MANUSCRIT

Présenté pour l'obtention de

L'HABILITATION À DIRIGER DES RECHERCHES

Délivrée par : *l'Université Toulouse 3 Paul Sabatier*

Présentée et soutenue le *10/05/2016* **par :**

Florian SIMATOS

Théorèmes limite fonctionnels, processus de branchement et réseaux stochastiques

Functional limit theorems, branching processes and stochastic networks

JURY

Patrick CATTIAUX	Professeur des universités	Président du jury
Jean-François DELMAS	Professeur des universités	Membre du jury
Thomas DUQUESNE	Professeur des universités	Rapporteur
Sergey FOSS	Full professor	Membre du jury
David GAMARNIK	Full professor	Rapporteur
Laurent MICLO	Directeur de recherche	Membre du jury

École doctorale et spécialité :

MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de Recherche :

ISAE-SUPAERO

Directeur de Thèse :

Laurent MICLO

Rapporteurs :

Maury BRAMSON, *Thomas DUQUESNE* et *David GAMARNIK*

This work is licensed under a Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License.

Acknowledgements

First and foremost, I would like to thank Professor Maury Bramson, Professor Thomas Duquesne and Professor David Gamarnik for accepting to review my manuscript. Some of their papers rank among the most influential in the fields of branching processes and stochastic networks, and it is an honor to have them as reviewers.

I would also like to thank Professor Laurent Miclo for accepting to act as director of my habilitation, as well as Professor Patrick Cattiaux, Professor Jean-François Delmas and Professor Sergey Foss for being part of my defense committee. I am learning on a daily basis how scarce a resource time is and it makes me even more thankful for the time everybody has devoted to my project.

Research is a weird business: it is done 90% alone, but these are the 10% remaining, made of lively discussions and interactions, which make it so exciting. It is therefore my sincere pleasure to thank here my various co-authors as well some colleagues who have played an important role in my career at some point even though it did not translate into a publication: the work presented here was only possible thanks to them. My warmest thanks thus go to Elie Aïdekon, François Baccelli, Vincent Bansaye, Sem Borst, Niek Bouman, Ed Coffman, A. Ganesh, Fabrice Guillemin, Tom Kurtz, Peter Lakner, Amaury Lambert, Marc Lelarge, Lasse Leskëla, Sarah Lilienthal, D. Manjunath, Alexandre Proutière, Josh Reed, Philippe Robert, Emmanuel Schertzer, Shuzo Tarumi, Danielle Tibi, Johan van Leeuwaarden, Remco van der Hofstad, Gil Zussman and Bert Zwart. I hope to have the chance – and, more importantly, the time – to pursue our rewarding collaborations.

The four-year period that I have spent in the Netherlands between CWI and TU/e represents the most exciting time I have had from a scientific standpoint. Although the previous list only contains a few Dutch names, I would also like to address a general thank to my formeer Dutch colleagues who have made these four years so enjoyable and productive: most of the work presented here has indeed been carried out while in the Netherlands. I would like to thank more specifically Bert Zwart, Sem Borst and Johan van Leeuwaarden for funding my post-doctoral positions.

My special thanks to Odile Riteau, our enthusiastic secretary at ISAE-SUPAERO, for her precious help in arranging all the practical details of the defense.

Finally, the writing of this manuscript has come at a turning point of my career and personal life: I was hired at ISAE-SUPAERO three months earlier and Gaëtan was barely two-month old. My habilitation and lecture notes being now written, I do not anticipate any more acknowledgements in the near future and so I seize this opportunity to thank one more time Sandrine for her love and support – may it last forever.

Contents

Acknowledgements **iii**

Summary **vii**

1 Introduction **1**

1.1 Functional law of large numbers and central limit theorem 1

1.2 Motivation 3

1.3 Notation used throughout the document 6

2 Two theoretical results on weak convergence **11**

2.1 Introduction 11

2.2 A sufficient condition for tightness 14

2.3 Weak convergence of regenerative processes 17

3 Branching processes **23**

3.1 Introduction 24

3.2 Galton–Watson processes in varying environments 30

3.3 Binary, homogeneous Crump–Mode–Jagers processes 37

3.4 Crump–Mode–Jagers trees with short edges 45

4 Stochastic networks **53**

4.1 Introduction 53

4.2 Processor-Sharing queue length process 56

4.3 A stochastic network with mobile customers 64

4.4 Lingering effect for queue-based scheduling algorithms 72

5 Branching processes in mathematical finance **81**

5.1 Study of a model of limit order book 81

6 Research perspectives **91**

6.1 Branching processes 91

6.2 Stochastic averaging via functional analysis 93

6.3 Stochastic networks 94

Summary

Overview

This manuscript describes some of the work I have been doing since 2010 and the end of my PhD. As the title suggests, it contains three main parts.

Functional limit theorems: Chapter 2 presents two theoretical results on the weak convergence of stochastic processes: one – Theorems 2.2.1 and 2.2.2 – is a sufficient condition for the tightness of a sequence of stochastic processes and the other – Theorem 2.3.1 – provides a sufficient condition for the weak convergence of a sequence of regenerative processes;

Branching processes: in Chapter 3, scaling limits of three particular types of branching processes are discussed: 1) Galton–Watson processes in varying environments, 2) binary and homogeneous Crump–Mode–Jagers processes and 3) Crump–Mode–Jagers processes with short edges;

Stochastic networks: Chapter 4 presents three results on stochastic networks: 1) scaling limits of the $M/G/1$ Processor-Sharing queue length process, 2) study of a model of stochastic network with mobile customers and 3) heavy traffic delay performance of queue-based scheduling algorithms.

These topics are closely intertwined: tightness of Galton–Watson processes in varying environments studied in Section 3.2 is obtained thanks to Theorem 2.2.1 of Chapter 2; results of Section 4.2 on the Processor-Sharing queue are essentially obtained by combining results on binary and homogeneous Crump–Mode–Jagers processes from Section 3.3 together with the sufficient condition for the convergence of regenerative processes from Chapter 2; and scaling limits of the model of mobile network of Section 4.3 rely on the same theoretical result. In order to give an overview of how these results relate to each other, Chapters 2, 3 and 4 are briefly presented next.

Chapter 5 is devoted to the study of a model of limit order books in mathematical finance: the main technical tool is a coupling with a branching random walk, and it also adopts an excursion point-of-view similar in spirit as in Section 2.3 whereby the limiting process is characterized by its excursion measure.

Finally, the manuscript is concluded in Chapter 6 with research perspectives.

Summary of Chapter 2

Section 2.2 presents the main result of [BS, BKS], namely a sufficient condition for the tightness of a sequence of stochastic processes (X_n) of the form

$$\mathbb{E}\left[1 \wedge d(X_n(t), X_n(t))^\beta \mid X_n(u), u \leq s\right] \leq F_n(t) - F_n(s),\ n \geq 1, 0 \leq s \leq t,$$

with d the distance, $\beta > 0$ and (F_n) a sequence of càdlàg functions with $F_n \xrightarrow{J_1} F$ in the Skorohod J_1 topology. This result extends a result from Kurtz [74] where F was assumed to be continuous: allowing general càdlàg F was motivated by the study of processes in varying environments that may exhibit accumulation of fixed points of discontinuity. This criterion is used in Section 3.2 to prove the tightness of a sequence of Galton–Watson processes in varying environments.

Section 2.3 presents the main results of [LS14]. This paper establishes a new method for proving the weak convergence of a sequence of regenerative processes from the convergence of their excursions. It is shown that for a tight sequence (X_n) of regenerative processes to converge, it is enough that the first excursion with "size" $> \varepsilon$ converges together with its left- and right-endpoints and its size. Here the size of an excursion is given by a general measurable mapping φ with values in $[0, \infty]$. This approach was motivated by the study of the Processor-Sharing queue in Section 4.2 and also proved useful to study a stochastic network with mobile users discussed in Section 4.3.

Summary of Chapter 3

Section 3.2 presents results on scaling limits of Galton–Watson processes in varying environments obtained in [BS15]. These results adopt the original approach of Grimvall [44] via the study of Laplace exponents and extend results of Kurtz [75] and Borovkov [17] by allowing offspring distributions to have infinite variance. The price to pay is a loss of generality in the drift term, corresponding to the first moments of the offspring distributions, which is assumed to have finite variation.

Section 3.3 presents results on scaling limits of binary and homogeneous Crump–Mode–Jagers processes obtained in [LSZ13, LS15]. The main tool is the connection with Lévy processes, whereby a binary and homogeneous Crump–Mode–Jagers process is seen as the local time process of a compound Poisson process with drift stopped upon hitting 0. Assuming that the sequence of Lévy processes converges, it is shown that this implies weak convergence of the sequence of associated local time processes – and thus of the binary and homogeneous Crump–Mode–Jagers processes – in the finite variance case where the limiting Lévy process is a Brownian motion. In the infinite variance case, this always implies convergence of the finite-dimensional distributions but extra assumptions are needed to get weak convergence.

Section 3.4 continues the study of Crump–Mode–Jagers processes and presents results from [SS]. Results in the binary and homogeneous case suggest a general invariance principle, namely that scaling limits of Crump–Mode–Jagers processes with "short" edges, i.e., where individuals do not live for "long" periods of time, should belong to the universality class of Galton–Watson processes. Results of this section quantify and make rigorous this intuition by shifting the viewpoint from the branching processes to the random chronological trees encoding them. In particular, the height and contour processes of such trees are studied and their scaling limits are derived under a “short” edge condition.

Summary of Chapter 4

Section 4.2 continues the presentation of results of [LSZ13]: here, the main object of study is the Processor-Sharing queue length process and the main tool, the Lamperti transformation which maps a binary, homogeneous Crump–Mode–Jagers process to an excursion of the queue length process of the $M/G/1$ Processor-Sharing queue. In particular, results of Section 3.3 together with continuity properties of the Lamperti transformation imply that “big” excursions of the Processor-Sharing queue length process converge. In view of the main result of Section 2.3, in order to get the convergence of the whole process, a control on the left- and right-endpoints and the size of excursions is needed, as well as tightness.

Section 4.3 presents results from [BS13, ST10] where a stochastic network with mobile users is studied. In this model, customers move within the network according to a Markovian dynamic independent from the service received, which is in sharp contrast with Jackson networks where movements are coupled with service completion. This model is characterized by the coexistence of two time scales: a “fast” time scale corresponding to the inner movements of customers and which accelerates as customers accumulate; and a “slow” time scale corresponding to arrivals to and departures from the network, which is bounded by the arrival and service rates and thus remains $O(1)$. The coexistence of these two time scales induces new and original queueing phenomena which are studied in this section, in particular concerning stability and heavy traffic behavior. Moreover, scaling limits in the heavy traffic regime are established by invoking results of Section 2.3 for regenerative processes. In this context, a control of “big” excursions is achieved precisely because of the coexistence of these two time scales which imply a spatial state space collapse phenomenon when the network is overloaded.

Section 4.4 presents results from [SBB13] concerning the delay performance in heavy traffic of queue-based algorithms. From a technical standpoint, this corresponds to studying CSMA in the classical interference model, with the particular feature that activation rates depend on the current workload through an activation ψ. Such queue-based algorithms have received considerable attention in recent years, and it is shown that a lingering effect lower bounds the delay performance which scales in the heavy traffic regime $\rho \uparrow 1$ at least as fast $1/(1-\rho)^2$. Thus, these systems are prone to inherent inefficiencies which proscribe achieving the optimal $1/(1-\rho)$ lower bound.

Chapter 1

Introduction

Contents

1.1 Functional law of large numbers and central limit theorem 1

1.2 Motivation 3

1.3 Notation used throughout the document 6

1.1 Functional law of large numbers and central limit theorem

Among the many important results from probability theory, two stand out prominently: the law of large numbers and the central limit theorem. Let throughout this introduction $(\xi_k, k \in \mathbb{N})$ be i.i.d. random variables and define the partial sums $X(k) = \xi_1 + \cdots + \xi_k$. Let $\overset{\text{a.s.}}{\to}$ and $\overset{\text{d}}{\to}$ denote almost sure convergence and convergence in distribution as $n \to \infty$, respectively.

Theorem 1.1.1 (Strong law of large numbers)**.** *If* $\mathbb{E}(|\xi_1|) < \infty$, *then* $n^{-1}X(n) \overset{\text{a.s.}}{\to} \mathbb{E}(\xi_1)$.

When $\mathbb{E}(\xi_1) \neq 0$, the strong law of large numbers therefore justifies the approximation $X(n) \approx n\mathbb{E}(\xi_1)$. If $\mathbb{E}(\xi_1) = 0$ however, it only states that $X(n)$ is negligible compared to n and leaves open the question of the right order of magnitude of $X(n)$ for large n: this question is, to some extent, settled by the central limit theorem.

Theorem 1.1.2 (Central limit theorem)**.** *Let N be a standard normal random variable: if,* $\mathbb{E}(\xi_1) = 0$ *and* $\mathbb{E}(\xi_1^2) = 1$, *then* $n^{-1/2}X(n) \overset{\text{d}}{\to} N$.

Although presented here as the first order asymptotic expansion of $X(n)$ in the *critical case* $\mathbb{E}(\xi_1) = 0$, the central limit theorem can also be seen as the second order asymptotic expansion of $X(n)$ in the non-critical case: indeed, Theorem 1.1.2 immediately entails that if ξ_1 has finite second moment, then $n^{1/2}(X(n) - n\mathbb{E}(\xi_1))$ converges in distribution to a normal random variable.

The problems considered in this manuscript deal with functional versions of these fundamental results. For the convergence of stochastic processes, we consider in this introduction the space of càdlàg functions endowed with the topology of uniform convergence on compact sets (as limit processes will be almost surely continuous). In a functional setting, the strong law of large numbers takes the following form.

Theorem 1.1.3 (Functional strong law of large numbers)**.** *For $n \in \mathbb{N}$ and $t \in \mathbb{R}_+$ define $\bar{X}_n(t) = n^{-1}X([nt])$: if $\mathbb{E}(|\xi_1|) < \infty$, then $\bar{X}_n \overset{\text{a.s.}}{\to} s$ with $s(t) = t\mathbb{E}(\xi_1)$.*

Similarly as in the scalar case, this result does not provide the right order of magnitude of $\bar{X}_n$ in the critical case $\mathbb{E}(\xi_1) = 0$. A refined result is provided by the following functional central limit theorem. Throughout this manuscript, W refers to a standard Brownian motion.

Theorem 1.1.4 (Functional central limit theorem)**.** *For $n \in \mathbb{N}$ and $t \in \mathbb{R}_+$ define $\hat{X}_n(t) = n^{-1/2}X([nt])$: if $\mathbb{E}(\xi_1) = 0$ and $\mathbb{E}(\xi_1^2) = 1$, then $\hat{X}_n \overset{\text{d}}{\to} W$.*

Similarly as in the scalar case, there are two ways to interpret the functional central limit theorem. On the one hand, it can be interpreted as saying that, in the critical case $\mathbb{E}(\xi_1) = 0$, fluctuations of $X([nt])$ are of the order of $n^{1/2}$:

$$(X([nt]), t \in \mathbb{R}_+) \approx \left(n^{1/2}W(t), t \in \mathbb{R}_+\right), \ \mathbb{E}(\xi_1) = 0. \tag{1.1}$$

On the other hand, it can be seen as the second order asymptotic expansion of the process $(X([nt]), t \in \mathbb{R}_+)$ as $n \to \infty$:

$$(X([nt]) - ns(t), t \in \mathbb{R}_+) \approx \left(n^{1/2}W(t), t \in \mathbb{R}_+\right). \tag{1.2}$$

Of course, these two viewpoints are equivalent, i.e., they lead to the same limit, but this is peculiar to the random walk case: in general, these two schemes will lead to different limits. This is for instance the case for the $M/M/1$ queue, and to see this let us begin by rewriting (1.1) in the following equivalent form:

$$\left(X([n^2 t]), t \in \mathbb{R}_+\right) \approx (nW(t), t \in \mathbb{R}_+), \ \mathbb{E}(\xi_1) = 0. \tag{1.3}$$

The interpretation of (1.1) is that one fixes the time-scale (n) and then looks for the correct order of magnitude of the fluctuations of X on this time-scale ($n^{1/2}$): this is actually the way we got to (1.1) in the first place. The interpretation for (1.3) reverses the viewpoint: one first fixes the space-scale (n) and then looks for the time-scale on which fluctuations are of this order (n^2). Said otherwise, both (1.1) and (1.3) aim at going beyond the crude convergence $\bar{X}_n \to \mathbf{0}$ (the constant function with value 0) given by the functional strong law of large numbers, but they achieve this in different ways: (1.1) asserts that $n^{1/2}\bar{X}_n(t) \overset{\text{d}}{\to} W(t)$, which amounts to amplifying fluctuations of $\bar{X}_n$; whereas (1.3) asserts that $\bar{X}_n(nt) \overset{\text{d}}{\to} W(t)$, which amounts to speeding up $\bar{X}_n$.

Let us now illustrate the fact that (1.2) and (1.3) are not necessarily equivalent by considering the $M/M/1$ queue. Recall that an $M/M/1$ queue with arrival rate λ and service rate μ is the Markov process with generator

$$\Omega(f)(k) = \lambda(f(k+1) - f(k)) + \mu\mathbb{1}(k \geq 1)(f(k-1) - f(k)), \ k \in \mathbb{Z}_+, f : \mathbb{Z}_+ \to \mathbb{R}.$$

Let $L^{(n)}(t)$ be the number of customers in a critical $M/M/1$ queue at time t, where the arrival and service rates λ, μ are equal and the superscript (n) refers to the initial state: $L^{(n)}(0) = n$. Note that this last relation fixes the scaling in space.

If $\bar{L}_n(t) = n^{-1}L^{(n)}(nt)$, then the functional strong law of large numbers states that $\bar{L}_n \xrightarrow{\text{a.s.}} \mathbf{1}$, the function which takes the constant value one, and there are now two kinds of central limit theorem. The first one deals with the second order asymptotics of $\bar{L}_n$, given by $n^{1/2}(\bar{L}_n - \mathbf{1}) \xrightarrow{\text{d}} (2\lambda)^{1/2}W$: this is the analog of (1.2) for the $M/M/1$ queue. The second one amounts to scaling $\bar{L}_n$ in time in order to get a more interesting limit: it states that $(\bar{L}_n(nt), t \in \mathbb{R}_+) \xrightarrow{\text{d}} |\mathbf{1} + (2\lambda)^{1/2}W|$. This result is the analog of (1.1) and (1.3) for the $M/M/1$ queue, and thus indeed illustrates that (1.2) and (1.3) are not in general equivalent.

The problems considered in this manuscript belong to the second category: we will be interested in the scaling in time and space, without centering terms, of stochastic processes in the critical regime. The term *scaling limits* is a generic term which usually refers to such normalization schemes. Since one usually ends up with diffusion processes in the limit, it is also sometimes called *diffusion approximation*.

Mainly two classes of stochastic processes will be investigated: **branching processes** and **stochastic networks**. For branching processes, the critical regime corresponds to the regime where the mean population size stays constant: it lies in-between the subcritical regime, where the mean population size decays exponentially fast, and the supercritical regime, where the mean population size increases exponentially fast.

For stochastic networks, the critical regime corresponds to the boundary of the stability region, defined as the set of parameters corresponding to a stable (e.g., positive recurrent in the case of Markov processes) process. It is usually expressible in the form $\rho = 1$, with ρ the *load* of the network, and is referred to in the queueing literature as *heavy traffic regime*.

1.2 Motivation

The motivation for the large amount of literature dedicated to functional limit theorems spans a wide range of reasons, both theoretical and applied.

Invariance principle

The deepest reason, in my view, takes its root in the so-called *invariance principle*. This corresponds to the striking feature of Theorems 1.1.3 and 1.1.4 that the limits only depend on the distribution of X_1 through its first and (in the central limit theorem setting) second moments. There are, of course, a wide variety of distributions with the same first and second moments but these differences are washed out in these asymptotic regimes. This principle has several profound implications.

From a theoretical standpoint, this makes it possible to classify processes according to their scaling limits: we talk about *universality class*. For instance, when looked through the suitable time-space lens, all finite variance random walks look like a Brownian motion, a fuzzy statement made precise by Theorem 1.1.4. More generally, the universality class of random walks is the class of Lévy processes, while the universality class of branching processes is the class of *continuous-state branching processes* (CSBP). In contrast to random walks and branching processes, stochastic networks do not form such a clear-cut class of stochastic processes, but as the heavy traffic limit of Jackson networks and more generally of a broad class of open queueing networks, semimartingale reflected Brownian motions probably stand out as the universality class of stochastic networks. Besides making it possible to classify broad classes of stochastic processes, the fact that different processes have the same scaling limits make the latter play a prominent role in probability theory. For instance, the central role played by the Brownian motion in probability theory is to a large extent due to its prevalence in functional limit theorems: this generally justifies the in-depth studies of these "universal" objects such as Lévy processes, CSBP and semimartingale reflected Brownian motions. In summary, invariance principles make it possible to reduce the dimensionality of the problem by putting forward a few canonical stochastic processes.

From an application standpoint, such invariance principles make it possible to identify the key parameters of the model under study. For instance, the fact that the semimartingale reflected Brownian motion arising as the heavy traffic approximation of open queueing networks only depends on the first two moments of the arrival and service distributions makes it possible to dimension such networks simply based on these two parameters. In particular, one does not need to infer the whole distribution of these processes. Similarly, if one wishes to use Feller diffusion to model some particular population dynamic, one only needs to learn the first and second moments of the corresponding offspring distribution. Also, it must be noted that, although the critical case may appear as a very special case (indeed, it usually corresponds to a set of parameters of zero Lebesgue measure), it often turns out to be a relevant one in practice. For communication networks, an operator seeks to ensure finite delay while avoiding waisting capacity; for branching processes, the critical case can explain the long-time persistence of a population which in the subcritical or supercritical cases would die out quickly or grow exponentially fast. These two simple examples illustrate the fact that there are often good "physical" reasons for a system to be in the critical regime.

Continuous mapping

Another motivation for establishing functional limit theorems comes from the strong notion of convergence considered. Indeed, we are considering convergence in distribution of stochastic processes in the Skorohod J_1 topology: this is much stronger than, say, the notion of convergence of finite-dimensional distributions. One fundamental advantage of this form of convergence is that it makes it possible to use of the continuous mapping theorem: if $X_n \xrightarrow{\text{d}} X$, then $\phi(X_n) \xrightarrow{\text{d}} \phi(X)$ for any functional ϕ almost surely continuous at X. For instance, in the real-valued case and if X is almost surely continuous, then the convergence $X_n \xrightarrow{\text{d}} X$ automatically implies the convergence $\sup_{[0,t]} X_n \xrightarrow{\text{d}} \sup_{[0,t]} X$ for any $t \in \mathbb{R}_+$, a conclusion which does not necessarily hold if X_n merely converges to X in the sense of finite-dimensional distributions. Thus once the convergence $X_n \xrightarrow{\text{d}} X$ is established, one gets for free the convergence of a vast number of other functionals of X_n. This optimistic statement must nonetheless be tempered: in practice, it turns out (at least, in my experience) that most functionals ϕ of interest are not continuous, or that it is challenging to prove that they are almost continuous at the limit point.

Nevertheless, such an approach has proved immensely successful for proving functional limit theorems. For instance, as will be explained in Section 3.1.2, the convergence $X_n \xrightarrow{\text{d}} X$ with X_n a random walk and X a Lévy process implies, by a direct application of the continuous mapping theorem, the convergence of a sequence of Galton–Watson processes. Another typical application is to describe the dynamic of the pre-limit processes by stochastic differential equations, typically driven by Poisson processes, and use continuity arguments to show convergence toward a diffusion process which satisfies the corresponding limiting stochastic differential equations, typically driven by Brownian motions arising by compensating the pre-limit Poisson processes.

Besides, results of the kind $\phi(X_n)\xrightarrow{\text{d}}\phi(X)$ are very useful for deriving approximations because, in most cases, the limit process is more tractable than the pre-limit processes. First of all, there is a whole technical apparatus developed to study continuous processes: when in the realm of the law of large numbers, the limits are usually described by ordinary differential equations for which an extensive theory is available; while when in the real of the functional central limit theorem, the theory of stochastic calculus makes it possible to perform explicit computation. Moreover, the limit processes are often by nature more tractable. For instance, although there is in general no tractable expression for the law of the supremum of a stochastic process on the time interval $[0,t]$, for a Brownian motion we have

$$\mathbb{P}\left(\sup_{0\le s\le t} W(s)\ge x\right)=\sqrt{\frac{2}{\pi t}}\int_x^\infty e^{-y^2/(2t)}\mathrm{d}y.$$

The functional convergence $X_n\xrightarrow{\text{d}}W$ combined with the continuous mapping theorem therefore justifies the approximation

$$\mathbb{P}\left(\sup_{0\le s\le t} X_n(s)\ge x\right)\approx\sqrt{\frac{2}{\pi t}}\int_x^\infty e^{-y^2/(2t)}\mathrm{d}y.$$

Stationary distributions

Functional central limit theorems are concerned with the asymptotic behavior of processes over finite time intervals (at least, when considering the J_1 topology): however, they can also turn out to be useful for studying stationary distributions. In this case, one is faced with the problem of interchanging limits: starting from $X_n(t)$, if one first lets $n\to\infty$ and then $t\to\infty$, one obtains the stationary distribution of the scaling limit; while if one first lets $t\to\infty$ and then $n\to\infty$, one obtains the asymptotic behavior of the stationary distribution of X_n. Typically, directly studying the asymptotic behavior of the stationary distribution of X_n can be a challenging task whereas, in part because scaling limits are usually more tractable than pre-limit processes, the stationary distribution of the scaling limit is a reasonable object. One thus would like to know when these two limits commute in order to gain insight into the asymptotic behavior of the stationary distributions associated to the sequence (X_n). This approach has for instance been carried out to study the stationary distribution of generalized Jackson networks in heavy traffic in [38], where the interchange of limits was guaranteed by the existence of uniform geometric Lyapunov functions.

Fluid limits and stability of Markov processes

Last but not least, functional law of large numbers have gained in queueing theory a tremendous impulse since the pioneering work of Rybko and Stolyar [103] and then

Dai [27], showing that the stability of a Markov process can be inferred from the stability of the associated fluid model, see also the earlier work by Malyshev and Menshikov [88]. In this setting, one is interested in determining whether some Markov process X is recurrent or not. The idea is to consider a sequence (x_n) of initial states with $\|x_n\| \to \infty$ and to consider

$$\bar{X}_n(t) = \frac{1}{\|x_n\|} X_n\big(\|x_n\| t\big)$$

where X_n is the process started at $X_n(0) = x_n$. In many cases, the process $\bar{X}_n$ will converge in distribution toward a deterministic function, called the fluid limit, and governed by a deterministic dynamical system, called fluid model. Rybko and Stolyar then observed that stability of the original Markov process could be deduced from the stability of its fluid model, which opened the way to study the stability of more complex queueing systems, see for instance Bramson [23].

1.3 Notation used throughout the document

General notation

Throughout this document, $\mathbb{R} = (-\infty, +\infty)$ denotes the set of real numbers, $\mathbb{R}_+ = [0, \infty)$ the set of non-negative real numbers, $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\} = [-\infty, +\infty]$ and $\bar{\mathbb{R}}_+ = \mathbb{R}_+ \cup \{\infty\} = [0, \infty]$ the compactified versions of $\mathbb{R}$ and $\mathbb{R}_+$, respectively, $\mathbb{Z}$ the set of integers, $\mathbb{Z}_+$ the set of non-negative integers and $\mathbb{N}$ the set of positive integers. For a set $C \subset \mathbb{R}$, $\overline{C}$ denotes its closure, C^c its complement and $|C| \in \mathbb{Z}_+ \cup \{\infty\}$ its cardinality.

For $x, y \in \mathbb{R}$ we write $x \wedge y = \min(x, y)$ and $x \vee y = \max(x, y)$ for the minimum and maximum between x and y, respectively, and we let $[x] = \max\{n \in \mathbb{Z} : n \leq x\}$ and $x^+ = x \vee 0$ denote the integer and positive parts of x, respectively.

For a function f defined on $\mathbb{R}_+$, for simplicity we will sometimes write $(f(t))$ for $(f(t), t \in \mathbb{R}_+)$; likewise, we will sometimes simply denote by (u_n) a sequence u_n indexed by $\mathbb{Z}$, $\mathbb{Z}_+$ or $\mathbb{N}$.

With a slight abuse in notation, the L_1 norm on $\mathbb{R}^d$ for any $d \in \mathbb{N}$ will be denoted by $\|\cdot\|_1$, i.e., $\|x\|_1 = |x_1| + \cdots + |x_d|$ for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Moreover, for $f \in \mathcal{D}(\mathbb{R}^d)$ we will write $\|f\|_1$ for the function $\|f\|_1(t) = \|f(t)\|_1$, and when no confusion can occur we will also use of bold notation for the L_1 norm, i.e., $\mathbf{x} = \|x\|_1$ and $\mathbf{f} = \|f\|_1$. The L_∞ norm will be denoted by $\|\cdot\|_\infty$.

Measures

For a topological set E endowed with its Borel σ-algebra, let $\mathcal{M}(E)$ be the set of locally finite measures on E and $\mathcal{M}_p(E) \subset \mathcal{M}(E)$ be the subset of finite point measures. Let $\epsilon_x \in \mathcal{M}(E)$ be the unit mass at $x \in E$, $|\nu| = \nu(E) \in \bar{\mathbb{R}}_+$ for $\nu \in \mathcal{M}(E)$, be the mass of ν and $\mathbf{z}$ be the empty measure, i.e., the only measure with $|\mathbf{z}| = 0$. When $E = \mathbb{R}$ we will simply write $\nu[a, b]$ instead of $\nu([a, b])$, and likewise $\nu(a, b)$, $\nu\{a\}$, etc, and we define $\pi(\mathbf{z}) = 0$ and, for $\nu \neq \mathbf{z}$,

$$\pi(\nu) = \inf\{x \in \mathbb{R} : \nu(x, \infty) = 0\}$$

the supremum of the support of ν. The set $\mathcal{M}(E)$ can be endowed with two standard topologies: the weak and the vague topologies. A sequence of measures (ν_n) is said to converge to ν:

- weakly if $\nu_n(f) \to \nu(f)$ for every bounded continuous function $f : E \to \mathbb{R}$;

- vaguely if $\nu_n(f) \to \nu(f)$ for every continuous function $f : E \to \mathbb{R}$ with a compact support.

Endowed with any of these topologies and provided E satisfies mild assumptions (e.g., being locally compact and completely separable), $\mathscr{M}(E)$ is a Polish space. Although weaker (i.e., the sequence (ν_n) may converge vaguely but not weakly), the interest of the vague topology is that simple criteria for relative compactness are available.

Càdlàg functions, excursions and some functional operators

For a complete, separable space $\mathfrak{X}$ with a metric d, $\mathscr{D}(\mathfrak{X})$ denotes the set of càdlàg functions $f : \mathbb{R}_+ \to \mathfrak{X}$. The space $\mathscr{D}(\mathfrak{X})$ is endowed with the Skorohod J_1 topology, which makes it a complete and separable metric space, and we will use the symbol $\xrightarrow{J_1}$ to denote the convergence of a sequence of functions in this topology. We will sometimes consider a different domain for the functions under consideration, and for an interval $I \subset \bar{\mathbb{R}}_+$ we will denote by $\mathscr{D}_I(\mathfrak{X})$ the set of càdlàg functions $f : I \to \mathfrak{X}$. For $f \in \mathscr{D}(\mathfrak{X})$ and $t \in \mathbb{R}_+$, we consider $\Delta f(t) = d(f(t), f(t-))$ the value of the jump of f at $t \in \mathbb{R}_+$, $\theta_t(f)$ the function f shifted at time t and $\sigma_t(f)$ the function f stopped at time t, i.e., $\theta_t(f) = (f(t+s), s \in \mathbb{R}_+)$ and $\sigma_t(f) = (f(t \wedge s), s \in \mathbb{R}_+)$.

Note that, in the sequel, we will consider vector-valued functions, i.e., $\mathfrak{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$, or measure-valued functions, i.e., $\mathfrak{X} = \mathscr{M}(\mathbb{R})$.

When $\mathfrak{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$, $T_0(f) \in [0,\infty]$ for $f \in \mathscr{D}(\mathbb{R}^d)$ is the first hitting time of $0 \in \mathbb{R}^d$: $T_0(f) = \inf\{t > 0 : f(t) = 0\}$. The stopping and shift operators σ_t and θ_t will most of the time be considered at $t = T_0$, and so we will use the notation $\sigma = \sigma_{T_0}$ and $\theta = \theta_{T_0}$, i.e., $\sigma(f) = \sigma_{T_0(f)}(f)$ and $\theta(f) = \theta_{T_0(f)}(f)$. We call *excursion* a function $f \in \mathscr{D}(\mathbb{R}^d)$ with $f(t) = 0$ for every finite $t \geq T_0(f)$. The set of excursions will be denoted by $\mathscr{E}(\mathbb{R}^d)$ and we will call $T_0(e)$ the *length* of $e \in \mathscr{E}(\mathbb{R}^d)$ and $\sup_{t\geq 0} \|f(t)\|_1$ its *height*. We also consider $\mathbf{0} \in \mathscr{E}(\mathbb{R}^d)$ the function which takes constant value 0, i.e., $\mathbf{0}(t) = 0$ for $t \in \mathbb{R}_+$ and for $\varepsilon > 0$ the mappings $T_\varepsilon^\uparrow, T_\varepsilon^\downarrow : \mathscr{D}(\mathbb{R}^d) \to \mathscr{E}(\mathbb{R}^d)$ defined by

$$T_\varepsilon^\uparrow(f) = \inf\{t \geq 0 : \|f(t)\|_1 > \varepsilon\} \text{ and } T_\varepsilon^\downarrow(f) = \inf\{t \geq 0 : \|f(t)\|_1 < \varepsilon\}.$$

For $f \in \mathscr{D}(\mathbb{R}^d)$, we call $\mathscr{Z}(f) = \{t \geq 0 : f(t) = 0\}$ the *zero set of* f. The right-continuity of f implies that its complement $\mathscr{Z}^c = \mathbb{R}_+ \setminus \mathscr{Z}$ is a countable union of disjoint intervals of the form (g, d) or $[g, d)$ called *excursion intervals*, see Kallenberg [61, Chapter 22]. With every such interval, we may associate an excursion $e \in \mathscr{E}(\mathbb{R}^d)$, defined as the function $\sigma(\theta_g(f))$, i.e., $\theta_g(f)$ stopped at its first hitting time $T_0(e) = d - g$ of 0. We call g and d its *left* and *right endpoints*, respectively. Finally, for $t \geq 0$ we consider the measurable mapping $E_t^S : \mathscr{D}(\mathbb{R}^d) \to \mathscr{E}(\mathbb{R}^d)$ which to a function f associates the excursion straddling t; when $f(t) = 0$ this is to be understood as $E_t^S(f) = \mathbf{0}$.

Finally, when $d = 1$ we will consider the reflection operator $R : \mathscr{D}(\mathbb{R}) \to \mathscr{D}(\mathbb{R})$ defined by

$$R(f)(t) = f(t) - \min\left(0, \inf_{0 \leq s \leq t} f(s)\right), \ t \geq 0.$$

Canonical notation

The problems presented in this manuscript all deal with the convergence in distribution of stochastic processes with values in $\mathbb{R}^d$ for some $d \in \mathbb{N}$ (vector-valued processes) or in $\mathscr{M}(\mathbb{R})$ or $\mathscr{M}((0,\infty))$ (measure-valued processes). Formally, we are considering the weak convergence of sequences of probability measures on $\mathscr{D}(\mathfrak{X})$, with $\mathfrak{X} = \mathbb{R}^d$ or $\mathscr{M}(\mathbb{R})$, and we will use the canonical notation.

By this, we mean that we consider the measurable space $(\Omega, \mathscr{F})$ with $\Omega = \mathscr{D}(\mathfrak{X})$ and $\mathscr{F}$ its Borel σ-field. The canonical process will be denoted by $X : \Omega \to \Omega$, which is the function with $X(\omega) = \omega$ for $\omega \in \Omega$, and $X(t) : \omega \in \Omega \mapsto \omega(t) \in \mathfrak{X}$ denotes the canonical projection. Then all functionals will implicitly be assumed to be considered at X and we consider the filtration $\mathscr{F}_t = \sigma(X(s), 0 \leq s \leq t)$ generated by X. We will then use the subscript n to refer to the sequence of systems considered and denote by $\mathbb{P}_n$ the law of the unscaled process, i.e., on the normal time and space scales. Systems can be indexed by the initial condition (such as in the fluid regime, where we typically consider a sequence of initial states x_n of size n) or by the system's parameters (such as in the heavy traffic regime). We will reserve bold notation, e.g., $\mathbf{P}_n$, for the law of the nth system after normalization. For instance, in a fluid regime where the scaling in both time and space is equal to n, $\mathbf{P}_n$ will be the law of $(X(nt)/n, t \geq 0)$ under $\mathbb{P}_n$ whereas in a diffusive regime $\mathbf{P}_n$ will typically be the law of $(X(n^2 t)/n, t \geq 0)$ under $\mathbb{P}_n$. In particular, the scaling at hand depends on the context. The corresponding expectation is then denoted by $\mathbb{E}_n$ and $\mathbf{E}_n$, respectively. For M a measurable mapping defined on Ω, we will simply say that M is tight, resp. converges, if $(\mathbf{P}_n \circ M^{-1})$ is tight, resp. if $\mathbf{P}_n \circ M^{-1}$ converges weakly to $\mathbf{P} \circ M^{-1}$.

We will use the symbols $\xrightarrow{\mathrm{d}}$, $\xrightarrow{\mathrm{w}}$ and $\xrightarrow{\mathrm{fdd}}$ to denote convergence in distribution, weak convergence and convergence of the finite-dimensional distributions, respectively: whether we consider these convergences under $\mathbb{P}_n$ or under $\mathbf{P}_n$ should always be clear from the context.

Finally, we will also use $\mathbb{P}$ and $\mathbb{E}$ to denote the probability and expectation of generic random variables and processes, and $\xrightarrow{\mathrm{a.s.}}$ to denote almost sure convergence, under a probability distribution which will always be specified.

Excursion measures and Itô's construction

Regeneration of Markov processes is well-studied since Itô's seminal paper [54], see for instance Blumenthal [12]. However, existence of excursion measures which describe the behavior in distribution of a regenerative process can be defined beyond this case. Namely, we will say that a probability distribution $\mathbb{P}$ on $\mathscr{D}(\mathbb{R}^d)$ is regenerative (at 0) if there exists a measure P_0 such that for any stopping time τ,

$$\mathbb{P}\big(\theta_\tau \in \cdot \mid \mathscr{F}_\tau\big) = P_0, \ \ \mathbb{P}\text{-almost surely on } \{\tau < \infty, \ X(\tau) = 0\}. \tag{1.4}$$

It is known, see Kallenberg [61] for rigorous statements, that for any regenerative process, the distributional behavior of its excursions away from 0 can be characterized by a σ-finite measure $\mathscr{N}$ on $\mathscr{E}(\mathbb{R}^d) \setminus \{\mathbf{0}\}$ called *excursion measure*. Note that if holding times at 0 are nonzero, then by the regeneration property they must be exponentially distributed. Also, the excursion measure is uniquely determined up to a multiplicative constant and satisfies $\mathscr{N}(1 \wedge T_0) < \infty$.

On the other hand, starting from a σ-finite measure $\mathscr{N}$ on $\mathscr{E}(\mathbb{R}^d) \setminus \{\mathbf{0}\}$ satisfying $\mathscr{N}(1 \wedge T_0) < \infty$, Itô's construction gives a way to construct a regenerative process with excursion measure $\mathscr{N}$. More precisely, the construction starts from a $\{\partial\} \cup (\mathscr{E}(\mathbb{R}^d) \setminus \{\mathbf{0}\})$-valued Poisson point process $(\alpha_t, t \in \mathbb{R}_+)$ with intensity measure $\mathscr{N}$, where ∂ is a cemetery point which by convention satisfies $T_0(\partial) = 0$. Let $\mathtt{d} \geq 0$

and

$$Y(t) = \mathtt{d}t + \sum_{0 \le s \le t} T_0(\alpha_s), \ t \in \mathbb{R}_+. \tag{1.5}$$

Since $\mathscr{N}(1 \wedge T_0) < \infty$, Y is well-defined and is a subordinator with drift $\mathtt{d}$ and Lévy measure $\mathscr{N}(T_0 \in \cdot)$. Let Y^{-1} be its right-continuous inverse and define the process X as follows:

$$X(t) = \alpha_{Y^{-1}(t-)}\left(t - Y(Y^{-1}(t)-)\right)$$

for $t \ge 0$ such that $\Delta Y(Y^{-1}(t)) > 0$ and 0 otherwise. Then it X is a regenerative process with excursion measure $\mathscr{N}$.

To get some intuition into this formula, one can explain where it comes from when starting from X: if $X(t) \neq 0$ and g is the left endpoint of the excursion E^S_t straddling t, then we obviously have

$$X(t) = E^S_t(t - g).$$

The above formula somehow unifies this relation by indexing excursions by the local time at 0, which is given by the process Y^{-1}. We then define $\alpha_{Y^{-1}(t)}$ as the excursion straddling t and since $Y(Y^{-1}(t)-)$ is its left endpoint, this gives the relation $X(t) = \alpha_{Y^{-1}(t-)}\left(t - Y(Y^{-1}(t)-)\right)$. The idea of Itô's construction is to revert this construction by using this relation.

Chapter 2

Two theoretical results on weak convergence

Contents

2.1 Introduction 11
2.1.1 Tightness 11
2.1.2 Characterizing accumulation points 13
2.2 A sufficient condition for tightness 14
2.2.1 Main result 14
2.2.2 Work in progress 16
2.3 Weak convergence of regenerative processes 17
2.3.1 Introduction 17
2.3.2 Main result 17
2.3.3 Extensions 18

2.1 Introduction

The standard machinery for proving the weak convergence of a sequence of probability measures on a Polish space is to show tightness and then uniquely characterize the possible accumulation points: for both steps, a wealth of methods have been devised. In this manuscript we are more specifically interested in probability measures on the space $\mathscr{D}(\mathfrak{X})$ of càdlàg functions on some Polish space $\mathfrak{X}$, endowed with the Skorohod J_1 topology which makes it a Polish space as well. Let in the rest of this discussion d be a distance on $\mathfrak{X}$ which makes it separable and complete and $(\mathbf{P}_n)$ be an arbitrary sequence of probability measures on $\mathscr{D}(\mathfrak{X})$.

2.1.1 Tightness

Given $0 \leq A \leq B$, we say that a finite sequence $\mathbf{b} = (b_\ell, \ell = 0, \dots, L)$ such that $b_0 = A < b_1 < \cdots < b_L = B$ is a *subdivision of* $[A, B]$, which is δ-sparse for $\delta > 0$ if $b_{\ell+1} - b_\ell > \delta$ for every $\ell = 0, \dots, L-2$. For $\delta > 0$, $0 \leq A \leq B$ and $f \in \mathscr{D}(\mathfrak{X})$ let

$$w'([A,B],\delta)(f) = \inf_{\mathbf{b}} \max_{\ell=0,\dots,L-1} \sup_{b_\ell \leq s,t < b_{\ell+1}} d(f(s), f(t))$$

where the infimum extends over all subdivisions $\mathbf{b} = (b_\ell, \ell = 0, \dots, L)$ of $[A,B]$ which are δ-sparse. When $A = 0$ we will simply write $w'(T,\delta) = w'([0,T],\delta)$. The function w' can be seen as a modulus of "continuity" for càdlàg functions in the sense that $f : \mathbb{R}_+ \to \mathfrak{X}$ is càdlàg if and only if $w'(T,\delta)(f) \to 0$ as $\delta \to 0$ for every T in some dense subset of $\mathbb{R}_+$. The Arzelà–Ascoli theorem for characterizing a relatively compact sequence of càdlàg functions translates in the following probabilistic terms.

