Stochastic Recurrences and their Applications to the Analysis of Partition-Valued Processes

Stochastische Recurrente Betrekkingen en hun Toepassingen in de Analyse van Partitie-Waardige Processen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 6 juni 2011 des middags te 2.30 uur

door

Oleksandr Marynych

geboren op 4 augustus 1986 te Kiev, Oekraïne

Promotoren:	Prof. dr. R. Fernandez
	Prof. dr. A. Iksanov
Co-promotor:	Dr. A. Gnedin

CONTENTS

INTRODUCTION

A *linear random recurrence* is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ which satisfies the distributional equality

$$X_1 = 0, \qquad X_n \stackrel{d}{=} V_n + \sum_{r=1}^{K} A_r(n) X^{(r)}_{I^n_{(r)}}, \quad n \geq 2, \tag{1}$$

where X_n is some characteristics of a structure of size n, which is split according to some rule into K smaller structures of random sizes $I^n_{(r)} \in \{1, \ldots, n\}$, V_n is a random toll term associated with the splitting, $A_r(n) > 0$ is a random weight of substructure r and K is a fixed positive integer. For every $r = 1, \ldots, K$, the variable $X^{(r)}_k$ which corresponds to the rth structure is assumed independent of $((I^n_{(1)}, \ldots, I^n_{(K)}, A_1(n), \ldots, A_K(n), V_n))_{n\geq 2}$ and distributed like X_k, for each $k \in \mathbb{N}$. Furthermore, the sequences $(X^{(1)}_n)_{n\in\mathbb{N}}, \ldots, (X^{(K)}_n)_{n\in\mathbb{N}}$ are assumed independent.

A very special case of (1) is the recurrence

$$X_1 = 0, \quad X_n \stackrel{d}{=} 1 + X'_{I_n}, \quad n \geq 2, \tag{2}$$

where $I_n \in \{1, \ldots, n-1\}$ is an arbitrary random index, and X'_k is assumed independent of (I_n) and distributed like X_k, for each $k \in \mathbb{N} := \{1, 2, \ldots\}$. Random recurrence (2) can be naturally interpreted in terms of a Markov chain. Let $(Y_n)_{n\in\mathbb{N}_0}$ be a decreasing Markov chain with state space $\mathbb{N}$ and transition probabilities

$$\begin{aligned}
\mathbb{P}\{Y_n = j | Y_{n-1} = i\} &= \pi_{i,j} > 0, \quad 1 \leq j < i, \quad i \geq 2, \\
\mathbb{P}\{Y_n = j | Y_{n-1} = i\} &= 0, \quad j \geq i, \quad i \geq 2, \\
\mathbb{P}\{Y_n = 1 | Y_{n-1} = 1\} &= 1.
\end{aligned}$$

For every fixed $n \in \mathbb{N}$, define a random variable

$$X_n := \inf\{k \in \mathbb{N}_0 := \mathbb{N} \cup \{0\} : Y_k = 1 \text{ given } Y_0 = n\}, \tag{3}$$

which is the absorption time of the chain starting at state $Y_0 = n$. Conditioning on the size of the first jump of the chain and using the Markov property it is seen that (2) holds, where $n - I_n$ is equal to the size of the first jump. Therefore, the distribution of I_n is given by

$$\mathbb{P}\{I_n = k\} = \pi_{n,k},\ 1 \le k < n,\ n \ge 2. \tag{4}$$

In this setting, recurrences (2) were investigated in [68, 82, 123, 132],

Random recurrences (1), often in the simple form with $K = 1$, arise in diverse areas of applied probability such as random regenerative structures [12, 56, 58, 59, 60, 61, 63], random trees [32, 38, 40, 42, 74, 77, 97, 107, 111, 112], coalescent theory [14, 41, 49, 50, 55, 62, 65, 100, 101], recursive algorithms [33, 66, 75, 109, 119, 121], random walks with barrier [78, 79, 104, 105], to name but a few.

The thesis consists of two parts. In the first part (Chapter 1) a new technique is developed for the asymptotic analysis of the moments of linear random recurrences, and some general results are presented for the absorption times of non-increasing Markov chains. In the second part (Chapters 2 and 3) we establish new results for two models where distributional recurrences (1) occur: these are the stick-breaking partitions (Chapter 2) and exchangeable coalescents (Chapter 3). We describe now the contents in more detail.

Chapter 1. The method of iterative functions presented in Section 1.1 can be applied to various linear recurrences of the form

$$a_1 = a,\ \ a_n = b_n + \sum_{k=1}^{n-1} c_{n,k} a_k,\ \ \ n \ge 2, \tag{5}$$

where $(c_{n,k})_{n\ge 2,\ 1\le k<n}$ is a given triangular array, and $(b_n)_{n\ge 2}$ a given sequence of real numbers. It particular, the moments $(\mathbb{E}X_n^k)_{n\in\mathbb{N}}$ of fixed order k for solutions of (1) satisfy such recurrences. Generally speaking, the asymptotic analysis of recurrences (5) is a hard analytical problem. Nevertheless, various methods, with different ranges of applicability, have been elaborated to date.

Among the existing approaches the most popular is the *singularity analysis of generating functions* [40, 47]. The method gives a very precise infor-

mation on the asymptotic behavior whenever there is a tractable functional relation between the generating functions of the sequences involved.

The idea of the *repertoire method* proposed in [66] can be briefly described as follows. First we build up a repertoire $(b_n^{(\alpha)})_{\alpha\in A}$, where A is a finite set, of inhomogeneous terms of (5) by choosing sequences $(a_n^{(\alpha)})$ such that the sum in (5) is tractable. Then we construct a solution a_n to (1.4) with inhomogeneous term b_n as a linear combination of solutions $a_n^{(\alpha)}$, $\alpha \in A$.

Yet another method, proposed in [28] and further developed in [121], is based on the harmonic analysis and potential theory.

We suggest in this thesis a new approach which we call the method of iterative functions. The development of the method was motivated by a conjecture of Martin Möhle (2008) regarding the number of collisions X_n in the coalescent process driven by the Poisson-Dirichlet random measure. In this thesis we settle the conjecture by proving (Theorem 55) that the moments $\mathbb{E}X_n^k$ for each $k \geq 1$ behave for large n like the powers of the *log-star* function which grows extremely slow, namely slower than any iteration of the logarithm. The analogous result is also shown for the moments $\mathbb{E}\tau_n^k$ of the absorption time τ_n of the same coalescent process. This kind of exotic behavior of the moments partly explains the fact that the methods available in the literature do not suffice to tackle the problem. The number of collisions and the absorption time in the Poisson-Dirichlet coalescent are interesting, yet special instances of the recurrence (1). The suggested method is also capable to tackle many other linear recurrences (1).

Theorem 19, which is the main result of Section 1.2, provides a weak convergence result for the absorption time X_n defined by (3). A curious by-product of this result is that beta (a, b) distributions with $a > 1$ and $b > 0$ occur as the laws of exponential functionals of killed subordinators (increasing Levý processes).

Chapter 2. In a classical occupancy scheme n balls are thrown independently in an infinite array of boxes with probability p_k of hitting box $k = 1, 2, \ldots$, where $(p_k)_{k\in\mathbb{N}}$ is a fixed sequence of positive frequencies sum-

ming up to one. The quantities of traditional interest are

- K_n the number of boxes occupied by at least one of n balls,
- $K_{n,r}$ the number of boxes occupied by exactly r out of n balls,
- M_n the range of occupancy, equal to the maximal index of occupied box,
- $L_n := M_n - K_n$ the number of empty boxes within the occupancy range,
- $Z_{n,k}$ the number of balls in the kth box in the last-to-first order of boxes.

In applications 'boxes' are clusters, species, types of data, etc. The quantities in the list characterize the sampling variability, which for large n is dominantly determined by the boxes occupied by a few balls, thus determined by the way the frequencies p_k approach zero, as $k \to \infty$. The first two variables are functionals of the induced partition of n, defined as the unordered collection of positive occupancy counts.

The *Bernoulli sieve* is a version of the occupancy scheme with random frequencies

$$P_k := W_1 W_2 \cdots W_{k-1}(1 - W_k), \quad k \in \mathbb{N},$$

where $(W_k)_{k\in\mathbb{N}}$ are independent copies of a random variable W taking values in $(0, 1)$. The name derives from the following recursive construction based on i.i.d. $q_k \stackrel{d}{=} 1 - W$: at round 1 a coin with probability q_1 for heads is flipped for each of n balls and every time it turns heads the ball is put in box 1, then at round 2 a coin with probability q_2 for heads is flipped for each of the remaining balls and every time it turns heads the ball is sent to box 2, and so on until all balls are allocated in boxes.

It is useful to identify frequencies (P_k) with the lengths of component intervals induced by splitting $[0, 1]$ at points of a multiplicative renewal process (sometimes called stick-breaking) $(Q_k)_{k\in\mathbb{N}_0}$, where

$$Q_0 := 1, \quad Q_j := \prod_{i=1}^{j} W_i, \quad j \in \mathbb{N}.$$

In the spirit of Kingman's 'paintbox representation' of exchangeable partitions [58], we may identify the boxes with open intervals (Q_k, Q_{k-1}), and mark the balls by independent points $U_1, \ldots, U_n$ sampled from the uniform $[0, 1]$ distribution, independently of (Q_k). The event $U_i \in (Q_{k-1}, Q_k)$ then means that ball i falls in box k. Keep in mind that in the natural order the intervals are indexed from the right to the left, thus the occupancy range is determined by the interval containing the leftmost mark $\min(U_1, \ldots, U_n)$. In view of this construction a random partition of n induced by the Bernoulli sieve can be called *stick-breaking partition*.

Theorem 23 which is the main result of Section 2 provides an ultimate result concerning the weak convergence of the number of occupied boxes in the Bernoulli sieve. To prove it we use a new approach which is based on the analysis of small frequencies P_k's and relies heavily upon certain weak convergence results for *perturbed random walks*. The latter topic is explored in Section 2.4 and may be of independent interest. Other results of Section 2 includes: the asymptotics of the first moment of the number L_n of empty boxes in the occupancy range of the Bernoulli sieve (Theorem 33), explicit determination of the law of L_n in a particular case (Proposition 34) and convergence of small parts to zero in probability under an infinite mean assumption (Proposition 38).

Chapter 3. Exchangeable *coalescents* have become a powerful tool for applications in the population genetics, and more precisely for analyzing the genealogy of DNA sequences. The coalescent process is a natural extension of the classical population genetics concept of neutral evolution and it appears as an approximation to the Fisher-Wright model when the population size is large. The basic idea of the coalescent (or more precisely n-coalescent), which goes back to J. F. C. Kingman [88, 89], can be briefly described as follows. We select n individuals from the present generation in a large population, and trace backwards in time their genealogical history. For every two individuals there will be an epoch when two lineages coalesce and their *most recent common ancestor* is encountered. Continuing this process backwards in time

we construct the coalescent tree, the root of this tree being the most recent common ancestor for the whole sample. The shape of coalescent tree depends heavily on the rate of coalescence of the lineages. In the simplest instance, when only binary collisions are possible, we obtain the *Kingman coalescent*. If multiple collisions are allowed, the corresponding coalescent is called the *lambda-coalescent*.

There are several important functionals of the coalescents which have become central objects of research in the past few years:

- X_n the number of collisions in the n-coalescent tree;
- τ_n the absorption time (the height of the n-coalescent tree or the time back to the most recent common ancestor);
- L_n the total branch length of the n-coalescent, which is the sum of lengths of all branches in the n-coalescent tree.

These functionals are particular examples of linear random recurrences (1) since the following distributional equalities hold

$$X_1 = 0, \quad X_n \stackrel{d}{=} 1 + X'_{I_n}, \quad n \geq 2; \tag{6}$$

$$\tau_1 = 0, \quad \tau_n \stackrel{d}{=} T_n + \tau'_{I_n}, \quad n \geq 2; \tag{7}$$

$$L_1 = 0, \quad L_n \stackrel{d}{=} nT_n + L_{I_n}, \quad n \geq 2, \tag{8}$$

where I_n denotes the number of remaining blocks after the first collision, and T_n is the time of the first collision; X'_k (respectively, τ'_k, L'_k) is assumed independent of I_n (respectively, (T_n, I_n)) and distributed like X_k (respectively, τ_k, L_k), for each $k \in \mathbb{N}$.

In Section 3.1 we consider the lambda-coalescents, also known as *coalescents with multiple collisions*, with positive frequency of singleton clusters. The class in focus covers, for instance, the beta (a, b)-coalescents with $a > 1$. We show that some large-sample properties of these processes can be derived by coupling with an increasing Lévy process (subordinator), and by exploiting parallels with the theory of regenerative composition structures.

In particular, we discuss the limiting distributions of the absorption time and the number of collisions. In Section 3.2 we provide expansions for the moments of the number of collisions in the beta $(2, b)$ - coalescents. Also we establish the strong law of large numbers and a central limit theorem for this functional. Finally in Section 3.3 we investigate the asymptotics of the functionals acting on the Poisson-Dirichlet coalescent. An interesting consequence of our results is that the asymptotics of X_n and τ_n are determined by a generalized 'log-star' function. A remarkable feature of this function is its extremely slow growth.

Chapter 1

Moments of random recurrences

The chapter is divided in two independent parts. The purpose of Section 1.1 is to propose a new method of obtaining the first-order asymptotics of the moments of linear random recurrences. In Section 1.2 we prove a result about the weak convergence of the absorption times of certain decreasing Markov chains, the main technical tool being the *method of moments*.

1.1 Method of iterative functions

Denote by $C^{(m)}(B)$ the space of functions which are m-times differentiable on the set B. If $B = [a, \infty)$ then the derivative at point a is assumed to be the right derivative. We use the notation

$$r^{\circ(0)}(x) \stackrel{def}{=} x, \;\; r^{\circ(k)}(x) \stackrel{def}{=} r(r^{\circ(k-1)}(x)), \;\; k \in \mathbb{N}.$$

Finally, we recall the standard notation $\lfloor x \rfloor = \sup\{k \in \mathbb{Z} : k \leq x\}$ and $\lceil x \rceil = \inf\{k \in \mathbb{Z} : k \geq x\}$ for the floor and ceiling function, respectively.

1.1.1 Definitions and basic properties.

This subsection introduces *iterative functions* and investigates some of their basic properties.

