure on a fluid network.

Theorem 13.12 *Under* **N1**–**N3**, *the solution to the Skorokhod problem of* Y *in* $\mathbb{R}_+^n$ *is given by*

$$\begin{aligned} L(t) &= 0 \vee \sup_{0 \leq s \leq t} \left[-(I - P')^{-1} Y(s) - (I - P')^{-1} w \right], \\ W(t) &= w + Y(t) + (I - P')L(t), \end{aligned}$$

where the supremum should be interpreted componentwise.

Proof. As W is determined by L and **S1**, we only have to prove the expression for L. By Theorem D.3 of Robert [272], we know that L_i satisfies the fixed-point equation

$$L_i(t) = 0 \vee \sup_{0 \leq s \leq t} [(P'L)_i(s) - w_i - Y_i(s)] \tag{13.5}$$

for $i = 1, \ldots, n$ and $t \geq 0$.

As a consequence of **N1**, we have $(I - P')^{-1} = I + P' + \ldots + P'^{n-1}$, and the j-th row of $(I - P')^{-1}$ is the j-th row of $I + P' + P'^2 + \ldots + P'^{j-1}$. Therefore, the theorem asserts that

$$L_i(t) = 0 \vee \sup_{0 \leq s \leq t} \left[-\sum_{k=0}^{i-1} \left[P'^k Y(s) + P'^k w \right]_i \right]. \tag{13.6}$$

The proof goes by induction. For $i = 1$, (13.6) is the same equation as (13.5). Let us now suppose that we know that (13.6) holds for $i = 1, \ldots, j-1$, where $j = 2, \ldots, n$. Furthermore,

let $j^* < j$ be such that $p_{j^*j} > 0$; it is unique by **N2**. Equation (13.5) shows that

$$\begin{aligned}
L_j(t) &= 0 \vee \sup_{0\le s\le t} [p_{j^*j} L_{j^*}(s) - w_j - Y_j(s)] \\
&= 0 \vee \sup_{0\le s\le t} \left[\left(0 \vee \sup_{0\le u\le s} - \sum_{k=0}^{j^*-1} p_{j^*j} \left[P'^k Y(u) + P'^k w \right]_{j^*} \right) - w_j - Y_j(s) \right] \\
&= 0 \vee \sup_{0\le s\le t} \left[\sup_{0\le u\le s} - \sum_{k=0}^{j^*-1} p_{j^*j} \left[P'^k Y(u) + P'^k w \right]_{j^*} - w_j - Y_j(s) \right] \qquad (13.7) \\
&= 0 \vee \sup_{0\le u\le t} \sup_{u\le s\le t} \left[- \sum_{k=0}^{j^*-1} \left[P'^{k+1} Y(u) + P'^{k+1} w \right]_j - w_j - Y_j(s) \right] \\
&= 0 \vee \sup_{0\le u\le t} \left[- \sum_{k=0}^{j^*} \left[P'^k Y(u) + P'^k w \right]_j \right], \qquad (13.8)
\end{aligned}$$

where we have used **N3** for the equalities (13.7) and (13.8).

The proof is completed after noting that the j-th row of P'^k only contains zeroes for $k = j^* + 1, \ldots, j - 1$. □

Instead of working directly with W, it is often convenient to work with a transformed version, $\widetilde{W} := (I - P')^{-1} W$. The process $\widetilde{W}$ lies in a cone $\mathcal{C}$, which is a polyhedron and a proper subset of the orthant $\mathbb{R}^n_+$. Under the present assumptions, at least one edge of $\mathcal{C}$ is in the interior of $\mathbb{R}^n_+$ and at least one is an axis. Below we give an interpretation of $\widetilde{W}$.

We next establish a correspondence between the event that $W_j(t) = 0$ and $\widetilde{W}_j(t) = 0$ under an additional condition.

Proposition 13.13 *Suppose that* **N1**–**N3** *hold, but with 'nondecreasing' replaced by 'strictly increasing' in* **N3**. *Then we have* $W_j(t) = 0$ *if and only if* $\widetilde{W}_j(t) = 0$, *for any* $j = 1, \ldots, n$ *and* $t \ge 0$.

Proof. For $j = 1$ we have $W_j(t) = \widetilde{W}_j(t)$, so the stated is satisfied; suppose therefore that $j > 1$. Since the matrix $(I - P')^{-1}$ is lower triangular and nonnegative, we straightforwardly get that $\widetilde{W}_j(t) = 0$ implies $W_j(t) = 0$.

For the converse, observe that under **N1**, **N2** (see the proof of Theorem 13.12; we use the same notation)

$$\widetilde{W}_j(t) = \sum_{k=0}^{j-1} \left[P'^k W \right]_j (t) = \sum_{k=0}^{j^*} \left[P'^k W \right]_j (t).$$

An induction argument shows that it suffices to prove that $W_j(t) = 0$ implies $W_{j^*}(t) = 0$. To see that this holds, we observe that by **S1** and (13.5), $W_j(t) = 0$ is equivalent to

$$p_{j^*j} L_{j^*}(t) - w_j - Y_j(t) = 0 \vee \sup_{0\le s\le t} \left[p_{j^*j} L_{j^*}(s) - w_j - Y_j(s) \right].$$

The right-hand side of this equality is clearly nondecreasing. Therefore, since Y_j is strictly increasing by assumption, we conclude that $dL_{j^*}(t) > 0$, which immediately yields $W_{j^*}(t) = 0$ by **S4**. This completes the proof. □

13.5.2 Lévy-driven tree fluid networks

In this subsection, we define a class of Lévy-driven fluid networks, which we call *tree fluid networks*. We are interested in the steady-state behavior of such networks.

Consider n (infinite-buffer) fluid queues, with external input to queue j in the time interval $[0,t]$ given by $A_j(t)$. We assume that $A := \{A(t) : t \geq 0\} = \{(A_1(t), \ldots, A_n(t))' : t \geq 0\}$ is a càdlàg Lévy process starting in $A(0) = 0 \in \mathbb{R}^n_+$. The buffers are continuously drained at a constant rate as long as there is content in the buffer. These drain rates are given by a vector r; for buffer j, the rate is $r_j > 0$. Note that the network introduced in Section 1.2 fits into this framework.

The interaction between the queues is modeled as follows. A fraction p_{ij} of the output of station i is immediately transferred to station j, while a fraction $1-\sum_{j\neq i} p_{ij}$ leaves the system. We set $p_{ii} = 0$ for all i, and suppose that $\sum_j p_{ij} \leq 1$. The matrix $P = \{p_{ij} : i,j = 1,\ldots,n\}$ is called the *routing matrix*. We assume that for any station i, there is at most one station feeding buffer i, and that $p_{ij} = 0$ for $j < i$. The resulting network can be represented by a (directed) tree. Indeed, the stations then correspond to nodes, and there is a vertex from station i and j if $p_{ij} > 0$. This motivates the name 'tree fluid networks'. We represent such a fluid network by the triplet (A, r, P). Note that P satisfies **N1**, **N2** by definition of a tree fluid network.

The *buffer content* process W and *regulator* L associated to the fluid network (A, r, P) are defined as the solution of the Skorokhod problem of

$$Y(t) := A(t) - (I - P')rt$$

with reflection matrix $I - P'$. The buffer content is sometimes called the *workload*, explaining the notation W. Importantly, the dynamics of the network are given by **S1**–**S4**, as the reader may verify. The process L_j can be interpreted as the cumulative unused capacity in station j.

Associated to the processes W and L, one can also define the process of the *age of the busy period*: for $j = 1,\ldots,n$, we set

$$B_j(t) := t - \sup\{s \leq t : W_j(s) = 0\}, \tag{13.9}$$

and let $B(t) = (B_1(t), \ldots, B_n(t))'$. Hence, if there is work in queue j at time t (that is, $W_j(t) > 0$), $B_j(t)$ is the time that elapsed after the last time that the j-th queue was empty. If there is no work in queue i at time t, then $B_i(t) = 0$. Similarly, one can also define the *age of the idle period* for $j = 1,\ldots,n$:

$$I_j(t) := t - \sup\{s \leq t : W_j(s) \neq 0\},$$

and the corresponding vector $I(t)$. As a result of these definitions, $I_j(t) > 0$ implies $B_j(t) = 0$ and $B_j(t) > 0$ implies $I_j(t) = 0$ for $j = 1,\ldots,n$. The quantities $\widetilde{B}_j(t)$ and $\widetilde{I}_j(t)$ are defined similarly, but with W_j replaced by the j-th element of $\widetilde{W} = (I - P')^{-1}W$.

The random variables $\widetilde{W}_j$, $\widetilde{B}_j$, and $\widetilde{I}_j$ have a natural interpretation. Indeed, let us consider all stations on a path from the root of the tree to station j. The *total* content of the buffers along this path is then given by $\widetilde{W}_j$. Consequently, $\widetilde{B}_j$ and $\widetilde{I}_j$ correspond to the ages of the busy and idle periods of this aggregate buffer.

In the rest of the chapter, we assume that the tree fluid network has the following additional properties:

T1 If $p_{ij} > 0$, then $p_{ij} > r_j/r_i$,

T2 $A_j(t)$ are nondecreasing for $j = 2,\ldots,n$,

T3 A is an n-dimensional Lévy process,

T4 A is integrable and $(I-P')^{-1}\mathbb{E}A(1) < r$.

Assumption **T1** can be interpreted as a *work-conserving* property, and **T4** ensures *stability* of the network. An important consequence of **T1** and **T2** is that Y is componentwise nondecreasing, except for Y_1. Consequently, if **T1** and **T2** hold for a tree fluid network, then **N1**–**N3** are automatically satisfied for the associated Skorokhod problem. Hence, Theorem 13.12 gives an explicit description of the buffer contents in the network.

In the next proposition, we find the steady-state behavior of the buffer content and the age of the busy (and idle) period for the Lévy-driven tree fluid network (A, r, P). We also consider the case where the inequality $p_{ij} > r_j/r_i$ in **T1** holds only weakly (i.e. $p_{ij} \geq r_j/r_i$), as this plays a role in priority fluid systems (see Section 13.6.3 below).

To this end, we define the process

$$X(t) := (I-P')^{-1}Y(t) = (I-P')^{-1}A(t) - rt.$$

Recall the definitions of $\overline{F}$ and $\overline{H}$ in Sections 13.2.1 and 13.2.2 respectively.

Proposition 13.14 *Suppose that* **T1**–**T4** *hold for the tree fluid network* (A, r, P).

(i) For any initial condition $W(0) = w$, the triplet of vectors $(W(t), B(t), I(t))$ converges in distribution to $((I-P')\overline{X}, \overline{F}, \overline{H})$ as $t \to \infty$.

(ii) If the second inequality in **T1** *holds only weakly, then for any initial condition $W(0) = w$, the triplet of vectors $(W(t), \widetilde{B}(t), \widetilde{I}(t))$ converges in distribution to $((I-P')\overline{X}, \overline{F}, \overline{H})$ as $t \to \infty$.*

Proof. Throughout this proof, a system of equations like (13.9) is abbreviated by $B(t) = t - \sup\{s \leq t : W(s) = 0\}$.

We start with the proof of (ii). By Theorem 13.12, we have for any $t > 0$

$$\widetilde{W}(t) = [x + X(t)] \vee \sup_{0 \leq s \leq t} [X(t) - X(s)],$$

where $x = (I-P')^{-1}w$. Moreover, as a consequence of Proposition 13.13, we have

$$\begin{aligned}
\widetilde{B}(t) &= t - \sup\{s \leq t : \widetilde{W}(s) = 0\} \\
&= t - \sup\left\{s \leq t : x + X(s) = 0 \wedge \inf_{0 \leq u \leq s}[x + X(u)]\right\} \\
&= t - \sup\left\{s \leq t : x + X(s) = 0 \wedge \inf_{0 \leq u \leq t}[x + X(u)]\right\},
\end{aligned}$$

where the last equality is best understood by sketching a sample path of X. The supremum over an empty set should be interpreted as zero.

This reasoning carries over to idle periods:

$$\widetilde{I}(t) = t - \sup\left\{s \leq t : x + X(s) \neq 0 \wedge \inf_{0 \leq u \leq s}[x + X(u)]\right\}.$$

Due to the stationarity of the increments of $\{X(t),\ t \geq 0\}$ {**T3**}, we may extend X to the two-sided process $\{X(t),\ t \in \mathbb{R}\}$. This leads to

$$\begin{pmatrix} \widetilde{W}(t) \\ \widetilde{B}(t) \\ \widetilde{I}(t) \end{pmatrix} \stackrel{\mathrm{d}}{=} \begin{pmatrix} [x - X(-t)] \vee \sup_{-t \leq s \leq 0}[-X(s)] \\ -\sup\left\{s : -t \leq s \leq 0, -X(s) = [x - X(-t)] \vee \sup_{-t \leq u \leq 0}[-X(u)]\right\} \\ -\sup\left\{s : -t \leq s \leq 0, -X(s) \neq [x - X(-t)] \vee \sup_{-t \leq u \leq s}[-X(u)]\right\} \end{pmatrix},$$

where '$\stackrel{\mathrm{d}}{=}$' stands for equality in distribution. Since $x - X(-t) \to -\infty$ almost surely by **T4**, this tends to

$$\begin{pmatrix} \sup_{s\le 0}[-X(s)] \\ -\sup\{s \le 0 : -X(s) = \sup_{u\le 0}[-X(u)]\} \\ -\sup\{s \le 0 : -X(s) \ne \sup_{u\le s}[-X(u)]\} \end{pmatrix},$$

a vector that is almost surely finite, again by **T4**. By time-reversibility (see Section 11.2), the latter vector is equal in distribution to $(\overline{X}, \overline{F}, \overline{H})$.

Claim (i) follows from (ii); use that $B(t) = \widetilde{B}(t)$ and $I(t) = \widetilde{I}(t)$ by Proposition 13.13. □

We remark that the above proof does not use **T3** to the fullest. Indeed, for the proposition to hold, it suffices that A has stationary increments and that it is time-reversible.

Let us now suppose that the initial buffer content w is random. Proposition 13.14 shows that $\{W(t)\}$ is a stationary process if $W(0) = w$ is distributed as μ^*, where μ^* is the distribution of $(I - P')\overline{X}$; this can be deduced from Representation (1.7) for the stationary buffer-content process $\widehat{W}$. We now show that this stationary distribution is unique.

Corollary 13.15 *Suppose that* **T1**–**T4** *hold for the tree fluid network* (A, r, P). *Then* μ^* *is the only stationary distribution.*

Proof. Suppose there exists another stationary distribution $\mu_0^* \ne \mu^*$. Let W_0^* be the corresponding stationary process. For any Borel set B in $\mathbb{R}_+^n$ and any $t \ge 0$, we then have $\mathbb{P}(W_0^*(0) \in B) = \mathbb{P}(W_0^*(t) \in B)$. Therefore,

$$\begin{aligned} \mathbb{P}(W_0^*(0) \in B) &= \lim_{t\to\infty} \mathbb{P}(W_0^*(t) \in B) \\ &= \lim_{t\to\infty} \int_0^\infty \mathbb{P}(W_0^*(t) \in B | W_0^*(0) = w)\mathbb{P}(W_0^*(0) \in dw) \\ &= \int_0^\infty \lim_{t\to\infty} \mathbb{P}(W_0^*(t) \in B | W_0^*(0) = w)\mathbb{P}(W_0^*(0) \in dw) \\ &= \int_0^\infty \mathbb{P}((I - P')\overline{X} \in B)\mathbb{P}(W_0^*(0) \in dw) = \mathbb{P}((I - P')\overline{X} \in B), \end{aligned}$$

where the second last equation is due to Proposition 13.14. This is clearly a contradiction. □

Corollary 13.15 answers, for the special case of tree fluid networks, a question from the paper of Konstantopolous *et al.* [195] on the uniqueness of the stationary distribution. Note that for the queueing problem related to (A, r, P), the uniqueness of the stationary distribution was discussed in Kella [182]. In contrast to the setting in [182], we allow for the first component of $A(t)$ to be a general Lévy process.

In the next section, we combine Proposition 13.14 with the results given in Sections 13.3 and 13.4 to study particular networks.

13.6 Tandem networks and priority systems

In this section, we analyze n fluid queues in tandem, which is a tree fluid network with a special structure. We also analyze a closely related priority system.

The tandem structure is specified by the form of the routing matrix: we suppose that P is such that $p_{i,i+1} > 0$ for $i = 1, \ldots, n-1$, and $p_{ij} = 0$ otherwise. Observe that we allow $p_{i,i+1} > 1$, and that it is not really a restriction to exclude $p_{i,i+1} = 0$; otherwise the queueing system splits into independent tandem networks.

In all of our results, we suppose that the tandem system (A, r, P) satisfies **T1**–**T4**. We rule out the degenerate case where the first $j \geq 1$ components of A are deterministic drifts, since an equivalent problem can then be studied with the first j stations removed. We also impose the following assumptions on the input Lévy process A:

T5 A has mutually independent components,

T6 the Lévy measure of A_1 is supported on $\mathbb{R}_+$.

Observe that under **T2** and **T3**, **T5** implies that $A_2, \ldots, A_n$ are independent nondecreasing subordinators.

This section consists of three parts. In Section 13.6.1, we are interested in the joint (steady-state) distribution of the buffer contents and the ages of the busy periods for fluid tandem networks, i.e., in the distribution of $(W(\infty), B(\infty))$. Section 13.6.2 considers the situation of a single compound Poisson input to the system. For that system, we are also interested in the ages of the idle periods, i.e., in the vector $I(\infty)$. In Section 13.6.3, we analyze buffer contents and busy periods in a priority system.

13.6.1 Generalities

To find the joint distribution of $W(\infty)$ and $B(\infty)$, throughout this section denoted by W and B respectively, we rely on Proposition 13.14. This motivates the analysis of $X(t) = (I - P')^{-1}A(t) - rt$. For $i = 2, \ldots, n$, we define the cumulant of $A_i(t)$ by $\theta_i^A(\beta) := -\log \mathbb{E}\mathrm{e}^{-\beta A_i(1)}$, $\beta \geq 0$. As in Section 13.4, we write ψ_i (defined by $\psi_i(\beta) = \log \mathbb{E}\mathrm{e}^{-\beta X_i(1)}$) for the Laplace exponent of $-X_i$. Its inverse is again denoted by Φ_i.

Under **T2** and **T6**, the Lévy measure of X is supported on $\mathbb{R}_+^n$. Moreover, as we ruled out trivial queues in the network, each of the components of $\overline{X}$ has a nondegenerate distribution. Therefore, let us recall that the following holds (cf. Proposition 11.6 as $q \to 0$): for $\alpha, \beta \geq 0$, $(\alpha, \beta) \neq (0, 0)$, $\beta \neq \Phi_i(\alpha)$, $i = 1, \ldots, n$, we have

$$\mathbb{E}e^{-\alpha\overline{F}_i - \beta\overline{X}_i} = -\mathbb{E}X_i(1)\frac{\Phi_i(\alpha) - \beta}{\alpha - \psi_i(\beta)}. \tag{13.10}$$

This identity plays a crucial role in the results of this section. For notational convenience, we shall write that (13.10) holds for any $\alpha, \beta \geq 0$, without the requirements $(\alpha, \beta) \neq (0, 0)$ and $\beta \neq \Phi_i(\alpha)$.

Now we can formulate the main result of this subsection. We remark that the first formula also holds if A_1 is not necessarily spectrally positive. For instance, it allows for phase-type downward jumps; see Section 12.3 for the joint transform of $\overline{X}_j$ and $\overline{F}_j$ in that case.

Theorem 13.16 *Consider a tandem fluid network (A, r, P) for which* **T1**–**T6** *holds. Then for $\omega, \beta \in \mathbb{R}_+^n$, the transform $\mathbb{E}\mathrm{e}^{-\langle\omega, W\rangle - \langle\beta, B\rangle}$ equals*

$$\prod_{j=1}^{n-1} \frac{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^n \theta_\ell^A(\omega_\ell) + \sum_{\ell=j+1}^n (p_{\ell-1,\ell}r_{\ell-1} - r_\ell)\omega_\ell + \sum_{\ell=j}^n \beta_\ell\right]\overline{F}_j - \omega_j\overline{X}_j}}{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^n \theta_\ell^A(\omega_\ell) + \sum_{\ell=j+1}^n (p_{\ell-1,\ell}r_{\ell-1} - r_\ell)\omega_\ell + \sum_{\ell=j+1}^n \beta_\ell\right]\overline{F}_j - p_{j,j+1}\omega_{j+1}\overline{X}_j}} \times \mathbb{E}e^{-\beta_n\overline{F}_n - \omega_n\overline{X}_n}.$$

Consequently, we have for $\omega, \beta \in \mathbb{R}_+^n$,

$$\mathbb{E}\mathrm{e}^{-\langle\omega, W\rangle - \langle\beta, B\rangle} = -\mathbb{E}X_n(1)\frac{\Phi_n(\beta_n) - \omega_n}{\beta_n - \psi_n(\omega_n)}$$
$$\times \prod_{j=1}^{n-1} \frac{\Phi_j\left(\sum_{\ell=j+1}^n \theta_\ell^A(\omega_\ell) + \sum_{\ell=j+1}^n (p_{\ell-1,\ell}r_{\ell-1} - r_\ell)\omega_\ell + \sum_{\ell=j}^n \beta_\ell\right) - \omega_j}{\Phi_j\left(\sum_{\ell=j+1}^n \theta_\ell^A(\omega_\ell) + \sum_{\ell=j+1}^n (p_{\ell-1,\ell}r_{\ell-1} - r_\ell)\omega_\ell + \sum_{\ell=j+1}^n \beta_\ell\right) - p_{j,j+1}\omega_{j+1}}$$

$$\times \prod_{j=1}^{n-1} \frac{\sum_{\ell=j+1}^{n} \theta_\ell^A(\omega_\ell) + \sum_{\ell=j+1}^{n}(p_{\ell-1,\ell} r_{\ell-1} - r_\ell)\omega_\ell + \sum_{\ell=j+1}^{n} \beta_\ell - \psi_j(p_{j,j+1}\omega_{j+1})}{\sum_{\ell=j+1}^{n} \theta_\ell^A(\omega_\ell) + \sum_{\ell=j+1}^{n}(p_{\ell-1,\ell} r_{\ell-1} - r_\ell)\omega_\ell + \sum_{\ell=j}^{n} \beta_\ell - \psi_j(\omega_j)}.$$

Proof. By Proposition 13.14(i), $(W, B) \stackrel{\mathrm{d}}{=} ((I - P')\overline{X}, \overline{F})$. Hence we have

$$\mathbb{E}\mathrm{e}^{-\langle\omega, W\rangle - \langle\beta, B\rangle} = \mathbb{E}\mathrm{e}^{-\langle (I-P)\omega, (I-P')^{-1}W\rangle - \langle\beta, B\rangle} = \mathbb{E}\mathrm{e}^{-\langle\beta, \overline{F}\rangle - \langle (I-P)\omega, \overline{X}\rangle}. \tag{13.11}$$

Now note that the stability condition **T4** for (A, r, P) implies **D** for X by the law of large numbers. Thus, in order to apply Theorem 13.7 for (13.11), it is enough to check that **F** holds. Standard algebraic manipulations give

$$X_1(t) = A_1(t) - r_1 t$$

and

$$X_{i+1}(t) = p_{i,i+1} X_i(t) + A_{i+1}(t) + (p_{i,i+1} r_i - r_{i+1})t$$

for $i = 1, \ldots, n-1$. Hence, **F** holds with $K_i = p_{i-1,i}$ and $\Upsilon_i(t) = A_i(t) + (p_{i-1,i} r_{i-1} - r_i)t$.

As a result, we know that from Theorem 13.7,

$$\begin{aligned}
&\mathbb{E}\mathrm{e}^{-\langle\beta, \overline{F}\rangle - \langle (I-P)\omega, \overline{X}\rangle} = \mathbb{E}\mathrm{e}^{-\langle\beta, \overline{F}\rangle - \langle\widetilde{\omega}, \overline{X}\rangle} \\
&= \prod_{j=1}^{n-1} \frac{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^{n} \theta_\ell^{\Upsilon}\left(\sum_{k=\ell}^{n} K_\ell^k \widetilde{\omega}_k\right) + \sum_{\ell=j}^{n} \beta_\ell\right]\overline{F}_j - \left(\sum_{k=j}^{n} K_j^k \widetilde{\omega}_k\right)\overline{X}_j}}{\mathbb{E}e^{-\left[\sum_{\ell=j+1}^{n} \theta_\ell^{\Upsilon}\left(\sum_{k=\ell}^{n} K_\ell^k \widetilde{\omega}_k\right) + \sum_{\ell=j+1}^{n} \beta_\ell\right]\overline{F}_j - \left(\sum_{k=j+1}^{n} K_j^k \widetilde{\omega}_k\right)\overline{X}_j}} \times \mathbb{E}e^{-\beta_n \overline{F}_n - \widetilde{\omega}_n \overline{X}_n},
\end{aligned}$$

where we have set $\widetilde{\omega} = (I - P)\omega$ for notational convenience.

The reader may check that $\sum_{k=j}^{n} K_j^k \widetilde{\omega}_k = \omega_j$ and $\sum_{k=j+1}^{n} K_j^k \widetilde{\omega}_k = p_{j,j+1}\omega_{j+1}$, leading to the first claim. The second assertion is a consequence of the first and (13.10). □

Theorem 13.16 extends several results from the literature on the steady-state distribution of the buffer content for tandem Lévy networks. In particular, if $A(t) = (A_1(t), 0)'$, $P = (p_{ij})$, with $p_{12} = 1$ and zeroes elsewhere, if one chooses $\beta_1 = \beta_2 = 0$ and $\omega_1 = 0$ in Theorem 13.16, then one obtains Theorem 3.2 of Dębicki *et al.* [92]. Additionally, if one chooses $\beta_1 = \beta_2 = 0$ and supposes that A_1 is a subordinator, we recover the results of Kella [180].

For use in Section 13.6.3, we point out that the expression for $\mathbb{E}e^{-\langle\omega, W\rangle - \langle\beta, B\rangle}$ in Theorem 13.16 equals $\mathbb{E}e^{-\langle\omega, W\rangle - \langle\beta, \widetilde{B}\rangle}$ if the second inequality in **T1** is weak, cf. Proposition 13.14(ii).

The lengths of the busy periods

Besides the Laplace transforms of the *ages* B of the busy periods, Theorem 13.16 also enables us to find the Laplace transforms of the *length* V of the steady-state running busy periods. Indeed, let D_i, $i = 1, \ldots, n$ denote the steady-state remaining lengths of the running busy period, so that $V_i = B_i + D_i$. We know that D_i and B_i are equal in distribution. In fact, following for instance [19, Sec. V.3], we have

$$(B_i, D_i) \stackrel{\mathrm{d}}{=} (U_i V_i, (1 - U_i)V_i), \tag{13.12}$$

where U_i are i.i.d. and uniform on $[0, 1]$.

For the Brownian (single) fluid queue, the following result is Corollary 3.8 of Salminen and Norros [281].

Corollary 13.17 *Consider a tandem fluid network (A, r, P) for which* **T1**–**T6** *holds. Then for $\alpha, \beta \geq 0$, $\alpha \neq \beta$,*

$$\mathbb{E}e^{-\alpha B_i - \beta D_i} = -\mathbb{E}X_i(1)\frac{\Phi_i(\alpha) - \Phi_i(\beta)}{\alpha - \beta}.$$

Moreover, we have for $\alpha \geq 0$,

$$\mathbb{E}e^{-\alpha V_i} = -\mathbb{E}X_i(1)\frac{d\Phi_i(\alpha)}{d\alpha}.$$

Proof. Since the second claim follows straightforwardly from the first, we only prove the first expression. Following (13.12), we have for $\alpha \neq \beta$,

$$\begin{aligned}(\alpha-\beta)\mathbb{E}e^{-\alpha B_i-\beta D_i} &= (\alpha-\beta)\mathbb{E}e^{-(\alpha-\beta)U_iV_i-\beta V_i} = (\alpha-\beta)\mathbb{E}\int_0^1 e^{-(\alpha-\beta)uV_i-\beta V_i}du \\ &= \mathbb{E}\int_\beta^\alpha e^{-uV_i}du = \mathbb{E}\int_0^\alpha e^{-uV_i}du - \mathbb{E}\int_0^\beta e^{-uV_i}du.\end{aligned}$$

The two identities that result upon setting $\beta = 0$ and $\alpha = 0$ can be used to express the first and second expectation in terms of the Laplace transform of B_i and D_i respectively; this yields for $\alpha \neq \beta$

$$\mathbb{E}e^{-\alpha B_i-\beta D_i} = \frac{1}{\alpha-\beta}\left[\alpha\mathbb{E}e^{-\alpha B_i} - \beta\mathbb{E}e^{-\beta B_i}\right],$$

where we have used the equality in distribution of B_i and D_i. Application of (13.10) completes the proof. □

13.6.2 A single compound Poisson input

In this subsection, we examine a tandem fluid network with a single compound Poisson input [186]. The following assumption formalizes our framework.

T7 $p_{i,i+1} = 1$ for $i = 1, \ldots, n-1$, while $p_{ij} = 0$ otherwise,

T8 A_1 is a compound Poisson process with positive drift $\mathtt{d}$ and intensity λ, and $A_j \equiv 0$ for $j = 2, \ldots, n$. Moreover, r_j decreases strictly in j and $\mathbb{E}A(1) < r_n$.

An important consequence of **T7** and **T8** is that

$$(r_j - r_k)\omega = \psi_j(\omega) - \psi_k(\omega), \tag{13.13}$$

which simplifies the resulting expressions in view of fact that we often deal with ratios of the fluctuation identity (13.10). Interestingly, it is also possible to study (joint distributions of) idle periods under these assumptions.

The following corollary collects some results that follow from **T7**, **T8**, and Theorem 13.16. Many more interesting formulas can be derived, but we have selected two examples for which the formulas are especially appealing.

Corollary 13.18 *Consider a tandem fluid network* (A, r, P) *for which* **T7** *and* **T8** *hold.*

(i) For $i = 1, \ldots, n$, *and* $\omega, \beta \geq 0$, *we have*

$$\mathbb{E}e^{-\omega W_i-\beta B_i} = -\mathbb{E}X_i(1)\frac{\Phi_i(\beta)-\omega}{\beta+(r_{i-1}-r_i)\omega} \times \frac{\Phi_{i-1}((r_{i-1}-r_i)\omega+\beta)}{\Phi_{i-1}((r_{i-1}-r_i)\omega+\beta)-\omega}.$$

Moreover, $\mathbb{P}(W_i = 0) = \mathbb{P}(B_i = 0) = \frac{\mathbb{E}X_i(1)}{\mathtt{d}-r_i}$.

(ii) For $i = 2, \ldots, n$, $\omega, \beta \geq 0$, *we have*

$$\mathbb{E}\left[e^{-\omega W_i-\beta B_i}; W_{i-1} = 0\right] = -\frac{\mathbb{E}X_i(1)}{\mathtt{d}-r_{i-1}}\frac{\Phi_i(\beta)-\omega}{\Phi_{i-1}((r_{i-1}-r_i)\omega+\beta)-\omega}.$$

Proof. To prove (i), apply Theorem 13.16 to obtain for $i = 1, \dots, n$,

$$\mathbb{E}e^{-\omega W_i - \beta B_i} = \frac{\mathbb{E}e^{-[(r_{i-1}-r_i)\omega+\beta]\overline{F}_{i-1}}}{\mathbb{E}e^{-[(r_{i-1}-r_i)\omega+\beta]\overline{F}_{i-1}-\omega\overline{X}_{i-1}}}\mathbb{E}e^{-\beta\overline{F}_i-\omega\overline{X}_i}.$$

With (13.10), this leads immediately to the given formula after invoking (13.13).

We find $\mathbb{P}(W_i = 0)$ upon choosing $\omega = 0$ and noting that

$$\mathbb{P}(W_i = 0) = \mathbb{P}(B_i = 0) = \lim_{\beta\to\infty} \mathbb{E}e^{-\beta\overline{F}_i} = -\mathbb{E}X_i(1) \lim_{\beta\to\infty} \frac{\Phi_i(\beta)}{\beta} = \frac{\mathbb{E}X_i(1)}{\mathtt{d} - r_i},$$

where the last equality follows from Proposition I.2 in [43].

The second claim uses a similar argument; it follows from Theorem 13.16 that for $i = 2, \dots, n$

$$\mathbb{E}e^{-\omega_i W_i - \beta_{i-1}B_{i-1} - \beta_i B_i} = \frac{\mathbb{E}e^{-[(r_{i-1}-r_i)\omega_i+\beta_{i-1}+\beta_i]\overline{F}_{i-1}}}{\mathbb{E}e^{-[(r_{i-1}-r_i)\omega_i+\beta_i]\overline{F}_{i-1}-\omega_i\overline{X}_{i-1}}}\mathbb{E}e^{-\beta_i\overline{F}_i-\omega_i\overline{X}_i},$$

and the numerator of the fraction tends to $\mathbb{P}(W_{i-1} = 0)$ as $\beta_{i-1} \to \infty$. Now apply (13.10) and (13.13). □

We end this subsection with an application of the theory in Section 13.4, which enables us to study the idle periods in a tandem fluid network satisfying **T7** and **T8**. For $\gamma \in \mathbb{R}_+^{k-1}$, we set

$$\mathcal{D}_j^k(\gamma) := c_j \sum_{\ell=j}^{k-1} \left(\frac{1}{c_{\ell+1}} - \frac{1}{c_\ell}\right) \left(\lambda + \sum_{p=1}^{\ell} \gamma_p\right),$$

which is similar to the definition of $\mathcal{C}_j^k$ in Section 13.4.

Proposition 13.19 *Consider a tandem fluid network (A, r, P) for which* **T7** *and* **T8** *holds. For $\gamma \in \mathbb{R}_+^n$, we have*

$$\mathbb{E}e^{-\langle\gamma, I\rangle} = 1 - \sum_{k=1}^{n} \mathbb{P}(W_k = 0)\mathbb{E}_k^{\downarrow}\left[e^{-\sum_{\ell=1}^{k-1}\gamma_\ell\overline{H}_\ell}\left(1 - e^{-\gamma_k\overline{H}_k}\right)\right],$$

where $\mathbb{P}(W_j = 0)$ is given in Corollary 13.18(i), and

$$\begin{aligned}\mathbb{E}_k^{\downarrow}e^{-\sum_{\ell=1}^{k}\gamma_\ell\overline{H}_\ell} &= \frac{\lambda + \sum_{\ell=1}^{k-1}\gamma_\ell\left(1 - \frac{c_k}{c_\ell}\right) - c_k\Phi_1(\mathcal{D}_1^k(\gamma))}{\lambda + \sum_{\ell=1}^{k}\gamma_\ell} \\ &\quad \times \prod_{j=1}^{k-1} \frac{\lambda + \sum_{\ell=1}^{k-1}\gamma_\ell - \sum_{\ell=j+1}^{k-1}\frac{c_k}{c_\ell}\gamma_\ell - c_k\Phi_{j+1}(\mathcal{D}_{j+1}^k(\gamma))}{\lambda + \sum_{\ell=1}^{k-1}\gamma_\ell - \sum_{\ell=j+1}^{k}\frac{c_k}{c_\ell}\gamma_\ell - c_k\Phi_j(\mathcal{D}_j^k(\gamma))}. \quad (13.14)\end{aligned}$$

Proof. Note that **T7** and **T8** imply **H**. The first claim follows from Proposition 13.14 and the facts that for $k = 2, \dots, n$,

$$\mathbb{E}e^{-\sum_{\ell=1}^{k}\gamma_\ell\overline{H}_\ell} = \mathbb{E}e^{-\sum_{\ell=1}^{k-1}\gamma_\ell\overline{H}_\ell} + \mathbb{E}_k^{\downarrow}\left[e^{-\sum_{\ell=1}^{k-1}\gamma_\ell\overline{H}_\ell}\left(1 - \gamma_k\overline{H}_k\right)\right]\mathbb{P}(\overline{X}_k = 0),$$

and $\mathbb{E}e^{\gamma_1\overline{H}_1} = 1 - \mathbb{E}_1^{\downarrow}\left[1 - e^{-\gamma_1\overline{H}_1}\right]\mathbb{P}(\overline{X}_1 = 0)$. These identities follow after observing that $\overline{H}_k$ vanishes on the event $\{\overline{X}_k = 0\}$, and that $\{\overline{X}_k = 0\}$ is the complement of $\{\overline{X}_k > 0\}$.

Let us now prove the expression for the $\mathbb{P}_k^{\downarrow}$-distribution of $(\overline{H}_1, \ldots, \overline{H}_k)'$. From Proposition 13.9 and Proposition 13.11, we know that

$$\mathbb{E}_k^{\downarrow} e^{-\sum_{\ell=1}^k \gamma_\ell \overline{H}_\ell} = \frac{\lambda \mathbb{E} e^{-\mathcal{D}_1^k(\gamma)\left(\rho_1^{(1)} - \sigma_1^{(1)}\right)}}{\lambda + \sum_{\ell=1}^k \gamma_\ell} \prod_{j=1}^{k-1} \frac{\sum_{\ell=1}^j \gamma_\ell + \lambda \mathbb{E} e^{-\mathcal{D}_{j+1}^k(\gamma)\left(\rho_{j+1}^{(j+1)} - \sigma_{j+1}^{(j+1)}\right)}}{\sum_{\ell=1}^j \gamma_\ell + \lambda \mathbb{E} e^{-\mathcal{D}_j^k(\gamma)\left(\rho_j^{(j)} - \sigma_j^{(j)}\right)}}.$$

The proof is finished after invoking (13.4) and noting that for $j = 1, \ldots, k-1$,

$$\frac{c_k}{c_j}\left[\lambda + \sum_{\ell=1}^j \gamma_\ell + \mathcal{D}_j^k(\gamma)\right] = \frac{c_k}{c_{j+1}}\left[\lambda + \sum_{\ell=1}^j \gamma_\ell + \mathcal{D}_{j+1}^k(\gamma)\right] = \lambda + \sum_{\ell=1}^{k-1} \gamma_\ell - \sum_{\ell=j+1}^{k-1} \frac{c_k}{c_\ell}\gamma_\ell,$$

and

$$\frac{c_k}{c_1}\left[\lambda + \mathcal{D}_1^k(\gamma)\right] = \lambda + \sum_{\ell=1}^{k-1} \gamma_\ell - \sum_{\ell=1}^{k-1} \frac{c_k}{c_\ell}\gamma_\ell,$$

as the reader readily verifies. □

13.6.3 A priority fluid system

In this subsection, we analyze a single station which is drained at a constant rate $\mathtt{r} > 0$. It is fed by n external inputs ('traffic classes') $A_1(t), \ldots, A_n(t)$, each equipped with its own (infinite-capacity) buffer. The queue discipline is (preemptive resume) *priority*, meaning that for each $i = 1, \ldots, n$, the i-th buffer is continuously drained only if first $i-1$ buffers do not require the full capacity $\mathtt{r}$. We call such a system a *priority fluid system*.

The aim of this section is to find the Laplace transform of (W, E), where $W_j = W_j(\infty)$ is the stationary buffer content of class-j input traffic, and $E_j = E_j(\infty)$ is the stationary age of the busy period for class j. We impose the following assumptions.

P1 A is an n-dimensional Lévy process with mutually independent components, and its Lévy measure is supported on $\mathbb{R}_+^n$, $A(0) = 0$,

P2 $A_j(t)$ are nondecreasing for $j = 2, \ldots, n$,

P3 A is integrable and $\sum_{i=1}^n \mathbb{E}A_i(1) < \mathtt{r}$.

The central idea is that W evolves in the same manner as the solution to the Skorokhod problem that corresponds to a tandem fluid network (A, r, P), with $r = (\mathtt{r}, \ldots, \mathtt{r})'$ and $P = (p_{ij})$ such that $p_{i,i+1} = 1$ for $i = 1, \ldots, n-1$ and $p_{ij} = 0$ otherwise. This equivalence has been noticed, for instance, by Elwalid and Mitra [130]. It allows us to use the notation of Section 13.6.1.