Theorem 2.1.1 (Theorem 16.8 in [11])**.** *The sequence* $(\mathbf{P}_n)$ *is tight if and only if the following two conditions hold:*

i) $X(t)$ *is tight for every* t *in a dense subset of*$\mathbb{R}_+$*;*

ii) *for every* $\eta > 0$ *and every* $T \geq 0$*,*

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbf{P}_n \left(w'(T,\delta) \geq \eta \right) = 0.$$

The sequence $(\mathbf{P}_n)$ is said to be C-tight if it is tight and any accumulation point only puts mass on continuous functions. A necessary and sufficient condition for $(\mathbf{P}_n)$ to be C-tight is provided by the above theorem, where w' needs to be replaced by w given by

$$w(T,\delta)(f) = \sup \left\{ d(f(t), f(s)) : 0 \leq s, t \leq T, |t-s| \leq \delta \right\}.$$

Again, w is a modulus of continuity for continuous functions in the sense that $f : \mathbb{R}_+ \to \mathfrak{X}$ is continuous if and only if $w(T,\delta)(f) \to 0$ as $\delta \to 0$ for every $T \in \mathbb{R}_+$. For $\beta > 0$ and $x, y \in \mathfrak{X}$ let $d_\beta(x,y) = 1 \wedge d(x,y)^\beta$. In presence of the following compact containment condition, Kurtz [74] proved another characterization of tightness.

Compact containment condition. *For every* $T, \eta > 0$*, there exists a compact set* K *such that*

$$\liminf_{n\to\infty} \mathbf{P}_n(X(t) \in K, 0 \leq t \leq T) \geq 1 - \eta.$$

Theorem 2.1.2 (Theorem 3.8.6 in [35])**.** *Assume that the compact containment condition holds. Then* $(\mathbf{P}_n)$ *is tight if and only if for every* $T > 0$*, there exist* $\beta > 0$ *and a family of random variables* $\{\gamma_n(\delta)\}$ *such that*

$$\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbf{E}_n(\gamma_n(\delta)) = \lim_{\delta\to 0} \limsup_{n\to\infty} \mathbf{E}_n \left(d_\beta(X(\delta), X(0)) \right) = 0$$

and

$$\mathbf{E}_n \left(d_\beta(X(u), X(t)) \mid \mathscr{F}_t \right) d_\beta(X(t), X(s)) \leq \mathbf{E}_n \left(\gamma_n(\delta) \mid \mathscr{F}_t \right)$$

for all $0 \leq s \leq t \leq u \leq T$ *satisfying* $u - t \leq \delta$ *and* $t - s \leq 2\delta$*.*

The above characterizations of tightness are of great conceptual importance, but are often difficult to work with in practice. For this reason, various more practical sufficient conditions for proving tightness have been proposed. For instance, by considering $\gamma_n(\delta) = w(T,\delta)(F_n)$ Theorem 2.1.2 entails the following useful result.

Theorem 2.1.3. *Assume that the compact containment condition holds. Then* $(\mathbf{P}_n)$ *is tight if for every* $T > 0$ *there exist* $\beta > 0$*, a sequence of non-decreasing, càdlàg functions* (F_n) *and* F *continuous such that* $F_n \xrightarrow{J_1} F$ *and*

$$\mathbf{E}_n \left(d_\beta(X(t), X(s)) \mid \mathscr{F}_s \right) \leq F_n(t) - F_n(s), \ n \in \mathbb{N}, 0 \leq s \leq t \leq T.$$

When $\mathbf{P}_n$ is known to converge in the sense of finite-dimensional distributions, the following result which can be found in Billingsley [11] states another simple condition on oscillations of X that ensure tightness, and hence weak convergence. We used this condition for proving the tightness of binary, homogeneous Crump–Mode–Jagers processes in [LS15], see Section 3.3.

Theorem 2.1.4 (Theorem 13.5 in [11])**.** *If the following two conditions hold, then* $\mathbf{P}_n \xrightarrow{\mathrm{w}} \mathbf{P}$*:*

i) $(X(t), t \in I)$ *converges for every finite subset* I *of* $\{t \in \mathbb{R}_+ : \mathbf{P}(\Delta X(t) \neq 0) = 0\}$*;*

ii) *there exist* $\beta \geq 0$*,* $\alpha > 1$ *and* F *non-decreasing and continuous such that for every* $0 \leq s \leq t \leq u$*,* $n \in \mathbb{N}$ *and* $\lambda > 0$*,*

$$\mathbf{P}_n\big(d(X(u), X(t)) \wedge d(X(t), X(s)) \geq \lambda\big) \leq \frac{1}{\lambda^\beta}\big(F(u) - F(s)\big)^\alpha. \tag{2.1}$$

Noting that the condition

$$\mathbf{E}_n\big(d_\beta(X(u), X(t)) d_\beta(X(t), X(s))\big) \leq \big(F(u) - F(s)\big)^\alpha, \; n \in \mathbb{N}, 0 \leq s \leq t \leq u,$$

implies (2.1) by Markov inequality, we see that Theorems 2.1.3 and 2.1.4 have a similar flavor. There are nonetheless noticeable differences, for instance the fact that the estimate in (2.1) is uniform in $n \in \mathbb{N}$, and also that one needs $\alpha > 1$. These two facts yielded significant technical difficulties in our study of Crump–Mode–Jagers processes in [LS15].

The above results aim at controlling the increments of the process at deterministic times. In a different line of thought, Aldous [1] proposed a sufficient condition in terms of stopping times, which became known as Aldous' criterion. This approach is especially useful when working with Markov processes and more generally with semimartingales, since there is a wealth of results available to control these processes stopped at stopping times. Moreover, Rebolledo [97] showed that the tightness of a sequence of semimartingales boils down to the tightness of the processes appearing in its canonical semimartingale decomposition, and so combining these two results yields an efficient way to prove tightness of a sequence of semimartingales. This approach is sometimes referred to as the Aldous–Rebolledo condition.

2.1.2 Characterizing accumulation points

To characterize accumulation points of $(\mathbf{P}_n)$, the standard approach consists in showing that finite-dimensional distributions converge. However, as pointed out in the introduction of Jacod and Shiryaev [56], this is "very often […] a very difficult (or simply impossible) task to accomplish" and so alternative approaches are called upon.

In [56], it is proposed to characterize accumulation points through the method of characteristic triplets: informally, a semimartingale is characterized by a characteristic triplet and, under proper assumptions, it is shown in [56] that a sequence of semimartingales converges if and only if the associated sequence of characteristic triplets converges. This approach is especially efficient for a sequence of Markov processes since their semimartingale structure is well understood. Typically, if $\mathbf{P}_n$ is the law of a Markov process with infinitesimal generator L_n, then for suitable test functions $f : \mathfrak{X} \to \mathbb{R}$ the process

$$\left(f(X(t)) - f(X(0)) - \int_0^t L_n(f)(X(u))\mathrm{d}u, t \in \mathbb{R}_+\right)$$

is a $\mathbf{P}_n$-local martingale with predictable quadratic covariation process

$$\left(\int_0^t \Gamma_n(f)(X(u))\mathrm{d}u, t \in \mathbb{R}_+\right)$$

where $\Gamma_n = L_n(f^2) - 2fL_n(f)$ is the so-called "carré du champ" operator. In my experience, this systematic approach to study Markov processes is very powerful but is not well known among "applied" researchers for whom it would be extremely useful. For instance, the only textbook that I am aware of where this is mentioned is Jacod and Shiryaev [56], and more precisely in Lemma VIII.3.68, where it is mentioned that this result is "well known to those familiar with Markov processes".

Another very important approach to characterize accumulation points (which often also takes care of tightness) is the continuous mapping approach. Typically, one expresses the dynamic of the pre-limit processes, i.e., X under $\mathbf{P}_n$, in the form $X(t) = \Phi_n(X(s), 0 \leq s \leq t)$ for some mapping Φ_n, and provided $\Phi_n \to \Phi$ in a suitable sense, one hopes that X under $\mathbf{P}_n$ converges in distribution to a process governed by the dynamic $Y(t) = \Phi(Y(s), 0 \leq s \leq t)$. For instance, if X is a continuous-time random walk, then $R(X)$ is an $M/M/1$ queue. Using continuity property of the reflection operator, one gets that the scaling limit of the $M/M/1$ queue is a reflected Brownian motion.

2.2 A sufficient condition for tightness

2.2.1 Main result

As already pointed out in [55], one of the limitations of Aldous' criterion is that accumulation points must be laws of processes which are quasi-left-continuous; due to the fact that the function F in Theorems 2.1.3 and 2.1.4 is required to be continuous, this is also the case for these two results. Assuming quasi-left-continuity is a reasonable assumption met in most practical cases, but the case of processes in varying environments is one example where this assumption is too demanding. Indeed, as explained in the introduction scaling limits are usually established in the critical regime. But when considering varying environments, there is a priori no reason to exclude non-critical environments to occur from time to time, as long as "most" environments are still critical.

To give a specific example, consider Galton–Watson processes in varying environments, studied in Section 3.2 and which actually motivated the forthcoming general result. For instance, if one starts with a critical Galton–Watson process and just changes one offspring distribution to make it non-critical, the strong law of large numbers shows that this will induce a deterministic and multiplicative jump for the limiting process. Typically, Aldous' criterion or Theorems 2.1.3 and 2.1.4 above fail in such case.

Of course, in this simple example the natural idea is to study the process before and after the fixed time of discontinuity and then "glue" the pieces together, invoking for instance Lemma 2.2 in Whitt [120]. However, one may push this example further and naturally be lead to consider processes where the set of fixed times of discontinuity $\{t \in \mathbb{R}_+ : \mathbf{P}(\Delta X(t) > 0) > 0\}$ is dense in $\mathbb{R}_+$, in which case it is not clear at all how to carry out such a program. For strong Markov processes, we proved in [BS] the following result which covers such a scenario.

Theorem 2.2.1. *Assume that $\mathbf{P}_n$ for each $n \in \mathbb{N}$ is the law of a strong Markov process, and that the compact containment condition holds. Then the sequence $(\mathbf{P}_n)$ is tight if for every $T > 0$ and every compact $K \subset \mathfrak{X}$ there exist $\eta_0 > 0$ and non-decreasing càdlàg functions F_n and F with $F_n \xrightarrow{J_1} F$ and such that for every $n \in \mathbb{N}$, every $x \in K$ and every $0 \le s \le t \le T$ with $F_n(t) - F_n(s) \le \eta_0$,*

$$\mathbf{E}_n\big(d_2(X(t), X(s)) \mid X(s) \in K\big) \le F_n(t) - F_n(s).$$

The key point of the above statement is that the function F is not assumed to be continuous which, as explained above, makes it possible for the set of fixed times of discontinuity of accumulation points of $\mathbf{P}_n$ to be dense in $\mathbb{R}_+$.

Moreover, the functions F_n and F are allowed to depend on the compact K chosen, and the oscillations need only be controlled for points s and t which are close, in the sense that $F_n(t) - F_n(s) \le \eta_0$ for some constant η_0 which is allowed to depend on K. As mentioned earlier, this result and in particular these two extensions made it possible to establish the tightness of a sequence of Galton–Watson processes in varying environments in [BS15].

In order to prove Theorem 2.2.1, we used the first-principle characterization of tightness via the modulus of continuity w'. The general idea to control $w'(T,\eta)$ is to build a good subdivision relying on the discontinuities of F. The construction involves two mains steps:

(a) identify a *deterministic* subdivision $\mathbf{b}^n = (b^n_\ell, \ell = 0, \dots, L-1)$ which, heuristically, avoids the deterministic jumps of F (recall that if F is continuous, we can essentially invoke Theorem 2.1.3);

(b) within each subinterval $[b^n_\ell, b^n_{\ell+1})$, build a *random* sequence $(\Upsilon^n_\ell(i), i \ge 0)$ which controls the oscillations of X under $\mathbf{P}_n$ on this time interval.

To give a flavor of the construction, let us explain how the subdivisions $\mathbf{b}^n$ are constructed. Throughout, we fix some $T, \varepsilon > 0$: then the càdlàg nature of F ensures the existence of a subdivision $\mathbf{b} = (b_\ell, \ell = 0, \dots, L)$ of $[0, T]$ such that

$$\max_{0 \le \ell < L} \big(F(b_{\ell+1}-) - F(b_\ell)\big) \le \frac{\varepsilon^3}{(1+F(T))^{1/3}} \quad \text{and} \quad \sum_{0 \le y < T : y \notin \mathbf{b}} \Delta F(y) \le \varepsilon^3.$$

This is what was meant earlier by "avoiding the deterministic jumps of F". Further, the convergence $F_n \xrightarrow{J_1} F$ ensures the existence of a sequence of subdivision $(\mathbf{b}^n)$ with $\mathbf{b}^n = (b^n_\ell, \ell = 0, \dots, L)$ such that for each $\ell = 0, \dots, L$, we have $b^n_\ell \to b_\ell$, $F_n(b^n_\ell) \to F(b_\ell)$ and $F_n(b^n_\ell -) \to F(b_\ell -)$.

Once we have the deterministic subdivision $\mathbf{b}^n$, the construction of each sequence $(\Upsilon^n_\ell(i), i \ge 0)$ borrows classical ideas to prove tightness using stopping time arguments. Namely, we define $\Upsilon^n_\ell(0) = b^n_\ell$ and for $i \in \mathbb{Z}_+$, $\Upsilon^n_\ell(i+1) = \infty$ if $\Upsilon^n_\ell(i) = \infty$, and otherwise

$$\Upsilon^n_\ell(i+1) = \inf\big\{y \ge \Upsilon^n_\ell(i) : d\big(X^n_\ell(y), X^n_\ell(\Upsilon^n_\ell(i))\big) \ge \varepsilon\big\}$$

where X^n_ℓ is the process defined by

$$X^n_\ell(t) = X^n(t \wedge b^n_{\ell+1}) - \Delta X^n(b^n_{\ell+1}) \mathbb{1}_{\{t \ge b^n_{\ell+1}\}} = \begin{cases} X^n(t) & \text{if } t < b^n_{\ell+1}, \\ X^n(b^n_{\ell+1}-) & \text{if } t \ge b^n_{\ell+1}. \end{cases}$$

From there, we essentially follow the proof of Theorem 16.10 in Kallenberg [61] taking care, along the way, of the new technical difficulties created by imposing the deterministic “skeleton” $\mathbf{b}^n$.

2.2.2 Work in progress

After [BS] was submitted, we came in touch with T. Kurtz who suggested a very short proof of the following result closely related to Theorem 2.2.1. It is stronger in some aspects – e.g., the process need not be Markov – and weaker in others – e.g., the functions F_n are not allowed to depend on K. As will be seen next, the proof is significantly shorter than our initial approach for proving Theorem 2.2.1, and the goal of the on-going work [BKS] is to extend the following proof to obtain a strict generalization of Theorem 2.2.1.

Theorem 2.2.2. *Assume that the compact containment condition holds. Then the sequence* $(\mathbf{P}_n)$ *is tight if for every* $T > 0$*, there exist* $\beta > 0$ *and a relatively compact sequence of non-decreasing càdlàg functions* (F_n) *such that for every* $n \in \mathbb{N}$ *and every* $0 \le s \le t \le T$*,*

$$\mathbf{E}_n\big(d_\beta(X(t), X(s)) \mid \mathscr{F}_s\big) \le F_n(t) - F_n(s). \tag{2.2}$$

The proof of this result uses the approach developed in Kurtz [76]. The general idea is to consider X time-changed by the inverse of F_n, i.e., $Y_n = X \circ F^{-1}$ with $F_n^{-1}(t) = \inf\{s \ge 0 : F_n(s) > t\}$. Then we actually have $X = Y_n \circ F_n$, and the goal is to show that Y_n under $\mathbf{P}_n$ is tight and to deduce the tightness of $\mathbf{P}_n$ from this.

For the tightness of Y_n, note that the bound (2.2) translates to

$$\mathbf{E}_n\big(d_\beta(Y_n(t), Y_n(s)) \mid \mathscr{F}_{F_n^{-1}(s)}\big) \le G_n(t) - G_n(s)$$

with $G_n = F_n \circ F_n^{-1}$, which implies in particular that $d_\beta(Y_n(t), Y_n(s)) \le \mathbb{1}_{\{G_n(t) > G_n(s)\}}$. For any $0 \le s \le t \le u$ we thus have

$$\mathbf{E}_n\left[d_\beta(Y_n(u), Y_n(t)) \mid \mathscr{F}_{F_n^{-1}(t)}\right] d_\beta(Y_n(t), Y_n(s)) \le (G_n(u) - G_n(t))\, \mathbb{1}_{\{G_n(t) > G_n(s)\}}.$$

Lemma 2.5 in Kurtz [76] implies that $G_n(t) \le t$ and that if $G_n(t) > G_n(s)$, then $G_n(t) > s$: therefore,

$$\mathbf{E}_n\left[d_\beta(Y_n(u), Y_n(t)) \mid \mathscr{F}_{F_n^{-1}(t)}\right] d_\beta(Y_n(t), Y_n(s)) \le u - s.$$

From this bound, it is then not hard to use Theorem 2.1.2 to show that Y_n under $\mathbf{P}_n$ is tight.

We now have the representation $X = Y_n \circ F_n$ with Y_n tight. To show the tightness of $\mathbf{P}_n$, assume without loss of generality (by working along appropriate subsequences and using Skorohod representation theorem) that $Y_n \overset{\text{a.s.}}{\to} Y$. Delving into the dynamic of F and its inverse and the definition of Y_n, it can then be shown that Y is constant on any interval $[u, v]$ on which F^{-1} is constant, except for maybe one jump. Lemma 2.3(b) in Kurtz [76] then implies the convergence $X_n = Y_n \circ F_n \overset{\text{a.s.}}{\to} Y \circ F$ which yields the desired tightness of $\mathbf{P}_n$.

2.3 Weak convergence of regenerative processes

In this section we present the method proposed in [LS14] for characterizing the weak convergence of regenerative processes through the asymptotic behavior of their excursion measures. For simplicity, we restrict ourselves in this manuscript to functions with values in $\mathbb{R}^d$ endowed with any norm $\|\cdot\|$, but the following results have been proved in [LS14] for functions taking values in an arbitrary Polish space $\mathfrak{X}$.

2.3.1 Introduction

The law of a regenerative process is characterized by its excursion measure together with some non-negative parameter: heuristically, the excursion measure characterizes the law of the excursions and the real-valued parameter the amount of time the process spends at 0 (through the parameter $\mathtt{d}$ in formula (1.5)). The question that we address here is the extent to which the convergence of the excursion measures associated to a sequence of regenerative processes characterizes the convergence of the processes themselves.

Some special care is needed when dealing with sequences of excursion measures. Indeed, although the framework for the weak convergence of measures is well-studied for finite measures, see for instance Billingsley [11], the technical apparatus available to study σ-finite measures such as the excursion measures we will be interested in is more limited. For this reason, we will be interested in the weak convergence of probability measures obtained by conditioning the excursion measures. This approach also has a natural sample-path interpretation that can be illustrated by considering a sequence of renormalized random walks converging to Brownian motion.

The first excursion of the random walks away from 0 converges in distribution to $\mathbf{0}$, the trivial excursion which takes the constant value 0. This behavior is typical in the case where the excursion measure of the limit process – here, the Brownian motion – has infinite mass. However, this trivial behavior disappears under a suitable conditioning: for instance, the first excursion of the random walk with length greater than $\varepsilon > 0$, which will be called *big excursion*, converges in distribution to the first big excursion of Brownian motion. Since this holds for any $\varepsilon > 0$, one may hope that this convergence characterizes the convergence of the whole process and Theorem 2.3.1 below provides sufficient conditions for such a statement to hold. It must be noted that the convergence of big excursions does not in general imply tightness, and so Theorem 2.3.1 below can be seen as a new way of characterizing accumulation points, see Section 2.3.3 for more details.

The following approach was initially motivated by the study of the Processor-Sharing queue, where one naturally controls excursions of the queue length process via intimate connections with Lévy and Crump–Mode–Jagers processes, see Sections 3.3 and 4.2. It was therefore a natural question to understand the extent to which this characterized the convergence of the whole process. This method also turned out to be useful for studying the stochastic network of Section 4.3.

2.3.2 Main result

We fix a measurable map $\varphi : \mathscr{E}(\mathbb{R}^d) \to [0, \infty]$ such that $e \neq \mathbf{0}$ if and only if $\varphi(e) > 0$ and call $\varphi(e)$ the *size* of the excursion e. For each $\varepsilon > 0$, we say that $e \in \mathscr{E}(\mathbb{R}^d)$ is ε*-big*, or just *big* if the context is unambiguous, if its size is strictly larger than ε, and ε*-small* or *small* otherwise. We denote by $\mathscr{D}_\varphi(\mathbb{R}^d) \subset \mathscr{D}(\mathbb{R}^d)$ the set of càdlàg functions f such that the number of ε-big excursions is locally finite for any $\varepsilon > 0$:

$$\mathscr{D}_\varphi(\mathbb{R}^d) = \Big\{ f \in \mathscr{D}(\mathbb{R}^d) : \forall \varepsilon > 0, \forall t > 0, \text{ the number of } \varepsilon\text{-big excursions of } f \\ \text{starting before } t \text{ is finite} \Big\}.$$

Then we can define $e_\varepsilon(f) \in \mathscr{E}(\mathbb{R}^d)$ for $f \in \mathscr{D}_\varphi(\mathbb{R}^d)$ as the first excursion e of f satisfying $\varphi(e) > \varepsilon$ and $g_\varepsilon(f) \in [0,\infty]$ as its left endpoint, with the convention $(g_\varepsilon, e_\varepsilon)(f) = (\infty, \mathbf{0})$ if no such excursion exists. The maps e_ε and g_ε are measurable maps from $\mathscr{D}_\varphi(\mathbb{R}^d)$ to $\mathscr{E}(\mathbb{R}^d)$ and $[0,\infty]$, respectively.

In this section, $\mathbf{P}_n$ and $\mathbf{P}$ are regenerative measures and we denote by $\mathscr{N}_n$ and $\mathscr{N}$ their respective excursion measures. Moreover, it is assumed that $\mathbf{P}_n(\mathscr{D}_\varphi(\mathbb{R}^d)) = \mathbf{P}(\mathscr{D}_\varphi(\mathbb{R}^d)) = 1$ and that $\mathscr{N}$ has infinite mass. Recall that, if M is a measurable mapping defined on $\mathscr{D}(\mathbb{R}^d)$, we say that (the sequence) M converges if $\mathbf{P}_n \circ M^{-1} \xrightarrow{\text{w}} \mathbf{P} \circ M^{-1}$, and that M is tight if $(\mathbf{P}_n \circ M^{-1})$ is tight.

Theorem 2.3.1. *Let $C \subset (0,\infty)$ be such that $0 \in \overline{C}$ and $\mathscr{N}(\varphi = \varepsilon) = 0$ for all $\varepsilon \in C$. If the sequence $(\mathbf{P}_n)$ is tight and for every $\varepsilon \in C$ each of g_ε, $T_0 \circ e_\varepsilon$ and $(e_\varepsilon, \varphi \circ e_\varepsilon)$ converges, then $\mathbf{P}_n \xrightarrow{\text{w}} \mathbf{P}$.*

The proof of Theorem 2.3.1 essentially relies on continuity properties of some truncation and concatenation operators which make it possible to identify accumulation points of the tight sequence $(\mathbf{P}_n)$. These operators are not continuous in general, so that it is not at all natural, and even less obvious, that the previous theorem holds under such minimal assumptions. For instance, unaware of our work Yano [122] proved a result similar to ours, but assuming specific continuity properties on accumulation points of $\mathbf{P}_n$. Before proceeding to a discussion of Theorem 2.3.1, we make three remarks.

First, the conditions of this result are sharp, in the sense that if one of the assumptions is removed the conclusion $\mathbf{P}_n \xrightarrow{\text{w}} \mathbf{P}$ does not necessary hold.

Second, in the case $\varphi = T_0$, then instead of assuming the convergence of $(e_\varepsilon, \varphi \circ e_\varepsilon)$ one only needs to assume the convergence of e_ε. Moreover, T_0 is in general not a continuous operator, and so the convergence of $T_0 \circ e_\varepsilon$ does not automatically follow from the convergence of e_ε.

Finally, the excursion measure characterizes the "law" of excursions of a regenerative process. As mentioned in the introduction, it does not characterize the law of the whole process, since for instance it contains no information on the amount of time spent at 0. The assumption on g_ε somehow contains this additional information.

2.3.3 Extensions

Tightness

Let $\mathbf{P}$ be the law of a standard Brownian motion, and for $n \in \mathbb{N}$ define $\mathbf{P}_n$ the law of X_n under $\mathbf{P}$, where X_n is obtained from X by replacing each excursion of X with length $\zeta \in [1/n, 2/n]$ by a triangle with height n and basis ζ. Then $(\mathbf{P}_n)$ is not tight whereas for every $\varepsilon > 0$ and for the choice $\varphi = T_0$, the three sequences g_ε, $T_0 \circ e_\varepsilon$ and e_ε converge. This example therefore shows that assuming $(\mathbf{P}_n)$ to be tight is not superfluous.

Moreover, this example intuitively suggests that when big excursions converge, one is essentially left with the problem of controlling the height of small excursions. Indeed, tightness of a sequence of càdlàg paths is essentially concerned with controlling oscillations, and for small excursions we do not lose much by upper bounding their oscillations by their height: Lemma 2.3.2 below formalizes this idea.

For $\varepsilon > 0$, let $\Phi_\varepsilon : \mathscr{D}(\mathbb{R}^d) \to \mathscr{D}(\mathbb{R}^d)$ be the mapping which to a function f associates the function where ε-big excursions are truncated. More precisely, for $f \in \mathscr{D}(\mathbb{R}^d)$ and $t \in \mathbb{R}_+$, we define (recall that $E^S_t(f)$ is the excursion of f straddling t)

$$\Phi_\varepsilon(f)(t) = \begin{cases} f(t) & \text{if } \varphi(E^S_t(f)) \le \varepsilon, \\ 0 & \text{else.} \end{cases}$$

For $f \in \mathscr{D}(\mathbb{R}^d)$ and $t \in \mathbb{R}_+$, let $\|f\|_1^{*,t} = \sup_{0 \le s \le t} \|f(s)\|_1$ and $\|f\|_1^* = \sup_{s \ge 0} \|f(s)\|_1$.

Lemma 2.3.2. *Consider some set $C \subset (0,\infty)$ such that $0 \in \overline{C}$ and assume that for every $\varepsilon \in C$, the three sequences g_ε, e_ε and $T_0 \circ e_\varepsilon$ are tight and that every accumulation point $(\gamma_\varepsilon, \tau_\varepsilon) \in [0,\infty]^2$ of the sequence $(g_\varepsilon, T_0 \circ e_\varepsilon)$ satisfies $\mathbb{P}(\gamma_\varepsilon = 0) = \mathbb{P}(\tau_\varepsilon = 0) = 0$. Then for any m and $\eta > 0$,*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbf{P}_n \left(w'(m, \delta) \ge 4\eta \right) \le \lim_{\varepsilon \to 0} \limsup_{n \to \infty} \mathbf{P}_n \left(\|\Phi_\varepsilon\|_1^{*,m} \ge \eta \right). \tag{2.3}$$

If in addition, for each $\varepsilon \in C$ the sequence e_ε is C-tight and all its accumulation points almost surely start at 0, then w' in the left hand side of (2.3) can be replaced by w.

Finally, if $\varphi(e) = \sup_{t \ge 0} \|e(t)\|$ for any norm $\|\cdot\|$, then the above assumptions imply that $(\mathbf{P}_n)$ is tight, and even C-tight if e_ε is C-tight and all its accumulation points almost surely start at 0.

Note that the right-hand side of (2.3) precisely means that the problem of controlling the modulus of continuity reduces to the problem of studying the height of small excursions, once we know that big excursions as well as their left endpoint and length are tight. Lemma 2.3.2 states in particular that the sequence $(\mathbf{P}_n)$ is automatically tight when considering φ of the form $\varphi(e) = \sup_{t \in \mathbb{R}_+} \|e(t)\|$ for some norm $\|\cdot\|$, so that we obtain the following simple version of Theorem 2.3.1.

Theorem 2.3.3. *Let φ be of the form $\varphi(e) = \sup_{t \in \mathbb{R}_+} \|e(t)\|$ for some norm $\|\cdot\|$ and $C \subset (0,\infty)$ such that $0 \in \overline{C}$ and $\mathscr{N}(\varphi = \varepsilon) = 0$ for all $\varepsilon \in C$. If for every $\varepsilon \in C$ the three sequences g_ε, e_ε and $T_0 \circ e_\varepsilon$ converge, then $\mathbf{P}_n \xrightarrow{\text{w}} \mathbf{P}$.*

Excursion measures

We now discuss various conditions on $\mathscr{N}_n$, an excursion measure of X_n, that guarantee that parts of the assumptions of Theorem 2.3.1 hold. In the rest of this subsection, we assume that $\mathscr{N}_n$ has finite mass and we denote by b_n the parameter of the exponentially distributed holding times at 0. We also assume that we are given a sequence (c_n) of positive real numbers such that $c_n \to \infty$ and (within this section) we define the set C as the complement of the set of atoms of $\mathscr{N} \circ \varphi^{-1}$:

$$C = \left\{ \varepsilon > 0 : \mathscr{N}(\varphi = \varepsilon) = 0 \right\}.$$

We will consider the following assumptions:

(H1) $c_n\mathcal{N}_n(\varphi > \varepsilon) \to \mathcal{N}(\varphi > \varepsilon)$ for every $\varepsilon \in C$;
(H2) $c_n\mathcal{N}_n(1 - e^{-\lambda T_0}; \varphi \le \varepsilon) \to \mathcal{N}(1 - e^{-\lambda T_0}; \varphi \le \varepsilon)$ for every $\lambda > 0$ and $\varepsilon \in C$;
(H3) $c_n\mathcal{N}_n\left(1 - e^{-\lambda T_0}\right) \to \mathcal{N}\left(1 - e^{-\lambda T_0}\right)$ for every $\lambda > 0$.

It is easy to see that (H1) and (H2) together imply (H3) when we have in addition $\mathcal{N}(\varphi = \infty) = 0$. Moreover, (H1), (H2) and (H3) together give the convergence, for $\varepsilon \in C$, of the two sequences $\varphi \circ e_\varepsilon$ and $T \circ e_\varepsilon$. Indeed, since

$$\mathbf{P}_n\left(\varphi \circ e_\varepsilon > \delta\right) = \frac{\mathcal{N}_n\left(\varphi > \delta \vee \varepsilon\right)}{\mathcal{N}_n\left(\varphi > \varepsilon\right)} \text{ and } \mathbf{E}_n\left(1 - e^{-\lambda T_0 \circ e_\varepsilon}\right) = \frac{\mathcal{N}_n\left(1 - e^{-\lambda T_0}; \varphi > \varepsilon\right)}{\mathcal{N}_n(\varphi > \varepsilon)}$$

we have the following elementary result.

Lemma 2.3.4. *If* (H1) *holds, then $\varphi \circ e_\varepsilon$ converges for every $\varepsilon \in C$. If* (H1), (H2) *and* (H3) *hold, then $T_0 \circ e_\varepsilon$ converges for every $\varepsilon \in C$.*

These assumptions also imply the convergence of g_ε under an additional assumption on b_n and c_n. Note that g_ε plays a particular role in the conditions of Theorem 2.3.1, since by regeneration g_ε is independent of e_ε, whereas in contrast $T \circ e_\varepsilon$ and $\varphi \circ e_\varepsilon$ are completely determined by e_ε. In particular, one has to prove the convergence of g_ε separately in order to check the assumptions of Theorem 2.3.1.

We stress that the excursion measure $\mathcal{N}$ is uniquely defined only after the local time has been normalized. A local time process $(L(t), t \ge 0)$ is a nondecreasing process with values in $\mathbb{R}_+$, which satisfies $L(0) = 0$ and for any $t > s$, $L(t) > L(s)$ if and only if there is $u \in (s, t)$ such that $X(u) = 0$. It is unique up to a multiplicative constant. It is known (see Kallenberg [61]) that the inverse of L is a subordinator, i.e., an increasing Lévy process. We let $\mathtt{d}$ be the drift coefficient of this subordinator. In particular, the zero set of X has positive Lebesgue measure if and only if $\mathtt{d} > 0$.

Lemma 2.3.5. *If* (H1) *and* (H2) *hold and $c_n/b_n \to \mathtt{d}$, then g_ε converges for every $\varepsilon \in C$.*

In summary, if assumptions (H1), (H2) and (H3) hold and $c_n/b_n \to \mathtt{d}$, then we get the convergence of g_ε, $T_0 \circ e_\varepsilon$ and $\varphi \circ e_\varepsilon$. We now state a result which shows how these assumptions can be used for tightness: recall that, according to Lemma 2.3.2, we are left with controlling the height of small excursions. In the following lemma, $\mathcal{N}_n(\cdot \mid T_0 = \infty)$ refers to the null measure when $\mathcal{N}_n(T_0 = \infty) = 0$.

Lemma 2.3.6. *For any $n \in \mathbb{N}$, $m \in \mathbb{R}_+$ and $\varepsilon, \eta, \lambda$ and $\alpha > 0$, it holds that*

$$\begin{aligned}\mathbf{P}_n\left(\|\Phi_\varepsilon\|_1^{*,m} \ge \eta\right) \le e^{\lambda m} \exp\left(-\frac{\lfloor \alpha c_n \rfloor \lambda}{\lambda + b_n} - \lfloor \alpha c_n \rfloor \mathcal{N}_n\left(1 - e^{-\lambda T_0} \mid T_0 < \infty\right)\right)\\ + \alpha c_n \mathcal{N}_n\left(\|X\|_1^* \ge \eta \mid \varphi \le \varepsilon, T_0 < \infty\right) + \mathcal{N}_n\left(\|X\|_1^{*,m} \ge \eta, \varphi \le \varepsilon \mid T_0 = \infty\right). \quad (2.4)\end{aligned}$$

Combining the previous results, we end up with the following result.

Theorem 2.3.7. *Assume that $\mathcal{Z}(X)$ has $\mathbf{P}$-almost surely zero Lebesgue measure, that $\mathcal{N}(\varphi = \infty) = 0$, that* (H1) *and* (H2) *hold and that $c_n/b_n \to 0$. Then for every $\varepsilon \in C$, each of g_ε, $T_0 \circ e_\varepsilon$ and $\varphi \circ e_\varepsilon$ converges.*

Assume in addition that $\mathcal{N}_n(T_0 = \infty) = 0$ for every $n \ge 1$, that $(e_\varepsilon, \varphi \circ e_\varepsilon)$ converges for every $\varepsilon \in C$ and that

$$\limsup_{n\to\infty} c_n \mathcal{N}_n\left(\|X\|_1^* \ge \eta, \varphi \le \varepsilon\right) \xrightarrow[\varepsilon \to 0]{} 0 \quad (2.5)$$

for every $\eta > 0$. Then $\mathbf{P}_n \xrightarrow{w} \mathbf{P}$.

Shifting

For a strong Markov process, it is natural to follow an excursion e conditioned on $\|e\|_1^* > \varepsilon$ only after the first time $T_\varepsilon^\uparrow(e)$ at which $\|e(t)\|_1 > \varepsilon$: indeed, the strong Markov property implies that the conditioning only affects the process e shifted at time $T_\varepsilon^\uparrow(e)$ through the value $e(T_\varepsilon^\uparrow(e))$ that e takes at this time. Following the excursion after time $T_\varepsilon^\uparrow(e)$ thus makes it possible to simplify the conditioning and should therefore be an easier task than studying the whole conditioned excursion. This approach is closely related to a remark by Aldous [5, Section 6.3] in the context of random graphs, and turned out to be useful for the study of the stochastic network with mobile customers of Section 4.3. To formalize it, for $\varepsilon > 0$ recall the mapping $T_\varepsilon^\uparrow(f) = \inf\{t \geq 0 : \|f(t)\|_1 > \varepsilon\}$ and introduce the mapping $e_\varepsilon^\uparrow : \mathscr{D}(\mathbb{R}^d) \to \mathscr{E}(\mathbb{R}^d)$ defined by

$$e_\varepsilon^\uparrow(f) = \theta_{T_\varepsilon^\uparrow \circ e_\varepsilon(f)}\left(e_\varepsilon(f)\right), \; f \in \mathscr{D}(\mathbb{R}^d).$$

Note that $e_\varepsilon^\uparrow$ is well-defined for $\varphi = \|\cdot\|_1^*$.

Proposition 2.3.8. *Consider $\varphi = \|\cdot\|_1^*$ and assume that $\mathbf{P}_n$ is C-tight. If $e_\varepsilon^\uparrow$ and $T_0 \circ e_\varepsilon^\uparrow$ converge for every ε such that $\mathscr{N}(\varphi = \varepsilon) = 0$, then $(e_\varepsilon, T_0 \circ e_\varepsilon)$ converges for every ε such that $\mathscr{N}(\varphi = \varepsilon) = 0$.*

Combined with Theorem 2.3.1 it has the following consequence.

Theorem 2.3.9. *Consider $\varphi = \|\cdot\|_1^*$ and assume that $\mathbf{P}_n$ is C-tight. If $e_\varepsilon^\uparrow$, $T_0 \circ e_\varepsilon^\uparrow$ and g_ε converge for every $\varepsilon > 0$, then $\mathbf{P}_n$ converges.*

Changing the conditioning

Lemma 2.3.2 shows that controlling excursions measured according to their height is particularly interesting, since it automatically implies tightness. It is therefore natural to ask whether such a control can be obtained from the control of excursions measured according to φ. More generally, if $e_\varepsilon^\phi(f)$ denotes the first excursion e of f that satisfies $\phi(e) > \varepsilon$, it is natural to ask under which conditions a control on $e_\varepsilon^{\varphi_1}$ gives a control on $e_\varepsilon^{\varphi_2}$, given two different maps $\varphi_1, \varphi_2 : \mathscr{E}(\mathbb{R}^d) \to [0,\infty]$.

Lemma 2.3.10. *For $i = 1,2$, let $\varphi_i : \mathscr{E}(\mathbb{R}^d) \to [0,\infty]$ be a measurable map such that $\varphi_i(e) = 0$ if and only if $e = \mathbf{z}$ and such that $\mathscr{N}(\varphi_i > \varepsilon)$ is finite for every $\varepsilon > 0$, and let $C_i = \{\varepsilon > 0 : \mathscr{N}(\varphi_i = \varepsilon) = 0\}$. If the following two conditions hold:*

- *$c_n \mathscr{N}_n(\varphi_i > \varepsilon_i) \to \mathscr{N}(\varphi_i > \varepsilon_i)$ for every $\varepsilon \in C_i$, $i = 1,2$;*
- *$(e_{\varepsilon_1}^{\varphi_1}, \varphi_2 \circ e_{\varepsilon_1}^{\varphi_1})$ converges for every $\varepsilon_1 \in C_1$;*

then $e_{\varepsilon_2}^{\varphi_2}$ converges for every $\varepsilon_2 \in C_2$.

Extended Itô's construction

In Section 1.3 we have recalled Itô's construction of a regenerative process from its excursion measure $\mathscr{N}$, a σ-finite measure on $\mathscr{E}(\mathbb{R}^d)$. Actually, this construction can be realized starting from any σ-finite measure on $\mathscr{D}(\mathbb{R}^d)$, as long as we are given a functional $\zeta : \mathscr{D}(\mathbb{R}^d) \to [0,\infty]$ such that $\zeta(f) = 0$ if and only if $f = \mathbf{0}$, and such that $\mathscr{N}(1 \wedge \zeta) < \infty$. The construction can then be repeated verbatim, replacing each time T_0 by ζ: this construction will be referred to as the extended Itô's construction.

The drawback of this construction is that it may not necessarily be invertible, namely, given a process X resulting from the extended Itô's construction, it may not necessarily be possible to isolate its "motifs" (that were glued one after the other in the construction). However, when this is possible, then all the previous results go through since the proofs rely on continuity properties of deterministic operators and go through without any change.

The interest of this extension is that the classical definition (1.4) of regeneration can be in some cases too restrictive. For instance, this definition excludes processes that stay at 0 for a duration that is not exponential. It also excludes processes (even Markovian ones) that have non-zero holding times but leave 0 continuously. Consider for instance the Markov process that stays at 0 for an exponential time, then increases at rate 1 for an exponential time and jumps back to 0: then (1.4) fails for $\tau = \inf\{t \geq 0 : X(t) > 0\}$.

The queue length process of the $G/G/1$ queue is another natural example where the condition (1.4) is not satisfied, but which can be constructed (and the construction inverted) according to the extended Itô's construction. Indeed, when interarrival times are not exponentially distributed, then the queue length process does not regenerate when it hits 0, because the amount of time it stays at 0 may depend on the excursion that just finished. On the other hand, it does regenerate when a customer initiates a busy cycle, i.e., when the queue length process jumps from 0 to 1.

Chapter 3

Branching processes

Contents

3.1 Introduction . . . 24
3.1.1 Basic definitions . . . 24
3.1.2 Scaling limits of Galton–Watson processes and beyond . . . 25
3.1.3 Crump–Mode–Jagers trees . . . 27
3.2 Galton–Watson processes in varying environments . . . 30
3.2.1 Context and general idea . . . 30
3.2.2 Main results . . . 34
3.2.3 Discussion . . . 36
3.3 Binary, homogeneous Crump–Mode–Jagers processes . . . 37
3.3.1 Binary, homogeneous Crump–Mode–Jagers processes and Lévy processes . . . 37
3.3.2 Discussion . . . 41
3.4 Crump–Mode–Jagers trees with short edges . . . 45
3.4.1 The case of short edges . . . 45
3.4.2 Fundamental formula for $\mathbb{H}(n)$ . . . 48

3.1 Introduction

3.1.1 Basic definitions

Branching processes encompass a wide class of stochastic processes modeling the evolution of a population over time. The simplest branching process is the **Galton–Watson process**, which informally records the number of individuals in each generation of a population where individuals behave in an i.i.d. manner. It is a discrete-time Markov chain with values in $\mathbb{Z}_+$ where every individual alive at time (generation) k begets a random number of children, independently from one another and according to the same *offspring distribution*. Formally, a stochastic process $(Z(k), k \in \mathbb{Z}_+)$ is a Galton–Watson process with offspring distribution q, a probability distribution on $\mathbb{Z}_+$, if it satisfies the recursion

$$Z(k+1) = \xi_k(1) + \cdots + \xi_k(Z(k)),\ k \in \mathbb{Z}_+, \tag{3.1}$$

where the random variables $(\xi_k(i), k, i \in \mathbb{Z}_+)$ are i.i.d. with common distribution q. One of the most important feature of Galton–Watson processes is their classification according to the mean number of children $m = \sum_{k \in \mathbb{Z}_+} kq\{k\}$: the *extinction time* $T_0(Z)$ is almost surely finite if and only $m \leq 1$. There is therefore a phase transition at the critical value $m = 1$, and a Galton–Watson process is said to be subcritical, critical or supercritical according to $m < 1$, $m = 1$ or $m > 1$, respectively.

This model can be varied indefinitely by incorporating more realistic features: for instance, time and/or space can be continuous, the offspring distribution can vary over time, individuals can undergo a spatial motion, there may be several types of individuals, resources constraining the growth of the population, etc. Branching processes capture a simple yet universal dynamic, relevant in a wide range of situations. Besides their obvious application in population dynamics, they also play a key role in many different fields, ranging from theoretical probability to nuclear physics through computer science and partial differential equations. We will for instance present applications of branching processes in queueing theory and mathematical finance in Sections 4.2 and 5.1, respectively.

Among all the possible extensions of Galton–Watson processes, **Crump–Mode–Jagers processes** are of special interest in this manuscript, namely in Sections 3.3, 3.4 and 4.2. Recall that $\mathscr{M}_p((0,\infty))$ is the set of finite point measures on $(0,\infty)$ and let in the sequel

$$\mathbb{S} = \left\{(v, \nu) \in (0,\infty) \times \mathscr{M}_p((0,\infty)) : v \geq \pi(\nu)\right\}.$$

An element of $\mathbb{S}$ will be called a *stick* or a *life descriptor*, depending on the context: the terminology stick will be used when considering chronological trees, which can be seen as sticks grafted to one another; and life descriptor when considering individuals.

Let $(V, \mathscr{P})$ be a random variable with values in $\mathbb{S}$. In a Crump–Mode–Jagers process with parameter $(V, \mathscr{P})$, each individual u is endowed with an independent copy $(V_u, \mathscr{P}_u)$ of $(V, \mathscr{P})$ such that each atom of $\mathscr{P}_u$ corresponds to the birth time (relatively to u's life) of a child of u. This new individual is endowed with a copy of $(V, \mathscr{P})$ independent from everything else and the process continues in this way until extinction (if any). A Crump–Mode–Jagers process is said to be **binary and homogeneous** if $\mathscr{P}$ can be expressed as a Poisson process independent from V and stopped at V, i.e., $\mathscr{P}[0,t] = P[0, t \wedge V]$ with P a Poisson process independent from V.