Definition 1. Suppose that the function $g : \mathbb{R}^+ \to \mathbb{R}^+$ is increasing, unbounded and continuous, and satisfies the following condition: for some $x_0 > 0$ and every $x_1 > x_0$ there exists $\varepsilon_{x_1} > 0$ such that

$$x - g(x) > \varepsilon_{x_1} \quad \text{for all} \quad x \in (x_0, x_1). \tag{1.1}$$

Assuming that $h : \mathbb{R}^+ \to \mathbb{R}^+$ and $k : [0, x_0] \to \mathbb{R}$ are continuous functions, define the function $g^* : \mathbb{R}^+ \to \mathbb{R}$ by the following equality

$$g^*(x) = \sum_{i=1}^{m_0(x)} h(g^{\circ(i-1)}(x)) + k(g^{\circ(m_0(x))}(x)), \tag{1.2}$$

where

$$m_0(x) := \inf\{k \geq 0 : g^{\circ(k)}(x) \leq x_0\}.$$

We call g^* the *iterative function generated by the quadruple* (h, g, x_0, k) and denote it by $g^* = \mathrm{Iter}(h, g, x_0, k)$.

Note that technical condition (1.1) is sufficient for $m_0(x)$ to be finite for every $x \in \mathbb{R}^+$. This follows from the estimate $m_0(x) \leq \lfloor \frac{x - x_0}{\varepsilon_x} \rfloor + 1$ which is implied by the inequality

$$x - k\varepsilon_x > g^{\circ(k)}(x), \quad x > x_0, \quad k = 0, \ldots, m_0(x),$$

which can be obtained by induction.

Remark 2. From the definition it follows that g^* satisfies the functional equation

$$g^*(x) \;=\; h(x) + g^*(g(x)),\ x > x_0, \tag{1.3}$$

with initial condition

$$g^*(x) \;=\; k(x), \qquad\qquad x \leq x_0.$$

Note that the last equality and (1.3) characterize g^*. Below are some examples of iterative functions.

Example 3. Let $h(x) \equiv 1$, $g(x) = \alpha x$, $\alpha \in [0,1)$, $x_0 = 1$, $k(x) \equiv 0$. Then $g^*(x) = 1+g^*(\alpha x)$, for $x > 1$, and $g^*(x) = 0$, for $x \in [0,1]$. The corresponding solution is $g^*(x) = \lceil \log_{\frac{1}{\alpha}} x \rceil$, $x > 1$.

By an *elementary function* we mean a function constructed from 'basic' functions (constants $x \mapsto c \in \mathbb{R}$; powers $x \mapsto x^\alpha$, $\alpha \in \mathbb{R}$; exponentials $x \mapsto a^x$, $a > 0$; logarithms $x \mapsto \log_a x$, $a > 0, a \neq 1$; trigonometric functions $x \mapsto \sin x$, $x \mapsto \cos x$, $x \mapsto \tan x$, $x \mapsto \cot x$ and their inverses) with the aid of finitely many elementary operations $(+, -, \times, \div)$ and compositions.

The next example is more general.

Example 4. Let $f(\cdot)$ be an arbitrary elementary function which is continuous on $[0, +\infty)$ and assumed to be unbounded and strictly increasing on $[x_0, +\infty)$, for some $x_0 > 0$. From the obvious equality

$$f(x) = 1 + f(f^{-1}((f(x) - 1)), \quad x > f^{-1}(x_0 + 1),$$

it follows that the function $f(\cdot)$ is iterative and that it is generated by the quadruple $(1, f^{-1}((f(x) - 1), x_0, f(x))$.

Example 5. Let $h(x) \equiv 1$, $g(x) = \log x$, $x_0 = 1$, $k(x) \equiv 0$. Then $g^*(x) = 1 + g^*(\log x)$, $x > 1$, or

$$g^*(x) = \log^* x,$$

the *log-star function* which is arguably the best known non-trivial iterative function. It is clear that $\mathrm{Iter}(1, g, x_0, 0) = m_0(x)$. In particular, this equality holds for the log-star function.

If $h(x_0) \neq 0$ then the iterative functions $\mathrm{Iter}(h, g, x_0, 0)$ are piecewise continuous. However we prefer to work with smooth iterative functions which was the main reason for introducing functions k in Definition 1. It turns out that $\mathrm{Iter}(h, g, x_0, 0)$ and $\mathrm{Iter}(h, g, x_0, k)$ are asymptotically equivalent, and an appropriate choice of k makes $\mathrm{Iter}(h, g, x_0, k)$ smooth enough. Below we formalize this statement and also describe how the mentioned smoothness can be obtained by the choice of k.

Introduce the equivalence relation $\approx$ on the set of iterative functions by the rule

$$g_1^* \approx g_2^* \Longleftrightarrow g_1^* = \mathrm{Iter}(h, g, x_0, k_1), \quad g_2^* = \mathrm{Iter}(h, g, x_0, k_2).$$

This relation induces partitioning of the set of iterative functions into the classes of equivalence.

Definition 6. The equivalence class

$$\mathcal{F} := \{F = \mathrm{Iter}(h, g, x_0, k), k \in C[0, x_0]\}$$

is called the *iterative function generated by the triple* (h, g, x_0). When it does not lead to ambiguity, we call an *iterative function generated by the triple* (h, g, x_0) an arbitrary element of this class.

Since $|g_1^*(x) - g_2^*(x)|$ is bounded on $\mathbb{R}^+$, for any $g_1^*, g_2^* \in \mathcal{F}$, all iterative function in the same equivalence class are asymptotically equivalent (provided they diverge to $+\infty$).

Definition 7. An *m-times differentiable modification* of iterative function g^* is an arbitrary iterative function $\hat{g}^*$ such that $\hat{g}^* \approx g^*$ and $\hat{g}^* \in C^{(m)}[x_0, +\infty)$.

Our first result which is a direct consequence of Lemma 16 and Lemma 17 given below shows that provided g and h are smooth enough one can find a function k such that the function $\mathrm{Iter}(h, g, x_0, k)$ is smooth. For a collection of functions $f_1, \ldots, f_n$ let $W(f_1, \ldots, f_n)$ denote its Wronskian.

Theorem 8. *Assume that* $g, h \in C^{(m)}[x_0, +\infty)$ *and that*

$$W\Big(x^i - g^i(x), i = 0, \ldots, m+1\Big)(x_0) \neq 0.$$

Then there exists a function k *of the form*

$$k(x) = \sum_{i=1}^{m+1} \alpha_i x^i,$$

such that the iterative function generated by the quadruple (h, g, x_0, k) *is* m*-times differentiable on* $[x_0, +\infty)$.

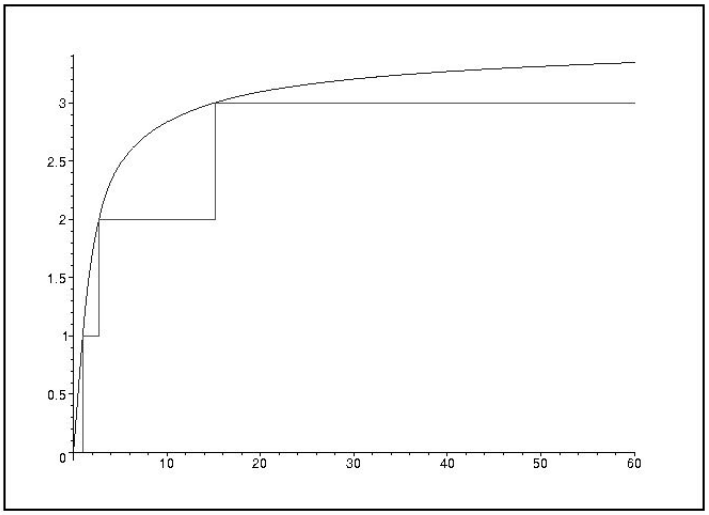
Remark 9. The vector of coefficients $(\alpha_1, \alpha_2, \ldots, \alpha_{m+1})$ is a solution to the system of linear equations (see Lemma 17) and can be calculated explicitly.

An example of a smoothed iterative function is given below.

Example 10. Recall that the log-star function is an iterative function generated by the quadruple $(1, \log x, 1, 0)$. A twice differentiable modification F of the log-star function can be constructed as follows. According to Lemma 17, the corresponding function k takes the form $k(x) = -\frac{2}{13}x^3 + \frac{3}{13}x^2 + \frac{12}{13}x$. Therefore

$$F(x) = \begin{cases} 1 + F(\log x), & x > 1, \\ -\frac{2}{13}x^3 + \frac{3}{13}x^2 + \frac{12}{13}x, x \in [0, 1]. \end{cases}$$

The graphs of functions $\log^* x$ and $F(x)$ for $x > 0$ are depicted below.



1.1.2 Asymptotics of moments. The first step of asymptotic analysis of recurrences (1) is to find the asymptotics of moments $\mathbb{E}X_n^k$ and central moments $\mathbb{E}(X_n - \mathbb{E}X_n)^k$, as $n \to \infty$.

This problem reduces to studying the recurrence equations of the form

$$a_1 = 0, \quad a_n = b_n + \sum_{k=1}^{n-1} c_{n,k} a_k, \ n \geq 2, \tag{1.4}$$

where $(b_n)_{n\in\mathbb{N}}$ and $(c_{n,k})_{n\in\mathbb{N},k<n}$ are given numeric sequences.

In the sequel, unless stated the contrary, we tacitly suppose that $b_n \geq 0$ and, hence, $a_n \geq 0$. However, a perusal of the proofs given below reveals that we could have assumed that b_n is only non-negative or non-positive for large enough n. Under this last assumption, formulations of results would get cumbersome which has forced us to keep less generality but more transparency.

While investigating recurrence (1.4), without loss of generality, we can assume that for every $n \geq 2$

$$\sum_{k=1}^{n-1} c_{n,k} = 1 \text{ and } c_{n,k} \geq 0,\ k = 1, \ldots, n-1 \tag{1.5}$$

(see, for instance, [121, p. 9]). Recurrences (1.4), with $b_n \geq 0$, which satisfy (1.5) will be referred to as *recurrences with weights reduced to probabilities*. If (1.5) holds, denote by I_n a random variable with distribution

$$\mathbb{P}\{I_n = k\} = c_{n,k}, \quad k = 1, \ldots, n-1.$$

Theorem 11. *Assume that the sequence $(a_n)_{n\in\mathbb{N}}$ satisfies recurrence (1.4) with weights reduced to probabilities. Let $g : \mathbb{R}^+ \to \mathbb{R}^+$ be continuous, increasing and unbounded function such that*

$$g(n) = \mathbb{E}I_n + o(\mathbb{E}I_n), \quad n \to \infty^1,$$

and $h : \mathbb{R}^+ \to \mathbb{R}^+$ a continuous function such that

$$h(n) \sim b_n, \quad n \to \infty.$$

If

- $\lim\limits_{n\to\infty} a_n = +\infty$,
- $g^*(\mathbb{E}I_n) - g^*(g(n)) = o(h(n))$, $n \to \infty$,

[1] Since $I_n \leq n-1$, it is always possible to choose g such that (1.1) holds.

where g^* is an iterative function generated by the triple (h, g, x_0) then the following implications are true

$$\mathbb{E}g^*(I_n) - g^*(\mathbb{E}I_n) = o(h(n)), \quad n \to \infty \Longrightarrow \tag{1.6}$$
$$a_n \sim g^*(n), \quad n \to \infty,$$

$$\mathbb{E}g^*(I_n) - g^*(\mathbb{E}I_n) \sim h(n)d, \quad n \to \infty, \quad \textit{for some } d < 1 \Longrightarrow \tag{1.7}$$
$$a_n \sim (1-d)^{-1}g^*(n), \quad n \to \infty.$$

Proof. Set $a'_n := a_n - g^*(n)$, $n \in \mathbb{N}$. The sequence (a'_n) satisfies the recurrence

$$a'_1 = -g^*(1), \;\; a'_n = b_n - g^*(n) + \mathbb{E}g^*(I_n) + \sum_{k=1}^{n-1} c_{n,k}a'_k, \quad n \geq 2. \tag{1.8}$$

If $\mathbb{E}g^*(I_n) - g^*(\mathbb{E}I_n) = o(h(n))$ and $g^*(\mathbb{E}I_n) - g^*(g(n)) = o(h(n))$ then the inhomogeneous term of (1.8) is $o(h(n))$. Therefore, applying Theorem 13 (II) yields $a'_n = o(a_n)$ which implies $a_n \sim g^*(n)$.

If $\mathbb{E}g^*(I_n) - g^*(\mathbb{E}I_n) \sim h(n)d$, for some $d \in (0,1)$, and $g^*(\mathbb{E}I_n) - g^*(g(n)) = o(h(n))$ then the inhomogeneous term of (1.8) is asymptotically equal to $h(n)d$. Therefore, applying Theorem 13 (I) yields $a'_n \sim a_n d$ which implies $a_n \sim (1-d)^{-1}g^*(n)$.

Finally, if $\mathbb{E}g^*(I_n) - g^*(\mathbb{E}I_n) \sim h(n)d$, for some $d < 0$, we can apply Theorem 13 (II) to the sequences $(g^*(n) - a_n)$ and (a_n) to conclude that $g^*(n) - a_n \sim -a_n d$. The latter is equivalent to $a_n \sim (1-d)^{-1}g^*(n)$. The proof is complete. □

Theorem 12. *Assume that the sequence $(a_n)_{n\in\mathbb{N}}$ satisfies recurrence* (1.4) *with weights reduced to probabilities. Let $g : \mathbb{R}^+ \to \mathbb{R}^+$ be a twice differentiable, increasing and unbounded function such that*

$$g(n) = \mathbb{E}I_n + o(\mathbb{E}I_n), \quad n \to \infty,$$

and $h : \mathbb{R}^+ \to \mathbb{R}^+$ a twice differentiable function such that

$$h(n) \sim b_n, \quad n \to \infty.$$

If the following conditions hold

(C1) $\lim\limits_{n\to\infty} a_n = +\infty$;

(C2) *There exists continuous function* k *such that the iterative function* F *generated by the quadruple* (h, g, x_0, k) *is twice differentiable;*

(C3) $F(\mathbb{E}I_n) - F(g(n)) = o(h(n))$, $n \to \infty$;

(C4) *There exists* $M > 0$ *such that for all* $n \in \mathbb{N}$

$$\operatorname{Var} I_n \le M\mathbb{E}I_n;$$

(C5) $\lim\limits_{n\to\infty} \sup_{x\ge\mathbb{E}I_n/2} |F''(x)| \dfrac{\operatorname{Var} I_n}{h(n)} = 0$;

(C6) $\lim\limits_{n\to\infty} \dfrac{\sup_{1\le x\le n} |F(x)|}{h(n)\operatorname{Var} I_n} = 0$;

(C7) $\lim\limits_{n\to\infty} \dfrac{F'(\mathbb{E}I_n)}{h(n)} = 0$,

then

$$a_n \sim F(n),\ n \to \infty.$$

Proof. Since conditions (C1) and (C3) hold, according to implication (1.6) in Theorem 11, it is enough to show that

$$\alpha_n := \mathbb{E}F(I_n) - F(\mathbb{E}I_n) = o(h(n)).$$

With $\kappa := \frac{1}{2M}$ and $A_n := \{|I_n - \mathbb{E}I_n| > \kappa \operatorname{Var} I_n\}$ we have

$$|\alpha_n| \le |\mathbb{E}(F(I_n) - F(\mathbb{E}I_n))1_{A_n}| + |\mathbb{E}(F(I_n) - F(\mathbb{E}I_n))1_{A_n^c}| =: \beta_n + \gamma_n.$$

An application of Chebyshev's inequality yields

$$\beta_n \le 2 \sup_{1\le x\le n} |F(x)|\mathbb{P}(A_n) \le \frac{2\operatorname{Var} I_n \sup_{1\le x\le n} |F(x)|}{(\kappa \operatorname{Var} I_n)^2},$$

which is $o(h(n))$ by condition (C6).