It is important to observe that **P1**–**P3** for the priority system implies **T1**–**T6** for the corresponding tandem fluid network, except that the second inequality in **T1** only holds as a weak inequality. However, as remarked in Section 13.6.1, the Laplace transform of the distribution of $(W, \widetilde{B})$ is then still given in Theorem 13.16.

The steady-state ages of the busy periods E can also be expressed in terms of the solution (W, L) to this Skorokhod problem, but it does *not always* equal $\widetilde{B}$ as in Section 13.6.1. To see this, notice that if class-1 traffic (highest priority) arrives to an empty system at time t, we have $W_2(t) = 0$, while $\widetilde{W}_2(t) > 0$ so that $\widetilde{B}_2(t) > 0$. However, it must hold that $E_2(t) = 0$.

Still, the following theorem shows that it is possible to express the distribution of (W, E) in terms of $(W, \widetilde{B})$.

Theorem 13.20 *Consider a priority fluid network for which* **P1**–**P3** *holds. Then for* $\omega, \beta \in \mathbb{R}^n_+$*, the transform* $\mathbb{E}\mathrm{e}^{-\langle\omega,W\rangle-\langle\beta,E\rangle}$ *equals*

$$\mathbb{E}e^{-\langle\omega,W\rangle-\langle\beta,\widetilde{B}\rangle} + \sum_{j=2}^{n} \mathbb{E}\left[e^{-\sum_{\ell=1}^{j-1}\omega_\ell W_\ell - \sum_{\ell=1}^{j-1}\beta_\ell \widetilde{B}_\ell}\left(1-e^{-\beta_j\widetilde{B}_j}\right); W_j = \ldots = W_n = 0\right].$$

Proof. In principle, E_j equals $\widetilde{B}_j$, except when $W_j = 0$. In fact, it follows from the above reasoning that

$$\begin{aligned}\mathbb{E}e^{-\langle\omega,W\rangle-\langle\beta,E\rangle} &= \mathbb{E}\left[e^{-\omega_1 W_1-\beta_1\widetilde{B}_1}; W_2 = \ldots = W_n = 0\right] \\ &\quad + \sum_{j=2}^{n}\mathbb{E}\left[e^{-\sum_{\ell=1}^{j}\omega_\ell W_\ell - \sum_{\ell=1}^{j}\beta_\ell\widetilde{B}_\ell}; W_j > 0, W_{j+1} = \ldots = W_n = 0\right].\end{aligned}$$

Now use the fact that $\{W_j > 0\}$ is the complement of $\{W_j = 0\}$ and rearrange terms. □

If the $A_2, \ldots, A_n$ are *strictly* increasing, it can be seen (for instance with Theorem 13.16) that

$$\mathbb{E}\left[e^{-\sum_{\ell=1}^{j-1}\omega_\ell W_\ell - \sum_{\ell=1}^{j-1}\beta_\ell\widetilde{B}_\ell}\left(1-e^{-\beta_j\widetilde{B}_j}\right); W_j = \ldots = W_n = 0\right] = 0.$$

Therefore, in that case, we have the equality in distribution $(W, E) \overset{\mathrm{d}}{=} (W, \widetilde{B})$.

Another important special case is when $A_1, \ldots, A_n$ are compound Poisson processes, say with intensities $\lambda_1, \ldots, \lambda_n$ respectively. Much is known about the resulting priority system, see for instance Jaiswal [171] for this and related models. To our knowledge, the distribution of (W, E) has not been investigated. However, it is given by Theorem 13.20 and Theorem 13.16 upon noting that $\theta^A_\ell(\omega) \to \lambda_\ell$ as $\omega \to \infty$. Since it is not so instructive to write out the resulting formulas, we leave this to the reader.

13.A Appendix: some calculations for a compound Poisson process with negative drift

In this appendix, we study a compound Poisson process Z with negative drift, and derive some results on the excursions of $Z - \underline{Z}$ from 0, just before its entrance to 0. These results are applied in Section 13.4.

Let us first fix the notation. Throughout this appendix, Z is a Lévy process on $(\Omega, \mathcal{F}, \mathbb{P})$ with Laplace exponent

$$\psi_{-Z}(\beta) := \log \mathbb{E}e^{-\beta Z(1)} = c\beta - \lambda\int_{\mathbb{R}_+}\left(1-e^{-\beta z}\right)F(dz),$$

where $c > 0$, $\lambda \in (0, \infty)$, and F is a probability distribution on $(0, \infty)$. That is, Z is a compound Poisson process under $\mathbb{P}$ with rate λ and negative drift $-c$, and its (positive) jumps are governed by F. We suppose that $\mathbb{E}Z(1) < 0$, so that Z drifts to $-\infty$. In analogy to Section 13.4, the inverse of ψ_{-Z} is denoted by Φ_{-Z}; it is uniquely defined since ψ_{-Z} is increasing. Observe that $\Phi_{-Z}(0) = 0$.

Set $T_0 = 0$, and let T_i denote the epoch of the i-th jump of Z. To the i-th jump of Z, we associate a vector of *marks*, denoted by $M_i \in \mathbb{R}^m_+$ (for some $m \in \mathbb{Z}_+$). We suppose that M_i is independent of the process $T \equiv \{T_n : n \geq 1\}$, and that it is also independent of $(Z(T_j) - Z(T_j-), M_j)$ for $j \neq i$. However, we allow for a dependency between M_i and $Z(T_i) - Z(T_i-)$. In fact, an interesting choice for M_i is $M_i = Z(T_i) - Z(T_i-)$ (so that $m = 1$).

Define τ_- as the first hitting time of zero, and N_- as the index of the last jump before τ_-, i.e.,

$$\tau_- := \inf\{t \geq 0 : Z(t) = 0\}, \quad N_- = \inf\{n \geq 0 : Z(T_{n+1}-) \leq 0\}.$$

Write $\mathbb{P}_\xi$ for the law of $Z + \xi$ under $\mathbb{P}$ with initial mark $M_0 = M$. We suppose that the initial condition (ξ, M) is independent of Z, and has the same distribution as $(Z(T_1) - Z(T_1-), M_1)$. Observe that both τ_- and N_- are $\mathbb{P}_\xi$-almost surely finite, and that (by the Markov property) the 'overshoot of the first excursion' $T_{N_-+1} - \tau_-$ has an exponential distribution with parameter λ.

In this appendix, it is our aim to characterize the $\mathbb{P}_\xi$-distribution of τ_- (excursion length), $\tau_- - T_{N_-}$ (excursion 'undershoot'), and M_{N_-} (mark of the last jump). Overshoots and undershoots have been studied extensively in the literature. However, as opposed to what we have here, these results are all related to the situation that a Lévy process can cross a boundary by jumping over it (strictly speaking, this is the only case where the terms 'overshoot' and 'undershoot' seem to be appropriate). See Doney and Kyprianou [121] for a recent contribution and for references.

In view of the results of Dufresne and Gerber [124], it is tempting to believe that $\tau_- - T_{N_-}$ has an exponential distribution. However, it turns out that this 'undershoot' has a completely different distribution.

Proposition 13.21 *We have for $\beta, \gamma \geq 0$ and $\kappa \in \mathbb{R}^m_+$,*

$$\begin{aligned}\mathbb{E}_\xi e^{-\beta(\tau_- - T_{N_-}) - \gamma\tau_- - \langle\kappa, M_{N_-}\rangle} &= \frac{[\beta + \gamma - c\Phi_{-Z}(\gamma) + \lambda]\,\mathbb{E}e^{-(\beta+\gamma+\lambda)\xi/c - \langle\kappa, M\rangle}}{\beta + \lambda\mathbb{E}e^{-(\beta+\gamma+\lambda)\xi/c}} \\ &= \frac{[\beta + \lambda\mathbb{E}_\xi e^{-\gamma\tau_-}]\,\mathbb{E}e^{-(\beta+\gamma+\lambda)\xi/c - \langle\kappa, M\rangle}}{\beta + \lambda\mathbb{E}e^{-(\beta+\gamma+\lambda)\xi/c}}.\end{aligned}$$

To prove this proposition, we need an auxiliary result on Poisson processes. Consider a Poisson point process $N(t)$ with parameter μ, and let ζ be a positive random variable, independent of N. Let $A(t)$ be the backward recurrence time process defined by N, that is the time from ζ to the nearest point to the left. The following lemma characterizes the joint distribution of $N(\zeta)$, $A(\zeta)$, and ζ.

Lemma 13.22 *We have for $\beta, \gamma \geq 0$ and $0 \leq s \leq 1$,*

$$\mathbb{E}s^{N(\zeta)}e^{-\beta A(\zeta) - \gamma\zeta} = \frac{\beta}{\beta + s\mu}\mathbb{E}e^{-(\beta+\gamma+\mu)\zeta} + \frac{s\mu}{\beta + s\mu}\mathbb{E}e^{-[\gamma+(1-s)\mu]\zeta}.$$

Proof. We only prove the claim for $\gamma = 0$; the general case follows by replacing the distribution of ζ by the (defective) distribution of $\tilde{\zeta}$ given by $\mathbb{E}e^{-\beta\tilde{\zeta}} = \mathbb{E}e^{-(\beta+\gamma)\zeta}$. Let $U_0 = 0$ and $U_1, U_2, \ldots$ be the location of consecutive points of N. Observe that

$$\begin{aligned}\mathbb{E}s^{N(\zeta)}e^{-\beta A(\zeta)} &= \sum_{n=0}^{\infty} s^n \mathbb{E}\left[e^{-\beta(\zeta - U_n)}; 0 \leq \zeta - U_n \leq U_{n+1} - U_n\right] \\ &= \sum_{n=0}^{\infty} s^n \int_0^\infty \int_0^t e^{-(\beta+\mu)(t-x)}\mathbb{P}_{U_n}(dx)\mathbb{P}_\zeta(dt) \\ &= \sum_{n=0}^{\infty} s^n \phi_n(\mu + \beta), \qquad (13.15)\end{aligned}$$

where

$$\phi_n(\beta) := \mathbb{E}\left[e^{-\beta(\zeta - U_n)}; \zeta \geq U_n\right].$$

Clearly, $\phi_0(\beta) = \mathbb{E}e^{-\beta\zeta}$. If we let B be the forward recurrence time process, we have for $n \geq 1$,

$$\begin{aligned}
\phi_n(\beta) &= \mathbb{E}\left[e^{-\beta(\zeta-U_n)}; \zeta \geq U_{n-1}\right] - \mathbb{E}\left[e^{-\beta(\zeta-U_n)}; U_{n-1} \leq \zeta < U_n\right] \\
&= \mathbb{E}\left[e^{-\beta(\zeta-U_{n-1})+\beta(U_n-U_{n-1})}; \zeta \geq U_{n-1}\right] - \mathbb{E}\left[e^{-\beta(\zeta-U_n)}; U_{n-1} \leq \zeta < U_n\right] \\
&= \mathbb{E}\left[e^{\beta(U_n-U_{n-1})}\right] \mathbb{E}\left[e^{-\beta(\zeta-U_{n-1})}; \zeta \geq U_{n-1}\right] \\
&\quad - \mathbb{E}\left[e^{\beta B(\zeta)} \middle| N(\zeta) = n-1\right] \mathbb{P}(N(\zeta) = n-1) \\
&= \frac{\mu}{\mu-\beta}\left[\phi_{n-1}(\beta) - \mathbb{P}(N(\zeta) = n-1)\right],
\end{aligned}$$

where we used the lack-of-memory property of the exponential distribution for the last equality. After iteration, we obtain

$$\phi_n(\beta) = \left(\frac{\mu}{\mu-\beta}\right)^n \mathbb{E}e^{-\beta\zeta} - \sum_{i=0}^{n-1} \left(\frac{\mu}{\mu-\beta}\right)^{n-i} \mathbb{P}(N(\zeta) = i).$$

Therefore, taking $0 < s < \beta/\mu$ (later we may use an analytic-continuation argument), we deduce from (13.15) that

$$\mathbb{E}\left[s^{N(\zeta)} e^{-\beta A(\zeta)}\right] = \mathbb{E}e^{-(\beta+\mu)\zeta} \sum_{n=0}^{\infty} \left(-\frac{s\mu}{\beta}\right)^n - \sum_{n=1}^{\infty} s^n \sum_{i=0}^{n-1} \left(-\frac{\mu}{\beta}\right)^{n-i} \mathbb{P}(N(\zeta) = i).$$

The double sum in this expression can be rewritten as

$$-\frac{s\mu}{\beta+s\mu} \sum_{i=0}^{\infty} s^i \mathbb{P}(N(\zeta) = i) = -\frac{s\mu}{\beta+s\mu} \mathbb{E}e^{-(1-s)\mu\zeta},$$

and the claim follows. □

Lemma 13.22 is the main ingredient to prove Proposition 13.21.

Proof of Proposition 13.21. The crucial yet simple observation is that

$$\begin{aligned}
&\mathbb{E}_\xi e^{-\beta(\tau_- - T_{N_-}) - \gamma\tau_- - \langle\kappa, M_{N_-}\rangle} \\
&\quad = \mathbb{E}_\xi\left[e^{-\beta(\tau_- - T_{N_-}) - \gamma\tau_- - \langle\kappa, M_{N_-}\rangle}; N_- = 0\right] + \mathbb{E}_\xi\left[e^{-\beta(\tau_- - T_{N_-}) - \gamma\tau_- - \langle\kappa, M_{N_-}\rangle}; N_- \geq 1\right] \\
&\quad = \mathbb{E}e^{-(\lambda+\beta+\gamma)\xi/c - \langle\kappa, M\rangle} + \mathbb{E}_\xi\left[e^{-\beta(\tau_- - T_{N_-}) - \gamma\tau_- - \langle\kappa, M_{N_-}\rangle}; N_- \geq 1\right]. \qquad (13.16)
\end{aligned}$$

To analyze the second term, we exploit the fact that there are several excursions of $Z - \underline{Z}$ from 0. Therefore, we set

$$C(t) := \inf\{s \geq 0 : Z(s) - Z(0) = -t\},$$

where an infimum over an empty set should be interpreted as infinity.

It is obvious that C is a subordinator with drift $1/c$, and that it jumps at rate λ/c with jumps distributed as τ_- under $\mathbb{P}_\xi$. This observation implies with Theorem VII.1 of Bertoin [43] that

$$\Phi_{-Z}(\gamma) = \frac{\gamma}{c} + \frac{\lambda}{c}\left(1 - \mathbb{E}_\xi e^{-\gamma\tau_-}\right). \qquad (13.17)$$

Lemma 13.22 can be applied to the Poisson process N constituted by the jump epochs of C, $\mu = \lambda/c$, and $\zeta = \xi$. Each jump of C corresponds to an excursion of $Z - \underline{Z}$ from 0, for which the 'excursion overshoot', the excursion length, and the marks of the last jump are of

interest. Observe that these quantities have the same distribution as $\tau_- - T_{N_-}$, τ_-, and M_{N_-} respectively. Using the notation of Lemma 13.22, this yields

$$\begin{aligned}&\mathbb{E}_\xi\left[e^{-\beta(\tau_- - T_{N_-})-\gamma\tau_- - \langle\kappa, M_{N_-}\rangle}; N_- \geq 1\right]\\ &\quad = \quad \mathbb{E}\left[\left(\mathbb{E}_\xi e^{-\gamma\tau_-}\right)^{N(\xi)-1} e^{-\beta A(\xi)/c-\gamma\xi/c}; N(\xi)\geq 1\right]\mathbb{E}_\xi e^{-\beta(\tau_- - T_{N_-})-\gamma\tau_- - \langle\kappa, M_{N_-}\rangle}.\end{aligned}$$

Therefore, Lemma 13.22 yields

$$\begin{aligned}\mathbb{E}\left[s^{N(\xi)-1}e^{-\beta A(\xi)/c-\gamma\xi/c}; N(\xi)\geq 1\right] &= \frac{\mathbb{E}\left[s^{N(\xi)}e^{-\beta A(\xi)/c-\gamma\xi/c}\right]-\mathbb{E}e^{-(\lambda+\beta+\gamma)\xi/c}}{s}\\ &= \frac{\lambda}{\lambda s+\beta}\left[\mathbb{E}e^{-((1-s)\lambda+\gamma)\xi/c}-\mathbb{E}e^{-(\lambda+\beta+\gamma)\xi/c}\right].\end{aligned}$$

Upon combining the preceding two displays with (13.16), we arrive at

$$\begin{aligned}&\mathbb{E}_\xi e^{-\beta(\tau_- - T_{N_-})-\gamma\tau_- - \langle\kappa, M_{N_-}\rangle}\\ &\quad = \quad \frac{\left[\beta+\lambda\mathbb{E}_\xi e^{-\gamma\tau_-}\right]\mathbb{E}e^{-(\lambda+\beta+\gamma)\xi/c-\langle\kappa,M\rangle}}{\lambda\mathbb{E}_\xi e^{-\gamma\tau_-}+\beta-\lambda\mathbb{E}e^{-(\lambda(1-\mathbb{E}_\xi e^{-\gamma\tau_-})+\gamma)\xi/c}+\lambda\mathbb{E}e^{-(\lambda+\beta+\gamma)\xi/c}},\end{aligned}$$

which, with the help of (13.17), reduces to

$$\frac{\left[\beta+\gamma-c\Phi_{-Z}(\gamma)+\lambda\right]\mathbb{E}e^{-(\beta+\gamma+\lambda)\xi/c-\langle\kappa,M\rangle}}{\beta+\gamma-c\Phi_{-Z}(\gamma)-\lambda\left(\mathbb{E}e^{-\Phi_{-Z}(\gamma)\xi}-1\right)+\lambda\mathbb{E}e^{-(\beta+\gamma+\lambda)\xi/c}}.$$

By definition of Φ_{-Z}, we have

$$\gamma = \psi_{-Z}(\Phi_{-Z}(\gamma)) = c\Phi_{-Z}(\gamma)+\lambda\left(\mathbb{E}e^{-\Phi_{-Z}(\gamma)\xi}-1\right),$$

and the claim follows. □

CHAPTER 14

Extremes of Markov-additive processes

This chapter investigates the extremes of a continuous-time Markov-additive process with one-sided jumps. These processes are generalizations of spectrally positive Lévy processes. To study the extremes jointly with the epochs at which they are 'attained', we investigate discrete-time Markov-additive processes and use an embedding in the spirit of Chapter 12 to relate these to a continuous-time setting.

Our results on extremes are first applied to determine the steady-state buffer-content distribution of several single-station queueing systems. We show that our framework comprises many models dealt with earlier, but that it also enables us to derive new results, for instance for the M/M/∞-driven fluid queue. At the same time, our setup offers interesting insights into the connections between the approaches developed so far, including matrix-analytic techniques, martingale methods, the rate-conservation approach, and the occupation-measure method.

Then we turn to tandem fluid networks driven by a Markov-additive process. For these networks, we partly extend the results of Chapter 13 by showing how the Laplace transform of the steady-state buffer-content vector can be found. Interestingly, this transform has a matrix quasi-product form. Fluid-driven priority systems also have this property.

14.1 Introduction

There exists a vast body of literature on Markov-modulated processes, and queueing systems with continuous-time Markov-additive input in particular. For instance, as a special case, so-called fluid-flow models have been under continuous investigation over the past three decades, a key reference being the work of Anick *et al.* [11]. The present chapter is motivated by an appealing formula of Asmussen and Kella [24] for (the Laplace transform of) the buffer content in a fluid queue driven by a Markov-additive process with nonnegative jumps and finitely many background states. Generalizing the Pollaczek-Khinchine formula, Asmussen and Kella show that the Laplace transform of the steady-state buffer content (jointly with the state of the background process) can be expressed in terms of the Laplace exponent of the Markov-additive process and a generally *unknown* vector. In view of the results on quasi-product forms for Lévy-driven networks in Chapter 13 and recent insights into fluid-flow networks by Kella [184], our initial aim was to understand how this formula generalizes to the setting of tandem networks

and priority systems.

We have seen in Section 1.1 that (fluid) queues can often be investigated through extremes. The results are then not only relevant for queueing theory, but also for risk theory and mathematical finance; see Section 1.1.5. The use of Markov-additive processes in a risk framework is exemplified by Asmussen [18, Ch. VI], Asmussen and Rolski [29], and Miyazawa [232]; see also the references therein. Fluctuations of Markov-additive processes have also been recently examined in the context of financial contracts, see Jobert and Rogers [176], and Pistorius [255].

By exploiting this relationship with extremes, Chapter 13 shows that the analysis of tandem networks with Lévy input requires knowledge of the maximum $\overline{X}$ of a Lévy process X, but importantly also of the epoch $\overline{F}^X$ at which this maximum is (first) 'attained'. More specifically, the arguments used in the preceding chapter indicate how to convert the transform of $(\overline{X}, \overline{F}^X)$ to the joint transform of *all* steady-state buffer contents in the network. In the Lévy context, the transform of $(\overline{X}, \overline{F}^X)$ is known, and can be deduced from Theorem VII.4 of Bertoin [43]: if X is a Lévy process with nonnegative jumps (but not an increasing subordinator) while $\mathsf{E}X(1) < 0$, we have for $\alpha, \beta \geq 0$,

$$\mathsf{E}e^{-\alpha\overline{F}^X - \beta\overline{X}} = -\mathsf{E}X(1)\frac{\beta - \Phi_{-X}(\alpha)}{\psi_{-X}(\beta) - \alpha}, \tag{14.1}$$

where $\psi_{-X}(\beta) := \log \mathsf{E}e^{-\beta X(1)}$ is the Laplace exponent of $-X$, and Φ_{-X} is its inverse (which exists since ψ_{-X} increases on $[0, \infty)$). Note that this formula follows from Proposition 11.6 by letting $q \to 0$, and that the choice $\alpha = 0$ yields a generalized Pollaczek-Khinchine formula since $\Phi_{-X}(0) = 0$. Therefore, before being able to treat fluid networks with Markov-additive input, first a better understanding of the corresponding *single* fluid queue is needed. More concretely, extension of the results for networks with Lévy inputs to our setting requires the 'Markov-additive counterpart' of (14.1). The primary goal of this chapter is to find this transform and to understand the relationship with results that have been obtained earlier.

Theorem 14.19, which is one of our main results, holds for a general Markov-additive process with nonnegative jumps and can be regarded as a true matrix version of (14.1). There are many more subtleties in the Markov-additive case than in the Lévy case. For instance, the identity $\psi_{-X}(\Phi_{-X}(\alpha)) = \alpha$ can be generalized to matrices in two different ways, and the two resulting matrices turn out to be both relevant for examining the extremes of X. In fact, we believe that these two matrix analogues of $\Phi_{-X}(\alpha)$ lie at the heart of fluctuation theory for Markov-additive processes, and that they also play a fundamental role in exit problems and transient queueing analysis.

As a by-product of our analysis, we gain more insight into the unknown vector in the formula of Asmussen and Kella, and give an alternative interpretation. Importantly, this enables us to relate their results, as well as our own, to recent developments in the literature on matrix-analytic methods for Markov-additive models.

The derivation of our results relies on Wiener-Hopf theory for an embedded process. Perhaps for historic reasons, the Wiener-Hopf technique is sometimes regarded as a complex-analysis tool from which probabilistic insight cannot be obtained. However, inspired by Kennedy's [189] proof of Theorem 11.1, we are able to give *interpretations* of all our results in terms of a last-passage process. This shows that our approach to Markov-additive processes is essentially different from the occupation-measure method of Asmussen [18], the martingale method of Asmussen and Kella [24], and the rate-conservation method of Miyazawa [232]. Still, we believe that many of our results can also be obtained with other methods, and that each method has its own advantages. We stress that our approach offers valuable insight into the connections between the aforementioned branches of research. At the same time, we are the first to characterize the distributions of extremes for general Markov-additive processes with one-sided jumps.

On the technical level, two steps are crucial. In the first place, we convert our continuous-time process to a discrete-time Markov-additive process by using an embedding. The maximum of the original, continuous-time process coincides with the maximum of the embedded process. In the special case of continuous Markov-additive processes, this idea has been applied by Asmussen [16]. However, by using this embedding we lose information on the epoch at which the extreme is 'attained', and we therefore also apply a second idea: we impose a step-dependent killing mechanism through which we keep track of the 'time' that passes in the continuous-time process between embedding epochs. The resulting procedure enables us to find the counterpart of (14.1). Note that this idea has also been used in Section 12.4.2 in the context of perturbed risk processes. The most important assumption (Assumption 14.1) underlying the results of this chapter entails that the downward jumps of the embedded process are 'memoryless' in a *space-time* sense (not only in a space sense!), which can be regarded as an analogue of the skip-free property in many matrix-analytic models. We remark that the killing technique is an alternative to other approaches that have been proposed for fluid-flow models [6, 15, 37].

Our results for discrete-time processes are of independent interest; they unify and extend (parts of) Section 1.12 and Chapter 5 of Prabhu [263]. This is exemplified (in Section 14.4.1) by analyzing a ramification of a queueing system with Markov-modulated ON/OFF input introduced by Cohen [78]. This system does not fall into the class of Markov-additive processes, but we are still able to study it directly with the help of our results on discrete-time processes. As a further application, we use this ON/OFF-type model to investigate the M/M/∞-driven fluid queue.

Although we give fixed-point and matrix equations for all matrices that play an important role in the theory, it is still an interesting and challenging issue to devise efficient algorithms for numerically calculating these matrices. Therefore, our work could serve as a first step towards the development of such new numerical methods; it could initiate an analysis in the spirit of many results in the matrix-analytic literature. We find this indispensable for a successful application of the theory.

This chapter is organized as follows. First, in Section 14.2, we start with the analysis of the extremes of a discrete-time Markov-additive process. The insight that we obtain is then applied to continuous-time Markov-additive processes in Section 14.3. Section 14.4 casts our results on extremes into the queueing setting, and some examples are given in Section 14.5. We address fluid networks in Section 14.6, and we end the chapter with some concluding remarks (Section 14.7).

14.2 A discrete-time process and its extremes

This section introduces the discrete-time three-dimensional process $(S,T,J) = \{(S_n,T_n,J_n) : n \geq 0\}$. Although this process may look quite specific at first sight, we show in Sections 14.4–14.7 that it is highly versatile: it can be used to study the steady-state buffer content (in conjunction with the steady-state age of the busy period) for a broad class of queueing systems, including networks and priority queues.

14.2.1 Definitions and assumptions

The process (S,T,J) takes values in $\mathbb{R}\times\mathbb{R}_+\times\mathcal{I}$, where $\mathcal{I}$ is a finite set with $N_+ + N_-$ elements. We write $\mathcal{I}_+$ for the first N_+ elements (which we call '+-points', as made clear below), and $\mathcal{I}_-$ for the last N_- elements (which we call '$-$-points'). The component J is interpreted as a 'random environment'. Furthermore, (S,T) is a two-dimensional random walk in this environment. We suppose that (S,T,J) is defined on some measurable space $(\Omega,\mathcal{F})$.

Of primary interest is the minimum $\underline{S}$ and the maximum $\overline{S}$ of the process S. After setting $\underline{F}^S := \inf\{n \geq 0 : S_n = \inf_{k\geq 0} S_k\}$ and $\overline{F}^S := \inf\{n \geq 0 : S_n = \sup_{k\geq 0} S_k\}$, these are defined as $\underline{S} := S_{\underline{F}^S}$ and $\overline{S} := S_{\overline{F}^S}$ respectively. The process T is interpreted as the 'real' time that passes between the (discrete) time epochs; it cannot decrease. Therefore, it is also of interest to study $\underline{T} := T_{\underline{F}^S}$, $\overline{T} := T_{\overline{F}^S}$, $\underline{J} := J_{\underline{F}^S}$ and $\overline{J} := J_{\overline{F}^S}$. The aim of this section is to fully characterize the joint distributions of the triplet $(\underline{S}, \underline{T}, \underline{J})$ if S drifts to $+\infty$, and $(\overline{S}, \overline{T}, \overline{J})$ if S drifts to $-\infty$, under a measure specified below.

Suppose that $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$ (with corresponding integration operator $\mathbb{E}$) such that (S, T, J) is a (discrete-time) Markov process on $\mathbb{R} \times \mathbb{R}_+ \times \mathcal{I}$. To describe the transition kernel, we need the vectors (specifically, their $\mathbb{P}$-distributions) $\{(\sigma^{jk}, U^{jk}) : j \in \mathcal{I}_+, k \in \mathcal{I}\}$, taking values in $[0, \infty)^2$, and also the vectors $\{(\tau^j, D^j) : j \in \mathcal{I}_-\}$, taking values in $[0, \infty)\times(0, \infty)$. Both sets of vectors are defined on $(\Omega, \mathcal{F})$. We remark that $\mathbb{P}(\sigma^{jk} = 0)$, $\mathbb{P}(U^{jk} = 0)$, and $\mathbb{P}(\tau^j = 0)$ are allowed to be strictly positive. The transition kernel corresponding to (S, T, J) is given by

$$p((s,t,j),(s+dv,t+dw,k)) = \begin{cases} p^J_{jk}\mathbb{P}\left(U^{jk} \in dv, \sigma^{jk} \in dw\right) & \text{if } j \in \mathcal{I}_+, k \in \mathcal{I}; \\ p^J_{jk}\mathbb{P}\left(-D^j \in dv, \tau^j \in dw\right) & \text{if } j \in \mathcal{I}_-, k \in \mathcal{I}. \end{cases}$$

The letters U and D stand for 'up' and 'down'. The U^{jk} and $-D^j$ can be interpreted as 'jump sizes', whereas the σ^{jk} and τ^j reflect 'sojourn times'.

The transition matrix of J is denoted by $\boldsymbol{P}^J$, and we suppose that it is irreducible. The unique stationary distribution of J is written as $\boldsymbol{\pi}_J$. For $k \in \mathcal{I}$, we write $\mathbb{P}_k$ for the law of (S, T, J) given $S_0 = T_0 = 0$ and $J_0 = k$. To avoid trivialities, we suppose throughout that both N_- and N_+ are nonzero, and that not all of the U^{jk} are degenerate at zero.

The following assumption is crucial in our analysis.

Assumption 14.1 *For any $j \in \mathcal{I}_-$, there exists some $\lambda^\alpha_j > 0$, $\mu^\alpha_j \in (0, 1]$ such that*

$$\mathbb{E}e^{-\alpha\tau^j - \beta D^j} = \mu^\alpha_j \frac{\lambda^\alpha_j}{\lambda^\alpha_j + \beta}, \quad \alpha, \beta \geq 0,$$

where $\mu^0_j = 1$.

Assumption 14.1 can be thought of as (a generalized version of) a memoryless property for the distribution of the jump sizes and sojourn times in the $-$-points. We suppose that this assumption holds *throughout this section.*

In many of the proofs in this section, an important role is played by a family of probability measures $\{\mathbb{P}^\alpha : \alpha \geq 0\}$. Under $\mathbb{P}^\alpha$, the distribution of U^{jk} is potentially defective, and the relation with $\mathbb{P}$ is given by $\mathbb{P}^\alpha(U^{jk} \in dv) = \mathbb{E}[e^{-\alpha\sigma^{jk}}; U^{jk} \in dv]$. Similarly, $\mathbb{P}^\alpha(D^j \in dv) = \mathbb{E}[e^{-\alpha\tau^j}; D^j \in dv]$. Furthermore, (S, J) is a discrete-time Markov process under $\mathbb{P}^\alpha$ with transition kernel

$$p^\alpha((s,j),(s+dv,k)) = \begin{cases} p^J_{jk}\mathbb{P}^\alpha(U^{jk} \in dv) & \text{if } j \in \mathcal{I}_+, k \in \mathcal{I}; \\ p^J_{jk}\mathbb{P}^\alpha(-D^j \in dv) & \text{if } j \in \mathcal{I}_-, k \in \mathcal{I}. \end{cases}$$

The $\mathbb{P}^\alpha$-law for which $S_0 = 0$ and $J_0 = k$ is denoted by $\mathbb{P}^\alpha_k$.

We note that $\{(S_n, J_n) : n \geq 0\}$ is a discrete-time *Markov-additive process* under each of the measures $\mathbb{P}_k$, $\mathbb{P}^\alpha_k$ for $k \in \mathcal{I}$ and $\alpha \geq 0$. As a result, the powerful Wiener-Hopf factorization for these processes is available. More details can be found in Asmussen [19, Sec. XI.2.2f].

In order to use this technique, we need some more notation related to *time-reversion*. Let us therefore introduce the time-reversed transition probabilities

$$\widehat{p}^J_{jk} = \frac{\pi_J(k)}{\pi_J(j)} p^J_{kj},$$

constituting the transition matrix $\widehat{\boldsymbol{P}}^J$; here $\pi_J(k)$ denotes the k-th element of $\boldsymbol{\pi}_J$.

Moreover, let $\widehat{\mathbb{P}}$ be a probability measure on $(\Omega, \mathcal{F})$ (with expectation operator $\widehat{\mathbb{E}}$) such that (S, T, J) is a Markov process with transition probability

$$\widehat{p}((s,t,j),(s+dv,t+dw,k)) = \begin{cases} \widehat{p}^J_{jk}\mathbb{P}\left(U^{kj} \in dv, \sigma^{kj} \in dw\right) & \text{if } j \in \mathcal{I}, k \in \mathcal{I}_+; \\ \widehat{p}^J_{jk}\mathbb{P}\left(-D^k \in dv, \tau^k \in dw\right) & \text{if } j \in \mathcal{I}, k \in \mathcal{I}_-. \end{cases}$$

This should be compared to the kernel p: the indices differ. The $\widehat{\mathbb{P}}$-law for which $S_0 = T_0 = 0$ and $J_0 = k$ is denoted by $\widehat{\mathbb{P}}_k$.

Finally, we also define the probability measures $\widehat{\mathbb{P}}^\alpha$ by requiring that (S, J) is a Markov process with transition kernel

$$\widehat{p}^\alpha((s,j),(s+dv,k)) = \begin{cases} \widehat{p}^J_{jk}\mathbb{E}[e^{-\alpha\sigma^{kj}}; U^{kj} \in dv] & \text{if } j \in \mathcal{I}, k \in \mathcal{I}_+; \\ \widehat{p}^J_{jk}\mathbb{E}[e^{-\alpha\tau^k}; -D^k \in dv] & \text{if } j \in \mathcal{I}, k \in \mathcal{I}_-, \end{cases}$$

and $\widehat{\mathbb{P}}^\alpha_k$ is defined as the $\widehat{\mathbb{P}}^\alpha$-law of this process given $S_0 = 0$ and $J_0 = k$.

14.2.2 Notation

We now introduce some convenient matrix notation. It is not our aim to define every single matrix that we use, but rather to present a set of notation rules that we follow throughout this chapter.

Vectors are always written as column vectors. By writing $\boldsymbol{E}$ instead of $\mathbb{E}$, we indicate that we deal a matrix or vector. For instance, we define

$$\boldsymbol{E}\left[S_1; J_1\right] := \left\{\mathbb{E}_j\left[S_1; J_1 = k\right] : j, k \in \mathcal{I}\right\},$$

and the j-th element of the vector $\boldsymbol{E}S_1$ is $\mathbb{E}_j S_1$. We use a similar convention for $\boldsymbol{P}$ and $\mathbb{P}$, thereby defining for instance the vector $\boldsymbol{P}(S_1 > 0)$.

A given $(\mathcal{I} \times \mathcal{I})$-matrix $\boldsymbol{A}$ (which is sometimes best thought of as a mapping from $\mathbb{R}^{|\mathcal{I}|}$ to $\mathbb{R}^{|\mathcal{I}|}$) is written in *block form* as

$$\boldsymbol{A} \equiv \begin{pmatrix} \boldsymbol{A}_{++} & \boldsymbol{A}_{+-} \\ \boldsymbol{A}_{-+} & \boldsymbol{A}_{--} \end{pmatrix},$$

where, for instance, $\boldsymbol{A}_{++}$ is an $(\mathcal{I}_+ \times \mathcal{I}_+)$-matrix corresponding to transitions from $+$-points to $+$-points. The $(\mathcal{I} \times \mathcal{I})$-identity matrix, denoted by $\boldsymbol{I}$, consists of the blocks $\boldsymbol{I}_{++}$, $\boldsymbol{0}_{+-}$, $\boldsymbol{0}_{-+}$, and $\boldsymbol{I}_{--}$ in self-evident notation. In conjunction with integration, the subscript '$-$' or '$+$' of $\boldsymbol{E}$ indicates the *row*, and '$\in +$' or '$\in -$' the *column* of the appropriate matrix block. For instance, $\boldsymbol{E}\left[S_1; J_1\right]$ consists of four blocks, which we write as $\boldsymbol{E}_+\left[S_1; J_1 \in +\right]$, $\boldsymbol{E}_+\left[S_1; J_1 \in -\right]$, $\boldsymbol{E}_-\left[S_1; J_1 \in +\right]$, and $\boldsymbol{E}_-\left[S_1; J_1 \in -\right]$. The matrix with the first and last two blocks is written as $\boldsymbol{E}_+\left[S_1; J_1\right]$ and $\boldsymbol{E}_-\left[S_1; J_1\right]$ respectively, and $\widehat{\boldsymbol{E}}_\pm\left[S_1; J_1\right]$ is defined analogously, but with $\mathbb{E}_j$ replaced by $\widehat{\mathbb{E}}_j$.

Similar conventions apply to vectors: the restriction of the vector $\boldsymbol{E}S_1$ to $\mathcal{I}_+$ (or $\mathcal{I}_-$) is written as $\boldsymbol{E}_+S_1$ (or $\boldsymbol{E}_-S_1$). Note that we have the relation $\boldsymbol{E}_+S_1 = \boldsymbol{E}_+\left[S_1; J_1\right]\boldsymbol{1} = \boldsymbol{E}_+\left[S_1; J_1 \in +\right]\boldsymbol{1}_+ + \boldsymbol{E}_+\left[S_1; J_1 \in -\right]\boldsymbol{1}_-$, where $\boldsymbol{1}$ stands for the $\mathcal{I}$-vector with ones, and similarly for $\boldsymbol{1}_+$ and $\boldsymbol{1}_-$. The $\mathcal{I}$-vector with zeroes is written as $\boldsymbol{0}$, and consists of $\boldsymbol{0}_+$ and $\boldsymbol{0}_-$.

The diagonal matrix with the vector $\boldsymbol{E}S_1$ on its diagonal is written as $\mathrm{diag}(\boldsymbol{E}S_1)$. For example, $\boldsymbol{I}_{++} = \mathrm{diag}(\boldsymbol{1}_+)$. We also write $\mathrm{diag}\left(\lambda^\alpha/(\lambda^\alpha + i\beta)\right)$ for the $(\mathcal{I}_- \times \mathcal{I}_-)$-diagonal matrix with element (j,j) equal to $\lambda_j^\alpha/(\lambda_j^\alpha + i\beta)$, and $\mathrm{diag}(\lambda^\alpha)$ is defined similarly. Moreover, we also set

$${}^{\#}\boldsymbol{A} := \mathrm{diag}\left(\frac{1}{\boldsymbol{\pi}_J}\right)\boldsymbol{A}'\ \mathrm{diag}\left(\boldsymbol{\pi}_J\right), \tag{14.2}$$

where '$'$' denotes matrix transpose. In conjunction with block notation, # has priority over block notation: ${}^{\#}\boldsymbol{A}_{++}$ is the $(+,+)$-block of ${}^{\#}\boldsymbol{A}$.

For instance, if we set for $\alpha \geq 0$, $\beta \in \mathbb{R}$

$$F_{jk}(\alpha,\beta) := \begin{cases} p^J_{jk}\mathbb{E}e^{-\alpha\sigma^{jk}+i\beta U^{jk}} & \text{if } j \in \mathcal{I}_+, k \in \mathcal{I}; \\ p^J_{jk}\mathbb{E}e^{-\alpha\tau^j - i\beta D^j} & \text{if } j \in \mathcal{I}_-, k \in \mathcal{I}, \end{cases}$$

this defines not only the matrix-transform of the transition kernel $\boldsymbol{F}(\alpha,\beta) := \{F_{jk}(\alpha,\beta) : j,k \in \mathcal{I}\}$, but also its four block matrices; note that Assumption 14.1 prespecifies the structure of $\boldsymbol{F}_{-+}(\alpha,\beta) = \boldsymbol{E}_-\left[e^{-\alpha T_1+i\beta S_1}; J_1 \in +\right]$ and $\boldsymbol{F}_{--}(\alpha,\beta) = \boldsymbol{E}_-\left[e^{-\alpha T_1+i\beta S_1}; J_1 \in -\right]$. The time-reversed counterpart is written as $\widehat{\boldsymbol{F}}(\alpha,\beta)$, i.e., $\widehat{\boldsymbol{F}}(\alpha,\beta) := {}^{\#}\boldsymbol{F}(\alpha,\beta)$. Note that in particular $\widehat{\boldsymbol{P}}^J = {}^{\#}\boldsymbol{P}^J$.