Crump–Mode–Jagers processes generalize Galton–Watson processes by considering a continuous-time dynamic and allowing individuals to beget their children at any point during their life length. Galton–Watson processes are thus special cases of Crump–Mode–Jagers processes obtained by taking $(V, \mathscr{P}) = (1, \xi\epsilon_1)$, in which case the law of ξ is the offspring distribution. Galton–Watson and Crump–Mode–Jagers processes are naturally coupled, since the process that keeps track of the number of individuals in each generation of a Crump–Mode–Jagers process is a Galton–Watson process. Crump–Mode–Jagers processes are the most general stochastic processes satisfying this property, and for this reason are also called *general branching processes.*

Crump–Mode–Jagers processes appear naturally in a wide variety of problems, such as percolation, analysis of bins and balls models or analysis of algorithms and particular cases of Crump–Mode–Jagers processes such as Yule processes, Bellman–Harris processes or age-dependent branching processes have been extensively studied. The study of non-Markovian branching processes is challenging and no exhaustive theory is currently available. Yet, owing to their deep connections with Lévy processes, binary and homogeneous Crump–Mode–Jagers processes form a large class of non-Markovian Crump–Mode–Jagers processes that can be treated in great generality.

3.1.2 Scaling limits of Galton–Watson processes and beyond

Returning to the Galton–Watson case, let $Z^{(n)}$ be a Galton–Watson process with offspring distribution q and started with n individuals. If q has finite mean m, then $(n^{-1}Z^{(n)}(k), k \in \mathbb{Z}_+)$ converges in distribution to the deterministic sequence $(m^k, k \in \mathbb{Z}_+)$: as in the random walk case discussed in the introduction, this functional law of large numbers is not very informative in the critical case $m = 1$ which suggests to scale time as well.

Analytical approach

Thus we are lead to consider the problem, in the case $m = 1$, to find a sequence $\Gamma_n \to \infty$ such that $X_n = (X_n(t), t \in \mathbb{R}_+)$ converges in distribution, where $X_n(t) = n^{-1}Z^{(n)}([\Gamma_n t])$. This problem was first considered by Feller [36] who showed, in the finite variance case $\sigma := \sum_{k \in \mathbb{Z}_+} k^2 q\{k\} < \infty$, that for the choice $\Gamma_n = n$, the finite-dimensional distributions of (X_n) converge to those of the so-called Feller diffusion, i.e., the solution of the stochastic differential equation $\mathrm{d}X(t) = (2\sigma X(t))^{1/2}\mathrm{d}W(t)$. Lamperti [81] relaxed the finite variance assumption, while under this assumption Lindvall [86] extended Feller's result to weak convergence.

The general case (assuming that the limit process is conservative) was solved by Grimvall [44] who provided a necessary and sufficient condition for the convergence in distribution of X_n in terms of the offspring distribution q. The main condition involves the convergence in distribution of the sum of $[n\Gamma_n]$ independent random variables equal in distribution to $(\xi - 1)/n$, where ξ is a random variable following the offspring distribution. For completeness, we state the following result in the triangular case where the offspring distribution q_n is allowed to depend on n. In particular, in the following statements $(\bar{\xi}_n(k), k \in \mathbb{Z}_+)$ are i.i.d. random variables equal in distribution to $(\xi_n - 1)/n$ where ξ_n has distribution q_n.

Theorem 3.1.1 (Theorem 3.1 in [44])**.** *The sequence (X_n) converges in the sense of finite-dimensional distributions toward a possibly non-conservative process if and* *only if the sequence* $(\sum_{k=1}^{[n\Gamma_n]}\bar{\xi}_n(k))$ *converges in distribution toward a finite random variable.*

Ethier and Kurtz [35, Theorem 9.1.3] actually showed that the convergence in distribution of $(\sum_{k=1}^{[n\Gamma_n]}\bar{\xi}_n(k))$ toward a finite random variable implies the convergence in distribution of (X_n) in $\mathscr{D}([0,\infty])$. The converse is presumably true, but is only known, as far as I am aware, in $\mathscr{D}([0,\infty))$, i.e., when the limit is conservative, cf. Theorem 3.4 in Grimvall [44].

This result thus reduces the problem of the convergence of (X_n) to the problem of the convergence of the sum of a triangular array. The latter problem is exhaustively studied in the book by Gnedenko and Kolmogorov [41], see also Jacod and Shiryaev [56] for a more recent treatment: the convergence of $(\sum_{k=1}^{[n\Gamma_n]}\bar{\xi}_n(k))$ is equivalent to the following convergence of some fundamental *characteristic triplet.*

Theorem 3.1.2. *The sequence* $(\sum_{k=1}^{[n\Gamma_n]}\bar{\xi}_n(k))$ *converges in distribution if and only if there exist* $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}_+$ *and* ν *a* σ*-finite measure on* $(0,\infty)$ *such that*

$$n\Gamma_n\mathbb{E}\left(\frac{\bar{\xi}_n}{1+\bar{\xi}_n^2}\right) \to \alpha,\ n\Gamma_n\mathbb{E}\left(\frac{\bar{\xi}_n^2}{1+\bar{\xi}_n^2}\right) \to \beta \text{ and } n\Gamma_n\mathbb{P}(\bar{\xi}_n \geq x) \to \nu[x,\infty),$$

where the last convergence holds for every $x > 0$ *such that* $\nu\{x\} = 0$.

Grimvall's proof of Theorem 3.1.1 is purely analytic, and relies on studying the Laplace transform of $X_n(t)$: this is the approach which we will follow in the next section in order to extend these results to the case of varying environments.

Probabilistic approach

There also exists a probabilistic proof of this result, which relies on the so-called **Lamperti transformation**. This transformation gives a profound probabilistic interpretation of Grimvall's result, and plays a key role in the forthcoming study of the Processor-Sharing queue in Section 4.2. The Lamperti transformation naturally arises by rewriting (3.1) in the following way: if $(\xi(k), k \in \mathbb{N})$ are i.i.d. random variables with common distribution q, $S(0) = Z(0)$ and $S(k) = Z(0) + \xi(1) + \cdots + \xi(k) - k$ for $k \in \mathbb{N}$, then an equivalent statistical description of (3.1) is

$$Z(k+1) = S\left(Z(0) + \cdots + Z(k)\right),\ k \in \mathbb{Z}_+.$$

Scaling in time and space and defining

$$X_n(t) = \frac{1}{n}Z([\Gamma_n t]) \text{ and } S_n(t) = \frac{1}{n}S([n\Gamma_n t]),\ n \in \mathbb{N}, t \in \mathbb{R}_+,$$

this leads to the relation

$$X_n(t) = S_n\left(\int_0^t X_n(s)\mathrm{d}s\right). \tag{3.2}$$

This naturally leads to the introduction of the Lamperti transformation, which acts on the Skorohod space $\bar{\mathscr{D}}$ of càdlàg functions $f : [0,\infty] \to [0,\infty]$ that tend either to 0 or ∞ at infinity and are absorbed at those values.

Definition 3.1.3 (Lamperti transformation)**.** *For any function* $f \in \bar{\mathscr{D}}$, *its Lamperti transformation* $\mathscr{T}(f) \in \bar{\mathscr{D}}$ *is the function defined by* $\mathscr{T}(f) = f \circ \varpi_f$ *with*

$$\varpi_f(t) = \inf\left\{u \in \mathbb{R}_+ : \int_0^u f(s)\mathrm{d}s > t\right\}.$$

Relation (3.2) then rewrites $S_n = \mathscr{T}(X_n)$, i.e., S_n is the Lamperti transformation of X_n. Since we want to study X_n, we need to invert $\mathscr{T}$: actually, $\mathscr{T}$ is a bijection whose inverse $\mathscr{T}^{-1}$ is given by $\mathscr{T}^{-1}(f) = f \circ \varpi'_f$ with

$$\varpi'_f(t) = \inf\left\{ u \in \mathbb{R}_+ : \int_0^u \frac{\mathrm{d}s}{f(s)} > t \right\}.$$

We thus have the relation $X_n = \mathscr{T}^{-1}(S_n)$ which suggests, provided $\mathscr{T}$ and its inverse are continuous, that the convergence of (S_n) and (X_n) are equivalent, and that if $S_n \overset{\mathrm{d}}{\to} S_\infty$ we should have $X_n \overset{\mathrm{d}}{\to} \mathscr{T}^{-1}(S_\infty)$. The point is that the convergence of (S_n), a sequence of random walks, is known: (S_n) converges in distribution if and only if $(S_n(1))$ does, which takes us back to the convergence of the triangular array $(\sum_{k=1}^{n\Gamma_n} \bar{\xi}_n(k))$.

Beyond the Galton–Watson case

Beyond the Galton–Watson case, results are much scarcer and we are still far from having a complete understanding of the scaling limits of the most general branching processes, namely Crump–Mode–Jagers processes. Jagers [58] established the convergence of finite-dimensional distributions of age-dependent branching processes in the finite variance case (more precisely, $(V_n, \xi_n) = (V, Y_n \epsilon_V)$ with Y_n independent from V, V has finite mean and X_n has uniformly bounded third moment) toward the finite-dimensional distributions of Feller diffusion; Helland [52] established the weak convergence of continuous-time Markov branching processes toward CSBP using the Lamperti transformation; and Sagitov [107, 109] studied some cases where a sequence of Crump–Mode–Jagers processes converge in the sense of finite-dimensional distributions toward a CSBP.

3.1.3 Crump–Mode–Jagers trees

Discrete and Galton–Watson trees

Galton–Watson processes keep track of the number of individuals in the generations of a population where individuals' lives are governed by i.i.d. random variables: they can therefore be seen as a particular functional of a random (discrete) tree. In the following we adopt the convention $\mathbb{N}^0 = \{\emptyset\}$ and for a finite sequence $u \in \mathbb{N}^n$ and $j \in \mathbb{N}$, we denote by $uj \in \mathbb{N}^{n+1}$ the sequence obtained by concatenating j at the end of u, with the convention $uj = (j)$ when $u = \emptyset$. Let in the sequel

$$\mathscr{U} = \bigcup_{n \in \mathbb{Z}_+ 0} \mathbb{N}^n$$

the set of finite sequences of integers.

Definition 3.1.4. *A discrete tree $\mathscr{T}$ is a subset of $\mathscr{U}$ such that:*

i) $\emptyset \in \mathscr{T}$;

ii) *if $u \in \mathscr{U}$ and $j \in \mathbb{N}$ are such that $uj \in \mathscr{T}$, then $u \in \mathscr{T}$;*

iii) *for every $u \in \mathscr{T}$, there exists $N_u \in \mathbb{N}$ such that $uj \in \mathscr{T}$ for every $j \in \{1, \dots, N_u\}$.*

A discrete can be embedded in the plane, in which case a typical representation is given in Figure 3.1. For a rigorous definition of this embedding and more generally of the following objects, the reader is referred to Lambert [79].

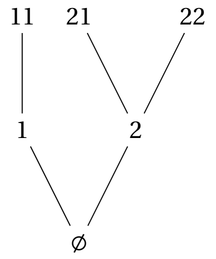


Figure 3.1: Typical representation of a discrete tree embedded in the plane.

Definition 3.1.5. *A Galton–Watson tree $\mathscr{T}$ with offspring distribution q is a random discrete tree such that $\emptyset$ has a random number ξ of children distributed according to q, and conditionally on ξ, the ξ subtrees of the root are i.i.d. and equal in distribution to $\mathscr{T}$.*

With this definition, it is clear that a Galton–Watson process is the process that keeps track of the number of nodes in every level (or depth, or generation) of a Galton–Watson tree with offspring distribution q. Likewise, the chronology of a Crump–Mode–Jagers process is coded by a random chronological tree.

Chronological and Crump–Mode–Jagers trees

Let in the sequel

$$\mathbb{U} = \bigcup_{n \in \mathbb{Z}_+} \mathbb{N}^n \times \mathbb{R}_+.$$

Definition 3.1.6. *A chronological tree $\mathbb{T}$ is a subset of $\mathbb{U}$ such that:*

i) $(\emptyset, 0) \in \mathbb{T}$;

ii) *the projection $\mathscr{T} := \{u \in \mathscr{U} : \exists \mu \in \mathbb{R}_+ \text{ s.t. } (u,\mu) \in \mathbb{T}\}$ of $\mathbb{T}$ on $\mathscr{U}$ is a discrete tree;*

iii) *for any $u \in \mathscr{T}$, there exist $0 \leq \alpha(u) < \omega(u) < \infty$ such that $(u,\sigma) \in \mathbb{T}$ if and only if $\alpha(u) \leq \sigma < \omega(u)$;*

iv) *for any $u \in \mathscr{T}$ and $j \in \mathbb{N}$ such that $uj \in \mathscr{T}$, $\alpha(uj) \in (\alpha(u), \omega(u)]$.*

A chronological tree can be embedded in the plane, in which case a typical representation is given in Figure 3.2. The projection $\mathscr{T}$ of $\mathbb{T}$ on the first coordinate will be called the *genealogical tree* associated to $\mathbb{T}$: for instance, the discrete tree of Figure 3.1 is the genealogical tree of the chronological tree of Figure 3.2. In the above definition, $\alpha(u)$ and $\omega(u)$ have to be thought of the time of birth and death of u. Given $u \in \mathscr{T}$, we define $\zeta_u = \omega(u) - \alpha(u) \in (0,\infty)$ the life length of u, and $\nu_u \in \mathscr{M}_p((0,\infty))$ the point process recording the birth time of u's children:

$$\nu_u = \sum_{j=1}^{N_u} \epsilon_{\alpha(uj)-\alpha(u)}.$$

Note that by construction we have $(\zeta_u, \nu_u) \in \mathbb{S}$.

Definition 3.1.7. *Let $(V, \mathscr{P})$ be a random variable in $\mathbb{S}$. A random chronological tree $\mathbb{T}$ satisfying the following two properties is called a Crump–Mode–Jagers tree with parameter $(V, \mathscr{P})$:*

i) *the genealogical tree $\mathscr{T}$ associated to $\mathbb{T}$ is a Galton–Watson tree;*

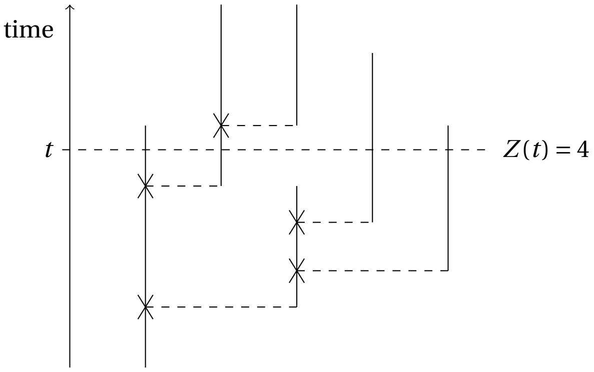


Figure 3.2: Typical representation of a chronological tree embedded in the plane. The horizontal axis bears no meaning, except for the case of atoms with multiplicity ≥ 2, in which case the leftmost subtree is explored first by the contour process, i.e., it is the first individual in lexicographic order.

ii) *conditionally on $\mathscr{T}$, (ζ_u, ν_u) is distributed like $(V, \mathscr{P})$ conditioned on $|\mathscr{P}| = N_u$ and these random variables are independent over $u \in \mathscr{T}$.*

A Crump–Mode–Jagers process is then the width process of the corresponding Crump–Mode–Jagers tree, i.e., the process which at time t counts the number of edges crossed by an horizontal line at level t as in Figure 3.2.

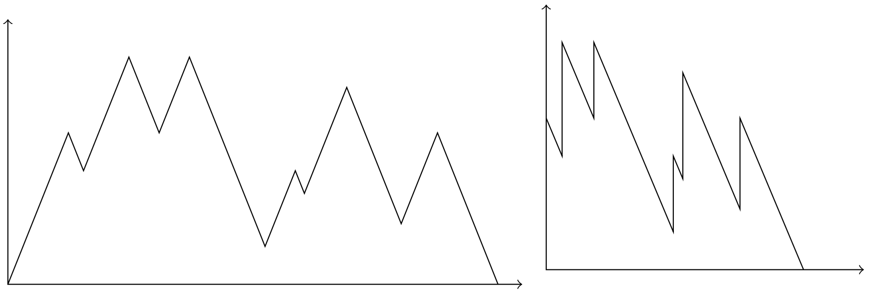
A chronological tree is encoded by the so-called **contour process** $\mathbb{C} = (\mathbb{C}(t), t \in \mathbb{R}_+)$, informally defined as follows. Imagine a walker moves at unit speed along the edges of the tree, going up on the left side and down on the right side, and begins exploring a new edge when it meets a cross. Then $\mathbb{C}(t)$ is equal to the height of the walker at time t, i.e., its distance to the root. The **jumping contour process** is defined similarly, except the walker goes at infinite speed along the left side of an edge, so the corresponding process undergoes a jump.

Both the contour and the jumping contour processes uniquely encode a given chronological tree. Although the jumping contour process is a less classical than the contour process, it turns out to be very useful in the study of binary and homogeneous Crump–Mode–Jagers processes as in this case, it is a Lévy process.

The contour process induces an order on nodes, called the *lexicographic order*, whereby $\ell(n) \in \mathscr{T}$ for $n \in \mathbb{Z}_+$ is the nth node to be visited (for the first time) by the contour process. The lexicographic order corresponds to the classical depth-first search of the discrete tree: on the tree of Figure 3.1, the lexicographic order of the 6 nodes is thus $\emptyset$, 1, 11, 2, 21, 22. If $\mathbb{H}(n)$ is the birth time of $\ell(n)$, the process $\mathbb{H} = (\mathbb{H}(n), n \in \mathbb{Z}_+)$ is called **height process**. A last object of fundamental importance is the **Lukasiewicz path** $S = (S(n), n \in \mathbb{Z}_+)$, defined by $S(0) = 0$ and

$$S(n+1) - S(n) = |\nu_{\ell(n)}| - 1.$$

The Lukasiewicz path depends only on the genealogical tree, the Lukasiewicz path of the tree of Figure 3.1 is for instance given by the sequence $(0, 1, 1, 0, 1, 0, -1)$: we visit nodes in the lexicographic order and each time make a jump given by the number of children of the current node minus one.



(a) Contour process.

(b) Jumping contour process.

Figure 3.3: Contour process (a) and jumping contour process (b) of the chronological tree of Figure 3.2.

Since a chronological tree with $\zeta_u = 1$ and $|\nu_u| = \nu_u\{1\}$ can naturally be seen as a discrete tree, the genealogical tree $\mathscr{T}$ associated to $\mathbb{T}$ also has associated height and contour processes. These processes will be denoted by $\mathscr{H}$ and $\mathscr{C}$, respectively, and in order to avoid any ambiguity, $\mathscr{H}$ and $\mathscr{C}$ will be called **genealogical height and contour processes**, and $\mathbb{H}$ and $\mathbb{C}$ will be called **chronological height and contour processes**.

Convergence of random trees

Although we will barely touch upon this topic, it is worthwhile to mention that scaling limits of random trees have recently been intensively studied. This line of work was initiated by the pioneering work of Aldous, Le Gall, Neveu and Pitman [2, 3, 4, 82, 90, 91] and the framework for the convergence of random trees is now well developed. In particular, a continuous tree is seen as a compact metric space. The space of continuous trees can then be equipped with the Gromov–Hausdorff topology, which defines a notion of convergence of a sequence of real trees (t_n). Although at first sight unworkable, one of the nice aspect of the theory is that in order for the sequence (of compact metric spaces) (t_n) to converge, it is actually enough for their associated (real-valued) contour processes to converge, see for instance Le Gall and Miermont [83, Corollary 3.7]. This fact provides an additional motivation for studying contour and height processes, as will be done in Section 3.3.

3.2 Galton–Watson processes in varying environments

In this section we discuss the convergence of Galton–Watson processes in varying environments. This is the only place in the manuscript where we start from a discrete-time stochastic process, and we denote by $Z = (Z(k), k \in \mathbb{Z}_+)$ the canonical process in this case. Results of this section can be found in [BS15].

3.2.1 Context and general idea

Context

For each $n \in \mathbb{N}$, consider a sequence of offspring distributions $(q_{i,n}, i \in \mathbb{Z}_+)$, the *environments*, and the corresponding law $\mathbb{P}_n$ under which $Z = (Z(k), k \in \mathbb{Z}_+)$ is a Galton–

Watson process started with n individuals and where individuals of the ith generation reproduce according to $q_{i,n}$. More precisely, if $\xi_{i,n}$ has distribution $q_{i,n}$, the random variables $(\xi_{i,n}(k), i, k \in \mathbb{Z}_+)$ are independent and $\xi_{i,n}(k)$ is equal in distribution to $\xi_{i,n}$, then the sequence $Z^{(n)}$ given by

$$Z^{(n)}(0) = n \text{ and } Z^{(n)}(i+1) = \sum_{k=1}^{Z^{(n)}(i)} \xi_{i,n}(k), \ i \in \mathbb{Z}_+,$$

is equal in distribution to Z under $\mathbb{P}_n$. We will study the asymptotic behavior of Z under $\mathbb{P}_n$ when time and space are scaled. The relation $Z(0) = n$ fixes the space scale to grow linearly with n and for the time scale, we consider an increasing, càdlàg and onto function $\gamma_n : [0,\infty) \to \mathbb{Z}_+$ and define $\mathbf{P}_n$ the law of the process $X_n = (X_n(t), t \geq 0)$ with

$$X_n(t) = \frac{1}{n} Z(\gamma_n(t)), \ t \in \mathbb{R}_+.$$

We are looking for suitable conditions on the $(q_{i,n})$ and on γ_n such that $\mathbf{P}_n$ converges weakly. In the finite variance case, general results for this class of processes were obtained by Kurtz [75] whose assumptions involve, for $n \in \mathbb{N}$, the functions

$$A_n(t) = \sum_{i=1}^{[nt]} \log a_{i,n}, \ B_n(t) = \frac{1}{n} \sum_{i=1}^{[nt]} b_{i,n} \text{ and } C_n(t) = \frac{1}{n^{3/2}} \sum_{i=1}^{[nt]} c_{i,n}, \ t \geq 0,$$

where for $i \in \mathbb{N}$,

$$a_{i,n} = \mathbb{E}(\xi_{i,n}), \ b_{i,n} = \mathbb{E}\left[\left(\frac{\xi_{i,n}}{a_{i,n}} - 1\right)^2\right] \text{ and } c_{i,n} = \mathbb{E}\left[\left|\frac{\xi_{i,n}}{a_{i,n}} - 1\right|^3\right].$$

Theorem 3.2.1 (Theorem 2.10 in [75]). *Let $\gamma_n(t) = [nt]$ and assume that $A_n \xrightarrow{J_1} A$ and that for each $t \geq 0$, $B_n(t) \to B(t)$ for some absolutely continuous and strictly increasing process B and $C_n(t) \to 0$. Then $\mathbf{P}_n$ converges weakly to the law of $e^A W \circ \tau$ where W is Feller diffusion and $\tau(t) = \int_0^t e^{-A(s)} B'(s) \mathrm{d}s$ for $t < \int_0^\infty e^{-A(s)} B'(s) \mathrm{d}s$.*

The assumption $C_n(t) \to 0$ somehow corresponds to assuming uniformly finite moments of order 3. In the case of constant environments where the law of $\xi_{i,n}$ does not depend on i, these assumptions correspond to assuming that the $\xi_{1,n}$ have uniformly bounded second moment and a third moment which grows slower than $n^{1/2}$. In this case, the assumptions of Theorem 3.1.2 hold with $\nu = 0$ and the limiting process is Feller diffusion.

By conditioning on the environment, this result also proved a conjecture of Keiding [62] on the scaling limits of finite variance Galton–Watson processes in random environments, see also Helland [53]. Kurtz' result was later strengthened by Borovkov [17] who assumed finite moments of order $2 + \varepsilon$ instead of 3 and only assumed B to be increasing (but assumed A to be continuous): the function $\tau(t)$ is then defined by $\tau(t) = \int_0^t e^{-A(s)} \mathrm{d}B(s)$. In both cases, the approach is to scale the pre-limit process $\mathbb{P}_n$ in both time and space in order to define a process converging to Feller diffusion, which explains the form of the limiting process obtained in 3.2.1 as a space-time transformation of Feller diffusion. However, Kurtz and Borovkov approaches differ in that Kurtz uses semigroup techniques developed in [74], while Borovkov uses earlier results [16] where a Galton–Watson process in varying environments is directly coupled with Feller diffusion.

Presentation of the approach

To go beyond the $2+\varepsilon$ finite moment assumption of Borovkov and deal with offspring distributions with possibly infinite variance, we adopted the original analytical approach of Feller, Grimvall and Lindvall [36, 44, 86]. Our main object of study is thus the Laplace exponent $u_n(s,t,\cdot)$ of $X(t)$ under $\mathbf{P}_n(\cdot \mid X(s)=1)$, defined for $0 \leq s \leq t$ and $\lambda \geq 0$ by

$$u_n(s,t,\lambda) = -\log \mathbf{E}_n\left(e^{-\lambda X(t)} \mid X(s)=1\right).$$

Our main result – Theorem 3.2.2 below – shows that u_n converges to the solution of a backward integro-differential equation, and that this characterizes the convergence in distribution of $\mathbf{P}_n$: the main ideas are presented now.

First of all, we note that the Markov property of $\mathbf{P}_n$ implies the following composition rule:

$$u_n(t_1,t_3,\lambda) = u_n\big(t_1,t_2,u_n(t_2,t_3,\lambda)\big), \quad 0 \leq t_1 \leq t_2 \leq t_3,\ \lambda \geq 0. \tag{3.3}$$

Further, our assumptions will be stated in terms of a characteristic triplet (α_n,β_n,ν_n) where α_n and β_n are càdlàg functions which essentially correspond to A_n and B_n of Theorem 3.2.1, and ν_n is a measure on $\mathbb{R}\times\mathbb{R}_+$. For $i \in \mathbb{Z}_+$ and $n \in \mathbb{N}$ let $\bar{\xi}_{i,n} = (\xi_{i,n}-1)/n$ and $\nu_{i,n}$ be the measure on $\mathbb{R}$ defined by $\nu_{i,n}[x,\infty) = n\mathbb{P}\big(\bar{\xi}_{i,n} \geq x\big)$ for $x \in \mathbb{R}$. We then define the triplet (α_n,β_n,ν_n) by the following relations:

$$\nu_n([x,\infty)\times(0,t]) = \sum_{i=0}^{\gamma_n(t)-1} \nu_{i,n}[x,\infty), \ x\in\mathbb{R}, t \in \mathbb{R}_+,$$

and

$$\alpha_n(t) = \int_{\mathbb{R}\times(0,t]} \frac{x}{1+x^2}\nu_n(\mathrm{d}x\mathrm{d}y) \ \text{ and } \ \beta_n(t) = \int_{\mathbb{R}\times(0,t]} \frac{x^2}{2(1+x^2)}\nu_n(\mathrm{d}x\mathrm{d}y), \ t \in \mathbb{R}_+.$$

We now present back-of-the-envelope computations to give some idea of the proof. Let in the sequel $t_i^n = \inf\{t \geq 0 : \gamma_n(t) = i\}$ and start by writing

$$u_n(s,t,\lambda) = u_n(t^n_{\gamma_n(s)},t,\lambda) = \lambda + \sum_{i=\gamma_n(s)}^{\gamma_n(t)-1} \big(u_n(t_i^n,t,\lambda) - u_n(t_{i+1}^n,t,\lambda)\big).$$

Using the composition rule (3.3) to write $u_n(t_i^n,t,\lambda) = u_n(t_i^n,t_{i+1}^n,u_n(t_{i+1}^n,t,\lambda))$, we continue with

$$u_n(s,t,\lambda) = \lambda + \sum_{i=\gamma_n(s)}^{\gamma_n(t)-1} \psi_{i,n}\big(u_n(t_i^n,t,\lambda)\big)$$

where we have defined $\psi_{i,n}(\lambda) = u_n(t_i^n,t_{i+1}^n,\lambda)-\lambda$ for $\lambda \in \mathbb{R}_+$. From the definition of u_n as Laplace exponent of $\mathbf{P}_n$, it can be showed that

$$\psi_{i,n}(\lambda) = -n\log\left(1-\frac{1}{n}\int\left(1-e^{-\lambda x}\right)\nu_{i,n}(\mathrm{d}x)\right), \ \lambda \in \mathbb{R}_+.$$

Provided the integral term $(1/n)\int\left(1-e^{-\lambda x}\right)\nu_{i,n}(\mathrm{d}x)$ vanishes, this last expression suggests the approximation $\psi_{i,n}(\lambda) \approx \int\left(1-e^{-\lambda x}\right)\nu_{i,n}(\mathrm{d}x)$ which leads to

$$\begin{aligned} u_n(s,t,\lambda) &\approx \lambda + \sum_{i=\gamma_n(s)}^{\gamma_n(t)-1}\int\left(1-e^{-xu_n(t_i^n,t,\lambda)}\right)\nu_{i,n}(\mathrm{d}x) \\ &= \lambda + \int_{\mathbb{R}\times(s,t]}\left(1-e^{-xu_n(y,t,\lambda)}\right)\nu_n(\mathrm{d}x\mathrm{d}y). \end{aligned}$$

In the context of the convergence of triangular arrays, one typically needs to integrate the tail distribution against functions which are negligible compared to x^2 as $x \to 0$, see for instance [56, Theorem VII.2.9]. For technical reasons which have a similar flavor, we introduce the function

$$g(x,\lambda) = 1 - e^{-\lambda x} - \frac{\lambda x}{1+x^2} + \frac{(\lambda x)^2}{2(1+x^2)}, \ x \in \mathbb{R}, \lambda \in \mathbb{R}_+,$$

which for fixed λ indeed satisfies $g(x,\lambda) = o(x^2)$ as $x \to 0$, and rewrite the previous equation as

$$u_n(s,t,\lambda) \approx \lambda + \int_{(s,t]} u_n(y,t,\lambda)\alpha_n(\mathrm{d}y) - \int_{(s,t]} u_n(y,t,\lambda)^2 \beta_n(\mathrm{d}y) \\ + \int_{\mathbb{R}\times(s,t]} g\big(x, u_n(y,t,\lambda)\big)\nu_n(\mathrm{d}x\,\mathrm{d}y).$$

The function g having the required behavior at 0, it is reasonable to pass to the limit in this last expression. Namely, if $(\alpha_n, \beta_n, \nu_n)$ converges to (α, β, ν) in a suitable sense, it is reasonable to expect that u_n will converge to u satisfying

$$u(s,t,\lambda) = \lambda + \int_{(s,t]} u(y,t,\lambda)\alpha(\mathrm{d}y) - \int_{(s,t]} u(y,t,\lambda)^2 \beta(\mathrm{d}y) \\ + \int_{\mathbb{R}\times(s,t]} g\big(x, u(y,t,\lambda)\big)\nu(\mathrm{d}x\,\mathrm{d}y). \quad (3.4)$$

Intrinsic limitations

The main contribution in [BS15] is to make the above line of arguments rigorous, and investigate general conditions on $(\alpha_n, \beta_n, \nu_n)$ under which $u_n \to u$ characterized by (3.4). In doing so, we have discovered some intrinsic limitation to this approach. First of all, in order for (3.4) to make sense, one needs α to be a suitable integrand: in [BS15] we have considered α to have finite variation, which is to put in regards to Theorem 3.2.1 where A (essentially the analog of α) is merely assumed to be càdlàg. It is conceivable that our approach could be pushed to more general integrands (e.g., functions with finite quadratic variation), but going all the way to càdlàg functions seems to require new ideas.

Another limitation of this approach arises due to the fact that, since subcritical offspring distributions are allowed, the process may go through *bottlenecks* where the process gets almost surely extinct. An extreme bottleneck would be to impose an offspring distribution with zero mean, but there are in principle many subtler ways to send the process to zero almost surely, for instance with an accumulation of slightly subcritical offspring distributions. In some cases, these bottlenecks make it impossible to study the convergence of $u_n(s,t,\lambda)$ as $n \to \infty$, see Section 2.3 in [BS15] for a simple example. This illustrates the need to study the convergence on a time-interval free of bottlenecks, and naturally leads us to study the process on $[\wp(t), t]$ with $\wp(t) \in [0,t]$ the last time a bottleneck occurs before time t. The most general definition for $\wp(t)$ is the following:

$$\wp(t) = \sup\left\{ s \le t : \lim_{\varepsilon\to 0} \liminf_{n\to\infty} \inf_{s\le y\le t} \mathbf{P}_n\big(X(t) > \varepsilon \mid X(y) = 1\big) = 0 \right\} \quad (3.5)$$

with the convention $\sup \emptyset = 0$. Intuitively, $\wp(t)$ has to be seen as the smallest time such that $u_n(s,t,\lambda)$ is bounded away from 0 uniformly in n and $s > \wp(t)$. Although this is not clear from this formula, this will become apparent in the alternative, simpler characterization (3.7) of $\wp(t)$ below, which holds under the Assumption A needed for our main results and which we introduce now.

3.2.2 Main results

We now state the main result from [BS15]. The statement of the theorem is lengthy, and so we first state the assumption which formalizes the convergence of the characteristic triplet $(\alpha_n, \beta_n, \nu_n)$. Essentially, the convergence $(\alpha_n, \beta_n, \nu_n) \to (\alpha, \beta, \nu)$ means pointwise convergence for the functions α_n and β_n and weak convergence for ν_n, with the additional requirement that the total variation of α_n also converges pointwisely and that, loosely speaking, jumps also converge (Condition (A2)).

In this section we denote by $\|f\|_{\mathrm{TV}}$ the total variation associated to f, i.e., $\|f\|_{\mathrm{TV}}(t)$ for $t \in \mathbb{R}_+$ is the total variation of f on the interval $[0, t]$.

Assumption A. *There exist a càdlàg function of locally finite variation α, an increasing càdlàg function β and a positive measure ν on $(0,\infty)^2$ such that the two following conditions hold:*

(A1) *For every $t \geq 0$ and every $x > 0$ such that $\nu(\{x\} \times (0, t]) = 0$,*

$$\alpha_n(t) \xrightarrow[n\to\infty]{} \alpha(t),\ \|\alpha_n\|_{\mathrm{TV}}(t) \xrightarrow[n\to\infty]{} \|\alpha\|_{\mathrm{TV}}(t),\ \beta_n(t) \xrightarrow[n\to\infty]{} \beta(t)$$
$$\text{and } \nu_n([x,\infty) \times (0, t]) \xrightarrow[n\to\infty]{} \nu([x,\infty) \times (0, t]);$$

(A2) *For every t such that $\Delta\alpha(t) \neq 0$, $\Delta\beta(t) \neq 0$ or $\nu((0,\infty) \times \{t\}) \neq 0$ and for every $x > 0$ such that $\nu(\{x\} \times \{t\}) = 0$,*

$$\alpha_{\gamma_n(t),n} \xrightarrow[n\to\infty]{} \Delta\alpha(t),\ \beta_{\gamma_n(t),n} \xrightarrow[n\to\infty]{} \Delta\beta(t) \text{ and } \nu_{\gamma_n(t),n}[x,\infty) \xrightarrow[n\to\infty]{} \nu([x,\infty) \times \{t\}).$$

Theorem 3.2.2 (Behavior on $[\wp(t), t]$)**.** *Assume that Assumption A holds, and let α, β and ν be the functions and measure defined there. Then, the following properties hold.*

I. *For every $t \geq 0$, we have $\Delta\alpha(t) \geq -1$ and $\int_{(0,\infty)\times(0,t]}(1 \wedge x^2)\nu(\mathrm{d}x\,\mathrm{d}y) < \infty$. Moreover, the following function $\tilde{\beta}$ is continuous and increasing:*

$$\tilde{\beta}(t) = \beta(t) - \int_{(0,\infty)\times(0,t]} \frac{x^2}{2(1+x^2)} \nu(\mathrm{d}x\,\mathrm{d}y),\ t \geq 0.$$

II. *For every $t, \lambda > 0$ and $s \in [\wp(t), t]$, there exists $u(s,t,\lambda) \in (0,\infty)$ such that for every $s_0 \geq 0$,*

$$\lim_{n\to\infty} \sup_{0\leq s\leq s_0} |u_n(s,t,\lambda) - u(s,t,\lambda)| = 0.$$

Moreover, the function $u_{t,\lambda} : s \in [\wp(t), t] \mapsto u(s,t,\lambda)$ is the unique càdlàg function that satisfies $\inf_{s\leq y\leq t} u_{t,\lambda}(y) > 0$ for every $\wp(t) < s \leq t$ and

$$u_{t,\lambda}(s) = \lambda + \int_{(s,t]} u_{t,\lambda}(y)\alpha(\mathrm{d}y) - \int_{(s,t]} u_{t,\lambda}(y)^2 \tilde{\beta}(\mathrm{d}y)$$
$$+ \int_{(0,\infty)\times(s,t]} \left(1 - e^{-xu_{t,\lambda}(y)} - \frac{xu_{t,\lambda}(y)}{1+x^2}\right) \nu(\mathrm{d}x\,\mathrm{d}y)$$

for every $\wp(t) \leq s \leq t$.

III. *Fix $t \geq 0$, $s \in [\wp(t), t]$ and $x \geq 0$. Then for every sequence of initial states (x_n) with $x_n \to x$, every $I \in \mathbb{Z}_+$, every $s \leq t_1 < \cdots < t_I \leq t$ and every $\lambda_1, \dots, \lambda_I > 0$,*

$$\lim_{n\to\infty} \mathbf{E}_n \left[\exp\left(-\lambda_1 X(t_1) - \cdots - \lambda_I X(t_I)\right) \mid X(s) = x_n \right] \\ = \exp\Big(-xu\Big(s, t_1, \lambda_1 + u\big(t_1, t_2, \lambda_2 + u(\cdots, u(t_{I-1}, t_I, \lambda_I)\cdots)\big)\Big)\Big). \quad (3.6)$$

IV. *Fix $t \geq 0$, $s \in [\wp(t), t]$ and $x \geq 0$. Then for every sequence of initial states (x_n) with $x_n \to x$, the sequence of processes $(X(y), s \leq y \leq t)$ under $\mathbf{P}_n(\cdot \mid X(s) = x_n)$ is tight on the space of càdlàg functions $f : [s,t] \to [0,\infty]$ endowed with the J_1 topology, where the space $[0,\infty]$ is equipped with the metric $d(x,y) = |e^{-x} - e^{-y}|$. In particular, weak convergence holds in view of* (3.6).

As promised earlier, under Assumption A the general definition (3.5) of the last bottleneck $\wp(t)$ before time t simplifies: under Assumption A, for any $\lambda > 0$ an alternative expression for $\wp(t)$ is

$$\wp(t) = \sup\left\{ s \leq t : \liminf_{n\to\infty} \inf_{s \leq y \leq t} u_n(y,t,\lambda) = 0 \right\}. \quad (3.7)$$

Moreover, we have identified a simple condition under which $\wp(t) = 0$. As discussed in Section 3.2.1, a bottleneck intuitively corresponds to the almost sure extinction of the process. Under Assumption A, the next proposition states that if offspring distributions have a mean uniformly bounded away from 0, then there is no bottleneck. In this case, Theorem 3.2.2 describes the asymptotic behavior of X_n on $[0,t]$.

Proposition 3.2.3 (No bottleneck)**.** *Let $t > 0$ and assume that Assumption A holds. If for every $C > 0$*

$$\liminf_{n\to\infty} \left(\inf_{0 \leq i \leq \gamma_n(t)} \mathbb{E}\left(\xi_{i,n}; \xi_{i,n} \leq Cn\right) \right) > 0,$$

then $\wp(t) = 0$.

We conclude this summary of results by stating a condition under which the convergence $u_n(s,t,\lambda) \to 0$ holds along a subsequence, for all $s < \wp(t)$ and $\lambda \geq 0$, which fits the intuition behind a bottleneck. This condition essentially states that any limit of X is conservative.

Proposition 3.2.4 (No explosion)**.** *Let $t > 0$ and assume that Assumption A holds. If the two sequences $(\|\alpha_n\|_{\mathrm{TV}}(t), n \geq 1)$ and $(\beta_n(t), n \geq 1)$ are bounded and*

$$\lim_{A\to\infty} \sup_{n \geq 1, 0 \leq s \leq y \leq t} \mathbf{P}_n(X(y) \geq A \mid X(s) = 1) = 0, \quad (3.8)$$

then there exists an increasing sequence of integers $n(k)$ such that $u_{n(k)}(s,t,\lambda) \to 0$ as $k \to \infty$ for all $s < \wp(t)$ and $\lambda \geq 0$.

Moreover, these assumptions are satisfied, i.e., $(\|\alpha_n\|_{\mathrm{TV}}(t), n \geq 1)$ and $(\beta_n(t), n \geq 1)$ are bounded and (3.8) *holds, if the following first moment condition is satisfied:*

$$\sup_{n \geq 1} \left(n \sum_{i=0}^{\gamma_n(t)-1} \mathbb{E}\left(|\bar{\xi}_{i,n}|\right) \right) < \infty.$$

3.2.3 Discussion

Let us conclude this section with a discussion of two interesting implications of Theorem 3.2.2: other implications, e.g., for Galton–Watson in random environment or Feller diffusion in random environment, can be found in [BS15].

On the correct time scale of Galton–Watson processes in random environment

Consider the simplest example of a Galton–Watson process in random environment where in each generation one chooses at random among one of two possible offspring distributions. More precisely, $(q_{i,n}, i \in \mathbb{Z}_+)$ is a sequence of i.i.d. probability distributions, so that conditionally on this sequence we have a Galton–Watson in varying environments, and their common distribution is $p_n^{(1)}\epsilon_{q^{(1)}} + p_n^{(2)}\epsilon_{q^{(2)}}$ where $p_n^{(j)} \in (0,1)$, $p_n^{(1)} + p_n^{(2)} = 1$ and $q^{(1)}$, $q^{(2)}$ are two deterministic probability distributions on $\mathbb{N}$. One of the questions which motivated us in the first place was to understand the correct time scale Γ_n on which to observe the process.

Let us call CSBP with characteristic (b, c, F) the CSBP whose branching mechanism is given by

$$\psi(\lambda) = \lambda b - \frac{1}{2}c\lambda^2 + \int (e^{-\lambda x} - 1 - \lambda x \mathbb{1}_{\{x \le 1\}})F(\mathrm{d}x),\ \lambda \in \mathbb{R}_+.$$

For each $j = 1,2$ let $Z_n^{(j)} = (Z_n^{(j)}(i), i \in \mathbb{Z}_+)$ be a Galton–Watson process with offspring distribution $q^{(j)}$ and consider $(\Gamma_n^{(j)})$ a sequence such that $(X_n^{(j)}, n \in \mathbb{N})$ converges weakly to the CSBP with characteristic $(b^{(j)}, c^{(j)}, F^{(j)})$, where $X_n^{(j)}(t) = n^{-1}Z_n^{(j)}(\lfloor \Gamma_n^{(j)} t \rfloor)$. If both $q^{(1)}$ and $q^{(2)}$ have finite variance, then the correct time scale is $\Gamma_n = n$ according to Feller's early result [36]. Thus when "mixing" these two processes, it is natural to speed up the resulting process by the common time scale and thus take $\Gamma_n = n$. To our knowledge, only such cases have been considered in the literature so far. When offspring distributions have infinite variance however, the situation becomes more delicate. Indeed, if for instance $q^{(1)}[x,\infty) \sim x^{-a}$ as $x \to \infty$ for some $a \in (1,2)$, then one needs to consider $\Gamma_n^{(1)} = n^{a-1}$. Thus there are now two "natural" time scales, namely $\Gamma_n^{(1)} = n^{a-1}$ and $\Gamma_n^{(2)} = n$.

Note that $\Gamma_n^{(j)}$ is the number of generations needed so that the variation of $Z_n^{(j)}$ may be of the order of n. Over Γ_n generations, the law of large numbers implies that $q^{(j)}$ has been used $p_n^{(j)}\Gamma_n$ times. Thus, if $p_n^{(j)}\Gamma_n \ll \Gamma_n^{(j)}$, the offspring distribution $q^{(j)}$ has not been picked sufficiently often in order to have any effect on the space scale n. This suggests that the correct time scale is $\Gamma_n = \min_j(\Gamma_n^{(j)}/p_n^{(j)})$ and indeed, the following result can be proved using Theorem 3.2.2:

- if $\Gamma_n^{(1)}/p_n^{(1)} \ll \Gamma_n^{(2)}/p_n^{(2)}$ and $\Gamma_n = \Gamma_n^{(1)}/p_n^{(1)}$, then X_n converges toward the CSBP with branching mechanism $(b^{(1)}, c^{(1)}, F^{(1)})$;
- if $\Gamma_n^{(1)} p_n^{(2)}/(\Gamma_n^{(2)} p_n^{(1)}) \to \ell \in (0,\infty)$ and $\Gamma_n = \Gamma_n^{(1)}/p_n^{(1)}$, then X_n converges toward the CSBP with characteristic $(b^{(1)} + \ell b^{(2)}, c^{(1)} + \ell c^{(2)}, F^{(1)} + \ell F^{(2)})$.