Using the Taylor expansion around $\mathbb{E}I_n$ leads to

$$\begin{aligned}\gamma_n &= \left|\mathbb{E}\Big(F'(\mathbb{E}I_n)(I_n-\mathbb{E}I_n)+\frac{1}{2}F''(\theta_n)(I_n-\mathbb{E}I_n)^2\Big)1_{A_n^c}\right| \\ &\le \left|F'(\mathbb{E}I_n)\mathbb{E}(I_n-\mathbb{E}I_n)1_{A_n}\right|+\frac{1}{2}\left|\mathbb{E}F''(\theta_n)(I_n-\mathbb{E}I_n)^2 1_{A_n^c}\right| = \gamma_{1,n}+\gamma_{2,n},\end{aligned}$$

where at the last line the equality $|\mathbb{E}(I_n-\mathbb{E}I_n)1_{A_n}| = |\mathbb{E}(I_n-\mathbb{E}I_n)1_{A_n^c}|$ has been utilized and $\theta_n \in [\mathbb{E}I_n-\kappa\mathrm{Var}\, I_n, \mathbb{E}I_n+\kappa\mathrm{Var}\, I_n]$. Consequently, by Cauchy-Schwarz and Chebyshev's inequalities, we obtain

$$\begin{aligned}\gamma_{1,n} &= |F'(\mathbb{E}I_n)\mathbb{E}(I_n-\mathbb{E}I_n)1_{A_n}| \le |F'(\mathbb{E}I_n)|\sqrt{\mathbb{E}(I_n-\mathbb{E}I_n)^2}\sqrt{\mathbb{P}(A_n)} \\ &\le |F'(\mathbb{E}I_n)|\sqrt{\mathrm{Var}\, I_n}\sqrt{\frac{\mathrm{Var}\, I_n}{(\kappa\mathrm{Var}\, I_n)^2}} = \frac{1}{\kappa}|F'(\mathbb{E}I_n)|,\end{aligned}$$

which is $o(h(n))$ by condition (C7).

Finally, an appeal to condition (C4) allows us to conclude that

$$\gamma_{2,n} \le \frac{1}{2}\sup_{x\ge \mathbb{E}I_n/2}|F''(x)|\mathrm{Var}\, I_n,$$

which is $o(h(n))$ by condition (C5). The proof is complete. □

Based on Theorem 11 and Theorem 12 we formulate the following algorithm which can be used to derive the asymptotic behavior of the moments of linear random recurrences.

Algorithm

1. Make a reduction to probabilities using, for example, the method described in [121]. As a result, we obtain the recurrence of the form

$$A_1 = 0, \quad A_n = B_n + \sum_{k=1}^{n-1} p_{n,k}A_k,$$

where $\sum_{k=1}^{n-1} p_{n,k} = 1$ for all $n \ge 2$ and $B_n \ge 0$. Let I_n be a random variable with distribution $\mathbb{P}\{I_n = k\} = p_{n,k}, n \ge 2, k < n$.

2. Prove the divergence of A_n using, for example, Proposition 14 or other methods.

3. Find a continuous, strictly increasing and unbounded function $g(x)$ defined on $\mathbb{R}^+$, such that $g(n) = \mathbb{E}I_n + o(\mathbb{E}I_n)$. Pick an x_0 as defined in (1.1). Find a continuous function $h(x)$ defined on $\mathbb{R}^+$ such that $h(n) \sim B_n$.

4. Find an iterative function g^* generated by the quadruple (h, g, x_0, k), where k is any continuous function on $[0, x_0]$.

5. If g^* is an elementary function go to the next step, otherwise, select k such that g^* is twice differentiable (see Theorem 8) and go to the next step.

6. If $g^*(\mathbb{E}I_n) - g^*(g(n)) = o(h(n))$ then go to the next step, otherwise go to step 3 and choose asymptotically smaller term $o(\mathbb{E}I_n)$.

7. If $\mathbb{E}g^*(I_n) - g^*(\mathbb{E}I_n) = o(h(n))$ (this can be checked using, for example, Theorem 12) then $A_n \sim g^*(n)$. If $\mathbb{E}g^*(I_n) - g^*(\mathbb{E}I_n) \sim h(n)d$ then $A_n \sim (1-d)^{-1}g^*(n)$.

1.1.3 Applications.

Number of collisions in beta$(a, 1)$-coalescents.

Let X_n be the number of collisions in beta$(a, 1)$-coalescent, $a > 0$, restricted to the set $\{1, \dots, n\}$ (see Chapter 3 for the definition and more details). Many results concerning the asymptotics of $\mathbb{E}X_n^k$, $k \in \mathbb{N}$, are known [36, 41, 42, 62, 76, 97, 111, 112] but we partly derive them again just in order to show how our method works.

It is known that $(X_n)_{n\in\mathbb{N}_0}$ satisfy recurrence (2) with I_n having distribution

$$\mathbb{P}\{I_n = n-k\} = \frac{\frac{(2-a)\Gamma(a+k-1)}{\Gamma(a)\Gamma(k+2)}}{1-\frac{\Gamma(a+n-1)}{\Gamma(a)\Gamma(n+1)}}, \quad k = 1, \dots, n-1, \quad n \geq 2,$$

if $a \neq 2$, and

$$\mathbb{P}\{I_n = n-k\} = \frac{1}{(h_n - 1)(k+1)}, \quad k = 1, \dots, n-1, \quad n \geq 2,$$

where $h_n = \sum_{k=1}^{n} k^{-1}$, if $a = 2$.

From these formulae we deduce the divergence of I_n to ∞ in probability which implies, by Proposition 14, that $\lim_{n\to\infty} \mathbb{E}X_n = +\infty$. It is also clear that no reduction of weights to probabilities is needed.

Case $0 < a < 1$ [36, 62, 78]. Since

$$\mathbb{E}I_n = n - (1-a)^{-1} + o(1),$$

we can choose

$$g(x) = x - \frac{1}{1-a} \quad \text{and} \quad h(x) = 1.$$

Then functional equation (1.3) has an elementary solution $g^*(x) = (1-a)x$. By Theorem 11, $\mathbb{E}X_n \sim g^*(n) \sim (1-a)n$, $n \to \infty$.

Case $a = 1$ (Bolthausen-Sznitman coalescent) [41, 78, 97, 111, 112]. Since

$$\mathbb{E}I_n = n - \log n + O(1), \quad \operatorname{Var} I_n = O(n)$$

we can choose

$$g(x) = x - \log x \quad \text{and} \quad h(x) = 1.$$

The relation

$$\frac{x}{\log x} = 1 + o(1) + \frac{x - \log x}{\log(x - \log x)}, \quad x \to \infty,$$

implies that $x \mapsto x/\log x$ is an iterative function generated by the quadruple $(1 + o(1), x - \log x, 2, x/\log x)$. An application of Theorem 12 with $F(x) = x/\log x$ gives[2] $\mathbb{E}X_n \sim \frac{n}{\log n}$, $n \to \infty$.

Case $a = 2$ (Theorem 46). From the expression for $\mathbb{P}\{I_n = k\}$, when $a = 2$, and the known asymptotics $h_n = \log n + \text{const} + O(1/n)$, $n \to \infty$, we deduce

$$\mathbb{E}I_n = \frac{1}{h_n - 1} \sum_{k=1}^{n-1} \frac{n-k}{k+1} = n - \frac{n}{\log n} + O\left(\frac{n}{\log^2 n}\right).$$

[2]The only thing which may require verification is condition $C3$. In the present situation, $\mathbb{E}I_n - g(n) = O(1)$, and the derivative of $F(x) = x/\log x$ tends to zero, as $x \to \infty$. Therefore, condition $C3$ follows by application of the mean value theorem.

Therefore, we can choose

$$g(x) = \left(x - \frac{x}{\log x}\right) 1_{(e,\infty)}(x) \quad \text{and} \quad h(x) = 1.$$

From the relation

$$\log^2 x = 2 + o(1) + \log^2(x - \frac{x}{\log x}), \quad x \to \infty,$$

it follows that $x \mapsto 2^{-1} \log^2 x$ is an iterative function generated by the quadruple $(1 + o(1), x - \frac{x}{\log x}, 2, 2^{-1} \log^2 x)$. Applying the mean value theorem to the differentiable function $x \mapsto \log^2 x$ we obtain

$$\log^2 \mathbb{E}I_n - \log^2 g(n) = O\left(\frac{1}{\log n}\right)$$

and

$$\lim_{n\to\infty} \frac{1}{2}\Big(\mathbb{E} \log^2 I_n - \log^2 \mathbb{E}I_n\Big) = 1 - \frac{\pi^2}{6},$$

which, in view of implication (1.7), yields $\mathbb{E}X_n \sim \frac{3}{\pi^2} \log^2 n$, $n \to \infty$.

Examples from the analysis of algorithms.

By using our method we give new proofs of the results from [109],[94] and [119], respectively.

The Quickselect algorithm. Let X_n be the number of comparisons that the Quickselect algorithm needs to find $\min(x_1, \ldots, x_n)$ of a sample $x_1, \ldots, x_n$. Then

$$X_1 = 0, \quad X_n \stackrel{d}{=} n - 1 + X'_{I_n}, \quad n \geq 2,$$

where $I_n = J_n \vee 1$, and J_n is uniformly distributed on $\{0, \ldots, n-1\}$. Since

$$\mathbb{E}I_n = \frac{n-1}{2} + \frac{1}{n},$$

we can choose

$$g(x) = \frac{x+1}{2} \quad \text{and} \quad h(x) = x - 1.$$

Then functional equation (1.3) has elementary solutions $g^*(x) = 2x + c$, $c \in \mathbb{R}$. By Theorem 11, $\mathbb{E}X_n \sim g^*(n) \sim 2n$, $n \to \infty$.

The depth of random node in a random binary search tree. The corresponding recurrence is

$$X_0 = -1,\ X_1 = 0,\quad X_n \stackrel{d}{=} 1 + X'_{I_n},\quad n \geq 2,$$

where $\mathbb{P}\{I_n = k\} = 2k/n^2$, for $k \in \{1, \ldots, n-1\}$ and $\mathbb{P}\{I_n = 0\} = 1/n$. Since

$$\mathbb{E}I_n = \frac{(n-1)(2n-1)}{3n},$$

we can choose

$$g(x) = 2x/3 \ \text{ and } \ h(x) = 1.$$

According to Example 1, the corresponding iterative function is

$$g^*(x) = \lceil \log_{\frac{3}{2}} x \rceil,\quad x > 1.$$

Since $\lim\limits_{n\to\infty} (\mathbb{E}\log^+ I_n - \log n) = -1/2$, it follows that $\lim\limits_{n\to\infty} (\mathbb{E}f(I_n) - f(\mathbb{E}I_n)) = 1 - \frac{1}{2\log(3/2)}$, where

$$f(x) = \frac{\log^+ x}{\log(3/2)}.$$

Finally, in view of $f(x) \sim g^*(x)$ we have $\mathbb{E}X_n \sim 2\log(3/2) f(n) \sim 2\log n$, $n \to \infty$, according to Theorem 11.

The Quicksort algorithm. Let X_n denote the random number of comparisons needed to sort a list of length n by the Quicksort. Then

$$X_0 = X_1 = 0 \ \text{ and } \ X_n \stackrel{d}{=} n - 1 + X'_{I_n - 1} + X''_{n - I_n},\quad n \geq 2,$$

where I_n has the uniform distribution on $\{1, \ldots, n\}$, and X'_k and X''_k are assumed independent of (I_n) and distributed like X_k, for each $k \in \mathbb{N}_0$.

Set $a_n := \mathbb{E}X_n$, then

$$a_0 = a_1 = 0 \ \text{ and } \ a_n = n - 1 + \sum_{k=0}^{n-1} \frac{2}{n} a_k,\quad n \geq 2.$$

The reduction of weights to probabilities can be made by the substitution $a'_n := a_n/(n+1)$ which yields

$$a'_n = \frac{n-1}{n+1} + \sum_{k=0}^{n-1} \frac{2(k+1)}{n(n+1)} a'_k,\quad n \geq 2.$$

Using the same arguments as in the previous example we obtain $a'_n \sim 2\log n$. Therefore, $\mathbb{E}X_n \sim 2n\log n$, $n \to \infty$, as is well-known[3].

Limitations of the method.

The method of iterative functions is not universal, it has some limitations which we would like to emphasize on.

(a) An indispensable requirement of our method to work is the divergence of (a_n), the solution to (1.4). In particular, our method cannot detect the convergence of a_n to a constant.

(b) It may be difficult to guess which elementary function has the same asymptotics as given iterative function.

(c) If condition (1.7) holds, for some $d \neq 0$, it may be hard to calculate the constant d explicitly. Therefore, it seems that a natural assumption for the method to work is (1.6) rather than (1.7). Condition (1.6) holds if the solution is nearly linear and the variance of index I_n grows not too fast (precise statements are made in Theorem 12). For instance, the expected number of collisions in the Bolthausen-Sznitman and Poisson-Dirichlet coalescents (see Theorem 55 in Chapter 3) exhibit the asymptotic behavior of this type.