14.2.3 The ladder heights of S

The goal of this subsection is to characterize the $\mathbb{P}_k$-distribution of (S,T,J) at the first strict ascending ladder epoch of S and at its first strict descending ladder epoch. We do not impose conditions on the drift of S yet.

The first strict ascending ladder epoch and the first weak descending ladder epoch of S are defined as

$$\tau_+ = \inf\{n \geq 1 : S_n > 0\}, \quad \tau_- = \inf\{n \geq 1 : S_n \leq 0\}.$$

Its first strict descending ladder epoch, for which the weak inequality is replaced by a strict inequality, is denoted by $\tilde{\tau}_-$.

The distribution of $(S_{\tau_+}, T_{\tau_+}, J_{\tau_+})$

In order to facilitate the investigation of the ascending ladder structure of (S,T,J), we first prove a useful lemma related to τ_-. For notational convenience, we define the matrix $\widehat{\boldsymbol{P}}^\alpha = \{\widehat{P}^\alpha_{jk} : j,k \in \mathcal{I}\}$ as

$$\widehat{\boldsymbol{P}}^\alpha := \widehat{\boldsymbol{E}}\left[e^{-\alpha T_{\tau_-}}; J_{\tau_-}\right].$$

This matrix admits a block form as described in Section 14.2.2. As before, when integrating a defective random variable, we only carry out the integration over the set where the random variable is both finite and well-defined: in the above definition of $\widehat{\boldsymbol{P}}^\alpha$, it is tacitly assumed that $\tau_- < \infty$.

Lemma 14.2 *Suppose that Assumption 14.1 holds. For $\alpha \geq 0, \beta \in \mathbb{R}$, we have*

$${}^{\#}\widehat{\boldsymbol{E}}\left[e^{-\alpha T_{\tau_-}+i\beta S_{\tau_-}}; J_{\tau_-}\right] = \begin{pmatrix} \boldsymbol{F}_{++}(\alpha, i\infty) & \boldsymbol{F}_{+-}(\alpha, i\infty) \\ \operatorname{diag}\left(\frac{\lambda^\alpha}{\lambda^\alpha + i\beta}\right) {}^{\#}\widehat{\boldsymbol{P}}^\alpha_{-+} & \operatorname{diag}\left(\frac{\lambda^\alpha}{\lambda^\alpha + i\beta}\right) {}^{\#}\widehat{\boldsymbol{P}}^\alpha_{--} \end{pmatrix}$$

Proof. After recalling that τ_- is a *weak* ladder epoch, it is immediate that for $\alpha \geq 0$, $j \in \mathcal{I}$, $k \in \mathcal{I}_+$,

$$\widehat{\mathbb{E}}_j\left[e^{-\alpha T_{\tau_-}+i\beta S_{\tau_-}}; J_{\tau_-} = k\right] = \widehat{p}^J_{jk}\mathbb{E}\left[e^{-\alpha\tau^{kj}}; U^{kj} = 0\right] = \widehat{F}_{jk}(\alpha, i\infty).$$

Hence, it remains to calculate

$$\widehat{\boldsymbol{E}}\left[e^{-\alpha T_{\tau_-}+i\beta S_{\tau_-}}; J_{\tau_-} \in -\right] = \widehat{\boldsymbol{E}}^\alpha\left[e^{i\beta S_{\tau_-}}; J_{\tau_-} \in -\right].$$

To find an expression for this quantity, we directly apply the idea of Lemma VIII.5.1 of Asmussen [19], as follows. Evidently, for $j \in \mathcal{I}$, $k \in \mathcal{I}_-$, we have

$$\widehat{\mathbb{P}}_j^\alpha \left(S_{\tau_-} < -x, J_{\tau_-} = k \right) = \sum_{n=1}^{\infty} \widehat{\mathbb{P}}_j^\alpha \left(S_{\tau_-} < -x, \tau_- = n, J_{\tau_-} = k \right).$$

Conditioning on S_{n-1} and using Assumption 14.1, we see that the summands equal

$$\widehat{\mathbb{E}}_j^\alpha \left[\mu_k^\alpha e^{-\lambda_k^\alpha (x + S_{n-1})}; \tau_- > n-1, J_{\tau_-} = k \right] = e^{-\lambda_k^\alpha x} \widehat{\mathbb{E}}_j^\alpha \left[\mu_k^\alpha e^{-\lambda_k^\alpha S_{n-1}}; \tau_- > n-1, J_{\tau_-} = k \right],$$

since the value of the n-th increment should (in absolute terms) be larger than $x + S_{n-1}$. Importantly, this is exponential in x, so that we obtain

$$\widehat{\mathbb{E}}_j^\alpha \left[e^{i\beta S_{\tau_-}}; J_{\tau_-} = k \right] = \frac{\lambda_k^\alpha}{\lambda_k^\alpha + i\beta} \sum_{n=1}^{\infty} \widehat{\mathbb{E}}_j^\alpha \left[\mu_k^\alpha e^{-\lambda_k^\alpha S_{n-1}}; \tau_- > n-1, J_{\tau_-} = k \right].$$

The latter sum is calculated by inserting $\beta = 0$ into this identity. □

The above lemma requires knowledge of $\widehat{\boldsymbol{P}}^\alpha$. The following proposition gives a fixed-point equation for this matrix, so that it can be found numerically. Write $\boldsymbol{F}_{++}^\alpha(dx)$ for the measure-valued $(\mathcal{I}_+ \times \mathcal{I}_+)$-matrix with element (j,k) equal to $p_{jk}^J \mathbb{P}^\alpha(U^{jk} \in dx)$ for $j,k \in \mathcal{I}_+$, and define $\boldsymbol{F}_{+-}^\alpha(dx)$ similarly.

Proposition 14.3 *For $\alpha \geq 0$, we have*

$$\begin{aligned} {}^{\#}\widehat{\boldsymbol{P}}_{--}^\alpha &= \operatorname{diag}(\mu^\alpha)\, \boldsymbol{P}_{--}^J + \int_{(0,\infty)} e^{{}^{\#}\widehat{\boldsymbol{Q}}_{--}^\alpha x}\, {}^{\#}\widehat{\boldsymbol{P}}_{-+}^\alpha \boldsymbol{F}_{+-}^\alpha(dx), \\ {}^{\#}\widehat{\boldsymbol{P}}_{-+}^\alpha &= \operatorname{diag}(\mu^\alpha)\, \boldsymbol{P}_{-+}^J + \int_{(0,\infty)} e^{{}^{\#}\widehat{\boldsymbol{Q}}_{--}^\alpha x}\, {}^{\#}\widehat{\boldsymbol{P}}_{-+}^\alpha \boldsymbol{F}_{++}^\alpha(dx), \end{aligned}$$

where the integral should be understood as componentwise integration, and ${}^{\#}\widehat{\boldsymbol{Q}}_{--}^\alpha$ is specified by

$${}^{\#}\widehat{\boldsymbol{Q}}_{--}^\alpha = -\left[\boldsymbol{I}_{--} - {}^{\#}\widehat{\boldsymbol{P}}_{-+}^\alpha \left(\boldsymbol{I}_{++} - \boldsymbol{F}_{++}(\alpha, i\infty) \right)^{-1} \boldsymbol{F}_{+-}(\alpha, i\infty) - {}^{\#}\widehat{\boldsymbol{P}}_{--}^\alpha \right] \operatorname{diag}(\lambda^\alpha).$$

Proof. Write $\tau_-(x) := \inf\{n > 0 : S_n \leq -x\}$ for $x \geq 0$. For $j \in \mathcal{I}$ and $k \in \mathcal{I}_-$, we have by the Markov property

$$\widehat{P}_{jk}^\alpha \equiv \widehat{\mathbb{P}}_j^\alpha(J_{\tau_-} = k) = \widehat{p}_{jk}^J \mu_k^\alpha + \sum_{\ell \in \mathcal{I}_+} \widehat{p}_{j\ell}^J \int_{(0,\infty)} \mathbb{P}^\alpha(U^{\ell j} \in dx) \widehat{\mathbb{P}}_\ell^\alpha(J_{\tau_-(x)} = k).$$

Note that the integration interval for $U^{\ell j}$ is $(0,\infty)$, because if $U^{\ell j}$ were 0, then J_{τ_-} would be in $\mathcal{I}_+$. It remains to show that

$$\widehat{\mathbb{P}}_\ell^\alpha(J_{\tau_-(x)} = k) = \sum_{j \in \mathcal{I}_-} \widehat{\mathbb{P}}_\ell^\alpha(J_{\tau_-} = j) \left[e^{\widehat{\boldsymbol{Q}}_{--}^\alpha x} \right]_{jk},$$

where

$$\widehat{\boldsymbol{Q}}_{--}^\alpha = -\operatorname{diag}(\lambda^\alpha) \left[\boldsymbol{I}_{--} - \widehat{\boldsymbol{F}}_{-+}(\alpha, i\infty) \left(\boldsymbol{I}_{++} - \widehat{\boldsymbol{F}}_{++}(\alpha, i\infty) \right)^{-1} \widehat{\boldsymbol{P}}_{+-}^\alpha - \widehat{\boldsymbol{P}}_{--}^\alpha \right].$$

To this end, note that $\tau_-(x)$ is nondecreasing in x. Also, the *first-passage* process $\{J_{\tau_-(x)} : x \geq 0\}$ given $J_{\tau_-} = j$ is a under $\widehat{\mathbb{P}}^\alpha_\ell$ a (defective) Markov process with values in $\mathcal{I}_-$, cf. Assumption 14.1. For ease we first concentrate on the case for which the distributions of the $U^{j\ell}$ do not have an atom at zero. After an exponentially distributed time with parameter λ^α_j, the first-passage process then jumps to a $-$-point $k \in \mathcal{I}_-$ with probability $\widehat{P}^\alpha_{jk}$ (where $j = k$ is allowed).

For the general case where $U^{j\ell}$ may have an atom at zero, we have to take into account the paths in which S stays at the same level for a while before entering $k \in \mathcal{I}_-$. This procedure leads to the given intensity matrix. □

There are several ways to extract an algorithm for determining ${}^{\#}\widehat{\boldsymbol{P}}^\alpha_{+-}$, ${}^{\#}\widehat{\boldsymbol{P}}^\alpha_{--}$, and ${}^{\#}\widehat{\boldsymbol{Q}}^\alpha_{--}$ from Proposition 14.3. For instance, it is possible to find an initial matrix ${}^{\#}\widehat{\boldsymbol{Q}}^{\alpha,0}_{--}$ and a matrix-function $\boldsymbol{\varphi}$ such that the recursion ${}^{\#}\widehat{\boldsymbol{Q}}^{\alpha,n+1}_{--} = \boldsymbol{\varphi}({}^{\#}\widehat{\boldsymbol{Q}}^{\alpha,n}_{--})$ characterizes a sequence of matrices that converges to ${}^{\#}\widehat{\boldsymbol{Q}}^\alpha_{--}$. We do not give further details here, but we refer to Asmussen [18, Sec. VI.2] and Miyazawa [232] instead; we have already encountered a similar algorithm in Proposition 12.4.

One difficulty that needs to be overcome is the calculation of matrix exponentials, see [236] for a survey of available methods. *Uniformization* is a useful tool in this respect, as noted, for instance, in [16, 36, 221, 233]. It is not our aim to devise fast algorithms for computing the matrix ${}^{\#}\widehat{\boldsymbol{Q}}^\alpha_{--}$, and we shall therefore not address these algorithmic properties here.

Our next result is a nonlinear system for the matrix $\boldsymbol{K}^\alpha_{--}$, where

$$\boldsymbol{K}^\alpha_{--} := \operatorname{diag}(\lambda^\alpha)\, {}^{\#}\widehat{\boldsymbol{Q}}^\alpha_{--} \operatorname{diag}(1/\lambda^\alpha). \tag{14.3}$$

To state the system, we define for $\beta \in \mathbb{R}$,

$$\boldsymbol{F}_{+\circlearrowleft-}(\alpha,\beta) := (\boldsymbol{I}_{++} - \boldsymbol{F}_{++}(\alpha,\beta))^{-1}\, \boldsymbol{F}_{+-}(\alpha,\beta),$$

and $\boldsymbol{F}^\alpha_{+\circlearrowleft-}(dx)$ is the measure for which $\beta \mapsto \boldsymbol{F}_{+\circlearrowleft-}(\alpha,\beta)$ is the characteristic function. These notions relate to the increment in the 'vertical direction', when starting in a $+$-point, until the epoch that a $-$-point is reached.

Corollary 14.4 *For $\alpha \geq 0$, the matrix $\boldsymbol{K}^\alpha_{--}$ satisfies*

$$\boldsymbol{K}^\alpha_{--} + \operatorname{diag}(\lambda^\alpha)\left(\boldsymbol{I}_{--} - \operatorname{diag}(\mu^\alpha)\boldsymbol{P}^J_{--}\right) - \int_{[0,\infty)} e^{\boldsymbol{K}^\alpha_{--}x} \operatorname{diag}(\mu^\alpha\lambda^\alpha)\boldsymbol{P}^J_{-+}\boldsymbol{F}^\alpha_{+\circlearrowleft-}(dx) = \boldsymbol{0}_{--}.$$

Proof. The idea of the proof is to slightly modify the process without changing the (time-reversed) first-passage process (and thus $\boldsymbol{K}^\alpha_{--}$). Indeed, interpret a sequence of $+$-points as a single $+$-point; one then obtains a different discrete-time process, with $\boldsymbol{F}_{+-}(\alpha,\beta)$ replaced by $\boldsymbol{F}_{+\circlearrowleft-}(\alpha,\beta)$. Importantly, for this 'new' J we have that $\boldsymbol{P}^J_{++} = \boldsymbol{0}_{++}$, so that ${}^{\#}\widehat{\boldsymbol{P}}^\alpha_{-+} = \operatorname{diag}(\mu^\alpha)\boldsymbol{P}^J_{-+}$ by Proposition 14.3. The formula for ${}^{\#}\widehat{\boldsymbol{Q}}^\alpha_{--}$ in this proposition then immediately leads to the desired matrix equation for $\boldsymbol{K}^\alpha_{--}$. □

The next proposition characterizes the $\mathbb{P}_k$-distribution of $(S_{\tau_+}, T_{\tau_+}, J_{\tau_+})$. The main ingredient is the celebrated Wiener-Hopf factorization, which has already been used in Section 12.4 in the context of perturbed risk models. For the random-walk case, we refer to Section 11.1.1.

Proposition 14.5 *For $\alpha \geq 0, \beta \in \mathbb{R}$ such that $(\alpha,\beta) \neq 0$, we have*

$$\boldsymbol{E}\left[e^{-\alpha T_{\tau_+} + i\beta S_{\tau_+}}; J_{\tau_+}\right] = \boldsymbol{I} - \left(\boldsymbol{I} - {}^{\#}\widehat{\boldsymbol{E}}\left[e^{-\alpha T_{\tau_-} + i\beta S_{\tau_-}}; J_{\tau_-}\right]\right)^{-1} (\boldsymbol{I} - \boldsymbol{F}(\alpha,\beta)),$$

where nonsingularity is implicit.

Proof. Write $\widehat{\boldsymbol{G}}(\alpha,\beta) := \widehat{\boldsymbol{E}}[e^{-\alpha T_{\tau_-}+i\beta S_{\tau_-}}; J_{\tau_-}]$. The statement is the Wiener-Hopf factorization (e.g., [19, Thm. XI.2.12]) for the Markov-additive process S under the measure $\mathbb{P}^\alpha$, provided $\boldsymbol{I} - {}^{\#}\widehat{\boldsymbol{G}}$ is nonsingular. This requirement is equivalent to nonsingularity of $\boldsymbol{I} - \widehat{\boldsymbol{G}}$.

To see that this matrix is nonsingular, we exploit the fact that $\widehat{\boldsymbol{G}}_{jk}$ is the transform of a nonlattice distribution for $j \in \mathcal{I}, k \in \mathcal{I}_-$. Therefore, we have $|\widehat{\boldsymbol{G}}_{jk}(\alpha,\beta)| < \widehat{P}^0_{jk}$ for $(\alpha,\beta) \neq (0,0)$, see, e.g., Theorem 6.4.7 of Chung [75]. As a result, $\boldsymbol{I}-\widehat{\boldsymbol{G}}$ is a strictly diagonally dominant matrix:

$$\sum_{k\in\mathcal{I}} \left|\widehat{\boldsymbol{G}}_{jk}(\alpha,\beta)\right| < \sum_{k\in\mathcal{I}_+} \widehat{p}^J_{jk}\mathbb{P}(U^{kj}=0) + \sum_{k\in\mathcal{I}_-} \widehat{P}^0_{jk} \leq 1,$$

where the last inequality follows from the fact that S_{τ_-} has a (possibly defective) distribution, see Lemma 14.2. □

The distribution of $(S_{\tilde{\tau}_-}, T_{\tilde{\tau}_-}, J_{\tilde{\tau}_-})$

We now turn to our second aim of this subsection, the characterization of the distribution of $(S_{\tilde{\tau}_-}, T_{\tilde{\tau}_-}, J_{\tilde{\tau}_-})$. This turns out to be simpler than the analysis of $(S_{\tau_+}, T_{\tau_+}, J_{\tau_+})$; particularly, Wiener-Hopf techniques are not required here. As the distribution of $(S_{\tilde{\tau}_-}, T_{\tilde{\tau}_-}, J_{\tilde{\tau}_-})$ is readily derived from $(S_{\tilde{\tau}_-}, T_{\tilde{\tau}_-}, J_{\tilde{\tau}_-1})$, we focus on the latter vector instead. We omit all proofs, since similar arguments apply as before.

Write $\boldsymbol{P}^\alpha$ for $\boldsymbol{E}[e^{-\alpha T_{\tilde{\tau}_-}}; J_{\tilde{\tau}_-1}]$; the indices should be compared to those in the definition of $\widehat{\boldsymbol{P}}^\alpha$. The analogue of Lemma 14.2 immediately gives the desired transform: for $\alpha \geq 0$, $\beta \in \mathbb{R}$, we have

$$\boldsymbol{E}\left[e^{-\alpha T_{\tilde{\tau}_-}+i\beta S_{\tilde{\tau}_-}}; J_{\tilde{\tau}_-1}\right] = \begin{pmatrix} \boldsymbol{0}_{++} & \boldsymbol{P}^\alpha_{+-}\operatorname{diag}\left(\frac{\lambda^\alpha}{\lambda^\alpha+i\beta}\right) \\ \boldsymbol{0}_{-+} & \operatorname{diag}\left(\frac{\lambda^\alpha}{\lambda^\alpha+i\beta}\right) \end{pmatrix}.$$

Let us therefore continue by giving a result in the spirit of Proposition 14.3. We set

$$\boldsymbol{P}^\alpha_{+\circlearrowleft-} := (\boldsymbol{I}_{++} - \boldsymbol{F}_{++}(\alpha,i\infty))^{-1}\left[\boldsymbol{P}^\alpha_{+-}\operatorname{diag}(1/\mu^\alpha) + \boldsymbol{F}_{+-}(\alpha,i\infty)\right]$$

for notational convenience.

Proposition 14.6 *For $\alpha \geq 0$, we have*

$$\boldsymbol{P}^\alpha_{+-} = \int_{(0,\infty)} \boldsymbol{F}^\alpha_{++}(dx)\boldsymbol{P}^\alpha_{+-}\operatorname{diag}(1/\mu^\alpha)e^{\boldsymbol{Q}^\alpha_{--}x}\operatorname{diag}(\mu^\alpha) + \int_{(0,\infty)} \boldsymbol{F}^\alpha_{+-}(dx)e^{\boldsymbol{Q}^\alpha_{--}x}\operatorname{diag}(\mu^\alpha),$$

where $\boldsymbol{Q}^\alpha_{--}$ is specified by

$$\boldsymbol{Q}^\alpha_{--} = -\operatorname{diag}(\lambda^\alpha)\left[\boldsymbol{I}_{--} - \operatorname{diag}(\mu^\alpha)\boldsymbol{P}^J_{--} - \operatorname{diag}(\mu^\alpha)\boldsymbol{P}^J_{-+}\boldsymbol{P}^\alpha_{+\circlearrowleft-}\right].$$

We next turn to the analogue of Corollary 14.4, which can be proven along the same lines. When inspecting the differences between the two corollaries, we first note that they are remarkably similar. Whereas the $\boldsymbol{K}^\alpha_{--}$-matrices are always the *first* matrices in each of the terms, the $\boldsymbol{Q}^\alpha_{--}$-matrices always appear *last*. In Section 14.2.5, we show that this has a specific reason.

Corollary 14.7 *For $\alpha \geq 0$, the matrix $\boldsymbol{Q}^\alpha_{--}$ satisfies*

$$\boldsymbol{Q}^\alpha_{--} + \operatorname{diag}(\lambda^\alpha)\left[\boldsymbol{I}_{--} - \operatorname{diag}(\mu^\alpha)\boldsymbol{P}^J_{--} - \int_{[0,\infty)} \operatorname{diag}(\mu^\alpha)\boldsymbol{P}^J_{-+}\boldsymbol{F}^\alpha_{+\circlearrowleft-}(dx)e^{\boldsymbol{Q}^\alpha_{--}x}\right] = \boldsymbol{0}_{--}.$$

14.2.4 The distribution of $(\overline{S}, \overline{T}, \overline{J})$

In this section, we study $\overline{S}$ (jointly with $\overline{T}, \overline{J}$), assuming that S drifts to $-\infty$. Therefore, *throughout this subsection*, we suppose that $\boldsymbol{\pi}'_J \boldsymbol{E} S_1 < 0$. We remark that, with the only exception of Lemma 14.13, all the results also hold under the weaker assumption that S drifts to $-\infty$. Our main tools are the ladder-height results obtained in the previous subsection.

The next theorem completely characterizes the distribution of $(\overline{S}, \overline{T}, \overline{J})$. It is formulated as an expression for $(\boldsymbol{I} - \boldsymbol{F}(\alpha, i\beta))\, \boldsymbol{E}[e^{-\alpha\overline{T}-\beta\overline{S}}; \overline{J}]$ and not for $\boldsymbol{E}[e^{-\alpha\overline{T}-\beta\overline{S}}; \overline{J}]$, since the nonsingularity of the matrix $\boldsymbol{I} - \boldsymbol{F}(\alpha, i\beta)$ is a delicate issue for real β; this is addressed in Section 14.2.5. One way to avoid these problems is to work with characteristic functions, and we shall often do so *in the proofs*. Still, our results are given in terms of Laplace transforms, since this is customary in the literature.

We express the aforementioned matrix in terms of the matrix characterized in Lemma 14.2 and the (still unknown) vector $\boldsymbol{P}(\overline{S} = 0)$. Observe that the matrices ${}^{\#}\widehat{\boldsymbol{P}}^{\alpha}_{--}$ and ${}^{\#}\widehat{\boldsymbol{P}}^{\alpha}_{-+}$ required in Lemma 14.2 can be found with Proposition 14.3.

Theorem 14.8 *For $\alpha, \beta \geq 0$, we have*

$$(\boldsymbol{I} - \boldsymbol{F}(\alpha, i\beta))\, \boldsymbol{E}\left[e^{-\alpha\overline{T}-\beta\overline{S}}; \overline{J}\right] = \left(\boldsymbol{I} - {}^{\#}\widehat{\boldsymbol{E}}\left[e^{-\alpha T_{\tau_-} - \beta S_{\tau_-}}; J_{\tau_-}\right]\right) \operatorname{diag}\left(\boldsymbol{P}(\overline{S} = 0)\right).$$

Proof. By the Markov property, we have for $\alpha \geq 0$, $\beta \in \mathbb{R}$ such that $(\alpha, \beta) \neq (0, 0)$,

$$\begin{aligned} \boldsymbol{E}\left[e^{-\alpha\overline{T}+i\beta\overline{S}}; \overline{J}\right] &= \left(\boldsymbol{I} - \boldsymbol{E}\left[e^{-\alpha T_{\tau_+} + i\beta S_{\tau_+}}; J_{\tau_+}\right]\right)^{-1} \operatorname{diag}\left(\boldsymbol{P}(\tau_+ = \infty)\right) \\ &= (\boldsymbol{I} - \boldsymbol{F}(\alpha, \beta))^{-1} \left(\boldsymbol{I} - {}^{\#}\widehat{\boldsymbol{E}}\left[e^{-\alpha T_{\tau_-} + i\beta S_{\tau_-}}; J_{\tau_-}\right]\right) \operatorname{diag}\left(\boldsymbol{P}(\tau_+ = \infty)\right), \end{aligned}$$

where the second equality follows from Proposition 14.5. The nonsingularity of $\boldsymbol{I} - \boldsymbol{F}(\alpha, \beta)$ follows from (strict) diagonal dominance, cf. the proof of Proposition 14.5. This proves the claim after an analytic-continuation argument. □

There is a direct, insightful interpretation of Theorem 14.8 in terms of a last-passage process, which is used on several occasions in this chapter. This interpretation is inspired by Kennedy's interpretation [189] of the Wiener-Hopf factorization. First note that the theorem states that $\boldsymbol{E}[e^{-\alpha\overline{T}+i\beta\overline{S}}; \overline{J}]$ equals

$$\sum_{n=0}^{\infty} \boldsymbol{F}^n(\alpha, \beta) \operatorname{diag}(\boldsymbol{P}(\overline{S} = 0)) - \sum_{k=0}^{\infty} \boldsymbol{F}^k(\alpha, \beta)\, {}^{\#}\widehat{\boldsymbol{E}}\left[e^{-\alpha T_{\tau_-} + i\beta S_{\tau_-}}; J_{\tau_-}\right] \operatorname{diag}(\boldsymbol{P}(\overline{S} = 0)).$$

Clearly, the n-th summand in the first term can be interpreted as the transform of (S_n, T_n, J_n) on the event $\{\sup_{m \geq n} S_m = S_n\}$. If the maximum is attained at T_n, this is precisely $\boldsymbol{E}[e^{-\alpha\overline{T}+i\beta\overline{S}}; \overline{J}]$. However, if this is not the case, we have to subtract the contribution due to the fact that there is an $\ell < n$ for which $S_\ell \geq S_n$. In that case, write $S_n = S_k + (S_n - S_k)$, where $k = \sup\{\ell < n : S_\ell \geq S_n\}$, so that n is now a so-called *last-passage* epoch for the process with (k, S_k) as the origin. Looking *backward* in time, starting from (n, S_n), k is a first weak descending ladder epoch. The argument is completed by exploiting the Markov property. Partitioning with respect to the last-passage epoch is sometimes called the Beneš-method [38].

It is insightful to give the complete argument for $\alpha = 0$ in formulas. The terms that need to be subtracted (because the maximum occurred earlier) are

$$\sum_{n=0}^{\infty} \boldsymbol{E}\left[e^{i\beta S_n}; \forall m \geq n : S_m \leq S_n, \exists m < n : S_m \geq S_n, J_n\right]$$

$$
\begin{aligned}
&= \sum_{k=0}^{\infty}\sum_{n=k+1}^{\infty} \boldsymbol{E}\left[e^{i\beta S_k + i\beta(S_n - S_k)}; \sup_{m\geq n} S_m = S_n, S_k \geq S_n, \sup_{k<\ell<n} S_\ell < S_n, J_n\right] \\
&= \sum_{k=0}^{\infty} \boldsymbol{E}\left[e^{i\beta S_k}; J_k\right] {}^{\#}\widehat{\boldsymbol{E}}\left[e^{i\beta S_{\tau_-}}; J_{\tau_-}\right] \operatorname{diag}(\boldsymbol{P}(\overline{S}=0)),
\end{aligned}
$$

where the first equality is justified by the fact that the events are disjoint as a result of the partitioning with respect to the last-passage epoch.

Interestingly, Theorem 14.8 implies that, to compute $\boldsymbol{E}[e^{-\alpha\overline{T}-\beta\overline{S}}]$, only the determination of the vector $\boldsymbol{P}(\overline{S}=0)$ is left. Before giving results on $\boldsymbol{P}(\overline{S}=0)$, however, we first show that Theorem 14.8 has a number of interesting consequences.

Let us define for $\alpha, \beta \geq 0$,

$$
\begin{aligned}
\boldsymbol{D}_{--}(\alpha,\beta) \;:=\; & \beta\boldsymbol{I}_{--} - \operatorname{diag}(\lambda^\alpha)\big[\boldsymbol{I}_{--} - \operatorname{diag}(\mu^\alpha)\boldsymbol{P}^J_{--} \\
& - \operatorname{diag}(\mu^\alpha)\boldsymbol{P}^J_{-+}\left(\boldsymbol{I}_{++} - \boldsymbol{F}_{++}(\alpha,i\beta)\right)^{-1}\boldsymbol{F}_{+-}(\alpha,i\beta)\big].
\end{aligned}
$$

After some elementary linear algebra for block-matrix inverses, we arrive at the following corollary. It is instructive to derive this result with the above interpretation of Theorem 14.8: consider the discrete-time process only at $-$-points. The corresponding statements for $\boldsymbol{E}_+$ can derived from those for $\boldsymbol{E}_-$ and the Markov property.

Corollary 14.9 *We have for $\alpha, \beta \geq 0$,*

$$
\begin{aligned}
&\boldsymbol{D}_{--}(\alpha,\beta)\boldsymbol{E}_-\left[e^{-\alpha\overline{T}-\beta\overline{S}}; \overline{J}\in +\right] \\
&\quad = \operatorname{diag}\left(\lambda^\alpha\right)\Big[{}^{\#}\widehat{\boldsymbol{P}}^{\alpha}_{-+} - \operatorname{diag}\left(\mu^\alpha\right)\boldsymbol{P}^J_{-+} \\
&\qquad\qquad \left(\boldsymbol{I}_{++} - \boldsymbol{F}_{++}(\alpha,i\beta)\right)^{-1}\left(\boldsymbol{I}_{++} - \boldsymbol{F}_{++}(\alpha,i\infty)\right)\Big]\operatorname{diag}\left(\boldsymbol{P}_+(\overline{S}=0)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
&\boldsymbol{D}_{--}(\alpha,\beta)\boldsymbol{E}_-\left[e^{-\alpha\overline{T}-\beta\overline{S}}; \overline{J}\in -\right] \\
&\quad = \Big[\beta\boldsymbol{I}_{--} - \operatorname{diag}\left(\lambda^\alpha\right)\Big(\boldsymbol{I}_{--} - \operatorname{diag}\left(\mu^\alpha\right)\boldsymbol{P}^J_{-+} \\
&\qquad\qquad \left(\boldsymbol{I}_{++} - \boldsymbol{F}_{++}(\alpha,i\beta)\right)^{-1}\boldsymbol{F}_{+-}(\alpha,i\infty) - {}^{\#}\widehat{\boldsymbol{P}}^{\alpha}_{--}\Big)\Big]\operatorname{diag}\left(\boldsymbol{P}_-(\overline{S}=0)\right).
\end{aligned}
$$

If $\boldsymbol{P}^J_{++} = \boldsymbol{0}_{++}$, the second claim of this corollary can be reformulated in the following interesting form:

$$
\boldsymbol{D}_{--}(\alpha,\beta)\boldsymbol{E}_-\left[e^{-\alpha\overline{T}-\beta\overline{S}}; \overline{J}\in -\right] = \left[\beta\boldsymbol{I}_{--} + \boldsymbol{K}^{\alpha}_{--}\right]\operatorname{diag}\left(\boldsymbol{P}_-(\overline{S}=0)\right). \tag{14.4}
$$

Our next aim is to find $\boldsymbol{P}(\overline{S}=0)$. The following lemma gives two matrix equations that must be satisfied by $\boldsymbol{P}(\overline{S}=0)$.

Lemma 14.10 *$\boldsymbol{P}(\overline{S}=0)$ satisfies the system*

$$
\begin{aligned}
\boldsymbol{P}_+(\overline{S}=0) &= \boldsymbol{F}_{++}(0,i\infty)\boldsymbol{P}_+(\overline{S}=0) + \boldsymbol{F}_{+-}(0,i\infty)\boldsymbol{P}_-(\overline{S}=0), \\
\boldsymbol{P}_-(\overline{S}=0) &= {}^{\#}\widehat{\boldsymbol{P}}^0_{-+}\boldsymbol{P}_+(\overline{S}=0) + {}^{\#}\widehat{\boldsymbol{P}}^0_{--}\boldsymbol{P}_-(\overline{S}=0).
\end{aligned}
$$

Proof. Let us write $\boldsymbol{A}$ for the matrix $\boldsymbol{I} - {}^{\#}\widehat{\boldsymbol{E}}[e^{i\beta S_{\tau_-}}; J_{\tau_-}]$, thereby suppressing the dependence on β. With Theorem 14.8 and a standard formula for block matrix inverses, this yields for any $\beta \in \mathbb{R}$,

$$\boldsymbol{P}(\overline{S} = 0) = \begin{pmatrix} \boldsymbol{I}_{++} & -\boldsymbol{A}_{++}^{-1}\boldsymbol{A}_{+-} \\ -\boldsymbol{A}_{--}^{-1}\boldsymbol{A}_{-+} & \boldsymbol{I}_{--} \end{pmatrix} \begin{pmatrix} \boldsymbol{B}_{++}^{-1} & \boldsymbol{0}_{+-} \\ \boldsymbol{0}_{-+} & \boldsymbol{B}_{--}^{-1} \end{pmatrix} (\boldsymbol{I} - \boldsymbol{F}(0,\beta))\boldsymbol{E}e^{i\beta\overline{S}},$$

where

$$\boldsymbol{B}_{++}^{-1} := \boldsymbol{A}_{++} - \boldsymbol{A}_{+-}\boldsymbol{A}_{--}^{-1}\boldsymbol{A}_{-+}, \quad \boldsymbol{B}_{--}^{-1} := \boldsymbol{A}_{--} - \boldsymbol{A}_{-+}\boldsymbol{A}_{++}^{-1}\boldsymbol{A}_{+-}.$$

After some elementary matrix algebra, this equation immediately yields that

$$\boldsymbol{A}_{++}\boldsymbol{P}_+(\overline{S} = 0) + \boldsymbol{A}_{+-}\boldsymbol{P}_-(\overline{S} = 0) = \begin{pmatrix} \boldsymbol{I}_{++} & \boldsymbol{0}_{+-} \end{pmatrix} (\boldsymbol{I} - \boldsymbol{F}(0,\beta))\boldsymbol{E}e^{i\beta\overline{S}}.$$

The right-hand side vanishes as $\beta \to 0$, and the first claim follows. The same argument works for the second assertion by rewriting $\boldsymbol{A}_{-+}\boldsymbol{P}_+(\overline{S} = 0) + \boldsymbol{A}_{--}\boldsymbol{P}_-(\overline{S} = 0)$. □

The first equation of this lemma can be alternatively derived by considering $\boldsymbol{P}_+(\overline{S} = 0)$ and conditioning on the first step. The interpretation of the second equation is slightly more complicated, and follows from arguments reminiscent of the interpretation of Theorem 14.8. Again, the idea is to partition with respect to the last-passage epoch $\ell := \inf\{n : S_n = \sup_{m \geq n} S_m\}$, which is either a $+$-point or a $-$-point. On the event $\{\overline{S} = 0\}$, starting from (ℓ, S_ℓ) and looking backward in time, zero is a first descending ladder epoch. On the other hand, looking forward in time from (ℓ, S_ℓ), the process cannot have a strict ascending ladder epoch. This explains the formula. Obviously, ℓ fails to be a stopping time; we refer to Chapter 13 for related arguments.

We briefly pause our analysis of $\boldsymbol{P}(\overline{S} = 0)$ to record the following consequence of Corollary 14.9 and Lemma 14.10. It can be regarded as a Pollaczek-Khinchine formula for $\overline{S}$.

Corollary 14.11 *For $\beta \geq 0$, we have*

$$\boldsymbol{D}_{--}(0,\beta)\boldsymbol{E}_-e^{-\beta\overline{S}} = \beta\boldsymbol{P}_-(\overline{S} = 0).$$

We now investigate to what extend the system of equations in Lemma 14.10 determines $\boldsymbol{P}(\overline{S} = 0)$. First, since $\boldsymbol{I}_{++} - \boldsymbol{F}_{++}(0, i\infty)$ is always nonsingular by assumption, the first formula shows that it suffices to find $\boldsymbol{P}_-(\overline{S} = 0)$ instead of the larger vector $\boldsymbol{P}(\overline{S} = 0)$. Unfortunately, the whole system of equations in Lemma 14.10 is *always* singular. More precisely, the equations can be combined into

$$\left(\boldsymbol{I}_{--} - {}^{\#}\widehat{\boldsymbol{P}}^0_{-+}(\boldsymbol{I}_{++} - \boldsymbol{F}_{++}(0,i\infty))^{-1}\boldsymbol{F}_{+-}(0,i\infty) - {}^{\#}\widehat{\boldsymbol{P}}^0_{--}\right)\boldsymbol{P}_-(\overline{S} = 0) = \boldsymbol{0}_-,$$

or, equivalently, by (14.3) and Proposition 14.3,

$$\boldsymbol{K}^0_{--}\boldsymbol{P}_-(\overline{S} = 0) = \boldsymbol{0}_-. \tag{14.5}$$

The following proposition shows that this determines $\boldsymbol{P}_-(\overline{S} = 0)$ (and therefore $\boldsymbol{P}(\overline{S} = 0)$) up to a constant.

Proposition 14.12 *The matrix $\boldsymbol{K}^0_{--}$ has the following properties:*

(i) zero is a simple eigenvalue of $\boldsymbol{K}^0_{--}$, and the other $N_- - 1$ eigenvalues have strictly negative real parts, and

(ii) if $N_- > 1$, then $\mathrm{diag}(1/\lambda^0)\boldsymbol{\pi}_J(-)$ and $\boldsymbol{P}_-(\overline{S}=0)$ are left and right eigenvectors of $\boldsymbol{K}^0_{--}$ respectively, corresponding to the eigenvalue zero.

Proof. For the first property, it suffices to consider the matrix $\widehat{\boldsymbol{Q}}^0_{--}$, which is similar to $\boldsymbol{K}^0_{--}$. The matrix $\widehat{\boldsymbol{Q}}^0_{--}$ inherits its irreducibility from $\boldsymbol{P}^J$, and since it is an intensity matrix of a (nondefective) Markov process, the assertion follows from standard Perron-Frobenius theory.

The 'right eigenvector' part of the second claim follows from (14.5), and the 'left eigenvector' part translates to $\widehat{\boldsymbol{Q}}^0_{--}\mathbf{1}_- = \mathbf{0}_-$. □

Proposition 14.12 shows that one more equation is needed to fully specify $\boldsymbol{P}_-(\overline{S}=0)$, and this equation is given in the following lemma. Let $\boldsymbol{\pi}_-$ be the unique $\mathcal{I}_-$-probability vector satisfying

$$\boldsymbol{\pi}'_- \,\mathrm{diag}(\lambda^0)\left(\boldsymbol{P}^J_{--} + \boldsymbol{P}^J_{-+}\left(\boldsymbol{I}_{++} - \boldsymbol{P}^J_{++}\right)^{-1}\boldsymbol{P}^J_{+-}\right) = \boldsymbol{\pi}'_- \,\mathrm{diag}(\lambda^0); \tag{14.6}$$

in fact, $\boldsymbol{\pi}_-$ is proportional to $\mathrm{diag}(1/\lambda^0)\boldsymbol{\pi}_J(-)$.