This discussion can be easily extended to the case of a finite number of offspring distributions that also vary with n, and it would be very interesting to understand the implications of Theorem 3.2.2 in more general settings, e.g., when we can choose among uncountably many offspring distributions.

A direct probabilistic proof?

As explained in the introduction, the convergence of a sequence of Galton–Watson processes is essentially equivalent to the convergence of a suitable sequence of random walks, and the Lamperti transformation provides a direct connection between these two objects. However, the Lamperti transformation breaks down in the case of varying environment, i.e., the image of a Galton–Watson process in varying environment by the Lamperti transformation does not have a simple probabilistic structure.

What is surprising is that Assumption A is almost equivalent to the convergence of some suitable process with independent increments, see for instance [56, Theorem VII.4.4]. This therefore suggests that some process with independent increments may be intrinsically tied to our problem, although we have been unable to identify such an object. If such it object exists, it would most probably lead to a simpler proof and a generalization of Theorem 3.2.2.

3.3 Scaling limits of binary and homogeneous Crump–Mode–Jagers processes

In this section we discuss the convergence of binary, homogeneous Crump–Mode–Jagers processes defined in Section 3.1.1. Results of this section can be found in [LS15] and, for the finite variance case, in [LSZ13].

3.3.1 Binary, homogeneous Crump–Mode–Jagers processes and Lévy processes

Link between binary, homogeneous Crump–Mode–Jagers processes and Lévy processes

For $n \in \mathbb{N}$ we consider $(V_n, \mathscr{P}_n)$ a random variable in $\mathbb{S}$ where V_n has finite mean and $\mathscr{P}_n$ is an independent Poisson process stopped at V_n and with intensity λ_n, and we call nth Crump–Mode–Jagers process the binary, homogeneous Crump–Mode–Jagers process with parameter $(V_n, \mathscr{P}_n)$. The following special case of Theorem 3.3 in [79] shows that the jumping contour process of a binary, homogeneous Crump–Mode–Jagers tree has an explicit probabilistic structure. To state this result, recall σ the operator that stops a function upon its first visit of 0, and define $\mathbb{P}_n^x$ for $x \in \mathbb{R}$ the law of the Lévy process started at x and with Laplace exponent ψ_n given by

$$\psi_n(u) = u - \lambda_n \mathbb{E}\left(1 - e^{-uV_n}\right), \; u \geq 0.$$

Thus X under $\mathbb{P}_n^0$ is of the form $P_n(t) - t$ with P_n a compound Poisson process with Lévy measure $\lambda_n \mathbb{P}(V_n \in \mathrm{d}x)$.

Proposition 3.3.1. *If X under $\mathbb{P}_n^0$ does not drift to $+\infty$, then $\mathbb{P}_n^x \circ \sigma^{-1}$ for any $x > 0$ is the law of the jumping contour process of the Crump–Mode–Jagers tree with parameter $(V_n, \mathscr{P}_n)$ started with one ancestor with deterministic life length x.*

Note that the assumption on $\mathbb{P}_n^0$ is equivalent to the (sub)critical assumption $\lambda_n \mathbb{E}(V_n) \leq 1$. Since a Crump–Mode–Jagers process is the width process of the corresponding Crump–Mode–Jagers tree, this result implies that a binary, homogeneous Crump–Mode–Jagers process is the local time process, in time, of a suitable Lévy process stopped at 0.

Definition 3.3.2 (Local time process)**.** *For $f \in \mathscr{D}(\mathbb{R})$ and $t \in \mathbb{R}_+$, let $\mu_{t,f}$ be the measure defined by*

$$\int_{\mathbb{R}} \varphi(x)\mu_{t,f}(\mathrm{d}x) = \int_0^t \varphi(f(s))\mathrm{d}s$$

for every measurable function $\varphi \geq 0$. When $\mu_{t,f}$ is absolutely continuous with respect to Lebesgue measure, we denote by $L(f)(\cdot,t)$ its Radon–Nikodym derivative restricted to $[0,\infty)$, satisfying the so-called occupation density formula

$$\int_0^t \varphi(f(s))ds = \int_0^\infty \varphi(x)L(f)(x,t)\mathrm{d}x \tag{3.9}$$

for any $t \geq 0$ and any measurable function $\varphi \geq 0$ with $\varphi(x) = 0$ for $x < 0$.

The measure $\mu_{t,f}$ is known as the occupation measure and the functional $L(f)$ as the local time process of f. It is uniquely determined up to sets of zero Lebesgue measure (in the space variable). Note that $L = L(X)$ under $\mathbb{P}_n^x$ is well defined and is simply given by $L(x,t) = \sum_{s \leq t} \mathbb{1}_{\{X(s)=x\}}$. In the sequel we will be especially interested in the local time process in space up to the time T_0 of the first visit of 0, which will be denoted by $L_0 = L(\cdot, T_0)$. Whenever convenient, we consider L_0 as an operator with domain and range $\mathscr{D}(\mathbb{R})$, which for instance makes it possible to write $\mathbb{P}_n^x \circ L_0^{-1}$ for the law of L_0 under $\mathbb{P}_n^x$.

Proposition 3.3.1 states that $\mathbb{P}_n^x \circ L_0^{-1}$ is the law of the nth Crump–Mode–Jagers process started with one ancestor with deterministic life length x. Studying the scaling limits of binary, homogeneous Crump–Mode–Jagers processes therefore reduces to studying the scaling limits of local time processes of Lévy processes.

Scaling limits of local time processes of Lévy processes

For the sequence of local time processes to converge, it is natural to assume that the Lévy processes themselves converge. For this reason, we fix some sequence $s_n \to \infty$ and consider $\mathbf{P}_n^x$ the law of the scaled process $X(nt)/s_n$ under $\mathbb{P}_n^x$. Throughout this section, we make the following assumption:

Assumptions on $\mathbf{P}_n^0$: for each $n \in \mathbb{N}$, X under $\mathbf{P}_n^0$ does not drift $+\infty$. Moreover, $\mathbf{P}_n^0 \xrightarrow{\mathrm{w}} \mathbf{P}^0$ where $\mathbf{P}^x$ is the law of a Lévy process with infinite variation started at $x \in \mathbb{R}$.

Note that the assumption that X under $\mathbf{P}_n^0$ does not drift $+\infty$ is equivalent to assuming that $\lambda\mathbb{E}(V_n) \leq 1$. Further, we will consider two specific excursion measures associated to $\mathbf{P}^0$:

- $\mathscr{N}$, the excursion measure of X under $\mathbf{P}^0$ normalized by considering the local time at 0 equal to $L(0,t)$;
- $\underline{\mathscr{N}}$, the excursion measure of $R(X)$ under $\mathbf{P}^0$ normalized by considering the local time at 0 equal to $-\min(0, \inf_{0\leq s\leq t} X(s))$.

Under the assumption $\mathbf{P}_n^0 \xrightarrow{\mathrm{w}} \mathbf{P}^0$, the occupation density formula (3.9) suggests that the convergence of, say, $L(\cdot,t)$ for fixed t is largely a matter of tightness. Indeed, if $\mathbf{P}_n^0$ converges, then so does $\int_0^t \varphi(X(s))\mathrm{d}s$ and so any accumulation point of $L(\cdot,t)$ under $\mathbf{P}_n^0$ should also satisfy (3.9). There are technical details to work out, but this is essentially the way to go and this is the approach we followed in the finite variance case where $\mathbf{P}^0$ has zero Lévy measure. In this case, tightness is easily established in great generality thanks to a queueing argument, see forthcoming Corollary 4.2.4.

However, tightness turns out to be a challenging issue in the infinite variance case where $\mathbf{P}^0$ has non-zero Lévy measure. This will be discussed in more details in Section 3.3.2, but consider the following simple fact: L_0 under $\mathbf{P}^0$ may fail to be càdlàg in space. More precisely, Barlow [6] derived a necessary and sufficient condition for L under $\mathbf{P}^0$ to be jointly continuous and showed that, when this condition fails, L is unbounded in space on any non-empty time-interval. Although Barlow's condition is in general hard to check, we showed in [LS15] that it takes a simple form for a spectrally positive Lévy process.

Lemma 3.3.3. *Let Ψ be the Laplace exponent of $\mathbf{P}^0$. The local time process of X under $\mathbf{P}^0$ is jointly continuous if and only if*

$$\int^{\infty} \frac{\mathrm{d}\lambda}{\Psi(\lambda)\sqrt{\log \lambda}} < \infty. \tag{3.10}$$

Thus when $\int^{\infty} \mathrm{d}\lambda / (\Psi(\lambda)\sqrt{\log \lambda}) = \infty$, there is no hope for the sequence of local time processes to be tight. It seems reasonable to conjecture that tightness will hold whenever (3.10) is satisfied, but we have only been able to prove this under more specific assumptions. Namely, we will consider two cases:

Finite variance case: $s_n = n^{1/2}$ and $\mathbf{P}^x$ is the law of a Brownian motion started at $x \in \mathbb{R}$ and drifting to $-\infty$;

Infinite variance case: there exists $\alpha \in (1,2)$ such that $s_n = n^{1/\alpha}$ and $\mathbb{P}(V_n \geq x) = (1+x)^{-\alpha}$ for every $n \in \mathbb{N}$ and $x \in \mathbb{R}_+$.

Moreover, as mentioned earlier we will be specifically interested in L_0, i.e., the local time process in space until the first hitting time of 0. Since X under $\mathbf{P}^0$ has infinite variation, we have $\mathbf{P}^0(T_0 = 0) = 1$ so that for any finite $x > 0$, $L_0 = L(\cdot, T_0)$ under $\mathbf{P}_n^x$ converges in distribution to the trivial excursion which takes the constant value 0. For this reason, some kind of conditioning is called upon to get a non-trivial limit.

Let in the sequel V_n^* be the forward-recurrence time of V_n, i.e., the random variable with density $\mathbb{P}(V_n \geq x)/\mathbb{E}(V_n)$ with respect to Lebesgue measure, and let $\mathbf{P}_n(\cdot) = \int \mathbf{P}_n^x(\cdot)\mathbb{P}(V_n \in \mathrm{d}x)$ and $\mathbf{P}_n^*(\cdot) = \int \mathbf{P}_n^x(\cdot)\mathbb{P}(V_n^* \in \mathrm{d}x)$ which correspond to the law of the Lévy process $\mathbf{P}_n^0$ started from a random initial condition, distributed according to V_n or V_n^*, respectively.

Theorem 3.3.4. *Let $\varepsilon > 0$. In the finite variance case, L_0 under $\mathbf{P}_n(\cdot \mid T_0 > \varepsilon)$ is tight. In the infinite variance case, L_0 under $\mathbf{P}_n(\cdot \mid \sup \sigma > \varepsilon)$ and $\mathbf{P}_n^*(\cdot \mid \sup \sigma > \varepsilon)$ are tight.*

Tightness in the finite variance case is proved thanks to queueing arguments, see forthcoming Lemma 4.2.2 and Theorem 4.2.3. These queueing arguments do not work anymore in the infinite variance case, and so to prove tightness in this case we control oscillations "by hand" and use Theorem 2.1.4 whereby we need to control

$$\mathbf{P}_n^{x_0}\left(|L_0(c) - L_0(b)| \wedge |L_0(b) - L_0(a)| \geq \lambda \mid L_0(\varepsilon) > 0\right).$$

The approach to control this probability will be explained in Section 3.2.2 below.

As mentioned earlier, the occupation density formula suggests that weak convergence holds as soon as tightness does. This is the path we have followed in [LSZ13] in the finite variance case, proving that L_0 under $\mathbf{P}_n(\cdot \mid T_0 > \varepsilon)$ converges in distribution to L_0 under $\underline{\mathcal{N}}(\cdot \mid T_0 > \varepsilon)$. We could also have followed this approach in the infinite variance case, but we have actually showed a stronger results, namely that finite-dimensional distributions always converge, even when the limiting process is not càdlàg. In the infinite variance case, whether we consider L_0 under $\mathbf{P}_n$ or $\mathbf{P}_n^*$ leads to different limits.

Theorem 3.3.5. *Let $\varepsilon > 0$. If $\mathbf{P}^0$ has non-zero Lévy measure, then:*

- *L_0 under $\mathbf{P}_n(\cdot \mid \sup\sigma > \varepsilon)$ converges in the sense of finite-dimensional distributions to L_0 under $\underline{\mathcal{N}}(\cdot \mid \sup\sigma > \varepsilon)$;*
- *L_0 under $\mathbf{P}^*_n(\cdot \mid \sup\sigma > \varepsilon)$ converges in the sense of finite-dimensional distributions to L_0 under $\mathcal{N}(\cdot \mid \sup\sigma > \varepsilon)$.*

The fact that $\underline{\mathcal{N}}$ appears when considering $\mathbf{P}_n$, and $\mathcal{N}$ when considering $\mathbf{P}^*_n$, has a simple explanation. Indeed, under $\mathbf{P}^0_n$, the reflected process $R(X)$ stays at 0 for an exponential duration and then starts a positive excursion, whose initial value is distributed as V_n; in particular, $\mathbf{P}_n(\cdot \mid \sup\sigma > \varepsilon)$ is the law of the first excursion with height $> \varepsilon$ of $R(X)$.

On the other hand, to understand the link between $\mathbf{P}^*_n$ and $\mathcal{N}$ one needs to recall that V^*_n is also the law of the overshoot of X under $\mathbb{P}_n$, i.e., V^*_n is equal in distribution to of $X(T^\uparrow)$ under $\mathbb{P}^0_n(\cdot \mid T^\uparrow < \infty)$ with $T^\uparrow = \inf\{t > 0 : X(t) > 0\}$. In particular, this implies that $\mathbf{P}^*_n(\cdot \mid \sup\sigma > \varepsilon)$ is the law of the first positive excursion with height $> \varepsilon$ of X under $\mathbf{P}^0_n(\cdot \mid T^\uparrow < \infty)$.

Combining Theorems 3.3.4 and 3.3.5, we have the following result.

Corollary 3.3.6. *Let $\varepsilon > 0$. In the finite variance case, L_0 under $\mathbf{P}_n(\cdot \mid T_0 > \varepsilon)$ converges in distribution to L_0 under $\underline{\mathcal{N}}(\cdot \mid T_0 > \varepsilon)$.*

In the infinite variance case, L_0 under $\mathbf{P}_n(\cdot \mid \sup\sigma > \varepsilon)$ and $\mathbf{P}^*_n(\cdot \mid \sup\sigma > \varepsilon)$ converge in distribution to L_0 under $\underline{\mathcal{N}}(\cdot \mid \sup\sigma > \varepsilon)$ and $\mathcal{N}(\cdot \mid \sup\sigma > \varepsilon)$, respectively.

Back to Crump–Mode–Jagers processes

Let us now go back to our initial object of interest, namely binary, homogeneous Crump–Mode–Jagers processes, and let $\mathbb{Z}^z_n$ (resp. $\mathbb{Z}^{z*}_n$) be the law of the nth Crump–Mode–Jagers process started with $z \in \mathbb{Z}_+$ individuals with i.i.d. life lengths equal in distribution to V_n (resp. V^*_n). As discussed earlier, Proposition 3.3.1 has the following fundamental consequence.

Lemma 3.3.7. *If X under $\mathbb{P}^0_n$ does not drift to $+\infty$, then $\mathbb{Z}^1_n$ (resp. $\mathbb{Z}^{1*}_n$) is the law of L_0 under $\mathbb{P}_n$ (resp. $\mathbb{P}^*_n$).*

Thus to understand the scaling limits of $\mathbb{Z}^1_n$ or $\mathbb{Z}^{1*}_n$, one needs to understand how the time/space scaling of the Lévy process affects the local time process. This is easily seen by defining $X_n(t) = X(nt)/s_n$ and considering the occupation density formula, which entails

$$\begin{aligned}
\int \varphi(x) L(X_n)(x,t)\mathrm{d}x &= \int_0^t \varphi(X_n(s))\mathrm{d}s && \text{(by definition of } L(X_n))\\
&= \frac{1}{n}\int_0^{nt} \varphi(X(s)/s_n)\mathrm{d}s && \text{(change of variables)}\\
&= \frac{1}{n}\int \varphi(x/s_n) L(x,nt)\mathrm{d}s && \text{(by definition of } L = L(X))\\
&= \int \varphi(x)\left(\frac{s_n}{n} L(s_n x, nt)\right)\mathrm{d}s && \text{(change of variables).}
\end{aligned}$$

This shows that the proper normalization of $\mathbb{Z}^1_n$ is obtained by scaling time by s_n and space by $r_n = n/s_n$, which leads us to define $\mathbf{Z}^z_n$ (resp. $\mathbf{Z}^{z*}_n$) the law of $X(s_n t)/r_n$ under $\mathbb{Z}^z_n$ (resp. $\mathbb{Z}^{z*}_n$). Note that a consequence of the assumption $\mathbf{P}^0_n \xrightarrow{\mathrm{d}} \mathbf{P}^0$ with $\mathbf{P}^0$ the law of a Lévy process with infinite variation is that $r_n \to \infty$ (see Lemma 3.4 in [LS15]), and so under this scaling the jumps of X under $\mathbf{Z}_n^{z*}$ are of vanishing size.

The following result is a mere rewriting of Theorems 3.3.4 and 3.3.5, where we introduce $\mathscr{L} = \mathscr{N} \circ L_0^{-1}$ and $\underline{\mathscr{L}} = \underline{\mathscr{N}} \circ L_0^{-1}$, i.e., $\mathscr{L}(f) = \mathscr{N}(f(L_0))$ and $\underline{\mathscr{L}}(f) = \underline{\mathscr{N}}(f(L_0))$ for any measurable function $f : \mathscr{D}(\mathbb{R}) \to \mathbb{R}_+$. Note that $\sup \sigma = T_0(L_0)$ and $T_0 = \int L_0$ (by the occupation density formula, and where $\int f = \int_0^\infty f(x)\mathrm{d}x$), which explains the different conditionings in the following statement.

Theorem 3.3.8. *Let $\varepsilon > 0$. In the finite variance case,*

$$\mathbf{Z}_n^1\left(\cdot \,\middle|\, \int X > \varepsilon\right) \xrightarrow{\mathrm{w}} \underline{\mathscr{L}}\left(\cdot \,\middle|\, \int X > \varepsilon\right).$$

If $\mathbf{P}^0$ has non-zero Lévy measure, then

$$\mathbf{Z}_n^1(\cdot \mid T_0 > \varepsilon) \xrightarrow{\mathrm{fdd}} \underline{\mathscr{L}}(\cdot \mid T_0 > \varepsilon) \;\; \textit{and} \;\; \mathbf{Z}_n^{1*}(\cdot \mid T_0 > \varepsilon) \xrightarrow{\mathrm{fdd}} \mathscr{L}(\cdot \mid T_0 > \varepsilon).$$

In the infinite variance case,

$$\mathbf{Z}_n^1(\cdot \mid T_0 > \varepsilon) \xrightarrow{\mathrm{w}} \underline{\mathscr{L}}(\cdot \mid T_0 > \varepsilon) \;\; \textit{and} \mathbf{Z}_n^{1*}(\cdot \mid T_0 > \varepsilon) \xrightarrow{\mathrm{w}} \mathscr{L}(\cdot \mid T_0 > \varepsilon).$$

Building on these convergence results at the excursion level, we can also consider macroscopic initial conditions, i.e., the convergence of $\mathbf{Z}_n^{z_n*}$ for some sequence $z_n \to \infty$. Let $T^L(\zeta)$ for $\zeta \in \mathbb{R}_+$ be the first time the amount of local time associated to X accumulated at level 0 exceeds ζ:

$$T^L(\zeta) = \inf\{t \in \mathbb{R}_+ : L(0,t) \geq \zeta\}.$$

Then under $\mathbb{P}_n^0(\cdot \mid T^L(z) < \infty)$ for $z \in \mathbb{Z}_+$, the process $L(\cdot, T^L(z))$ is equal in distribution to the sum of z i.i.d. copies of L_0 under $\mathbb{P}_n^* = \int \mathbb{P}_n^x(\cdot)\mathbb{P}(V_n^* \in \mathrm{d}x)$: this is a direct consequence of the strong Markov property, and this also corresponds to the branching property of $\mathbb{Z}_n^{z*}$. In particular, after scaling in time and space we obtain by Lemma 3.3.7 that if X under $\mathbb{P}_n^0$ does not drift to $+\infty$, then $\mathbf{Z}_n^{z*}$ for $z \in \mathbb{Z}_+$ is the law of $L(\cdot, T^L(z/r_n))$ under $\mathbf{P}_n^0(\cdot \mid T^L(z/r_n) < \infty)$. Passing to the limit, we then obtain the following result.

Theorem 3.3.9. *Let $(z_n, n \in \mathbb{Z}_+)$ be an integer-valued sequence such that $z_n/r_n \to \zeta \in \mathbb{R}_+$.*

In the finite variance case, $\mathbf{Z}_n^{z_n*}$ converges weakly to the law of $L(\cdot, T^L(\zeta))$ under $\mathbf{P}^0(\cdot \mid T^L(\zeta) < \infty)$.

If $\mathbf{P}^0$ has non-zero Lévy measure, $\mathbf{Z}_n^{z_n*}$ converges in the sense of finite-dimensional distributions to $L(\cdot, T^L(\zeta))$ under $\mathbf{P}^0(\cdot \mid T^L(\zeta) < \infty)$.

In the infinite variance case, $\mathbf{Z}_n^{z_n*}$ converges weakly to the law of $L(\cdot, T^L(\zeta))$ under $\mathbf{P}^0(\cdot \mid T^L(\zeta) < \infty)$.

3.3.2 Discussion

Generalizations

The results presented above actually merge two sets of results, obtained in [LSZ13] in the finite variance case and in [LS15] when $\mathbf{P}^0$ has non-zero Lévy measure. This "historical" reason explains why different conditionings are considered in the finite and infinite variance cases (e.g., in Theorem 3.3.8), but there is little doubt that the results hold in all cases considered for any of the two conditionings. Moreover, the techniques of [LS15] could most probably be used to remove the assumption that X drifts to $-\infty$ in the finite variance case, at least regarding the convergence of finite-dimensional distributions. Finally, because of their interpretation in queueing theory and for branching processes, we have restricted our attention to local time processes in $\mathbb{R}_+$ but similar results should hold when considering local time processes in $\mathbb{R}$.

Tightness

As explained after Theorem 3.3.4, tightness in the finite variance case is proved thanks to queueing arguments and in the infinite variance case, we need to control the probability

$$\mathbf{P}_n^{x_0}\left(|L_0(c) - L_0(b)| \wedge |L_0(b) - L_0(a)| \geq \lambda \mid L_0(\varepsilon) > 0\right)$$

in order to invoke Theorem 2.1.4. Such a control is achieved by giving an explicit description of the law of $|L_0(c) - L_0(b)| \wedge |L_0(b) - L_0(a)|$ under $\mathbb{P}_n^{x_0}(\cdot \mid L_0(a) > 0)$ in terms of geometric random variables (it is then not difficult to change the conditioning to $\{L_0(\varepsilon) > 0\}$). To illustrate this point, let $T_a = \inf\{t > 0 : X(t) = a\}$ for $a \in \mathbb{R}$ and recall that $L_0(a)$ counts the number of visits of X to a up to T_0 so that conditionally on $\{L_0(a) > 0\}$, X visits a before 0. Then, starting from a, the probability of hitting a before 0 is equal to $p_n(a) = \mathbb{P}_n^a(T_a < T_0)$ and to the strong Markov property implies that $L_0(a) - 1$ under $\mathbb{P}_n^a$ is a geometric random variable with parameter $p_n(a)$.

To compute the joint law of $(L_0(b), L_0(a))$, we then decompose the path of X in-between successive visits to a to write

$$L_0(b) = \sum_{k=1}^{L_0(a)-1} \xi_k$$

where ξ_k counts the number of visits of X to b in-between two successive visits to a. In particular, conditionally on $L_0(a)$ the ξ_k are i.i.d. with $\xi_k = 0$ with probability $\mathbb{P}_n^a(T_a < T_b \mid T_a < T_0)$ – starting from a and conditionally on returning to a before visiting 0, X returns to a before visiting b – while with the remaining probability $\xi_k - 1$ is a geometric random variable with parameter $\mathbb{P}_n^b(T_b < T_a)$ which is the probability that, starting from b, X returns to b before visiting a.

Thanks to such path decompositions, we get an explicit expression for the law of $|L_0(c) - L_0(b)| \wedge |L_0(b) - L_0(a)|$. However explicit, it remains a difficult problem to derive a bound which has the required uniformity in n, a, b, c and λ as per Theorem 2.1.4, which is the reason why we restricted ourselves in the infinite variance case to the case where the law of V_n has a very specific expression. It must be noted that similar difficulties were faced in previous works on scaling limits of local time processes [13, 14, 15, 26, 67, 93], which we briefly discuss now, see [LS15] for more details.

First of all, the present case, i.e., scaling limits of local time processes associated to finite variation Lévy processes, has barely been considered: we are only aware of a paper by Khoshnevisan [67], who assumed $\mathbb{E}(V_n^4) < \infty$ and derived strong invariance principles with explicit convergence rates thanks to embedding techniques.

In contrast, scaling limits of local time processes associated to random walks have been intensively studied in the non-triangular and arithmetic case, i.e., the step distribution ξ of the random walk does not depend on n and is arithmetic. In this case, complete results have been obtained by Borodin [13, 14], who proved invariance principle if $\mathbb{E}(\xi^2) < \infty$ or if ξ is in the domain of attraction of a stable law with index $1 < \alpha < 2$.

On the other hand, the picture is far to be as complete in the non-arithmetic case. First of all, in this case the very definition of the local time process is unclear since it cannot be defined by keeping track of the number of visits to different points in space. In Csörgő and Révész [26] for instance, five different definitions are discussed. In the (non-arithmetic and) finite variance case, Perkins [93] proved convergence of the finite-dimensional distributions if $\mathbb{E}(\xi^2) < \infty$, and weak convergence if $\mathbb{E}(\xi^4) < \infty$ and $\limsup_{|t| \to \infty} |\mathbb{E}(e^{it\xi})| < 1$; see also [15, 26]. In the non-arithmetic and infinite variance case similar to the one considered here (let alone in a triangular setting), I am not aware of any previous result.

This discrepancy between the arithmetic and non-arithmetic cases, and the fact that Perkins needs stronger moment assumptions in the non-arithmetic case for weak convergence, reflects the fact that tightness is significantly more difficult in the non-arithmetic case. Indeed, in the non-arithmetic case one has to control oscillations of the local time process for points which can be arbitrarily close, whereas in the arithmetic case the local time process stays constant in-between two points of the support of the step distribution. It turns out that these small oscillations are the most difficult to control.

Finite-dimensional convergence

In contrast to tightness which we have only been able to prove under specific assumptions, Theorems 3.3.5 and 3.3.9 show that finite-dimensional distributions converge in great generality, even when the limiting process is not càdlàg. Such results rely on explicit description of the law of $(L_0(a), a \in A)$ with A a finite subset of $(0, \infty)$.

More precisely, the main idea is to decompose the path of X in a suitable way and to reduce the problem to the convergence of some explicit Markov chain on A. Namely, let $M = (M_k, k = 1, \dots, K)$ be the finite sequence with values in A which keeps track of the successively distinct elements of A visited by X before the first visit to 0, and let

$$S(a) = \sum_{k=1}^{\infty} \mathbb{1}_{\{M_k = a\}}$$

be the number of visits of M to $a \in A$. The idea to compute $(L_0(a), a \in A)$ is then to sum up the amount of local time accumulated at a in-between two visits of M to a: for the pre-limit processes, this takes the form of a sum of geometric random variables, and in the limit this takes the form of a sum of exponential random variables, by excursion theory. Since the parameters of the geometric random variables involved converge after proper normalization to the parameter of the exponential random variables, this indeed reduces the problem to the convergence of the Markov chain M, which is done by showing convergence of the initial distributions and transition probabilities.

Finite vs. infinite variance

Theorem 3.3.9 states that the scaling limits of binary, homogeneous Crump–Mode–Jagers processes are the local time processes of spectrally positive Lévy processes. On the other hand, it is known for a long time that the scaling limit of Markovian

Crump–Mode–Jagers processes (which corresponds to the case where V_n is exponentially distributed) is Feller diffusion, see for instance Helland [52]. Fortunately, these two results are consistent since it is known by a variation of the original Ray–Knight theorems [73, 96] that $L(\cdot, T^L(\zeta))$ under $\mathbf{P}^0(\cdot \mid T^L(\zeta) < \infty)$ is Feller diffusion. In particular, Theorem 3.3.9 implies that Feller diffusion is the universality class of binary, homogeneous Crump–Mode–Jagers processes whose life length distribution has, essentially, a finite variance.

In contrast, $L(\cdot, T^L(\zeta))$ under $\mathbf{P}^0(\cdot \mid T^L(\zeta))$ does not seem to have any simple probabilistic description when $\mathbf{P}^0$ has non-zero Lévy measure. In this case, X under $\mathbf{P}^0$ is discontinuous and so results of Eisenbaum and Kaspi [32] imply that $L(\cdot, T^L(\zeta))$ under $\mathbf{P}^0(\cdot \mid T^L(\zeta))$ is not Markov. In particular, in this case the scaling limits of binary, homogenous Crump–Mode–Jagers processes do not belong to the class of CSBP's. To my knowledge, this is the first result of this kind, although Sagitov [108] has some results of a similar flavor.

This dichotomy, finite variance/Markovian scaling limit and infinite variance/non-Markovian scaling limit, has a simple intuitive explanation. When V_n has a finite variance, individuals do not live for a long time and so everything happens as if each individual was giving birth to all its children simultaneously. In this case, although the pre-limit processes are not Markov due to the possible residual life times, in the limit this does not matter and the process becomes Markovian. When V_n has infinite variance however, this picture is no longer true: some individuals live for a macroscopic duration (this corresponds to the jumps of the limiting jumping contour process $\mathbf{P}^0$), these macroscopic residual life times matter and the limiting process cannot be Markov.

This intuition has a particularly clear interpretation in terms of trees: in the finite variance case, edges of the Crump–Mode–Jagers tree are short which makes it “close” to its genealogical tree, which is simply a Galton–Watson tree. This suggests that the results of Theorem 3.3.9 could be pushed further and that, more generally, provided V_n is “small”, scaling limits of Crump–Mode–Jagers processes should belong to the class of CSBP's. In the next section such ideas will be developed, where instead of focusing on the Crump–Mode–Jagers processes themselves we will study the height and contour processes of the corresponding Crump–Mode–Jagers trees.

Discontinuity at 0

Another interesting feature of Theorem 3.3.9 is the fact that in order to obtain a proper limit one needs for the initial individuals to start with the size-biased life time distribution V_n^*, rather than the “regular” life time distribution V_n. As far as I know, this fact is unusual in the context of branching processes. As explained earlier, in our case it comes from the fact that V_n^* is equal in distribution to the overshoot of X under $\mathbb{P}_n^0$.

To complete the picture, let us informally discuss the asymptotic behavior of $\mathbf{Z}_n^{z_n}$ in the finite variance case. Let a and b be the drift and Gaussian coefficient of $\mathbf{P}^0$, and assume moreover that $\lambda_n \to \lambda \in (0,\infty)$. Then Theorem 3.3.9 implies that for any integer sequence (z_n) with $z_n/n^{1/2} \to \zeta \in (0,\infty)$, $\mathbf{Z}_n^{z_n *}$ converges weakly to the law of Feller diffusion started at ζ, the drift and Gaussian coefficients of which can be computed to be equal to $2a/b$ and $4/b$, respectively. In contrast, it can be proved that, under the same assumptions, $\mathbf{Z}_n^{z_n}$ converges in the sense of finite-dimensional distributions to a process Z such that $Z(0) = \zeta$, while $(Z(t), t > 0)$ is equal in distribution to Feller diffusion with the same drift and Gaussian coefficient, but started at $2\zeta/(b\lambda)$. Thus when $b\lambda \neq 2$, starting with the "regular" life length V_n for the ancestors creates a discontinuity at time 0. Although surprising from a branching perspective, this phenomenon has a natural queueing interpretation in terms of the state-space collapse phenomenon which will be discussed in Section 4.2.3.

3.4 Crump–Mode–Jagers trees with short edges

In this section we discuss the convergence of the height and contour processes associated to Crump–Mode–Jagers trees with "short" edges: results of this section can be found in [SS].

3.4.1 The case of short edges

In the previous section we were interested in binary, homogeneous Crump–Mode–Jagers processes. Despite their explicit probabilistic description as local time processes of "simple" Lévy processes, proving weak convergence turned out to be a challenging issue in the infinite variance case: at this point, it is therefore not clear how to study directly more general cases.

However, the tree interpretation for the dichotomy between the finite and infinite variance cases provided in Section 3.3.2 suggests that, if individuals do not live for a long time, the corresponding Crump–Mode–Jagers tree should be "close" to its genealogical Galton–Watson counterpart. Guided by this intuition, we have studied in [SS] the asymptotic behavior of the height and contour processes of Crump–Mode–Jagers trees and could establish, in the "short" edge case, the following general sufficient conditions for the convergence of their finite-dimensional distributions.

In this section, we consider as canonical space the space of chronological trees, and $\mathscr{H}$, $\mathscr{C}$, $\mathbb{H}$, $\mathbb{C}$ and S introduced in Section 3.1.3 denote the genealogical and chronological height and contour processes and the Lukasiewicz path associated to the canonical tree. The intuition that a Crump–Mode–Jagers tree should be "close" to its genealogical Galton–Watson counterpart means that $\mathbb{H}$ and $\mathscr{H}$ as well as $\mathbb{C}$ and $\mathscr{C}$ should be close.

Probabilistic set-up

For each $n \in \mathbb{N}$, we consider $(V_n, \mathscr{P}_n)$ in $\mathbb{S}$ with $\mathbb{E}(|\mathscr{P}_n|) \leq 1$, and we consider $\mathbb{P}_n$ the law of the Crump–Mode–Jagers tree with parameter $(V_n, \mathscr{P}_n)$. We further assume that the trees are near-critical, in the sense that

$$\lim_{n\to\infty} \mathbb{E}(|\mathscr{P}_n|) = 1.$$

For $\nu \in \mathscr{M}_p((0,\infty))$ and $r \in \{1, \ldots, |\nu|\}$, we define

$$\Lambda(\nu, r) = \sup\{x \geq 0 : \nu[x,\infty) \leq r\}$$

which is the location of the rth atom of ν, where atoms are ranked from largest to smallest; for instance, $\pi(\nu) = \Lambda(\nu, 1)$. We finally define A_n the random variable with distribution

$$\mathbb{E}\left[f(A_n)\right] = \frac{1}{\mathbb{E}(|\mathscr{P}_n|)} \mathbb{E}\left(\sum_{r=1}^{|\mathscr{P}_n|} f\big(\Lambda(\mathscr{P}_n, r)\big)\right),$$

so that in particular,

$$\mathbb{E}(A_n) = \mathbb{E}\left(\int x\mathscr{P}_n(\mathrm{d}x)\right).$$

Convergence of the chronological height process

We now state our main results concerning the convergence of the chronological height process: we fix a sequence $\varepsilon_n \to 0$ and consider the rescaled processes

$$\mathscr{H}_n(t) = \varepsilon_n \mathscr{H}([nt]) \text{ and } \mathbb{H}_n(t) = \varepsilon_n \mathbb{H}([nt]). \tag{3.11}$$

We exhibit a simple and explicit condition under which $\mathscr{H}_n$ and $\mathbb{H}_n$ are asymptotically proportional to one another. Our results will involve the following condition, the second part of which is automatically satisfied in the non-triangular case where the law of A_n does not depend on n.

Condition T-H. *For every $n \in \mathbb{N}$, $\mathbb{E}(A_n) < \infty$. Moreover, there exists an integrable random variable ξ with $\mathbb{E}\xi = 0$ such that $A_n - \mathbb{E}(A_n) \xrightarrow{\mathrm{d}} \xi$ and $\mathbb{E}[(A_n - \mathbb{E}(A_n))^+] \to \mathbb{E}(\xi^+)$.*

Theorem 3.4.1. *Let $t > 0$. If Condition* T-H *holds, $\mathscr{H}([nt]) \xrightarrow{\mathrm{d}} \infty$ and the sequence $(\mathscr{H}_n(t), n \geq \mathbb{N})$ is tight, then $\mathbb{H}_n(t) - \mathbb{E}(A_n)\mathscr{H}_n(t) \xrightarrow{\mathrm{d}} 0$.*

The following immediate corollary of this result states that under mild conditions on the A_n's, $\mathscr{H}_n$ and $\mathbb{H}_n$ are indeed asymptotically proportional to one another, at least in the sense of finite-dimensional distributions.

Corollary 3.4.2. *Assume that Condition* T-H *holds and that:*
(H1) *$n\varepsilon_n \to \infty$;*
(H2) *$\mathbb{E}(A_n) \to \alpha$ for some $\alpha \in (0,\infty)$;*
(H3) *$\mathscr{H}_n \xrightarrow{\mathrm{d}} \mathscr{H}_\infty$ for some $\mathscr{H}_\infty$ satisfying $\mathbb{P}(\mathscr{H}_\infty(t) > 0) = 1$ for every $t > 0$.*
Then $(\mathscr{H}_n, \mathbb{H}_n) \xrightarrow{\mathrm{d}} (\mathscr{H}_\infty, \alpha\mathscr{H}_\infty)$.

Note that conditions under which $\mathscr{H}_n \xrightarrow{\mathrm{d}} \mathscr{H}_\infty$ are well understood. If $S_n(t) = (1/n\varepsilon_n)S([nt])$, then the assumption $S_n \xrightarrow{\mathrm{d}} S_\infty$ for some Lévy process S_∞ with infinite variation and whose Lévy exponent satisfies $\int_1^\infty \mathrm{d}u/\psi(u) < \infty$, together with some mild assumption, ensures (H3), see Theorem 2.3.1 in Duquesne and Le Gall [31].

Moreover, this result highlights the key role played by the random variable A_n, and in particular its mean. Actually, A_n has a natural genealogical interpretation. Take a typical individual, say u, and consider one of its ancestors: then A_n is the age of this ancestor when giving birth to the next ancestor of u. In the critical case this interpretation was stated in Nerman [89], and the assumption $\mathbb{E}(A_n) < \infty$ is made in every previous work on scaling limits of Crump–Mode–Jagers processes that I am aware of, see for instance [104, 105, 106, 107, 108, 109, 110].

Convergence of the chronological contour process

Under the assumption $\mathbb{E}(A_n) \to \alpha < \infty$ and other mild conditions, Corollary 3.4.2 states that the genealogical and chronological height processes are essentially proportional to one another. We now state a result on the contour process when this assumption is not enforced, which makes it possible for the chronological and genealogical processes to scale in different ways. We thus consider two sequences ε_n and $\bar{\varepsilon}_n$, both converging to 0, rescale the genealogical processes using $\bar{\varepsilon}_n$ as

$$\mathscr{H}_n(t) = \bar{\varepsilon}_n \mathscr{H}([nt]),\ \mathscr{C}_n(t) = \bar{\varepsilon}_n \mathscr{C}(nt) \text{ and } S_n(t) = \frac{1}{n\bar{\varepsilon}_n} S([nt]),$$

and the chronological processes using ε_n as

$$\mathbb{H}_n(t) = \varepsilon_n \mathbb{H}([nt]) \text{ and } \mathbb{C}_n(t) = \varepsilon_n \mathbb{C}(nt).$$

In the Galton–Watson case, it is well-known that $\mathscr{C}_n$ is essentially obtained from $\mathscr{H}_n$ by a deterministic time-change under rather mild assumptions (essentially conditions (C2)–(C3) below). We now show that a similar statement holds at the chronological level under the following two assumptions.

Condition T-C1. *$(V_n, \mathscr{P}_n) \xrightarrow{\text{d}} (V_\infty, \mathscr{P}_\infty)$ for some random $(V_\infty, \mathscr{P}_\infty)$ in $\mathbb{S}$ such that $\mathbb{E}(V_\infty) < \infty$ and $\mathbb{E}(|\mathscr{P}_\infty|) = 1$.*

Let $V > 0$ be some random variable and G the additive subgroup generated by the support of its distribution. In the sequel we say that V is *non-arithmetic* if G is dense in $\mathbb{R}$; otherwise, we say that V is *arithmetic* and in this case, there exists a unique $h > 0$, called the *span* of V, such that $G = h\mathbb{Z}$. For a random variable $V > 0$ with finite mean, we define V^* as follows:

- if V is non-arithmetic, we define

$$\mathbb{P}(V^* \geq x) = \frac{1}{\mathbb{E}(V)} \int_x^\infty \mathbb{P}(V \geq y) \mathrm{d}y,\ x \in \mathbb{R}_+;$$

- if V is arithmetic with span h, we define

$$\mathbb{P}(V^* = kh) = \frac{1}{\mathbb{E}(V)} \mathbb{P}(V \geq kh),\ k \in \mathbb{N}.$$

Condition T-C2. *$V_n^* \xrightarrow{\text{d}} V_\infty^*$ with V_∞ as in Condition* T-C1*, and:*

- *if V_∞ is non-arithmetic, then V_n for each n is non-arithmetic;*
- *if V_∞ is arithmetic, then V_n for each n is arithmetic.*

In the sequel, we will refer to the first case as the *non-arithmetic case* and to the second case as the *arithmetic case*. Note that Conditions T-C1 and T-C2 (except for $\mathbb{E}(V_\infty) < \infty$ and $\mathbb{E}(|\mathscr{P}_\infty|) = 1$) are automatically satisfied in the non-triangular case where the law of $(V_n, \mathscr{P}_n)$ does not depend on n.

Theorem 3.4.3. *Assume that Conditions* T-C1 *and* T-C2 *hold and that:*

(C1) $\mathbb{E}(V_n) \to \mathbb{E}(V_\infty)$*;*

(C2) $\lim_{n\to\infty} n\varepsilon_n = \lim_{n\to\infty} n\bar{\varepsilon}_n = \infty$*;*

(C3) *$S_n \xrightarrow{\text{d}} S_\infty$ for some Lévy process S_∞ with infinite variation;*

(C4) *$(\mathscr{H}_n, \mathscr{C}_n) \xrightarrow{\text{d}} (\mathscr{H}_\infty, \mathscr{C}_\infty)$ with $\mathscr{H}_\infty, \mathscr{C}_\infty$ almost surely continuous and satisfying the condition $\mathbb{P}(\mathscr{H}_\infty(t), \mathscr{C}_\infty(t) > 0) = 1$ for every $t > 0$;*

(C5) *$\mathbb{H}_n \xrightarrow{\text{fdd}} \mathbb{H}_\infty$ with $\mathbb{H}_\infty$ almost surely continuous at* 0 *and satisfying the condition $\mathbb{P}(\mathbb{H}_\infty(t) > 0) = 1$ for every $t > 0$.*

Then $(\mathbb{H}_n, \mathbb{C}_n) \xrightarrow{\text{fdd}} \left(\mathbb{H}_\infty, \mathbb{H}_\infty \circ \varphi_\infty\right)$ *where* $\varphi_\infty(t) = t/(2\mathbb{E}(V_\infty))$.

Note that the assumptions discussed after Corollary 3.4.2 actually imply (C4) with $\mathscr{C}_\infty(t) = \mathscr{H}_\infty(t/2)$. Combining Theorems 3.4.1 and 3.4.3, we obtain the following joint convergence.

Corollary 3.4.4. *Assume that except for* (C5)*, the conditions of Theorem 3.4.1 and Theorem 3.4.3 hold with* $\bar{\varepsilon}_n = \varepsilon_n$*: then*

$$(\mathscr{H}_n, \mathscr{C}_n, \mathbb{H}_n, \mathbb{C}_n) \xrightarrow{\text{fdd}} \left(\mathscr{H}_\infty, \mathscr{H}_\infty(\cdot/2), \alpha\mathscr{H}_\infty, \alpha\mathscr{H}_\infty \circ \varphi_\infty\right).$$

In [109], Sagitov investigated in the non-triangular setting the size of a population generated from a Crump–Mode–Jagers tree conditioned to survive at large time under the short edge assumption, i.e., when $\mathbb{E}(V_1), \mathbb{E}(A_1) < \infty$ (see also Green [43]). It was shown that in the limit, the population size is described in terms of a CSBP where space and time are scaled analogously as in Corollary 3.4.4. As a consequence, the previous corollary can be seen as a genealogical version of [109]. We also note that in [109], the results are obtained through an entirely different approach, namely complex analytic computation involving some non-trivial extension of the renewal theorem.