1.1.4 Auxiliary results.

Properties of recurrences (1.4)**.**

Theorem 13. *Suppose that* $(a_n)_{n\in\mathbb{N}}$ *and* $(a'_n)_{n\in\mathbb{N}}$ *satisfy the recurrences*

$$a_n = b_n + \sum_{k=1}^{n-1} p_{n,k}a_k, \ n \geq N, \tag{1.9}$$

[3]The first result concerning the complexity of (non-randomized) Quicksort algorithm with $O(n\log n)$ asymptotic goes back to the pioneering work by Hoare [72]. For complete analysis of Quicksort and its different modifications we refer to a survey [129].

and

$$a'_n = b'_n + \sum_{k=1}^{n-1} p_{n,k} a'_k, \ n \geq N, \tag{1.10}$$

respectively. Suppose that $b_n \geq 0$, *for* $n \geq N$, *and* $\lim\limits_{n\to\infty} a_n = +\infty$. *Then*

I. $b'_n \sim b_n$, $n \to \infty$ *implies* $a'_n \sim a_n$, $n \to \infty$, *and*

II. $b'_n = o(b_n)$, $n \to \infty$ *implies* $a'_n = o(a_n)$, $n \to \infty$.

Proof of (I). We exploit the idea of proof of [63, Proposition 3]. Suppose there exists $\varepsilon_0 > 0$ such that $a_n > (1+\varepsilon_0)a'_n$ for infinitely many n. Since $\lim\limits_{n\to\infty} a_n = +\infty$, we can pick $\varepsilon \in (0, \varepsilon_0]$ such that for any $c > 0$ the inequality $a_n > (1+\varepsilon)a'_n + c$ holds for infinitely many n. Let n_c be minimal such n. Since $\lim\limits_{c\to\infty} n_c = +\infty$, without loss of generality we can assume that $n_c > N$. For $n \leq n_c - 1$, we have $a_n \leq (1+\varepsilon)a'_n + c$ which implies

$$(1+\varepsilon)a'_{n_c} + c < a_{n_c} = b_{n_c} + \sum_{k=1}^{n_c-1} p_{n_c,k} a_k \leq b_{n_c} + c + (1+\varepsilon) \sum_{k=1}^{n_c-1} p_{n_c,k} a'_k.$$

Simplifying the last expression gives $1 + \varepsilon < b_{n_c}/b'_{n_c}$. Sending $c \to \infty$ leads to $\varepsilon < 0$, which is a contradiction. Thus we have proved that

$$\limsup_{n\to\infty} \frac{a_n}{a'_n} \leq 1.$$

A symmetric argument proves the converse inequality for the lower limit.
Proof of (II) proceeds by applying the already established part (I) to the sequences (a_n) and $(a_n - a'_n)$ and noting that the relation $b_n \sim b_n - b'_n$ implies $a_n \sim a_n - a'_n$. The proof is complete. □

A simple sufficient condition for $\lim\limits_{n\to\infty} a_n = +\infty$ is given below.

Proposition 14. *Assume that the sequence* $(a_n)_{n\in\mathbb{N}}$ *satisfies (1.4) with weights reduced to probabilities. If* $I_n \xrightarrow{P} \infty$ *and* $\liminf\limits_{n\to\infty} b_n = b \in (0, \infty]$, *then* $\lim\limits_{n\to\infty} a_n = +\infty$.

Proof. From recurrence (1.4) we obtain

$$\begin{aligned} a_n &= b_n + \sum_{k=1}^{n-1} p_{n,k} a_k = b_n + \sum_{k=1}^{M-1} p_{n,k} a_k + \sum_{k=M}^{n-1} p_{n,k} a_k \\ &\geq b_n + \Big(\inf_{1 \leq k < M} a_k \Big) \sum_{k=1}^{M-1} p_{n,k} + \Big(\inf_{M \leq k \leq n-1} a_k \Big) \sum_{k=M}^{n-1} p_{n,k}. \end{aligned}$$

Sending $n \to \infty$ gives $\liminf_{n\to\infty} a_n \geq b + \inf_{k \geq M} a_k$. Letting $M \to \infty$ leads to $\liminf_{n\to\infty} a_n \geq b + \liminf_{n\to\infty} a_n$ which completes the proof. □

The moments of different functionals of partition-valued processes appearing in Sections 2 and 3 satisfy the following recurrence

$$a_0 := a, \;\; a_n = b_n + \sum_{k=0}^{n} p_{n,k} a_k, \; n \in \mathbb{N}, \tag{1.11}$$

where $(p_{n,k})_{0 \leq k \leq n}$ is a probability distribution with $p_{n,n} < 1$ and $(b_n)_{n \in \mathbb{N}}$ is a given sequence of positive real numbers. With a view towards subsequent applications Proposition 15 investigates the rate of growth of the so defined a_n and particularly provides conditions of its boundedness.

Proposition 15. *Suppose there exists a sequence $(\psi_n)_{n\in\mathbb{N}}$ such that*

(C1) $\liminf_{n\to\infty} \psi_n \sum_{k=0}^{n} (1 - k/n) p_{n,k} > 0$,

(C2) *the sequence $(\psi_k b_k / k)_{k \in \mathbb{N}}$ is non-increasing.*

Then the sequence $(a_n)_{n \in \mathbb{N}_0}$ defined by (1.11) satisfies

$$a_n = O\Big(\sum_{k=1}^{n} \frac{b_k \psi_k}{k} \Big), \;\; n \to \infty. \tag{1.12}$$

In particular, (a_n) is bounded if the series $\sum_{k=1}^{\infty} \frac{b_k \psi_k}{k}$ converges.

Proof. Write for simplicity p_k for $p_{n,k}$ and let $\pi_k = \sum_{j=0}^{k} p_j$. Using (C2) we have

$$\sum_{k=1}^{n} \frac{b_k \psi_k}{k} \pi_{k-1} \geq \frac{b_n \psi_n}{n} \sum_{k=1}^{n} \pi_{k-1} = b_n \psi_n \sum_{j=0}^{n-1} (1 - j/n) p_j.$$

By (C1) there exist $n_0 \in \mathbb{N}$ and $c > 0$ such that

$$c \sum_{k=1}^{n} \frac{b_k \psi_k}{k} \pi_{k-1} \geq b_n \,, \qquad n \geq n_0. \tag{1.13}$$

From this, $x_n := c \sum_{k=1}^{n} b_k \psi_k / k$ satisfies

$$x_n \geq b_n + \sum_{k=1}^{n} x_k p_k \,, \quad n \geq n_0 \tag{1.14}$$

To check the latter, write

$$\begin{aligned} b_n + \sum_{k=1}^{n} x_k p_k &= b_n + c \sum_{j=1}^{n} \sum_{k=j}^{n} \frac{b_j \psi_j}{j} p_k \\ &= b_n + c \sum_{j=1}^{n} \frac{b_j \psi_j}{j} (1 - \pi_{j-1}) \\ &= b_n + c \sum_{j=1}^{n} \frac{b_j \psi_j}{j} - c \sum_{j=1}^{n} \frac{b_j \psi_j}{j} \pi_{j-1} \\ &= x_n + b_n - c \sum_{j=1}^{n} \frac{b_j \psi_j}{j} \pi_{j-1} \overset{(1.13)}{\leq} x_n. \end{aligned}$$

Set $x_0 := 0$. Subtracting (1.11) from (1.14) we see that $y_n := x_n + c_0 - a_n$ satisfies $y_n \geq \sum_{k=0}^{n} p_k y_k$ for $n \geq n_0$ and arbitrary c_0. By choosing $c_0 \geq \max_{n \leq n_0} a_n$ it is easily shown by induction that $y_n \geq 0$ for all $n \in \mathbb{N}$, which implies the desired estimate of a_n. □

Properties of iterative functions.

For the given strictly increasing continuous function g, there exists the unique inverse function g^{-1} which defines the sequence $(A_n)_{n \in \mathbb{N}_0}$ as follows

$$A_0 = 0, \;\; A_n := (g^{-1})^{\circ(n-1)}(x_0), \;\; n \in \mathbb{N}. \tag{1.15}$$

Lemma 16. *Assume that $g, h \in C^{(m)}[x_0, +\infty)$, $k \in C^{(m)}[0, x_0]$ and $F = \mathrm{Iter}(h, g, x_0, k)$ is m-times differentiable at x_0. Then F is m-times differentiable on $[x_0, +\infty)$.*

Proof. We only treat the case $m = 1$, as, for $m = 2, 3, \ldots$, the proof is the same. Since F is a sum of compositions of $C^{(1)}[x_0, +\infty)$ functions, it is differentiable on $[x_0, +\infty)\backslash(A_i)_{i\in\mathbb{N}}$. Therefore, we only have to check the continuity and differentiability at the points (A_i).
First step. Proof of continuity. By the assumption, F is continuous at $A_1 = x_0$, i.e.,

$$k(x_0) = h(x_0) + k(g(x_0)). \tag{1.16}$$

For fixed $k \geq 2$, we have from (1.2)

$$F(A_k - 0) = \sum_{i=1}^{k-1} h(g^{\circ(i-1)}(A_k - 0)) + k(g^{\circ(k-1)}(A_k - 0)), \tag{1.17}$$

and

$$F(A_k + 0) = \sum_{i=1}^{k} h(g^{\circ(i-1)}(A_k + 0)) + k(g^{\circ(k)}(A_k + 0)). \tag{1.18}$$

Now use (1.16) and continuity of h and g to obtain

$$\begin{aligned} &F(A_k + 0) - F(A_k - 0) \\ &\quad = h(g^{\circ(k-1)}(A_k)) + k(g^{\circ(k)}(A_k)) - k(g^{\circ(k-1)}(A_k)) \\ &\quad = h(x_0) + k(g(x_0)) - k(x_0) \overset{(1.16)}{=} 0. \end{aligned}$$

Second step. Proof of differentiability. The differentiability of F at x_0 implies that

$$k'(x_0) = h'(x_0) + k'(g(x_0))g'(x_0). \tag{1.19}$$

For $k \geq 2$, using (1.17) and (1.18) yields

$$F'_-(A_k) = \lim_{x\to A_k - 0} \frac{d}{\mathrm{d}x}\left(\sum_{i=1}^{k-1} h(g^{\circ(i-1)}(x)) + k(g^{\circ(k-1)}(x)) \right),$$

$$F'_+(A_k) = \lim_{x\to A_k + 0} \frac{d}{\mathrm{d}x}\left(\sum_{i=1}^{k} h(g^{\circ(i-1)}(x)) + k(g^{\circ(k)}(x)) \right).$$

Consequently,

$$\begin{aligned} & F'_+(A_k) - F'_-(A_k) \\ &= \lim_{x\to A_k+0} \frac{d}{dx} h(g^{\circ(k-1)}(x)) + k(g^{\circ(k)}(x)) - \lim_{x\to A_k-0} \frac{d}{dx} k(g^{\circ(k-1)}(x)). \end{aligned}$$

Set $u(x) := g^{\circ(k-1)}(x)$, then $u(A_k+0) = u(A_k-0) = u(A_k) = x_0$ and

$$\begin{aligned} & F'_+(A_k) - F'_-(A_k) \\ &= \lim_{x\to A_k+0} \frac{d}{dx} h(u(x)) + k(g(u(x))) - \lim_{x\to A_k-0} \frac{d}{dx} k(u(x)) \\ &= \lim_{x\to A_k+0} (h'(u(x)) + k'(g(u(x)))g'(u(x)))u'(x) - \lim_{x\to A_k-0} k'(u(x)))u'(x) \\ &= (h'(x_0) + k'(g(x_0))g'(x_0) - k'(x_0))u'(x_0) = 0, \end{aligned}$$

by (1.19). The proof is complete. □

From Lemma 16 it follows that the function F is m-times differentiable provided it satisfies the conditions

$$\begin{aligned} & k(x_0) = h(x_0) + k(g(x_0)), \\ & k'(x_0) = h'(x_0) + k'(g(x_0))g'(x_0), \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & k^{(m)}(x_0) = h^{(m)}(x_0) + (k(g(x_0)))^{(m)}. \end{aligned} \tag{1.20}$$

The following lemma proves the existence of a function $k(x)$ such that (1.20) holds.

Lemma 17. *Assume that* $W\Big(x-g(x), \ldots, x^{m+1} - g^{m+1}(x)\Big)\Big|_{x=x_0} \neq 0$. *Then there exists a function* $k(x) = \sum_{i=1}^{m+1} \alpha_i x^i$ *which satisfies (1.20).*

Proof. Plugging the representation $k(x) = \sum_{i=1}^{m+1} \alpha_i x^i$ into (1.20) gives the system of linear equations

$$\begin{aligned} & \Big(\alpha_1(x_0 - g(x_0)) + \ldots + \alpha_{m+1}(x_0^{m+1} - g^{m+1}(x_0))\Big) = h(x_0), \\ & \Big(\alpha_1 \tfrac{d}{dx}(x - g(x)) + \ldots + \alpha_{m+1} \tfrac{d}{dx}(x^{1+m} - g^{m+1}(x))\Big)\Big|_{x=x_0} = h'(x_0), \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & \Big(\alpha_1 \tfrac{d^m}{dx^m}(x - g(x)) + \ldots + \alpha_{m+1} \tfrac{d^m}{dx^m}(x^{m+1} - g^{m+1}(x))\Big)\Big|_{x=x_0} = h^{(m)}(x_0). \end{aligned}$$

The determinant of this system is $W(x_0)$ which is not equal to zero by the assumption. Therefore, the system has a unique solution which implies that the function k is well defined and satisfies conditions (1.20). □

Using a reasoning similar to that used in the proof of Theorem 13 one can check the following.

Theorem 18. *Let the triples (h_1, g, x_0) and (h_2, g, x_0) generate the iterative functions f_1 and f_2, respectively. Assume that $\lim_{x\to\infty} f_1(x) = +\infty$. Then*

I. $h_2(x) \sim h_1(x)$, $x \to \infty$ implies $f_2(x) \sim f_1(x)$, $x \to \infty$, and

II. $h_2(x) = o(h_1(x))$, $x \to \infty$ implies $f_2(x) = o(f_1(x))$, $x \to \infty$.

1.2 Absorption time of decreasing Markov chains

1.2.1 Main result. In this subsection we investigate the weak convergence of the absorption time X_n defined by (3). Recall that the distribution of the corresponding I_n was given in (4).

Theorem 19. *Assume that, for each $x > 0$, there exists a limit*

$$\Phi(x) := \lim_{n\to\infty} n(1 - \mathbb{E}(I_n/n)^x), \tag{1.21}$$

and that $\Phi(x)$ is finite, for some $x > 0$. Then, as $n \to \infty$,

$$X_n/n \xrightarrow{d} \int_0^\infty e^{-\sigma_t} \mathrm{d}t, \tag{1.22}$$

where $(\sigma_t)_{t\geq 0}$ is a subordinator with the Lévy exponent $\Phi(x)$.

Remark 20. According to the Lévy-Khintchine formula, $\Phi(x)$ can be represented as follows

$$\Phi(x) = \lambda + \kappa x + \int_0^\infty (1 - e^{-tx})\rho(\mathrm{d}t),\ x \geq 0,$$

where $\lambda \geq 0$ is the killing rate, $\kappa \geq 0$ is the drift and ρ is the Lévy measure of (σ_t).