Lemma 14.13 *We have*

$$\boldsymbol{\pi}'_-\boldsymbol{P}_-(\overline{S}=0) = 1 - \boldsymbol{\pi}'_- \,\mathrm{diag}(\lambda^0)\boldsymbol{P}^J_{-+}\left(\boldsymbol{I}_{++} - \boldsymbol{P}^J_{++}\right)^{-1}\boldsymbol{E}_+S_1.$$

This equation is independent of the $N_- - 1$ independent linear equations stemming from (14.5).

Proof. The idea is to premultiply the expression for $\boldsymbol{P}_-(\overline{S}=0)$ in Corollary 14.11 by $\boldsymbol{\pi}'_-$, to divide both sides by β, and then let $\beta \to 0$. By definition of $\boldsymbol{\pi}_-$, this immediately yields that $\boldsymbol{\pi}'_-\boldsymbol{P}_-(\overline{S}>0)$ equals

$$\lim_{\beta\to 0}\frac{1}{\beta}\boldsymbol{\pi}'_- \,\mathrm{diag}\left(\lambda^0\right)\boldsymbol{P}^J_{-+}\left[\left(\boldsymbol{I}_{++} - \boldsymbol{P}^J_{++}\right)^{-1}\boldsymbol{P}^J_{+-} - \left(\boldsymbol{I}_{++} - \boldsymbol{F}_{++}(0,i\beta)\right)^{-1}\boldsymbol{F}_{+-}(0,i\beta)\right]\boldsymbol{E}_-e^{-\beta\overline{S}}.$$

It is not hard to see that this equals $\boldsymbol{\pi}'_- \,\mathrm{diag}\left(\lambda^0\right)\boldsymbol{P}^J_{-+}\boldsymbol{E}_+S_{\gamma_-}$, where $\gamma_- := \inf\{n \geq 1 : S_n \in \mathcal{I}_-\}$. To compute $\boldsymbol{E}_+S_{\gamma_-}$, we condition on the first step to see that

$$\boldsymbol{E}_+S_{\gamma_-} = \boldsymbol{E}_+S_1 + \boldsymbol{P}^J_{++}\boldsymbol{E}_+S_{\gamma_-},$$

and the claim follows.

The independence of the other $N_- - 1$ equations follows from the fact that

$$\boldsymbol{\pi}'_- \,\mathrm{diag}(\lambda^0)\boldsymbol{P}^J_{-+}\left(\boldsymbol{I}_{++} - \boldsymbol{P}^J_{++}\right)^{-1}\boldsymbol{E}_+S_1 < 1,$$

due to the stability constraint $\boldsymbol{\pi}'_J\boldsymbol{E}S_1 < 0$. □

Let us summarize the results of this section by providing a 'recipe' how the joint transform $\boldsymbol{E}[e^{-\alpha\overline{T}-\beta\overline{S}};\overline{J}]$ can be found.

- To obtain $\boldsymbol{P}(\overline{S}=0)$:
  - Calculate ${}^{\#}\widehat{\boldsymbol{P}}^0_{+-}$ and ${}^{\#}\widehat{\boldsymbol{P}}^0_{--}$ through the fixed point of Proposition 14.3.
  - Compute $\boldsymbol{K}^0_{--}$, and find the unique $\mathcal{I}_-$-probability vector $\boldsymbol{h}_-$ satisfying $\boldsymbol{K}^0_{--}\boldsymbol{h}_- = \mathbf{0}_-$.
  - Compute $\boldsymbol{\pi}_-$ with (14.6), and set

$$\kappa := \frac{1 - \boldsymbol{\pi}'_- \,\mathrm{diag}(\lambda^0)\boldsymbol{P}^J_{-+}\left(\boldsymbol{I}_{++} - \boldsymbol{P}^J_{++}\right)^{-1}\boldsymbol{E}_+S_1}{\boldsymbol{\pi}'_-\boldsymbol{h}_-}.$$

– Set $\boldsymbol{P}_-(\overline{S}=0)=\kappa\boldsymbol{h}_-$ and

$$\boldsymbol{P}_+(\overline{S}=0)=\kappa(\boldsymbol{I}_{++}-\boldsymbol{F}_{++}(0,i\infty))^{-1}\boldsymbol{F}_{+-}(0,i\infty)\boldsymbol{h}_-.$$

- To obtain $\boldsymbol{E}[e^{-\alpha\overline{T}-\beta\overline{S}};\overline{J}]$ for fixed $\alpha\geq 0$:
  - Calculate ${}^{\#}\widehat{\boldsymbol{P}}^{\alpha}_{+-}$ and ${}^{\#}\widehat{\boldsymbol{P}}^{\alpha}_{--}$ through the fixed point of Proposition 14.3.
  - Determine the transform with Corollary 14.9 for every value of $\beta\geq 0$ for which the transform is needed.

14.2.5 The spectral method for the distribution of $(\overline{S},\overline{T},\overline{J})$

In Section 14.2.4, we have used Wiener-Hopf theory and nonlinear matrix equations to study the distribution of $(\overline{S},\overline{T},\overline{J})$. Our probabilistic reasoning culminated in a number of results, including the Pollaczek-Khinchine formula for $\overline{S}$ in Corollary 14.11. However, our approach is not the only way to prove this statement; it is usually possible to take alternative approaches.

Other methods to obtain the Pollaczek-Khinchine formula are (Kolmogorov) differential equations [297], martingales [24], rate-conservation laws [232], or conditioning on the first step [79, Sec. II.4.5]. Importantly, these methods express the Laplace transform in terms of an *unknown* vector, which is the equivalent of $\boldsymbol{P}_-(\overline{S}=0)$ from the previous subsection. For a discrete-state model, Gail *et al.* [138] show how this vector can be found. Cast into the present setting, they show that $\operatorname{adj}\boldsymbol{D}_{--}(0,\beta)\boldsymbol{P}_-(\overline{S}=0)$ must vanish to the order at least r at $\beta=\nu$ if $\nu\neq 0$ is a singularity of $\boldsymbol{D}_{--}(0,\nu)$ with algebraic multiplicity r. Here $\operatorname{adj}\boldsymbol{D}_{--}(0,\beta)$ denotes the adjoint matrix of $\boldsymbol{D}_{--}(0,\beta)$, i.e., the transpose of the matrix formed by taking the cofactor of each element of $\boldsymbol{D}_{--}(0,\beta)$.

Alternatively, one can use a technique that we call the *spectral method*. It is the goal of this approach to find $N_- - 1$ linear independent vectors $\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_{N_--1}$ such that $\boldsymbol{\ell}_j'\boldsymbol{P}_-(\overline{S}=0)=0$. As in Section 14.2.4, the remaining equation is given in Lemma 14.13 (note that its proof only relies on Corollary 14.11). To determine the $\boldsymbol{\ell}_j$, one determines a root ν_j of the equation $\det\boldsymbol{D}_{--}(0,\beta)=0$, and identifies the $\boldsymbol{\ell}_j$ with a left eigenvector of $\boldsymbol{D}_{--}(0,\nu_j)$ corresponding to the eigenvalue zero.

The main question is whether the spectral approach is feasible, i.e., whether *enough* eigenvectors $\boldsymbol{\ell}_j$ can be found. The spectral method is then an alternative for the 'matrix-analytic' approach taken in the previous subsection. As a first step to answering this question, we investigate the relationship between the matrices $\boldsymbol{D}_{--}(\alpha,\cdot)$, $\boldsymbol{Q}^{\alpha}_{--}$, and $\boldsymbol{K}^{\alpha}_{--}$. We omit a proof of the next proposition, since we prove a more general statement in Theorem 14.15 below.

To gain some intuition for the result, note that the identity in (14.4) immediately shows the equivalence of (ii) and (iii) if $\boldsymbol{P}^{J}_{++}=\boldsymbol{0}_{++}$.

Proposition 14.14 *Suppose that* $\boldsymbol{\pi}_J'\boldsymbol{E}S_1<0$.

For any ν *with* $\Re(\nu)\geq 0$, *the following are equivalent:*

(i) $-\nu$ *is an eigenvalue of* $\boldsymbol{Q}^{\alpha}_{--}$,

(ii) $-\nu$ *is an eigenvalue of* $\boldsymbol{K}^{\alpha}_{--}$, *and*

(iii) zero is an eigenvalue of $\boldsymbol{D}_{--}(\alpha,\nu)$.

Moreover, the geometric multiplicities of these eigenvalues coincide.

This proposition shows why the recursions in Corollaries 14.4 and 14.7 are necessarily matrix versions of the equation $\boldsymbol{D}_{--}(\alpha,\beta)=\boldsymbol{0}_{--}$. Indeed, suppose that $(-\nu,\boldsymbol{\ell})$ is a left eigenpair for $\boldsymbol{K}^{\alpha}_{--}$, so that $\boldsymbol{\ell}'\boldsymbol{K}^{\alpha}_{--}=-\nu\boldsymbol{\ell}'$. Since then $\boldsymbol{\ell}'e^{\boldsymbol{K}^{\alpha}_{--}x}=e^{-\nu x}\boldsymbol{\ell}'$, it follows from the recursion

for $\boldsymbol{K}^{\alpha}_{--}$ in Corollary 14.4 that $\boldsymbol{\ell}'\boldsymbol{D}_{--}(\alpha,\nu) = \boldsymbol{0}'_-$. The same reasoning goes through for the recursion in Corollary 14.7, but one then has to work with the right eigenpair.

Proposition 14.14 has interesting consequences for the location of singularities of $\boldsymbol{D}_{--}(\alpha,\beta)$ in the right complex halfplane (that is, the values of β for which this matrix is singular). First, since $\boldsymbol{Q}^{\alpha}_{--}$ is a real matrix, these singularities come in conjugate pairs. Moreover, as a result of Proposition 14.12, zero is a simple singularity and the real parts of the other singularities are strictly positive. In fact, all nonzero singularities must be in the open disc with radius and center $\max_j \lambda_j^{\alpha}$. For $\alpha = 0$, this claim has recently been proven with different methods by Tzenova *et al.* [297]. In [297], it is also shown that $\beta \mapsto \det \boldsymbol{D}_{--}(0,\beta)$ has exactly N_- zeroes (counting multiplicities).

Let us now explain how Proposition 14.14 relates to our aim to find vectors that are orthogonal to $\boldsymbol{P}_-(\overline{S} = 0)$. It follows from Proposition 14.14 that $\boldsymbol{K}^{0}_{--}$ (or $\boldsymbol{Q}^{0}_{--}$) is diagonalizable if and only if there exist $N_- - 1$ pairs $(-\nu_j, \boldsymbol{\ell}_j)$ with $\nu_j \neq 0$ such that the $\boldsymbol{\ell}_j$ are linearly independent and $\boldsymbol{\ell}_j'\boldsymbol{D}_{--}(0,\nu_j) = \boldsymbol{0}'_-$. By Corollary 14.11, we then have $\boldsymbol{\ell}_j'\boldsymbol{P}_-(\overline{S} = 0) = 0$. The remaining equation can be determined with Lemma 14.13. Therefore, the spectral method can be successfully applied in the diagonalizable case and several relatively explicit results can be derived, cf. Kella [184]. In fact, the spectral method not only yields $\boldsymbol{P}_-(\overline{S} = 0)$, but it is also possible to reconstruct $\boldsymbol{K}^{0}_{--}$; see for instance Section 5 of Asmussen [16].

However, if $\boldsymbol{K}^{0}_{--}$ is *not* diagonalizable, Proposition 14.14 shows that it is *impossible* to find enough pairs $(-\nu_j, \boldsymbol{\ell}_j)$ with the above properties. This raises the questions whether the spectral method is still feasible, and whether it yields $\boldsymbol{K}^{0}_{--}$ (or $\boldsymbol{Q}^{0}_{--}$) as a by-product. We address these questions in the remainder of this subsection for *general* but *fixed* $\alpha \geq 0$.

In fact, we show that it is *always* possible to construct $\boldsymbol{K}^{\alpha}_{--}$ and $\boldsymbol{Q}^{\alpha}_{--}$ with the spectral method. For $\alpha = 0$, the procedure also gives exactly $N_- - 1$ vectors orthogonal to $\boldsymbol{P}_-(\overline{S} = 0)$. Based on Proposition 14.14, one might guess that these vectors are the generalized left eigenvectors of $\boldsymbol{D}_{--}(\alpha,\nu)$ if ν is a singularity of $\boldsymbol{D}_{--}(\alpha,\nu)$. Interestingly, it turns out that this is not the case.

It is most insightful to present the procedure in an algorithmic form:

- Locate the singularities of $\boldsymbol{D}_{--}(\alpha,\beta)$ in the right complex halfplane (if $\alpha = 0$, then $\beta = 0$ is such a singularity).
- For every nonzero singularity ν, find as many independent vectors $\boldsymbol{\ell}$ with $\boldsymbol{\ell}'\boldsymbol{D}_{--}(\alpha,\nu) = \boldsymbol{0}'_-$ as possible (if $\alpha = 0$, then $\boldsymbol{\pi}_-$ is such a vector for $\nu = 0$, see (14.6)).
- This results in s pairs $(-\nu_j, \boldsymbol{\ell}_j)$, for some $s \leq N_-$, $j = 1,\ldots,s$ (the ν_j need not be distinct). If $s = N_-$, then stop; $\boldsymbol{K}^{\alpha}_{--}$ is diagonalizable.
- $\boldsymbol{K}^{\alpha}_{--}$ is not diagonalizable. If $\alpha > 0$, execute the following subroutine for each $j = 1,\ldots,s$. If $\alpha = 0$, set $d_s = 1$ and $\boldsymbol{\ell}_s^{(1)} = \boldsymbol{\pi}_-$, and execute the following subroutine for each $j = 1,\ldots,s-1$:
  - Set $p := 1$ and write $\boldsymbol{\ell}_j^{(1)} := \boldsymbol{\ell}_j$.
  - If possible, find a vector $\boldsymbol{\ell}$, independent of $\boldsymbol{\ell}_j^{(1)},\ldots,\boldsymbol{\ell}_j^{(p)}$, such that

$$\boldsymbol{\ell}'\boldsymbol{D}_{--}(\alpha,\nu_j) = \boldsymbol{\ell}_j^{(p)'} - \sum_{q=1}^{p} \int_{[0,\infty)} \frac{x^q}{q!} e^{-\nu_j x} \boldsymbol{\ell}_j^{(p-q+1)'} \operatorname{diag}(\mu^{\alpha}\lambda^{\alpha}) \boldsymbol{P}^{J}_{-+} \boldsymbol{F}^{\alpha}_{+\circlearrowleft -}(dx).$$

  - If the previous step was successful, set $\boldsymbol{\ell}_j^{(p+1)} := \boldsymbol{\ell}$, $p = p+1$, and repeat the previous step. If it was unsuccessful, set $d_j := p$ and stop the subroutine.

The following theorem shows that this algorithm yields both $\boldsymbol{K}^{\alpha}_{--}$ for $\alpha \geq 0$ and $\boldsymbol{P}_{-}(\overline{S}=0)$. The matrix $\boldsymbol{Q}^{\alpha}_{--}$ can be found in a similar fashion, but the reasoning in the proof must then be applied to the time-reversed process (i.e., the process (S,T,J) under $\widehat{\mathbb{P}}$). For notational convenience, we only write down the nonzero elements of the matrices. Note that the $\boldsymbol{J}_j$-matrices are Jordan blocks.

Theorem 14.15 *For $\alpha \geq 0$, the matrix $\boldsymbol{K}^{\alpha}_{--}$ is constructed as follows:*

$$\boldsymbol{K}^{\alpha}_{--} = \begin{pmatrix} \boldsymbol{L}_1 \\ \vdots \\ \boldsymbol{L}_s \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{J}_1 & & \\ & \ddots & \\ & & \boldsymbol{J}_s \end{pmatrix} \begin{pmatrix} \boldsymbol{L}_1 \\ \vdots \\ \boldsymbol{L}_s \end{pmatrix}, \tag{14.7}$$

where the $(d_j \times d_j)$-matrices $\boldsymbol{J}_j$ and $(d_j \times N_-)$-matrices $\boldsymbol{L}_j$ are defined as

$$\boldsymbol{J}_j := \begin{pmatrix} -\nu_j & & & \\ 1 & -\nu_j & & \\ & \ddots & \ddots & \\ & & 1 & -\nu_j \end{pmatrix}, \qquad \boldsymbol{L}_j = \begin{pmatrix} \boldsymbol{\ell}_j^{(1)'} \\ \vdots \\ \boldsymbol{\ell}_j^{(d_j)'} \end{pmatrix}.$$

Moreover, if $\alpha = 0$, then the rows of $\boldsymbol{L}_1, \ldots, \boldsymbol{L}_{s-1}$ constitute exactly $N_- - 1$ independent vectors that are orthogonal to $\boldsymbol{P}_{-}(\overline{S}=0)$.

Proof. The first step is to 'regroup' the +-points as in the proof of Corollary 14.4. That is, any sequence of +-points is replaced by a single +-point; the matrix $\boldsymbol{F}_{+-}(\alpha,\beta)$ then needs to be replaced by $\boldsymbol{F}_{+\circlearrowright -}(\alpha,\beta)$, and we have $\boldsymbol{P}^{J}_{++} = \boldsymbol{0}_{++}$ for the new process.

Denote the first (strict) ascending ladder epoch of this process by $\bar{\tau}_+$. Equation (14.4), which applies due to $\boldsymbol{P}^{J}_{++} = \boldsymbol{0}_{++}$, now factorizes $\boldsymbol{D}_{--}(0,\beta)$ into two matrices:

$$\boldsymbol{D}_{--}(\alpha,\beta) = \left[\beta \boldsymbol{I}_{--} + \boldsymbol{K}^{\alpha}_{--}\right] \left(\boldsymbol{I}_{--} - \boldsymbol{E}_{-}\left[e^{-\alpha T_{\bar{\tau}_+} - \beta S_{\bar{\tau}_+}}; J_{\bar{\tau}_+} \in -\right]\right). \tag{14.8}$$

This equation can be regarded as a *factorization identity*. Indeed, the first matrix in this identity has singularities in the right complex halfplane, and the second matrix in the left complex halfplane. For similar factorizations in a discrete-state framework, we refer to Zhao *et al.* [313]. The second matrix is abbreviated as $\boldsymbol{M}(\alpha,\beta)$ for convenience.

To prove the theorem, write $\boldsymbol{K}^{\alpha}_{--}$ in the Jordan form $\boldsymbol{L}^{-1}_{--}\boldsymbol{J}_{--}\boldsymbol{L}_{--}$, cf. (14.7). If $\alpha = 0$, we know that zero is a simple eigenvalue and that its corresponding left eigenvector is $\boldsymbol{\pi}_-$, cf. Proposition 14.12. The above factorization identity shows that

$$\operatorname{adj}\,(\beta \boldsymbol{I}_{--} + \boldsymbol{J}_{--})\,\boldsymbol{L}_{--}\boldsymbol{D}_{--}(\alpha,\beta) = \det\,(\beta \boldsymbol{I}_{--} + \boldsymbol{J}_{--})\,\boldsymbol{L}_{--}\boldsymbol{M}(\alpha,\beta). \tag{14.9}$$

Now observe that $\beta \boldsymbol{I}_{--} + \boldsymbol{J}_{--}$ is a block-diagonal matrix, and that for (square) block matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of arbitrary size,

$$\operatorname{adj} \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{pmatrix} = \begin{pmatrix} \det \boldsymbol{B} \operatorname{adj} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \det \boldsymbol{A} \operatorname{adj} \boldsymbol{B} \end{pmatrix}.$$

This shows that (14.9) is equivalent to the s systems

$$\operatorname{adj}\,(\beta \boldsymbol{I}_{d_j d_j} + \boldsymbol{J}_j)\boldsymbol{L}_j\boldsymbol{D}_{--}(\alpha,\beta) = (\beta - \nu_j)^{d_j}\boldsymbol{L}_j\boldsymbol{M}(\alpha,\beta). \tag{14.10}$$

If $\alpha = 0$, the equation for $j = s$ plays no role and is redundant. In the rest of the proof, we consider this system for fixed j and drop the subscripts j from the notation.

It remains to show that our algorithm constructs the matrix $\boldsymbol{L}$ ($\equiv \boldsymbol{L}_j$). First observe that (14.10) is equivalent to the d equations

$$\sum_{n=1}^{k} \frac{(-1)^{n-1}}{(\beta-\nu)^{k-n}} \boldsymbol{\ell}^{(n)'} \boldsymbol{D}_{--}(\alpha,\beta) = (\beta-\nu)\boldsymbol{\ell}^{(k)'} \boldsymbol{M}(\alpha,\beta), \tag{14.11}$$

for $k = 1, \ldots, d$. If we consider this equation for $k = 1$ and let $\beta \to \nu$, then it becomes clear that $\boldsymbol{\ell}^{(1)'} \boldsymbol{D}_{--}(\alpha,\nu) = \boldsymbol{0}_-$. Using this fact in the same equation for $k = 2$, we obtain

$$\boldsymbol{\ell}^{(2)'} \boldsymbol{D}_{--}(\alpha,\beta) - \frac{1}{\beta-\nu} \boldsymbol{\ell}^{(1)'} \left[\boldsymbol{D}_{--}(\alpha,\beta) - \boldsymbol{D}_{--}(\alpha,\nu)\right] = (\beta-\nu)\boldsymbol{\ell}^{(2)'} \boldsymbol{M}(\alpha,\beta).$$

Upon letting $\beta \to \nu$, we see (with dominated convergence and $\Re(\nu) > 0$) that $\boldsymbol{\ell}^{(2)'} \boldsymbol{D}_{--}(\alpha,\nu) = \boldsymbol{\ell}^{(1)'} - \boldsymbol{\ell}^{(1)'} \boldsymbol{D}^{(1)}_{--}(\alpha,\nu)$, where

$$\boldsymbol{D}^{(q)}_{--}(\alpha,\nu) := \int_{[0,\infty)} \frac{x^q}{q!} e^{-\nu x} \operatorname{diag}(\mu^\alpha \lambda^\alpha) \boldsymbol{P}^J_{-+} \boldsymbol{F}^\alpha_{+\circlearrowleft -}(dx).$$

Conversely, if there exists no vector $\boldsymbol{\ell}^{(2)}$ independent of $\boldsymbol{\ell}^{(1)}$ such that $\boldsymbol{\ell}^{(2)'} \boldsymbol{D}_{--}(\alpha,\nu) = \boldsymbol{\ell}^{(1)'} - \boldsymbol{\ell}^{(1)'} \boldsymbol{D}^{(1)}_{--}(\alpha,\nu)$, then (14.11) cannot hold. All essential ideas of the proof have now been used, and it is completed with an induction argument.

Suppose that we know that for $n = 2, \ldots, k$,

$$\boldsymbol{\ell}^{(n)'} \boldsymbol{D}_{--}(\alpha,\nu) = \boldsymbol{\ell}^{(n-1)'} - \sum_{q=1}^{n-1} \boldsymbol{\ell}^{(q)'} \boldsymbol{D}^{(n-q)}_{--}(\alpha,\nu). \tag{14.12}$$

Note that we have just proven this assertion for $n = 2$, and we now show that it also holds for $n = k+1$. For this, first multiply the $k-1$ equations in (14.12) by $(-1)^{n-1}(\beta-\nu)^{n-k-1}$, and substitute them in Equation (14.11) for $n = k+1$ such that terms $\boldsymbol{D}_{--}(\alpha,\beta) - \boldsymbol{D}_{--}(\alpha,\nu)$ appear everywhere; also use $\boldsymbol{\ell}^{(1)'} \boldsymbol{D}_{--}(\alpha,\nu) = \boldsymbol{0}'_-$. After some algebra, one then obtains

$$\begin{aligned}(\beta-\nu)\boldsymbol{\ell}^{(k+1)'} \boldsymbol{M}(\alpha,\beta) = (-1)^k \boldsymbol{\ell}^{(k+1)'} \boldsymbol{D}_{--}(\alpha,\beta) + \frac{(-1)^{k-1}}{\beta-\nu} \boldsymbol{\ell}^{(k)'} \left[\boldsymbol{D}_{--}(\alpha,\beta) - \boldsymbol{D}_{--}(\alpha,\nu)\right] \\ + \sum_{n=1}^{k-1} \frac{(-1)^{n-1}}{(\beta-\nu)^{k-n+1}} \boldsymbol{\ell}^{(n)'} \Bigg[\boldsymbol{D}_{--}(\alpha,\beta) - \boldsymbol{D}_{--}(\alpha,\nu) - (\beta-\nu)\boldsymbol{I}_{--} \\ - \sum_{q=1}^{k-n} (-(\beta-\nu))^q \boldsymbol{D}^{(q)}_{--}(\alpha,\nu)\Bigg].\end{aligned}$$

Upon letting $\beta \to \nu$, this leads to (14.12) for $n = k+1$. □

Two elements of the preceding proof deserve special attention. First, we emphasize the appealing form of the factorization (14.8); we encounter similar forms in the remainder. Another interesting point is the connection between the system (14.12) and Corollary 14.4, which is used in Section 14.3. To explain this, we rewrite this system and the equation $\boldsymbol{\ell}_j^{(1)'} \boldsymbol{D}_{--}(\alpha,\nu_j) = \boldsymbol{0}'_-$ as

$$\begin{aligned}\boldsymbol{0}_{d_j -} &= -\boldsymbol{J}_j \boldsymbol{L}_j - \boldsymbol{L}_j \operatorname{diag}(\lambda^\alpha) + \boldsymbol{L}_j \operatorname{diag}(\mu^\alpha \lambda^\alpha) \boldsymbol{P}^J_{--} \\ &\quad + \sum_{k=0}^{d_j-1} e^{-\nu_j x} \frac{x^k}{k!} \left(\nu_j \boldsymbol{I}_{d_j d_j} + \boldsymbol{J}_j\right)^k \boldsymbol{L}_j \operatorname{diag}(\mu^\alpha \lambda^\alpha) \boldsymbol{P}^J_{-+} \boldsymbol{F}^\alpha_{+-}(dx),\end{aligned} \tag{14.13}$$

for $j = 1, \ldots, s$.

In the proof of the preceding theorem, we showed that there is some s such that (14.13) holds for a unique d_j and unique matrices $\boldsymbol{J}_j$ and $\boldsymbol{L}_j$ (with independent rows; uniqueness holds up to multiplication by a constant). We now argue that a solution to (14.13) immediately gives a solution to the equation in Corollary 14.4. To see this, stack the s matrix equations of (14.13) into a single system, premultiply by $\boldsymbol{L}_{--}^{-1}$, note that

$$\sum_{k=0}^{d_j-1} e^{-\nu_j x} \frac{x^k}{k!} \left(\nu_j \boldsymbol{I}_{d_j d_j} + \boldsymbol{J}_j\right)^k = e^{\boldsymbol{J}_j x},$$

and use (14.7). The argument can also be reversed: given a solution to the equation in Corollary 14.4, the 'building blocks' for the Jordan form must solve (14.13). This proves the following.

Corollary 14.16 *The matrix equations in Corollaries 14.4 and 14.7 have a unique solution.*

14.2.6 The distribution of $(\underline{S}, \underline{T}, \underline{J})$

In this subsection, we study the minimum of S if it drifts to $+\infty$. Therefore, *throughout this subsection*, we suppose that $\boldsymbol{\pi}_J' \boldsymbol{E} S_1 > 0$. We remark that the matrix $\beta \boldsymbol{I}_{--} - \boldsymbol{Q}_{--}^{\alpha}$ is always nonsingular for $\beta \geq 0$, since $\boldsymbol{Q}_{--}^{\alpha}$ is a defective intensity matrix.

Theorem 14.17 *For $\alpha, \beta \geq 0$, we have $\underline{J} \in \mathcal{I}_+$ and*

$$\begin{aligned} \boldsymbol{E}\left[e^{-\alpha \underline{T} + \beta \underline{S}}; \underline{J} \in +\right] &= \left[\begin{pmatrix} \boldsymbol{I}_{++} \\ \boldsymbol{0}_{-+} \end{pmatrix} + \begin{pmatrix} \boldsymbol{P}_{+\circlearrowleft -}^{\alpha} \\ \boldsymbol{I}_{--} \end{pmatrix} (\beta \boldsymbol{I}_{--} - \boldsymbol{Q}_{--}^{\alpha})^{-1} \operatorname{diag}\left(\mu^{\alpha} \lambda^{\alpha}\right) \boldsymbol{P}_{-+}^{J} \right] \\ &\quad \times \operatorname{diag}\left(\boldsymbol{1}_+ - \boldsymbol{P}_{+\circlearrowleft -}^{0} \boldsymbol{1}_-\right). \end{aligned}$$

In particular, for $j \in \mathcal{I}$ and $k \in \mathcal{I}_+$, we have the matrix-exponential form

$$\mathbb{P}_j(\underline{S} < x; \underline{J} = k) = \left(1 - \boldsymbol{e}_k' \boldsymbol{P}_{+\circlearrowleft -}^{0} \boldsymbol{1}_-\right) \boldsymbol{e}_j' \begin{pmatrix} \boldsymbol{P}_{+\circlearrowleft -}^{0} \\ \boldsymbol{I}_{--} \end{pmatrix} e^{-\boldsymbol{Q}_{--} x} \operatorname{diag}(\lambda^0) \boldsymbol{P}_{-+}^{J} \boldsymbol{e}_k,$$

where $x \leq 0$.

Proof. The Markov property shows that for $\alpha, \beta \geq 0$,

$$\boldsymbol{E}_+\left[e^{-\alpha \underline{T} + \beta \underline{S}}; \underline{J} \in +\right] = \boldsymbol{P}_{+\circlearrowleft -}^{\alpha} \boldsymbol{E}_-\left[e^{-\alpha \underline{T} + \beta \underline{S}}; \underline{J} \in +\right] + \operatorname{diag}(\boldsymbol{P}_+(\underline{S} = 0))$$

and

$$\begin{aligned} \boldsymbol{E}_-\left[e^{-\alpha \underline{T} + \beta \underline{S}}; \underline{J} \in +\right] &= \operatorname{diag}\left(\frac{\mu^{\alpha} \lambda^{\alpha}}{\lambda^{\alpha} + \beta}\right) \boldsymbol{P}_{-+}^{J} \boldsymbol{E}_+\left[e^{-\alpha \underline{T} + \beta \underline{S}}; \underline{J} \in +\right] \\ &\quad + \operatorname{diag}\left(\frac{\mu^{\alpha} \lambda^{\alpha}}{\lambda^{\alpha} + \beta}\right) \boldsymbol{P}_{--}^{J} \boldsymbol{E}_-\left[e^{-\alpha \underline{T} + \beta \underline{S}}; \underline{J} \in +\right]. \end{aligned}$$

Substitution of the first equation in the second yields, with the expression for $\boldsymbol{Q}_{--}^{\alpha}$ in Proposition 14.6,

$$\boldsymbol{E}_-\left[e^{-\alpha \underline{T} + \beta \underline{S}}; \underline{J} \in +\right] = \left(\beta \boldsymbol{I}_{--} - \boldsymbol{Q}_{--}^{\alpha}\right)^{-1} \operatorname{diag}\left(\mu^{\alpha} \lambda^{\alpha}\right) \boldsymbol{P}_{-+}^{J} \operatorname{diag}(\boldsymbol{P}_+(\underline{S} = 0)).$$

The proof is finished after observing that $\boldsymbol{P}_+(\underline{S} = 0) = \boldsymbol{1}_+ - \boldsymbol{P}_{+\circlearrowleft -}^{0} \boldsymbol{1}_-$. Note that this vector is nonzero as a result of the drift condition. □

14.3 Markov-additive processes and their extremes

In this section, we study the extremes of a continuous-time Markov-additive process X with nonnegative jumps and finitely many background states. Loosely speaking, such a process is characterized by a number of Lévy processes (with nonnegative jumps) $Z^1, \ldots, Z^N$ and a continuous-time Markov process with state space $\{1, \ldots, N\}$; X behaves as Z^j when the Markov process is in state j. Our goal is to find the Laplace transform of the maximum and minimum of X, jointly with the epoch at which they are attained and the state of the Markov process at that moment.

We first give a precise definition of the process under study (Section 14.3.1). Section 14.3.2 introduces an *embedded process* that falls in the framework of Section 14.2, so that the maximum of the embedded process equals the maximum $\overline{X}$ of the original process. This embedding facilitates the computation of the desired transform, see Section 14.3.3. For the minimum, a similar procedure can be followed; the analysis of $\underline{X}$ may be found in Section 14.3.4.

14.3.1 Definitions and assumptions

A continuous-time Markov-additive process $\{(X(t), \mathsf{J}(t)) : t \geq 0\}$ is defined on some probability space $(\Omega', \mathcal{F}', \mathsf{P})$ and has càdlàg paths with values in $(\mathbb{R}, \{1, \ldots, N\})$. We only define Markov-additive processes with nonnegative jumps and a finite number of background states, but we refer to the classical papers by Çinlar [76] and by Ney and Nummelin [243] for the construction and properties of general Markov-additive processes.

Under P, $\{\mathsf{J}(t) : t \geq 0\}$ is a (finite-state) continuous-time Markovian background process, which stays in state j for an exponentially(q_j) distributed amount of time, and then jumps according to some transition matrix $\mathbf{P}^{\mathsf{J}}$. We allow J to jump to the *same* state. We assume that J is irreducible, so that there is a unique stationary distribution $\boldsymbol{\pi}_{\mathsf{J}}$ (i.e., $\boldsymbol{\pi}'_{\mathsf{J}} \mathrm{diag}(q) \mathbf{P}^{\mathsf{J}} = \boldsymbol{\pi}'_{\mathsf{J}} \mathrm{diag}(q)$). While $\mathsf{J}(t) = j$, the process $X(t)$ behaves under P as a spectrally positive (i.e., without negative jumps) Lévy process Z^j, with *Laplace exponent*

$$\psi_{-Z^j}(\beta) := \log \mathsf{E} \exp(-\beta Z^j(1)) = \frac{1}{2}\sigma_j^2 \beta^2 - c_j \beta - \int_{(0,\infty)} \left(1 - e^{-\beta y} - \beta y \mathbf{1}_{(0,1)}(y)\right) \Pi_j(dy),$$

where $\int_{(0,\infty)} (1 \wedge y^2) \Pi_j(dy) < \infty$ and $\beta, \sigma_j \geq 0$. In particular, $X(0) = 0$. The reason for writing ψ_{-Z^j} instead of ψ_{Z^j} is that we try to follow the notation of Bertoin [43, Ch. VII] as closely as possible. Let $\psi_{-Z}(\beta)$ be the vector with elements $\psi_{-Z^j}(\beta)$, $j = 1, \ldots, N$.

We need some further notation related to ψ_{-Z^j}, for which we need to suppose that $\sigma_j > 0$. By Hölder's inequality, ψ_{-Z^j} is then strictly convex. Let $\Phi_{-Z^j}(0)$ be the largest solution of the equation $\psi_{-Z^j}(\beta) = 0$, and define Φ_{-Z^j} (the 'inverse' of ψ_{-Z^j}) as the unique increasing function $\Phi_{-Z^j} : [0, \infty) \to [\Phi_{-Z^j}(0), \infty)$ such that $\psi_{-Z^j}(\Phi_{-Z^j}(\beta)) = \beta$ for $\beta > 0$. We write Φ_{-Z} for the vector with elements Φ_{-Z^j}, which is only defined for elements with $\sigma_j > 0$.

When the background process J jumps from j to k, the process X jumps according to some distribution H_{jk} on $[0, \infty)$. The matrix of the Laplace transforms corresponding to these 'environmental jumps' is written as $\mathbf{H}$, i.e., element (j, k) of the matrix $\mathbf{H}(\beta)$ equals $\int_{[0,\infty)} e^{-\beta x} H_{jk}(dx)$.

In the spirit of Section 14.2.2, we use the matrix notation

$$\mathbf{E}\left[e^{-\beta X(t)}; \mathsf{J}(t)\right] := \left\{\mathsf{E}_j\left[e^{-\beta X(t)}; \mathsf{J}(t) = k\right] : j, k = 1, \ldots, N\right\},$$

and similarly for other quantities than $X(t)$. We draw attention on the difference between $\mathbf{E}$, the matrix version of the 'continuous-time' mean E corresponding to P, and $\boldsymbol{E}$, the matrix version of the 'discrete-time' mean $\mathbb{E}$ corresponding to $\mathbb{P}$.

Using this matrix notation, the definition of (X, J) entails that $\mathsf{E}\left[e^{-\beta X(t)}; \mathsf{J}(t)\right]$ is given by $e^{t\boldsymbol{\psi}_{-X}(\beta)}$, where

$$\boldsymbol{\psi}_{-X}(\beta) = \operatorname{diag}(\psi_{-Z}(\beta)) - \operatorname{diag}(q)\left(\boldsymbol{I} - \mathbf{P}^{\mathsf{J}} \circ \mathbf{H}(\beta)\right), \tag{14.14}$$

with $\circ$ denoting componentwise (Hadamard) matrix multiplication. Note that for instance Asmussen [19] uses a slightly different (yet equivalent) representation, but ours is more convenient in the context of this chapter. The representation in (14.14) can be proven along the lines of the proof of Proposition XI.2.2 in [19], by setting up a differential equation for $\mathsf{E}_j[e^{-\beta X(t)}; \mathsf{J}(t) = k]$.

Each of the states $j = 1, \ldots, N$ can be classified as follows. When, for some j, $\sigma_j = 0$ and $c_j \geq 0$, we call j a *subordinator state*. Special cases are *zero-drift states* ($\sigma_j = c_j = 0$ and $\Pi_j \equiv 0$), *compound Poisson states* ($\sigma_j = c_j = 0$, $\Pi_j(\mathbb{R}_+) \in (0, \infty)$), and *strict subordinator states*[1] (all other subordinator states). If $\sigma_j = 0$ and $c_j < 0$, we call j a *negative-drift compound Poisson state*. A special case is a *negative-drift* state, for which $\Pi_j \equiv 0$. The other states are called *Brownian states*; these are characterized by $\sigma_j > 0$ (note that for these Brownian states Z^j is not necessarily a Brownian motion with drift; it may also be that $\Pi_j \not\equiv 0$).

There is no one-to-one correspondence between $\boldsymbol{\psi}_{-X}$ and tuples $(\psi_{-Z}, q, \mathbf{P}^{\mathsf{J}}, \mathbf{H})$. For instance, consider the situation that Z^j corresponds to the sum of a Brownian motion and a compound Poisson process. Then one could equivalently do as if there are environmental jumps at the jump epochs of the Poisson process; by also adapting the transition matrix, one obtains an alternative description of the same stochastic process.

Consequently, since J is allowed to make self-transitions, without loss of generality we can assume that there are no compound Poisson states nor negative-drift compound Poisson states. Indeed, these states can be replaced by zero-drift or negative-drift states, provided the H_{jj} and q_j are changed appropriately. Throughout, we suppose that there is at least one negative-drift state or Brownian state after this simplification (if X drifts to $-\infty$, then this is a consequence of the spectral positivity).

The above observations allow a partitioning of the states $1, \ldots, N$ of the background process into

(i) the strict subordinator states, labeled 's';

(ii) the zero-drift states, labeled 'z';

(iii) the negative-drift states, labeled 'n'; and

(iv) the Brownian states, labeled 'B'.

In the following, we *always* assume that the state space $\{1, \ldots, N\}$ of J is partitioned in the order s–z–n–B. Sometimes, it is unnecessary to distinguish between s- and z-states, and it is therefore convenient to refer to s- and z-states as $\underline{s}$-states. If we use this $\underline{s}$-notation in block matrices, we suppose that the order is s–z. Similarly, we refer to n- and B-states as $\sim$-states, again preserving the order.

We also need another probability measure on $(\Omega', \mathcal{F}')$, denoted by $\widehat{\mathsf{P}}$. Under $\widehat{\mathsf{P}}$, (X, J) is a Markov-additive process with Laplace exponent

$$\widehat{\boldsymbol{\psi}}_{-X}(\beta) := \operatorname{diag}(1/\boldsymbol{\pi}_{\mathsf{J}})\boldsymbol{\psi}'_{-X}(\beta)\operatorname{diag}(\boldsymbol{\pi}_{\mathsf{J}}). \tag{14.15}$$

That is, working with (X, J) under $\widehat{\mathsf{P}}$ amounts to working with the *time-reversed* Markov-additive process under the measure P, and vice versa.