3.4.2 Fundamental formula for $\mathbb{H}(n)$

To conclude this chapter on branching processes, we present a fundamental formula for $\mathbb{H}(n)$ which lays the foundation for proving the results stated above. In the Galton–Watson case, it is well-known that $\mathscr{H}(n)$ can be expressed in terms of the dual Lukasiewicz path, and one of the main contribution of [SS] is to extend this formula to $\mathscr{H}(n)$.

Iterative construction of a chronological forest from a sequence of sticks

In Section 3.1.3 we have started from a chronological tree and explained how to construct the sequence $((\zeta_{\ell(k)}, \nu_{\ell(k)}), k \in \mathbb{Z}_+)$ of sticks ranked in lexicographic order. To understand the forthcoming formula for $\mathbb{H}(k)$, we need to revert the viewpoint, i.e., we start from a sequence $(\mathbb{S}_k, k \in \mathbb{Z}_+)$ with $\mathbb{S}_k = (\zeta_k, \nu_k)$ from which we build a chronological tree in a consistent way, i.e., such that $(\mathbb{S}_k)$ is the sequence of sticks ranked in lexicographic order. This sequential construction is illustrated in Figure 3.4.

At time $k = 0$ we start with the empty forest and we add the stick $\mathbb{S}_0$ at time $k = 1$. In the case considered in Figure 3.4, ν_0 has two atoms which correspond to birth times of individuals, but these two atoms are not yet matched with the sticks corresponding to these individuals. These unmatched atoms are called *stubs*, and when there is at least one stub we apply the following rule:

Rule #1: if there is at least one stub, we graft the next stick to the highest stub.

Thus, we iteratively apply this rule until there is no more stub, at which point we have built a complete chronological tree. In Figure 3.4 we can apply this rule at time 1 and 2 and at time 3 there is no stub anymore. In such a case, we apply the following rule:

Rule #2: if there is no stub, we start a new tree with the next stick.

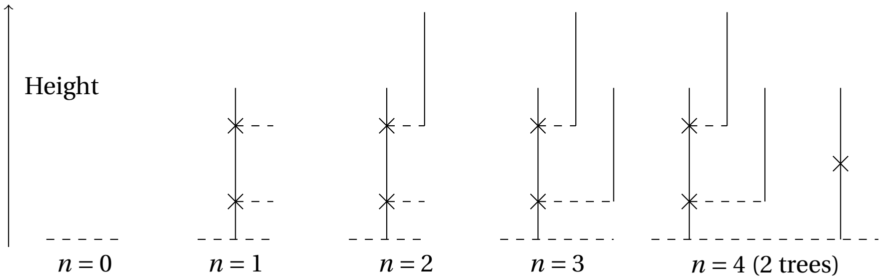


Figure 3.4: We start at $n=0$ with nothing, then add $\mathbb{S}_0$ at time $k=1$. At this time, there are two stubs and so the next stick $\mathbb{S}_1$ is grafted to the highest stub. At $k=2$, there is only one stub where the next stick $\mathbb{S}_2$ is grafted, leading to $\mathbb{F}^3$ which has no stub. The next step then starts the construction of the next tree.

Thus, starting at time $k=0$ from the empty forest and iterating $k+1$ times these two rules, we end up with a chronological forest which we denote $\mathbb{F}^k$, and is uniquely determined by $(\mathbb{S}_i, i=0,\dots,k)$. Starting from the infinite sequence $(\mathbb{S}_k, k\geq 0)$, we build in this way a forest $\mathbb{F}^\infty$, possibly consisting of infinitely many chronological trees.

Fundamental formula for $\mathbb{H}(n)$

This construction makes it possible to recover all the objects of interest in a transparent way from the so-called exploration process $(\rho_k, k\in\mathbb{Z}_+)$. Informally, ρ_k is the measure on $(0,\infty)$ which records the positions of the stubs in the kth stage of the above construction: formally, it is defined through the recursion

$$\rho_0=\mathbf{z} \text{ and } \rho_{k+1}=\rho_k-\epsilon_{\pi(\rho_k)}+\theta_{\pi(\rho_k)}(\nu_k),\ k\in\mathbb{Z}_+,$$

where by convention $\pi(\mathbf{z})=0$ and $\epsilon_0=\mathbf{z}$. From the exploration process we then recover the chronological height process $\mathbb{H}=(\mathbb{H}(k),k\in\mathbb{Z}_+)$ by

$$\mathbb{H}(k)=\pi(\rho_k)$$

and the Lukasiewicz path S by

$$S(0)=0 \text{ and } S(k)=|\rho_k|-\sum_{i=1}^k \mathbb{1}_{\{|\rho_i|=0\}}.$$

When all the objects are constructed in this way, it is natural to view them not as a function of the chronological tree but directly as a function of the sequence $(\mathbb{S}_k, k\in\mathbb{Z}_+)$. Actually, to avoid boundary issues when reverting time in the forthcoming arguments, we will view $\mathbb{H}$ as a mapping $\mathbb{H}:\mathbb{S}^\mathbb{Z}\to\mathbb{R}_+^\mathbb{N}$, $\mathbb{H}(k)$ as a mapping $\mathbb{H}(k):\mathbb{S}^\mathbb{Z}\to\mathbb{R}_+$, etc. This makes it possible to define the genealogical height process efficiently, namely, if $\mathscr{G}:\mathbb{S}^\mathbb{Z}\to\mathbb{S}^\mathbb{Z}$ is the genealogical operator which maps a sequence $((\zeta_k,\nu_k),k\in\mathbb{Z})$ to its associated genealogical sequence $((1,|\nu_k|\epsilon_1),k\in\mathbb{Z})$, then $\mathscr{H}=\mathbb{H}\circ\mathscr{G}$. Similarly, it is natural to view $\mathbb{P}_n$ as a probability distribution on $\mathbb{S}^\mathbb{Z}$, namely, under $\mathbb{P}_n$ the canonical sequence $((\zeta_k,\nu_k),k\in\mathbb{Z})$ is i.i.d. with common distribution $(V_n,\mathscr{P}_n)$.

Further, for $n \in \mathbb{Z}$ we consider the dual operator $\vartheta_n : \mathbb{S}^{\mathbb{Z}} \to \mathbb{S}^{\mathbb{Z}}$ defined by

$$\vartheta_n\left((s_k, k \in \mathbb{Z})\right) = (s_{n-1-k}, k \in \mathbb{Z}).$$

In the Galton–Watson case, it is well-known that many quantities of interest have an explicit expression in terms of the dual processes. For instance, let $(T(k))$ be the sequence of weak ascending ladder height times associated to S:

$$T(0) = 0 \ \text{ and } \ T(k+1) = \inf\{i > T(k) : S(i) \geq S(T(k))\}$$

Then the set

$$\mathscr{A}(n) = \left\{n - T(k) : 0 \leq T(k) \leq n\right\} \circ \vartheta_n = \left\{n - T(k) \circ \vartheta_n : 0 \leq T(k) \circ \vartheta_n \leq n\right\}, \ n \in \mathbb{Z}_+,$$

is the set of ancestors of n and in particular

$$\mathscr{H}(n) = |\mathscr{A}(n)|, \ n \in \mathbb{Z}_+,$$

see for instance Duquesne and Le Gall [31]. More precisely, $n - T(k) \circ \vartheta_n$ is the index of the kth ancestor of n, assuming that ancestors are ordered from highest to lowest date of birth (or height).

Entering now the realm of chronological trees, one can actually prove that $A(k) \circ \vartheta_n$ is the age of the kth ancestor when giving birth to the $(k-1)$st ancestor, where for $k \in \mathbb{N}$ with $T(k) < \infty$,

$$A(k) = \Lambda\left(\nu_{T(k)-1}, S(T(k-1)) - S(T(k)-1) + 1\right).$$

The following formula thus merely expresses the date of birth of an individual as the sum over its ancestors of their age when giving birth to the next ancestor. It may therefore seem obvious but combined with Lemma 3.4.6 which describes the probabilistic structure of the sequence $(A(k))$, it immediately entails Theorem 3.4.1.

Proposition 3.4.5. *For every $n \in \mathbb{Z}_+$, we have*

$$\mathbb{H}(n) = \left(\sum_{k:0<T(k)\leq n} A(k)\right) \circ \vartheta_n = \sum_{k:0<T(k)\circ\vartheta_n \leq n} \left(A(k) \circ \vartheta_n\right). \tag{3.12}$$

Let $\tau_\ell^- = \inf\{k \geq 0 : S(k) = -\ell\}$ for $\ell \geq 0$ be the record time associated to the strong descending height process of the Lukasiewicz path S: the following result explains the correlation structure, under $\mathbb{P}_n$, between the $T(k)$'s and the $A(k)$'s.

Lemma 3.4.6. *Let $\chi = \inf\{k \geq 0 : T(k) = \infty\}$. Then under $\mathbb{P}_n$, the sequence*

$$\left(\left(T(k) - T(k-1), A(k)\right), k = 1, \dots, \chi - 1\right)$$

is equal in distribution to $((\Delta T_n(k), A_n(k)), k = 1, \dots, \chi_n - 1)$, where the random variables $((\Delta T_n(k), A_n(k)), k \geq 1)$ are i.i.d. with common distribution $(\Delta T_n, A_n)$ satisfying:

$$\begin{aligned}\mathbb{E}&\left[f(A_n) \cdot g(\Delta T_n) \mid \Delta T_n(1) < \infty\right] \\ &= \frac{1}{\mathbb{E}(|\mathscr{P}_n|)} \sum_{t\geq 1} \sum_{x \geq 1} \mathbb{E}\left[f(\Lambda(\mathscr{P}_n, x))\, g(t); |\mathscr{P}_n| \geq x\right] \mathbb{P}_n\left(\tau_{x-1}^- = t-1\right).\end{aligned} \tag{3.13}$$

for every test functions f, g, and χ_n is an independent geometric random variable with parameter $1 - \mathbb{E}(|\mathscr{P}_n|)$.

In general, the correlation between the $T(k)$'s and the $A(k)$'s makes it difficult to exploit (3.12). However, when $\mathbb{E}(A_n) < \infty$, then the law of large numbers suggests the approximation

$$\left(\sum_{k:0<T(k)\leq n} A(k) \right) \circ \vartheta_n \approx \mathbb{E}(A_n) \times |\{k : 0 < T(k) \leq n\}| \circ \vartheta_n$$

which, in view of the relation $\mathscr{H}(k) = |\{k : 0 < T(k) \leq n\}| \circ \vartheta_n$ which stems from Proposition 3.4.5, gives an intuitive explanation for Theorem 3.4.1. The actual proof consists in making these arguments rigorous.

Chapter 4

Stochastic networks

Contents

4.1 Introduction 53

4.1.1 Brief overview 54

4.1.2 Common techniques 55

4.2 Processor-Sharing queue length process 56

4.2.1 Context 56

4.2.2 Main results 57

4.2.3 Discussion 62

4.3 A stochastic network with mobile customers 64

4.3.1 Model 64

4.3.2 Homogenization property and implications 66

4.3.3 Ideas 70

4.4 Lingering effect for queue-based scheduling algorithms 72

4.4.1 Context and illustrative example 73

4.4.2 Model description 75

4.4.3 Presentation of the lingering effect and main results 77

4.4.4 An intriguing heavy-tail phenomenon for $\beta \in (1,2)$ 79

4.1 Introduction

Stochastic networks do not form such a clear-cut class of stochastic processes as branching processes. My personal definition of stochastic networks is the class of stochastic models dealing with the problem of sharing resources in an uncertain environment. Telecommunication is a natural application area, a non-comprehensive list of which also includes manufacturing, logistics, transport, health care, biology and social networks. Both the source of uncertainty and the resources being shared depend on the application in mind: in telecommunication, the uncertainty may come from the customers' behavior or the radio conditions, and the resources can be the frequency spectrum or bandwidth; in transport, the uncertainty may come from the customers' behavior or weather conditions, and the resources can be the capacity of a public transport, train or highway; etc.

Down the road, the ultimate goal of the study of such models is to help design better networks, either through a better dimensioning or through a more efficient use of existing infrastructure. The dimensioning problem consists in dimensioning a network in order to meet target criteria, e.g., in terms of throughput, delay or fairness; in this case customers' behavior and the way resources are shared are typically input of the problem. Conversely, one can try to optimize a network: given a network architecture and customers' behavior, come up with efficient ways of sharing resources. Although the actual impact of the study of stochastic networks on real-world system is actively debated, there is no question that the general problematic of resource sharing spurred fascinating mathematical problems and solutions.

4.1.1 Brief overview

The study of stochastic networks can probably be traced back to the study of telephone networks and the foundation of queueing theory by Erlang [33, 34]. It became rapidly apparent that except in few special cases, explicit formulas could not be obtained which triggered the need for approximation techniques. In particular, the study of the heavy traffic regime was advocated by Kingman [68, 70] who studied the $G/G/1$ FIFO queue and also the Join-the-Shortest-Queue policy [69]. In contrast to branching processes, it is hard to pinpoint one canonical model of stochastic network which would play for stochastic networks the equivalent of the Galton–Watson process for branching processes. Classical models include the $M/M/1$ and $M/M/\infty$ queues and Jackson networks, as well as their generalizations to more general inter-arrival and service time distributions. Of course, many other models could naturally be included in this list.

The theory of single-server queues has received considerable attention since the foundation of queueing theory and a great deal is known about the $G/G/1$ queue, with finite or infinite buffer, customer abandonment, retrial, several classes of customers, etc. The performance of various service disciplines has also been intensively studied, in particular for First In First Out but also, for instance, for Last In First Out, Earliest Deadline First, Processor-Sharing, Shortest Remaining Processing Time and multiclass service disciplines such as Generalized Processor-Sharing and queues with priority. These various service disciplines have been studied in details by various authors, in particular in terms of delay since Loynes' foundational paper [87] settled the question of throughput for any work-conserving service discipline. It is for instance now well understood that the service distribution, in particular in terms of its tail being light or heavy, greatly impacts the delay performance of the service discipline considered. More recently, heavy traffic results at the process level have been obtained for several of these service disciplines, see for instance [46, 84, 85]. It is interesting to note that such kinds of results were initially advocated by Kingman in [70] in order to understand the scaling limits of the stationary distributions.

The study of the many-server queue was initiated by Erlang in order to dimension wired telephone networks, yielding the celebrated Erlang loss formula latter generalized in several directions. This topic was recently revived due to its application to the modeling of call centers. For such models, the (relatively) important number of parameters makes it possible to consider various asymptotic regimes, one of which is the famous Halfin–Whitt regime [49] which has attracted (and continues to attract) considerable attention in recent years.

Turning our attention to networks of interconnected queues/servers, we first note the seminal work of Reiman [100] concerning generalized Jackson networks and proving convergence in the heavy traffic regime of the queue length process toward a reflected Brownian motion in the orthant. This object was previously introduced by Harrison and Reiman [51]. More recently, Gamarnik and Zeevi [38] established the validity of the interchange of limits, whereby the stationary distributions of the generalized Jackson networks are proved to converge in heavy traffic toward the stationary distribution of the limiting reflected Brownian motion. Such results motivated a large amount of work dedicated to studying reflected Brownian motions in the orthant and also in more complex geometrical structures such as wedges. They were later generalized to a broader class of networks, called stochastic processing networks, and this approach proved to be very useful in control problems. Typically, optimal solutions to the control problem considered are intractable for the original stochastic networks, while they become tractable and sometimes even explicit for the limiting diffusion. In such case, solutions to the asymptotic control problem can be used to propose a control policy for the original network, which in the most optimistic cases can be shown to be asymptotically optimal in the heavy traffic regime, see for instance Bell and Williams [7].

4.1.2 Common techniques

Some techniques for proving convergence in the heavy traffic regime were already mentioned in Section 2.1.2, namely the approach leveraging the semimartingale decomposition and also the continuous mapping theorem. These two viewpoints have a lot in common, especially when representing the processes of interest via stochastic differential equations, and they have been extensively used to study stochastic networks. However, in order to address some features characteristic to stochastic networks, these methods have been pushed in some particular directions and new methods have been proposed. We discuss here briefly two topics which have played an important role in the recent development of stochastic networks: the Skorohod map and the state space collapse approach.

The difficulty in analyzing stochastic networks usually lies at the boundary of the state space. Consider for instance the $M/M/1$ queue: away from 0, it behaves like a random walk and thus converges to a Brownian motion. But what happens, in the limit, when the process hits 0? Although apparently innocuous and maybe even negligible at first sight, addressing such questions actually usually concentrates most of the technical work. The Skorohod map, sometimes referred to as reflection map, has emerged in recent years as a powerful approach to deal with this problem. Indeed, the Skorohod map makes it possible to represent a process constrained to live in some particular domain as the image of a “free” process pushed back to the interior of the state space when at the boundary. For instance, the $M/M/1$ queue length process can be represented as the reflection of a continuous-time random walk. Invoking the continuity of the Skorohod map together with the continuous mapping theorem, one then gets that the scaling limit of the $M/M/1$ queue is a reflected Brownian motion. Several extensions of the Skorohod map have been studied and applied to various stochastic networks, see for instance [29, 30, 95, 98].

Another approach developed in the context of stochastic networks is the state space collapse approach. Early on, it was observed that the scaling limits of many stochastic networks exhibit a form of reduction of the dimension of the state space: although the original stochastic network is multidimensional, typically one dimension for each queue of the network, the scaling limit lives in a lower-dimensional space, sometimes even one-dimensional. This was for instance observed by Kingman [69] in the context of two queues under the Join-the-Shortest-Queue service discipline, and later by Reiman [101] on various examples of networks. This phenomenon was systematically investigated by Bramson [22] and Williams [121] for open multiclass queueing networks and then observed in a wide variety of contexts, see for instance [115, 118].

4.2 Processor-Sharing queue length process

This section presents the results of [LSZ13, LS15] concerning the scaling limits of the single-server $M/G/1$-PS queue.

4.2.1 Context

The Processor-Sharing queue is a single-server queue in which the server splits its service capacity equally among all the customers present. For instance, if the server has a service capacity c and if there are exactly $q \geq 1$ customers in the queue during the time interval $[t, t+h]$ (in particular there is no arrival or departure), then the residual service requirement of each customer is decreased by ch/q during this time interval while the total workload is decreased by ch. Processor-Sharing is an idealization of the Round-Robin service discipline and plays a fundamental role in the modeling of communication networks such as the Internet or wireless networks.

Because of its importance in applications, many of its properties have been extensively studied in the literature, see for instance the survey by Yashkov [123] for results up to 1987. Expressions for Laplace transforms of various quantities of interest, both for the transient and stationary regimes, have been derived but they are often difficult to work with. In that context, heavy traffic approximations have provided valuable insight, see for instance [57, 124].

At the process level, fluid limits of the queue length process were investigated in Chen et al. [24], see also Jean-Marie and Robert [59] for the supercritical regime. Gromoll et al. [47] studied in detail the near-critical regime and proved fluid limits for the entire measure-valued descriptor of the queue. This opened the way to use the state space collapse approach of Bramson [22] and Williams [121] to derive heavy traffic approximation for the measure-valued $G/G/1$-PS queue in [45]. Note also that variants of the Processor-Sharing service discipline have been considered, for instance, in the multiclass setting, the Generalized Processor-Sharing service disciplines [30, 95] and recently the Limited Processor-Sharing service discipline [125, 126, 127].

Of particular interest to us are the results in the heavy traffic regime at the process level. In this context, Gromoll [45] considered the $G/G/1$-PS queue and extended the state space collapse approach of Bramson and Williams to establish scaling limits, assuming finite moments of order $4+\varepsilon$. Gromoll's results imply in particular, in this case, the convergence of the queue length process toward a reflected Brownian. Although this approach could probably be pushed to a finite variance assumption, an intrinsic limitation of the state space collapse approach is that it cannot work in the infinite variance case where the workload and queue length processes have different orders of magnitude. In contrast, although our approach only works in the case of Poisson arrivals, it makes it possible in the infinite variance case to establish scaling limits of excursions and to uniquely identify accumulation points, see Theorems 4.2.8 and 4.2.9 below for precise statements.

Results of this section are a continuation of those of Section 3.3: in Section 3.3 results concerning binary, homogeneous Crump–Mode–Jagers processes are presented. Here we will argue that an excursion of the Processor-Sharing queue length process is obtained by applying the Lamperti transformation to a binary, homogeneous Crump–Mode–Jagers process and so using the characterization for the convergence of regenerative processes presented in Section 2.3, we will obtain convergence result for the queue length process of the $M/G/1$-PS queue. In particular, we will use the same notation as in Section 3.3, which will be recalled below for clarity.

4.2.2 Main results

Notation

For $\lambda \in \mathbb{R}_+$ and V a positive random variable, the Processor-Sharing queue with parameter (V,λ) denotes the $M/G/1$-PS queue where arrivals occur according to a Poisson process with intensity λ and service requirements are i.i.d. equal in distribution to V. For each $n \in \mathbb{N}$ and $\chi \in \mathscr{M}_p((0,\infty))$, we consider $\mathbb{Q}_n^\chi$ the law of the queue length process of the Processor-Sharing queue with parameter (V_n,λ_n) started with the initial condition χ, i.e., there are initially $|\chi|$ customers in the queue with initial service requirements given by the atoms of χ.

For the most, we adopt the notation of Section 3.3, namely, P_n denotes a Poisson process with intensity λ_n independent from V_n and $\mathscr{P}_n[0,t] = P_n[0,t\wedge V_n]$: $P_n[0,t]$ represents the potential number of children begot up to time t and $\mathscr{P}[0,t]$ the actual number number of children, where we only take into consideration those born before the individual dies at age V_n. We fix a normalizing sequence $s_n \to \infty$ and define $r_n = n/s_n$ as well as:

- $\mathbf{P}_n^x$ the law of $X(nt)/s_n$ under $\mathbb{P}_n^x$, the law of the Lévy process started at $x \in \mathbb{R}$ and with Laplace exponent ψ_n given by
$$\psi_n(u) = u - \lambda_n \mathbb{E}\left(1 - e^{-uV_n}\right), \ u \geq 0;$$
- $\mathbf{Z}_n^\chi$ the law of $X(s_n t)/r_n$ under $\mathbb{Z}_n^\chi$, the law of the binary, homogeneous Crump–Mode–Jagers process with parameter $(V_n,\mathscr{P}_n)$ and initial condition $\chi \in \mathscr{M}_p((0,\infty))$.

For $z \in \mathbb{Z}_+$, we use the notation $\mathbf{Z}_n^z$ and $\mathbf{Z}_n^{z*}$ to denote a random initial condition with z initial customers with i.i.d. life lengths distributed according to V_n and V_n^*, respectively. Further, we make similar assumptions and consider similar cases as in Section 3.3.

Assumptions on $\mathbf{P}_n^0$: for each $n \in \mathbb{N}$, X under $\mathbf{P}_n^0$ does not drift $+\infty$. Moreover, $\mathbf{P}_n^0 \xrightarrow{\text{W}} \mathbf{P}^0$ where $\mathbf{P}^x$ is the law of a Lévy process with infinite variation started at $x \in \mathbb{R}$.

Finite variance case: $s_n = n^{1/2}$ and $\mathbf{P}^x$ is the law of a Brownian motion started at $x \in \mathbb{R}$ and with drift $-\alpha$ for some $\alpha \in (0,\infty)$;

Infinite variance case: there exists $\alpha \in (1,2)$ such that $s_n = n^{1/\alpha}$ and $\mathbb{P}(V_n \geq x) = (1+x)^{-\alpha}$ for every $n \geq 1$ and $x \geq 0$.

In the finite variance case, the fact that X under $\mathbf{P}^0$ drifts to $-\infty$ together with $\mathbf{P}_n^0 \xrightarrow{\text{W}} \mathbf{P}^0$ implies that X under $\mathbf{P}_n^0$ for n large enough drifts to $-\infty$. Since $\mathbb{P}_n^0 \circ R^{-1}$ is the law of the workload associated to the Processor-Sharing queue with parameters (V_n,λ_n) started empty, this implies that the corresponding Processor-Sharing queue is stable, i.e., the measure-valued Markov process describing the Processor-Sharing queue with parameter (V_n,λ_n) is positive Harris recurrent. In the sequel we will denote by $\mathbb{Q}_n^*$ the stationary version of $\mathbb{Q}_n^\chi$. Note that $\mathbb{P}_n^0 \circ R^{-1}$ is also the law of the jumping contour process associated with the binary and homogeneous Crump–Mode–Jagers process with parameter $(V_n, \mathscr{P}_n)$, which was actually the basis for the results in Section 3.3. This double interpretation of $\mathbb{P}_n^0 \circ R^{-1}$ is quite intriguing.

Link between the Processor-Sharing queue and Crump–Mode–Jagers processes

Imagine incoming customers graft a stick whose length is equal to their service requirement to the stick of a customer in service chosen uniformly at random, and at a distance from the origin of this stick equal to the service already received by this customer.

Then one builds in this way a random chronological tree, and a minute thought reveals that it is a Crump–Mode–Jagers tree. Moreover, time is "slowed down" for the Processor-Sharing representation, in that because of the Processor-Sharing discipline, each customer in service ages only at rate $1/q$ if there are q customers in service. Thus if time is instantaneously sped up by q, customers age at rate one, while arrivals now occur at rate λq: this is exactly the description of a binary, homogeneous Crump–Mode–Jagers process. Moreover, the mapping that locally speeds up time by a factor equal to the current state of the system is nothing but the inverse Lamperti transformation $\mathscr{T}^{-1}$. And indeed, we have formally the following result first noted by Kitayev and Yashkov [72].

Proposition 4.2.1. *For any $n \in \mathbb{N}$ and initial condition χ, we have $\mathbb{Q}_n^\chi \circ \sigma^{-1} = \mathbb{Z}_n^\chi \circ \mathscr{T}^{-1}$, i.e., the image of the binary, homogeneous Crump–Mode–Jagers process with parameter $(V_n, \mathscr{P}_n)$ and initial condition χ by the Lamperti transformation is the first busy cycle of the $M/G/1$-PS queue with parameter (V_n, λ_n) and initial condition χ.*

Note that the idea to couple a queueing system and a branching process by creating a genealogy between customers dates back to Kendall [66], and have proved extremely successful, for instance in the study of polling systems [102]. A similar idea underlies the coupling between a model of limited order book and branching random walks laid down in the next chapter.

Since we know that time and space need to be scaled by r_n and s_n, respectively, under $\mathbb{Z}_n^\chi$ to obtain interesting results, this result also tells us how to scale the process under $\mathbb{Q}_n^\chi$. Namely, let $Z = \mathscr{T}^{-1}(X)$, $Z_n(t) = Z(s_n t)/r_n$ and $X_n(t) = \mathscr{T}(Z_n)$: then, with the notation of Definition 3.1.3, we can prove that $\varpi_{Z_n}(t) = \frac{1}{s_n}\varpi_Z(nt)$ and so by definition of the Lamperti transformation,

$$X_n(t) = Z_n \circ \varpi_{Z_n}(t) = \frac{1}{r_n} Z\left(s_n \times \frac{1}{s_n}\varpi_Z(nt)\right) = \frac{1}{r_n}\mathscr{T}(Z)(nt) = \frac{1}{r_n}X(nt).$$

This computation therefore shows that the correct scaling of X under $\mathbb{Q}_n^\chi$ is to scale space by r_n and time by n: if $\mathbf{Q}_n^\chi$ is the law of $X(nt)/r_n$ under $\mathbb{Q}_n^\chi$, then the relation of Proposition 4.2.1 remains for the scaled processes:

$$\mathbf{Q}_n^\chi \circ \sigma^{-1} = \mathbf{Z}_n^\chi \circ \mathscr{T}^{-1}.$$

Similarly as for $\mathbf{Z}$, we use $\mathbf{Q}_n^z$ and $\mathbf{Q}_n^{z*}$ to denote a random initial condition with z initial customers with i.i.d. service requirements distributed according to V_n and V_n^*, respectively. We will finally denote by $\mathbf{Q}_n^*$ the law of $X(nt)/s_n$ under $\mathbb{Q}_n^*$, i.e., $\mathbf{Q}_n^*$ is the stationary version of $\mathbf{Q}_n^\chi$.

We know by Theorems 3.3.8 and 3.3.9 that $\mathbf{Z}_n^\chi$, with suitable conditioning and initial condition, converges weakly: in view of the relation $\mathbf{Q}_n^\chi \circ \sigma^{-1} = \mathbf{Z}_n^\chi \circ \mathscr{T}^{-1}$, transferring this result to excursions of the Processor-Sharing queue length process is just a matter of continuity properties of the Lamperti transformation and its inverse. Note that properties similar to the following ones are proved in Helland [52], but only for excursions started away from 0. In [LSZ13] we have extended these results to allow for excursions starting at 0. Let $\mathscr{E}'(\mathbb{R}) \subset \mathscr{E}(\mathbb{R})$ be the subset of excursions $e \in \mathscr{E}(\mathbb{R})$ such that $T_0(e)$ and $\int_0^{T_0(e)} \mathrm{d}u/e(u)$ are finite.

Lemma 4.2.2. *Let P_n, P be probability measures on $\mathscr{D}(\mathbb{R})$ such that (P_n) is C-tight, $(P_n \circ T_0^{-1})$ is tight and $P_n(\mathscr{E}(\mathbb{R})) = P(\mathscr{E}(\mathbb{R})) = 1$.*

If $P_n \xrightarrow{\text{w}} P$ then $P_n \circ \mathscr{T}^{-1} \xrightarrow{\text{w}} P \circ \mathscr{T}^{-1}$.

If $P_n(\mathscr{E}'(\mathbb{R})) = 1$, then the sequence $(P_n \circ \mathscr{T})$ is C-tight.

If $P_n \xrightarrow{\text{w}} P$ and $P_n(\mathscr{E}'(\mathbb{R})) = P(\mathscr{E}'(\mathbb{R})) = 1$, then $P_n \circ \mathscr{T} \xrightarrow{\text{w}} P \circ \mathscr{T}$.

Finite variance case

Although tightness is usually a technical issue, in the finite variance case it comes from a simple queueing argument. Theorem 4.2.3 below may look naive to an experienced reader, but to the best of our knowledge its implications in terms of tightness have not been used before: similar arguments could for instance have been used in Limic [85] for the LIFO queue.

Theorem 4.2.3. *Under $\mathbb{Q}_n^*$, the departure process of the queue length process is a Poisson process with parameter λ_n.*

Indeed, Kelly [64] proved that the departure process of any quasi-reversible queue is Poisson (Theorem 3.6 in [64]), and that the Processor-Sharing service discipline, being a symmetric service discipline, is quasi-reversible (Theorem 3.10 in [64]). To see the implication of this result in terms of tightness, write X in the form

$$X(t) = X(0) + A(t) - D(t)$$

with A and D the arrival and departure processes, respectively. Under $\mathbb{Q}_n^*$, $X(0)$ is a geometric random variable with parameter $1 - \lambda_n \mathbb{E}(V_n)$, A under $\mathbb{Q}_n^*$ is a Poisson process with parameter λ_n and so is D according to Theorem 4.2.3. Scaling in time and space and compensating the Poisson processes, we get the expression

$$\frac{1}{n^{1/2}} X(nt) = \frac{1}{n^{1/2}} X(0) + A_n(t) - D_n(t)$$

with $A_n(t) = n^{-1/2}(A(nt) - nt)$ and $D_n(t) = n^{-1/2}(D(nt) - nt)$. Classical results imply that A_n and D_n under $\mathbb{Q}_n^*$ converge in distribution to a Brownian motion. Note that A_n and D_n are not independent, nonetheless their convergence toward a continuous process implies that their difference is tight. Since in the finite variance case we have $n^{1/2}(1 - \lambda_n \mathbb{E}(V_n)) \to -\alpha$ (the convergence of the drift of X under $\mathbf{Q}_n^0$ converges to the drift of X under $\mathbf{P}_n^0$), we obtain that $n^{-1/2} X(0)$ under $\mathbf{Q}_n^*$ converges in distribution to an exponential random variable, which implies the tightness of $\mathbf{Q}_n^*$, and even its C-tightness since jumps of X under $\mathbf{Q}_n^*$ are of amplitude $1/r_n \to 0$. Note that the assumption $\alpha > 0$, i.e., X under $\mathbf{P}^0$ drifts to $-\infty$, is essential for this proof to hold.

Corollary 4.2.4. *In the finite variance case, the sequence $\mathbf{Q}_n^*$ is C-tight.*

Although this result seems quite peculiar, it actually encompasses much more tightness results thanks to the following continuity properties of the shift and stopping operators.

Lemma 4.2.5. *If $f_n, f \in \mathscr{D}(\mathbb{R})$ and $t_n, t \in \mathbb{R}_+$ are such that $f_n \xrightarrow{J_1} f$, $t_n \to t$ and $\Delta f_n(t_n) \to \Delta f(t)$, then $\theta_{t_n} f_n \to \theta_t f$ and $\sigma_{t_n} f_n \to \sigma_t f$.*

This result has the following probabilistic consequence: if P_n is C-tight and $P_n \circ \tau_n^{-1}$ is tight, where the τ_n's are arbitrary random times, then $P_n \circ \theta_{\tau_n}^{-1}$ and $P_n \circ \sigma_{\tau_n}^{-1}$ are C-tight. This result makes it possible to isolate pieces of interest in the sample path of X.

A fundamental observation at this point is that since $\mathbb{P}_n^0 \circ R^{-1}$ is the law of the workload process associated to $\mathbb{Q}_n^0$ and since the zero sets of the workload and queue length processes coincide, any question on the endpoints of excursions of X under $\mathbb{Q}_n^0$ selected on their length is actually a question on endpoints of excursions of X under $\mathbb{P}_n^0 \circ R^{-1}$ selected on the same criterion. The asymptotic behavior as $n \to \infty$ of the zero set of X under $\mathbb{P}_n^0 \circ R^{-1}$ is well known issue and in particular, we will take for granted the results pertained to it which we need along the way.

For instance, one easily gets that $\mathbf{Q}_n^* \circ T_0^{-1}$ is tight, which implies that $\mathbf{Q}_n^* \circ \sigma^{-1}$ is tight by the previous lemma. Leveraging the occupation density formula, we get as explained in the discussion in Section 3.3.1 that $\mathbf{Q}_n^* \circ \sigma^{-1}$ converges weakly to the law of $\mathscr{T}(L(\cdot, T^L(\zeta)))$ under $\mathbf{P}^0(\cdot \mid T^L(\zeta) < \infty)$ for ζ some suitable random variable independent from X.

Likewise, we get convergence of excursions suitably conditioned. First of all, the tightness of $\mathbf{Q}_n^* \circ T_0^{-1}$ implies the tightness of $\mathbf{Q}_n^* \circ \theta^{-1} = \mathbf{Q}_n^0$. We now use the notation of Section 2.3 with $\varphi = T_0$, so that e_ε is the first excursion with length $> \varepsilon$ and g_ε is its left endpoint. Because the zero sets of the workload and queue length processes coincide, we get that $\mathbf{Q}_n^0 \circ g_\varepsilon^{-1}$ and $\mathbf{Q}_n^0 \circ (T_0 \circ e_\varepsilon)^{-1}$ are tight, and so since $e_\varepsilon = \sigma \circ \theta_{g_\varepsilon}$ we get the tightness of $\mathbf{Q}_n^0 \circ e_\varepsilon^{-1}$, i.e., the first excursion of the Processor-Sharing queue length process with length $> \varepsilon$. Again, the occupation density formula implies that there is actually weak convergence, and more precisely that $\mathbf{Q}_n^0 \circ e_\varepsilon^{-1}$ converges weakly to the law of $\mathscr{T} \circ L^0$ under $\underline{\mathscr{N}}(\cdot \mid T_0 > \varepsilon)$.

Now that we have convergence of excursions, their length and their endpoint, and also the tightness of the whole process, we can invoke Theorem 2.3.1 to get convergence of the whole process.

Theorem 4.2.6. *Let $z_n \in \mathbb{Z}_+$ such that $z_n / n \to \zeta \in \mathbb{R}_+$. Then in the finite variance case, $\mathbf{Q}_n^{z_n *} \xrightarrow{\text{w}} \mathbf{Q}^\zeta$ which is the only distribution such that:*

- *$\mathbf{Q}^\zeta$ is regenerative with excursion measure $\underline{\mathscr{N}} \circ (\mathscr{T} \circ L^0)^{-1}$;*
- *$\mathscr{Z}(X)$ under $\mathbf{Q}^\zeta \circ \theta^{-1}$ has almost surely zero Lebesgue measure;*
- *$\mathbf{Q}^\zeta \circ \sigma^{-1}$ is the law of $\mathscr{T}(L(\cdot, T^L(\zeta)))$ under $\mathbf{P}^0(\cdot \mid T^L(\zeta) < \infty)$.*

A corollary to Gromoll's results in [45] is that if $\sup_n \mathbb{E}(V_n^{4+\varepsilon}) < \infty$, then $\mathbf{Q}_n^0$ converges weakly to the law of a reflected Brownian motion: fortunately, we recover this result.

Theorem 4.2.7. *If $\mathbf{P}^x$ is the law of a Brownian motion started at $x \in \mathbb{R}$ and which does not drift to $+\infty$, then $\underline{\mathscr{N}} \circ (\mathscr{T} \circ L^0)^{-1}$ is the excursion measure of a reflected Brownian motion.*

For the identification of the coefficients of this Brownian motion, see [LSZ13]. As mentioned in Section 3.3, this result follows from a variation of the original Ray–Knight theorems [73, 96], which state that the local time process of a Brownian excursion started away from 0 is Feller diffusion, combined with the fact that the Lamperti transformation of Feller diffusion is again Brownian motion. In particular, it is reasonable to believe that this result also holds if X under $\mathbf{P}^x$ is a Brownian motion with positive drift, although we have not tried to prove this.

Note also that the above reasoning fills in the missing gap in the proof of Theorem 3.3.4 and indeed shows that L^0 under $\mathbf{P}_n(\cdot \mid T_0 > \varepsilon)$ is tight: indeed, on the one hand we have

$$\mathbf{P}_n(\cdot \mid T_0 > \varepsilon) \circ (L^0)^{-1} = \left[\mathbf{P}_n \circ (L^0)^{-1}\right](\cdot \mid T_0 \circ L^0 > \varepsilon) = \mathbf{Z}^1_n\left(\cdot \mid \int X > \varepsilon\right) = \mathbf{Z}^1_n(\cdot \mid T_0 \circ \mathscr{T} > \varepsilon)$$

where we have used the identity $T_0 \circ L^0 = \int X$, while on the other hand,

$$\left[\mathbf{Q}^1_n \circ \sigma^{-1}\right](\cdot \mid T_0 > \varepsilon) = \left[\mathbf{Z}^1_n \circ \mathscr{T}^{-1}\right](\cdot \mid T_0 > \varepsilon) = \left[\mathbf{Z}^1_n(\cdot \mid T_0 \circ \mathscr{T} > \varepsilon)\right] \circ \mathscr{T}^{-1} \qquad (4.1)$$

which shows that

$$\mathbf{Z}^1_n(\cdot \mid T_0 \circ \mathscr{T} > \varepsilon) = \left[\mathbf{Q}^0_n \circ (\theta_{g_\varepsilon} \circ \sigma)^{-1}\right] \circ \mathscr{T}.$$

Since $\mathbf{Q}^0_n \circ (\theta_{g_\varepsilon} \circ \sigma)^{-1} = \mathbf{Q}^0_n \circ e_\varepsilon^{-1}$ has been argued to be tight, continuity properties of $\mathscr{T}$ imply the tightness of $\mathbf{Z}^1_n(\cdot \mid T_0 \circ \mathscr{T} > \varepsilon)$ as desired.

Infinite variance case

In the infinite variance, the above arguments go through almost step by step, with the major difference that the arguments leading to Corollary 4.2.4 do not work anymore: the departure process is still Poisson, but compensating the Poisson processes now leads to processes that blow up because of the scaling in space by $r_n = n^{1-1/\alpha}$ with $\alpha \in (1,2)$: then $r_n \ll n^{1/2}$, where $n^{1/2}$ is the correct scaling of the compensated Poisson processes. In the infinite variance case, X under $\mathbf{Q}^*_n$ is the difference of two processes, each of which blows up, but they are correlated in such a way that their jumps compensate each other.

If we cannot use the same arguments to get tightness, we can still transfer results from Crump–Mode–Jagers processes thanks to Lemma 4.2.2. Indeed, according to Theorem 3.3.8, we have $\mathbf{Z}^1_n(\cdot \mid T_0 > \varepsilon) \xrightarrow{w} \underline{\mathscr{L}}(\cdot \mid T_0 > \varepsilon)$ in the infinite variance case. Using Lemma 2.3.10, we can change the conditioning and show instead that $\mathbf{Z}^1_n(\cdot \mid T_0 \circ \mathscr{T} > \varepsilon) \xrightarrow{w} \underline{\mathscr{L}}(\cdot \mid T_0 \circ \mathscr{T} > \varepsilon)$. Since $\left[\mathbf{Q}^1_n \circ \sigma^{-1}\right](\cdot \mid T_0 > \varepsilon) = \left[\mathbf{Z}^1_n(\cdot \mid T_0 \circ \mathscr{T} > \varepsilon)\right] \circ \mathscr{T}^{-1}$ by (4.1), we can finally use the continuity properties of $\mathscr{T}^{-1}$ of Lemma 4.2.2 to get that the first excursion of $\mathbf{Q}^0_n$ with length $\varepsilon > 0$ converges weakly to $[\underline{\mathscr{L}}(\cdot \mid T_0 > \varepsilon)] \circ \mathscr{T}^{-1}$.

Theorem 4.2.8. *Let $\varepsilon > 0$. In the infinite variance case, we have*

$$\left[\mathbf{Q}^1_n \circ \sigma^{-1}\right](\cdot \mid T_0 > \varepsilon) \xrightarrow{w} \left[\underline{\mathscr{L}} \circ \mathscr{T}^{-1}\right](\cdot \mid T_0 > \varepsilon),$$

Because the zero set of X under $\mathbf{Q}^0_n$ and $\mathbf{P}^0_n \circ R^{-1}$ have the same law, we still have the control on $\mathbf{Q}^0_n \circ g_\varepsilon^{-1}$ and $\mathbf{Q}^0_n \circ (T_0 \circ g_\varepsilon)^{-1}$ required to apply Theorem 2.3.1. The difference with the finite variance case is that we do not know whether $\mathbf{Q}^0_n$ is tight. Nonetheless, except for (2.5) all the assumptions of Theorem 2.3.7 can be shown to be satisfied, and so we have the following result.

Theorem 4.2.9. *Consider the infinite variance case, and assume that at least one of the following two conditions is met:*

- *the sequence* $(\mathbf{Q}_n^0)$ *is tight;*
- *for any* $\eta > 0$,

$$\lim_{\varepsilon\to 0}\limsup_{n\to\infty}\left[s_n\mathbf{Q}_n^1\left(\sup\sigma \geq \eta, T_0 \leq \varepsilon\right)\right] = 0. \tag{4.2}$$

Then the conclusions of Theorem 4.2.6 hold.

In words, the condition (4.2) reduces the problem of convergence of the scaled Processor-Sharing queue length process $\mathbf{Q}^{z_n*}$ to the problem of showing that the height of “short” excursions is small. Besides, a major difference with the finite variance case is that now, it is not clear whether the process with excursion measure $\underline{\mathcal{N}} \circ (\mathcal{T} \circ L^0)^{-1}$ has any simpler probabilistic description. However, there is one intriguing fact we can record about this process, namely that its one-dimensional distributions coincide with those of the height process associated to $\mathbf{P}^0$.

More precisely, it is shown in Kella et al. [63] that the one-dimensional distributions of the Processor-Sharing and Last-In-First-Out (LIFO) queue started empty are equal, while Limic [85] argued that the queue length process of the LIFO queue converges to the height process associated to $\mathbf{P}^0$ (see Duquesne and Le Gall [31]). Taken together with the above convergence result, this suggests that the one-dimensional distributions of the regenerative process with excursion measure $\underline{\mathcal{N}} \circ (\mathcal{T} \circ L^0)^{-1}$ coincide with those of the height process associated to $\mathbf{P}^0$.