1.2.2 Proof of Theorem 19.

For $n \in \mathbb{N}$ and $k \in \mathbb{N}_0$, set

$$m_n^{(k)} := \mathbb{E}X_n^k \quad \text{and} \quad M_n^{(k)} := \mathbb{E}(X_n/n)^k.$$

We start with an auxiliary result.

Lemma 21. *For fixed $k \in \mathbb{N}$,*

$$m_1^{(k)} = 0, \quad m_n^{(k)} = \sum_{j=0}^{k-1} \binom{k}{j} (-1)^{k-j-1} m_n^{(j)} + \sum_{i=1}^{n-1} \mathbb{P}\{I_n = i\} m_i^{(k)}, \quad n \geq 2.$$

Proof. For fixed $k \in \mathbb{N}$ and $n \geq 2$, using (2) we obtain

$$\begin{aligned}
m_n^{(k)} &= \mathbb{E}(1 + X_{I_n})^k = \sum_{i=0}^{k} \binom{k}{i} \mathbb{E}X_{I_n}^i = \sum_{i=0}^{k-1} \binom{k}{i} \mathbb{E}(X_n - 1)^i \\
&+ \mathbb{E}X_{I_n}^k = \sum_{i=0}^{k-1} \binom{k}{i} \sum_{j=0}^{i} \binom{i}{j} m_n^{(j)} (-1)^{i-j} + \sum_{i=1}^{n-1} \mathbb{P}\{I_n = i\} m_i^{(k)} \\
&= \sum_{j=0}^{k-1} m_n^{(j)} \sum_{i=j}^{k-1} \binom{k}{i} \binom{i}{j} (-1)^{i-j} + \sum_{i=1}^{n-1} \mathbb{P}\{I_n = i\} m_i^{(k)} \\
&= \sum_{j=0}^{k-1} \binom{k}{j} (-1)^{k-j-1} m_n^{(j)} + \sum_{i=1}^{n-1} \mathbb{P}\{I_n = i\} m_i^{(k)},
\end{aligned}$$

where the last equality follows from the formula

$$\begin{aligned}
\sum_{i=j}^{k-1} \binom{k}{i} \binom{i}{j} (-1)^{i-j} &= \frac{k!}{j!} \sum_{i=j}^{k-1} \frac{(-1)^{i-j}}{(k-i)!(i-j)!} \\
&= \binom{k}{j} \sum_{i=0}^{k-j-1} \binom{k-j}{i} (-1)^i \\
&= \binom{k}{j} \left(\sum_{i=0}^{k-j} \binom{k-j}{i} (-1)^i - (-1)^{k-j} \right) \\
&= \binom{k}{j} (-1)^{k-j-1}.
\end{aligned}$$

□

Proof of Theorem 19. Assuming that (1.21) holds set

$$\Psi_n(x) := \mathbb{E}(I_n/n)^x = \mathbb{E}\exp(-x(\log n - \log I_n)), \quad n \in \mathbb{N}.$$

Then, for each $x > 0$,

$$\Phi(x) = \lim_{n\to\infty} n(1 - \Psi_n(x)) = \lim_{n\to\infty}\Big(- n\log \Psi_n(x)\Big).$$

This implies that $\Phi(x) = -\log \phi(x)$, where $\phi(x)$ is the Laplace-Stieltjes transform of, possibly improper, probability distribution $\mu \neq \delta_\infty$ on $[0, \infty]$. Hence $\Phi(x)$ is the Laplace exponent of a subordinator (σ_t), say. Furthermore,

$$\mathbb{E}\bigg(\int_0^\infty e^{-\sigma_t}\mathrm{d}t\bigg)^k = \frac{k!}{\Phi(1)\cdots\Phi(k)} =: l^{(k)}, \quad k \in \mathbb{N}_0,$$

and the law of the last integral is uniquely determined by its moments. While in the case $\lambda = 0$ (no killing) this directly follows from [21, Theorem 2(i)], in the case $\lambda > 0$ one can argue as follows. We have

$$\int_0^\infty e^{-\sigma_t}\mathrm{d}t = \int_0^T e^{-\sigma_t}\mathrm{d}t \stackrel{d}{=} \int_0^T e^{-\sigma^*_t}\mathrm{d}t,$$

where (σ^*_t) is a subordinator with the Laplace exponent $\Phi^*(x) = \Phi(x) - \lambda$, and T is a random variable with exponential distribution with parameter λ which is independent of both subordinators. Now another appeal to [21, Theorem 2(i)] allows us to conclude that

$$\mathbb{E}\bigg(\int_0^T e^{-\sigma^*_t}\mathrm{d}t\bigg)^k = \frac{k!}{(\lambda + \Phi^*(1))\cdots(\lambda + \Phi^*(k))} = l^{(k)}, \quad k \in \mathbb{N}_0.$$

Therefore, to prove (1.22) it suffices to show that, for every $k \in \mathbb{N}$,

$$\lim_{n\to\infty} M_n^{(k)} = l^{(k)}. \tag{1.23}$$

For every $k \in \mathbb{N}_0$, set

$$a_n^{(k)} := n^k(M_n^{(k)} - l^{(k)}), \quad n \in \mathbb{N}.$$

By using induction on k we will prove that $a_n^{(k)} = o(n^k)$, as $n \to \infty$, for every $k \in \mathbb{N}_0$. Since $a_n^{(0)} = 0$ the hypothesis holds for $k = 0$. Assume that

$a_n^{(j)} = o(n^j)$ for $j \le k-1$, in particular, that (1.23) holds with k replaced by j, $j \le k-1$. This together with Lemma 21 implies that, for every fixed $k \in \mathbb{N}$, the sequence $(M_n^{(k)})$ satisfies the equality

$$M_1^{(k)} = 0, \quad M_n^{(k)} = kM_n^{(k-1)}/n + o(1/n) + \sum_{i=1}^{n-1} \mathbb{P}\{I_n = i\}(i/n)^k M_i^{(k)}, \quad n \ge 2.$$

Hence

$$a_n^{(k)} = c_n^{(k)} + \sum_{i=1}^{n-1} \mathbb{P}\{I_n = i\} a_i^{(k)}, \ n \in \mathbb{N},$$

where

$$c_n^{(k)} := n^{k-1}\left(n\Big(\mathbb{E}(I_n/n)^k - 1\Big)l^{(k)} + kl^{(k-1)}\right) + o(n^{k-1}).$$

As $n \to \infty$, the expression in the large parantheses goes to $-\Phi(k)l^{(k)} + kl^{(k-1)} = 0$. Therefore, $c_n^{(k)} = o(n^{k-1})$. Now it can be checked that $|a_n^{(k)}| \le \sum_{i=1}^n |c_i^{(k)}|$, $n \in \mathbb{N}$. Since the usual convergence entails the convergence in the sense of Cezàro this further leads to

$$\frac{|a_n^{(k)}|}{n^k} \le \frac{1}{n}\sum_{i=1}^{n} \frac{|c_i^{(k)}|}{i^{k-1}} \to 0,$$

as $n \to \infty$. The proof is complete. □

1.2.3 Example. Consider an example in which X_n/n converges in distribution to a random variable with beta(a, b) distribution with density

$$x \mapsto \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} 1_{(0,1)}(x). \tag{1.24}$$

Example 22. Let $\alpha > -1$ and δ, γ be non-negative numbers such that $\gamma/(1+\alpha) \le \delta$. For $n \in \mathbb{N}$ large enough, set

$$\mathbb{P}\{I_n = n-1\} = 1 - \frac{\delta}{n}, \quad \mathbb{P}\{I_n = k\} = \left(\frac{k}{n}\right)^{\alpha} \frac{\gamma}{n^2}, \quad k = 2, 3, \ldots, n-2,$$

and

$$\mathbb{P}\{I_n = 1\} = \frac{\delta}{n} - \frac{\gamma}{n^2}\sum_{k=2}^{n-2}\left(\frac{k}{n}\right)^{\alpha}.$$

Then (1.21) holds with

$$\begin{aligned}\Phi(x) &= x+\delta-\frac{\gamma}{x+\alpha+1}=\left(\delta-\frac{\gamma}{\alpha+1}\right)+x+\frac{\gamma}{\alpha+1}-\frac{\gamma}{x+\alpha+1}\\ &= \left(\delta-\frac{\gamma}{\alpha+1}\right)+x+\gamma\int_0^\infty(1-e^{-xt})e^{-(\alpha+1)t}\mathrm{d}t.\end{aligned}$$

Hence $\Phi(x)$ is the Laplace exponent of a subordinator with killing rate $\delta-\gamma/(\alpha+1)$, unit drift and the Lévy measure ρ defined by

$$\rho(\mathrm{d}t)=\gamma e^{-(\alpha+1)t}1_{(0,\infty)}(t)\mathrm{d}t.$$

If we take $\delta:=\gamma/(\alpha+1)>0$ then $X_n/n\xrightarrow{d}\mathrm{beta}(\alpha+2,\delta)$, as $n\to\infty$, whereas if we take $\delta>0$ and $\gamma:=0$ then $X_n/n\xrightarrow{d}\mathrm{beta}(1,\delta)$, as $n\to\infty$. Thus the set of limiting laws for X_n/n contains $\mathrm{beta}(a,b)$ laws with parameters $a\geq 1$ and $b>0$, and these are the laws of the exponential functionals of certain subordinators.

Note that there does not exist a subordinator for which the law of exponential functional is $\mathrm{beta}(a,b)$ law with $a<1$. Indeed, assuming the contrary the corresponding Laplace exponent would take the form $x\mapsto\frac{x(x+a+b-1)}{x+a-1}$. This, however, is impossible since this function is not monotone on $(0,\infty)$ when $a<1$.

1.3 Bibliographic comments

Methods of asymptotic analysis of recurrences. Some known approaches to investigating *deterministic recurrences* were partially surveyed in Introduction. On the other hand, the *contraction method* is a quite universal probabilistic technique for the asymptotic analysis of linear *random recurrences* (1). The method was invented by Uwe Rösler in [119] for the analysis of the Quicksort algorithm and further developed in [106, 117, 120, 121]. The picture up to 1998 is surveyed in [122]. It seems that the most general result on the contraction method with *nondegenerate limit equation* is derived in [108]. An important extension to the linear random recurrences with

degenerate limit equation can be found in [109]. The main result of [109] is reformulated as Proposition 51 of the present work.

Iterative functions and "log star"-function. The iterative functions which were the main ingredient of the method described in Section 1.1 have already been used in the context of divide-and-conquer paradigm [85]. The cited paper is concerned with stochastic processes $(T(x))_{x\in\mathbb{R}^+}$ whose marginal distributions are given by the equality

$$T(x) \stackrel{d}{=} a(x) + T'(t(x)), \quad x \in \mathbb{R}^+,$$

where $a(\cdot)$ is a non-negative (deterministic) function, and $t(\cdot)$ is a random variable taking values in $[0, \cdot]$ which is independent of $(T'(x))_{x\in\mathbb{R}^+}$, an independent copy of $(T(x))_{x\in\mathbb{R}^+}$.

To our knowledge, the "log star" asymptotics arises not often. In particular, we are only aware of three applied models which exhibit such a behavior:

(a) the number of distinguishable alleles according to the Ohta-Kimura model of neutral mutation [86];

(b) the average complexity of Delaunay triangulation of the Euclidian minimum spanning tree [37];

(c) the total number of particles at time $t > 0$ in the spatial Kingman coalescent on graph with bounded degrees [4].

Absorption time in Markov chains.

There are several articles dealing with the absorption times of nonincreasing Markov chains. From the results of these papers it follows that the weak asymptotic behavior of the absorption times can be substantially different. For instance, in [132] it is proved that under certain monotonicity assumptions, the absorption times exhibit the same weak asymptotic behavior as the first-passage time through a level by a random walk with finite variance of steps (which means that the limit law is normal). In [68] conditions are found which ensure that, suitably normalized, absorption times weakly converge to

the law of exponential functional of a subordinator. There are examples in which the limit law does not exist at all (see, for example, [82]).

Assuming that the distribution of size of the first jump of a Markov chain is of the form

$$\mathbb{P}\{I_n = k\} = \frac{p_{n-k}}{p_1 + \ldots + p_{n-1}},\ k = 1, \ldots, n-1,\ n \geq 2,$$

where $(p_k)_{k\in\mathbb{N}}$ is a probability distribution with $p_1 > 0$, or

$$I_n = [n\eta] + 1, \tag{1.25}$$

where $\eta \in (0, 1)$ is a random variable with distribution not supported by a geometric sequence, in [78] and [80], respectively, the whole spectra are obtained of possible limiting laws for, properly centered and normalized, absorption times. In the case when η in (1.25) has uniform $[0, 1]$ law, the law of corresponding absorption time X_n arises in a number of diverse applications. For instance, the X_n has the same law as (a) the number of upper records in a sample of size $n+1$ from a continuous distribution, (b) the number of cycles in random permutations of $n + 1$ objects, (c) the number of collision events that take place in beta $(3, 1)$ coalescent restricted to the set $\{1, 2, \ldots, n+1\}$ until there is just a single block. A longer list of examples can be found in [10].

Section 1.1 and Section 1.2 are based on [95] and [96], respectively.

Chapter 2

Stick-breaking partitions

2.1 Definition and discussion

In a classical occupancy scheme balls are thrown independently in an infinite array of boxes with probability p_k of hitting box $k = 1, 2, \ldots$, where $(p_k)_{k\in\mathbb{N}}$ is a fixed collection of positive frequencies summing up to unity. A quantity of traditional interest is the number K_n of boxes occupied by at least one of n balls. In concrete applications 'boxes' correspond to distinguishable species or types, and K_n is the number of distinct species represented in a random sample of size n.

Less explored are the mixture models in which frequencies are themselves random variables $(P_k)_{k\in\mathbb{N}}$, so that the balls are allocated independently conditionally given the frequencies. The model is important for many applications related to sampling from random discrete distributions, and may be interpreted as the occupancy scheme in random environment. The variability of allocation of the balls is then affected by both randomness in sampling and randomness of the environment. With respect to K_n, the environment may be called *strong* if the randomness in (P_k) has dominating effect. One way to capture this idea is to consider the conditional expectation

$$R_n^* := \mathbb{E}\big(K_n \,|\, (P_k)\big) = \sum_{k=1}^{\infty} \big(1 - (1 - P_k)^n\big)$$

and to compare fluctuations of K_n about R_n^* with fluctuations of R_n^* itself. By Karlin's [84] law of large numbers, one always has $K_n \sim R_n^*$ a.s., as $n \to \infty$, so the environment may be regarded as strong if the sampling variability is negligible to the extent that R_n^* and K_n, normalized by the same constants, has the same limiting distributions, see [60] for examples.