[1] It is customary in the literature to use the term *strict subordinator* for a subordinator with an infinite lifetime; here, it stands for a strictly increasing subordinator.

As before, we define

$$\begin{aligned}
\overline{X}(t) &:= \sup\{X(s) : 0 \le s \le t\},\\
\overline{F}^X(t) &:= \inf\{s < t : X(s) = \overline{X}(t) \text{ or } X(s-) = \overline{X}(t)\},\\
\underline{X}(t) &:= \inf\{X(s) : 0 \le s \le t\},\\
\underline{F}^X(t) &:= \inf\{s < t : X(s) = \underline{X}(t) \text{ or } X(s-) = \underline{X}(t)\}.
\end{aligned}$$

We also set $\overline{\mathsf{J}}(t) := \mathsf{J}(\overline{F}^X(t))$ and $\underline{\mathsf{J}}(t) = \mathsf{J}(\underline{F}^X(t))$. It is our aim to study these quantities as $t \to \infty$, in which case we omit the time index. We study the joint P-distributions of $(\overline{X}, \overline{F}^X, \overline{\mathsf{J}})$ (in Section 14.3.3) and $(\underline{X}, \underline{F}^X, \underline{\mathsf{J}})$ (in Section 14.3.4). We rely extensively on two fundamental properties of Lévy processes, which we recall in the next subsection.

14.3.2 Intermezzo on Lévy processes

In this intermezzo, we consider a Lévy process Z (i.e., there is no background process) with killing at an exponentially distributed epoch. We let e_q denote the killing epoch with mean $1/q$, and suppose that it is independent of Z. We also suppose that the process does not have negative jumps, and that it is 'Brownian' in the terminology of the previous subsection, i.e., that Z has a Brownian component. Note that consequently the inverse Φ_{-Z} of the Laplace exponent is well-defined.

We start with two observations that have already been proven useful in the preceding two chapters. First, if we define quantities $\overline{Z}, \overline{F}^Z, \underline{Z}$, and $\underline{F}^Z$ similarly as for X, we have the following interesting identities: for $\alpha, \beta \ge 0$,

$$\begin{aligned}
\mathsf{E}e^{-\alpha\underline{F}^Z(e_q)+\beta\underline{Z}(e_q)} &= \mathsf{E}e^{-\alpha\underline{F}^Z(e_q)}\mathsf{E}e^{-\beta\underline{Z}(e_{q+\alpha})},\\
\mathsf{E}e^{-\alpha\left(e_q-\overline{F}^Z(e_q)\right)-\beta\left(\overline{Z}(e_q)-Z(e_q)\right)} &= \mathsf{E}e^{-\alpha\left(e_q-\overline{F}^Z(e_q)\right)}\mathsf{E}e^{-\beta\left(\overline{Z}(e_{q+\alpha})-Z(e_{q+\alpha})\right)},
\end{aligned}$$

see also (12.3).

Moreover, the PRS identity in Proposition 11.5 shows there are two ways of decomposing $(e_q, Z(e_q))$ into two *independent* vectors:

(i)
- a vector $(\sigma, U) := (\overline{F}^Z(e_q), \overline{Z}(e_q))$ related to the process till time $\overline{F}^Z(e_q)$, and
- an *independent* second vector $(\tau, -D) := (e_q - \overline{F}^Z(e_q), Z(e_q) - \overline{Z}(e_q))$ related to the process between $\overline{F}^Z(e_q)$ and e_q.

(ii)
- a vector $(\underline{F}^Z(e_q), \underline{Z}(e_q))$ related to the process till time $\underline{F}^Z(e_q)$ (this vector has the same distribution as $(\tau, -D)$), and
- an *independent* second vector $(e_q - \underline{F}^Z(e_q), Z(e_q) - \underline{Z}(e_q))$ related to the process between time $\underline{F}^Z(e_q)$ and e_q (this vector has the same distribution as (σ, U)).

In the special case of no jumps, Asmussen [16] exploits the first *splitting property* in the context of Markov-additive processes.

Due to the assumptions that Z is spectrally positive and that it has a Brownian component, $\overline{Z}(e_{q+\alpha}) - Z(e_{q+\alpha})$ has an exponential distribution; see Proposition 11.6 (applied for $X = -Z$). Importantly, Proposition 11.6 also gives an expression for the Laplace transform of (σ, U), and shows that (τ, D) satisfies Assumption 14.1 with $\mu^\alpha = \Phi_{-Z}(q + \alpha)$. The latter property facilitates application of the results of Section 14.2 in the context of continuous-time Markov-additive processes, as we demonstrate in the next subsection.

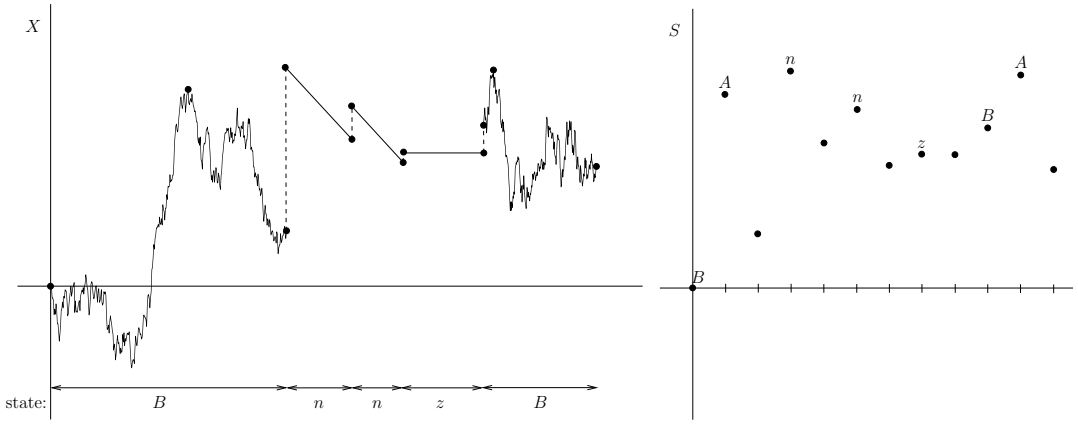

Figure 14.1: The left-hand diagram represents the process X with its embedding points, along with the state *labels. The discrete-time embedded process S is given in the right-hand diagram, along with the* point *labels.*

14.3.3 The distribution of $(\overline{X}, \overline{F}^X, \overline{\mathsf{J}})$

We have collected all the necessary prerequisites for an embedding that allows us to characterize the distribution of $(\overline{X}, \overline{F}^X, \overline{\mathsf{J}})$, provided X drifts to $-\infty$. *Throughout this subsection*, we impose the stronger requirement that $\boldsymbol{\pi}'_{\mathsf{J}}\mathbf{E}X(1) < 0$, but, as in Section 14.2, the majority of our results only requires the weaker assumption that X drifts to $-\infty$ almost surely; this holds in particular for our main result, Theorem 14.19.

To find the distribution of $(\overline{X}, \overline{F}^X, \overline{\mathsf{J}})$, we do not monitor the full process (X, J), but we just record time and position at 'special' epochs only. For $\underline{s}$-states and n-states, these epochs are

- whenever a sojourn time in these states starts; we call these *$\underline{s}$-points* and *n-points*, respectively, and
- immediately before this sojourn time ends (where the environmental jump at that epoch is *not* included).

For B-states, these epochs are

- whenever a sojourn time in these states starts; we call these *B-points*,
- whenever the maximum within this sojourn time is attained; we call these *A-points*, and
- immediately before this sojourn time ends (again without the environmental jump at that epoch).

Note that we have thus constructed a discrete-time stochastic process from X that still contains all information on the maximum of X. We call this process the *embedded process*. Importantly, as a result of the independence discussed in Section 14.3.2, the embedded process fits into the framework of Section 14.2, when the space-component of the embedded points is recorded in S and the time-component in T. The embedding is illustrated in Figure 14.1; in the realization of X, a negative-drift compound Poisson state has been replaced by a negative-drift state with environmental jumps and self-transitions. Note that some of the embedding points remain unlabeled.

Motivated by this embedding, we label n-points and A-points as $-$-points (as from these points the process moves down), in accordance with the terminology of Section 14.2.2. The order is $n - A$. Moreover, we refer to n-points and B-points as $\sim$-points. The (environmental-jump) points immediately after the $-$-points are called $\uparrow$-points. These points are *always* incorporated into the embedded process, even if there are no environmental jumps; the step size is then simply zero. To be able to perform arithmetics, the order is again preserved: the j-th $\uparrow$-point comes after the j-th $-$-point, which in turn corresponds to the j-th $\sim$-point.

Application of this labeling shows that we have

$$\lambda^\alpha := \begin{pmatrix} \operatorname{vec}\left(\frac{q_n+\alpha}{-c_n}\right) \\ \operatorname{vec}(\Phi_{-Z}(q_B+\alpha)) \end{pmatrix}, \quad \mu^\alpha := \begin{pmatrix} \operatorname{vec}\left(\frac{q_n}{q_n+\alpha}\right) \\ \operatorname{vec}\left(\frac{\Phi_{-Z}(q_B)}{\Phi_{-Z}(q_B+\alpha)}\right) \end{pmatrix}. \tag{14.16}$$

The notation in (14.16) should be interpreted as follows. First, q_n is the block vector of q that corresponds to n; similarly c_n is the block vector of the drift vector c corresponding to n. Then $(q_n + \alpha)/c_n$ is the vector with element j equal to $(q_{n,j} + \alpha)/c_{n,j}$. The vector q_B is defined analogously to q_n. With $k = 1, \ldots, N$ being the index of the j-th B-state, the j-th element of $\Phi_{-Z}(q_B + \alpha))$ is $\Phi_{-Z^k}(q_{B,j} + \alpha))$. The notation used in the definition of μ^α should be read in a similar fashion.

Note that an explicit expression for Φ_{-Z} can be given for Brownian states without jumps; this is exploited by Asmussen [16].

With the theory of Section 14.2 at hand, the embedding argument shows that the key quantities for studying the distribution of $(\overline{X}, \overline{F}^X, \overline{\mathsf{J}})$ are the matrices ${}^{\#}\widehat{\mathbf{P}}^\alpha_{-z}$ and ${}^{\#}\widehat{\mathbf{P}}^\alpha_{--}$ that contain the last-passage transforms of the embedded process. Here, # refers to time-reversal with respect to the Markov chain underlying the *embedded* (discrete-time) Markov-additive process, cf. (14.2). However, we shall not rely on this exact definition, since it is easier to interpret the ${}^{\#}\widehat{\mathbf{P}}^\alpha$-matrices directly as last-passage transforms.

We next focus on the calculation of these last-passage matrices. To this end, we set for $\alpha, \beta \geq 0$,

$$\mathbf{F}_{\underline{s}_\circlearrowleft}(\alpha, \beta) := \left(\alpha \boldsymbol{I}_{\underline{s}\underline{s}} - \boldsymbol{\psi}_{-X^{\underline{s}\underline{s}}}(\beta)\right)^{-1} \operatorname{diag}(q_{\underline{s}}),$$

where $\boldsymbol{\psi}_{-X^{\underline{s}\underline{s}}}$ is the $(\underline{s}, \underline{s})$-block in the matrix $\boldsymbol{\psi}_{-X}$. The matrices $\mathbf{F}_{s_\circlearrowleft}(\alpha, \beta)$ and $\mathbf{F}_{z_\circlearrowleft}(\alpha, \beta)$ are defined similarly, with $\underline{s}$ replaced by s and z respectively. The idea behind this definition is that $\mathbf{F}_{\underline{s}_\circlearrowleft}$ characterizes the displacement in time and space when we start in an $\underline{s}$-state, then stay in $\underline{s}$-states, until a $\sim$-state is reached. It is important to realize that the change in the position due to the environmental jump before reaching the $\sim$-state is not included; these jumps appear in the following formulas. We define for $\alpha, \beta \geq 0$,

$$\mathbf{F}_{\uparrow \underline{s}_\circlearrowleft \cdot}(\alpha, \beta) := \mathbf{P}^{\mathsf{J}}_{\sim \underline{s}} \circ \mathbf{H}_{\sim \underline{s}}(\beta) \mathbf{F}_{\underline{s}_\circlearrowleft}(\alpha, \beta) \mathbf{P}^{\mathsf{J}}_{\underline{s}\cdot} \circ \mathbf{H}_{\underline{s}\cdot}(\beta) + \mathbf{P}^{\mathsf{J}}_{\sim\cdot} \circ \mathbf{H}_{\sim\cdot}(\beta),$$

where '$\cdot$' can be replaced by any of the blocks s, z, n, or B. The first term should be interpreted as zero if there are no $\underline{s}$-states. Importantly, we have now defined $\mathbf{F}_{\uparrow \underline{s}_\circlearrowleft \sim}(\alpha, \beta)$, which corresponds to the displacement in time and space between the *end* of a sojourn time in a $\sim$-state and the *beginning* of a sojourn time in the next $\sim$-state, including both environmental jumps.

We also set

$$\mathbf{F}_{\uparrow \underline{s}_\circlearrowleft A}(\alpha, \beta) := \mathbf{F}_{\uparrow \underline{s}_\circlearrowleft B}(\alpha, \beta) \operatorname{diag}\left(\mathbf{E}_B e^{-\alpha \overline{F}^Z(e_q) - \beta \overline{Z}(e_q)}\right),$$

where the diagonal matrix in the right-hand side should be read as follows. Denote q_B as before, and let $k = 1, \ldots, N$ be the index of the j-th B-state (so that $q_k = q_{B,j}$). Then

$$\operatorname{diag}\left(\mathbf{E}_B e^{-\alpha \overline{F}^Z(e_q) - \beta \overline{Z}(e_q)}\right)_{jj} := \mathsf{E}_k\left[e^{-\alpha \overline{F}^{Z^k}(e_{q_k}) - \beta \overline{Z^k}(e_{q_k})}\right].$$

We now have all prerequisites to show how ${}^{\#}\widehat{\mathbf{P}}^{\alpha}_{-z}$ and ${}^{\#}\widehat{\mathbf{P}}^{\alpha}_{--}$ can be calculated. A proof is given later, since it relies on a slightly different, 'sparser' embedding than the one that we have just described. The measure-valued matrices $\mathbf{F}^{\alpha}_{\uparrow \underline{s}_{\circlearrowleft}-}(dx)$ and $\mathbf{F}^{\alpha}_{\uparrow s_{\circlearrowleft} z}(dx)$ are defined similarly as in Section 14.2.3.

Proposition 14.18 *For $\alpha \geq 0$, let ${}^{\sharp}\widehat{\boldsymbol{\mathcal{P}}}^{\alpha}_{--}$, ${}^{\sharp}\widehat{\boldsymbol{\mathcal{Q}}}^{\alpha}_{--}$ satisfy the nonlinear system*

$$\begin{aligned} {}^{\sharp}\widehat{\boldsymbol{\mathcal{P}}}^{\alpha}_{--} &= \int_{(0,\infty)} e^{\,{}^{\sharp}\widehat{\boldsymbol{\mathcal{Q}}}^{\alpha}_{--} x} \operatorname{diag}(\mu^{\alpha}) \mathbf{F}^{\alpha}_{\uparrow \underline{s}_{\circlearrowleft}-}(dx), \\ {}^{\sharp}\widehat{\boldsymbol{\mathcal{Q}}}^{\alpha}_{--} &= -\left[\boldsymbol{I}_{--} - \operatorname{diag}(\mu^{\alpha}) \mathbf{F}_{\uparrow z_{\circlearrowleft}-}(\alpha,\infty) - {}^{\sharp}\widehat{\boldsymbol{\mathcal{P}}}^{\alpha}_{--} \right] \operatorname{diag}(\lambda^{\alpha}). \end{aligned}$$

We then have

$$\begin{aligned} {}^{\#}\widehat{\mathbf{P}}^{\alpha}_{-z} &= \int_{(0,\infty)} e^{\,{}^{\sharp}\widehat{\boldsymbol{\mathcal{Q}}}^{\alpha}_{--} x} \operatorname{diag}(\mu^{\alpha}) \mathbf{F}^{\alpha}_{\uparrow s_{\circlearrowleft} z}(dx), \\ {}^{\#}\widehat{\mathbf{P}}^{\alpha}_{--} &= {}^{\sharp}\widehat{\boldsymbol{\mathcal{P}}}^{\alpha}_{--} - \begin{pmatrix} {}^{\#}\widehat{\mathbf{P}}^{\alpha}_{-z} \mathbf{F}_{z_{\circlearrowleft}}(\alpha,\infty) \mathbf{P}^{\mathsf{J}}_{zn} \circ \mathbf{H}_{zn}(\infty) & \mathbf{0}_{-B} \end{pmatrix}. \end{aligned}$$

The analogue of the matrix $\boldsymbol{K}^{\alpha}_{--}$ (see Section 14.2.3) is

$$\boldsymbol{\mathcal{K}}^{\alpha}_{--} := \operatorname{diag}\left(\frac{q_{\sim}}{\mu^{\alpha}}\right) {}^{\sharp}\widehat{\boldsymbol{\mathcal{Q}}}^{\alpha}_{--} \operatorname{diag}\left(\frac{\mu^{\alpha}}{q_{\sim}}\right),$$

and this matrix plays a prominent role in the fluctuation theory for Markov-additive processes.

To formulate our next result, we need to define closely related last-passage matrices; their precise relationship to $\boldsymbol{\mathcal{K}}^{\alpha}_{--}$ is investigated below. Compared to Section 14.2, it is somewhat more involved to work with last-passage matrices in the general Markov-additive setting, due to the presence of subordinator states and Brownian states. We set

$$\begin{aligned} \mathbf{K}^{\alpha}_{--} &:= -\operatorname{diag}\left(\frac{q_{\sim}}{\mu^{\alpha}}\right) \left[\boldsymbol{I}_{--} - \operatorname{diag}(\mu^{\alpha}) \mathbf{F}_{\uparrow z_{\circlearrowleft}-}(\alpha,\infty) - {}^{\#}\widehat{\mathbf{P}}^{\alpha}_{--} \right] \operatorname{diag}\left(\frac{\mu^{\alpha}\lambda^{\alpha}}{q_{\sim}}\right), \\ \mathbf{K}^{\alpha}_{-z} &:= \operatorname{diag}\left(\frac{q_{\sim}}{\mu^{\alpha}}\right) {}^{\#}\widehat{\mathbf{P}}^{\alpha}_{-z}, \end{aligned}$$

and define the α-independent matrices

$$\mathbf{K}_{zz} := -\operatorname{diag}(q_z) \left[\boldsymbol{I}_{zz} - \mathbf{P}^{\mathsf{J}}_{zz} \circ \mathbf{H}_{zz}(\infty) \right], \quad \mathbf{K}_{zn} := \operatorname{diag}(q_z) \mathbf{P}^{\mathsf{J}}_{zn} \circ \mathbf{H}_{zn}(\infty),$$

and $\mathbf{K}_{zA} := \mathbf{0}_{zB}$. We remark that we have lost the interpretation of these matrices as intensity matrices related to the last-passage process.

The following theorem, which is the main result of this subsection and the matrix version of (14.1), should be compared with (14.4). Note that the presence of the matrix $\boldsymbol{\psi}_{-X}(\beta) - \alpha \boldsymbol{I}$ is already apparent from the results of Kaspi [179].

Theorem 14.19 *For $\alpha, \beta \geq 0$, we have*

$$\begin{aligned} &(\boldsymbol{\psi}_{-X}(\beta) - \alpha \boldsymbol{I}) \, \mathsf{E}\left[e^{-\alpha \overline{F}^{X} - \beta \overline{X}}; \mathsf{J} \right] \\ &= \begin{pmatrix} \mathbf{0}_{ss} & \mathbf{0}_{sz} & \mathbf{0}_{s-} \\ \mathbf{0}_{zs} & \mathbf{K}_{zz} - \alpha \boldsymbol{I}_{zz} & \mathbf{K}_{z-} \\ \mathbf{0}_{-s} & \mathbf{K}^{\alpha}_{-z} & \beta \boldsymbol{I}_{--} + \mathbf{K}^{\alpha}_{--} \end{pmatrix} \operatorname{diag}\begin{pmatrix} \mathbf{0}_{s} \\ \mathbf{P}_{z}(\overline{X} = 0) \\ \boldsymbol{v}_{-} \end{pmatrix}, \end{aligned}$$

where $\boldsymbol{v}_{-} = \left(-c_n \mathbf{P}_n(\overline{X} = 0)); \frac{q_B}{\Phi_{-Z}(q_B)} \mathbf{P}_A(\overline{X} = 0) \right)$.

Proof. Write

$$\boldsymbol{C}(\alpha,\beta) := \operatorname{diag}\left(\mathsf{E}e^{-\alpha e_q + i\beta Z(e_q)}\right)\mathbf{P}^{\mathsf{J}} \circ \mathbf{H}(-i\beta).$$

A formal proof could be based on direct application of Theorem 14.8; due to the presence of many auxiliary states, a substantial amount of matrix algebra is then needed. However, we feel that it is more insightful to follow the lines of the arguments used earlier, when interpreting Theorem 14.8.

For instance, let us consider the scenario that both $\mathsf{J}(0)$ and $\overline{\mathsf{J}}$ are Brownian states. Consider the embedded process: it starts in a B-point. In terms of the embedded process, the maximum S^* should be in an A-point, after some number of steps, say n^*. In the first place, it should be that *after* n^*, the embedded process never exceeds S^*. This leads to the expression

$$(\boldsymbol{I} - \boldsymbol{C}(\alpha,\beta))^{-1}_{BB} \operatorname{diag}\left(\mathsf{E}_B e^{-\alpha \overline{F}^Z(e_q) + i\beta \overline{Z}(e_q)}\right) \operatorname{diag}\left(\mathbf{P}_A(\overline{X}=0)\right).$$

However, there are paths for which the process exceeds S^* *before* n^*; these need to be subtracted from the above expression. Looking back in time, starting from n^*, this means that there is a weak descending ladder epoch, necessarily an A-point or n-point. If it is an A-point, it must be preceded by a B-point. Hence, the contribution of the paths that need to be subtracted is

$$\begin{aligned}&(\boldsymbol{I} - \boldsymbol{C}(\alpha,\beta))^{-1}_{BB} \operatorname{diag}\left(\mathsf{E}_B e^{-\alpha \overline{F}^Z(e_q) + i\beta \overline{Z}(e_q)} \circ \frac{\Phi_{-Z}(q_B+\alpha)}{\Phi_{-Z}(q_B+\alpha)+i\beta}\right) {}^{\#}\widehat{\mathbf{P}}^{\alpha}_{AA} \operatorname{diag}\left(\mathbf{P}_A(\overline{X}=0)\right)\\ &\quad + (\boldsymbol{I} - \boldsymbol{C}(\alpha,\beta))^{-1}_{Bn} \operatorname{diag}\left(\frac{q_n+\alpha}{q_n+\alpha-c_n i\beta}\right) {}^{\#}\widehat{\mathbf{P}}^{\alpha}_{nA} \operatorname{diag}\left(\mathbf{P}_A(\overline{X}=0)\right).\end{aligned}$$

A similar formula can be found if J starts in a subordinator state or a negative-drift state. Finally, the fact that

$$\boldsymbol{I} - \boldsymbol{C}(\alpha,\beta) = \operatorname{diag}(q+\alpha-\psi_{-Z}(-i\beta))^{-1}\left[\alpha \boldsymbol{I} - \boldsymbol{\psi}_{-X}(-i\beta)\right],$$

yields the asserted B-column after some elementary algebra.

This argument can be repeated if $\overline{\mathsf{J}}$ is a zero-drift state or a negative-drift state, but some additional arguments are needed. In that case, we also need to subtract paths that go from a z-point to a z-point (or n-point) without a strictly positive environmental jump, and then have no strict ascending ladder height. This accounts for the terms $\mathbf{K}_{zz} - \alpha \boldsymbol{I}_{zz}$ and $\mathbf{K}_{zn}$. □

We now show that the unknown vectors $\mathbf{P}_z(\overline{X}=0)$ and $\boldsymbol{v}_-$ can be found in almost exactly the same way as in Section 14.2.4. The following lemma casts Lemma 14.10 and Proposition 14.12 into the general Markov-additive setting. It shows that two important properties carry over to this general framework. First, $\mathbf{P}_z(\overline{X}=0)$ can be expressed in terms of $\boldsymbol{v}_-$ (more precisely, in terms of $\mathbf{P}_n(\overline{X}=0)$). Furthermore, there is a simple relationship between $\boldsymbol{v}_-$ and the right eigenvector of ${}^{\sharp}\widehat{\boldsymbol{\mathcal{Q}}}^{0}_{--} \operatorname{diag}(1/q_\sim)$ corresponding to its (simple) eigenvalue zero.

Lemma 14.20 *The* $\mathbf{K}$*-matrices have the following properties:*

(i) $\mathbf{P}_z(\overline{X}=0) = -\mathbf{K}^{-1}_{zz}\mathbf{K}_{z-}\boldsymbol{v}_-$, *and*

(ii) if there is more than one $\sim$*-state, then* $\boldsymbol{v}_-$ *is a right eigenvector of* $\boldsymbol{\mathcal{K}}^0_{--} = \mathbf{K}^0_{--} - \mathbf{K}^0_{-z}\mathbf{K}^{-1}_{zz}\mathbf{K}_{z-}$.

Next we formulate a result in the same spirit as Corollary 14.11, which immediately follows from Theorem 14.19 and Lemma 14.20. It is the Markov-additive version of (14.1) for $\alpha = 0$. A closely related formula has been obtained by Asmussen and Kella [24, Eq. (4.1)], who phrase their result in terms of the reflected process and a local-time vector. The precise relationship between the two formulas is further investigated in Section 14.4.2.

Corollary 14.21 *For $\beta \geq 0$, we have*

$$\boldsymbol{\psi}_{-X}(\beta)\mathbf{E}e^{-\beta\overline{X}} = \beta \begin{pmatrix} \mathbf{0}_{\underline{s}} \\ \boldsymbol{v}_- \end{pmatrix}.$$

The vector $\boldsymbol{v}_-$ is determined by Lemma 14.20 and the next lemma, which is an analogue of Lemma 14.13. Note that this lemma corrects Equation (4.2) in [24].

Lemma 14.22 *We have*

$$-\boldsymbol{\pi}_{\mathsf{J}}'\mathbf{E}X(1) = \boldsymbol{\pi}_{\mathsf{J}}(\sim)'\boldsymbol{v}_-.$$

Proof. Since $\boldsymbol{\pi}_{\mathsf{J}}$ satisfies $\boldsymbol{\pi}_{\mathsf{J}}'\mathrm{diag}(q)\mathbf{P}^{\mathsf{J}} = \boldsymbol{\pi}_{\mathsf{J}}'\mathrm{diag}(q)$, Corollary 14.21 shows that for $\beta > 0$,

$$\frac{1}{\beta}\boldsymbol{\pi}_{\mathsf{J}}'\left[\mathrm{diag}(\psi_{-Z}(\beta)) - \mathrm{diag}(q)\left(\mathbf{P}^{\mathsf{J}} - \mathbf{P}^{\mathsf{J}} \circ \mathbf{H}(\beta)\right)\right]\mathbf{E}e^{-\beta\overline{X}} = \boldsymbol{\pi}_{\mathsf{J}}(\sim)'\boldsymbol{v}_-.$$

Now let $\beta \to 0$ to obtain

$$-\boldsymbol{\pi}_{\mathsf{J}}'\left[\mathbf{E}Z(1) + \mathrm{diag}(q)\mathbf{P}^{\mathsf{J}} \circ \int x\mathbf{H}(dx)\right] = \boldsymbol{\pi}_{\mathsf{J}}(\sim)'\boldsymbol{v}_-.$$

Using Corollary XI.2.9(b) and (the second equality in) Corollary XI.2.5 of Asmussen [19], it is not hard to see that the left-hand side equals $-\boldsymbol{\pi}_{\mathsf{J}}'\mathbf{E}X(1)$. □

Censored embedding and spectral considerations

Let us consider the embedded process (only) on $-$-points and $\uparrow$-points. We refer to the resulting process as the *censored embedded process*. In the censored embedded process, one always jumps from a $-$-point to a $\uparrow$-point and vice versa. Using the notation of Section 14.2, this means that $|\mathcal{I}_+| = |\mathcal{I}_-|$, $\boldsymbol{F}_{++}(\alpha,\beta) = \mathbf{0}_{++}$, and $\boldsymbol{F}_{--}(\alpha,\beta) = \mathbf{0}_{--}$, while

$$\boldsymbol{F}_{+-}(\alpha,\beta) = \mathbf{F}_{\uparrow\underline{s}_{\bigcirc}-}(\alpha,-i\beta), \quad \boldsymbol{F}_{-+}(\alpha,\beta) = \mathrm{diag}\left(\frac{\mu^\alpha\lambda^\alpha}{\lambda^\alpha + i\beta}\right).$$

It is left to the reader to check that this leads to

$$\begin{aligned}\boldsymbol{D}_{--}(\alpha,\beta) \;=\; & \mathrm{diag}\left(\frac{\mu^\alpha\lambda^\alpha}{q_\sim}\right)\left[\mathrm{diag}(\psi_{-Z^{\sim\sim}}(\beta)) - \mathrm{diag}(q_\sim)\left(\boldsymbol{I}_{--} - \mathbf{F}_{\uparrow\underline{s}_{\bigcirc}\sim}(\alpha,\beta)\right) - \alpha\boldsymbol{I}_{--}\right] \\ & \times \mathrm{diag}\left(\mathbf{E}_\sim e^{-\alpha\overline{F}^Z(e_q) - \beta\overline{Z}(e_q)}\right).\end{aligned}$$

Therefore, the factorization identity (14.8) can be rewritten as

$$\mathrm{diag}(\psi_{-Z^{\sim\sim}}(\beta)) - \mathrm{diag}(q_\sim)\left(\boldsymbol{I}_{--} - \mathbf{F}_{\uparrow\underline{s}_{\bigcirc}\sim}(\alpha,\beta)\right) - \alpha\boldsymbol{I}_{--} = \left[\beta\boldsymbol{I}_{--} + \boldsymbol{\mathcal{K}}^\alpha_{--}\right]\boldsymbol{M}'_{--}(\alpha,\beta), \tag{14.17}$$

for some matrix $\boldsymbol{M}'_{--}(\alpha,\beta)$ which is nonsingular if $\Re(\beta) \geq 0$.

This identity immediately shows that the reasoning in Section 14.2.5 can be repeated verbatim in the general Markov-additive case, showing how $\boldsymbol{\mathcal{K}}^\alpha_{--}$ can be found based on 'spectral' considerations. We also note that $\boldsymbol{\mathcal{K}}^0_{--}$ determines $\boldsymbol{v}_-$ up to a constant, since the nullspace of $\boldsymbol{\mathcal{K}}^0_{--}$ has dimension one. We have thus provided answers to the questions raised in Section 4 of Asmussen and Kella [24].

With the censored process at our disposal, we can prove Proposition 14.18; the matrix ${}^\sharp\widehat{\boldsymbol{\mathcal{P}}}^\alpha_{--}$ contains the last-passage transforms for the *censored* embedded process.

Proof of Proposition 14.18. By applying Proposition 14.3 to the censored embedded process, the system for ${}^{\sharp}\widehat{\boldsymbol{\mathcal{P}}}^{\alpha}_{--}$ and ${}^{\sharp}\widehat{\boldsymbol{\mathcal{Q}}}^{\alpha}_{--}$ is readily found. The expressions for ${}^{\#}\widehat{\mathbf{P}}^{\alpha}_{-z}$ and ${}^{\#}\widehat{\mathbf{P}}^{\alpha}_{--}$ follow by considering the *whole* embedded process again. The first formula is a consequence of the fact that there must be a (censored) $\uparrow$-point before a z-point if it is a last-passage point. The matrix ${}^{\#}\widehat{\mathbf{P}}^{\alpha}_{--}$ is found upon noticing that if an n-point is a last-passage point for the censored process, then either a z-point or (the same) n-point is the corresponding last-passage point in the *whole* embedded process. □

From a theoretical point of view, there are two main reasons why we feel that the system in Proposition 14.18 is unsatisfactory:

- It contains the transform $\mathbf{E}_B e^{-\alpha \overline{F}^Z(e_q)-\beta \overline{Z}(e_q)}$, and is therefore not really the analogue of Corollary 14.4. In other words, the presence of the function Φ_{-Z} is undesirable, since it arises from a nonlinear system itself.
- It cannot be viewed as a matrix analogue of $\alpha = \psi_{-Z}(\Phi_{-Z}(\alpha))$.

The formula in the following proposition has the 'right' form in view of the above two issues. In the light of recent progress made by Pistorius [256], this result may have attractive numerical features as well. To see that it is a matrix version of $\alpha = \psi_{-Z}(\Phi_{-Z}(\alpha))$, recall the representation of the Laplace exponent of X in (14.14).

Proposition 14.23 *For $\alpha \geq 0$, $\boldsymbol{\mathcal{K}}^{\alpha}_{--}$ is the unique matrix that solves the nonlinear system*

$$\begin{aligned}\alpha \boldsymbol{I}_{--} \;=\; & \left(\boldsymbol{\mathcal{K}}^{\alpha}_{--}\right)^2 \operatorname{diag}\left(\frac{\sigma^2_{\sim}}{2}\right) + \boldsymbol{\mathcal{K}}^{\alpha}_{--} \operatorname{diag}(c_{\sim}) \\ & - \int_{(0,\infty)} \left(\boldsymbol{I}_{--} - e^{\boldsymbol{\mathcal{K}}^{\alpha}_{--} y} + \boldsymbol{\mathcal{K}}^{\alpha}_{--} y \mathbf{1}_{(0,1)}(y)\right) \operatorname{diag}(\Pi_{\sim}(dy)) \\ & - \operatorname{diag}(q_{\sim}) + \int_{[0,\infty)} e^{\boldsymbol{\mathcal{K}}^{\alpha}_{--} y} \operatorname{diag}(q_{\sim}) \mathbf{F}^{\alpha}_{\uparrow \underline{s}_{\circlearrowleft} \sim}(dy).\end{aligned}$$

Moreover,

$$\mathbf{K}^{\alpha}_{-z} = \int_{(0,\infty)} e^{\boldsymbol{\mathcal{K}}^{\alpha}_{--} x} \operatorname{diag}(q_{\sim}) \mathbf{F}^{\alpha}_{\uparrow s_{\circlearrowleft} z}(dx)$$

and

$$\mathbf{K}^{\alpha}_{--} = \boldsymbol{\mathcal{K}}^{\alpha}_{--} - \left(\begin{array}{cc} \mathbf{K}^{\alpha}_{-z} \mathbf{F}_{z_{\circlearrowleft}}(\alpha,\infty) \mathbf{P}^{\mathsf{J}}_{zn} \circ \mathbf{H}_{zn}(\infty) \operatorname{diag}(-1/c_n) & \mathbf{0}_{-B} \end{array} \right).$$

Proof. The first claim is a consequence of (14.17), after repeating the proof of Theorem 14.15 and the subsequent reasoning. The last two formulas follow from the corresponding expressions in terms of the ${}^{\#}\widehat{\mathbf{P}}^{\alpha}$-matrices, see Proposition 14.18. □

14.3.4 The distribution of $(\underline{X}, \underline{F}^X, \underline{\mathsf{J}})$

In this subsection, we study the minimum of X if it drifts to $+\infty$. We suppose *throughout this subsection* that $\boldsymbol{\pi}'_{\mathsf{J}} \mathbf{E} X(1) > 0$.

As in the previous subsection, we do not monitor the full process (X, J), but we only record for $\underline{s}$-states and n-states the time and position at the start (leading to $\underline{s}$-points and n-points, respectively) and immediately before the end of the sojourn time, and for B-states in addition the minimum within the sojourn times (leading to A-points).

While an A-point was a $-$-point in the previous subsection, the situation is now different. In order to still preserve our conventions of Section 14.2.2, it is therefore necessary to group the points differently. Hence, the $-$-points in this subsection *are not the same as in Section 14.3.3*:

we now say that n-points and B-points are $-$-points. The points immediately after $-$-points are still called $\uparrow$-points; these are also different from before, since for instance A-points are now $\uparrow$-points. Again, when using this notation in block matrices, we adopt the indicated order.

Despite the relabeling, as a result of the theory in Section 14.3.2 and the new definition of $-$-states, λ^α and μ^α are still given by (14.16). We keep the notation $\mathbf{F}_{\underline{s}_{\circlearrowleft}}$ of the previous subsection, but we also define for $\alpha, \beta \geq 0$,

$$\begin{aligned}
\mathbf{F}_{\underline{s}_{\circlearrowleft}\cdot}(\alpha,\beta) &:= \mathbf{F}_{\underline{s}_{\circlearrowleft}}(\alpha,\beta)\mathbf{P}^{\mathsf{J}}_{\underline{s}\cdot} \circ \mathbf{H}_{\underline{s}\cdot}(\beta), \\
\mathbf{F}_{\uparrow\underline{s}_{\circlearrowleft}-}(\alpha,\beta) &:= \operatorname{diag}\left(\mathsf{E}_{\sim} e^{-\alpha\overline{F}^{Z}(e_q)-\beta\overline{Z}(e_q)}\right)\left[\mathbf{P}^{\mathsf{J}}_{\sim\underline{s}} \circ \mathbf{H}_{\sim\underline{s}}(\beta)\mathbf{F}_{\underline{s}_{\circlearrowleft}-}(\alpha,\beta) + \mathbf{P}^{\mathsf{J}}_{\sim\sim} \circ \mathbf{H}_{\sim\sim}(\beta)\right],
\end{aligned}$$

where again '$\cdot$' stands for any of the blocks s, z, n, or B, and the first matrix should be interpreted as zero if there are no $\underline{s}$-states.

As in the previous subsection, it is useful to study a censored embedded process with only $-$-points and $\uparrow$-points. The next proposition shows how the first-passage matrices of the censored embedded process ($\boldsymbol{\mathcal{P}}$-matrices) as well as the first-passage matrices of the *whole* embedded process ($\mathbf{P}$-matrices) can be found.

Proposition 14.24 *We have*

$$\begin{aligned}
\boldsymbol{\mathcal{P}}^{\alpha}_{--} &= \int_{(0,\infty)} \mathbf{F}^{\alpha}_{\uparrow\underline{s}_{\circlearrowleft}-}(dx) e^{\boldsymbol{\mathcal{Q}}^{\alpha}_{--}x} \operatorname{diag}(\mu^\alpha), \\
\boldsymbol{\mathcal{Q}}^{\alpha}_{--} &= -\operatorname{diag}(\lambda^\alpha)\left[\boldsymbol{I}_{--} - \operatorname{diag}(\mu^\alpha)\boldsymbol{\mathcal{P}}^{\alpha}_{--}\operatorname{diag}(1/\mu^\alpha) - \operatorname{diag}(\mu^\alpha)\mathbf{F}_{\uparrow z_{\circlearrowleft}-}(\alpha,\infty)\right].
\end{aligned}$$

Moreover, $\mathbf{P}^{\alpha}_{--} = \boldsymbol{\mathcal{P}}^{\alpha}_{--}$ *and*

$$\mathbf{P}^{\alpha}_{s-} = \int_{(0,\infty)} \mathbf{F}^{\alpha}_{s_{\circlearrowleft}-}(dx) e^{\boldsymbol{\mathcal{Q}}^{\alpha}_{--}x} \operatorname{diag}(\mu^\alpha).$$

Proof. Consider the censored embedded process. In the notation of Section 14.2, we have $|\mathcal{I}_+| = |\mathcal{I}_-|$, $\boldsymbol{F}_{++}(\alpha,\beta) = \mathbf{0}_{++}$, and $\boldsymbol{F}_{--}(\alpha,\beta) = \mathbf{0}_{--}$, while

$$\boldsymbol{F}_{+-}(\alpha,\beta) = \mathbf{F}_{\uparrow\underline{s}_{\circlearrowleft}-}(\alpha,\beta), \quad \boldsymbol{F}_{-+}(\alpha,\beta) = \operatorname{diag}\left(\frac{\mu^\alpha\lambda^\alpha}{\lambda^\alpha + i\beta}\right).$$

The nonlinear system then follows from Proposition 14.6. The first-passage matrices for the whole embedded process follow readily. □

With the $\mathbf{P}$-matrices and $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$ at our disposal, it is straightforward to find the Laplace transform of $(\underline{X}, \underline{F}^X, \underline{J})$ along the lines of the proof of Theorem 14.17.