However, these processes can be dramatically different. For instance, it follows from Lemma 3.10 in Duquesne and Le Gall [31] that if the Laplace exponent Ψ of $\mathbf{P}^0$ is such that

$$\int^{\infty} \frac{\mathrm{d}\lambda}{\Psi(\lambda)\sqrt{\log\lambda}} < \infty \text{ while } \int^{\infty} \frac{\mathrm{d}\lambda}{\Psi(\lambda)} = \infty$$

then the height process is not càdlàg while the process with excursion measure $\underline{\mathcal{N}} \circ (\mathcal{T} \circ L^0)^{-1}$ is continuous.

To conclude this presentation of results, we state in view of Lemma 3.3.3 a natural conjecture concerning scaling limits of the Processor-Sharing queue length process.

Conjecture. *If condition* (3.10) *holds, i.e., if*

$$\int^{\infty} \frac{\mathrm{d}\lambda}{\Psi(\lambda)\sqrt{\log\lambda}} < \infty$$

where Ψ *is the Laplace exponent of* $\mathbf{P}^0$, *then the conclusions of Theorem 4.2.6 hold.*

4.2.3 Discussion

Let us go back to the discussion “Discontinuity at 0” of Section 3.3.2. There, we explained that in the finite variance case, $\mathbf{Z}_n^{z_n}$ and $\mathbf{Z}_n^{z_n*}$ restricted to $(0,\infty)$ both converge to Feller diffusion with the same drift and Gaussian coefficients, but with a different initial condition. Namely, although $X(0)$ under $\mathbf{Z}_n^{z_n}$ converges in distribution to ζ, $\mathbf{Z}_n^{z_n}$ restricted to $(0,\infty)$ converges to Feller diffusion started at $\zeta' \neq \zeta$. This phenomenon, unusual in the branching literature, has a simple queueing interpretation in terms of the well-known state-space collapse phenomenon.

First of all, by applying the Lamperti transformation we obtain the same behavior for the Processor-Sharing queue: $\mathbf{Q}_n^{z_n}$ and $\mathbf{Q}_n^{z_n *}$ restricted to $(0,\infty)$ both converge to reflected Brownian motion with the same drift and Gaussian coefficients, but with a different initial condition. Namely, the scaling limit of $\mathbf{Q}_n^{z_n}$ does not take the value ζ at time 0+. To understand what happens, consider, under $\mathbb{Q}_n^{\chi}$ for $\chi \in \mathcal{M}_p((0,\infty))$, $(Y(t), t \geq 0)$ the workload process and define, for $n \in \mathbb{N}$ and $t \in \mathbb{R}_+$, the fluid scaled processes

$$\overline{X}_n(t) = \frac{1}{n^{1/2}} X\left(n^{1/2} t\right) \text{ and } \overline{Y}_n(t) = \frac{1}{n^{1/2}} Y\left(n^{1/2} t\right)$$

and the diffusion scaled processes

$$\widehat{X}_n(t) = \overline{X}_n\left(n^{1/2} t\right) = \frac{1}{n^{1/2}} X(nt) \text{ and } \widehat{Y}_n(t) = \overline{Y}_n\left(n^{1/2} t\right) = \frac{1}{n^{1/2}} Y(nt).$$

Note that $\mathbb{Q}_n^{z*} \circ \widehat{X}_n^{-1} = \mathbf{Q}_n^{z*}$ and $\mathbb{Q}_n^{z*} \circ \widehat{Y}_n^{-1} = \mathbf{P}_n^{z*} \circ R^{-1}$, so that $(\widehat{X}_n, \widehat{Y}_n)$ under $\mathbb{Q}_n^{z_n *}$ is tight and any accumulation point is a pair of reflected Brownian motions. Note that $\widehat{Y}_n(t)$ can be expressed as the sum over the customers in the queue of their residual service time, which in stationarity are i.i.d. with common distribution V_n^*: the law of large numbers thus suggests that in the limit, $\widehat{Y}_n$ will be proportional to $\widehat{X}_n$ with proportionality coefficient the limit of $\mathbb{E}(V_n^*)$, say β^*. This property is exactly the state-space collapse property, which states that $(\widehat{X}_n, \widehat{Y}_n) \xrightarrow{\text{d}} (R(W), \beta^* R(W))$ for some suitable Brownian motion W.

To understand the aforementioned discontinuity at 0 when considering $\widehat{X}_n$ under $\mathbb{Q}_n^{z_n}$, let us zoom in around time 0 and consider the fluid scaled processes $\overline{X}_n$ and $\overline{Y}_n$. Let $\beta = \lim_n \mathbb{E}(V_n)$: by definition, at the critical point the amount of work that enters the queue is equal to the amount of work that exits it. Hence by the law of large numbers, it is not surprising that $\overline{Y}_n$ under $\mathbb{Q}_n^{z_n}$ converges in distribution to w, the deterministic function with constant value $w(0) = \zeta\beta$. Note that the workload process does not fluctuate on the fluid time scale $n^{1/2}$, while it does on the diffusion time scale n.

Let on the other hand q be the limit of $\overline{X}_n$ under $\mathbb{Q}_n^{z_n}$, so that $q(0) = \zeta$ (see Gromoll et al. [47]). It is known that as t goes to infinity, $q(t)$ converges to an equilibrium point q_∞. In steady state the residual service requirement of each customer has mean β^*, which suggests thanks to the law of large numbers that q_∞ must satisfy $q_\infty \beta^* = w(0)$.

So it takes a time of order of $n^{1/2}$ for the (scaled) queue length process to go from ζ to $q_\infty = w(0)/\beta^* = \zeta\beta/\beta^*$. Since the time scale n of the heavy traffic approximation is orders of magnitude larger, this happens instantaneously on the diffusion time scale and causes a discontinuity when $\beta \neq \beta^*$.

Once the process has reached the equilibrium point of the fluid limit, the state space collapse property applies. In particular, this shows that $\widehat{X}_n$ under $\mathbb{Q}_n^{z_n}$ should converge in distribution to a process Q such that $Q(0) = \zeta$ and $Q(t) = (1/\beta^*) R(Y)(t)$ for $t > 0$. In particular, $Q(0+) = \zeta\beta/\beta^*$ is different from $Q(0)$ when $\beta \neq \beta^*$, which provides yet another interpretation of the discontinuity of local times mentioned above. This separation of fluid and diffusion time scales is at the heart of state space collapse, see for instance Bramson [22].

4.3 A stochastic network with mobile customers

This section presents the results of [BS13, ST10] on the study of a stochastic network with mobile customers.

4.3.1 Model

Introduction

Recent wireless technologies have triggered interest in a new class of stochastic networks, called *mobile networks.* In contrast with Jackson networks where customers only move upon completion of service at some node, in these mobile networks, transitions of customers within the network occur *independently* of the service received: customers move and receive some service where they are.

Moreover, at any given time, each node capacity is divided between the customers present, whose service rate thus depends on the capacity and on the state of occupancy of the node. Once his initial service requirement has been fulfilled, a customer definitively leaves the network. In [19], elaborate capacity sharing policies are considered but in the simplest setting, which will be of interest to us, nodes implement the Processor-Sharing discipline by dividing their capacity equally between all the customers present. Previous works [19, 48] have mainly focused on determining the stability region of such networks, and it has been commonly observed that the customers' mobility represents an opportunity for the network to increase this region. Indeed, because of their mobility, customers offer a diversity of channel conditions to the base stations (in charge of allocating the resources of the nodes), thus making it possible to select the customers in the most favorable state. Such a scheduling strategy is sometimes referred to as an *opportunistic scheduling* strategy, see [18] and the references therein for more details. The results presented next show that mobility entails an optimal use of the networks resource.

Model

In this section, we are interested in the $\mathbb{Z}_+^K$-valued Markov process, for $K \in \mathbb{N}$, whose dynamic is described as follows:

- customers arrive at node $k = 1, \dots, K$ according to a Poisson process with intensity λ_k, and arrival streams are independent;
- customers have i.i.d. service requirements, exponentially distributed with parameter one, independent of the arrival processes;
- node $k = 1, \dots, K$ serves customers at rate μ_k according to the Processor-Sharing service discipline;
- while still in service, customers move independently of everything else according to a Markov process with Q-matrix $Q = (q_{k\ell}, k, \ell = 1, \dots, K)$ and leave the network when their service requirement has been completed.

We stress that this model is in sharp contrast with classical queueing networks, such as Jackson networks, where customers only move upon completion of service. Here, customers arrive with a single service requirement and they are served *along their route*: customers receive some service where they are and their trajectory is governed by some random dynamics independent of the service. In the model we consider, the trajectory process ξ appearing in (4.5) is a Markov process with Q-

matrix Q, independent of the arrival and potential departure processes.

Because service requirements are exponentially distributed, the memoryless property of the exponential distribution implies that the process keeping track of the number of customers in every node over time is a Markov process with values in $\mathbb{Z}_K^+$ and with generator Ω given, for $f : \mathbb{R}^K \to \mathbb{R}$ and $x \in \mathbb{Z}_+^K$, by

$$\begin{aligned}\Omega(f)(x) = \sum_{k=1}^K \lambda_k \left[f(x+e_k) - f(x)\right] + \sum_{k=1}^K \mathbb{1}(x_k > 0)\mu_k \left[f(x-e_k) - f(x)\right]&\\ + \sum_{1\le k,\ell\le K} q_{k\ell} x_k \left[f(x-e_k+e_\ell) - f(x)\right]& \quad (4.3)\end{aligned}$$

with $e_k \in \mathbb{Z}_+^K$ denoting the kth unit vector, with 0's everywhere except at the kth coordinate where there is a one. Note that the Processor-Sharing assumption has no impact on this generator, which would remain the same for other service disciplines. However, this assumption will have an impact on sojourn time distributions, as can be guessed from the forthcoming formula (4.5), specific to the Processor-Sharing service discipline.

We will denote by $\mathbb{P}^x$ the law of this Markov process started at $x \in \mathbb{Z}_+^K$, by A_k the arrival process to node k and by D_k the potential departure process from node k. Thus under $\mathbb{P}^x$ the $2K$ processes $(A_k, D_k, k = 1, \dots, K)$ are independent and A_k, resp. D_k, is a Poisson process with intensity λ_k, resp. μ_k. The special case where all arrival and service rates are equal to 0, i.e., $\lambda_k = \mu_k = 0$, will be denoted by $\mathbb{P}_c^x$ and will be referred to as the *closed system.* Thus under $\mathbb{P}_c^x$, X simply records the positions of $\|x\|_1$ independent walkers moving according to Q.

In order to ease the notation, when no confusion can arise we will use bold typeset to denote the L_1 norm of a vector or function, e.g., $\mathbf{x} = \|x\|_1$ and $\mathbf{f} = \|f\|_1$. With this notation, $\mathbf{X}(t)$ is the total number of customers in the system at time t and $\mathbf{A}$ is the total arrival process. In particular, $\mathbf{X}$, $\mathbf{A}$, the X_k's and the D_k's are linked by the following relation:

$$\mathbf{X}(t) = \mathbf{X}(0) + \mathbf{A}(t) - \sum_{k=1}^K \int_0^t \mathbb{1}(X_k(s-) > 0) D_k(\mathrm{d}s).$$

Note that, since $\mathbb{1}(X_k(s-) > 0) \le \mathbb{1}(\mathbf{X}(s-) > 0)$, we obtain

$$\mathbf{X}(t) \ge \mathbf{X}(0) + \mathbf{A}(t) - \int_0^t \mathbb{1}(\mathbf{X}(s-) > 0)\mathbf{D}(\mathrm{d}s)$$

from which it follows by induction that

$$\mathbf{X}(t) \ge L(t) \text{ where } L(t) = \mathbf{X}(0) + \mathbf{A}(t) - \int_0^t \mathbb{1}(L(s-) > 0)\mathbf{D}(\mathrm{d}s). \quad (4.4)$$

Note that L under $\mathbb{P}^x$ is an $M/M/1$ queue started at $\mathbf{x}$ with input rate $\boldsymbol{\lambda}$ and output rate $\boldsymbol{\mu}$.

According to the Processor-Sharing service discipline, the server splits its service capacity equally among all the customers present at any point in time, and so each customer present at node k at time t is instantaneously served at rate $\mu_k / X_k(t)$. In particular, if a customer has been present in the network between time s and t and was at node $\xi(u)$ at time $u \in [s, t]$, then between $[s, t]$ it received equal to

$$\int_s^t \frac{\mu_{\xi(s)}}{X_{\xi(s)}(s)} \mathrm{d}s. \quad (4.5)$$

4.3.2 Homogenization property and implications

Throughout this section, we assume that Q is irreducible, which ensures the existence of a unique stationary distribution π satisfying $\pi Q = 0$. Assume that the networks is initially overloaded, i.e., $\mathbf{X}(0) \gg 1$, and consider the network at a fixed time t. Since the arrival and departure rates are bounded, by this time the total number of customers will have varied by at most a constant amount, of the order of t, so that $\mathbf{X}(t) \approx \mathbf{X}(0)$. Furthermore, if (loosely speaking) τ denotes the mixing time of Q, then at time τ every customer initially present and still present at time τ will be approximately distributed according to π in the network. Combining these two observations, the strong law of large numbers suggests the approximation

$$X(t) \approx \mathbf{X}(t) \times \pi, \ t \geq \tau. \tag{4.6}$$

This property will be called *homogenization property* and states that when the network is overloaded, customers get spread in the network according to π. This is a spatial form of state space collapse induced by mobility.

Rigorously, we can formalize the above discussion as follows. Let in the sequel

$$\tilde{X}(t) = \frac{X(t)}{\mathbf{X}(t)}, \ \mathbf{X}(t) > 0,$$

with the convention $\tilde{X}(t) = \pi$ when $\mathbf{X}(t) = 0$. If ξ is a Markov process with Q-matrix Q, let moreover

$$\tau(\varepsilon) = \sup\left\{ t \geq 0 : \max_{1 \leq k, \ell \leq K} |\mathbb{P}(\xi(t) = \ell \mid \xi(0) = k) - \pi_\ell| \geq \varepsilon \right\}.$$

Let finally Pois(z) denote a Poisson random variable with parameter $z \in \mathbb{R}_+$.

Lemma 4.3.1. *There exists $\delta_0 > 0$, depending on π and K, such that for any $\delta < \delta_0$, $t \geq \tau(\delta/(4K))$ and $x \in \mathbb{Z}_+^K$, we have*

$$\mathbb{P}^x\left(\left\|\tilde{X}(t) - \pi\right\|_1 \geq \delta\right) \leq \mathbb{P}\left(\mathrm{Pois}((\boldsymbol{\lambda} + \boldsymbol{\mu})\mathrm{t}) \geq \delta \mathbf{x}/(4\mathrm{K})\right) + 2K \exp\left(-\frac{\delta^2 \mathbf{x}}{16K^2}\right).$$

In the above upper bound, the first term controls the number of arrivals and departures up to time t, and the second term controls the probability of the event $\|\tilde{X}(t) - \pi\|_1 \geq \delta$ for the closed system and follows from Chernoff's inequality.

Pushing the intuition behind the homogenization property further, we see that the relation $X(t) \approx \mathbf{X}(t) \times \pi$ should, or at least could, hold not only at a fixed time but also for a long time interval, in the most optimistic case as long as $\mathbf{X}(t) \gg 1$. This is of course a subtler property than just proving homogenization at a fixed time, and indeed the proof of the following result requires more care: some ideas for the proof will be given next. In the sequel we define $\gamma = -\mathrm{Tr(Q)}$ and $\pi_\wedge = \min_k \pi_k$.

Proposition 4.3.2. *Assume that the constants $t > 0$, $0 < \delta < 1$, $\phi \in \mathbb{N}$ and $x \in \mathbb{Z}_+^K$ satisfy*

$$\eta < \delta_0, \quad \frac{\phi\eta}{8K(\boldsymbol{\lambda} + \boldsymbol{\mu})} > \max\left(\tau\left(\eta/(4K)\right), \frac{1}{\gamma}\right), \quad \mathbf{x} > \phi \quad \textit{and} \quad \left\|\frac{x}{\mathbf{x}} - \pi\right\|_1 \leq \eta,$$

where $\eta = \delta^2 \pi_\wedge / 32$. Then it holds that

$$\mathbb{P}^x \left(T^\uparrow(\tilde{X} - \pi, \delta) \leq t \wedge T^\downarrow(X, \phi) \right) \leq c_2(\delta) \exp\left(\log t - c_1(\delta^4 \phi - \log \phi)\right),$$

where in the above bound, c_1 does not depend on t, δ, ϕ and x but only on K, λ, μ and Q; and $c_2(\delta)$ does not depend t, ϕ and x but only on K, λ, μ, Q and δ.

Informally, this bound essentially states that if the process starts homogenized (condition $\|x/\mathbf{x} - \pi\|_1 \leq \eta$), then it stays homogenized as long as there are many customers in the system. Let us now investigate the consequence of the homogenization property.

Implications for the long-time behavior of the process

First of all, the bound of Lemma 4.3.1 can be leveraged to prove stability (i.e., positive recurrence) whenever $\boldsymbol{\lambda} < \boldsymbol{\mu}$. Intuitively, this result is clear: if there are many customers in the network, then the probability of having an empty node is small and so the drift of $\mathbf{X}$ should be approximately $\boldsymbol{\lambda} - \boldsymbol{\mu} < 0$. More formally, we have

$$\mathbb{E}^x\left(\mathbf{X}(t) - \mathbf{X}(0)\right) = (\boldsymbol{\lambda} - \boldsymbol{\mu}) t + \sum_{k=1}^K \mu_k \int_0^t \mathbb{P}^x \left(X_k(s) = 0\right) \mathrm{d}s.$$

But since $X_k(s) = 0$ implies that $\|\tilde{X}(s) - \pi\|_1 \geq \pi_\wedge$, we get

$$\mathbb{E}^x\left(\mathbf{X}(t) - \mathbf{X}(0)\right) \leq (\boldsymbol{\lambda} - \boldsymbol{\mu}) t + \boldsymbol{\mu} \int_0^t \mathbb{P}^x \left(\left\| \tilde{X}(s) - \pi \right\|_1 \geq \pi_\wedge \right) \mathrm{d}s.$$

Together with Lemma 4.3.1 and Foster's criterion, we can deduce thanks to this bound the positive recurrence of $(\mathbb{P}^x)$ when $\boldsymbol{\lambda} < \boldsymbol{\mu}$. Since on the other hand $\mathbf{X} \geq L$ (by (4.4)) and L is transient when $\boldsymbol{\lambda} > \boldsymbol{\mu}$ (recall that L under $\mathbb{P}^x$ is an $M/M/1$ queue with input rate $\boldsymbol{\lambda}$ and output rate $\boldsymbol{\mu}$), we obtain the following result.

Proposition 4.3.3. *If $\boldsymbol{\lambda} < \boldsymbol{\mu}$, then $(\mathbb{P}^x)$ is positive recurrent. If $\boldsymbol{\lambda} > \boldsymbol{\mu}$, then $(\mathbb{P}^x)$ is transient.*

The case $\boldsymbol{\lambda} = \boldsymbol{\mu}$ is subtler. Since L is null-recurrent in this case, the bound $\mathbf{X} \geq L$ only entails that X cannot be positive recurrent. In contrast to a critical $M/M/1$ queue whose drift is $= 0$ away from 0, the drift of $\mathbf{X}$ is always strictly positive due to the fact that there is always a positive probability that a potential departure finds an empty node, creating a slack between the arrival and departure rates. However, the drift should go to 0 as the size of the initial state becomes large (because this probability becomes smaller and smaller) and a result due to Lamperti [80] asserts that if the drift goes sufficiently fast to 0, then the process is actually (null) recurrent. Because of the bound of Lemma 4.3.1, this will the case for us and so we get the following result which completes the picture.

Proposition 4.3.4. *If $\boldsymbol{\lambda} = \boldsymbol{\mu}$, then $(\mathbb{P}^x)$ is null recurrent.*

In the transient case, we can moreover give a precise description on the way the process goes to ∞.

Theorem 4.3.5. *If $\boldsymbol{\lambda} > \boldsymbol{\mu}$, then under $\mathbb{P}^x$ for any $x \in \mathbb{Z}_+^K$ we have as $t \to \infty$*

$$\frac{X(t)}{t} \xrightarrow{\text{a.s.}} (\boldsymbol{\lambda} - \boldsymbol{\mu}) \times \pi.$$

This result has a simple intuitive explanation, which actually lays down a sketch of proof based on the Borel–Cantelli lemma. Namely, the bound of Proposition 4.3.2 asserts that the process stays homogenized as long as there are many customers in the system. More precisely, as $n \to \infty$ the upper bound vanishes for, say, $\phi = n^{1/2}$ and $t = e^n$ which essentially implies, when starting with n customers in the system spread according to π, that the process will remain homogenized for at least e^n units of time. But since the process is transient, by this time there will be e^n customers in the system, and so by the strong Markov property the process will stay homogenized at least for e^{e^n} additional units of time, at which point there will be e^{e^n} customers in the system, etc. The process is thus bound to a reinforcing effect such that the homogenization property actually holds on an *infinite* time-horizon.

Functional convergence on the constant time scale

We now turn our attention to implications of the homogenization property in terms of functional convergence. As a warm-up, consider the process on the constant time scale. Under $\mathbb{P}^x_c$, $\tilde{X}$ is the empirical measure of a system of $\mathbf{x}$ i.i.d. particles moving according to Q and with initial distribution $x/\mathbf{x}$. Since for $\tilde{x} \in [0,1]^K$ with $\tilde{\mathbf{x}} = 1$, $\tilde{x}e^{tQ}$ is the law at time t of the position of a single particle with initial location distributed according to $\tilde{x}$, we get that, if $x_n/\mathbf{x}_n \to \tilde{x}$, then $\tilde{X}(t)$ under $\mathbb{P}^{x_n}_c$ converges in distribution to $\tilde{x}e^{tQ}$. It is not difficult to extend this at a functional level, and also to allow for arrivals and departures (negligible on the constant time scale) to get the following result.

Lemma 4.3.6. *Consider a sequence of initial states (x_n) such that $\mathbf{x}_n \to \infty$ and $x_n/\mathbf{x}_n \to \tilde{x}$: then the processes $\tilde{X}$ and $X/\mathbf{x}_n$ under $\mathbb{P}^{x_n}$ have the same asymptotic behavior, namely they both converge in distribution to $(\tilde{x}e^{tQ}, t \geq 0)$.*

Fluid limits

The previous result describes the behavior of the process on the constant time scale and describes, since $\tilde{x}e^{tQ} \to \tilde{\mathbf{x}}\pi$ as $t \to \infty$, how the system reaches an homogenized state where $\tilde{X}(t) \approx \pi$. We now consider fluid limits, i.e., the process on the time scale commensurate with the size of the initial state.

Theorem 4.3.7. *Consider a sequence of initial states (x_n) such that $\mathbf{x}_n \to \infty$ and $x_n/\mathbf{x}_n \to \tilde{x}$, and define $\overline{X}_n(t) = X(\mathbf{x}_n t)/\mathbf{x}_n$. Then for every $\varepsilon > 0$,*

$$\theta_\varepsilon \circ \overline{X}_n \xrightarrow{\mathrm{d}} \left(\left(1 + (\boldsymbol{\lambda} - \boldsymbol{\mu})(t+\varepsilon)\right)^+ \times \pi,\ t \in \mathbb{R}_+\right)$$

where $\xrightarrow{\mathrm{d}}$ refers to the convergence in distribution under $\mathbb{P}^{\mathbf{x}_n}$. Moreover, if $\tilde{x} = \pi$, then the same result holds with $\varepsilon = 0$.

When $\tilde{x} \neq \pi$, the discontinuity at 0 comes from the fact that the process gets homogenized in a constant time, which happens instantaneously on the fluid time scale. This has a similar flavor as the discussion in Section 4.2.3 on the Processor-Sharing queue: in both cases, a discontinuity at 0 is created by the coexistence of two time scales, where the fast time scale drives the process to an equilibrium point reached instantaneously in the slow time scale.

Heavy traffic regime

We now consider scaling limits in the heavy traffic regime where $\boldsymbol{\lambda} \approx \boldsymbol{\mu}$. To this end, we consider two sequences (λ^n) and (μ^n) in $\mathbb{R}_+^K$ and denote by $\mathbb{P}_n^x$ the law of the Markov process with parameters λ^n and μ^n (the other parameters K and Q being held fixed) started at x. We consider the following heavy traffic assumption, where as previously we have $\boldsymbol{\lambda}^n = \sum_{k=1}^K \lambda_k^n$ and $\boldsymbol{\mu}^n = \sum_{k=1}^K \mu_k^n$ and we define in addition $\rho_n = \boldsymbol{\lambda}^n / \boldsymbol{\mu}^n$.

Heavy traffic assumption: $\boldsymbol{\lambda}^n \leq \boldsymbol{\mu}^n$ for each $n \in \mathbb{N}$ and there exist $\ell \in (0,\infty)$ and $\alpha \in \mathbb{R}_+$ such that $\boldsymbol{\lambda}^n \to \ell$ and $n\big(1-\rho_n\big) \to \alpha$.

The homogenization property provides a picture of what happens when there is a large number of customers in the network: customers are spread across the various nodes according to the stationary distribution π, and in particular it is unlikely for any of the nodes to be empty. Thus, the full aggregate service rate is likely to be used and the total number of customers evolves as in a single $M/M/1$ queue with the combined service rate of all nodes. This property provides a useful handle on the processes of interest far away from zero. Imagine for instance that the network starts empty: it will eventually become highly loaded, at which point the homogenization property kicks in and holds until the network becomes empty (or close to) again. In other words, the homogenization property should give us control over excursions that reach a certain height, which makes it natural to use the approach presented in Section 2.3, and in particular Theorem 2.3.9, to prove the following result.

Theorem 4.3.8. *Assume that the heavy traffic assumption holds and for $t \in \mathbb{R}_+$ let $X_n(t) = X(n^2 t)/n$. Then X_n under $\mathbb{P}_n^0$ converges in distribution to the K-dimensional process $R(W) \times \pi$, where W is a Brownian motion with drift $-\ell\alpha$ and variance 2ℓ started at 0.*

Thus the joint queue length process asymptotically concentrates on a line whose angle corresponds to the stationary distribution π, thus exhibiting a form of state space collapse. The total number of customers, after scaling, behaves asymptotically as in a single $M/M/1$ queue, and thus evolves as a reflected Brownian motion. These characteristics are strongly reminiscent of the heavy traffic behavior of the joint queue length process in various queueing networks, see for instance [22, 101, 115, 118, 121]. However, to the best of our knowledge, this is the first result which shows that mobility of customers, rather than scheduling, routing or load balancing, can act as a mechanism producing state space collapse.

To conclude the presentation of our results, we now mention implications of the homogenization property for the stationary distribution and for sojourn times. In the following statements, we assume that $\alpha > 0$: since $n(1-\rho_n) \to \alpha$, this implies in particular that for n large enough, we have $\rho_n < 1$ in which case $\mathbb{P}_n$ has a unique stationary version which we denote by $\mathbb{P}_n^*$.

Moreover, we define E_α and E_1' are independent exponential random variables with parameters α and 1, respectively, and Υ denotes the sojourn time of one of the $\mathbf{X}(0)$ initial customers chosen uniformly at random, with $\Upsilon = 0$ if $\mathbf{X}(0) = 0$. Since the stationary distribution of the process $R(W)$ appearing in the previous statement is the law of E_α, the following result can be interpreted as an interchange of limits.

Theorem 4.3.9. *Assume that the heavy traffic assumption holds and that $\alpha > 0$. Then*

$X(0)/n$ under $\mathbb{P}_n^*$ converges in distribution as $n \to \infty$ to the K-dimensional vector $E_\alpha \times \pi$.

We now state a result describing the sojourn time of a typical customer. Recall that by assumption, $\ell \in (0,\infty)$ is the limit of the total arrival rate $\boldsymbol{\lambda}^n$ which has thus to be thought of as the total arrival rate in the asymptotic regime $n \to \infty$.

Theorem 4.3.10. *Assume that the heavy traffic assumption holds and that $\alpha > 0$ and let (x_n) in $\mathbb{Z}_+^K$ with $x_n/n \to b\pi$ for some $b \in (0,\infty)$: then $n^{-1}\Upsilon$ under $\mathbb{P}_n^{x_n}$ converges in distribution to bE_1'/ℓ.*

Averaging over the stationary distribution of $\mathbb{P}_n$ under which $\mathbf{X}(0)/n \xrightarrow{\text{d}} E_\alpha$, this readily entails the following result.

Corollary 4.3.11. *Assume that the heavy traffic assumption holds and that $\alpha > 0$: then $n^{-1}\Upsilon$ under $\mathbb{P}_n^*$ converges in distribution to $E_\alpha E_1'/\ell$.*

This corollary is similar to heavy traffic results for the sojourn time distribution of ordinary Processor-Sharing queue, see for instance [111, 124, 128]. This result may be intuitively explained by the snapshot principle, see Reiman [99]: in heavy traffic conditions the total number of customers in the system hardly varies over the time scale of a sojourn time. Thus each individual customer sees a service rate that is random, determined by the inverse of the total number of customers in stationarity which has an asymptotically exponential distribution in heavy traffic, but nearly constant over the duration of its sojourn time.

It is worth emphasizing that although in the present model the customers within each of the individual nodes are served in a Processor-Sharing manner, at any given time the service rates of customers may strongly vary across nodes. Due to the homogenization property, however, the empirical distribution of the location of each individual customer over the course of a long sojourn time in a heavy traffic regime will be close to the stationary distribution π. Hence each individual customer will see a π-weighted average of the service rates in the various nodes, which is only affected by the exponentially distributed total number of customers in the entire system, just like in an ordinary Processor-Sharing queue. More formally, the asymptotic behavior of Υ comes from the following approximations, starting from the formula (4.5) for the service received by a customer initially present and still present at time t:

$$\begin{aligned}
\int_0^t \frac{\mu_{\xi(s)}}{X_{\xi(s)}(s)}\mathrm{d}s &\approx \int_0^t \frac{\mu_{\xi(s)}}{\mathbf{X}(s)\pi_{\xi(s)}}\mathrm{d}s && \text{(homogenization property)}\\
&\approx \frac{1}{\mathbf{X}(0)}\int_0^t \frac{\mu_{\xi(s)}}{\pi_{\xi(s)}}\mathrm{d}s && \text{(snapshot principle)}\\
&\approx \frac{1}{\mathbf{X}(0)} \times t \times \sum_{k=1}^K \frac{\mu_k}{\pi_k} \times \pi_k && \text{(ergodic theorem).}
\end{aligned}$$

Thus asymptotically, the service received by a typical customer grows linearly in time, with proportionality coefficient equal to $\mu/\mathbf{X}(0)$. Since $\mu^n \to \ell$, this provides an intuitive explanation for Theorem 4.3.10.

4.3.3 Ideas

We conclude this section with a presentation of some ideas underlying the results presented above.

Martingale argument for the proof of Proposition 4.3.2

The proof of Proposition 4.3.2 relies on the construction of a martingale which decouples time and space. There have been two constructions of this martingale, one directly for the open system in [ST10] and one restricted to the closed system in [BS13]. In the latter approach, where the martingale is only constructed for the closed system, a bound similar to the bound of Proposition 4.3.2 is first proved for the closed system using this martingale, and then transferred to the open system via a coupling argument between the open and closed systems.

Martingale for the open system. In [ST10] this martingale is built directly for the open system, i.e., with general λ and μ, but needs the assumption that Q is diagonalizable. This martingale is a multidimensional generalization of the martingale built in [37] for the $M/M/\infty$ queue, which is not completely surprising given the form of the generator Ω in (4.3). The approach relies on building a family of space-time harmonic functions indexed by some parameter $c \in \mathbb{R}^n$, and then on integrating over c in such a way as to preserve the harmonic property;

Martingale for the closed system. In [BS13] this martingale is built only for the closed system. In this case, the construction is much simpler and can be carried through without assuming Q to be diagonalizable.

Let us give some details on the martingale and how it is used in the latter case, in particular in the sequel we do not assume that Q is diagonalizable.

Let $\mathscr{S}_\le = \{u \in [0,1]^{K-1} : \mathbf{u} \le 1\}$ and $\mathscr{S} = \{u \in [0,1]^K : \mathbf{u} = 1\}$ be the K-dimensional simplex. Let $P : \mathscr{S}_\le \to \mathscr{S}$ be the function that completes $u \in \mathscr{S}_\le$ into a probability distribution, i.e., $(Pu)_k = u_k$ if $1 \le k \le K-1$ and $(Pu)_K = 1 - \mathbf{u}$ for $u \in \mathscr{S}_\le$. Note that P is invertible with inverse $P^{-1} : \mathscr{S} \to \mathscr{S}_\le$ being the projection of the $K-1$ first coordinates. Let also $\Pi = \operatorname{diag}(\pi_1, \dots, \pi_K)$ be the diagonal matrix with entries (π_k) on the diagonal.

Moreover, we consider J the Jordan normal form corresponding to Q with change of basis matrix ω. If ω_i is ω's ith column and $I \in \{1, \dots, K\}$ is the number of distinct eigenvalues of Q, say $(\vartheta_1, \dots, \vartheta_I)$, for each $i = 1, \dots, I$ we consider $k(i) \in \{1, \dots, K\}$ such that $\omega_{k(i)}$ is associated to ϑ_i and we consider the following function $F : \mathbb{R}^K \to \mathbb{R}_+$:

$$F(u) = \prod_{i=1}^{I-1} \left|(\omega u)_{k(i)}\right|^{m_i}, \ u \in \mathbb{R}^K,$$

where m_i is the algebraic multiplicity of ϑ_i. For $c \in (0,\infty)$ and $t \in \mathbb{R}_+$ let

$$M_c(t) = e^{-c\gamma t} \int_{\mathscr{S}_\le} \prod_{k=1}^{K} \left(\frac{(Pu)_k}{\pi_k}\right)^{X_k(t)} \left(F(\Pi^{-1}Pu)\right)^{c-1} \mathrm{d}u.$$

Proposition 4.3.12. *For any $c > 0$ and $x \in \mathbb{Z}_+^K$, the process M_c is a bounded martingale under $\mathbb{P}_c^x$.*

The interest of this martingale is that it takes the form $e^{-c\gamma t} \times \Gamma(X(t))$ for some explicit function $\Gamma : \mathbb{R}_+^K \to \mathbb{R}_+$, and thus decouples the time and space variables. It is especially suited to stopping time arguments which make it possible to control Laplace transforms of hitting times. To illustrate this approach, consider the stopping time $T = T^\uparrow(\tilde{X} - \pi, \delta)$: then it can be shown that there exists a constant $C \in (0,\infty)$, which only depends on δ, such that

$$M_{1/(\gamma t)}(0) \le C \exp\left(K \log t + \frac{\mathbf{X}(0)}{\pi_\wedge} \left\| \frac{X(0)}{\mathbf{X}(0)} - \pi \right\|_1 \right) \tag{4.7}$$

and

$$\Gamma(X(T)) \geq C^{-1}\exp\left(-\frac{\mathbf{X}(0)\delta^2}{4}\right). \tag{4.8}$$

Armed with these two inequalities, we can then prove the bound

$$\mathbb{P}_c^x(T \leq t) \leq eC^2\exp\left(K\log t + \frac{\mathbf{x}\delta^2}{4} + \frac{\mathbf{x}}{\pi_\wedge}\left\|\frac{x}{\mathbf{x}} - \pi\right\|_1\right) \tag{4.9}$$

as follows:

$$\begin{aligned}
\mathbb{P}_c^x(T \leq t) &\leq e\mathbb{E}_c^x\left(e^{-T/t}\right) && \text{(Markov inequality)}\\
&= e\mathbb{E}_c^x\left(M_{1/(\gamma t)}(T) \times \frac{1}{\Gamma(X(T))}\right) && \text{(definition of } M_c\text{)}\\
&\leq eC\exp\left(\frac{\mathbf{x}\delta^2}{4}\right)\mathbb{E}_c^x\left(M_{1/(\gamma t)}(T)\right) && \text{(by (4.8))}\\
&= eC\exp\left(\frac{\mathbf{x}\delta^2}{4}\right)\mathbb{E}_c^x\left(M_{1/(\gamma t)}(0)\right) && \text{(optional stopping theorem),}
\end{aligned}$$

from which we get (4.9) using finally (4.7). This entails a bound similar to the one of Proposition 4.3.2 for the closed system, which can be transferred to the open system (and thus prove Proposition 4.3.2) via a coupling argument.

Heavy traffic through the control of excursions

Another set of ideas useful to prove Theorem 4.3.8 is that of focusing on excursions and use Theorem 2.3.9. As explained above, focusing on excursions is natural in the present context: indeed, the homogenization property gives a control on excursions big in the sense that their supremum is large. More precisely, once there are many customers, the homogenization property kicks and makes it possible to control the excursion until it is close to 0.

However, in order to invoke Theorem 2.3.9 one must control big excursions until they hit 0 exactly. Inspecting the bound in (4.3.2) reveals that it only makes it possible to control an excursion started with n customers as long as there are of the order of $\log n$ customers. The idea is to repeat inductively this argument, i.e., control the process started from $\log n$ customers and control it until there are $\log\log n$ customers, etc, and prove that the network empties shortly (on the diffusion time scale) after having only $\log\mathbf{X}(0)$ customers in the network: as $n \to \infty$,

$$\max_{x:\mathbf{x}=\lfloor\log n\rfloor} \mathbb{P}_n^x\left(T_0 \geq n^{1/2}\right) \to 0.$$

Note that, in order to control the asymptotic behavior of $e_\varepsilon^\uparrow$, thanks to the strong Markov property we only need to control the initial condition $X(T_\varepsilon^\uparrow)$, see the discussion preceding Proposition 2.3.8. This, again, is done thanks to the homogenization property which suggests that $X(T_\varepsilon^\uparrow) \approx \varepsilon\pi$.

4.4 Lingering effect for queue-based scheduling algorithms

This section presents the results of [SBB13] (see also [SBB14] for more technical details) on the delay performance of queue-based CSMA algorithms.

4.4.1 Context and illustrative example

Context

Constrained queueing networks consist of networks of queues with constraints on the set of queues that can be simultaneously active, i.e., transmitting a packet. These constraints may come from wireless interference, limited shared resources in a factory, access to objects in a database, etc, and can be modeled by a constraint graph, with nodes representing queues/servers and edges representing constraints. These networks constitute a versatile class of models that have recently triggered an intense research activity. One of the centerpieces in the scheduling literature is the celebrated MaxWeight algorithm as proposed in the seminal work [116, 117]. The MaxWeight algorithm provides throughput optimality and maximum queue stability in a variety of scenarios, and has emerged as a powerful paradigm in cross-layer control and resource allocation problems [39].

While not strictly optimal in terms of delay performance, MaxWeight algorithms do achieve so-called equivalent workload minimization and offer favorable scaling characteristics in heavy traffic conditions [114, 115]. As a further key appealing feature, MaxWeight algorithms only need information on the queue lengths and instantaneous service rates, and do not rely on any explicit knowledge of the underlying system parameters. On the downside, solving the maximum-weight problem tends to be challenging and potentially NP-hard. This is exacerbated in a network setting, where a centralized control entity may be lacking or require global state information, creating a substantial communication overhead in addition to the computational burden. This concern is especially relevant as the maximum-weight problem needs to be solved at a high pace, commensurate with the fast time scale on which scheduling algorithms typically need to operate.

This issue has provided a strong impetus for devising algorithms that entail lower computational complexity and communication overhead but retain the maximum stability of the MaxWeight algorithm. An exciting breakthrough in this quest was recently achieved by Shah and Shin [113] who proposed a low complexity, decentralized algorithm and proved that it achieved similar performance as MaxWeight in terms of throughput. The essence of Shah and Shin's algorithm is a queue-based scheduling algorithm that can be described as follows. At rate one, each node i that is not blocked (i.e., none of its neighbors is active) flips a coin and becomes inactive with probability $\psi(Q_i)$, with Q_i being the number of packets in i's buffer at this moment and $\psi : \mathbb{Z}_+ \to [0,1]$ a parameter of the algorithm, and becomes active with the complementary probability. Shah and Shin proved that if ψ decays sufficiently slowly, namely $\psi(x) \sim 1/\log x$ as $x \to \infty$, then this algorithm guarantees maximum stability for any constraint graph.

However, the picture becomes different when we consider performance metrics such as expected queue lengths or delays. Performance lower bounds in [21] indicate that the "cautious" back-off functions involved in establishing maximum stability tend to produce extremely large delays, typically growing *exponentially* in $1/(1-\rho)$, with ρ the load of the system, in contrast to the usual *linear* growth. More specifically, the bounds show that the expected queue lengths grow as $\psi^{-1}(1-\rho)$ as $\rho \uparrow 1$. Here ψ^{-1} represents the inverse of the decreasing function ψ. For $\psi = 1/\log$, this entails the exponential increase of the delay in $1/(1-\rho)$.

Besides, the reasoning behind the above lower bound suggests that the delay performance may be improved when the function ψ decays faster, e.g., inverse-polynomially: $\psi(a) \sim a^{-\beta}$, with $\beta > 0$, so that $\psi^{-1}(1-\rho) \sim 1/(1-\rho)^{1/\beta}$ as $\rho \uparrow 1$. The lower bound $\psi^{-1}(1-\rho)$ on delay would now only scales polynomially in $1/(1-\rho)$, and the larger the value of the exponent β, the slower the growth rate of $\psi^{-1}(1-\rho)$. In particular, this lower bound makes it possible that for $\beta \geq 1$, the expected queue lengths will only exhibit the usual linear growth in $1/(1-\rho)$ as $\rho \uparrow 1$. Note that a larger value of β means that a node is more "aggressive", in the sense that it is less likely to enter a back-off and more inclined to hold on to the medium, and hence the coefficient β will be referred to as the aggressiveness parameter. It is worth mentioning that maximum stability for the above back-off functions is not guaranteed by existing results, which do not apply for any $\beta > 0$. In fact, for $\beta > 1$, maximum stability has been shown not to hold in certain topologies [40].

Results of this section aim to gain fundamental insight to what extent a larger aggressiveness parameter can improve the delay performance. Our main finding is that for large values of β, a *lingering effect* can cause the mean stationary delay to increase in heavy traffic as $1/(1-\rho)^2$.

Illustrative example

Consider a network consisting of four queues which are split into two groups, in such a way that if either queue of one group is transmitting a packet, no queue of the other group may transmit, and vice versa. In terms of constraint graph, this corresponds to the complete bipartite case. A group is said to be *active* if one of its queues is transmitting a packet, and a queue is said to be active if it belongs to the active group. The other group and queues are said to be *inactive*.

In our model, time is slotted and active queues adhere to the following algorithm: after each transmission, each active queue flips a coin and *advertises a release* with probability $(1+a)^{-\beta}$, with a the number of packets that this queue has to transmit and $\beta > 0$. If the two active queues advertise a release simultaneously, then active queues become inactive and vice versa: such a time is called a *switching time*. This simple distributed algorithm gives rise to dynamics as illustrated in Figure 4.1, where the system is considered over three consecutive switching times t_1, t_2 and t_3.

Between switching times, the packet buffers at the active queues are drained, while packets accumulate at the inactive queues. The dynamics shown in Figure 4.1 are representative of the case $\beta > 1$ where a switch does not occur until both active queues are close to being empty, see Theorem 4.4.3 and Corollary 4.4.5.

Let us now give a flavor of the lingering effect. Imagine that the two active queues start with initial queue lengths of the same order, say n. As just mentioned, queues retain the shared resource until the time τ^* at which both queues are close to being empty, thus preventing other queues from activating until this time. The law of large numbers suggests that τ^* is of the order of n (i.e., active queues are drained linearly as in Figure 4.1), but the central limit theorem suggests that up to time τ^*, the first active queue to have been completely drained will be empty of the order of $n^{1/2}$ units of time, while waiting for the other queue to empty. This lingering effect is illustrated in Figure 4.1, which will be explained in greater detail in Section 4.4.5.