In this chapter we focus on the limiting distributions of K_n for the *Bernoulli sieve* [52, 56, 57, 63], which is the infinite occupancy scheme with random frequencies

$$P_k := W_1 W_2 \cdots W_{k-1}(1 - W_k), \quad k \in \mathbb{N}, \tag{2.1}$$

where $(W_k)_{k\in\mathbb{N}}$ are independent copies of a random variable W taking values in $(0, 1)$. From a viewpoint, K_n is the number of blocks of a regenerative composition structure [11, 58, 60] induced by a compound Poisson process with jumps $|\log W_k|$. Discrete probability distributions with random masses (2.1) are sometimes called residual allocation models, the best known being the instance associated with Ewens' sampling formula, when $W \stackrel{d}{=} \text{beta}(c, 1)$ for $c > 0$. Following [56, 63], frequencies (2.1) can be considered as sizes of the component intervals obtained by splitting $[0, 1]$ at points of a stick-breaking process (multiplicative renewal process) $(Q_k)_{k\in\mathbb{N}_0}$, where

$$Q_0 := 1, \quad Q_j := \prod_{i=1}^{j} W_i, \quad j \in \mathbb{N}. \tag{2.2}$$

Accordingly, boxes can be identified with open intervals (Q_k, Q_{k-1}), and balls with points of an independent sample $U_1, \ldots, U_n$ from the uniform distribution on $[0, 1]$ which is independent of (Q_k). In this representation balls i and j occupy the same box iff points U_i and U_j belong to the same component interval.

Throughout we assume that the distribution of $|\log W|$ is non-lattice, and we use the following notation for the moments

$$\mu := \mathbb{E}|\log W|, \quad \sigma^2 := \text{Var}\,(\log W) \quad \text{and} \quad \nu := \mathbb{E}|\log(1 - W)|,$$

which may be finite or infinite.

In this chapter we derive the limiting distributions of K_n directly from the properties of the counting process

$$\begin{aligned} N^*(x) &:= \#\{k \in \mathbb{N} : P_k \geq e^{-x}\} \\ &= \#\{k \in \mathbb{N} : W_1 \cdots W_{k-1}(1 - W_k) \geq e^{-x}\}, \quad x > 0, \end{aligned} \tag{2.3}$$

in the range of small frequencies (large x). This allows us to treat the cases of finite and infinite ν in a unified way, and to see how the centering of K_n needs to be adjusted in the case $\nu = \infty$. We emphasize here that the connection between K_n and $N^*(x)$ remains veiled unless we consider the Bernoulli sieve as the occupancy scheme with random frequencies (a random environment), and the process of occupancy counts K_n is analyzed conditionally on the environment. Thus we believe that the approach presented here offers a natural way to study the occupancy problem, since the method is based on a direct analysis of frequencies and calls for generalizations.

2.2 Markov chains and distributional recurrences

A random combinatorial structure which captures the occupancy of boxes by n indistinguishable balls is the *weak composition* C_n^* comprised of nonnegative integer parts summing up to n. The adjective 'weak' means that zero parts are allowed, for instance, the sequence $(2, 3, 0, 1, 0, 0, 1, 0, 0, 0, \ldots)$ (padded by infinitely many 0's) is a possible value of C_7^*. A related structure which contains less information is a composition C_n of n obtained by discarding zero parts of C_n^*. Arranging further the parts of C_n in non-increasing order yields a random partition of n.

If C_n^* is generated by the Bernoulli sieve, the corresponding partition can be called *stick-breaking partition*. In this case, the parts of C_n^* can be represented (see [58, p. 452]) as the magnitudes of jumps of a time-homogeneous nonincreasing Markov chain $Q_n^* = (Q_n^*(k))_{k \in \mathbb{N}_0}$ on integers, which

starts at n and moves from n to m with transition probabilities

$$q^*(n,m) = \binom{n}{m}\mathbb{E}(1-W)^{n-m}W^m, \quad m = 0,\dots,n.$$

In the same direction, parts of the composition C_n are the magnitudes of jumps of a Markov chain $Q_n = (Q_n(k))_{k\in\mathbb{N}_0}$ with transition probabilities

$$q(n,m) = \binom{n}{m}\frac{\mathbb{E}(1-W)^{n-m}W^m}{1-\mathbb{E}W^n}, \quad m = 0,\dots,n-1.$$

Define M_n to be the index of the last occupied box in the Bernoulli sieve, which is the value of k satisfying $Q_k < \min(U_1,\dots,U_n) < Q_{k-1}$, and let $L_n := M_n - K_n$ be the number of empty boxes with indices not exceeding M_n. The Markovian realization implies (see [56, Section 3]) the following distributional recurrences the first two of which are of the form (2), and the third is of the form (1):

$$M_0 = 0, \quad M_n =_d M'_{Q^*_n(1)} + 1, \quad n \in \mathbb{N},$$

$$K_0 = 0, \quad K_n =_d K'_{Q_n(1)} + 1, \quad n \in \mathbb{N}, \tag{2.4}$$

$$L_0 = 0, \quad L_n =_d L'_{Q^*_n(1)} + 1_{\{Q^*_n(1)=n\}}, \quad n \in \mathbb{N}, \tag{2.5}$$

where in the right-hand sides M'_j and L'_j are assumed independent of $Q^*_n(1)$ and distributed like M_j and L_j, respectively, for each $j \in \mathbb{N}$, and K'_j is assumed independent of $Q_n(1)$ and distributed like K_j, for each $j \in \mathbb{N}$. Analysis of these recurrences by known direct methods is difficult, as these impose restrictive conditions on the moments of $Q_n(1)$ or $Q^*_n(1)$. Nevertheless, coupling with the multiplicative random walk allows one to gain enough information about the compositions. For instance, let $g_{n,m}$ be the potential function, equal to the probability that Q_n ever visits the state m:

$$g_{n,m} = \sum_{j=0}^{\infty}\mathbb{P}\{Q_n(j) = m\}.$$

The coupling implies that ([63, Proposition 5])

$$\lim_{n\to\infty} g_{n,m} = \frac{1-\mathbb{E}W^m}{\mu m}, \tag{2.6}$$

which is 0 if $\mu = \infty$.

2.3 Number of occupied boxes

Set

$$\rho^*(x) := \inf\{k \in \mathbb{N} : W_1 W_2 \dots W_k < e^{-x}\}, \quad x > 0. \tag{2.7}$$

One of the main results of this chapter is the following theorem.

Theorem 23. *If there exist functions $f : \mathbb{R}_+ \to \mathbb{R}_+$ and $g : \mathbb{R}_+ \to \mathbb{R}$ such that $(\rho^*(x) - g(x))/f(x)$ converge weakly, as $x \to \infty$, to some non-degenerate and proper distribution, then also $(X_n - b_n)/a_n$ converge weakly, as $n \to \infty$, to the same distribution, where X_n can be either K_n or $N^*(\log n)$, and the constants are given by*

$$b_n = \int_0^{\log n} g(\log n - y)\, \mathbb{P}\{|\log(1-W)| \in \mathrm{d}y\}, \quad a_n = f(\log n).$$

In more detail, possible limits for ρ^* and the convergence criteria summarized in [103, Proposition 27] lead to the following characterization.

Corollary 24. The assumption of Theorem 23 holds iff either the distribution of $|\log W|$ belongs to the domain of attraction of a stable law, or the function $\mathbb{P}\{|\log W| > x\}$ slowly varies at ∞. Accordingly, there are five possible types of convergence:

(a) If $\sigma^2 < \infty$ then, with

$$b_n = \mu^{-1}\left(\log n - \int_0^{\log n} \mathbb{P}\{|\log(1-W)| > x\}\mathrm{d}x\right) \tag{2.8}$$

and $a_n = (\mu^{-3}\sigma^2 \log n)^{1/2}$, the limiting distribution of $(K_n - b_n)/a_n$ is standard normal.

(b) If $\sigma^2 = \infty$, and

$$\int_0^x y^2\, \mathbb{P}\{|\log W| \in \mathrm{d}y\} \;\sim\; \ell(x), \quad x \to \infty,$$

for some ℓ slowly varying at ∞, then, with b_n given in (2.8) and $a_n = \mu^{-3/2} c_{[\log n]}$, where (c_n) is any positive sequence satisfying $\lim_{n\to\infty} n\ell(c_n)/c_n^2 = 1$, the limiting distribution of $(K_n - b_n)/a_n$ is standard normal.

(c) If

$$\mathbb{P}\{|\log W| > x\} \sim x^{-\alpha}\ell(x), \quad x \to \infty, \tag{2.9}$$

for some ℓ slowly varying at ∞ and $\alpha \in (1,2)$ then, with b_n given in (2.8) and $a_n = \mu^{-(\alpha+1)/\alpha} c_{[\log n]}$, where (c_n) is any positive sequence satisfying $\lim_{n\to\infty} n\ell(c_n)/c_n^\alpha = 1$, the limiting distribution of $(K_n - b_n)/a_n$ is α-stable with characteristic function

$$t \mapsto \exp\{-|t|^\alpha \Gamma(1-\alpha)(\cos(\pi\alpha/2) + i\sin(\pi\alpha/2)\operatorname{sgn}(t))\}, \ t \in \mathbb{R}.$$

(d) Assume that the relation (2.9) holds with $\alpha = 1$. Let $r : \mathbb{R} \to \mathbb{R}$ be any nondecreasing function such that $\lim_{x\to\infty} x\mathbb{P}\{|\log W| > r(x)\} = 1$ and set

$$m(x) := \int_0^x \mathbb{P}\{|\log W| > y\}\mathrm{d}y, \quad x > 0.$$

Then, with

$$b_n := \int_0^{\log n} \frac{\log n - y}{m(r((\log n - y)/m(\log n - y)))}\, \mathbb{P}\left\{\left|\log(1-W)\right| \in \mathrm{d}y\right\}$$

and

$$a_n := \frac{r(\log n/m(\log n))}{m(\log n)},$$

the limiting distribution of $(K_n - b_n)/a_n$ is 1-stable with characteristic function

$$t \mapsto \exp\{-|t|(\pi/2 - i\log|t|\operatorname{sgn}(t))\}, \ t \in \mathbb{R}.$$

(e) If the relation (2.9) holds for $\alpha \in [0,1)$ then, with $b_n \equiv 0$ and $a_n := \log^\alpha n/\ell(\log n)$, the limiting distribution of K_n/a_n is the Mittag-Leffler law θ_α with moments

$$\int_0^\infty x^k\, \theta_\alpha(\mathrm{d}x) \ = \ \frac{k!}{\Gamma^k(1-\alpha)\Gamma(1+\alpha k)}, \quad k \in \mathbb{N}.$$

According to [56], the number L_n of empty boxes is regulated by μ and ν via the relation $\lim_{n\to\infty} \mathbb{E}L_n = \nu/\mu$ (provided at least one of these is finite), and the weak asymptotics of M_n (the index of the last occupied box) coincides

with that of $\rho^*(\log n)$, i.e., $(M_n - b_n)/a_n$ and $(\rho^*(\log n) - b_n)/a_n$ have the same proper and non-degenerate limiting distribution (if any). In [56, 63] it is shown that under the condition $\nu < \infty$ the weak asymptotics of K_n coincides with that of M_n, hence with that of $\rho^*(\log n)$. That is to say, when $\nu < \infty$, the way L_n varies does not affect the asymptotics of K_n through the representation $K_n = M_n - L_n$. Clearly, this result is a particular case of Theorem 23 because when $\nu < \infty$

$$\lim_{x\to\infty} \frac{g(x) - \int_0^x g(x-y)\,\mathbb{P}\{|\log(1-W)| \in \mathrm{d}y\}}{f(x)} = 0 \qquad (2.10)$$

(see Remark 31 for the proof).

Theorem 23 says that in the case $\nu = \infty$ the asymptotics of L_n may affect the asymptotics of K_n, and this is indeed the case whenever (2.10) fails, hence a two-term centering of K_n is indispensable. The following example illustrates the phenomenon.

Example. Assume that, for some $\gamma \in (0, 1/2)$,

$$\mathbb{P}\{W > x\} = \frac{1}{1 + |\log(1-x)|^\gamma}, \quad x \in [0, 1).$$

Then

$$\mathbb{E}\log^2 W < \infty \ \text{ and } \ \mathbb{P}\{|\log(1-W)| > x\} \sim x^{-\gamma}, \quad x \to \infty,$$

and in this case,

$$a_n = \mathrm{const}\log^{1/2} n \ \text{ and } \ b_n = \mu^{-1}(\log n - (1-\gamma)^{-1}\log^{1-\gamma} n + o(\log^{1-\gamma} n)).$$

Thus we see that the second term $b_n - \mu^{-1}\log n$ of centering cannot be ignored (as it is not killed by the scaling). Moreover, one can check that indeed

$$\mathbb{E}L_n \sim \frac{1}{\mu}\sum_{k=1}^{n} \frac{\mathbb{E}W^k}{k} \ \sim \ b_n - \mu^{-1}\log n \ \sim \ \frac{1}{\mu(1-\gamma)}\log^{1-\gamma} n,$$

which demonstrates the substantial contribution of L_n.

We will make use of the poissonized version of the occupancy model, in which balls are thrown in boxes in continuous time, at the epochs of a unit

rate Poisson process $(\pi_t)_{t\geq 0}$. The variables associated with time $t \geq 0$ will be denoted $K(t)$,$R^*(t)$ etc., i.e., $K(t) = K_{\pi_t}$. For instance, the expected number of occupied boxes within time interval $[0, t]$ conditionally given (P_k) is

$$R^*(t) = \sum_{n=0}^{\infty} \left(e^{-t}t^n/n!\right) R_n^* = \sum_{k=1}^{\infty} \left(1 - e^{-tP_k}\right).$$

The advantage of the poissonized model is that, given (P_k), the allocation of balls in boxes $1, 2, \ldots$, as t varies, occurs by *independent* Poisson processes of rates $P_1, P_2, \ldots$.

The variable $N^*(x)$ is the number of sites on $[0, x]$ visited by a perturbed random walk with the generic components $(|\log W|, |\log(1 - W)|)$. We will develop some general renewal theory for perturbed random walks, which we believe might be of some independent interest. The approach based on perturbed random walks is more general than the one exploited in [56] and is well adapted to treat the cases $\nu < \infty$ and $\nu = \infty$ in a unified way.