Theorem 14.25 *For $\alpha, \beta \geq 0$, we have*

$$\begin{aligned}
\mathsf{E}\left[e^{-\alpha\underline{F}^X + \beta\underline{X}}; \underline{J}\right] &= \operatorname{diag}\begin{pmatrix} \mathbf{1}_s - \mathbf{P}^0_{s-}\mathbf{1}_- \\ \mathbf{1}_z - \mathbf{F}_{z_{\circlearrowleft}-}(0,\infty)\mathbf{1}_- - \mathbf{F}_{z_{\circlearrowleft}s}(0,\infty)\mathbf{P}^0_{s-}\mathbf{1}_- \\ \mathbf{0}_- \end{pmatrix} \\
&\quad + \begin{pmatrix} \mathbf{P}^{\alpha}_{s-}\operatorname{diag}(1/\mu^\alpha) \\ \mathbf{F}_{z_{\circlearrowleft}s}(\alpha,\infty)\mathbf{P}^{\alpha}_{s-}\operatorname{diag}(1/\mu^\alpha) + \mathbf{F}_{z_{\circlearrowleft}-}(\alpha,\infty) \\ \boldsymbol{I}_{--} \end{pmatrix} \left(\beta\boldsymbol{I}_{--} - \boldsymbol{\mathcal{Q}}^{\alpha}_{--}\right)^{-1} \\
&\quad \times \operatorname{diag}(\mu^\alpha\lambda^\alpha)\begin{pmatrix} \mathbf{0}_{-\underline{s}} & \operatorname{diag}\left(\mathbf{1}_- - \mathbf{P}^0_{--}\mathbf{1}_- - \mathbf{F}_{\uparrow z_{\circlearrowleft}-}(0,\infty)\mathbf{1}_-\right) \end{pmatrix}.
\end{aligned}$$

We remark that it is possible to derive a system for $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$ in the spirit of Proposition 14.23. This generalizes the results in Section 5.3 of Miyazawa and Takada [233] and Proposition 2(i) of Pistorius [256].

Indeed, the reader may check that $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$ solves a similar system as in Proposition 14.23, but the place of the matrices $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$ and $\exp\left(\boldsymbol{\mathcal{Q}}^{\alpha}_{--}x\right)$ is different: instead of premultiplied, they should now be postmultiplied. This is in line with the correspondence between Corollaries 14.4 and 14.7.

We conclude this section with a simple relationship between $\boldsymbol{Q}^{\alpha}_{--}$ and $\widehat{\boldsymbol{\mathcal{K}}}^{\alpha}_{--}$, which can be regarded as the analogue of (14.3). The matrix $\widehat{\boldsymbol{\mathcal{K}}}^{\alpha}_{--}$ is defined as $\boldsymbol{\mathcal{K}}^{\alpha}_{--}$, but with the dynamics of the Markov-additive process specified by the time-reversed Laplace exponent $\widehat{\boldsymbol{\psi}}_{-X}$ instead of $\boldsymbol{\psi}_{-X}$. The next lemma formalizes the intuition that the last-passage matrices under the measure $\widehat{\mathsf{P}}$ are closely related to the first-passage matrices under the measure P.

Lemma 14.26 *For $\alpha \geq 0$, we have*

$$\boldsymbol{\mathcal{Q}}^{\alpha}_{--} = \operatorname{diag}\left(\frac{1}{\boldsymbol{\pi}_{\mathsf{J}}(\sim)}\right)\left[\widehat{\boldsymbol{\mathcal{K}}}^{\alpha}_{--}\right]' \operatorname{diag}(\boldsymbol{\pi}_{\mathsf{J}}(\sim)). \tag{14.18}$$

Proof. The matrix $\widehat{\boldsymbol{\mathcal{K}}}^{\alpha}_{--}$ satisfies the system given in Proposition 14.23, but with $\mathbf{F}^{\alpha}_{\uparrow \underline{s}_{\circlearrowleft}\sim}(dx)$ replaced by $\widehat{\mathbf{F}}^{\alpha}_{\uparrow \underline{s}_{\circlearrowleft}\sim}(dx)$ (in self-evident notation; however, the embedding is different!). It can be checked that

$$\widehat{\mathbf{F}}_{\uparrow \underline{s}_{\circlearrowleft}\sim}(\alpha,\beta) = \operatorname{diag}\left(\frac{1}{q_{\sim}\boldsymbol{\pi}_{\mathsf{J}}(\sim)}\right)\mathbf{F}'_{\uparrow \underline{s}_{\circlearrowleft}\sim}(\alpha,\beta)\operatorname{diag}\left(q_{\sim}\boldsymbol{\pi}_{\mathsf{J}}(\sim)\right),$$

and that the matrix on the right-hand side of (14.18) satisfies the same matrix equation as $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$. Uniqueness of its solution proves the claim. □

14.4 The fluid queue: theory

In this section, we use the theory developed in the previous sections to analyze a single fluid queue. Recall from Section 1.1 that this means that work (fluid) arrives at a storage facility, where it is gradually drained; if the input temporarily exceeds the output capacity, then work can be stored in a buffer. As usual, we let $W(t)$ be the buffer content at time t and $B(t)$ be the age of the busy period. The input process A is governed by a background process J.

It is our aim to study the distribution of $(W(t), B(t), \mathsf{J}(t))$ in steady-state, i.e., as $t \to \infty$, for a number of different input processes. We abbreviate $W(\infty)$, $B(\infty)$, and $\mathsf{J}(\infty)$ as W, B, and J respectively; their existence follows from assumptions that we impose later on.

14.4.1 Markov-modulated ON/OFF input

Suppose that the input process corresponds to a single source that is driven by a background process J that switches between N states. The transitions of the background process are governed by an irreducible Markov chain J, defined through the transition probability matrix $\boldsymbol{P}^{J} := \{p^{J}_{jk} : j,k = 1,\ldots,N\}$; the sojourn times in the each of the N states are specified below. Suppose that J and all other random objects in this subsection are defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

If the background process is in state j for $j = 1,\ldots,N-1$, it feeds work into the reservoir at a constant rate $R_j < r$. Since the fluid level decreases during these periods, we call the corresponding states *OFF-states*. The lengths of the sojourn times in these states are all

mutually independent. Moreover, the sojourn time in OFF-state j is exponentially distributed with parameter q_j.

If the source is in state N, the so-called *ON-state*, the source generates work according to a generic stochastic process $\{A_{\rm ON}(t) : t \geq 0\}$. In order to ensure that the buffer content does not decrease (strictly) while the source emits fluid, we suppose that $A_{\rm ON}(t) \geq rt$ for any $t \geq 0$ almost surely. The ON-period is terminated after some period distributed as the generic random variable $\mathtt{k} > 0$ ('killing time'), independent of $A_{\rm ON}$. After this ON-period, $\mathtt{J}$ always makes a transition to an OFF-state (i.e., J has no self-transitions in state N). We suppose that $\mathtt{k}$ is integrable. In principle, the probability distribution governing the transitions to OFF-states may depend on (the whole trajectory of) $A_{\rm ON}$ and $\mathtt{k}$, but we suppose for simplicity that this is not the case. The ON-periods are mutually independent, and also independent of the OFF-periods.

We emphasize the versatility of this Markov-modulated ON/OFF model. For instance, by redefining ON-periods, it is possible to incorporate multiple (and distributionally different) subsequent ON-periods. Moreover, as observed by Cohen [78] in a special case of our model, the superposition of a number of (independent) ON/OFF sources can be regarded as a single source. Indeed, one considers the aggregate of the sources, which has the same structure as a single ON/OFF source, but with ON-periods corresponding to so-called inflow periods (i.e., periods in which at least one of the sources is in the ON-state). This shows that, if for each source it holds that it emits work at a rate of at least r while ON, then we may restrict our attention to the single-source model. We mention that our formulation of the model has been inspired by work of Kella and Whitt [185] and Scheinhardt and Zwart [284], who consider the (more specific) situation of a single OFF-state and strictly alternating ON-periods and OFF-periods. A closely related model has been recently examined by Boxma *et al.* [60].

To characterize the distribution of $(W, B, \mathtt{J})$, we use an embedding and the theory from Section 14.2. Let $\mathtt{k}^*$ be distributed as the elapsed time that the source is ON, if we observe the system in steady state in an ON-state. That is, it has the integrated-tail distribution

$$\mathbb{P}(\mathtt{k}^* > y) = \frac{1}{\mathbb{E}\mathtt{k}} \int_y^\infty \mathbb{P}(\mathtt{k} > x) dx,$$

where $y \geq 0$. We also need the expected sojourn time *between* ON-states, $\mathbb{E}V_{\rm OFF}$. Standard formulas for moments of phase-type distributions show that

$$\mathbb{E}V_{\rm OFF} = \boldsymbol{P}^J_{N-} \left(\boldsymbol{I}_{--} - \boldsymbol{P}^J_{--}\right)^{-1} \mathrm{vec}\left(\frac{1}{q_-}\right),$$

where the beginnings of the OFF-sojourn times and ON-sojourn times are labeled as $-$-points and $+$-points respectively, as in Section 14.2. The quantity $\mathbb{E}V_{\rm OFF}$ plays an important role for the probability p_k that the source is in state k when the system is in steady state. For $k = 1, \ldots, N-1$, we find that

$$p_k = \frac{\mathbb{E}V_{\rm OFF}}{\mathbb{E}V_{\rm OFF} + \mathbb{E}\mathtt{k}} \frac{\pi_J(k)}{\boldsymbol{\pi}_J(-)' \,\mathrm{vec}(q_k/q_-)},$$

and $p_N = \mathbb{E}\mathtt{k}/(\mathbb{E}V_{\rm OFF} + \mathbb{E}\mathtt{k})$. The stability condition of this model is

$$\frac{\mathbb{E}A_{\rm ON}(\mathtt{k})}{\mathbb{E}V_{\rm OFF} + \mathbb{E}\mathtt{k}} + \boldsymbol{R}'_- \boldsymbol{p}_- < r.$$

We write $\widehat{\boldsymbol{P}}^J = \{\widehat{p}^J_{jk} : j, k = 1, \ldots, N\}$ for the time-reversed transition matrix of the Markov process J, and we define $\widehat{\mathbb{P}}$ such that (S, T, J) has the transition kernel

$$\widehat{p}((s,t,j),(s+dv,t+dw,k)) = \begin{cases} \widehat{p}^J_{jk}\mathbb{P}(U \in dv, \sigma \in dw) & \text{if } j = N \text{ and } k = 1, \ldots, N; \\ \widehat{p}^J_{jk}\mathbb{P}\big(-D^j \in dv, \tau^j \in dw\big) & \text{if } j = 1, \ldots, N-1 \text{ and } k = 1, \ldots, N, \end{cases}$$

with

$$\mathbb{E}e^{-\alpha\sigma-\beta U} = \mathbb{E}\left[e^{-\alpha\mathtt{k}-\beta[A_{\rm ON}(\mathtt{k})-r\mathtt{k}]}\right], \quad \mathbb{E}e^{-\alpha\tau^j-\beta D^j} = \frac{q_j}{q_j+\alpha+\beta(r-R_j)}.$$

We next express the distribution of (W, B, J) in terms of the distribution of (S, T).

Proposition 14.27 *For $k = 1, \ldots, N-1$, $\omega, \beta \geq 0$, we have*

$$\mathbb{E}\left[e^{-\omega W-\beta B}; \mathsf{J} = k\right] = p_k \widehat{\mathbb{E}}_k e^{-\omega\overline{S}-\beta\overline{T}},$$

and

$$\mathbb{E}\left[e^{-\omega W-\beta B}; \mathsf{J} = N\right] = p_N \mathbb{E}\left[e^{-(\beta-\omega r)\mathtt{k}^*-\omega A_{\rm ON}(\mathtt{k}^*)}\right] \widehat{\boldsymbol{P}}^J_{N-}\widehat{\boldsymbol{E}}_- e^{-\omega\overline{S}-\beta\overline{T}}.$$

This proposition can be proven with regenerative-processes theory [19, Ch. VI]. The construction borrows its key elements from Theorem 4 in [185]. Specializing to just W, it relies on two principles:

- The classical Reich formula says that W is distributed as $\sup_{t\geq 0} -\widehat{A}(-t) - rt$, with $\widehat{A}$ being the version of the input process A with stationary increments and with time indexed by $\mathbb{R}$. That is, $-\widehat{A}(-t)$ can be thought of as the work generated in the interval $[-t, 0]$ given that the system started in steady state at time $-\infty$ (it is then always in steady state, in particular at time zero). This entails that the process $-\widehat{A}(-t) - rt$ (thus looking *backward* in time!) needs to be analyzed.
- To construct $\sup_{t\geq 0} -\widehat{A}(-t) - rt$, the state of the background process at time zero is sampled from $\boldsymbol{p}$. Two possibilities arise.
  - The initial state is N. The background process stays in this state for a period that has the integrated-tail distribution of $\mathtt{k}$; the increment is $A_{\rm ON}(\mathtt{k}^*) - r\mathtt{k}^* \geq 0$. The next state, say j, is sampled from $\widehat{\boldsymbol{P}}^J_{N-}$, and the process $\{-\widehat{A}(-t-\mathtt{k}^*) - rt - r\mathtt{k}^* : t \geq 0\}$ behaves exactly in the same way as the process $\{A(t) - rt\}$ with initial state j (independently of the initial increment), except for the following two changes. The background states are chosen according to the time-reversed probabilities $\{\widehat{p}^J_{jk}\}$, and the trajectories during ON-periods are 'reversed'. Still, the distribution of the increment during such a period remains the same. As a result, the embedded process, which is governed by the kernel $\widehat{p}$, can be used to express the remaining contribution to the supremum.
  - The initial state is $k = 1, \ldots, N-1$. It stays in this initial state for a period that has the integrated-tail distribution of τ^k, which is again exponential with parameter q_k; as a consequence we could do as if the background process had just jumped to k at time zero. The supremum can thus immediately be expressed in terms of the time-reversed embedded process.

We emphasize that the distribution of $(W, B, \mathsf{J}) \equiv (W(\infty), B(\infty), \mathsf{J}(\infty))$ is not affected by the initial state of the system. More precisely, the steady-state solution is independent of the buffer content $W(0)$ at time zero, the state of the background process $\mathsf{J}(0)$ at that epoch, and the time spent already in this state before time zero.

The $\widehat{\mathbb{E}}_k e^{-\omega\overline{S}-\beta\overline{T}}$ for $k = 1, \ldots, N-1$ can be found as explained in Section 14.2. Hence, in order to use the above theorem, an expression for the transform of $(\mathtt{k}^*, A_{\rm ON}(\mathtt{k}^*))$ is needed; from Scheinhardt and Zwart [284] we have

$$\mathbb{E}\left[e^{-\alpha\mathtt{k}^*-\beta A_{\rm ON}(\mathtt{k}^*)}\right] = \frac{1}{\mathbb{E}\mathtt{k}}\,\mathbb{E}\left[\int_0^{\mathtt{k}} e^{-\alpha t-\beta A_{\rm ON}(t)}dt\right]. \tag{14.19}$$

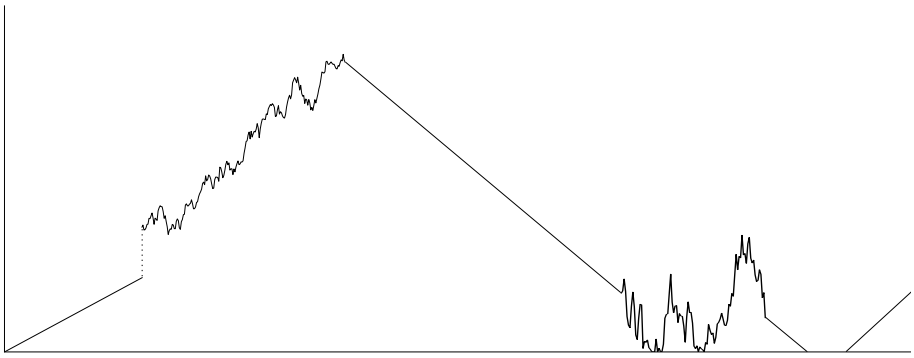

Figure 14.2: A realization of W.

For the special case of A_{ON} being just a drift, i.e., $A_{\mathrm{ON}}(t) = R_N t$ almost surely for some $R_N > r$, it evidently reads

$$\mathbb{E}[e^{-(\alpha+\beta R_N)\mathtt{k}^*}] = \frac{1 - \mathbb{E}[e^{-(\alpha+\beta R_N)\mathtt{k}}]}{(\alpha + \beta R_N)\mathbb{E}\mathtt{k}}.$$

When specialized to the distribution of W and using (14.19), Proposition 14.27 reduces to

$$\mathbb{E}e^{-\omega W} = \left(\boldsymbol{p}'_- + \frac{p_N}{\mathbb{E}\mathtt{k}}\,\mathbb{E}\left[\int_0^{\mathtt{k}} e^{-\omega[A_{\mathrm{ON}}(t)-rt]}dt\right]\widehat{\boldsymbol{P}}^J_{N-}\right)\widehat{\boldsymbol{E}}_- e^{-\omega\overline{S}}.$$

In Boxma *et al.* [60], a similar expression has been interpreted as a decomposition of W in terms of a clearing process and an independent dam process.

14.4.2 Markov-additive input

In this subsection, we suppose that there is an irreducible Markov process J such that (A, J) is a Markov-additive process on some probability space $(\Omega', \mathcal{F}', \mathsf{P})$. We define $X(t) := A(t) - rt$, the *free process*. Clearly, (X, J) is a Markov-additive process as well. Even though Proposition 14.28 below holds in much greater generality, we suppose throughout that X does not have negative jumps. Consequently, this subsection relies extensively on Theorem 14.19. We do not analyze the spectrally negative case, but it could be analyzed with Theorem 14.25; further details can be found in Miyazawa and Takada [233].

In Figure 14.2, we have plotted a possible realization of the process W. Note that there are Brownian states, subordinator states, and negative-drift states.

We now establish the precise relationship between the buffer-content process and extremes of the free process, which follows from the reasoning in Section II.3 and Section VI.7 of Asmussen [18]; see also Section 4 of Miyazawa and Takada [233]. Again, $(B(0), W(0), \mathsf{J}(0)$ does not have influence on the behavior of $(B(t), W(t), \mathsf{J}(t))$ as $t \to \infty$, a property that is intuitively clear. The result follows by the same arguments as those used for Markov-modulated ON/OFF input, but no 'residual' (or 'clearing-model') quantities are needed since the sojourn times of J are exponential. We write $\widehat{\mathsf{P}}_k$ for the law of the Markov-additive process (X, J) with $\mathsf{J}(0) = k$ and Laplace exponent $\widehat{\boldsymbol{\psi}}_{-X}$ defined in (14.15).

Proposition 14.28 *Suppose that $\boldsymbol{\pi}'_{\mathsf{J}}\mathbf{E}X(1) < 0$. Then (W, B) is a finite random vector, and for any $\omega, \beta \geq 0$, $k = 1, \ldots, N$, we have*

$$\mathsf{E}\left[e^{-\omega W - \beta B}; \mathsf{J} = k\right] = \pi_{\mathsf{J}}(k)\widehat{\mathsf{E}}_k e^{-\beta\overline{F}^X - \omega\overline{X}}.$$

We now work out the preceding proposition for the distribution of (W, J), since the resulting formula is particularly appealing. Corollary 14.21 shows that for $\omega \geq 0$, provided $\widehat{\psi}_{-X}(\omega)$ is nonsingular,

$$\operatorname{diag}(\boldsymbol{\pi}_{\mathsf{J}})\widehat{\mathsf{E}}e^{-\omega\overline{X}} = \omega \operatorname{diag}(\boldsymbol{\pi}_{\mathsf{J}})\widehat{\psi}_{-X}^{-1}(\omega) \begin{pmatrix} \mathbf{0}_{\underline{s}} \\ \widehat{\boldsymbol{v}}_{-} \end{pmatrix} = \omega \left[\boldsymbol{\psi}'_{-X}(\omega)\right]^{-1} \begin{pmatrix} \mathbf{0}_{\underline{s}} \\ \boldsymbol{u}_{-} \end{pmatrix},$$

where we set $\boldsymbol{u}_{-} := \boldsymbol{\pi}_{\mathsf{J}}(\sim) \circ \widehat{\boldsymbol{v}}_{-}$ (recall that $\sim$-states stand for n-states and B-states). The vector $\widehat{\boldsymbol{v}}_{-}$ is defined in the same way as the vector $\boldsymbol{v}_{-}$-vector, but with P replaced by $\widehat{\mathsf{P}}$. With Proposition 14.28, this leads immediately to the identity

$$\mathsf{E}\left[e^{-\omega W}; \mathsf{J}\right] \boldsymbol{\psi}_{-X}(\omega) = \omega \begin{pmatrix} \mathbf{0}'_{\underline{s}} & \boldsymbol{u}'_{-} \end{pmatrix} \tag{14.20}$$

for arbitrary $\omega \geq 0$. This formula is Equation (4.1) of Asmussen and Kella [24], who interpret $\boldsymbol{u}_{-}$ in terms of local times. The following observation, however, is new. By combining Lemma 14.26 with Lemma 14.20, it readily follows that $\boldsymbol{u}_{-}$ must be a left eigenvector of $\boldsymbol{\mathcal{Q}}^{0}_{--}$ (corresponding to the simple eigenvalue zero); this uniquely determines $\boldsymbol{u}_{-}$ up to a constant. This constant can be found by writing down the formula for $\mathsf{E}e^{-\omega W}$ from (14.20), using $\mathbf{1} = \mathbf{P}^{\mathsf{J}}\mathbf{1}$, and letting $\omega \to 0$ in the resulting expression.

Motivated by Proposition 14.28, we next characterize the $\widehat{\mathsf{P}}$-distribution of $(\overline{X}, \overline{F}^{X}, \overline{\mathsf{J}})$ (the last component is not required here, but it is needed in Section 14.6). To avoid the introduction of yet more matrices, we suppose that there are no zero-drift states. The following result then follows immediately from Theorem 14.19 and Lemma 14.26.

Corollary 14.29 *Suppose that $\boldsymbol{\pi}'_{\mathsf{J}}\mathsf{E}X(1) < 0$ and that there are no zero-drift states. We then have for $\alpha, \beta \geq 0$,*

$$\left(\boldsymbol{\psi}'_{-X}(\beta) - \alpha\boldsymbol{I}\right) \operatorname{diag}(\boldsymbol{\pi}_{\mathsf{J}})\widehat{\mathsf{E}}\left[e^{-\alpha\overline{F}^{X} - \beta\overline{X}}; \overline{\mathsf{J}}\right] = \begin{pmatrix} \mathbf{0}_{ss} & \mathbf{0}_{s-} \\ \mathbf{0}_{-s} & \left(\beta\boldsymbol{I}_{--} + \left[\boldsymbol{\mathcal{Q}}^{\alpha}_{--}\right]'\right) \operatorname{diag}(\boldsymbol{u}_{-}) \end{pmatrix}.$$

In conclusion, if X is spectrally positive, the matrix $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$ plays a similar role for the steady-state buffer-content process as the matrix $\boldsymbol{\mathcal{K}}^{\alpha}_{--}$ for the maximum of the free process.

14.5 The single queue: examples

Many known models can be incorporated into the framework of the preceding section. To emphasize the versatility of our framework, we now give some examples.

The M/G/1 queue

Consider a single-server queue with Poisson arrivals at rate λ and i.i.d. service requirements; a generic service requirement is denoted by U. Throughout, we assume stability, i.e., $\lambda\mathbb{E}U < 1$. A standard result ('PASTA') for the M/G/1 queue is that the steady-state buffer content ('virtual waiting time') has the same distribution as the steady-state waiting time under the First-In-First-Out discipline, see, e.g., [19, Cor. II.9.2]. In turn, these quantities have the same distribution as the maximum of a Lévy process X with unit negative drift and jumps distributed as U.

Two approaches are possible to derive this distribution. The first relies on Proposition 14.28 and the remarks thereafter. Since the free process X is a Lévy process with unit negative drift and jumps distributed as U, it immediately yields the desired Pollaczek-Khinchine formula

$$\mathbb{E}e^{-\beta\overline{X}} = \frac{\beta(1 - \lambda\mathbb{E}U)}{\beta - \lambda\left(1 - \mathbb{E}e^{-\beta U}\right)}. \tag{14.21}$$

To gain some intuition for the embedding technique, it is insightful to give a second derivation of this formula based on the results in Section 14.2. Essentially, we are interested in the maximum of a discrete-time Markov-additive process S, for which the modulating part takes values in $\{+,-\}$. The process S increases by amounts that are distributed as U (corresponding to a $+$-point) and decreases by exponentially(λ) distributed amounts (corresponding to a $-$-point). This is summarized as

$$\boldsymbol{F}(0,\beta) = \begin{pmatrix} 0 & \mathbb{E}e^{i\beta U} \\ \frac{\lambda}{\lambda+i\beta} & 0 \end{pmatrix}.$$

Note that if the maximum is attained, the background process must be in a $-$-point.

First note that $\mathbb{P}_-(\overline{S}=0) = 1-\lambda\mathbb{E}U$ by Lemma 14.13. Some further straightforward computations based on Corollary 14.11 show that $\mathbb{E}_-[e^{-\beta\overline{S}};\overline{J}=-]$ indeed equals the right-hand side of (14.21).

We next analyze the length of the busy period. Note that the Laplace transform of the length of a busy cycle (which consists of a busy period and an idle period) is $P^\alpha_{+-} = \mathbb{E}_+[e^{-\alpha T_{\tau_-}}; J_{\tau_- -1} = -]$. By the Markov property, these two parts of a busy cycle are independent, and the part corresponding to the idle period has Laplace transform $\mu^\alpha = \lambda/(\lambda+\alpha)$. As a result, the transform of the length of the busy period is given by $G^\alpha := P^\alpha_{+-}/\mu^\alpha + \mathbb{P}(U=0)$ for $\alpha \geq 0$. Since $\lambda^\alpha = \lambda+\alpha$, the recursion of Proposition 14.6 reduces in the present setting to

$$P^\alpha_{+-} = \frac{\lambda}{\lambda+\alpha}\mathbb{E}\left[e^{-(\lambda+\alpha)[1-P^\alpha_{+-}-\mu^\alpha\mathbb{P}(U=0)]U}; U>0\right]. \tag{14.22}$$

This shows that G^α satisfies Takács' fixed-point equation

$$G^\alpha = \mathbb{E}e^{-[\alpha+\lambda(1-G^\alpha)]U}, \tag{14.23}$$

in accordance with, for instance, Prabhu [263, Thm. 1.37].

The BMAP/G/1 queue

The BMAP/G/1 queue is a generalization of the M/G/1 queue. Here BMAP is shorthand for *batch Markovian arrival process*. Special cases include the MMPP/G/1 queue, where MMPP stands for *Markov modulated Poisson process*, and the PH/G/1 queue, where PH stands for *phase-type renewal process*. For further special cases, we refer to Latouche and Ramaswami [210, Sec. 3.5]. The BMAP/G/1 queue has been studied in detail by Lucantoni [221], and it is our present aim to relate his results to ours. This is particularly relevant since our notation does not always agree with the standard notation in the matrix-analytic literature as used in [221].

The virtual waiting time in a BMAP/G/1 queue is defined as the buffer content in a fluid queue with special Markov-additive input; we describe this below. More precisely, as observed by Tzenova *et al.* [297], the BMAP/G/1 queue can be viewed as a fluid-flow model with jumps (fluid-flow models are discussed below). It is important to note that the setting of Section 14.4.2 can therefore be used.

In a BMAP/G/1 queue, the arrival process is governed by a Markovian background process J that can take $N<\infty$ values. The sojourn time of J in state j has an exponential distribution with parameter q_j. At the end of a sojourn time in state j, with probability $p^{(n)}_{jk}$, $n \geq 0$ customers arrive (that all bring in a generic amount of work $U>0$) and a transition of J to state k occurs. These transition probabilities satisfy $\sum_{n=0}^\infty \sum_{k=1}^N p^{(n)}_{jk} = 1$ for $j = 1,\ldots,N$. We write H for the distribution of U, and the stationary distribution of J is denoted by $\boldsymbol{\pi}_{\mathsf{J}}$ as usual.

Let us now define the free process X such that (X, J) becomes a Markov-additive process. Since the amount of work in the system decreases at unit rate, it readily follows that the Laplace exponent of X is given by

$$\boldsymbol{\psi}_{-X}(\beta) = \beta \boldsymbol{I} - \operatorname{diag}(q) \left(\boldsymbol{I} - \sum_{n=0}^{\infty} \boldsymbol{P}^{(n)} \left[\mathsf{E} e^{-\beta U} \right]^n \right), \tag{14.24}$$

where $\boldsymbol{P}^{(n)}$ is the matrix with elements $p_{jk}^{(n)}$. We suppose that the system is stable, i.e., $\boldsymbol{\pi}'_{\mathsf{J}} \mathbf{E} X(1) < 0$. It is an immediate consequence of Proposition 14.28 and the remarks thereafter that

$$\mathsf{E} e^{-\omega W} = \omega \boldsymbol{u}'_{-} \boldsymbol{\psi}_{-X}^{-1}(\omega) \mathbf{1},$$

at least for $\omega \geq 0$ for which the matrix $\boldsymbol{\psi}_{-X}(\omega)$ is nonsingular. This formula, in the present context due to Ramaswami, is Equation (45) in [221]. In the matrix-analytic literature, it is customary to use the notation $\boldsymbol{y}_0$ for $\boldsymbol{u}_{-}$. Note that we have shown in Section 14.4.2 that $\boldsymbol{u}'_{-} \boldsymbol{\mathcal{Q}}^0_{--} = \mathbf{0}'_{-}$.

This motivates the investigation of the matrix $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$ for $\alpha \geq 0$. We have argued in Section 14.3.4 that $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$ satisfies the 'postmultiplication version' of (14.24), as opposed to the 'premultiplication version' in Proposition 14.23. That is, we have

$$\boldsymbol{\mathcal{Q}}^{\alpha}_{--} + \alpha \boldsymbol{I} = -\operatorname{diag}(q) \left(\boldsymbol{I} - \sum_{n=0}^{\infty} \boldsymbol{P}^{(n)} \int_{[0,\infty)} e^{\boldsymbol{\mathcal{Q}}^{\alpha}_{--} x} H^{(n)}(dx) \right),$$

with $H^{(n)}(dx)$ denoting the n-fold convolution of $H(dx)$. Upon setting

$$\boldsymbol{G}^{\alpha} := \int_{[0,\infty)} e^{\boldsymbol{\mathcal{Q}}^{\alpha}_{--} x} H(dx), \tag{14.25}$$

the fixed-point system for $\boldsymbol{\mathcal{Q}}^{\alpha}$ reduces to

$$\boldsymbol{\mathcal{Q}}^{\alpha}_{--} + \alpha \boldsymbol{I} = -\operatorname{diag}(q) \left(\boldsymbol{I} - \sum_{n=0}^{\infty} \boldsymbol{P}^{(n)} \left[\boldsymbol{G}^{\alpha} \right]^n \right).$$

Substitution of this expression in (14.25) leads to a fixed-point system for $\boldsymbol{G}^{\alpha}$:

$$\boldsymbol{G}^{\alpha} = \int_{[0,\infty)} e^{-\alpha x} e^{-\operatorname{diag}(q)\left(\boldsymbol{I} - \sum_{n=0}^{\infty} \boldsymbol{P}^{(n)} [\boldsymbol{G}^{\alpha}]^n\right) x} H(dx),$$

which is the matrix version of (14.23) if $\boldsymbol{P}^{(1)}$ is the only nonzero matrix in the sequence $\{\boldsymbol{P}^{(n)} : n \geq 0\}$. Based on this formula, Lucantoni [221] gives an algorithm that serves as an efficient alternative for Neuts' approach to M/G/1-type queueing systems [242]. Importantly, it is not necessary to compute $\boldsymbol{\mathcal{Q}}^0_{--}$ in order to find $\boldsymbol{u}_{-}$: the definition of $\boldsymbol{G}^0$ in (14.25) shows that $\boldsymbol{u}_{-}$ is necessarily proportional to the unique probability vector $\boldsymbol{g}$ satisfying $\boldsymbol{g}' \boldsymbol{G}^0 = \boldsymbol{g}'$. The normalizing constant is found as in Section 14.4.2.

The G/M/1 queue and its ramifications

Let us now suppose that the interarrival times have a general distribution (the generic service time is written as U), while the service requirements are exponentially distributed (say with rate λ). Throughout, we assume stability, i.e., $\lambda \mathbb{E} U > 1$.

Again, we are interested in the waiting-time distribution. As opposed to the M/G/1 queue, however, the buffer-content process cannot be represented by a (reflected) Markov-additive process. Still, by exploiting a connection with random walks, we can use the results of Section 14.2

to find the waiting-time distribution. Similar arguments can be applied to the G/PH/1 queue. We rely on the fact that the steady-state waiting time is equal in distribution to $-\underline{S}$ if $S_0 = +$, where the process S is the process that we defined in the context of M/G/1 queues.

It is interesting to see that we have essentially found the waiting-time distribution while examining the M/G/1 queue in detail. Note that we know from Proposition 14.6 that $Q_{--} = -\lambda(1 - P^0_{-+} - \mathbb{P}(U = 0))$. Therefore, Theorem 14.17 immediately yields

$$\mathbb{E}_+ e^{\beta \underline{S}} = \left(1 + \frac{\lambda(P^0_{-+} + \mathbb{P}(U=0))}{\beta - Q_{--}}\right)\left[1 - P^0_{-+} - \mathbb{P}(U=0)\right] = \frac{-Q_{--}}{\lambda} + \frac{\lambda + Q_{--}}{\lambda}\frac{-Q_{--}}{\beta - Q_{--}}.$$

This shows that the waiting-time distribution has an atom at zero with mass $-Q_{--}/\lambda$, and that its density over $(0, \infty)$ is proportional to the density of an exponential distribution with parameter $-Q_{--}$. To relate this to well-known formulas for $-Q_{--}$, note that (14.22) implies that

$$\frac{\lambda}{\lambda + Q_{--}} \mathbb{E} e^{Q_{--}U} = 1,$$

which is consistent with (for instance) Theorem VIII.5.8 of Asmussen [19].

The same result can also be obtained with the theory of Markov-additive processes developed in Section 14.4.2, and even certain many-server queues are covered by the results. One then needs to look at the problem in a slightly different way, due to Sengupta; see [19, Sec. XI.3d]. The Sengupta approach leads to the waiting-time distributions in the G/PH/s and MAP/PH/s queue, as detailed by Asmussen and Møller [28].

Fluid-flow models

A fluid-flow model is a fluid queue with a special type of Markov-additive input: the free process X is not allowed to have jumps nor Brownian states. They constitute undoubtedly the most well-studied fluid queues; we do not attempt to give a full bibliography, but refer to [16, 203, 274] for more details.

Recently, there has been some interest in deriving the Laplace transform of the busy period in fluid-flow models [6, 37]; see also [15] for an earlier contribution. It is our present aim to show how some of the main results are reproduced in our general theory. We remark that we allow states with zero drifts.

Even though fluid models are special Markov-additive processes, we shall work within the framework of Section 14.2 to derive formulas that are familiar from the fluid-flow literature. To facilitate the use of our discrete-time results, we use an embedding that records the time and position at the *beginning* of a sojourn time of the underlying background process J. In self-evident notation, we partition the state space into $+$-points, 0-points, and $-$-points. The intensity matrix of J is written as $\mathbf{Q}^{\mathsf{J}}$; this also defines $\mathbf{Q}^{\mathsf{J}}_{++}$, for instance.

Let $\mathbf{\Psi}^{\alpha}_{+-}$ be the matrix with the transforms of the busy-period lengths. That is, if $c_j > 0$ and $c_k < 0$, then element (j, k) of this matrix is the Laplace transform of the length of the first positive excursion of X on the event that it ends this excursion in state k. In other words, it corresponds to the amount of time that X spends above zero on the event that it starts in state j and it first hits zero in state k.

Let us use the notation $\mathrm{vec}(c_+)$ and $\mathrm{vec}(c_-)$ for the vector of strictly positive and strictly negative drifts respectively. We also set $\mu^{\alpha}_{\pm} := \mathrm{diag}(q_{\pm}/(q_{\pm} + \alpha))$, $\lambda^{\alpha}_{\pm} := \mathrm{diag}((q_{\pm} + \alpha)/c_{\pm})$, and

$$\begin{aligned}
\mathbf{T}^{\alpha}_{\pm\pm} &:= \pm \mathrm{diag}\left(\frac{1}{c_{\pm}}\right)\left[\mathbf{Q}^{\mathsf{J}}_{\pm\pm} - \alpha \boldsymbol{I}_{\pm\pm} - \mathbf{Q}^{\mathsf{J}}_{\pm 0}(\mathbf{Q}^{\mathsf{J}}_{00} - \alpha \boldsymbol{I}_{00})^{-1}\mathbf{Q}^{\mathsf{J}}_{0\pm}\right], \\
\mathbf{T}^{\alpha}_{\pm\mp} &:= \pm \mathrm{diag}\left(\frac{1}{c_{\pm}}\right)\left[\mathbf{Q}^{\mathsf{J}}_{\pm\mp} - \mathbf{Q}^{\mathsf{J}}_{\pm 0}(\mathbf{Q}^{\mathsf{J}}_{00} - \alpha \boldsymbol{I}_{00})^{-1}\mathbf{Q}^{\mathsf{J}}_{0\mp}\right].
\end{aligned}$$

Note that, in the notation of Section 14.2, we are interested in $\mathbf{\Psi}^{\alpha}_{+-} = \boldsymbol{P}^{\alpha}_{+-}\,\mathrm{diag}(1/\mu^{\alpha})$. As in the proof of Corollary 14.4, we consider a sequence of +- and 0-points as a single +-point, so that $\boldsymbol{F}_{+-}(\alpha,\beta) = (\beta\boldsymbol{I}_{++} - \mathbf{T}^{\alpha}_{++})^{-1}\mathbf{T}^{\alpha}_{+-}$. Then Proposition 14.6 immediately yields that

$$\mathbf{\Psi}^{\alpha}_{+-} = \int_{(0,\infty)} e^{\mathbf{T}^{\alpha}_{++}x}\mathbf{T}^{\alpha}_{+-}e^{\boldsymbol{\mathcal{Q}}^{\alpha}_{--}x}dx,$$

where $\boldsymbol{\mathcal{Q}}^{\alpha}_{--} = \mathbf{T}^{\alpha}_{--} + \mathbf{T}^{\alpha}_{-+}\mathbf{\Psi}^{\alpha}_{+-}$. Since the eigenvalues of $\mathbf{T}^{\alpha}_{++}$ have a strictly negative real part and those of $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$ have a nonpositive real part, the integral in the above representation for $\mathbf{\Psi}^{\alpha}_{+-}$ converges. This implies the identity (see Bean *et al.* [36] for references)

$$\mathbf{T}^{\alpha}_{++}\mathbf{\Psi}^{\alpha}_{+-} + \mathbf{\Psi}^{\alpha}_{+-}\boldsymbol{\mathcal{Q}}^{\alpha}_{--} = -\mathbf{T}^{\alpha}_{+-}.$$

After some rearranging and substitution of $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$, we obtain the matrix equation

$$\mathbf{T}^{\alpha}_{+-} + \mathbf{\Psi}^{\alpha}_{+-}\mathbf{T}^{\alpha}_{-+}\mathbf{\Psi}^{\alpha}_{+-} + \mathbf{T}^{\alpha}_{++}\mathbf{\Psi}^{\alpha}_{+-} + \mathbf{\Psi}^{\alpha}_{+-}\mathbf{T}^{\alpha}_{--} = \mathbf{0}_{+-},$$

which is Theorem 1 of Bean *et al.* [37] and, for $\alpha = 0$, Theorem 2 of Rogers [274]. Note that no drift condition was imposed to derive this equation.