This leads to a fraction of the slots of the order of $n^{-1/2}$ where the shared resource is used inefficiently. This may at first sight seem negligible when n is large, and indeed queues seem to empty at the same time on the coarse time scale of Figure 4.1. However, we will actually establish in Theorem 4.4.3 that it has a significant impact

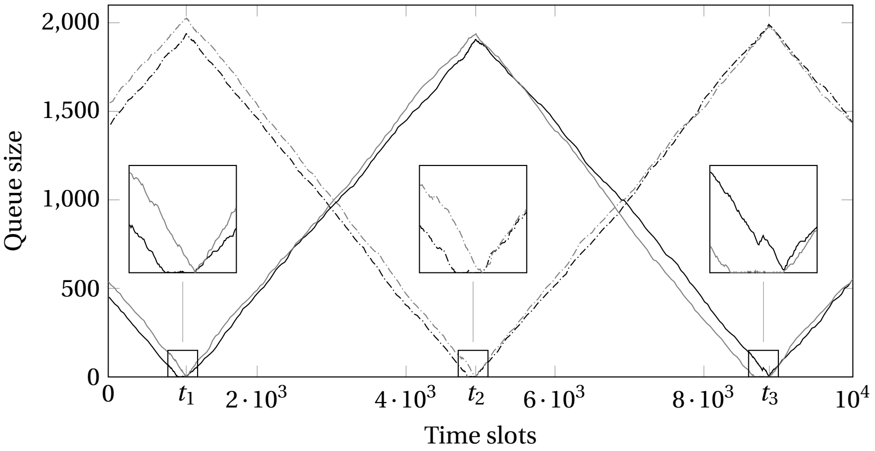


Figure 4.1: A sample path on the normal time scale with $\beta = 2$, representative of the case $\beta > 1$. The three boxes zoom in to show the lingering effect. The first queue to empty lingers around zero while the other queue is yet to empty, resulting in an inefficient use of the resource.

in heavy traffic, causing queue lengths to grow at a rate $1/(1-\rho)^2$ as $\rho \uparrow 1$, instead of the optimal $1/(1-\rho)$.

4.4.2 Model description

Informal description

Let us now give a more precise definition of our model, which as mentioned earlier corresponds to a complete bipartite constraint graph. In order to analyze the lingering effect in the simplest possible setting, we focus on a symmetric system consisting of two groups of $J \geq 2$ queues. At any given point in time, one of the groups is *active* while the other is *inactive*.

Time is slotted and inactive queues have simple dynamics, driven by independent and identically distributed numbers of packet arrivals in each slot, so that they each simply grow according to random walks with step size distribution denoted by ξ. During each time slot, active queues increase by independent amounts distributed as ξ as well but if at least one packet is present at the start of the slot, an active queue also flushes exactly one packet.

Moreover, at the end of each time slot, each active queue tosses a coin and advertises a *momentary release* with probability $\psi(a)$, with a the number of packets in the queue at the end of the time slot. This (momentary) release gives inactive queues an opportunity to become active: if all the active queues *simultaneously* advertise a release, then inactive queues become active and active queues become inactive. Such a time is called a *switching time*. We will assume in the sequel that $\psi(a) = (1+a)^{-\beta}$ for some parameter $\beta > 0$, called the aggressiveness parameter. In particular, $\psi(a) \to 0$ as $a \to \infty$, and so active queues are less likely to advertise a release when they are highly loaded; this mechanism thus gives priority to highly loaded queues in a distributed fashion. This model qualitatively resembles the canonical models for queue-based medium access control mechanisms [21, 92, 94, 112, 113], see [SBB13] for a detailed comparison.

Formal description and heavy traffic regime

The model is characterized by four parameters:

- $J \in \mathbb{N} \setminus \{1\}$, the number of queues in each group;
- a random variable $\xi \in \mathbb{Z}_+$ whose law governs the number of arrivals in each queue in each time slot;
- the aggressiveness parameter $\beta \in (0, \infty]$ from which the $[0,1]$-valued sequence $(\psi(a), a \geq 0)$ is defined via $\psi(a) = (1+a)^{-\beta}$ for $a \geq 0$ when $\beta \in (0,\infty)$, and $\psi(0) = 1$ and $\psi(a) = 0$ for $a > 0$ when $\beta = \infty$.

Note that we have been able to rigorously study the case $\beta = \infty$, while for $\beta < \infty$ we only have partial theoretical results. Our main results concern the process embedded at switching times, which lives in $\mathbb{Z}_+^J \times \mathbb{Z}_+^J$: the (discrete-time) canonical process on this space will be denoted by $X = (X^a, X^i)$, the first coordinate $X^a = (X_j^a)$ recording the state of active queues and the second coordinate $X^i = (X_j^i)$ the state of inactive queues.

To define the process of interest, we first need to describe the process in-between two switching times, which lives in $\mathbb{Z}_+^J \times \mathbb{Z}_+^J \times \{0,1\}^J$, where the additional 0/1 coordinate keeps track of whether active queues advertise a release or not. In order to avoid confusion, when considering $\mathbb{Z}_+^J \times \mathbb{Z}_+^J \times \{0,1\}^J$ as the canonical space, we will use the notation (A, I, B) to refer to the canonical process. For $a \in \mathbb{Z}_+^J$, let P^a be the law under which (A, I, B) obeys the following dynamic:

- the $2J$ processes (A_j, B_j) and I_j are mutually independent;
- for each $j = 1, \ldots, J$, I_j is a random walk started at 0 and where jumps are equal in distribution to ξ;
- for each $j = 1, \ldots, J$, (A_j, B_j) is a Markov chain started at $(a, 0)$ and governed by the following dynamics: for any $\alpha \in \mathbb{Z}_+$ and any function $f : \mathbb{Z}_+ \times \{0,1\} \to [0,\infty)$, we have
$$E^\alpha\left[f(A_j(1), B_j(1))\right] = \mathbb{E}\left[f\big(Y(\alpha), B(\alpha)\big)\right] \tag{4.10}$$
where:
  - $Y(\alpha) = \alpha + \xi - \mathbb{1}_{\{\alpha > 0\}}$ is the number of packets at the end of the time slot;
  - $B(\alpha) = \mathbb{1}(U < \psi(Y(\alpha)))$ indicates whether the queue advertises a release or not, with U uniformly distributed in $[0,1]$ and independent from ξ.

The switching time is then defined as $\tau^* = \inf\{k \in \mathbb{Z}_+ : B(k) = 1\}$, and the law $\mathbb{P}^x$ of the process started at $x = (x^a, x^i)$ and embedded at switching times is then given by

$$\mathbb{E}^x\left[f\big(X^a(1), X^i(1)\big)\right] = E^{x^a}\left[f\big(x^i + I(\tau^*), A(\tau^*)\big)\right]. \tag{4.11}$$

Note that, from the above definition of P^a, I only records the increments of inactive queues, which justifies the term x^i in the above definition. Furthermore, when $\beta = \infty$ we have $\tau^* = T_0(A)$ and so $A(\tau^*) = 0$, so that the above dynamic simply reduces to

$$\mathbb{E}^{x^a}\left[f\big(X^a(1)\big)\right] = E^{x^a}\left[f\left(I(\tau^*)\right)\right].$$

In the sequel, we finally consider $T^j = T_0(A_j)$ for $j = 1, \ldots, J$ and $T^{(j)}$ the T^j ranked in increasing order, i.e., $\{T^j\} = \{T^{(j)}\}$ and $T^{(1)} \leq \cdots \leq T^{(J)}$.

4.4.3 Presentation of the lingering effect and main results

Lingering effect

When $\beta > 1$, an active queue only advertises releases when it is close to being empty: more precisely, the number of packets in the queue at the time when it advertises the first release is of the order of one, even if the initial condition is very large, see Corollary 4.4.5 below. In particular, once a queue gains possession of the resource, it holds onto it, even when some or all of the other queues in the same group are empty and it would be more efficient for the queues in the other group to receive the resource. This causes a lingering effect as discussed in the introduction and illustrated in Figure 4.1 for a scenario with $J = 2$, $\beta = 2$. We have been able to rigorously prove this lingering effect for $\beta = \infty$ (see Theorem 4.4.3 below), while for $\beta > 1$ this remains an open problem.

It may appear that the two queues in the same group drain around the same time (as can indeed be shown to be the case on a "fluid scale"). When we zoom in, however, we see that there is actually a time period where one of the queues is already empty, while the other one clings to the resource and prevents the two queues in the other group from activating. In this section we give a heuristic explanation of how this inefficient use of the resource leads to a quadratic growth of the mean stationary delay in heavy traffic.

Consider a regime in which active queues only advertise a release when they are close to empty, i.e., $A(\tau^*) \approx 0$. Applying (4.11) with $f(x^a, x^i) = \mathbf{x}^a$ (recall the notation $\mathbf{x} = \|x\|_1$), we obtain

$$\mathbb{E}^x\left(\mathbf{X}^a(1)\right) = \mathbf{x}^i + E^{x^a}\left(\mathbf{I}(\tau^*)\right).$$

Thus integrating over $\mathbb{P}^\infty$, the stationary distribution of $\mathbb{P}^x$ assumed to exist, and defining the probability measure $P^\infty(\cdot) = \sum_a \mathbb{P}^a(\cdot)\mathbb{P}^\infty(X^a(0) = a)$, we obtain

$$\mathbb{E}^\infty(\mathbf{X}^a(0)) = \mathbb{E}^\infty(\mathbf{X}^a(1)) = \mathbb{E}^\infty(\mathbf{X}^i(0)) + E^\infty(\mathbf{I}(\tau^*)) = E^\infty(\mathbf{A}(\tau^*)) + E^\infty(\mathbf{I}(\tau^*))$$

where the last equality follows by applying (4.11) in stationarity and with $f(x) = \|x^i\|_1$. In particular, since $\mathbf{A}(\tau^*) \approx 0$, it follows that $\mathbb{E}^\infty(\mathbf{X}^a(0)) \approx E^\infty(\mathbf{I}(\tau^*))$ and by symmetry,

$$\mathbb{E}^\infty(X_1^a(0)) \approx E^\infty(I_1(\tau^*)).$$

Next, we have $E^\infty(I_1(\tau^*)) = \mathbb{E}(\xi)E^\infty(\tau^*)$ since I_1 is a random walk where jumps are equal in distribution to ξ and is independent from τ^*. Further, we have by definition $\mathbb{E}^\infty(X_1^a(0)) = E^\infty(A_1(0))$, so that

$$E^\infty(A_1(0)) \approx \mathbb{E}(\xi)E^\infty(\tau^*). \tag{4.12}$$

The goal is now to relate τ^* to $A_1(0)$ in a different way, see (4.13) below. Remember that active queues only advertise a release when they are close to empty; moreover, active queues are stable and so once all active queues are close to 0, it is only a matter of constant time for them to simultaneously advertise a release. This suggests that the switching time should occur around the largest time at which an active queue empties, i.e., this suggests the approximation $\tau^* \approx T^{(J)}$. The law of large numbers combined with the central limit theorem show that

$$T^j \approx \frac{A_j(0)}{1 - \mathbb{E}(\xi)} + A_j(0)^{1/2},$$

where we neglect multiplicative constants, possibly random, appearing in front of first- or second-order terms and that do not influence the order of magnitude of the final result. This leads to the approximation

$$T^{(J)} = \max_{j=1,\dots,J} T^j \approx \frac{\|A(0)\|_\infty}{1-\mathbb{E}(\xi)} + \|A(0)\|_\infty^{1/2}.$$

Since under $\mathbb{P}^\infty$ queues are symmetric, we have $\|A(0)\|_\infty \approx A_1(0) + A_1(0)^{1/2}$ which finally leads to

$$\tau^* \approx \frac{A_1(0)}{1-\mathbb{E}(\xi)} + A_1(0)^{1/2}. \tag{4.13}$$

Combining (4.12) and (4.13), we thus obtain

$$E^\infty(A_1(0)) \approx \frac{\mathbb{E}(\xi)}{1-\mathbb{E}(\xi)} E^\infty(A_1(0)) + E^\infty\left[A_1(0)^{1/2}\right].$$

Thus upon a concentration-like result of the kind $\mathbb{E}^\infty[A_1(0)^{1/2}] \approx [\mathbb{E}^\infty(A_1(0))]^{1/2}$ it is reasonable to expect

$$\left(1 - \frac{\mathbb{E}(\xi)}{1-\mathbb{E}(\xi)}\right)\mathbb{E}^\infty(A_1(0)) \approx \left[\mathbb{E}^\infty(A_1(0))\right]^{1/2}.$$

i.e., since $1-\mathbb{E}(\xi)/(1-\mathbb{E}(\xi)) \approx 1-\rho$,

$$\mathbb{E}^\infty(A_1(0)) \approx \frac{1}{(1-\rho)^2}.$$

In summary, $E^\infty(A_1(0))$, and hence $\mathbb{E}^\infty(\|X(0)\|_1)$, should grow as $1/(1-\rho)^2$ in the heavy traffic regime. While admittedly crude, the above heuristic arguments provide the correct estimates and can be turned into a rigorous proof when $\beta = \infty$.

As reflected in the above computations, the square factor really stems from the relation $\tau^* \approx T^{(1)} + \|A(0)\|_\infty^{1/2}$, i.e., τ^* occurs somehow long after $T^{(1)}$, the time at which it would be optimal to switch in order to avoid inefficient use of the resource. But it is difficult to make the system switch exactly at $T^{(1)}$ in a distributed fashion, and here the penalty incurred is a square root. The penalty may seem negligible but this small inefficiency has a significant impact in heavy traffic.

Main results

Let $\rho = 2\mathbb{E}(\xi)$. Before studying delay performance, one must first prove stability. The following result settles this question, it can be proved by showing that, in the terminology of Gamarnik and Zeevi [38], the L_∞ norm is a geometric Lyapunov function, and more precisely that when $\beta = \infty$ and $\mathbb{E}(\xi) < 1/2$, then

$$\lim_{K\to\infty} \sup_{x^a : \|x^a\|_\infty \geq K} \left(\frac{1}{\|X^a(0)\|_\infty}\mathbb{E}^{x^a}\left[\|X^a(2)\|_\infty - \|X^a(0)\|_\infty\right]\right) < 0.$$

Proposition 4.4.1. *If $\beta = \infty$ and $\rho < 1$, then $(\mathbb{P}^x)$ is positive recurrent. If $\mathbb{P}^\infty$ denotes its stationary version, then $\mathbb{E}^\infty(\|X(0)\|_1) < \infty$.*

Concerning the lingering effect, the previous discussion has highlighted that it essentially comes from the fact that the switching time τ^* is much larger than the first time $T^{(1)}$ at which a queue empties. Technically, the following result provides the necessary estimate. Note that by the central limit theorem, $T^{(J)} - T^{(1)}$ is of the order of $\|A(0)\|_\infty^{1/2}$. Also, note that when $\beta = \infty$ we simply have $\tau^* = T_0(A)$: although the hitting time of 0 by an $M/M/1$ queue has an explicit Laplace transform, it is more challenging to prove anything explicit on the hitting time of $0 \in \mathbb{Z}_+^J$ for J i.i.d. $M/M/1$ queues. The following result only shows that in this multi-dimensional setting, the hitting time of 0 happens a constant time after every queue has hit 0 at least once.

Proposition 4.4.2. *Assume that $\beta = \infty$ and let $\rho_0 < 2$ and $M > 0$: then*

$$\sup_{a,\xi} E^a \left(\tau^* - T^{(J)} \right) < \infty,$$

where the supremum is taken over vectors $a \in \mathbb{Z}_+^J$ and random variables ξ with $\rho < \rho_0$ and $\mathbb{E}(\xi^2) \leq M$.

Once we have these two results, then the arguments laid down above can essentially be carried through under the following heavy traffic regime. For $n \in \mathbb{N}$ we consider ξ_n with $\mathbb{E}(\xi_n) < 1/2$, $\mathbb{E}(\xi_n) \to 1/2$ and $\sup_n \mathbb{E}(\xi_n^2) < \infty$, and we define $\rho_n = 2\mathbb{E}(\xi_n)$. We denote by $\mathbb{P}_n^x$ the law of the process with these parameters, and by $\mathbb{P}_n^\infty$ its stationary version which exists by the previous result.

Theorem 4.4.3. *If $\beta = \infty$, then*

$$0 < \liminf_{n\to\infty} \left[(1-\rho_n)^2 \mathbb{E}_n^\infty (\|X(0)\|_1) \right] \leq \limsup_{n\to\infty} \left[(1-\rho_n)^2 \mathbb{E}_n^\infty (\|X(0)\|_1) \right] < \infty. \tag{4.14}$$

In the case $\beta > 1$, we have only proved the following result which shows that $\tau^* - T^{(J)}$ is, as in the case $\beta = \infty$, of constant order. However, this result is not strong enough in order to be able to carry through all the arguments presented above. Let $\tau_1 = \inf\{k \in \mathbb{Z}_+ : B_1(k) = 1\}$ be the first time at which the first active queue advertises a release and P_1^a for $a \in \mathbb{Z}_+$ be the law of (A_1, B_1) under $\mathbb{P}^\alpha$, for any $\alpha \in \mathbb{Z}_+^J$ with $\alpha_1 = a$. The following result shows that the first time the first active queue advertises a release is close to the time at which this queue hits 0 and that at this time this queue is of constant order.

Proposition 4.4.4. *If $\mathbb{E}(\xi) < 1$ and $\beta > 1$, then $(A_1(\tau_1), \tau_1 - T^1)$ under P_1^n converges in distribution as $n \to \infty$ to a finite random variable.*

This result, which deals with the behavior of one active queue, immediately implies that the switching time is close to the time at which the last active queue hits 0 and that at the switching time all active queues are of constant order.

Corollary 4.4.5. *If $\mathbb{E}(\xi) < 1$ and $\beta > 1$, then for any sequence $a_n \in \mathbb{Z}_+^J$ with $\|a_n\|_1 \to \infty$, $(A(\tau^*), \tau^* - T^{(J)})$ under P^{a_n} converges in distribution to a finite random variable.*

4.4.4 An intriguing heavy-tail phenomenon for $\beta \in (1,2)$

Although Corollary 4.4.5 suggests that Theorem 4.4.3 remains valid in the case $\beta > 1$ in light of the heuristic arguments presented above, the result as such is not strong enough for the proofs to go through. For one, the previous heuristic reasoning is based on the mean behavior of the random variables, whereas Corollary 4.4.5 only addresses the behavior in law. And actually, both empirical and theoretical evidence suggests that when $\beta > 1$ is small enough, presumably $\beta \in (1,2)$, heavy-tail phenomena appear which make these two behaviors – in the mean and in law – different.

On the one hand, although $A(\tau^*)$ under P_1^n converges in distribution as $n \to \infty$ to a finite random variable as soon as $\beta > 1$, simulation results suggest that, for $\beta = 1.2$, $J = 2$ and a symmetric initial state $a_n = (n, n)$, $E^{a_n}(A_1(\tau^*))$ grows like $n^{0.4}$ as $n \to \infty$. The following theoretical result may explain this phenomenon.

It is proved in [SBB13] that, for any function ψ, the limit in distribution of $\tau_1 - T^1$ under P_1^n, as $n \to \infty$, is a random variable $\mu \in \mathbb{Z}$ having the following property:

$$\mathbb{E}\left(|\mu|; \mu \leq 0\right) = \sum_{k \geq 0} k \mathbb{E}\left(\psi(S(k)) \prod_{i=0}^{k-1} \left(1 - \psi(S(i))\right) \mid I \geq 1\right)$$

where S is a random walk started at 0 with jumps distributed according to $1 - \xi$ and $I = \min_{k \geq 1} S(k)$. This relation comes from an intuitive path reversal argument, see Lemma 5.3 in [SBB13]. As $i \to \infty$, we have $S(i) \sim (1 - \rho/2)i$ by the strong law of large numbers, which suggests the approximation as $k \to \infty$

$$\mathbb{E}\left(\psi(S(k)) \prod_{i=0}^{k-1} \left(1 - \psi(S(i))\right) \mid I \geq 1\right) \approx \psi((1 - \rho/2)k) \prod_{i=0}^{k-1} \left(1 - \psi((1 - \rho/2)i)\right),$$

up to multiplicative constants. For $\psi(x) \approx x^{-\beta}$ with parameter $\beta > 1$, the infinite product $\prod_{i \geq 0}(1 - \psi(i))$ is finite and so

$$\mathbb{E}\left(|\mu|; \mu \leq 0\right) \approx \sum_{k \geq 1} k \times \psi(k) \approx \sum_{k \geq 1} \frac{1}{k^{\beta - 1}}.$$

This suggests that $\mathbb{E}(|\mu|; \mu \leq 0) = \infty$ when $\beta \in (1, 2)$ which thus corroborates the previous simulation result. If correct, these arguments would imply, by Fatou's lemma, that although $A(\tau^*)$ converges in distribution to a finite random variable, its mean diverges to ∞. In this case, it is not clear at all whether Theorem 4.4.3 would remain true or not when $\beta \in (1, 2)$.

Chapter 5

Application of branching processes in mathematical finance

Contents

5.1 Study of a model of limit order book 81
5.1.1 Models and main results 81
5.1.2 Long-time behavior of the price process in the case $\lambda > \mu$ . . 83
5.1.3 Scaling limit in the case $\lambda = \mu$, $\mathbb{E}(J) > 0$ 86

In this chapter we present an application of the theory of branching processes to the study of a stochastic model for the limit order book. Results of this chapter can be found in [LRS, Sim14].

5.1 Study of a model of limit order book

5.1.1 Models and main results

Context

A limit order book is a financial trading mechanism that keeps track of orders made by traders, thereby making it possible to execute trades in the future. Typically, a trader places an order to buy a security at a certain level x. If the price of the security is higher than x when the order is placed, then the order is kept in the book and may be fulfilled later in the future, as the price of the security fluctuates and falls below x. Similarly, traders may place sell orders, which gives rise to two-sided order books. Because of the importance of limit order books in financial markets, there has been a lot of research on these models, see for instance the survey by Gould et al. [42].

There are many variants for the information on the book to which traders have access. For instance, traders may only have access to the current so-called bid and ask prices, that correspond to the lowest sell order and the highest buy order. In this case, traders have an incentive to place orders in the vicinity of these prices in order for their order to be fulfilled quickly while at the same time buying (or selling) at a reasonable price. More generally, the dynamic of a limit order book is intricate because its current state influences its future evolution. Stochastic models capturing this dynamic have for instance been proposed in Cont et al. [25], Lakner et al. [78] and Kelly and Yudovina [65]. In [LRS, Sim14] we have studied the one-sided limit order book model of Lakner et al. [78] and exhibited unexpected connections with the theory of branching processes. We will study two closely related models which we introduce now.

Model on $\mathbb{R}$

The first model is a Markov process on the space $\mathcal{M}_p(\mathbb{R})$ of finite point measure on $\mathbb{R}$. Because of the financial motivation behind the model, we will sometimes call a measure $\nu \in \mathcal{M}_p(\mathbb{R})$ a *book*, an atom an *order* and the supremum $\pi(\nu)$ of the support of ν the *price*. The model is characterized by three parameters: two real numbers $\lambda, \mu > 0$ and the law of some real random variable J. Throughout, J is assumed to be integrable with $\mathbb{E}(J) \in \mathbb{R}$. Given the current state $\nu \in \mathcal{M}_p(\mathbb{R})$, there are two possible transitions:

- at rate μ and provided that $|\nu| > 0$, an atom located at $\pi(\nu)$ is removed;
- at rate λ, an atom is added to ν at location $\pi(\nu) + J$ with J independent from everything else.

This corresponds to a Markov process whose infinitesimal generator is given by

$$\omega(f)(\nu) = \lambda \mathbb{E}\left[f(\nu + \epsilon_{\pi(\nu)+J}) - f(\nu) \right] + \mu \left[f(\nu - \epsilon_{\pi(\nu)}) - f(\nu) \right] \mathbb{1}_{\{|\nu|>0\}}$$

for any measurable function $f : \mathcal{M}_p(\mathbb{R}) \to \mathbb{R}_+$ and any $\nu \in \mathcal{M}_p(\mathbb{R})$. Note the boundary condition that, when an order is added to an empty book, its location is distributed according to J. Actually, we will be interested in the case $\lambda > \mu$ so that this happens only a finite number of times since the total number of orders in the book is an $M/M/1$ queue with arrival rate λ and service rate μ. Let in this section $\rho = \lambda/\mu$ and $\mathbb{P}^\nu$ be the law of the Markov process with infinitesimal generator ω started at $\nu \in \mathcal{M}_p(\mathbb{R})$.

Theorem 5.1.1. *Assume that $\rho > 1$ and let $\rho_0^{-1} = \inf_{\eta \geq 0} \mathbb{E}(e^{\eta J}) \in [1, \infty]$. Then under $\mathbb{P}^\nu$ for any $\nu \in \mathcal{M}_p(\mathbb{R})$, the asymptotic behavior of $\pi(X(t))$ as $t \to \infty$ is governed by the following cases:*

- *if $\mathbb{E}(J) > 0$, then $\pi(X(t)) \overset{\text{a.s.}}{\to} +\infty$;*
- *if $\mathbb{E}(J) < 0$, $\mathbb{P}(J > 0) > 0$ and $\rho > \rho_0$, then $\pi(X(t)) \overset{\text{a.s.}}{\to} +\infty$;*
- *if $\mathbb{E}(J) < 0$, $\mathbb{P}(J > 0) > 0$ and $\rho < \rho_0$, then $\pi(X(t)) \overset{\text{a.s.}}{\to} -\infty$.*

The remaining cases are $\mathbb{E}(J) < 0$ and $\mathbb{P}(J > 0) = 0$ which is trivial, and the boundary case $\mathbb{E}(J) = 0$ which is left open. This theorem is proved thanks to a coupling with a branching random walk explained below.

Reflected model on $\mathbb{R}_+$

The above result is concerned with the case $\rho > 1$. We will also be interested in the case $\rho = 1$, where the total number of orders is a critical $M/M/1$ queue. In this case, the boundary condition matters and for convenience, we will study a reflected version of $\mathbb{P}^\nu$ where each negative order is “pushed back” to 0. Given the current state $\nu \in \mathcal{M}_p(\mathbb{R}_+)$, there are two possible transitions for this reflected model:

- similarly as for the unreflected model, at rate μ and provided that $|\nu| > 0$, an atom located at $\pi(\nu)$ is removed;
- at rate λ, an atom is added to ν at location $(\pi(\nu) + J)^+$ with J independent from everything else.

This corresponds to a Markov process whose infinitesimal generator $\underline{\omega}$ is given by

$$\underline{\omega}(f)(\nu) = \lambda\mathbb{E}\left[f(\nu + \epsilon_{(\pi(\nu)+J)^+}) - f(\nu)\right] + \mu\left[f(\nu - \epsilon_{\pi(\nu)}) - f(\nu)\right]\mathbb{1}_{\{|\nu|>0\}}$$

for any measurable function $f : \mathcal{M}_p(\mathbb{R}_+) \to \mathbb{R}_+$ and any $\nu \in \mathcal{M}_p(\mathbb{R}_+)$. Let in the sequel $\vartheta_n : \mathcal{M}(\mathbb{R}_+) \to \mathcal{M}(\mathbb{R}_+)$ act on measures as follows:

$$\vartheta_n(\nu)[y,\infty) = \frac{1}{n}\nu[ny,\infty),\ y \geq 0, \tag{5.1}$$

and let $\underline{\mathbb{P}}^\nu$ be the law of the Markov process with generator $\underline{\omega}$ and started at ν, and $\underline{\mathbf{P}}_n^\nu$ be the law of $(\vartheta_n \circ X(n^2 t), t \in \mathbb{R}_+)$ under $\underline{\mathbb{P}}^\nu$. We will omit the superscript when the initial state is the empty measure $\mathbf{z}$, i.e., we will write $\underline{\mathbb{P}}$ and $\underline{\mathbf{P}}_n$ for $\underline{\mathbb{P}}^\mathbf{z}$ and $\underline{\mathbf{P}}_n^\mathbf{z}$, respectively, with their corresponding expectations $\underline{\mathbb{E}}$ and $\underline{\mathbf{E}}_n$. Let W be a standard Brownian motion and $\alpha = (2\lambda)^{1/2}$.

Theorem 5.1.2. *Assume that $\lambda = \mu$, $\mathbb{E}(J) > 0$ and $\mathbb{P}(J \in \{-j^*, -j^*+1, \ldots, 0, 1\}) = 1$ for some $j^* \in \mathbb{N}$. Then $\underline{\mathbf{P}}_n$ converges weakly to the unique probability measure under which X satisfies the following two properties:*

a. $\pi \circ X$ is equal in distribution to $\alpha\mathbb{E}(J)R(W)$;

b. $X(t)$ for each $t \geq 0$ is absolutely continuous with respect to Lebesgue measure with density $\mathbb{1}_{\{0\leq y\leq\pi(X(t))\}}/\mathbb{E}(J)$, i.e.,

$$X(t)[0,y] = \frac{1}{\mathbb{E}(J)}\min\bigl(y, \pi(X(t))\bigr),\ t, y \geq 0. \tag{5.2}$$

This theorem is proved thanks to a connection with Lévy trees explained below.

5.1.2 Long-time behavior of the price process in the case $\lambda > \mu$

Asymmetric behavior of the price process

Before explaining the coupling underlying the proof of Theorem 5.1.1, let us gain some intuition. In general, the behavior of the price is asymmetric due to the system's dynamic. On the one hand, the price increases when an order is added to the right of the current price, and so an increase of the price is distributed according to J given that $J > 0$. On the other hand, the price decreases when an order is removed from the book, in which case the decrease of the price depends on the distance between the price and the location of the second rightmost order. In particular, orders to the left of the price act as a barrier that slow down the price as it wants to drift downward.

Thus, although $\mathbb{E}(J) < 0$ seems at first a natural condition for the price to drift to $-\infty$, it seems plausible that if ρ is sufficiently large, there will be so many orders sitting to the left of the price that they will eventually make the price drift to $+\infty$. This intuition turns out to be correct, and Theorem 5.1.1 explicitly exhibits the threshold ρ_0 below which the price drifts to $-\infty$ and above which it drifts to $+\infty$.

Moreover, it is very likely that the asymmetric behavior of the price process also translates to the speed at which it goes to $\pm\infty$. Indeed, the above argument suggests that the price process goes at linear speed to $+\infty$ and only at a logarithmic speed to $-\infty$.

This kind of behavior is strongly reminiscent of the behavior of extremal particles in branching random walks. There, although a typical particle drifts to $-\infty$ when $\mathbb{E}(J) < 0$, one may still observe atypical trajectories due to the exponential explosion in the number of particles, see the classical references by Hammersley [50], Kingman [71] and Biggins [9]. This analogy has actually been our initial motivation to investigate the relation between $\mathbb{P}$ and branching random walks. And indeed, we will show in Theorem 5.1.5 that $\mathbb{P}$ can be realized as some functional of a branching random walk, and this coupling will make the proof of Theorem 5.1.1 quite intuitive.

Coupling with a branching random walk

From a high-level perspective, the coupling adds a new dimension to the initial limit order book model by creating a filiation between the orders: typically, we will declare an order x to be a child of some other order y if x was added to the book at a time where y corresponded to the price of the book, i.e., x is added at a time t where $y = \pi(X(t))$. As mentioned earlier, similar ideas were developed by Kendall [66] in queueing theory where they have proved extremely successful: in this context, a customer is typically declared to be the child of a customer in service.

If an order is the unique rightmost atom of the current configuration, then with probability $\mu/(\lambda+\mu)$ this order will be removed at the next transition, and with probability $\lambda/(\lambda+\mu)$ it will be assigned a new child. In particular, it is intuitively clear that this construction will give each order (at most) a geometric number of offspring. By labeling the edge between two orders with the displacement of the child with respect to its parent, which has distribution J, we end up with a Galton–Watson tree with geometric offspring distribution and i.i.d. real-valued labels on the edges, i.e., a branching random walk. The idea of the coupling is to reverse this construction and start from the branching random walk to build the book process $\mathbb{P}^\nu$. To do so, we will realize X under $\mathbb{P}^\nu$ as the iteration of a deterministic tree operator Φ on a random tree, thus encoding all the randomness in the tree.

We now present the formal coupling, which involves discrete trees where each node is endowed with a label and a color. Formally, let in the sequel

$$\mathcal{U}_{\ell,c} = \mathcal{U} \times \mathbb{R} \times \{\mathrm{g}, \mathrm{r}, \mathrm{w}\}.$$

Definition 5.1.3. *A discrete, colored and labelled tree* $\mathtt{t}$ *is a subset of* $\mathcal{U}_{\ell,c}$ *such that:*

- *the projection* $\mathcal{T} = \{u \in \mathcal{U} : \exists(\ell, c) \in \mathbb{R} \times \{g, r, w\} \text{ s.t. } (u, \ell, c) \in \mathtt{t}\}$ *of* $\mathtt{t}$ *on* $\mathcal{U}$ *is a discrete tree, called the genealogical structure of* $\mathtt{t}$*;*
- *for every* $u \in \mathcal{T}$*, there exists a unique pair* $(\ell_{\mathtt{t}}(u), c_{\mathtt{t}}(u)) \in \mathbb{R} \times \{g,r,w\}$ *such that* $(u, \ell_{\mathtt{t}}(u), c_{\mathtt{t}}(u)) \in \mathtt{t}$*;*
- *the sets* $\mathcal{G}(\mathtt{t}) = \{(u, \ell, c) \in \mathtt{t} : c_{\mathtt{t}}(u) = g\}$ *and* $\mathcal{R}(\mathtt{t}) = \{(u, \ell, c) \in \mathtt{t} : c_{\mathtt{t}}(u) = r\}$ *are finite;*
- *the set* $\mathcal{G}(\mathtt{t}) \cup \mathcal{R}(\mathtt{t})$ *contains the root and is connected.*

In the sequel we extend the domains of the mappings $\ell_{\mathtt{t}}$ and $c_{\mathtt{t}}$ to $\mathtt{t}$, so that $\ell_{\mathtt{t}}(v) \in \mathbb{R}$ has to be thought of the label of $v \in \mathtt{t}$ and $c_{\mathtt{t}}(v) \in \{\mathrm{g}, \mathrm{r}, \mathrm{w}\}$ as its color: either g(reen), r(ed) or w(hite). Labels on the nodes induce labels on the edges in the following way: the label of the edge $e = (u, uj)$ for any $u \in \mathcal{T}$ and $j \in \mathbb{N}$ such that

$uj \in \mathcal{T}$ is equal to $\ell_{\mathtt{t}}(uj) - \ell_{\mathtt{t}}(u)$. Let in the sequel $\widetilde{\mathcal{U}}_{\ell,c}$ the set of discrete, colored and labelled trees. For $\mathtt{t} \in \widetilde{\mathcal{U}}_{\ell,c}$ let $\Gamma(\mathtt{t}) \in \mathcal{M}_p(\mathbb{R})$ be the point measure recording the labels of the green nodes of $\mathtt{t}$:

$$\Gamma(\mathtt{t}) = \sum_{v \in \mathcal{G}(\mathtt{t})} \epsilon_{\ell_{\mathtt{t}}(v)}.$$

Let $\widetilde{\mathcal{U}}^*_{\ell,c} = \{\mathtt{t} \in \widetilde{\mathcal{U}}_{\ell,c} : |\mathcal{G}(\mathtt{t})| > 0\}$ be the set of trees with at least one green node, and for $\mathtt{t} \in \widetilde{\mathcal{U}}^*_{\ell,c}$ let $\gamma(\mathtt{t})$ be the green node with largest label and $w(\mathtt{t})$ the number of white children of $\gamma(\mathtt{t})$. If there are several green nodes with maximal labels, we choose the last one where in the sequel, nodes are ordered according to the lexicographic order. Note that the label of $\gamma(\mathtt{t})$ in $\mathtt{t}$ is simply the supremum of the support of $\Gamma(\mathtt{t})$, i.e., $\ell_{\mathtt{t}} \circ \gamma(\mathtt{t}) = \pi \circ \Gamma(\mathtt{t})$. We now introduce the mapping which, when iterated on a suitable random tree, makes it possible to recover $\mathbb{P}^{\nu}$.

Definition 5.1.4. *Let $\Phi : \widetilde{\mathcal{U}}_{\ell,c} \to \widetilde{\mathcal{U}}_{\ell,c}$ be the following operator: if $|\mathcal{G}(\mathtt{t})| = 0$ then $\Phi(\mathtt{t}) = \mathtt{t}$, while if $\mathtt{t} \in \widetilde{\mathcal{U}}^*_{\ell,c}$, then Φ changes the color of one node according to the following rule:*

- *if $w(\mathtt{t}) > 0$, Φ transforms the first white child of $\gamma(\mathtt{t})$ into a green node;*
- *if $w(\mathtt{t}) = 0$, then Φ transforms $\gamma(\mathtt{t})$ into a red node.*

We also define Φ_n as the nth iterate of Φ, defined by Φ_0 being the identity map and $\Phi_{n+1} = \Phi \circ \Phi_n$ for $n \geq 0$, and we define $\kappa(\mathtt{t}) = \inf\{n \geq 0 : |\mathcal{G}(\Phi_n(\mathtt{t}))| = 0\} \in \mathbb{Z}_+ \cup \{\infty\}$ the first time where iterating Φ on $\mathtt{t}$ creates a tree with no green node. We can now state our main result.

Theorem 5.1.5. *Let $\mathtt{T}_\varrho$ for $\varrho \in \mathbb{R}$ be the following random $\widetilde{\mathcal{U}}_{\ell,c}$-valued tree:*

- *the genealogical structure of $\mathtt{T}$ is a Galton–Watson tree with geometric offspring distribution with parameter $\lambda/(\lambda+\mu)$;*
- *labels on the edges are i.i.d. equal in distribution to J, independently from the genealogical structure, and the root has label ϱ;*
- *all nodes are white, except for the root which is green.*

Then for any $\varrho \in \mathbb{R}$, $(\Gamma \circ \Phi_n(\mathtt{T}), 0 \leq n \leq \kappa(\mathtt{T}_\varrho))$ is equal in distribution to the process $(X(t), 0 \leq t \leq T_0(|X|))$ under $\mathbb{P}^{\epsilon_\varrho}$ embedded at jump epochs.

In words, the process $(X(t), 0 \leq t \leq T_0(|X|))$ under $\mathbb{P}^{\epsilon_0}$ can be realized by starting from the random tree $\mathtt{T}$, iterating the map Φ and only keeping track of the locations of the green nodes. Thus in order to determine the asymptotic behavior of $\pi(X(t))$ as $t \to \infty$, one needs to understand the asymptotic behavior of $\pi \circ \Gamma \circ \Phi_n(\mathtt{T})$, conditionally on $|\mathtt{T}| = \infty$ (because of the boundary condition when the book empties, and because $\kappa(\mathtt{t}) = \infty$ if and only $|\mathtt{t}| = \infty$), as $n \to \infty$.

The proof of this coupling is very intuitive, although a rigorous proof requires to delve into the details of the dynamic induced by Φ. Indeed, the memoryless property of the geometric random variable suggests that for any n, $\gamma \circ \Phi_n(\mathtt{T})$ has at least one white child with probability $\lambda/(\lambda+\mu)$. If this is the case, then applying Φ adds an atom to $\Gamma \circ \Phi_n(\mathtt{T})$, whose location is given by the label of $\gamma \circ \Phi_n(\mathtt{T})$ plus the value of the label of the newly added edge. Since this label did not play a role thus far, it is independent from everything else and distributed according to J. If on the other hand $\gamma \circ \Phi_n(\mathtt{T})$ has no more white child, then applying Φ removes an atom of $\Gamma \circ \Phi_n(\mathtt{T})$ located, by definition of γ, at $\pi \circ \Gamma \circ \Phi_n(\mathtt{T})$.

Case $\pi(X(t)) \to -\infty$

Assume that $\mathbb{E}(J) < 0$, $\mathbb{P}(J > 0) > 0$ and $\rho < \rho_0$: then Theorem 5.1.5 gives a transparent proof of the fact that $\pi(X(t)) \to -\infty$. Since $\rho > 1$ by assumption, we have in particular $\rho_0 > 1$ and so there must exist $\eta > 0$ such that $\mathbb{E}(e^{\eta J}) < \infty$. Let M_n be the rightmost point of the branching random walk $\mathtt{T}$ at time n, i.e.,

$$M_n = \max\{\ell_{\mathtt{T}}(u) : v \in \mathtt{T} \text{ and } |v| = n\}$$

where $|v|$ is the distance from v to the root. Under the assumptions made on J and ρ, Theorem 4 in Biggins [9] shows that $M_n \to -\infty$ almost surely in the event of non-extinction, and in particular, $\ell_{\mathtt{T}}(v_n) \to -\infty$ for any sequence of nodes (v_n) such that $|v_n| \to \infty$. In order to conclude the proof it remains to observe that the dynamic induced by Φ is such that $\gamma \circ \Phi_n(\mathtt{T})$ can only spend a finite amount of time at any fixed distance from the root, i.e., $|\gamma \circ \Phi_n(\mathtt{T})| \to \infty$.

Case $\pi(X(t)) \to +\infty$

Proving that $\pi \circ \Gamma \circ \Phi_n(\mathtt{T}) \to +\infty$ when either $\mathbb{E}(J) > 0$, or $\mathbb{E}(J) < 0$, $\mathbb{P}(J > 0) > 0$ and $\rho > \rho_0$ requires more care, and we only sketch the main idea: the details are worked out in [Sim14].

The key observation is that as long as an order is in the book, the behavior of the price does not depend on the state of the book to the left of this order. In particular, in order to compute the probability that the order sitting initially in the book at 0 is never removed from the book, we may as well assume that all orders that are placed in $(-\infty, 0)$ are instantaneously removed, or killed.

In terms of the underlying tree $\mathtt{T}$, removing all orders that are placed in $(-\infty, 0)$ amounts to removing all nodes v, together with all their descendants, with label $\ell_{\mathtt{T}}(v) < 0$. We thus obtain a new tree $B(\mathtt{T})$, a subtree of the original tree $\mathtt{T}$, which is a well-known object: this is precisely a branching random walk with a barrier at 0. Under our assumptions on J and ρ, a result of Biggins et al. [10] shows that the probability $\mathbb{P}(|B(\mathtt{T})| = \infty)$ of $B(\mathtt{T})$ being infinite is strictly positive.

Going back to the limit order book, $B(\mathtt{T})$ being infinite means exactly that the initial order sitting at 0 will never be removed. Since this happens with positive probability, an order is eventually added to the book that is never removed. This order then constitutes a barrier under which the price never falls. Then, a renewal type argument shows that this phenomenon repeats itself: at regular intervals, an order is added to the book that constitutes a new barrier under which the price never falls. Eventually, this barrier moves up and forces to price to diverge to $+\infty$, and this argument even suggests that the speed is linear.

5.1.3 Scaling limit in the case $\lambda = \mu$, $\mathbb{E}(J) > 0$

Connection with Lévy trees

For $f \in \mathscr{D}(\mathbb{R})$, $a \geq 0$ and $g \leq d$ we say that the function $e = (f((g + t) \wedge d) - a, t \geq 0)$ is an *excursion of f above level a* if $e \in \mathscr{E}(\mathbb{R})$, $e(t) \geq 0$ for every $t \geq 0$ and $f(g-) \leq a$. The following lemma provides the key connection with Lévy trees and is at the heart of our proof of Theorem 5.1.2.

Lemma 5.1.6. *Let $a \in \mathbb{Z}_+$. Then under $\mathbb{P}$, the sequence of successive excursions of $\pi \circ X$ above level a are i.i.d. with common distribution the first excursion of $\pi \circ X$ away from 0 under $\mathbb{P}^{\epsilon_1}$.*

Proof. Fix some $a \in \mathbb{Z}_+$, and consider X under $\mathbb{P}^\nu$ for any $\nu \in \mathcal{M}_p(\mathbb{R}_+)$ with $\pi(\nu) \le a$ that only puts mass on integers. Then when the first excursion of $\pi \circ X$ above a begins, the price is at a and an order is added at $a+1$. Thus if g is the left endpoint of the first excursion above a, $X(g)$ must be of the form $X(g) = X(g-) + \epsilon_{a+1}$ with $\pi(X(g-)) = a$. This excursion lasts as long as at least one order sits at $a+1$, and if d is the right endpoint of the first excursion above a, then what happens during the time interval $[g, d]$ above a is independent from $X(g-)$ and is the same as what happens above 0 during the first excursion of $\pi \circ X$ away from 0 under $\mathbb{P}^{\epsilon_1}$. Moreover, $X(d)$ only puts mass on integers and satisfies $\pi(X(d)) \le a$, so that thanks to the strong Markov property we can iterate this argument and obtain the result. □

This regenerative property shows that the first excursion of $\pi \circ X$ away from 0 is almost the contour process of Galton–Watson tree. More precisely, consider a stochastic process $H \in \mathcal{E}$ with finite length and continuous sample paths, that starts at 1, increases or decreases with slope ± 1 and only changes direction at integer times.

For integers $a \in \mathbb{Z}_+$ and $p \in \mathbb{N}$ and conditionally on H having p excursions above level a, let $(e^k_{a,p}, k = 1, \dots, p)$ be these p excursions. Then H is the contour function of a Galton–Watson tree if and only if for each a and p, the $(e^k_{a,p}, k = 1, \dots, p)$ are i.i.d. with common distribution H. Indeed, H can always be seen as the contour function of some discrete tree. With this interpretation, the successive excursions above a of H code the subtrees rooted at nodes at depth $a+1$ in the tree. The $(e^k_{a,p}, k = 1, \dots, p)$ being i.i.d. therefore means that the subtrees rooted at a node at depth $a+1$ are i.i.d., which, for $a = 0$, is precisely the definition of a Galton–Watson tree.