2.4 Renewal theory for perturbed random walks

2.4.1 Preliminaries. Let $(\xi_k, \eta_k)_{k\in\mathbb{N}}$ be independent copies of a random vector (ξ, η) with arbitrarily dependent components $\xi > 0$ and $\eta \geq 0$. We assume that the law of ξ is nonlattice, although extension to the lattice case is possible. For $(S_k)_{k\in\mathbb{N}_0}$ a random walk with $S_0 = 0$ and increments ξ_k, the sequence $(T_k)_{k\in\mathbb{N}}$ with

$$T_k := S_{k-1} + \eta_k, \quad k \in \mathbb{N},$$

is called a *perturbed random walk*. Since $\lim_{k\to\infty} T_k = \infty$ a.s., there is some finite number

$$N(x) := \#\{k \in \mathbb{N} : T_k \leq x\}, \quad x \geq 0,$$

of sites visited on the interval $[0, x]$. Let

$$R(x) := \sum_{k=0}^{\infty} \left(1 - \exp\left(-xe^{-T_k}\right)\right), \quad x \geq 0. \tag{2.11}$$

Our aim is to find conditions for the weak convergence of, properly normalized and centered, $N(x)$ and $R(x)$, as $x \to \infty$.

It is natural to compare $N(x)$ with the number of renewals

$$\rho(x) := \#\{k \in \mathbb{N}_0 : S_k \le x\} \;=\; \inf\{k \in \mathbb{N} : S_k > x\}, \quad x \ge 0.$$

In the case $\mathbb{E}\eta < \infty$ the weak convergence of one of the variables $(\rho(x) - g(x))/f(x)$ and $(N(x) - g(x))/f(x)$ (with suitable f, g) implies the weak convergence of the other to the same distribution. We thus mainly focus on the cases when the contribution of the η_k's does affect the asymptotics of $N(x)$. Remark 25 collects some relevant properties of the functions f.

Remark 25. Assume that $\frac{\rho(x)-g(x)}{f(x)}$ weakly converges to a non-degenerate and proper distribution, and that the centering $g(x)$ cannot be ignored. Then, according to [103, Proposition 27], the function f can be taken to satisfy $f(\log n) = a_n$ with a_n defined in Theorem 23 and Corollary 24. Using a result on asymptotic inverses of regularly varying functions [24, Theorem 1.5.12] we conclude that $f(x)$ is regularly varying at ∞ with index $\beta \in [1/2, 1]$. In particular, $\lim_{x\to\infty} f(x) = +\infty$, and $f(\log x)$ is slowly varying at ∞. Finally, we note that $f(x)$ grows not slower than $\sqrt{x}$. This is obvious if either $\beta \in (1/2, 1]$ (in which case $\mathbb{P}\{\xi > x\}$ is slowly varying at ∞ with index $-\beta$), or $f(x) \sim \mathrm{const}\sqrt{x}$ (in which case $\operatorname{Var}\xi < \infty$). In the remaining case $\int_0^x y^2 \mathbb{P}\{\xi \in \mathrm{d}y\} \sim \ell(x)$ and $\lim_{x\to\infty} \ell(x) = \infty$, where ℓ is slowly varying at ∞, $f(x)$ satisfies $\lim_{x\to\infty} \frac{x\ell(f(x))}{f^2(x)} = 1$. Since $\lim_{x\to\infty} \ell(f(x)) = +\infty$ we conclude that $\lim_{x\to\infty} (x/f^2(x)) = 0$.

Recall the following easy observation: for $x, y \ge 0$

$$\rho(x+y) - \rho(x) \overset{a.s.}{\le} \rho'(x, y) \overset{d}{=} \rho(y), \tag{2.12}$$

where $\rho'(x, y) := \inf\{k - \rho(x) \in \mathbb{N} : S_k - S_{\rho(x)} > y\}$, and $(\rho'(x, y))_{y\ge 0}$ is independent of $\rho(x)$ and has the same distribution as $(\rho(y))_{y\ge 0}$. Denote by $U(x) := \mathbb{E}\rho(x) = \sum_{k=0}^{\infty} \mathbb{P}\{S_k \le x\}$ the renewal function of (S_k). From (2.12) and Fekete's lemma we have

$$U(x+y) - U(x) \le C_1 y + C_2, \quad x, y \ge 0, \tag{2.13}$$

for some positive constants C_1 and C_2.

For fixed function $f > 0$ we say that functions g_1, g_2 are *f-equivalent* if

$$\lim_{x\to\infty} \frac{g_1(x) - g_2(x)}{f(x)} = 0.$$

In what follows we will consider functions involved in centering of random variables up to this kind of equivalence. For instance, if we write $g = 0$, this means that g is f-equivalent to zero for some (context-dependent) f involved in scaling of some random variables.

The next lemma will be used in the proof of Theorem 30.

Lemma 26. *If $\frac{\rho(x)-g(x)}{f(x)}$ weakly converges then*

$$\lim_{x\to\infty} \frac{g(x) - g(x-y)}{f(x)} = 0 \quad \textit{locally uniformly in } y, \tag{2.14}$$

and, for every $\lambda \in \mathbb{R}$,

$$\lim_{x\to\infty} \frac{\int_0^x g(x-y)\,\mathrm{d}G(y) - \int_0^{x+\lambda} g(x+\lambda-y)\,\mathrm{d}G(y)}{f(x)} = 0, \tag{2.15}$$

for arbitrary distribution function G with $G(0) = 0$.

Proof. First of all, note that, according to Remark 25, $\lim_{x\to\infty} f(x) = +\infty$. Clearly (2.14) is a property of the class of f-equivalent functions g. We refer to [103, Proposition 27] for the list of possible limiting laws and corresponding normalizations for $\rho(x)$. Relation (2.14) trivially holds when $g(x) \equiv 0$. It is known that $g(x)$ cannot be chosen as zero if the law of ξ belongs to the domain of attraction of an α-stable law for $\alpha \in [1, 2]$. It is known that for the law of ξ in the domain of attraction of an α-stable law with $\alpha \in (1, 2]$ one can take $g(x) = x/\mathbb{E}\xi$ which satisfies (2.14).

Thus the only troublesome case is the stable domain of attraction for $\alpha = 1$. According to [3, Theorem 3], one can take

$$g(x) = \frac{x}{m(r(x/m(x)))},$$

where $m(x) := \int_0^x \mathbb{P}\{\xi > y\}\mathrm{d}y$, and $r(x)$ is any nondecreasing function such that $\lim_{x\to\infty} x\mathbb{P}\{\xi > r(x)\} = 1$. The concavity of $m(x)$ implies that $x \mapsto x/m(x)$

is nondecreasing. Thus $x \mapsto m(r(x/m(x)))$ is nondecreasing too as superposition of three nondecreasing functions. Hence, for every $\gamma \in (0,1)$,

$$g(\gamma x) \geq \gamma g(x), \quad x > 0,$$

which readily implies the subadditivity of g via

$$g(x) + g(z) \geq \left(\frac{x}{x+z} + \frac{z}{x+z}\right) g(x+z) = g(x+z).$$

Thus,

$$\limsup_{x\to\infty} \frac{g(x) - g(x-y)}{f(x)} \leq 0.$$

For the converse inequality for the lower limit it is enough to choose non-increasing g from the f-equivalence class, and by [3, Theorem 2] this indeed can be done by taking inverse function to $x \mapsto xm(r(x))$.

The stated uniformity of convergence is checked along the same lines, and (2.15) follows from the subadditivity of g and easy estimates. □

2.4.2 The case without centering.

We start with criteria for the weak convergence of $\rho(x)$ and $R(x)$ in the case when no centering is needed.

Theorem 27. *For $Y(x)$ any of the variables $\rho(x)$, $N(x)$ or $R(e^x)$ the following conditions are equivalent:*

(a) *there exists function $f(x) : \mathbb{R}_+ \to \mathbb{R}_+$ such that, as $x \to \infty$, $Y(x)/f(x)$ weakly converges to a proper and non-degenerate law,*

(b) *for some $\alpha \in [0,1)$ and some function ℓ slowly varying at ∞,*

$$\mathbb{P}\{\xi > x\} \sim x^{-\alpha}\ell(x), \quad x \to \infty. \tag{2.16}$$

Furthermore, if (2.16) holds then the limiting law is the Mittag-Leffler distribution θ_α, and one can take $f(x) = x^\alpha/\ell(x)$.

The assertion of Theorem 27 regarding $\rho(x)$ follows from [103, Proposition 27]. For the other two variables the result is a consequence of the following lemma.

Lemma 28. *We have*

$$\lim_{x\to\infty} \frac{N(x)}{\rho(x)} = 1 \quad \textit{in probability}$$

and

$$\lim_{x\to\infty} \frac{R(x)}{\rho(\log x)} = 1 \quad \textit{in probability.}$$

Proof. By definition of the perturbed random walk

$$\rho(x-y) - \sum_{j=1}^{\rho(x)} 1_{\{\eta_j > y\}} \le N(x) \le \rho(x) \tag{2.17}$$

for $0 < y < x$. Clearly, $\rho(x) \uparrow \infty$ a.s. and

$$\rho(x-y) \ge \rho(x) - \rho'(x-y, y) \ \text{ a.s.} \tag{2.18}$$

with ρ' as in (2.12), from which

$$\frac{\rho(x-y)}{\rho(x)} \ \xrightarrow{P}\ 1, \quad x \to \infty. \tag{2.19}$$

Finally, by the strong law of large numbers we have

$$\lim_{x\to\infty} \frac{\sum_{j=1}^{\rho(x)} 1_{\{\eta_j > y\}}}{\rho(x)} = \mathbb{P}\{\eta > y\} \ \text{ a.s.}$$

Therefore, dividing (2.17) by $\rho(x)$ and letting first $x \to \infty$ and then $y \to \infty$ we obtain the first part of the lemma.

In what follows the record $\int_a^b$ with $b < \infty$ means $\int_{[a,b]}$. For the second assertion, we use the representation

$$\begin{aligned} R(x) &= \int_1^\infty (1 - e^{-x/y}) \mathrm{d}\, N(\log y) \\ &= \int_0^x N(\log x - \log y) e^{-y} \mathrm{d}y - (1 - e^{-x}) N(0). \end{aligned} \tag{2.20}$$

Since $N(x)$ is a.s. non-decreasing in x we have, for any $a < x$,

$$\begin{aligned} \int_0^x N(\log x - \log y) e^{-y} \mathrm{d}y &\ge \int_0^a N(\log x - \log y) e^{-y} \mathrm{d}y \\ &\ge N(\log x - \log a)(1 - e^{-a}). \end{aligned}$$

Dividing this inequality by $\rho(\log x)$, sending $x \to \infty$ along with using (2.19) and the already established part of the lemma, and finally letting $a \to \infty$, we obtain the half of desired conclusion.

To get the other half, write

$$\begin{aligned}\int_0^x N(\log x - \log y)e^{-y}\mathrm{d}y &\overset{\text{a.s.}}{\leq} \rho(\log x)(1-e^{-x}) \\ &+ \int_0^1 (\rho(\log x - \log y) - \rho(\log x))e^{-y}\mathrm{d}y,\end{aligned} \tag{2.21}$$

where (2.12), the inequality $N(x) \leq \rho(x)$ a.s., and the fact that $\rho(y)$ is a.s. non-decreasing in y have been used. Since, by (2.13),

$$\mathbb{E}\int_0^1 (\rho(\log x - \log y) - \rho(\log x))e^{-y}\mathrm{d}y \leq \int_0^1 (C_1|\log y| + C_2)e^{-y}\mathrm{d}y < \infty,$$

then dividing (2.21) by $\rho(\log x)$ and sending $x \to \infty$ completes the proof. □

2.4.3 The case with nonzero centering.

Now we turn to a more intricate case when some centering is needed. We denote by $F(x)$ the distribution function of η, and by $U(x)$ the renewal function of (S_k).

We will see that a major part of the variability of $N(x)$ is absorbed by the *renewal shot-noise* process $(M(x))_{x\geq 0}$, where

$$M(x) := \sum_{k=0}^{\rho(x)-1} F(x - S_k), \quad x \geq 0,$$

is the conditional expectation of $N(x)$ given (S_k).

Lemma 29. *We have*

$$\mathbb{E}\Big(N(x) - M(x)\Big)^2 = \int_0^x F(x-y)(1-F(x-y))\mathrm{d}\,U(y),$$

which implies that, as $x \to \infty$,

$$\mathbb{E}\Big(N(x) - M(x)\Big)^2 = O\bigg(\int_0^x (1-F(y))\mathrm{d}y\bigg) = o(x). \tag{2.22}$$

Proof. For integer $i < j$,

$$\begin{aligned}
&\mathbb{E}\Big(1_{\{S_i \le x\}}(1_{\{S_i+\eta_{i+1}\le x\}} - F(x-S_i)) \times \\
&\quad \times \;\; 1_{\{S_j \le x\}}(1_{\{S_j+\eta_{j+1}\le x\}} - F(x-S_j))\Big| (\xi_k,\eta_k)_{k=1}^{j} \Big) \\
&\quad = \;\; 1_{\{S_i \le x\}}(1_{\{S_i+\eta_{i+1}\le x\}} - F(x-S_i)) 1_{\{S_j \le x\}} \Big(F(x-S_j) - F(x-S_j) \Big) \\
&\quad = \;\; 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}\Big(N(x) - M(x) \Big)^2 &= \mathbb{E}\Big(\sum_{k=0}^{\infty} 1_{\{S_k \le x\}} \Big(1_{\{S_k+\eta_{k+1}\le x\}} - F(x-S_k) \Big) \Big)^2 \\
&= \mathbb{E} \sum_{n=0}^{\infty} 1_{\{S_k \le x\}} \Big(1_{\{S_k+\eta_{k+1}\le x\}} - F(x-S_k) \Big)^2 \\
&= \mathbb{E} \sum_{k=0}^{\infty} 1_{\{S_k \le x\}} \Big(F(x-S_k) - F^2(x-S_k) \Big) \\
&= \int_0^x F(x-y)(1-F(x-y)) \mathrm{d}\, U(y).
\end{aligned}$$

If $\mathbb{E}\eta < \infty$, then by the key renewal theorem, as $x \to \infty$,

$$\lim_{x\to\infty} \mathbb{E}\Big(N(x) - M(x) \Big)^2 \;=\; a^{-1} \int_0^{\infty} F(y)(1-F(y)) \mathrm{d}y < \infty,$$

where $a := \mathbb{E}\xi$ may be finite or infinite. If $\mathbb{E}\eta = \infty$ and $a < \infty$, the generalization of the key renewal theorem due to Sgibnev [130, Theorem 4] yields

$$\mathbb{E}\Big(N(x) - M(x) \Big)^2 \;\sim\; a^{-1} \int_0^x (1-F(y)) \mathrm{d}y.$$

Finally, if $\mathbb{E}\eta = \infty$ and $a = \infty$ a modification of Sgibnev's proof gives

$$\mathbb{E}\Big(N(x) - M(x) \Big)^2 \;=\; o\Big(\int_0^x (1-F(y)) \mathrm{d}y \Big).$$

Thus (2.22) follows in any case.