Importantly, the theory of Section 14.4.2 shows that the matrix $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$ is a key quantity for fluid-flow models. For instance, under a stability assumption, a left eigenvector of $\boldsymbol{\mathcal{Q}}^{0}_{--}$ (corresponding to the simple eigenvalue zero) appears in the representation of W as a phase-type distribution. The matrix $\boldsymbol{\mathcal{Q}}^{\alpha}_{--}$ plays a prominent role in many system characteristics of fluid queues, see also Section 14.7.

M/M/∞-driven fluid queues

Although it was assumed that the state space of the background process be finite, we now give an example with a countably infinite state space that still fits into our framework. The model is a fluid-flow model, but we show that we can translate it in terms of the queue with Markov-modulated ON/OFF input of Section 14.4.1.

Consider the following queueing model. A buffer is emptied at a constant service rate r, and jobs arrive according to a Poisson process (with rate λ). They stay active for an exponentially distributed period of time (without loss of generality, we set its mean equal to 1); while active they feed work into the buffer at unit rate. Notice that the number of (active) jobs in the system follows an M/M/∞-model, therefore it has a Poisson distribution with mean λ; denote $p_k := e^{-\lambda}\lambda^k/k!$. This leads to the stability condition $\lambda < r$.

The buffer level increases when the number of active jobs exceeds r, whereas the buffer is drained (or remains empty) when the number of jobs is below r. Let $X(t)$ denote the free process at time t as before, and let $N(t)$ the number of active flows at time t. For ease we assume that $r \notin \mathbb{N}$; $r_- := \lfloor r \rfloor$ and $r_+ := \lceil r \rceil$. Define for $\ell > r$

$$\sigma_\ell := \inf\{t \geq 0 : N(t) = r_- \mid N(0) = \ell\}, \quad U_\ell := X(\sigma_\ell).$$

An explicit formula for $\xi_\ell(\alpha,\beta) := \mathbb{E}[e^{-\alpha\sigma_\ell - \beta U_\ell}]$ is provided by Preater [264].

Due to exponentiality and reversibility properties, we have that the steady-state buffer content W is distributed as $\sup_{t\geq 0} X(t)$. To study this supremum, it suffices to consider an embedding. One embedding could be the position of the free process at epochs jobs arrive and leave, but this has drawback that the dimension of the background process is (countably) infinite. Evidently, we could alternatively opt for the 'sparser' embedding that lumps together the states $r_+, r_+ + 1, \ldots$ into state r_+; the supremum of the embedded process coincides with the supremum of the full free process. Then the sojourn time in state $k = 0, \ldots, r_-$ is exponential with parameter $\lambda + k$, whereas the Laplace transform of the time spend in

r_+, jointly with the net amount of work generated, is $\xi_{r_+}(\alpha,\beta)$. With $q_j := \lambda + j$, it is easy to verify that corresponding discrete-time Markov chain on $\{0,\ldots,r_+\}$ has the following transition probabilities: $p^J_{j,j+1} = \lambda/q_j$, if $j = 0,\ldots,r_-$; $p^J_{j,j-1} = j/q_j$, if $j = 1,\ldots,r_-$; $p^J_{r_+,r_-} = 1$; $p^J_{jk} = 0$, otherwise. Define $\mathbb{P}$ such that (S,T,J) has the transition kernel

$$p((s,t,j),(s+dv,t+dw,k)) = \begin{cases} p^J_{jk}\mathbb{P}(U \in dv, \sigma \in dw) & \text{if } j = r_+ \text{ and } k = 0,\ldots,r_+; \\ p^J_{jk}\mathbb{P}\big(-D^j \in dv, \tau^j \in dw\big) & \text{if } j = 0,\ldots,r_- \text{ and } k = 0,\ldots,r_+, \end{cases}$$

with

$$\mathbb{E}e^{-\alpha\sigma-\beta U} = \mathbb{E}e^{-\alpha\sigma_{r_+}-\beta U_{r_+}} = \xi_{r_+}(\alpha,\beta), \quad \mathbb{E}e^{-\alpha\tau^j-\beta D^j} = \frac{q_j}{q_j+\alpha+\beta(r-j)}.$$

A procedure analogous to that for Markov-modulated ON/OFF input now yields for $k = 0,\ldots,r_-$ and $\omega,\beta \geq 0$,

$$\mathbb{E}\left[e^{-\omega W-\beta B}; \mathsf{J} = k\right] = p_k \mathbb{E}_k e^{-\omega\overline{S}-\beta\overline{T}},$$

and

$$\mathbb{E}\left[e^{-\omega W-\beta B}; \mathsf{J} = r_+\right] = \left[\sum_{k=r_+}^{\infty} p_k \xi_k(\alpha,\beta)\right] \mathbb{E}_{r_-} e^{-\omega\overline{S}-\beta\overline{T}}.$$

14.6 Tandem networks with Markov-additive input

In this section, we consider the single queueing systems of Section 14.4 in a network setting. Networks with Markov-modulated ON/OFF input [284] or Markov-additive input [184] can be studied with nearly the same techniques, so we choose to only continue the investigation of queueing systems with Markov-additive input.

Even though our framework offers an appealing approach to such networks, we do not strive for the greatest possible generality. Instead, we only give the main ideas without proofs, since the results can be proven along the lines of Chapter 13. Several extensions are discussed in the next section.

One of the simplest networks is a tandem network, in which n fluid reservoirs are lined up in series. Queue j is drained at rate r_j as long as there is content in buffer j. After fluid is released from queue j, it immediately flows to queue $j+1$, unless $j = n$; then it leaves the system.

We suppose that the input to the first queue is governed by the same Markov-additive process (A,J) as in Section 14.4.2, i.e., its input process A is spectrally positive. Furthermore, we suppose for simplicity that J has no zero-drift states and that there is no external input to queues $2,\ldots,n$. To avoid 'invisible' stations, we impose the condition $r_1 > \ldots > r_n$.

We define $W_j(t)$ as the content in buffer j at time t, and let $\boldsymbol{W}(t)$ be the vector of buffer contents. The evolution of the process $\boldsymbol{W}$ is completely determined by A and the initial buffer-content vector $\boldsymbol{W}(0)$. Formally, this can be made precise by using Skorokhod reflection mappings; see Section 13.5. It is our aim to study the steady-state vector of buffer contents in this network, which we denote by $\boldsymbol{W} := \boldsymbol{W}(\infty)$. The inclusion of the ages of the busy periods raises no additional difficulties, but we focus here on the simplest possible situation.

We define for $j = 1,\ldots,n$, $X_j(t) := A(t) - r_j t$ and $\boldsymbol{X}(t) = (X_1(t),\ldots,X_n(t))'$. Note that $(\boldsymbol{X},\mathsf{J})$ is a multidimensional Markov-additive process on $\mathbb{R}^n \times \{1,\ldots,N\}$ under P. We also set

$$\overline{X}_j := \sup_{t\leq 0} X_j(t), \quad \overline{F}^X_j := \inf\{t \geq 0 : X_j(t) = \overline{X}_j(\infty) \text{ or } X_j(t-) = \overline{X}_j(\infty)\},$$

and $\overline{\mathsf{J}}_j := \mathsf{J}(\overline{F}_j^X)$. Throughout, we suppose that $\boldsymbol{\pi}'_{\mathsf{J}}\mathbf{E}X_n(1) < 0$, so that each component of $\boldsymbol{X}$ drifts to $-\infty$.

Our analysis consists of three steps. First, the queueing problem is formulated in terms of free processes. The splitting technique of Section 14.3.2 can be used, in a different form, to characterize the extremes of these free processes. This is reminiscent of the analysis of Lévy-driven fluid networks in Chapter 13. The final step converts the results back to the queueing setting.

We start by giving the analogue of Proposition 14.28, thereby establishing the connection between fluid networks and extremes of X. It can be proven along the lines of Proposition 13.14. Note that the distribution of $\boldsymbol{W} = \boldsymbol{W}(\infty)$ is independent of $\boldsymbol{W}(0)$ and $\mathsf{J}(0)$.

Proposition 14.30 *The vector $\boldsymbol{W}$ is finite, and for any $\boldsymbol{\omega} \in \mathbb{R}_+^n$, we have*

$$\mathsf{E}\left[e^{-\langle \boldsymbol{\omega}, \boldsymbol{W}\rangle}; \mathsf{J} = k\right] = \boldsymbol{\pi}_{\mathsf{J}}(k)\widehat{\mathsf{E}}_k\left[e^{-\sum_{i=1}^{n-1}(\omega_i-\omega_{i+1})\overline{X}_i-\omega_n\overline{X}_n}; \overline{\mathsf{J}}_n\right]\mathbf{1}.$$

We use splitting to calculate the transform in this expression. In Chapter 13, splitting is distinguished from splitting from the left, but this is irrelevant for the arguments and the results. Modulo this remark, the following lemma can be proven along the lines of Lemma 13.2.

Lemma 14.31 *For any j, $\{(\boldsymbol{X}(t), \mathsf{J}(t)) : 0 \le t \le \overline{F}_j^X\}$ and $\{(\boldsymbol{X}(\overline{F}_j^X + t) - \boldsymbol{X}(\overline{F}_j^X), \mathsf{J}(\overline{F}_j^X + t)) : t \ge 0\}$ are $\widehat{\mathsf{P}}$-conditionally independent given $\mathsf{J}(\overline{F}_j^X)$.*

With this proposition at our disposal, the joint distribution of $\overline{\boldsymbol{F}}^X$ and $\overline{\boldsymbol{X}}$ can be derived in only a few lines. The key element in this analysis is the observation $\overline{F}_1^X \le \ldots \le \overline{F}_n^X$. In the following theorem, we give the resulting Laplace transform; in the terminology of Chapter 13, this transform has a *quasi-product form*. The proof requires only minor modifications in comparison with the proof of Theorem 13.16, and is therefore omitted. We emphasize that the product is taken from 1 to $n-1$; the order is important, since the matrices do not commute.

Corollary 14.32 *We have for $\boldsymbol{\beta} \in \mathbb{R}_+^n$,*

$$\begin{aligned}\widehat{\mathsf{E}}\left[e^{-\langle \boldsymbol{\beta}, \overline{\boldsymbol{X}}\rangle}; \overline{\mathsf{J}}_n \in \sim\right] \;=\; & \widehat{\mathsf{E}}\left[e^{-\left[\sum_{k=2}^n (r_1-r_k)\beta_k\right]\overline{F}_1^X - \left[\sum_{k=1}^n \beta_k\right]\overline{X}_1}; \overline{\mathsf{J}}_1 \in \sim\right] \\ & \times \prod_{j=1}^{n-1}\left\{\left(\widehat{\mathsf{E}}_\sim\left[e^{-\left[\sum_{k=j+1}^n (r_j-r_k)\beta_k\right]\overline{F}_j^X - \left[\sum_{k=j+1}^n \beta_k\right]\overline{X}_j}; \overline{\mathsf{J}}_j \in \sim\right]\right)^{-1}\right. \\ & \left. \times\, \widehat{\mathsf{E}}_\sim\left[e^{-\left[\sum_{k=j+2}^n (r_{j+1}-r_k)\beta_k\right]\overline{F}_{j+1}^X - \left[\sum_{k=j+1}^n \beta_k\right]\overline{X}_{j+1}}; \overline{\mathsf{J}}_{j+1} \in \sim\right]\right\},\end{aligned}$$

whenever the appropriate matrices are nonsingular.

Corollary 14.32 expresses the transform of the $\widehat{\mathsf{P}}$-distribution of $(\overline{\boldsymbol{X}}, \overline{\mathsf{J}}_n)$ in terms of the marginals $(\overline{X}_j, \overline{\mathsf{J}}_j)$ for $j = 1, \ldots, n$. Importantly, the transforms of these marginals can be found with Corollary 14.29. As a final step, we therefore cast the results back into the queueing setting. For notational convenience, we define

$$\eta_j(\boldsymbol{\omega}) := \sum_{\ell=j+1}^{n} (r_{\ell-1} - r_\ell)\omega_\ell,$$

so that we obtain the main result of this section, which is a generalization of (14.20). The simplicity of the expression for the Laplace transform is remarkable, especially in view of the transform-free solution of Kroese and Scheinhardt [202] for the two-station fluid-flow tandem with a two-dimensional background state space.

Theorem 14.33 *For $\boldsymbol{\omega} \in \mathbb{R}_+^n$, we have*

$$\begin{aligned}&\mathsf{E}\left[e^{-\langle \boldsymbol{\omega}, \boldsymbol{W}\rangle}; \mathsf{J}\right](\boldsymbol{\psi}_{-X_1}(\omega_1) - \eta_1(\boldsymbol{\omega})\boldsymbol{I}) \\ &\quad = \left(\mathbf{0}_s' \quad \omega_n \left[\boldsymbol{u}_-^n\right]' \prod_{j=1}^{n-1} \left\{ \left[\omega_{j+1}\boldsymbol{I}_{--} + \boldsymbol{\mathcal{Q}}_{--}^{(j)}(\eta_j(\boldsymbol{\omega}))\right]^{-1} \left[\omega_j \boldsymbol{I}_{--} + \boldsymbol{\mathcal{Q}}_{--}^{(j)}(\eta_j(\boldsymbol{\omega}))\right] \right\} \right),\end{aligned}$$

whenever the appropriate matrices are nonsingular.

Importantly, this theorem shows that the joint buffer-content distribution for a fluid network can immediately be established from *known* results about the single (fluid) queue discussed in Section 14.5. For instance, Lucantoni's algorithm for the BMAP/G/1 immediately yields $\boldsymbol{\mathcal{Q}}_{--}^{(j)}(\cdot)$, and similarly for algorithms that efficiently solve the matrix-quadratic equation in fluid-flow models.

Specializing Theorem 14.33 to the marginal distribution of W_n for $n > 1$, we obtain the interesting formula

$$\mathsf{E}\left[e^{-\omega W_n}; \mathsf{J} \in \sim\right] = \frac{\left[\boldsymbol{u}_-^n\right]'}{r_n - r_{n-1}} \left[\omega \boldsymbol{I}_{--} + \boldsymbol{\mathcal{Q}}_{--}^{(n-1)}((r_{n-1} - r_n)\omega)\right]^{-1} \boldsymbol{\mathcal{Q}}_{--}^{(n-1)}((r_{n-1} - r_n)\omega),$$

which should be compared with Theorem 3.2 of [92] or Corollary 13.18(i).

14.7 Concluding remarks

In the course of writing this chapter, we have bypassed several interesting questions. It is the aim of this section to sketch how some additional features can be incorporated into our framework. These features are mainly inspired by models that have been recently studied in the literature.

Markov-additive processes under exponential killing

The approach taken in this chapter can also be used to characterize the distributions of $(\overline{X}(t), \overline{F}^X(t), \overline{\mathsf{J}}(t))$ and $(\overline{X}(t), \overline{F}^X(t), \mathsf{J}(t))$ for any $t \geq 0$. By taking Laplace transforms with respect to time, this amounts to investigating $(\overline{X}(e_\lambda), \overline{F}^X(e_\lambda), \overline{\mathsf{J}}(e_\lambda))$ and $(\overline{X}(e_\lambda), \overline{F}^X(e_\lambda), \mathsf{J}(e_\lambda))$ for some $\lambda > 0$. The resulting identities can be viewed as the analogue of the second formula in Proposition 11.6.

The vector $(\overline{X}(e_\lambda), \mathsf{J}(e_\lambda))$ plays a role in a number of problems in applied probability. First, it completely specifies the solution to the one-sided exit problem [207]. We remark that, if there are no subordinator states, the nonnegative matrix $-(\boldsymbol{\mathcal{K}}_{--}^\lambda)^{-1}$ plays a prominent role in this solution; it can be interpreted as a local-time matrix. Moreover, the distribution of $(\overline{X}(e_\lambda), \mathsf{J}(e_\lambda))$ also immediately specifies the transient behavior of a queue with Markov-additive input, see [6] for a special case.

Ramifications of the tandem network in Section 14.6; priority systems

In Section 14.6, there are no external inputs to the stations $2, \ldots, n$ of a tandem fluid network. As long as these external inputs are increasing subordinators, i.e., if they do not depend on the state of the background process J, our reasoning immediately carries over to this more general setting.

Kella [184] *does* allow for a dependence of this external input (or the drain rates) on the background state, and we now outline how our framework should be modified to be able to

derive expressions under this assumption. In terms of the one-dimensional Markov-additive process X of Section 14.3, it is not sufficient to study $\overline{F}^X$ (jointly with $(\overline{X}, \overline{\mathsf{J}})$), but knowledge is required about the amount of time spent *in each of the states* till time $\overline{F}^X$.

The last-passage (or Wiener-Hopf) approach that we have used in this chapter can still be applied, but the matrices $\boldsymbol{\mathcal{K}}^{\alpha}_{--}$ now depend on a *vector* $\mathrm{vec}(\alpha)$ instead of a single value. An expression such as $\boldsymbol{\psi}_{-X}(\beta) - \alpha\boldsymbol{I}$ in Theorem 14.19 then changes to $\boldsymbol{\psi}_{-X}(\beta) - \mathrm{diag}(\alpha)$. However, the reasoning essentially requires no further new ideas. As for tandem networks, the only remaining assumption is that the components of $\overline{\boldsymbol{F}}^X$ are ordered (note that a similar assumption is needed in [184]).

Recently, there has been an interest in fluid-driven priority systems [296]. As seen in Section 13.6.3, these systems are closely related tandem queues with external inputs and equal drain rates. Although equal drain rates are not covered in Section 14.6, the techniques still apply. Indeed, if the external inputs are nondecreasing processes (with the first station as the only possible exception, see Chapter 13), the components of $\overline{\boldsymbol{F}}^X$ are ordered. In particular, our theory can be used to analyze priority fluid systems with Markov-additive input.

Phase-type jumps in the opposite direction

All Markov-additive processes in this chapter have one-sided jumps. Given the tractability of Lévy processes with phase-type jumps in the opposite direction (see Chapter 12 and [256]), it seems plausible that a similar analysis can be carried out for Markov-additive processes. Indeed, one can then again use an embedded process to which the theory of Section 14.2 can be applied.

Bibliography

[1] J. Abate, G. Choudhury, and W. Whitt, *An introduction to numerical transform inversion and its application to probability models*, Computational Probability (W. K. Grassmann, ed.), Kluwer, Boston, MA, 2000, pp. 257–323. *[11]*

[2] R. Addie, P. Mannersalo, and I. Norros, *Most probable paths and performance formulae for buffers with Gaussian input traffic*, European Trans. Telecomm. **13** (2002), 183–196. *[37, 43, 47, 147]*

[3] R. J. Adler, *An introduction to continuity, extrema, and related topics for general Gaussian processes*, Institute of Mathematical Statistics, Hayward, CA, 1990. *[27, 39, 45]*

[4] R. J. Adler and J. E. Taylor, *Random fields and their geometry*, preliminary version, 2005. *[28, 29, 39, 42]*

[5] R. Agrawal, A. Makowski, and Ph. Nain, *On a reduced load equivalence for fluid queues under subexponentiality*, Queueing Syst. **33** (1999), 5–41. *[86, 88]*

[6] S. Ahn and V. Ramaswami, *Efficient algorithms for transient analysis of stochastic fluid flow models*, J. Appl. Probab. **42** (2005), 531–549. *[207, 240, 244]*

[7] J. M. P. Albin, *On extremal theory for self-similar processes*, Ann. Probab. **26** (1998), 743–793. *[60]*

[8] D. Aldous, *Probability approximations via the Poisson clumping heuristic*, Springer-Verlag, New York, 1989. *[31]*

[9] D. Aldous and A. Bandyopadhyay, *A survey of max-type recursive distributional equations*, Ann. Appl. Probab. **15** (2005), 1047–1110. *[4]*

[10] V. Anantharam, *How large delays build up in a GI/G/1 queue*, Queueing Syst. **5** (1989), 345–367. *[34]*

[11] D. Anick, D. Mitra, and M. M. Sondhi, *Stochastic theory of a data-handling system with multiple sources*, Bell System Tech. J. **61** (1982), 1871–1894. *[8, 205]*

[12] E. Arjas and T. P. Speed, *A note on the second factorization identity of A. A. Borovkov*, Theory Probab. Appl. **18** (1973), 576–578. *[171]*

[13] S. Asmussen, *Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the GI/G/1 queue*, Adv. in Appl. Probab. **14** (1982), 143–170. *[34, 100]*

[14] ———, *Risk theory in a Markovian environment*, Scand. Actuar. J. (1989), 69–100. *[100]*

[15] ———, *Busy period analysis, rare events and transient behavior in fluid flow models*, J. Appl. Math. Stochastic Anal. **7** (1994), 269–299. *[207, 240]*

[16] ———, *Stationary distributions for fluid flow models with or without Brownian noise*, Stoch. Models **11** (1995), 21–49. *[160, 207, 212, 219, 225, 227, 240]*

[17] ———, *Subexponential asymptotics for stochastic processes: extremal behavior, stationary distributions and first passage probabilities*, Ann. Appl. Probab. **8** (1998), 354–374. *[177]*

[18] ———, *Ruin probabilities*, World Scientific Publishing Co. Inc., 2000. *[6, 34, 100, 114, 206, 212, 236]*

[19] ———, *Applied probability and queues*, second ed., Springer, New York, 2003. *[4, 5, 111, 164, 165, 169, 170, 196, 208, 211, 213, 224, 230, 235, 237, 240]*

[20] S. Asmussen, F. Avram, and M. Pistorius, *Russian and American put options under exponential phase-type Lévy models*, Stochastic Process. Appl. **109** (2004), 79–111. *[6, 160, 164]*

[21] S. Asmussen, K. Binswanger, and B. Højgaard, *Rare events simulation for heavy-tailed distributions*, Bernoulli **6** (2000), 303–322. *[99]*

[22] S. Asmussen, S. Foss, and D. Korshunov, *Asymptotics for sums of random variables with local subexponential behaviour*, J. Theoret. Probab. **16** (2003), 489–518. *[22, 154]*

[23] S. Asmussen, P. Fuckerieder, M. Jobmann, and H.-P. Schwefel, *Large deviations and fast simulation in the presence of boundaries*, Stochastic Process. Appl. **102** (2002), 1–23. *[114]*

[24] S. Asmussen and O. Kella, *A multi-dimensional martingale for Markov additive processes and its applications*, Adv. in Appl. Probab. **32** (2000), 376–393. *[205, 206, 218, 229, 230, 237]*

[25] S. Asmussen, C. Klüppelberg, and K. Sigman, *Sampling at subexponential times, with queueing applications*, Stochastic Process. Appl. **79** (1999), 265–286. *[88]*

[26] S. Asmussen and D. P. Kroese, *Improved algorithms for rare event simulation with heavy tails*, preprint, 2005. *[99]*

[27] S. Asmussen, D. P. Kroese, and R. Y. Rubinstein, *Heavy tails, importance sampling and cross-entropy*, Stoch. Models **21** (2005), 57–76. *[100]*

[28] S. Asmussen and J. R. Møller, *Calculation of the steady state waiting time distribution in GI/PH/c and MAP/PH/c queues*, Queueing Syst. **37** (2001), 9–29. *[240]*

[29] S. Asmussen and T. Rolski, *Computational methods in risk theory: a matrix-algorithmic approach*, Insurance Math. Econom. **10** (1991), no. 4, 259–274. *[206]*

[30] S. Asmussen and R. Y. Rubinstein, *Steady state rare events simulation in queueing models and its complexity properties*, Advances in queueing (J. Dshalalow, ed.), CRC, Boca Raton, FL, 1995, pp. 429–461. *[98, 100]*

[31] J.-M. Azaïs and M. Wschebor, *On the distribution of the maximum of a Gaussian field with d parameters*, Ann. Appl. Probab. **15** (2005), 254–278. *[31]*

[32] R. Azencott, *Grandes déviations et applications*, Ecole d'été de probabilités de Saint-Flour VIII — 1978, Springer, Berlin, 1980, pp. 1–176. *[28]*

[33] F. Baccelli and P. Brémaud, *Elements of queueing theory*, Springer, Berlin, 2003. *[4]*

[34] R. R. Bahadur and S. L. Zabell, *Large deviations of the sample mean in general vector spaces*, Ann. Probab. **7** (1979), 587–621. *[28]*

[35] P. Baldi and B. Pacchiarotti, *Importance sampling for the ruin problem for general Gaussian processes*, preprint, 2004. *[135, 136]*

[36] N. G. Bean, M. M. O'Reilly, and P. G. Taylor, *Algorithms for return probabilities for stochastic fluid flows*, Stoch. Models **21** (2005), 149–184. *[212, 241]*

[37] ———, *Hitting probabilities and hitting times for stochastic fluid flows*, Stochastic Process. Appl. **115** (2005), 1530–1556. *[207, 240, 241]*

[38] V. E. Beneš, *General stochastic processes in the theory of queues*, Addison-Wesley, Reading, MA, 1963. *[4, 214]*

[39] S. M. Berman, *Maximum and high level excursion of a Gaussian process with stationary increments*, Ann. Math. Statist. **43** (1972), 1247–1266. *[52]*

[40] ———, *An asymptotic bound for the tail of the distribution of the maximum of a Gaussian process*, Ann. Inst. H. Poincaré Probab. Statist. **21** (1985), 47–57. *[29]*

[41] ———, *Sojourns and extremes of stochastic processes*, Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1992. *[31, 53]*

[42] J. Bertoin, *Sur la décomposition de la trajectoire d'un processus de Lévy spectralement positif en son infimum*, Ann. Inst. H. Poincaré Probab. Statist. **27** (1991), 537–547. *[184]*

[43] ———, *Lévy processes*, Cambridge University Press, Cambridge, 1996. *[151, 156, 157, 161, 163, 169, 175, 181, 182, 183, 185, 186, 198, 202, 206, 223]*

[44] ———, *Subordinators: examples and applications*, Lectures on probability theory and statistics (Saint-Flour, 1997), Springer, Berlin, 1999, pp. 1–91. *[156]*

[45] J. Bertoin and R. A. Doney, *Cramér's estimate for Lévy processes*, Statist. Probab. Lett. **21** (1994), 363–365. *[174]*

[46] ———, *On conditioning a random walk to stay nonnegative*, Ann. Probab. **22** (1994), 2152–2167. *[34]*

[47] ———, *Some asymptotic results for transient random walks*, Adv. in Appl. Probab. **28** (1996), 207–226. *[154]*

[48] J. D. Biggins, *Large deviations for mixtures*, Electron. Comm. Probab. **9** (2004), 60–71. *[111]*

[49] P. Billingsley, *Convergence of probability measures*, John Wiley & Sons Inc., New York, 1968. *[15, 17]*

[50] ———, *Convergence of probability measures*, second ed., John Wiley & Sons Inc., New York, 1999. *[15, 42, 151]*

[51] N. H. Bingham, C. M. Goldie, and J. L. Teugels, *Regular variation*, Cambridge University Press, Cambridge, 1989. *[13, 15, 48, 49, 56]*

[52] V. I. Bogachev, *Gaussian measures*, American Mathematical Society, Providence, RI, 1998. *[26, 27, 39]*

[53] N. K. Boots and M. Mandjes, *Fast simulation of a queue fed by a superposition of many (heavy-tailed) sources*, Probab. Engrg. Inform. Sci. **16** (2002), 205–232. *[128, 130, 137]*

[54] N. K. Boots and P. Shahabuddin, *Simulating tail probabilities in GI/GI/1 queues and insurance risk processes with subexponential distributions*, preprint, 2002. *[99, 100]*

[55] A. A. Borovkov, *Stochastic processes in queueing theory*, Springer-Verlag, New York, 1976. *[4, 154]*

[56] ———, *Large deviations of sums of random variables of two types*, Siberian Adv. Math. **11** (2001), no. 4, 1–24. *[88]*

[57] A. A. Borovkov and A. A. Mogul′skiĭ, *Large deviations and testing statistical hypotheses. I. Large deviations of sums of random vectors*, Siberian Adv. Math. **2** (1992), 52–120. *[17]*

[58] R. J. Boucherie, O. J. Boxma, and K. Sigman, *A note on negative customers, $GI/G/1$ workload, and risk processes*, Probab. Engrg. Inform. Sci. **11** (1997), 305–311. *[164]*

[59] O. J. Boxma and V. Dumas, *Fluid queues with long-tailed activity period distributions*, Comp. Comm. **21** (1998), 1509–1529. *[8]*

[60] O. J. Boxma, O. Kella, and D. Perry, *An intermittent fluid system with exponential on-times and semi-Markov input rates*, Probab. Engrg. Inform. Sci. **15** (2001), 189–198. *[234, 236]*

[61] M. Braverman, T. Mikosch, and G. Samorodnitsky, *Tail probabilities of subadditive functionals of Lévy processes*, Ann. Appl. Probab. **12** (2002), 69–100. *[177]*

[62] M. Braverman and G. Samorodnitsky, *Functionals of infinitely divisible stochastic processes with exponential tails*, Stochastic Process. Appl. **56** (1995), 207–231. *[21]*

[63] J. A. Bucklew, *Large deviation techniques in decision, simulation, and estimation*, John Wiley & Sons Inc., New York, 1990. *[17]*

[64] J. A. Bucklew, P. Ney, and J. S. Sadowsky, *Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains*, J. Appl. Probab. **27** (1990), 44–59. *[99]*

[65] V. V. Buldygin and V. V. Zaiats, *A global asymptotic normality of the sample correlogram of a stationary Gaussian process*, Random Oper. Stochastic Equations **7** (1999), 109–132. *[44]*

[66] N. R. Chaganty, *Large deviations for joint distributions and statistical applications*, Sankhyā Ser. A **59** (1997), 147–166. *[111]*

[67] T. Chan, *Some applications of Lévy processes in insurance and finance*, preprint, 2005. *[177]*

[68] L. Chaumont and R. A. Doney, *On Lévy processes conditioned to stay positive*, Electron. J. Probab. **10** (2005), 948–961. *[180, 184]*

[69] P. Cheridito, *Mixed fractional Brownian motion*, Bernoulli **7** (2001), 913–934. *[89]*

[70] S. Chevet, *Gaussian measures and large deviations*, Probability in Banach spaces, IV (Oberwolfach, 1982), Springer, Berlin, 1983, pp. 30–46. *[26, 28]*

[71] J. Choe and N. B. Shroff, *On the supremum distribution of integrated stationary Gaussian processes with negative linear drift*, Adv. in Appl. Probab. **31** (1999), 135–157. *[51]*

[72] ———, *Use of the supremum distribution of Gaussian processes in queueing analysis with long-range dependence and self-similarity*, Stoch. Models **16** (2000), 209–231. *[51]*

[73] G. Choudhury, *Some aspects of an M/G/1 queueing system with optional second service*, Top **11** (2003), 141–150. *[169]*

[74] J. Chover and J. Feldman, *On positive-definite integral kernels and a related quadratic form*, Trans. Amer. Math. Soc. **89** (1958), 92–99. *[27]*

[75] K. L. Chung, *A course in probability theory*, third ed., Academic Press, San Diego, 2001. *[17, 213]*

[76] E. Çinlar, *Markov additive processes. I, II*, Z. Wahrsch. Verw. Gebiete **24** (1972), 85–93; 95–121. *[223]*

[77] B. S. Cirel′son, I. A. Ibragimov, and V. N. Sudakov, *Norms of Gaussian sample functions*, Proceedings of the Third Japan-USSR Symposium on Probability Theory (Tashkent, 1975) (Berlin), Springer, 1976, pp. 20–41. *[28]*

[78] J. W. Cohen, *Superimposed renewal processes and storage with gradual input*, Stochastic Process. Appl. **2** (1974), 31–57. *[8, 207, 234]*

[79] ———, *The single server queue*, second ed., North-Holland Publishing Co., Amsterdam, 1982. *[4, 154, 218]*

[80] J. F. Collamore, *Importance sampling techniques for the multidimensional ruin problem for general Markov additive sequences of random vectors*, Ann. Appl. Probab. **12** (2002), 382–421. *[100, 111]*

[81] J. B. Conway, *A course in functional analysis*, Springer-Verlag, New York, 1990. *[17]*

[82] H. E. Daniels, *The first crossing-time density for Brownian motion with a perturbed linear boundary*, Bernoulli **6** (2000), 571–580. *[62]*

[83] R. B. Davies and D. S. Harte, *Tests for Hurst effect*, Biometrika **74** (1987), 95–102. *[128, 140]*

[84] A. de Acosta, *Large deviations for vector-valued Lévy processes*, Stochastic Process. Appl. **51** (1994), 75–115. *[116]*

[85] K. Dębicki, *A note on LDP for supremum of Gaussian processes over infinite horizon*, Statist. Probab. Lett. **44** (1999), 211–219. *[36, 51]*

[86] ———, *Asymptotics of the supremum of scaled Brownian motion*, Probab. Math. Statist. **21** (2001), 199–212. *[51, 63]*

[87] ———, *Ruin probability for Gaussian integrated processes*, Stochastic Process. Appl. **98** (2002), 151–174. *[29, 51, 52, 53, 59, 64, 67]*

[88] ———, *Some remarks on properties of generalized Pickands constants*, Tech. Report PNA-R0204, CWI, the Netherlands, 2002. *[31]*

[89] K. Dębicki, A. B. Dieker, and T. Rolski, *Quasi-product forms for Lévy-driven fluid networks*, preprint, 2005. *[158]*

[90] K. Dębicki and M. Mandjes, *Exact overflow asymptotics for queues with many Gaussian inputs*, J. Appl. Probab. **40** (2003), 704–720. *[52, 130, 144]*

[91] ———, *Traffic with an fBm limit: convergence of the stationary workload process*, Queueing Syst. **46** (2004), 113–127. *[8]*

[92] K. Dębicki, M. Mandjes, and M. van Uitert, *A tandem queue with Lévy input: a new representation of the downstream queue length*, to appear in Probab. Engrg. Inform. Sci., 2006. *[180, 196, 244]*

[93] K. Dębicki, Z. Michna, and T. Rolski, *On the supremum from Gaussian processes over infinite horizon*, Probab. Math. Statist. **18** (1998), 83–100. *[36, 51]*

[94] ———, *Simulation of the asymptotic constant in some fluid models*, Stoch. Models **19** (2003), 407–423. *[31]*

[95] K. Dębicki and Z. Palmowski, *On-off fluid models in heavy traffic environment*, Queueing Syst. **33** (1999), 327–338. *[9]*

[96] K. Dębicki and T. Rolski, *A Gaussian fluid model*, Queueing Syst. **20** (1995), 433–452. *[51, 63]*

[97] ———, *A note on transient Gaussian fluid models*, Queueing Syst. **41** (2002), 321–342. *[52]*

[98] K. Dębicki, B. Zwart, and S. Borst, *The supremum of a Gaussian process over a random interval*, Statist. Probab. Lett. **68** (2004), 221–234. *[43]*

[99] A. Dembo and T. Zajic, *Large deviations: from empirical mean and measure to partial sums process*, Stochastic Process. Appl. **57** (1995), 191–224. *[116]*

[100] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, second ed., Springer, New York, 1998. *[17, 18, 19, 20, 98, 99, 105, 109, 110, 118, 122]*

[101] F. den Hollander, *Large deviations*, American Mathematical Society, Providence, RI, 2000. *[17]*

[102] P. den Iseger, *Numerical transform inversion using Gaussian quadrature*, Probab. Engrg. Inform. Sci. **20** (2006), 1–44. *[11]*

[103] D. Denisov, S. Foss, and D. Korshunov, *Tail asymptotics for the supremum of a random walk when the mean is not finite*, Queueing Syst. **46** (2004), 15–33. *[22]*

[104] J.-D. Deuschel and D. W. Stroock, *Large deviations*, Academic Press Inc., Boston, MA, 1989. *[17, 28]*

[105] A. B. Dieker, *Applications of factorization embeddings for Lévy processes*, Tech. Report PNA-E0505, CWI, the Netherlands, 2005. *[158]*

[106] ———, *Conditional limit theorems for queues with Gaussian input, a weak convergence approach*, Stochastic Process. Appl. **115** (2005), 849–873. *[32]*

[107] ———, *Extremes of Gaussian processes over an horizon*, Stochastic Process. Appl. **115** (2005), 207–248. *[32]*

[108] ———, *Reduced-load equivalence for queues with Gaussian input*, Queueing Syst. **49** (2005), 405–414. *[32]*

[109] A. B. Dieker and M. Lelarge, *Tails for (max,plus) recursions under subexponentiality*, to appear in Queueing Syst., 2006. *[i]*

[110] A. B. Dieker and M. Mandjes, *On spectral simulation of fractional Brownian motion*, Probab. Engrg. Inform. Sci. **17** (2003), 417–434. *[128]*

[111] ———, *Fast simulation of overflow probabilities in a queue with Gaussian input*, Tech. Report PNA-E0404, CWI, the Netherlands, 2004. *[101]*

[112] ———, *On asymptotically efficient simulation of large deviation probabilities*, Adv. in Appl. Probab. **37** (2005), 539–552. *[101]*

[113] ———, *Efficient simulation of random walks exceeding a nonlinear boundary*, to appear in Stoch. Models, 2006. *[101]*

[114] ———, *Extremes of Markov-additive processes with one-sided jumps, with queueing applications*, preprint, 2006. *[158]*

[115] C. R. Dietrich and G. N. Newsam, *Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix*, SIAM J. Scient. Comput. **18** (1997), no. 4, 1088–1107. *[140]*

[116] I. H. Dinwoodie and S. L. Zabell, *Large deviations for exchangeable random vectors*, Ann. Probab. **20** (1992), 1147–1166. *[111]*

[117] ———, *Large deviations for sequences of mixtures*, Statistics and Probability: A Raghu Raj Bahadur Festschrift (J. K. Ghosh, S. K. Mitra, K. R. Parthasarathy, and B. L. S. Prakasa Roa, eds.), Wiley Eastern, New Delhi, 1993, pp. 171–190. *[111]*

[118] R. A. Doney, *On Wiener-Hopf factorisation and the distribution of extrema for certain stable processes*, Ann. Probab. **15** (1987), 1352–1362. *[157, 164]*

[119] ———, *Stochastic bounds for Lévy processes*, Ann. Probab. **32** (2004), 1545–1552. *[160, 174]*

[120] ———, *Tanaka's construction for random walks and Lévy processes*, Séminaire de Probabilités XXXVIII, Springer, Berlin, 2005, pp. 1–4. *[183, 184]*

[121] R. A. Doney and A. E. Kyprianou, *Overshoots and undershoots of Lévy processes*, to appear in Ann. Appl. Probab., 2006. *[163, 174, 201]*

[122] P. Doukhan, G. Oppenheim, and M. S. Taqqu (eds.), *Theory and applications of long-range dependence*, Birkhäuser Boston, Boston, MA, 2003. *[62]*

[123] N. Duffield and N. O'Connell, *Large deviations and overflow probabilities for the general single server queue, with applications*, Math. Proc. Cam. Phil. Soc. **118** (1995), 363–374. *[36, 51]*

[124] F. Dufresne and H. Gerber, *The surpluses immediately before and at ruin, and the amount of the claim causing ruin*, Insurance Math. Econom. **7** (1988), 193–199. *[201]*

[125] T. Duncan, Y. Yan, and P. Yan, *Exact asymptotics for a queue with fractional Brownian input and applications in ATM networks*, J. Appl. Probab. **38** (2001), 932–945. *[52]*

[126] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*, John Wiley & Sons Inc., New York, 1997. *[17]*

[127] P. Dupuis and H. Wang, *Importance sampling, large deviations, and differential games*, Stoch. Stoch. Rep. **76** (2004), 481–508. *[104, 128, 133, 138, 139, 148]*

[128] J. Durbin, *The first-passage density of the Brownian motion process to a curved boundary (with an appendix by D. Williams)*, J. Appl. Probab. **29** (1992), 291–304. *[62]*

[129] R. S. Ellis, *Entropy, large deviations, and statistical mechanics*, Springer-Verlag, New York, 1985. *[17]*