The difference between this regenerative property and the regenerative property satisfied by $\pi \circ X$ under $\mathbb{P}$ and described in Lemma 5.1.6 is that, *when conditioned to belong to the same excursion away from* 0, consecutive excursions of $\pi \circ X$ above some level are neither independent, nor identically distributed. If for instance we condition some excursion above level a to be followed by another such excursion within the same excursion away from 0, this biases the number of orders put in $\{0, \dots, a\}$ during the first excursion above a. Typically, one may think that more orders are put in $\{0, \dots, a\}$ in order to increase the chance of the next excursion above a to start soon, i.e., before the end of the current excursion away from 0.

However, this bias is weak and will be washed out in the asymptotic regime of Theorem 5.1.2. Thus it is natural to expect that $\pi \circ X$ under $\underline{\mathbf{P}}_n$ will converge to a process satisfying a continuous version of the discrete regenerative property satisfied by the contour function of Galton–Watson trees.

Such a regenerative property has been studied in Weill [119], who has showed that it characterizes the contour process of Lévy trees (see for instance Duquesne and Le Gall [31] for a background on this topic). Thus upon showing that this regenerative property passes to the limit, we will have drastically reduced the possible limit points for $\pi \circ X$, and it will remain to show that, among the contour processes of Lévy trees, the limit that we have is actually a reflected Brownian motion. From there, an argument based on local time considerations makes it possible to conclude that Theorem 5.1.2 holds.

The main steps for proving Theorem 5.1.2 are: (1) showing tightness of $\underline{\mathbf{P}}_n$; (2) showing, based on Lemma 5.1.6, that for any accumulation point $\mathbf{P}$, $\pi \circ X$ under $\mathbf{P}$ satisfies the regenerative property studied in Weill [119]; (3) arguing that among the contour processes of Lévy trees, $\pi \circ X$ under $\mathbf{P}$ must actually be a reflected Brownian motion; (4) showing that $X(t)$ under $\mathbf{P}$ has density $\mathbb{1}_{\{y \le \pi \circ X(t)\}} / \mathbb{E}(J)$ with respect to Lebesgue measure. Although this proof strategy is very attractive at a conceptual level, many technical details need to be taken care of along the way. By and large, the second point of the above program is the most challenging one, and the coupling with a branching random walk turns out to provide key estimates to control the asymptotic behavior of some random times.

More on the second step

In order to give an idea of the technical details involved and how the coupling with a branching random walk of Section 5.1.2 is used, let us give a flavor of the arguments involved in the second step. Let $\mathbf{P}$ be an arbitrary accumulation point of the sequence $\underline{\mathbf{P}}_n$, and let $\mathbf{Q} = \mathbf{P} \circ \pi^{-1}$ be the law of $\pi \circ X$ under $\mathbf{P}$. Then $\mathbf{Q}$ is a probability measure on $\mathscr{D}(\mathbb{R})$, and we will denote by E the canonical process on this space. The first thing to do is to show that $\mathbf{Q}$ has an excursion measure. In order to do so, we need, among other things, to show that the Lebesgue measure of the zero set of E, namely $\ell(t) = \int_0^t \mathbb{1}_{\{E(s)=0\}} \mathrm{d}s$, is $\mathbf{Q}$-almost surely equal to 0. Using a continuity argument and other arguments from queueing and excursion theory, we can reduce this problem to showing that

$$\limsup_{n \to \infty} \frac{1}{n} \underline{\mathbb{E}}^{\epsilon_1} \left(\int_0^{T_0 \circ \pi} \mathbb{1}_{\{0 < \pi \circ X(s) \le n\varepsilon\}} \mathrm{d}s \right) \xrightarrow[\varepsilon \to 0]{} 0.$$

Recall the random tree $\mathtt{T}_1$ of Theorem 5.1.5, and let B the operator that acts on trees by removing all nodes, together with their descendants, with label < 0. Then a slight variation of Theorem 5.1.5, incorporating the different boundary condition, implies that the integral term $\int_0^{T_0 \circ \pi} \mathbb{1}_{\{0 < \pi \circ X(s) \le n\varepsilon\}} \mathrm{d}s$ is approximately equal to the number of nodes in the tree $B(\mathtt{T}_1)$ with label $\le \varepsilon n$. Since we are in the ballistic regime $\mathbb{E}(J) > 0$, this property is easily established for the tree $\mathtt{T}_1$ and can then be transferred to $B(\mathtt{T}_1)$.

Once the existence of an excursion measure of $\mathbf{Q}$, say $\mathscr{N}$, is established, we need to prove that it satisfies the regenerative property of Weill [119]. Let in the sequel $\xi_{a,u}$ for $a, u \in \mathbb{N}$ denote the number of excursions of E before time T_0 with height $> u$, and $E_{a,u}^k$ for $k \in \{1, \dots, \xi_{a,u}\}$ the kth excursion of E above a with height $> u$. Then we need to prove that for any $f : \mathscr{D}(\mathbb{R}) \to \mathbb{R}_+$ continuous and bounded,

$$\mathscr{N} \left(\prod_{k=1}^{p} f_k(E_{a,u}^k) \mid \xi_{a,u} = p \right) = \prod_{k=1}^{p} \mathscr{N} \big(f_k(E) \mid \sup E > u \big).$$

This equality means that, under $\mathscr{N}(\cdot \mid \xi_{a,u} = p)$, the p excursions of E above a with height $> u$ are i.i.d. with common distribution $\mathscr{N}(\cdot \mid \sup E > u)$. In order to do so, we prove that

$$\underline{\mathbf{E}}_n \left(\prod_{k=1}^{p} f_k(E_{a,u}^k) \right) \xrightarrow[n \to \infty]{} \prod_{k=1}^{p} \mathscr{N} \big(f_k(E) \mid \sup E > u \big)$$

and that

$$\underline{\mathbf{E}}_n \left(\prod_{k=1}^{p} f_k(E_{a,u}^k) \right) \xrightarrow[n \to \infty]{} \mathscr{N} \left(\prod_{k=1}^{p} f_k(E_{a,u}^k) \mid \xi_{a,u} = p \right).$$

The first convergence stems from the fact, discussed after Lemma 5.1.6, that the $E^k_{a,u}$ are almost i.i.d., and in order to show that they are indeed asymptotically i.i.d., we use a coupling argument. The second convergence on the other hand follows by a continuity argument: since $\underline{\mathbf{P}}_n \xrightarrow{\text{w}} \mathbf{P}$, it is natural to expect the excursions of $\pi \circ X$ (above level a and with height $> u$) to converge to the corresponding excursions of E under $\mathscr{N}$, which is the meaning of this statement. To prove such a continuity argument, we need to control the convergence of particular random times, namely the endpoints of the excursions of interest. The control of such random times is controlled thanks to the coupling with the branching random.

Chapter 6

Research perspectives

Contents

6.1 Branching processes **91**

6.1.1 A general candidate for the scaling limit of the chronological height process 92

6.1.2 Scaling limits beyond the case of short edges 93

6.2 Stochastic averaging via functional analysis **93**

6.3 Stochastic networks **94**

6.3.1 Delay performance of queue-based CSMA protocols 94

6.3.2 Performance of CSMA protocols in a dynamic setting 95

To conclude this manuscript, this chapter contains my research perspectives for the next three years or so. I am only presenting the research topics in line with the results presented here, which will account for most of my research activity. I believe that my shared interests in branching processes and stochastic networks represent a fruitful balance which I will try to maintain.

6.1 Branching processes

Concerning branching processes, I intend to delve deeper into the scaling limits of the height and contour processes of Crump–Mode–Jagers trees, aiming to extend the results of [SS] presented in Section 3.4. The main idea is to leverage the general decomposition results of [SS] to go beyond the case of short cases. To give a flavor of the approach considered, recall the formula (3.12):

$$\mathbb{H}(n) = \left(\sum_{k: 0 < T(k) \leq n} A(k) \right) \circ \vartheta_n.$$

This purely deterministic formula holds for any chronological tree. In the case of a Crump–Mode–Jagers tree it expresses the height process at a fixed time as a simple functional of two explicit renewal processes with an intricate correlation structure.

As seen in Section 3.4, the case of short edges essentially corresponds to the case where one of these renewal processes is asymptotically deterministic: this makes these two processes asymptotically independent and entails that, in the limit, the chronological and genealogical height (and contour) processes are proportional to one another. I believe that this formula can be leveraged in at least two other promising directions:

i) define and study a general candidate of the scaling limit of the height process;

ii) prove convergence toward this candidate in some cases which go beyond the case of short edges.

Note that these two research directions concern scaling limits of the height process. Once (if) these steps are successfully carried out, the next natural step will then be to investigate the convergence of the chronological contour process. It seems plausible that, in general, the height and contour processes are still linked by a time-change relation, possibly random in the case of long edges, which represents an interesting research direction in the longer term.

6.1.1 A general candidate for the scaling limit of the chronological height process

The formula (3.12) naturally suggests a general candidate for the scaling limit of the height process. More precisely, all the objects appearing in (3.12) are deterministically built from the sequence $(\nu_k, k \in \mathbb{Z}_+)$ of offspring point processes of individuals ranked in lexicographic order, cf. Section 3.4.2. Provided in the limit we have a continuous analog of this sequence, it should be possible to carry on the same construction in a continuous setting. To shed some light on this idea, define $M(k) = \max_{\{0,\dots,n\}} S$, L the local time process of $M - S$ at 0 and $t_0 = 0 < t_1 < \cdots$ the points of discontinuity of L^{-1}. With this definition, we then have $T(k) = L^{-1}(t_k)$ and $S(T(k-1)) = S(L^{-1}(t_k-))$ from which it follows, since

$$A(k) = \Lambda\left(\nu_{T(k)-1}, S(T(k-1)) - S(T(k)-1)+1\right),$$

that

$$\sum_{k:0<T(k)\le n} A(k) = \sum_{s\le L(n):\Delta L^{-1}(s)>0} \Lambda\left(\nu_{L^{-1}(s)-1}, S(L^{-1}(s-)) - S(L^{-1}(s)-1)+1\right). \tag{6.1}$$

Intuitively, it seems that this construction can be performed when S is a Lévy process with infinite variation and no Gaussian component since in this case it can be constructed directly from the Poisson point process of jumps, see for instance Bertoin [8, Theorem I.1].

More precisely, $(\nu(t), t \in \mathbb{R}_+)$ would now be a Poisson point process on $\mathcal{M}(\mathbb{R}_+)$, say with characteristic measure N. We would then consider $\pi = N \circ |\cdot|^{-1}$, where $|\cdot| : \nu \in \mathcal{M}(\mathbb{R}_+) \mapsto |\nu|$, and assume that π satisfies $\int_0^\infty (x \wedge x^2)\pi(\mathrm{d}x) < \infty$: then for any $\mathtt{d} \in \mathbb{R}$ there is a deterministic construction of the Lévy process with Laplace exponent

$$\mathtt{d}u + \int_0^\infty \left(e^{-ux} - 1 + ux\right)\pi(\mathrm{d}x)$$

from the Poisson point process $|\cdot| \circ \nu$ which we will denote by S. Let $M(t) = \sup_{[0,t]} S$ and L be the local time process of $M - S$ at 0: then in view of (6.1), a natural continuous analog of the sum $\sum_{k:0<T(k)\le n} A(k)$ would be

$$\sum_{s\le L(t):\Delta L^{-1}(s)>0} \Lambda\left(\nu(L^{-1}(s)), S(L^{-1}(s-)) - S(L^{-1}(s)-)\right).$$

Let $H(t)$ be this sum, which we can thus view as a mapping applied at the killed Poisson point process $(\nu(s), 0 \le s \le t)$. Finally, the continuous analog of composing with the dual operator ϑ_n would be to apply the above construction, for each fixed t, not to $(\nu(s), s \le t)$ but rather to $\vartheta^c_t = (\nu(t-s), 0 \le s \le t)$. Thus upon correct definition and notation, the natural candidate for the scaling limit of the discrete chronological height process in this case is the process

$$\left(\left(\sum_{s \le L(t) : \Delta L^{-1}(s) > 0} \Lambda\left(\nu(L^{-1}(s)), S(L^{-1}(s-)) - S(L^{-1}(s)-)\right) \right) \circ \vartheta^c_t, \ t \in \mathbb{R}_+ \right).$$

I intend to try and make the above arguments rigorous, and to look at several questions that naturally arise along the way: can we perform this construction in a more general setting, i.e., when the Lévy process has a Gaussian component? what are the properties of this limiting object? For instance, if one thinks of a Crump–Mode–Jagers process as the local time process of the contour process, it is natural to consider the existence of a local time process associated to this object.

6.1.2 Scaling limits beyond the case of short edges

As explained above, the case of short edges corresponds to the case where the two renewal processes $(\sum_{i=0}^k A(i), k \in \mathbb{Z}_+)$ and $(T(k), k \in \mathbb{Z}_+)$ are asymptotically independent due to the fact that the first process properly scaled converges to a deterministic limit by the law of large numbers. As a second research topic, I intend to consider other cases where the correlation structure becomes asymptotically tractable, in which case it seems plausible to be able to show convergence to the general object defined in Section 6.1.1.

In collaboration with Emmanuel Schertzer, we have already identified one such case which we plan to study in the short term. Namely, consider the critical and non-triangular case: then assuming in addition that the offspring distribution has a finite second moment, i.e., $\mathbb{E}(|\mathscr{P}|^2) < \infty$, back-of-the-envelope calculation suggests that the processes $(\sum_{i=0}^k A(i), k \in \mathbb{Z}_+)$ and $(T(k), k \in \mathbb{Z}_+)$ properly scaled converge to two independent Poisson processes. This asymptotic independence suggests that this case should be tractable.

6.2 Stochastic averaging via functional analysis

The research topics presented next on stochastic networks rely to a large extent on the stochastic averaging principle. This principle is a general phenomenon when we have a Markov process which can be decomposed into a fast process, say σ, and a slow one, say Q. Under appropriate scaling regimes, the fast process mixes quickly which makes the slow process interacts with it only through its stationary distribution. When this stationary distribution depends itself on the slow process, we typically get a dynamical system (deterministic or stochastic) for the slow process.

From the above description, it should be clear that mixing time considerations play a key role for proving stochastic averaging. Roughly speaking, there will be stochastic averaging as soon as the mixing time of the fast process is much smaller than the typical time scale on which the slow process evolves. This suggests a general result that would make it possible to prove stochastic averaging from assumptions on mixing times, but I am not aware of any such result in the literature. Rather, the standard approach to prove stochastic averaging relies on considering a particular occupation measure and then using analytical arguments, see for instance Kurtz [77]. In an on-going collaboration with Laurent Miclo, we intend to try and formalize the above intuition based on mixing time thanks to results from functional analysis. More precisely, the stochastic averaging principle amounts to justifying the approximation

$$\int_0^t F(Q(s),\sigma(s))\mathrm{d}s \approx \int_0^t \pi_{Q(s)}[F(Q(s),\cdot)]\mathrm{d}s$$

where π_x is the stationary distribution of the fast process σ when the slow process Q is fixed to x, and $\mu[f] = \int f \mathrm{d}\mu$ denotes the integral of f against the measure μ. We thus have to control the difference $F(Q(s),\sigma(s)) - \pi_{Q(s)}[F(Q(s),\cdot)]$ which can be written as $\Omega_{Q(s)}(\phi(Q(s),\cdot))$ for some function ϕ, and where Ω_x is the generator of the slow process when the fast process is fixed to x. For each fixed x, the function $\phi(x,\cdot)$ is thus the solution to the Poisson equation associated to Ω_x and logarithmic Sobolev inequalities such as in [28] make it possible to relate it – and in particular its supremum – to the spectral gap of the fast process. Gathering all the pieces, this paves the way to relating the stochastic averaging principe directly to spectral gap considerations, which was precisely our initial goal. We expect this new approach to be particularly useful to study CSMA protocols which we discuss now.

6.3 Stochastic networks

Concerning stochastic networks, I intend to work on two topics related to CSMA protocols:

i) study the delay performance of queue-based CSMA protocols;

ii) study the performance of CSMA protocols in a dynamic setting with user arrivals and departures.

As alluded to above, both problems will in part aim at establishing stochastic averaging principles and thus rely on the new approach which we are developing with Laurent Miclo.

6.3.1 Delay performance of queue-based CSMA protocols

This research proposal is a direct continuation of the work initiated in [SBB13] and presented in Section 4.4. The general problem is the following: we have a constraint graph G and a function $\psi : \mathbb{Z}_+ \to [0,1]$ such that at rate one, each node i that is not blocked flips a coin and becomes inactive with probability $\psi(Q_i)$, with Q_i be number of packets in i's buffer at this moment, and becomes active with the complementary probability. This is in essence the protocol proposed in Shah and Shin [113], many variations of which exist in the literature.

Essentially, the results presented in Section 4.4 suggest that a more aggressive function ψ, i.e., that decays faster to 0, should improve the delay performance of such an protocol; this is also the conclusion to be drawn from [20]. However, the only general theoretical result guaranteeing maximum stability is due to Shah and Shin [113], who showed that the choice of $\psi(x) = 1/\log x$, i.e., decaying to 0 quite slowly, guarantees a maximal throughput. For this choice of functions, the delay in heavy traffic scales *exponentially* in $1/(1-\rho)$, which prohibits practical applications of such an protocol. Together with the result of Ghaderi et al. [40] who showed that a more aggressive ψ could result in diminishing the stability region, this is more or less the state of the art concerning these queue-based protocols.

The folklore has it that it is not possible to have a universally good function ψ, i.e., a function ψ which ensures maximum stability for any constraint graph, and that decays essentially faster than $1/\log(x)$. More precisely, the conjecture is that for any $\alpha > 0$, there exists a constraint graph G such that the protocol with $\psi(q) \approx q^{-\alpha}$ is not throughput optimal. I would like to work on this question, and more specifically to try and address the following one.

i) Given a network G, what are the functions ψ that give maximum stability?

For instance, we could be interested in the quantity

$$\alpha^*(G) = \sup\big\{\alpha > 0 : \text{the protocol with } \psi(a) = (1+a)^{-\alpha} \\ \text{is throughput optimal for the constraint graph } G\big\}.$$

This question is probably quite ambitious, and a first step toward answering it may be to restrict oneself to functions ψ that ensure *timescale separation.*

More precisely, the Markov process describing the above network can be written in the form (σ, Q) with σ the process of schedules and Q the process of queue lengths. The key property of the function $\psi(x) = 1/\log x$ which entails maximum stability for any constraint graph is that with this choice, the stochastic averaging principle holds for the process (σ, Q). In other words, for this choice of ψ the process σ evolves slowly compared to the queue length process Q, and this property is sometimes referred to as timescale separation in the queueing literature.

It is commonly believed that if timescale separation holds for ψ, then the protocol will stabilize the network whenever possible. I intend to work on this question, which naturally suggests to consider the following problem:

i') Given a network G, what are the functions ψ that ensure timescale separation?

I expect that the approach we are developing with Laurent Miclo will be particularly useful to answer this question, since it will probably make it possible to directly relate mixing time properties of Glauber dynamic on the hard-core model to the timescale separation property.

Finally, although there is a consensus in the community that more aggressive functions ψ should yield a better delay performance, there is no theoretical result in that vein. I therefore also intend to work on the following question.

ii) Among all the functions ψ that ensure maximum stability, do more aggressive ones yield a lower delay?

6.3.2 Performance of CSMA protocols in a dynamic setting

Almost every work on the performance of CSMA protocols considers a static setting, with a fixed number of static users with infinite flows to transmit. This corresponds to the fact that, in the previous description, the constraint graph G is fixed. This framework is very relevant for networks with a fixed infrastructure, such as Internet routers, but is more questionable in the case of ad-hoc networks where users come and go in a dynamic way. For this reason, the last problem on which I intend to work in the coming years is concerned with the performance analysis of CSMA protocols in a dynamic setting.

As a warm-up, I have begun to work on the following model in collaboration with Sem Borst and Fabio Cecchi. Users arrive uniformly at random on a circle according to a Poisson process with intensity λ, and try to activate at rate ν. If, when a user tries to activate, no active user is within interference range, then the user activates and leaves the network after an exponential random variable with parameter μ. If L is the length of the circle and R is the interference range, then the maximal number of users simultaneously active is $N^* = \lfloor L/R \rfloor$. The first natural conjecture is that the network is stable if $\lambda < N^* \mu$: indeed, if the network is overloaded after some time it should converge to a configuration with N^* active users and then drain. Moreover, timescale separation should clearly hold for such a network, and so we are trying to prove that the measure-valued process which describes the positions of users in the system converges, under some suitable asymptotic regime, toward an explicit measure-valued ordinary differential equation.

After dealing with this model, I intend to pursue several other interesting questions. First of all, I would like to investigate a potential *geometric heavy traffic regime*. Namely, the stability condition $\lambda < N^* \mu$ involves both the arrival and service rates λ and μ, but also the geometry of the network through the constant N^*. Consider for instance the case where $L = N^* R + \varepsilon$ and $N^* - 1 < \lambda/\mu < N^*$, where N^*, R, λ and μ are fixed: then as $\varepsilon \downarrow 0$, it should take a longer and longer time for the process to reach a state with exactly N^* active users, which is however necessary in order to drain the system. I therefore expect the stationary number of users to blow up under this regime, a phenomenon which would be very interesting to study in more details. Moreover, this phenomenon seems to be closely related to adsorption-desorption systems studied by physicists, see for instance Jin et al. [60], and it is an exciting problem to be able to draw formal connections. Finally, it would be interesting to go beyond the circle and study similar models where users arrive in a more complex geometric space, and understand the impact of the geometry on the system's performance.

Publications discussed in this manuscript

[BKS] Vincent Bansaye, Thomas G. Kurtz, and Florian Simatos. Tightness for processes with fixed points of discontinuities and applications in varying environment. Submitted.

[BS] Vincent Bansaye and Florian Simatos. A sufficient condition for the tightness of time-inhomogeneous Markov processes. arXiv:1409.5215.

[BS13] Sem Borst and Florian Simatos. A stochastic network with mobile users in heavy traffic. *Queueing Syst.*, 74(1):1–40, 2013.

[BS15] Vincent Bansaye and Florian Simatos. On the scaling limits of Galton–Watson processes in varying environments. *Electron. J. Probab.*, 20(75):1–36, 2015.

[LRS] Peter Lakner, Josh Reed, and Florian Simatos. Scaling limit of a limit order book model via the regenerative characterization of Lévy trees. arXiv:1312.2340.

[LS14] Amaury Lambert and Florian Simatos. The weak convergence of regenerative processes using some excursion path decompositions. *Ann. Inst. H. Poincaré Probab. Statist.*, 50(2):492–511, 2014.

[LS15] Amaury Lambert and Florian Simatos. Asymptotic Behavior of Local Times of Compound Poisson Processes with Drift in the Infinite Variance Case. *J. Theoret. Probab.*, 28(1):41–91, 2015.

[LSZ13] Amaury Lambert, Florian Simatos, and Bert Zwart. Scaling limits via excursion theory: Interplay between Crump-Mode-Jagers branching processes and Processor-Sharing queues. *Ann. Appl. Probab.*, 23(6):2357–2381, 2013.

[SBB13] Florian Simatos, Niek Bouman, and Sem Borst. Lingering issues in distributed scheduling. In *Proc. ACM SIGMETRICS '13*, pages 141–152, 2013.

[SBB14] Florian Simatos, Niek Bouman, and Sem Borst. Lingering issues in distributed scheduling. *Queueing Syst.*, 77(2):243–273, 2014.

[Sim14] Florian Simatos. Coupling limit order books and branching random walks. *J. Appl. Probab.*, 51(3):625–639, 2014.

[SS] Emmanuel Schertzer and Florian Simatos. Height and contour processes of Crump-Mode-Jagers forests (I): general distribution and scaling limits in the case of short edges. arXiv 1506.03192.

[ST10] Florian Simatos and Danielle Tibi. Spatial homogenization in a stochastic network with mobility. *Ann. Appl. Probab.*, 20(1):312–355, 2010.

References

[1] David Aldous. Stopping times and tightness. *Ann. Probab.*, 6(2):335–340, 1978.

[2] David Aldous. The continuum random tree. I. *Ann. Probab.*, 19(1):1–28, 1991.

[3] David Aldous. The continuum random tree. II. An overview. In *Stochastic analysis (Durham, 1990)*, volume 167 of *London Math. Soc. Lecture Note Ser.*, pages 23–70. Cambridge Univ. Press, Cambridge, 1991.

[4] David Aldous. The continuum random tree. III. *Ann. Probab.*, 21(1):248–289, 1993.

[5] David Aldous. Brownian excursions, critical random graphs and the multiplicative coalescent. *Ann. Probab.*, 25(2):812–854, 1997.

[6] M. T. Barlow. Necessary and sufficient conditions for the continuity of local time of Lévy processes. *Ann. Probab.*, 16(4):1389–1427, 1988.

[7] S. L. Bell and R. J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy. *Ann. Appl. Probab.*, 11(3):608–649, 2001.

[8] Jean Bertoin. *Lévy processes*, volume 121 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1996.

[9] J. D. Biggins. The first- and last-birth problems for a multitype age-dependent branching process. *Adv. in Appl. Probab.*, 8(3):446–459, 1976.

[10] J. D. Biggins, Boris D. Lubachevsky, Adam Shwartz, and Alan Weiss. A branching random walk with a barrier. *Ann. Appl. Probab.*, 1(4):573–581, 1991.

[11] Patrick Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999.

[12] Robert M. Blumenthal. *Excursions of Markov processes*. Probability and its Applications. Birkhäuser Boston Inc., Boston, MA, 1992.

[13] A. N. Borodin. The asymptotic behavior of local times of recurrent random walks with finite variance. *Teor. Veroyatnost. i Primenen.*, 26(4):769–783, 1981.

[14] A. N. Borodin. Asymptotic behavior of local times of recurrent random walks with infinite variance. *Teor. Veroyatnost. i Primenen.*, 29(2):312–326, 1984.

[15] A. N. Borodin. On the character of convergence to Brownian local time. I, II. *Probab. Theory Relat. Fields*, 72(2):231–250, 251–277, 1986.

[16] K. Borovkov. On the convergence of branching processes to a diffusion process. *Theory Probab. Appl.*, 30(3):496–506, 1986.

[17] K. Borovkov. A note on diffusion-type approximation to branching processes in random environments. *Theory Probab. Appl.*, 47(1):132–138, 2003.

[18] Sem Borst. User level performance of channel-aware scheduling algorithms in wireless data networks. *IEEE/ACM Trans. Netw.*, 13(3):636–647, June 2005.

[19] Sem Borst, Alexandre Proutiere, and Nidhi Hegde. Capacity of wireless data networks with intra- and inter-cell mobility. In *Proc. IEEE INFOCOM '06*, pages 1058–1069, 2006.

[20] N. Bouman, S. Borst, and J. van Leeuwaarden. Achievable delay performance in CSMA networks. In *Proc. 49th Annual Allerton Conference*, pages 384–391, September 2011.

[21] N. Bouman, S. C. Borst, and J. S. H. van Leeuwaarden. Stability of spatial wireless systems with random admissible-set scheduling. In *Proc. VALUETOOLS '11*, pages 57–65, ICST, Brussels, Belgium, Belgium, 2011. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[22] Maury Bramson. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Syst.*, 30(1-2):89–148, 1998.

[23] Maury Bramson. *Stability of queueing networks*, volume 1950 of *Lecture Notes in Mathematics*. Springer, Berlin, 2008. Lectures from the 36th Probability Summer School held in Saint-Flour, July 2–15, 2006.

[24] Hong Chen, Offer Kella, and Gideon Weiss. Fluid approximations for a processor-sharing queue. *Queueing Syst.*, 27(1-2):99–125, 1997.

[25] Rama Cont, Sasha Stoikov, and Rishi Talreja. A stochastic model for order book dynamics. *Oper. Res.*, 58(3):549–563, 2010.

[26] M. Csörgő and P. Révész. On strong invariance for local time of partial sums. *Stochastic Process. Appl.*, 20(1):59–84, 1985.

[27] J. G. Dai. On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.*, 5(1):49–77, February 1995.

[28] P. Diaconis and L. Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *Ann. Appl. Probab.*, 6(3):695–750, 1996.

[29] A. B. Dieker and X. Gao. Sensitivity analysis for diffusion processes constrained to an orthant. *Ann. Appl. Probab.*, 24(5):1918–1945, 2014.

[30] Paul Dupuis and Kavita Ramanan. A Skorokhod problem formulation and large deviation analysis of a processor sharing model. *Queueing Syst.*, 28(1-3):109–124, 1998.

[31] Thomas Duquesne and Jean-François Le Gall. Random trees, Lévy processes and spatial branching processes. *Astérisque*, (281):vi+147, 2002.

[32] Nathalie Eisenbaum and Haya Kaspi. A necessary and sufficient condition for the Markov property of the local time process. *Ann. Probab.*, 21(3):1591–1598, 1993.

[33] A. K. Erlang. Sandsynlighedsregning og Telefonsamtaler. *Nyt Tidsskrift for Matematik B*, 20:33, 1909.

[34] A. K. Erlang. Løsning af nogle Problemer fra Sandsynlighedsregningen af Betydning for de automatiske Telefoncentraler. *Elektroteknikeren*, 13:5, 1917.

[35] Stewart N. Ethier and Thomas G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1986. Characterization and convergence.

[36] William Feller. Diffusion processes in genetics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 227–246, Berkeley and Los Angeles, 1951. University of California Press.

[37] Christine Fricker, Philippe Robert, and Danielle Tibi. On the rates of convergence of Erlang's model. *J. Appl. Probab.*, 36(4):1167–1184, 1999.

[38] David Gamarnik and Assaf Zeevi. Validity of heavy traffic steady-state approximation in generalized Jackson networks. *Ann. Appl. Probab.*, 16(1):56–90, 2006.

[39] Leonidas Georgiadis, Michael J. Neely, and Leandros Tassiulas. Resource allocation and cross-layer control in wireless networks. *Found. Trends Netw.*, 1(1):1–144, April 2006.

[40] Javad Ghaderi, Sem Borst, and Phil Whiting. Backlog-based random-access in wireless networks: Fluid limits and instability. In *Proc. WiOpt*, pages 15–22, 2012.

[41] B. V. Gnedenko and A. N. Kolmogorov. *Limit distributions for sums of independent random variables*. Translated from the Russian, annotated, and revised by K. L. Chung. With appendices by J. L. Doob and P. L. Hsu. Revised edition. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills., Ont., 1968.

[42] Martin D. Gould, Mason A. Porter, Stacy Williams, Mark McDonald, Daniel J. Fenn, and Sam D. Howison. Limit order books. *Quant. Finance*, 13(11):1709–1742, 2013.

[43] P. J. Green. Conditional limit theorems for general branching processes. *J. Appl. Probability*, 14(3):451–463, 1977.

[44] Anders Grimvall. On the convergence of sequences of branching processes. *Ann. Probab.*, 2:1027–1045, 1974.

[45] H. Christian Gromoll. Diffusion approximation for a processor sharing queue in heavy traffic. *Ann. Appl. Probab.*, 14(2):555–611, 2004.

[46] H. Christian Gromoll, Łukasz Kruk, and Amber L. Puha. Diffusion limits for shortest remaining processing time queues. *Stoch. Syst.*, 1(1):1–16, 2011.

[47] H. Christian Gromoll, Amber L. Puha, and Ruth J. Williams. The fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab.*, 12(3):797–859, 2002.

[48] M. Grossglauser and D. Tse. Mobility increases the capacity of ad-hoc wireless networks. In *Proc. IEEE INFOCOM '01*, volume 3, pages 1360–1369, 2001.

[49] Shlomo Halfin and Ward Whitt. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Oper. Res.*, 29(3):567–588, May–June 1981.

[50] J. M. Hammersley. Postulates for subadditive processes. *Ann. Probab.*, 2:652–680, 1974.

[51] J. Michael Harrison and Martin I. Reiman. Reflected Brownian motion on an orthant. *Ann. Probab.*, 9(2):302–308, 1981.

[52] Inge S. Helland. Continuity of a class of random time transformations. *Stochastic Process. Appl.*, 7(1):79–99, 1978.

[53] Inge S. Helland. Minimal conditions for weak convergence to a diffusion process on the line. *Ann. Probab.*, 9(3):429–452, 1981.

[54] Kiyosi Itô. Poisson point processes attached to Markov processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. III: Probability theory*, pages 225–239, Berkeley, Calif., 1972. Univ. California Press.

[55] J. Jacod, J. Mémin, and M. Métivier. On tightness and stopping times. *Stochastic Process. Appl.*, 14(2):109–146, 1983.

[56] Jean Jacod and Albert N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, second edition, 2003.

[57] David L. Jagerman and Bhaskar Sengupta. The $GI/M/1$ processor-sharing queue and its heavy traffic analysis. *Comm. Statist. Stochastic Models*, 7(3):379–395, 1991.

[58] Peter Jagers. Diffusion approximations of branching processes. *Ann. Math. Statist.*, 42:2074–2078, 1971.

[59] Alain Jean-Marie and Philippe Robert. On the transient behavior of the processor-sharing queue. *Queueing Syst.*, 17:129–136, 1994.

[60] X. Jin, G. Tarjus, and J. Talbot. An adsorption-desorption process on a line: kinetics of the approach to closest packing. *J. Phys. A: Math. Gen.*, 27(7):L195, 1994.

[61] Olav Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

[62] Niels Keiding. Extinction and exponential growth in random environments. *Theoretical Population Biology*, 8(1):49–63, 1975.

[63] Offer Kella, Bert Zwart, and Onno Boxma. Some time-dependent properties of symmetric $M/G/1$ queues. *J. Appl. Probab.*, 42(1):223–234, 2005.

[64] F. P. Kelly. *Reversibility and Stochastic Networks.* Wiley, Chichester, 1979, reprinted 1987, 1994.

[65] Frank Kelly and Elena Yudovina. A Markov model of a limit order book: thresholds, recurrence, and trading strategies. Submitted.

[66] David G. Kendall. Some problems in the theory of queues. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):151–185, 1951.

[67] Davar Khoshnevisan. An embedding of compensated compound Poisson processes with applications to local times. *Ann. Probab.*, 21(1):340–361, 1993.

[68] J. F. C. Kingman. The single server queue in heavy traffic. *Proc. Cambridge Philos. Soc.*, 57:902–904, 1961.

[69] J. F. C. Kingman. Two similar queues in parallel. *Ann. Math. Statist.*, 32:1314–1323, 1961.

[70] J. F. C. Kingman. On queues in heavy traffic. *J. Roy. Statist. Soc. Ser. B*, 24:383–392, 1962.

[71] J. F. C. Kingman. The first birth problem for an age-dependent branching process. *Ann. Probab.*, 3(5):790–801, 1975.

[72] M. Yu. Kitayev and S. F. Yashkov. Analysis of a single-channel queueing system with the discipline of uniform sharing of a device. *Engineering Cybernetics*, 17:42–49, 1979.

[73] F. B. Knight. Random walks and a sojourn density process of Brownian motion. *Trans. Amer. Math. Soc.*, 109:56–86, 1963.

[74] Thomas G. Kurtz. Semigroups of conditioned shifts and approximation of Markov processes. *Ann. Probab.*, 3(4):618–642, 1975.

[75] Thomas G. Kurtz. Diffusion approximations for branching processes. In *Branching processes (Conf., Saint Hippolyte, Que., 1976)*, volume 5 of *Adv. Probab. Related Topics*, pages 269–292. Dekker, New York, 1978.

[76] Thomas G. Kurtz. Random time changes and convergence in distribution under the Meyer-Zheng conditions. *Ann. Probab.*, 19(3):1010–1034, 1991.

[77] Thomas G. Kurtz. Averaging for martingale problems and stochastic approximation. In *Applied stochastic analysis (New Brunswick, NJ, 1991)*, volume 177 of *Lecture Notes in Control and Inform. Sci.*, pages 186–209. Springer, Berlin, 1992.

[78] Peter Lakner, Josh Reed, and Sasha Stoikov. High frequency asymptotics for the limit order book. Preprint available on the authors' webpage.

[79] Amaury Lambert. The contour of splitting trees is a Lévy process. *Ann. Probab.*, 38(1):348–395, 2010.

[80] John Lamperti. Criteria for the recurrence or transience of stochastic process. I. *J. Math. Anal. Appl.*, 1:314–330, 1960.

[81] John Lamperti. Limiting distributions for branching processes. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 225–241. University of California Press, 1967.

[82] Jean-François Le Gall. The uniform random tree in a Brownian excursion. *Probab. Theory Related Fields*, 96(3):369–383, 1993.

[83] Jean-François Le Gall and Grégory Miermont. Scaling limits of random trees and planar maps. In *Probability and statistical physics in two and more dimensions*, volume 15 of *Clay Math. Proc.*, pages 155–211. Amer. Math. Soc., Providence, RI, 2012.

[84] Vlada Limic. On the behavior of LIFO preemptive resume queues in heavy traffic. *Electron. Comm. Probab.*, 5:13–27 (electronic), 2000.

[85] Vlada Limic. A LIFO queue in heavy traffic. *Ann. Appl. Probab.*, 11(2):301–331, 2001.

[86] Torgny Lindvall. Convergence of critical Galton-Watson branching processes. *J. Appl. Probability*, 9:445–450, 1972.

[87] R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58(03):497–520, 1962.

[88] V. A. Malyshev and M. V. Menshikov. Ergodicity, continuity and analyticity of countable Markov chains. *Transactions of the Moscow Mathematical Society*, (1):1–48, 1981.

[89] Olle Nerman. The stable pedigrees of critical branching populations. *J. Appl. Probab.*, 21(3):447–463, 1984.

[90] J. Neveu and J. Pitman. Renewal property of the extrema and tree property of the excursion of a one-dimensional Brownian motion. In *Séminaire de Probabilités, XXIII*, volume 1372 of *Lecture Notes in Math.*, pages 239–247. Springer, Berlin, 1989.

[91] J. Neveu and J. W. Pitman. The branching process in a Brownian excursion. In *Séminaire de Probabilités, XXIII*, volume 1372 of *Lecture Notes in Math.*, pages 248–257. Springer, Berlin, 1989.

[92] Jian Ni, Bo Tan, and R. Srikant. Q-CSMA: Queue-Length-Based CSMA/CA Algorithms for Achieving Maximum Throughput and Low Delay in Wireless Networks. *IEEE/ACM Trans. Netw.*, 20(3):825–836, June 2012.

[93] Edwin Perkins. Weak invariance principles for local time. *Z. Wahrsch. Verw. Gebiete*, 60(4):437–451, 1982.

[94] Shreevatsa Rajagopalan, Devavrat Shah, and Jinwoo Shin. Network adiabatic theorem: an efficient randomized protocol for contention resolution. In *Proc. ACM SIGMETRICS '09*, SIGMETRICS '09, pages 133–144, New York, NY, USA, 2009. ACM.

[95] Kavita Ramanan and Martin I. Reiman. Fluid and heavy traffic diffusion limits for a generalized processor sharing model. *Ann. Appl. Probab.*, 13(1):100–139, 2003.

[96] Daniel Ray. Sojourn times of diffusion processes. *Illinois J. Math.*, 7:615–630, 1963.

[97] Rolando Rebolledo. Sur l'existence de solutions à certains problèmes de semimartingales. *C. R. Acad. Sci. Paris Sér. A-B*, 290(18):A843–A846, 1980.

[98] Josh Reed, Amy Ward, and Dongyuan Zhan. On the generalized drift Skorokhod problem in one dimension. *J. Appl. Probab.*, 50(1):16–28, 2013.

[99] Martin I. Reiman. The heavy traffic diffusion approximation for sojourn times in Jackson networks. In *Applied probability—computer science: the interface, Vol. II (Boca Raton, Fla., 1981)*, volume 3 of *Progr. Comput. Sci.*, pages 409–421. Birkhäuser Boston, Boston, MA, 1982.

[100] Martin I. Reiman. Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9(3):441–458, 1984.

[101] Martin I. Reiman. Some diffusion approximations with state space collapse. In *Modelling and performance evaluation methodology (Paris, 1983)*, volume 60 of *Lecture Notes in Control and Inform. Sci.*, pages 209–240. Springer, Berlin, 1984.

[102] J. A. C. Resing. Polling systems and multitype branching processes. *Queueing Syst.*, 13(4):409–426, 1993.

[103] A. N. Rybko and A. L. Stolyar. On the ergodicity of random processes that describe the functioning of open queueing networks. *Problems on Information Transmission*, 28(3):3–26, 1992.

[104] S. M. Sagitov. Limit theorem for a critical branching process of general type. *Mat. Zametki*, 34(3):453–461, 1983.

[105] S. M. Sagitov. Limit behavior of general branching processes. *Mat. Zametki*, 39(1):144–155, 159, 1986.

[106] S. M. Sagitov. A multidimensional critical branching process generated by a large number of particles of a single type. *Teor. Veroyatnost. i Primenen.*, 35(1):98–109, 1990.

[107] S. M. Sagitov. General branching processes: convergence to Irzhina processes. *J. Math. Sci.*, 69(4):1199–1206, 1994. Stability problems for stochastic models (Kirillov, 1989).

[108] Serik Sagitov. Measure-branching renewal processes. *Stochastic Process. Appl.*, 52(2):293–307, 1994.

[109] Serik Sagitov. A key limit theorem for critical branching processes. *Stochastic Process. Appl.*, 56(1):87–100, 1995.

[110] Serik Sagitov. Limit skeleton for critical Crump-Mode-Jagers branching processes. In *Classical and modern branching processes (Minneapolis, MN, 1994)*, volume 84 of *IMA Vol. Math. Appl.*, pages 295–303. Springer, New York, 1997.

[111] Bhaskar Sengupta. An approximation for the sojourn-time distribution for the $GI/G/1$ processor-sharing queue. *Comm. Statist. Stochastic Models*, 8(1):35–57, 1992.

[112] D. Shah, Jinwoo Shin, and P. Tetali. Medium access using queues. In *Proc. IEEE FOCS '11*, pages 698–707, 2011.

[113] Devavrat Shah and Jinwoo Shin. Randomized scheduling algorithm for queueing networks. *Ann. Appl. Probab.*, 22(1):128–171, 2012.

[114] Devavrat Shah and Damon Wischik. Switched networks with maximum weight policies: fluid approximation and multiplicative state space collapse. *Ann. Appl. Probab.*, 22(1):70–127, 2012.

[115] Alexander L. Stolyar. Maxweight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *Ann. Appl. Probab.*, 14(1):1–53, 2004.

[116] Leandros Tassiulas and Anthony Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Trans. Automat. Control*, 37(12):1936–1948, 1992.

[117] Leandros Tassiulas and Anthony Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Trans. Inform. Theory*, 39(2):466–478, 1993.

[118] I. M. Verloop, U. Ayesta, and R. Núñez-Queija. Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing. *Oper. Res.*, 59(3):648–660, 2011.

[119] Mathilde Weill. Regenerative real trees. *Ann. Probab.*, 35(6):2091–2121, 2007.

[120] Ward Whitt. Some useful functions for functional limit theorems. *Math. Oper. Res.*, 5(1):67–85, 1980.

[121] R. J. Williams. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Syst.*, 30(1-2):27–88, 1998.

[122] Kouji Yano. Functional limit theorems for processes pieced together from excursions. *J. Math. Soc. Japan*, to appear.

[123] S. F. Yashkov. Processor-sharing queues: some progress in analysis. *Queueing Syst.*, 2(1):1–17, 1987.

[124] S. F. Yashkov. On a heavy traffic limit theorem for the $M/G/1$ processor-sharing queue. *Comm. Statist. Stochastic Models*, 9(3):467–471, 1993.

[125] Jiheng Zhang, J. G. Dai, and Bert Zwart. Law of large number limits of limited processor-sharing queues. *Math. Oper. Res.*, 34(4):937–970, 2009.

[126] Jiheng Zhang, J. G. Dai, and Bert Zwart. Diffusion limits of limited processor sharing queues. *Ann. Appl. Probab.*, 21(2):745–799, 2011.

[127] Jiheng Zhang and Bert Zwart. Steady state approximations of limited processor sharing queues in heavy traffic. *Queueing Syst.*, 60(3-4):227–246, 2008.

[128] A. P. Zwart and O. J. Boxma. Sojourn time asymptotics in the $M/G/1$ processor sharing queue. *Queueing Syst.*, 35(1-4):141–166, 2000.