[130] A. Elwalid and D. Mitra, *Analysis, approximations and admission control of a multi-service multiplexing system with priorities*, Proc. of IEEE INFOCOM, 1995, pp. 463–472. *[199]*

[131] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling extremal events*, Springer, Berlin, 1997. *[6, 13]*

[132] W. Feller, *An introduction to probability theory and its applications. Vol. II.*, second ed., John Wiley & Sons Inc., New York, 1971. *[13, 151, 154]*

[133] J. Feng and T. Kurtz, *Large deviations for stochastic processes*, preprint, 2005. *[17, 111]*

[134] S. Foss and T. Konstantopoulos, *An overview of some stochastic stability methods*, J. Oper. Res. Soc. Japan **47** (2004), 275–303. *[4]*

[135] S. Foss and D. Korshunov, *Sampling at a random time with a heavy-tailed distribution*, Markov Process. Related Fields **6** (2000), 543–568. *[88]*

[136] S. Foss and S. Zachary, *Asymptotics for the maximum of a modulated random walk with heavy-tailed increments*, Analytic methods in applied probability, Amer. Math. Soc., Providence, RI, 2002, pp. 37–52. *[178]*

[137] H. J. Furrer, *Risk processes perturbed by α-stable Lévy motion*, Scand. Actuar. J. (1998), 59–74. *[166]*

[138] H. R. Gail, S. L. Hantler, and B. A. Taylor, *Spectral analysis of M/G/1 and G/M/1 type Markov chains*, Adv. in Appl. Probab. **28** (1996), 114–165. *[218]*

[139] A. Ganesh and N. O'Connell, *A large deviation principle with queueing applications*, Stoch. Stoch. Rep. **73** (2002), 25–35. *[100]*

[140] A. Ganesh, N. O'Connell, and D. Wischik, *Big queues*, Springer, Berlin, 2004. *[17, 99, 100]*

[141] A. Ganesh and D. Wischik, *The calculus of Hurstiness*, preprint, 2005. *[35]*

[142] J. L. Geluk and L. de Haan, *Regular variation, extensions and Tauberian theorems*, Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam, 1987. *[13]*

[143] H. Gerber and E. Shiu, *From ruin theory to pricing reset guarantees and perpetual put options*, Insurance Math. Econom. **24** (1999), 3–14. *[6]*

[144] P. Glasserman, Ph. Heidelberger, and P. Shahabuddin, *Asymptotically optimal importance sampling and stratification for pricing path-dependent options*, Math. Finance **9** (1999), 117–152. *[104]*

[145] P. Glasserman, Ph. Heidelberger, P. Shahabuddin, and T. Zajic, *Multilevel splitting for estimating rare event probabilities*, Oper. Res. **47** (1999), 585–600. *[96]*

[146] P. Glasserman and S. Kou, *Analysis of an importance sampling estimator for tandem queues*, ACM Trans. Modeling Comp. Simulation **4** (1995), 22–42. *[104]*

[147] P. Glasserman and J. Li, *Importance sampling for portfolio credit risk*, preprint, 2003. *[104, 111]*

[148] P. Glasserman and Y. Wang, *Counterexamples in importance sampling for large deviations probabilities*, Ann. Appl. Probab. **7** (1997), 731–746. *[103, 112, 133]*

[149] P. Glynn, *Efficiency improvement techniques*, Ann. Oper. Res. **53** (1994), 175–197. *[96, 98]*

[150] P. Glynn and D. Iglehart, *Importance sampling for stochastic simulations*, Management Science **35** (1989), 1367–1392. *[98]*

[151] P. Glynn and W. Whitt, *The asymptotic efficiency of simulation estimators*, Oper. Res. **40** (1992), 505–520. *[98]*

[152] C. M. Goldie and C. Klüppelberg, *Subexponential distributions*, A practical guide to heavy tails (Santa Barbara, CA, 1995), Birkhäuser Boston, Boston, MA, 1998, pp. 435–459. *[21, 22]*

[153] P. Greenwood and J. Pitman, *Fluctuation identities for Lévy processes and splitting at the maximum*, Adv. in Appl. Probab. **12** (1980), 893–902. *[183]*

[154] J. M. Harrison and M. I. Reiman, *Reflected Brownian motion on an orthant*, Ann. Probab. **9** (1981), 302–308. *[190]*

[155] J. M. Harrison and R. J. Williams, *Brownian models of open queueing networks with homogeneous customer populations*, Stochastics **22** (1987), 77–115. *[179]*

[156] ———, *Brownian models of feedforward queueing networks: quasireversibility and product form solutions*, Ann. Appl. Probab. **2** (1992), 263–293. *[179]*

[157] Ph. Heidelberger, *Fast simulation of rare events in queueing and reliability models*, ACM Trans. Modeling Comp. Simulation **5** (1995), 43–85. *[96, 98]*

[158] D. G. Hobson, D. Williams, and A. T. A. Wood, *Taylor expansions of curve-crossing probabilities*, Bernoulli **5** (1999), 779–795. *[62]*

[159] T. Höglund, *A unified formulation of the central limit theorem for small and large deviations from the mean*, Z. Wahrsch. Verw. Gebiete **49** (1979), 105–117. *[19]*

[160] G. Hooghiemstra, *Conditioned limit theorems for waiting-time processes of the M/G/1 queue*, J. Appl. Probab. **20** (1983), 675–688. *[33]*

[161] C. Huang, M. Devetsikiotis, I. Lambadaris, and A. R. Kaye, *Fast simulation of queues with long-range dependent traffic*, Stoch. Models **15** (1999), no. 3, 429–460. *[128, 135]*

[162] J. Hüsler, *Extremes of a Gaussian process and the constant H_α*, Extremes **2** (1999), 59–70. *[31]*

[163] J. Hüsler and V. Piterbarg, *Extremes of a certain class of Gaussian processes*, Stochastic Process. Appl. **83** (1999), 257–271. *[51, 53, 57, 60, 61]*

[164] ———, *On the ruin probability for physical fractional Brownian motion*, Stochastic Process. Appl. **113** (2004), 315–332. *[51, 64]*

[165] J. Hüsler and Ch. Schmid, *Extreme values of a portfolio of Gaussian processes and a trend*, preprint, 2004. *[86]*

[166] M. Huzak, M. Perman, H. Šikić, and Z. Vondraček, *Ruin probabilities and decompositions for general perturbed risk processes*, Ann. Appl. Probab. **14** (2004), 1378–1397. *[85, 161, 166, 169, 173]*

[167] D. Iglehart, *Extreme values in the GI/G/1 queue*, Ann. Math. Statist. **43** (1972), 627–635. *[154]*

[168] M. Iltis, *Sharp asymptotics of large deviations in* $\mathbf{R}^{\mathbf{d}}$, J. Theoret. Probab. **8** (1995), 501–522. *[19]*

[169] M. Jacobsen, *Splitting times for Markov processes and a generalised Markov property for diffusions*, Z. Wahrsch. Verw. Gebiete **30** (1974), 27–43. *[180]*

[170] J. Jacod and A. N. Shiryaev, *Limit theorems for stochastic processes*, second ed., Springer, Berlin, 2003. *[8, 151, 164]*

[171] N. K. Jaiswal, *Priority queues*, Academic Press, New York, 1968. *[200]*

[172] P. Jelenković and A. Lazar, *Asymptotic results for multiplexing subexponential on-off processes*, Adv. in Appl. Probab. **31** (1999), 394–421. *[8, 86]*

[173] P. Jelenković and P. Momčilović, *Large deviation analysis of subexponential waiting times in a processor-sharing queue*, Math. Oper. Res. **28** (2003), 587–608. *[88]*

[174] ———, *Large deviations of square root insensitive random sums*, Math. Oper. Res. **29** (2004), 398–406. *[88]*

[175] P. Jelenković, P. Momčilović, and B. Zwart, *Reduced load equivalence under subexponentiality*, Queueing Syst. **46** (2004), 97–112. *[86, 88]*

[176] A. Jobert and L. C. G. Rogers, *Option pricing with Markov-modulated dynamics*, preprint, 2004. *[6, 206]*

[177] S. Juneja and P. Shahabuddin, *Simulating heavy tailed processes using delayed hazard rate twisting*, ACM Trans. Modeling Comp. Simulation **12** (2002), 94–118. *[99]*

[178] I. Kaj and M. S. Taqqu, *Convergence to fractional Brownian motion and to the Telecom process: the integral representation approach*, preprint, 2004. *[8, 10]*

[179] H. Kaspi, *On the symmetric Wiener-Hopf factorization for Markov additive processes*, Z. Wahrsch. Verw. Gebiete **59** (1982), 179–196. *[228]*

[180] O. Kella, *Parallel and tandem fluid networks with dependent Lévy inputs*, Ann. Appl. Probab. **3** (1993), 682–695. *[179, 180, 196]*

[181] ———, *Stability and nonproduct form of stochastic fluid networks with Lévy inputs*, Ann. Appl. Probab. **6** (1996), 186–199. *[179]*

[182] ———, *Stochastic storage networks: stationarity and the feedforward case*, J. Appl. Probab. **34** (1997), 498–507. *[179, 194]*

[183] ———, *Non-product form of two-dimensional fluid networks with dependent Lévy inputs*, J. Appl. Probab. **37** (2000), 1117–1122. *[179]*

[184] ———, *Markov-modulated feedforward fluid networks*, Queueing Syst. **37** (2001), 141–161. *[205, 219, 242, 244, 245]*

[185] O. Kella and W. Whitt, *A storage model with a two-state random environment*, Oper. Res. **40** (1992), S257–S262. *[234, 235]*

[186] ———, *A tandem fluid network with Lévy input*, Queueing and related models, Oxford Univ. Press, New York, 1992, pp. 112–128. *[179, 197]*

[187] ———, *Useful martingales for stochastic storage processes with Lévy input*, J. Appl. Probab. **29** (1992), 396–403. *[179, 180]*

[188] J. Kennedy, *A probabilistic view of some algebraic results in Wiener-Hopf theory for symmetrizable Markov chains*, Stochastics and quantum mechanics, World Sci. Publishing, River Edge, NJ, 1992, pp. 165–177. *[160]*

[189] ———, *Understanding the Wiener-Hopf factorization for the simple random walk*, J. Appl. Probab. **31** (1994), 561–563. *[152, 206, 214]*

[190] G. Kersting and K. Memişoğlu, *Path decompositions for Markov chains*, Ann. Probab. **32** (2004), 1370–1390. *[180]*

[191] L. Kleinrock, *Queueing systems*, vol. I, II, John Wiley & Sons Inc., New York, 1975, 1976. *[4]*

[192] C. Klüppelberg, A. E. Kyprianou, and R. A. Maller, *Ruin probabilities and overshoots for general Lévy insurance risk processes*, Ann. Appl. Probab. **14** (2004), 1766–1801. *[161, 174, 176]*

[193] S. G. Kobelkov, *The ruin problem for the stationary Gaussian process*, Theory Probab. Appl. **49** (2005), 155–163. *[31, 52, 64]*

[194] T. Konstantopoulos and V. Anantharam, *Optimal flow control schemes that regulate the burstiness of traffic*, IEEE/ACM Trans. on Networking **3** (1995), 423–432. *[86]*

[195] T. Konstantopoulos, G. Last, and S.-J. Lin, *On a class of Lévy stochastic networks*, Queueing Syst. **46** (2004), 409–437. *[179, 180, 194]*

[196] T. Konstantopoulos and S.-J. Lin, *Macroscopic models for long-range dependent network traffic*, Queueing Syst. **28** (1998), 215–243. *[8, 10]*

[197] D. Korshunov, *On distribution tail of the maximum of a random walk*, Stochastic Process. Appl. **72** (1997), 97–103. *[155]*

[198] ———, 2005, personal communication. *[22]*

[199] L. Kosten, *Stochastic theory of a multi-entry buffer (I)*, Delft Progr. Rep., Series F **1** (1974), 10–18. *[8]*

[200] S. Kou and H. Wang, *First passage times of a jump diffusion process*, Adv. in Appl. Probab. **35** (2003), 504–531. *[164]*

[201] Yu. Kozachenko, O. Vasylyk, and T. Sottinen, *Path space large deviations of a large buffer with Gaussian input traffic*, Queueing Syst. **42** (2002), 113–129. *[33, 34, 35, 36, 38, 44, 46, 48, 51]*

[202] D. P. Kroese and W. R. W. Scheinhardt, *Joint distributions for interacting fluid queues*, Queueing Syst. **37** (2001), 99–139. *[243]*

[203] V. Kulkarni, *Fluid models for single buffer systems*, Frontiers in queueing, CRC, Boca Raton, FL, 1997, pp. 321–338. *[8, 240]*

[204] H. Kunita, *Stochastic flows and stochastic differential equations*, Cambridge University Press, Cambridge, 1990. *[66]*

[205] H. H. Kuo, *Gaussian measures in Banach spaces*, Springer-Verlag, Berlin, 1975. *[26]*

[206] T. Kurtz, *Limit theorems for workload input models*, Stochastic networks: theory and applications (F. Kelly, S. Zachary, and I. Ziedins, eds.), Oxford University Press, 1996, pp. 119–140. *[8]*

[207] A. E. Kyprianou and Z. Palmowski, *Fluctuations of spectrally negative Markov additive processes*, preprint, 2005. *[244]*

[208] ———, *A martingale review of some fluctuation theory for spectrally negative Lévy processes*, Séminaire de Probabilités XXXVIII, Springer, Berlin, 2005, pp. 16–29. *[157, 180]*

[209] J. Lamperti, *Semi-stable stochastic processes*, Trans. Amer. Math. Soc. **104** (1962), 62–78. *[41, 60]*

[210] G. Latouche and V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*, SIAM, Philadelphia, PA, 1999. *[238]*

[211] M. R. Leadbetter, G. Lindgren, and H. Rootzén, *Extremes and related properties of random sequences and processes*, Springer-Verlag, New York, 1983. *[29, 31, 74]*

[212] P. L'Ecuyer, *Efficiency improvement and variance reduction*, Proceedings of the 1994 Winter Simulation Conference, 1994, pp. 122–132. *[96]*

[213] M. Ledoux, *The concentration of measure phenomenon*, American Mathematical Society, Providence, RI, 2001. *[28, 29]*

[214] M. Ledoux and M. Talagrand, *Probability in Banach spaces*, Springer-Verlag, Berlin, 1991. *[28]*

[215] T. Lehtonen and H. Nyrhinen, *On asymptotically efficient simulation of ruin probabilities in a Markovian environment*, Scand. Actuar. J. (1992), 60–75. *[100]*

[216] ———, *Simulating level-crossing probabilities by importance sampling*, Adv. in Appl. Probab. **24** (1992), 858–874. *[100, 101, 114, 121]*

[217] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, *On the self-similar nature of Ethernet traffic (extended version)*, IEEE/ACM Trans. on Networking **2** (1994), no. 1, 1–15. *[10, 140]*

[218] W. V. Li and Q.-M. Shao, *Gaussian processes: inequalities, small ball probabilities and applications*, Stochastic processes: theory and methods, North-Holland, Amsterdam, 2001, pp. 533–597. *[28]*

[219] M. A. Lifshits, *Gaussian random functions*, Kluwer Academic Publishers, Dordrecht, 1995. *[26, 28, 32, 45]*

[220] N. Likhanov and R. R. Mazumdar, *Cell loss asymptotics for buffers fed with a large number of independent stationary sources*, J. Appl. Probab. **36** (1999), 86–96. *[128, 144]*

[221] D. M. Lucantoni, *New results on the single server queue with a batch Markovian arrival process*, Stoch. Models **7** (1991), 1–46. *[212, 238, 239]*

[222] K. Majewski, *Large deviations for multi-dimensional reflected fractional Brownian motion*, Stoch. Stoch. Rep. **75** (2003), 233–257. *[44]*

[223] M. Mandjes and S. Borst, *Overflow behavior in queues with many long-tailed inputs*, Adv. in Appl. Probab. **32** (2000), 1150–1167. *[8]*

[224] M. Mandjes, P. Mannersalo, I. Norros, and M. van Uitert, *Large deviations of infinite intersections of events in Gaussian processes*, Tech. Report PNA-E0409, CWI, the Netherlands, 2004. *[38]*

[225] L. Massoulié and A. Simonian, *Large buffer asymptotics for the queue with fractional Brownian input*, J. Appl. Probab. **36** (1999), 894–906. *[51]*

[226] K. Maulik and B. Zwart, *Tail asymptotics for exponential functionals of Lévy processes*, Stochastic Process. Appl. **116** (2006), 156–177. *[177]*

[227] Z. Michna, *On tail probabilities and first passage times for fractional Brownian motion*, Math. Methods Oper. Res. **49** (1999), 335–354. *[128, 135]*

[228] T. Mikosch, S. I. Resnick, H. Rootzén, and A. Stegeman, *Is network traffic approximated by stable Lévy motion or fractional Brownian motion?*, Ann. Appl. Probab. **12** (2002), 23–68. *[8, 10]*

[229] P. W. Millar, *Zero-one laws and the minimum of a Markov process*, Trans. Amer. Math. Soc. **226** (1977), 365–391. *[180, 183]*

[230] ———, *A path decomposition for Markov processes*, Ann. Probab. **6** (1978), 345–348. *[180, 183]*

[231] D. S. Mitrinović, *Analytic inequalities*, Springer-Verlag, New York, 1970. *[131]*

[232] M. Miyazawa, *Hitting probabilities in a Markov additive process with linear movements and upward jumps: applications to risk and queueing processes*, Ann. Appl. Probab. **14** (2004), 1029–1054. *[206, 212, 218]*

[233] M. Miyazawa and H. Takada, *A matrix exponential form for hitting probabilities and its application to a Markov-modulated fluid queue with downward jumps*, J. Appl. Probab. **39** (2002), 604–618. *[212, 233, 236]*

[234] A. A. Mogul′skiĭ, *Large deviations for the trajectories of multidimensional random walks*, Theory Probab. Appl. **21** (1976), 309–323. *[116]*

[235] ———, *Large deviations for processes with independent increments*, Ann. Probab. **21** (1993), 202–215. *[116]*

[236] C. Moler and Ch. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev. **45** (2003), 3–49. *[212]*

[237] E. Mordecki, *The distribution of the maximum of a Lévy processes with positive jumps of phase-type*, Proceedings of the Conference Dedicated to the 90th Anniversary of Boris Vladimirovich Gnedenko (Kyiv, 2002), vol. 8, 2002, pp. 309–316. *[160, 161, 164, 166]*

[238] ———, *Optimal stopping and perpetual options for Lévy processes*, Finance Stoch. **6** (2002), 473–493. *[6]*

[239] E. S. Murphree, *Some new results on the subexponential class*, J. Appl. Probab. **26** (1989), 892–897. *[22]*

[240] S. V. Nagaev, *Large deviations of sums of independent random variables*, Ann. Probab. **7** (1979), 745–789. *[17]*

[241] O. Narayan, *Exact asymptotic queue length distribution for fractional Brownian traffic*, Adv. Perform. Anal. **1** (1998), 39–63. *[51]*

[242] M. F. Neuts, *Structured stochastic matrices of $M/G/1$ type and their applications*, Marcel Dekker Inc., New York, 1989. *[239]*

[243] P. Ney and E. Nummelin, *Markov additive processes. I. Eigenvalue properties and limit theorems*, Ann. Probab. **15** (1987), 561–592. *[223]*

[244] I. Norros, *Busy periods of fractional Brownian storage: a large deviations approach*, Adv. Perform. Anal. **2** (1999), 1–19. *[34, 38, 48]*

[245] H. Nyrhinen, *On the typical level crossing time and path*, Stochastic Process. Appl. **58** (1995), 121–137. *[34]*

[246] G. O'Brien, *Sequences of capacities, with connections to large-deviation theory*, J. Theoret. Probab. **9** (1996), 19–35. *[17]*

[247] G. O'Brien and W. Vervaat, *Compactness in the theory of large deviations*, Stochastic Process. Appl. **57** (1995), 1–10. *[18]*

[248] A. G. Pakes, *Convolution equivalence and infinite divisibility*, J. Appl. Probab. **41** (2004), 407–424. *[21, 174, 176, 177]*

[249] K. R. Parthasarathy, *Probability measures on metric spaces*, Academic Press Inc., New York, 1967. *[40]*

[250] E. Parzen, *Probability density functionals and reproducing kernel Hilbert spaces*, Proc. Sympos. Time Series Analysis (Brown Univ., 1962), Wiley, New York, 1963, pp. 155–169. *[27]*

[251] V. V. Petrov, *Sums of independent random variables*, Springer-Verlag, New York, 1975. *[17]*

[252] J. Pickands, III, *Asymptotic properties of the maximum in a stationary Gaussian process*, Trans. Amer. Math. Soc. **145** (1969), 75–86. *[29]*

[253] ———, *Upcrossing probabilities for stationary Gaussian processes*, Trans. Amer. Math. Soc. **145** (1969), 51–73. *[29]*

[254] F. J. Piera, R. R. Mazumdar, and F. M. Guillemin, *On product-form stationary distributions for reflected diffusions with jumps in the positive orthant*, Adv. in Appl. Probab. **37** (2005), 212–228. *[179, 180]*

[255] M. Pistorius, *Pricing American and barrier options under a stochastic volatility model with jumps*, preprint, 2005. *[6, 206]*

[256] ———, *On maxima and ladder processes for a dense class of Lévy processes*, to appear in J. Appl. Probab., 2006. *[160, 161, 231, 233, 245]*

[257] V. Piterbarg, *Asymptotic methods in the theory of Gaussian processes and fields*, American Mathematical Society, Providence, RI, 1996. *[28, 29, 30, 31, 52, 61, 65, 67, 75, 79]*

[258] ———, *Large deviations of a storage process with fractional Brownian motion as input*, Extremes **4** (2001), 147–164. *[52]*

[259] V. Piterbarg and V. Fatalov, *The Laplace method for probability measures in Banach spaces*, Russian Math. Surveys **50** (1995), 1151–1239. *[29]*

[260] V. Piterbarg and V. Prisjažnjuk, *Asymptotic behavior of the probability of a large excursion for a nonstationary Gaussian process*, Teor. Verojatnost. i Mat. Statist. (1978), 121–134, 183. *[52]*

[261] V. Piterbarg and S. Stamatovic, *Limit theorem for high level a-upcrossings by χ-process*, Theory Probab. Appl. **48** (2004), 734–741. *[31]*

[262] N. U. Prabhu, *On the ruin problem of collective risk theory*, Ann. Math. Statist. **32** (1961), 757–764. *[5]*

[263] ———, *Stochastic storage processes*, Springer, New York, 1998. *[4, 162, 170, 207, 238]*

[264] J. Preater, *$M/M/\infty$ transience revisited*, J. Appl. Probab. **34** (1997), 1061–1067. *[241]*

[265] A. Puhalskii, *Large deviation analysis of the single server queue*, Queueing Syst. **21** (1995), 5–66. *[17]*

[266] ———, *Large deviations and idempotent probability*, Chapman & Hall/CRC, Boca Raton, FL, 2001. *[17]*

[267] C. Qualls and H. Watanabe, *Asymptotic properties of Gaussian processes*, Ann. Math. Statist. **43** (1972), 580–596. *[53]*

[268] E. Reich, *On the integrodifferential equation of Takács. I*, Ann. Math. Statist. **29** (1958), 563–570. *[4]*

[269] S. I. Resnick, *Extreme values, regular variation, and point processes*, Springer-Verlag, New York, 1987. *[13]*

[270] ———, *A probability path*, Birkhäuser Boston, Boston, MA, 1999. *[15, 17]*

[271] D. Revuz and M. Yor, *Continuous martingales and Brownian motion*, second ed., Springer-Verlag, Berlin, 1994. *[9, 40]*

[272] Ph. Robert, *Stochastic networks and queues*, Springer, Berlin, 2003. *[4, 100, 180, 190]*

[273] R. T. Rockafellar, *Convex analysis*, Princeton University Press, Princeton, N.J., 1970. *[19, 100, 145, 146]*

[274] L. C. G. Rogers, *Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains*, Ann. Appl. Probab. **4** (1994), 390–413. *[240, 241]*

[275] T. Rolski, H. Schmidli, V. Schmidt, and J. L. Teugels, *Stochastic processes for insurance and finance*, John Wiley & Sons Ltd., Chichester, 1999. *[6, 166]*

[276] H. L. Royden, *Real analysis*, The Macmillan Co., New York, 1968. *[116]*

[277] R. Y. Rubinstein and D. P. Kroese, *The cross-entropy method*, Springer, New York, 2004. *[98]*

[278] J. S. Sadowsky, *On the optimality and stability of exponential twisting in Monte Carlo estimation*, IEEE Trans. Inform. Theory **39** (1993), 119–128. *[99]*

[279] ———, *On Monte Carlo estimation of large deviations probabilities*, Ann. Appl. Probab. **6** (1996), 399–422. *[101, 103, 104, 106, 107, 108, 111, 114, 116, 119, 120, 124]*

[280] J. S. Sadowsky and J. A. Bucklew, *On large deviations theory and asymptotically efficient Monte Carlo estimation*, IEEE Trans. Inform. Theory **36** (1990), 579–588. *[99, 104, 107, 128, 137]*

[281] P. Salminen and I. Norros, *On busy periods of the unbounded Brownian storage*, Queueing Syst. **39** (2001), 317–333. *[196]*

[282] G. Samorodnitsky and M. S. Taqqu, *Stable non-Gaussian random processes, stochastic models with infinite variance*, Chapman & Hall, New York, 1994. *[157]*

[283] K. Sato, *Lévy processes and infinitely divisible distributions*, Cambridge University Press, Cambridge, 1999. *[151]*

[284] W. R. W. Scheinhardt and B. Zwart, *A tandem fluid queue with gradual input*, Probab. Engrg. Inform. Sci. **16** (2002), 29–45. *[234, 235, 242]*

[285] H. Schmidli, *Distribution of the first ladder height of a stationary risk process perturbed by α-stable Lévy motion*, Insurance Math. Econom. **28** (2001), 13–20. *[85, 166, 173]*

[286] E. Seneta, *Regularly varying functions*, Springer-Verlag, Berlin, 1976. *[13]*

[287] Q.-M. Shao, *Bounds and estimators of a basic constant in extreme value theory of Gaussian processes*, Statist. Sinica **6** (1996), 245–257. *[31]*

[288] A. Shwartz and A. Weiss, *Large deviations for performance analysis*, Chapman & Hall, London, 1995. *[17]*

[289] D. Siegmund, *Importance sampling in the Monte Carlo study of sequential tests*, Ann. Statist. **4** (1976), 673–684. *[103]*

[290] ———, *Sequential analysis*, Springer-Verlag, New York, 1985. *[6]*

[291] D. W. Stroock, *Probability theory, an analytic view*, Cambridge University Press, Cambridge, 1993. *[15]*

[292] R. Szechtman and P. Glynn, *Rare-event simulation for infinite server queues*, Proceedings of the 2002 Winter Simulation Conference, 2002, pp. 416–423. *[99]*

[293] L. Takács, *Introduction to the theory of queues*, Oxford University Press, New York, 1962. *[4]*

[294] M. S. Taqqu, W. Willinger, and R. Sherman, *Proof of a fundamental result in self-similar traffic modeling*, Comp. Comm. Review **27** (1997), 5–23. *[9]*

[295] H. C. Tijms, *A first course in stochastic models*, John Wiley & Sons Ltd., Chichester, 2003. *[4]*

[296] E. Tzenova, I. Adan, and V. Kulkarni, *A two-priority fluid flow model: joint steady state distribution of the buffer content processes*, Tech. Report SPOR 2004-04, Technical University Eindhoven, 2004. *[245]*

[297] ———, *Fluid models with jumps*, Stoch. Models **21** (2005), 37–55. *[218, 219, 238]*

[298] V. V. Uchaikin and V. M. Zolotarev, *Chance and stability*, VSP, Utrecht, 1999. *[157]*

[299] A. S. Üstünel and M. Zakai, *Transformation of measure on Wiener space*, Springer-Verlag, Berlin, 2000. *[26]*

[300] A. van der Vaart and H. van Zanten, *Donsker theorems for diffusions: necessary and sufficient conditions*, Ann. Probab. **33** (2005), 1422–1451. *[42]*

[301] A. van der Vaart and J. Wellner, *Weak convergence and empirical processes*, Springer-Verlag, New York, 1996. *[17]*

[302] H. van Zanten, *When is a linear combination of independent fBm's equivalent to a single fBm?*, Tech. Report WS2005-11, Vrije Universiteit Amsterdam, Department of Mathematics, 2005. *[89]*

[303] S. R. S. Varadhan, *Large deviations and applications*, SIAM, Philadelphia, PA, 1984. *[17]*

[304] N. Veraverbeke, *Asymptotic behaviour of Wiener-Hopf factors of a random walk*, Stochastic Process. Appl. **5** (1977), 27–37. *[154]*

[305] W. Whitt, *Stochastic-process limits*, Springer, New York, 2002. *[8, 15]*

[306] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, *Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level*, IEEE/ACM Trans. on Networking **5** (1997), 71–86. *[8, 62]*

[307] D. Wischik, *Moderate deviations in queueing theory*, preprint, 2001. *[8]*

[308] ———, *Sample path large deviations for queues with many inputs*, Ann. Appl. Probab. **11** (2001), 379–404. *[37, 131]*

[309] A. T. A. Wood and G. Chan, *Simulation of stationary Gaussian processes in* $[0,1]^d$, J. Comp. Graph. Stat. **3** (1994), no. 4, 409–432. *[140]*

[310] H. Yang and L. Zhang, *Spectrally negative Lévy processes with applications in risk theory*, Adv. in Appl. Probab. **33** (2001), 281–291. *[85]*

[311] S. Zachary, *A note on Veraverbeke's theorem*, Queueing Syst. **46** (2004), 9–14. *[155]*

[312] A. Zeevi and P. Glynn, *On the maximum workload of a queue fed by fractional Brownian motion*, Ann. Appl. Probab. **10** (2000), 1084–1099. *[52]*

[313] Y. Q. Zhao, W. Li, and W. J. Braun, *Censoring, factorizations, and spectral analysis for transition matrices with block-repeating entries*, Methodol. Comput. Appl. Probab. **5** (2003), 35–58. *[220]*

[314] B. Zwart, S. Borst, and K. Dębicki, *Reduced-load equivalence for Gaussian processes*, Oper. Res. Lett. **33** (2005), 502–510. *[86, 88]*

[315] ———, *Subexponential asymptotics of hybrid fluid and ruin models*, Ann. Appl. Probab. **15** (2005), 500–517. *[86]*

Author Index

Samenvatting (Summary)

Extremen en vloeistofmodellen

In de wachtrijtheorie worden systemen bestudeerd waar klanten aankomen, eventueel wachten op hun bediening, en weggaan nadat ze zijn geholpen. Bij sommige toepassingen van deze theorie, bijvoorbeeld in moderne communicatienetwerken, zijn individuele klanten (datapakketjes) zo klein dat het eenvoudiger is om het binnengebrachte werk als een vloeistofstroom op te vatten. De inputstroom bij een vloeistofmodel is aan een kansmechanisme onderhevig, vergelijkbaar aan dat in een systeem met klanten. Zowel de aankomstmomenten als de hoeveelheden werk die klanten meebrengen zijn immers onzeker.

Bij het standaard vloeistofmodel komt vloeistof in een buffer terecht die met een bepaalde vaste snelheid leegstroomt. Als er tijdelijk meer vloeistof de buffer instroomt dan kan worden verwerkt, dan wordt de overtollige vloeistof opgeslagen en neemt de bufferinhoud dus toe. De bufferinhoud neemt weer af zodra er minder vloeistof instroomt dan er wegstroomt.

Dit roept de vraag op of we de kans dat de bufferinhoud een bepaalde drempelwaarde overschrijdt kunnen uitrekenen. De kans dat vloeistof niet in de buffer past indien deze een vaste (eindige) capaciteit zou hebben hangt hiermee nauw samen. Inzichten in deze kansen zijn nuttig bij het ontwerpen van systemen waarin wachtrijen een rol spelen, aangezien vaak kosten verbonden zijn aan buffercapaciteit.

Bij het bestuderen van (de kansverdeling van) de bufferinhoud kan gebruik gemaakt worden van een opvallende identiteit, die de niet-negatieve bufferinhoud koppelt aan een extreem (het maximum) van een gerelateerd proces. Dit gerelateerde proces is een 'vrij proces' in de zin dat het zowel positieve als negatieve waarden kan aannemen. Het belangrijkste voordeel van het werken met extremen is dat de opgebouwde theorie ook gebruikt kan worden in risicotheorie en financiële wiskunde.

Dit proefschrift bestaat uit de delen A tot en met C, voorafgegaan door twee hoofdstukken van algemene aard. Het eerste hoofdstuk behandelt het verband tussen extremen en wachtrijsystemen, in het bijzonder vloeistofmodellen. Ook netwerken van deze systemen worden kort beschouwd, en de relevantie van vloeistofmodellen met zogenaamde Gaussische inputstromen en Lévy inputstromen wordt besproken. Hoofdstuk 2 gaat vervolgens in op enkele onderwerpen uit de wiskundige analyse en kansrekening die een belangrijke rol spelen in dit proefschrift. De drie delen van dit proefschrift beginnen elk met een hoofdstuk waarin specifieke achtergrondinformatie wordt besproken die van belang is in de resterende hoofdstukken van deze delen.

Deel A: Vloeistofmodellen met Gaussische inputstromen

In deel A, dat de hoofdstukken 3 tot en met 6 omvat, worden vloeistofmodellen met Gaussische inputstromen geanalyseerd. Bij deze modellen bestaat de mogelijkheid dat er hevige correlaties

zijn in de inputstroom tussen verschillende tijdsintervallen.

Hoofdstuk 4 richt zich op de volgende twee vragen: (i) Als de bufferinhoud een hoog peil heeft bereikt, op welke manier is dit gebeurd? en (ii) Als de buffer lang niet-leeg is geweest, op welke manier is dit tot stand gekomen? Met behulp van de theorie van grote afwijkingen formuleren we conditionele limietstellingen als antwoord op deze vragen. Zo beschrijft de limietstelling corresponderend met vraag (i) volgens welk pad een hoog bufferniveau bereikt wordt. Een bijproduct van de analyse is een goede benadering voor de logaritme van de kans dat de bufferinhoud groot is.

In hoofdstuk 5 verkrijgen we een nauwkeurigere benadering van deze kans, namelijk een zogenaamde asymptotisch exacte benadering. Gebruik makend van het verband tussen vloeistofmodellen en extremen, analyseren we de kans op een grote bufferinhoud door bepaalde bestaande technieken voor extremen van Gaussische processen uit te breiden.

De resultaten van hoofdstuk 5 worden gebruikt in hoofdstuk 6 om een vloeistofmodel te analyseren waarbij meerdere heterogene Gaussische inputstromen in dezelfde buffer samenvloeien. We zijn geïnteresseerd in voorwaarden waaronder een deel van deze inputstromen de anderen domineert in de zin dat dit deel, in zekere asymptotische zin, de verdeling van de bufferinhoud bepaalt.

Deel B: Simulatie

Deel B, bestaande uit de hoofdstukken 7 tot en met 10, heeft tot doel de simulatie van kleine kansen te onderzoeken. In de context van dit proefschrift is dit in het bijzonder relevant ter bepaling van verlieskansen in vloeistofmodellen. Er zijn verschillende simulatie-aanpakken denkbaar bij het schatten van dit soort kleine kansen, zodat de vraag rijst welke hiervan het meest efficiënt is.

Hoofdstuk 8 behandelt de simulatie van een belangrijke klasse van kleine kansen, namelijk kansen die verband houden met de theorie van grote afwijkingen. Met behulp van deze theorie formuleren we voorwaarden om te bepalen of een schatter zekere optimaliteitseigenschappen bezit, en laten we zien dat deze de bestaande voorwaarden uit de literatuur verbeteren. De toepasbaarheid van deze verscherpte voorwaarden beperkt zich overigens niet tot de vloeistofmodellen en extremen die in dit proefschrift centraal staan.

In hoofdstuk 9 geven we een eerste toepassing van bovengenoemde voorwaarden in de context van extremen. We bestuderen de kans dat een zogenaamde stochastische wandeling ooit een bepaalde waarde bereikt, waarbij deze waarde afhangt van het tijdstip van de wandeling. Hiertoe onderzoeken we twee simulatiemethoden; de eerste simuleert op padniveau, en de ander op stapniveau. Voor beide methoden leiden we condities af waaronder optimaliteit bereikt wordt, en vervolgens vergelijken we deze condities.

De voorwaarden van hoofdstuk 8 worden ook gebruikt in hoofdstuk 10, waar we terugkeren naar de Gaussische vloeistofmodellen uit deel A. Wederom zijn we geïnteresseerd in (de kansverdeling van) de bufferinhoud, maar nu willen we deze kansen simuleren in plaats van benaderingen vinden. Hiertoe analyseren we de optimaliteitseigenschappen van vier verschillende simulatiemethoden en voeren we computerexperimenten uit om te zien hoe goed elke methode werkt in de praktijk.

Deel C: Vloeistofmodellen en netwerken met Lévy inputstromen

In deel C, bestaande uit de hoofdstukken 11 tot en met 14, worden vloeistofmodellen met Lévy inputstromen bestudeerd. De geheugenloosheid van Lévy processen manifesteert zich in een zogenaamde splitseigenschap. Er ontstaat een krachtige combinatie van technieken door deze eigenschap te gebruiken na het herformuleren van vloeistofmodellen in termen van extremen.

Hoofdstuk 12 bevat een drietal toepassingen van de splitstechniek. Eerst gebruiken we deze techniek om de extremen van een Lévy proces met een bepaalde sprongstructuur te analyseren, en vervolgens bestuderen we geperturbeerde risicoprocessen. Bij de derde toepassing maken we gebruik van resultaten voor stochastische wandelingen om het asymptotische gedrag te vinden van de kans dat het maximum van een Lévy proces een zekere grote waarde overschrijdt.

In hoofdstuk 13 beschrijven we hoe de splitstechniek gebruikt kan worden om inzicht te verschaffen in netwerken van vloeistofbuffers met Lévy inputstromen. We laten zien dat de verdeling van de bufferinhoud-vector (preciezer gezegd: de Laplace getransformeerde hiervan) een zogenaamde quasi-productvorm kan hebben. Deze quasi-productvormen verschijnen ook bij het bestuderen van de periodes dat de buffer leeg respectievelijk vol is. We werken de resulterende formules uit voor netwerken waarbij de Lévy inputstromen niet-negatieve sprongen hebben.

De resultaten uit hoofdstuk 13 worden uitgebreid in hoofdstuk 14, waar Lévy processen worden vervangen door algemenere processen, de zogenaamde Markov-additieve processen. Aannemende dat de inputstroom niet-negatieve of niet-positieve sprongen heeft, onderzoeken we eerst het corresponderende vloeistofmodel met een enkele buffer. In de literatuur is een breed scala aan technieken beschreven om speciale gevallen van dit model te analyseren. Met behulp van de splitseigenschap kan het algemene model opgelost worden en kunnen tegelijkertijd verbanden gelegd worden met andere methoden. Gebruik makend van dit resultaat laten we zien dat matrix-quasi-productvormen een cruciale rol spelen in netwerken met Markov-additieve inputstromen.

About the author

A. B. (Ton) Dieker was born in Amsterdam (the Netherlands) on August 23, 1979. In the same city, he completed grammar school at the Fons Vitae Lyceum in June 1997. As a part of the Operations Research program of the Vrije Universiteit Amsterdam (VU), he spent four months at the Humboldt-Universität zu Berlin (Germany) in the summer of 2001. Less than a year later, in April 2002, he graduated from the VU. His master's thesis was entitled "Simulation of fractional Brownian motion". Subsequently, he became a Ph.D. student under Michel Mandjes at Twente University, with the Center for Mathematics and Computer Science (CWI) as his daily working environment. He visited the Laboratoire de Probabilités, Université Pierre et Marie Curie, Paris (France) for three months in 2004. Ton defends his thesis at the University of Amsterdam on March 9, 2006. After this, he will join Neil O'Connell's research group at University College Cork (Ireland) as a postdoctoral researcher.