□

Theorem 30. *If for some random variable Z*

$$\frac{\rho(x)-g(x)}{f(x)} \stackrel{d}{\to} Z, \quad x\to\infty, \tag{2.23}$$

then also

$$\frac{M(x)-\int_0^x g(x-y)\mathrm{d}\,F(y)}{f(x)} \stackrel{d}{\to} Z, \quad x\to\infty, \tag{2.24}$$

$$\frac{N(x)-\int_0^x g(x-y)\mathrm{d}\,F(y)}{f(x)} \stackrel{d}{\to} Z, \quad x\to\infty, \tag{2.25}$$

and

$$\frac{R(x)-\int_0^{\log x} g(\log x-y)\mathrm{d}\,F(y)}{f(\log x)} \stackrel{d}{\to} Z, \quad x\to\infty. \tag{2.26}$$

Proof. Integration by parts yields

$$M(x)=\int_0^x F(x-y)\mathrm{d}\rho(y) = -F(x)+\int_0^x \rho(x-y)\mathrm{d}F(y),$$

so to prove (2.24) it is enough to show that, as $x\to\infty$,

$$T(x) := \int_0^x \frac{\rho(x-y)-g(x-y)}{f(x)}\,\mathrm{d}F(y) \stackrel{d}{\to} Z.$$

For any fixed $\delta\in(0,x)$ we may decompose $T(x)$ as

$$\begin{aligned} T_1(x)+T_2(x) &:= \int_0^\delta \frac{\rho(x-y)-g(x-y)}{f(x)}\,\mathrm{d}F(y) \\ &+ \int_\delta^x \frac{\rho(x-y)-g(x-y)}{f(x)}\,\mathrm{d}F(y). \end{aligned} \tag{2.27}$$

From the proof of Lemma 26 we know that without loss of generality it can be assumed that $g(x)$ is nondecreasing. Thus, almost surely,

$$\begin{aligned} \frac{\rho(x)-g(x)}{f(x)}F(\delta) &- \frac{\rho(x)-\rho(x-\delta)}{f(x)}F(\delta) \\ &\le T_1(x) \\ &\le \frac{\rho(x)-g(x)}{f(x)}F(\delta)+\frac{g(x)-g(x-\delta)}{f(x)}F(\delta). \end{aligned}$$

In view of (2.12) and (2.14), we have the convergence $\lim_{\delta\to\infty}\lim_{x\to\infty} T_1(x)=Z$ in distribution.

For $x > 0$ set

$$Z_x(t) := \frac{\rho(tx) - g(tx)}{f(x)}, \quad t \geq 0$$

and $\mathcal{Z}_x := (Z_x(t))_{t\geq 0}$. We will establish next that there exists a random process $\mathcal{Z} = (Z(t))_{t\geq 0}$ such that

$$\frac{\sup\limits_{y\in[0,x]} (\rho(y) - g(y))}{f(x)} = \sup_{t\in[0,1]} Z_x(t) \xrightarrow{d} \sup_{t\in[0,1]} Z(t), \quad x \to \infty, \tag{2.28}$$

and, similarly,

$$\frac{\inf\limits_{y\in[0,x]} (\rho(y) - g(y))}{f(x)} = \inf_{t\in[0,1]} Z_x(t) \xrightarrow{d} \inf_{t\in[0,1]} Z(t), \quad x \to \infty. \tag{2.29}$$

Case 1: If $g(x) = x/\mathbb{E}\xi$ then Z is an α-stable random variable for some $\alpha \in (1, 2]$. Denote by $\mathcal{Z} = (Z(t))_{t\geq 0}$ a stable Lévy process such that $Z(1) \stackrel{d}{=} Z$. Regard $\mathcal{Z}_x$ and $\mathcal{Z}$ as random elements of Skorohod's space $D[0, \infty)$ endowed with the M_1-topology.

By [23, Theorem 1b],

$$\mathcal{Z}_x \Rightarrow \mathcal{Z}, \quad x \to \infty. \tag{2.30}$$

Since the supremum and infimum functionals are M_1-continuous, we obtain (2.28) and (2.29) using the continuous mapping theorem.

Case 2: If $g(x)$ cannot be chosen to be f-equivalent to $x \to x/\mathbb{E}\xi$ (which is just zero in the case $\mathbb{E}\xi = \infty$), then Z is a 1-stable random variable. Set $\mathcal{Z} = (Z(t))_{t\geq 0}$, where

$$Z(t) = \hat{Z}(t) - t \log t, \quad t \geq 0,$$

and $(\hat{Z}(t))_{t\geq 0}$ is a stable Lévy process such that $\hat{Z}(1) \stackrel{d}{=} Z$. With this notation we derive (2.30) from [35, Theorem 2], from which (2.28), (2.29) follow along the above lines.

Now it remains to estimate

$$\begin{aligned}\frac{\inf\limits_{y\in[0,x]}(\rho(y)-g(y))}{f(x)}(F(x)-F(\delta)) &\le \frac{\inf\limits_{y\in[0,x-\delta]}(\rho(y)-g(y))}{f(x)}(F(x)-F(\delta))\\ &\le T_2(x)\\ &\le \frac{\sup\limits_{y\in[0,x]}(\rho(y)-g(y))}{f(x)}(F(x)-F(\delta)).\end{aligned}$$

Using (2.28) and (2.29), we conclude that $\lim\limits_{\delta\to\infty}\lim\limits_{x\to\infty} T_2(x)=0$ in probability. The proof of (2.24) is complete.

In view of (2.22), $\mathbb{E}(M(x)-N(x))^2=o(x)$. Since $f^2(x)$ grows not slower than x (see Remark 25), Chebyshev's inequality yields

$$\frac{N(x)-M(x)}{f(x)} \xrightarrow{P} 0, \quad x\to\infty.$$

Now (2.25) follows from (2.24).

It remains to establish (2.26). To this end, introduce for $x>1$

$$\begin{aligned}Q_1(x) &:= \int_1^x e^{-y}(N(\log x)-N(\log x-\log y))\mathrm{d}y \ge 0,\\ Q_2(x) &:= \int_0^1 e^{-y}(N(\log x-\log y)-N(\log x))\mathrm{d}y \ge 0.\end{aligned}$$

Using

$$\mathbb{E}N(x)=\int_0^x F(x-y)\mathrm{d}\,U(y) = -F(x)+\int_0^x U(x-y)\mathrm{d}\,F(y)$$

and (2.13), we conclude that for $y\in(1,x)$,

$$\mathbb{E}N(\log x)-\mathbb{E}N(\log x-\log y)\le C_1(1+F(0))\log y + C_2(1+F(0)).$$

Therefore, $\mathbb{E}Q_1(x)=O(1)$, as $x\to\infty$, whence $\frac{Q_1(x)}{f(\log x)}\xrightarrow{P}0$. Similarly, $\frac{Q_2(x)}{f(\log x)}\xrightarrow{P}0$. Thence, recalling (2.20)

$$\frac{Q_1(x)-Q_2(x)}{f(\log x)} = \frac{(1-e^{-x})N(\log x)-R(x)-(1-e^{-x})N(0)}{f(\log x)} \xrightarrow{P} 0,$$

as $x \to \infty$. From the inequality $N(x) \le \rho(x)$ a.s. and the weak law of large numbers for ρ it follows that $N(\log x)$ grows in probability not faster than $\log x$. Therefore

$$\frac{N(\log x) - R(x)}{f(\log x)} \xrightarrow{P} 0, \quad x \to \infty.$$

Now an appeal to (2.25) completes the proof. □

2.5 Proof of Theorem 23

Set

$$S_0^* := 0 \quad \text{and} \quad S_k^* := |\log W_1| + \ldots + |\log W_k|, \quad k \in \mathbb{N},$$

and

$$T_k^* := S_{k-1}^* + |\log(1 - W_k)|, \quad k \in \mathbb{N}.$$

The sequence $(T_k^*)_{k\in\mathbb{N}}$ is a perturbed random walk. Since

$$\rho^*(x) = \inf\{k \in \mathbb{N} : S_k^* > x\}, \quad N^*(\log x) := \#\{k \in \mathbb{N} : T_k^* \le \log x\},$$

an appeal to Theorem 27 in the case $g = 0$, and to Theorem 30 in the case $g \neq 0$ proves the result for $N^*(\log n)$. To prove the statement for K_n we will use the poissonization.

Step 1. We first check that

$$\lim_{t\to\infty} \mathbb{E}\,\mathrm{Var}(K(t)|(P_k)) = \frac{\log 2}{\mu}, \tag{2.31}$$

which is 0 for $\mu = \infty$. Plainly, this will imply that

$$\frac{K(t) - \mathbb{E}(K(t)|(P_k))}{q(t)} \xrightarrow{P} 0, \tag{2.32}$$

for any function $q(t)$ such that $\lim_{t\to\infty} q(t) = \infty$.

According to [84, formula (25)],

$$\mathrm{Var}(K(t)|(P_k)) = \sum_{k=1}^{\infty} \left(e^{-tP_k} - e^{-2tP_k}\right).$$

With $U^*(x) := \sum_{k=0}^{\infty} \mathbb{P}\{S_k^* \le x\}$ and $\varphi(t) := \mathbb{E}e^{-t(1-W)}$ we obtain

$$\begin{aligned}\mathbb{E}\operatorname{Var}(K(t)|(P_k)) &= \mathbb{E}\sum_{k=1}^{\infty}\Big(\varphi(te^{-S_{k-1}^*}) - \varphi(2te^{-S_{k-1}^*})\Big)\\ &= \int_0^{\infty}\Big(\varphi(te^{-x}) - \varphi(2te^{-x})\Big)\mathrm{d}U^*(x),\end{aligned}$$

which is the same as

$$\mathbb{E}\operatorname{Var}(K(e^x)|(W_k)) = \int_0^{\infty} A(x-y)\mathrm{d}U^*(y). \tag{2.33}$$

for $A(t) := \varphi(e^t) - \varphi(2e^t)$, $t \in \mathbb{R}$. To proceed, observe that

$$\int_0^{\infty} \frac{e^{-z(1-W)} - e^{-2z(1-W)}}{z}\,\mathrm{d}z = \log 2,$$

which implies that $A(t)$ is integrable, since by Fubini's theorem,

$$\begin{aligned}\int_{\mathbb{R}} A(t)\mathrm{d}t &= \int_0^{\infty}\frac{\varphi(z)-\varphi(2z)}{z}\,\mathrm{d}z\\ &= \mathbb{E}\int_0^{\infty}\frac{e^{-z(1-W)} - e^{-2z(1-W)}}{z}\,\mathrm{d}z = \log 2.\end{aligned}$$

Furthermore, arguing in the same way as in [56, Section 5] we can prove that $A(t)$ is directly Riemann integrable. Therefore, application of the key renewal theorem on $\mathbb{R}$ to (2.33) yields (2.31).

Chebyshev's inequality together with (2.31) imply that, for every $\varepsilon > 0$,

$$\lim_{t\to\infty}\mathbb{P}\{|K(t) - \mathbb{E}(K(t)|(P_k))| > \varepsilon q(t)|(P_k)\} = 0 \ \text{ in probability},$$

which proves (2.32) upon taking expectation and invoking the Lebesgue bounded convergence theorem.

Step 2. Step 1 implies that $(K(t) - g(t))/f(t)$ weakly converges to a proper and nondegenerate probability law if and only if

$$\frac{\mathbb{E}(K(t)|(P_k)) - g(t)}{f(t)} = \frac{R^*(t) - g(t)}{f(t)}$$

weakly converges to the same law.

Using this observation and exploiting Theorem 27 (in the case $g = 0$) or formula (2.26) of Theorem 30 (in the case $g \neq 0$) we conclude that weak convergence of $\frac{\rho^*(x)-g(x)}{f(x)}$ to some distribution θ implies the weak convergence of both

$$\frac{R^*(t) - \int_0^{\log t} g(\log t - y)\, \mathbb{P}\{|\log(1-W)| \in \mathrm{d}y\}}{f(\log t)}$$

and

$$\frac{K(t) - \int_0^{\log t} g(\log t - y)\, \mathbb{P}\{|\log(1-W)| \in \mathrm{d}y\}}{f(\log t)}$$

to θ.

Step 3. It remains to pass from the poissonized occupancy model to the fixed-n model. In view of (2.15),

$$b(t) := \int_0^{\log t} g(\log t - y)\, \mathbb{P}\{|\log(1-W)| \in \mathrm{d}y\}$$

satisfies

$$\lim_{t\to\infty} \frac{b(t) - b([t(1\pm\varepsilon)])}{f(\log t)} = 0$$

for every $0 < \varepsilon < 1$. Since $f(\log t)$ is slowly varying (see Remark 25 and Theorem 27), we have

$$X_\pm(t) := \frac{K(t) - b(\lfloor t(1\pm\varepsilon)\rfloor)}{f(\log(\lfloor t(1\pm\varepsilon)\rfloor))} \;\Rightarrow\; \theta.$$

Let C_t be the event that the number of balls thrown before time t lies in the limits from $\lfloor(1-\varepsilon)t\rfloor$ to $\lfloor(1+\varepsilon)t\rfloor$. By the monotonicity of K_n, we have

$$\begin{aligned} X_-(t) &\geq X_-(t)1_{C_t} \\ &\geq \frac{K_{\lfloor(1-\varepsilon)t\rfloor} - b(\lfloor t(1-\varepsilon)\rfloor)}{f(\log(\lfloor t(1-\varepsilon)\rfloor))} 1_{C_t}. \end{aligned}$$

Since $\mathbb{P}(C_t) \to 1$, we conclude that

$$\theta(x,\infty) \geq \limsup_{n\to\infty} \mathbb{P}\left\{\frac{K_n - b(n)}{f(\log n)} > x\right\},$$

for all $x \in \mathbb{R}$. To prove the converse inequality for the lower limit, one has to note that

$$X_+(t)1_{(C_t)^c} \;\xrightarrow{P}\; 0,$$

and proceed in the same manner. The proof of the theorem is complete.

Remark 31. Here is the promised verification of (2.10). We use terminology introduced in the proof of Lemma 26.

Lemma 32. *Relation* (2.10) *is a property of the class of f-equivalent functions g.*

Proof. Assume that g satisfies (2.10). We have to show that any g_1 such that $\lim_{x\to\infty} \frac{g(x)-g_1(x)}{f(x)} = 0$ satisfies (2.10), as well.

Plainly, it is enough to check that

$$A(x) := \frac{\int_0^x (g(x-y) - g_1(x-y))\mathrm{d}\, F(y)}{f(x)} \ \to 0, \quad x \to \infty. \tag{2.34}$$

For any $\varepsilon > 0$ there exists $x_0 > 0$ such that for all $x > x_0$ $\frac{|g(x)-g_1(x)|}{f(x)} < \varepsilon$. Since f is regularly varying with index $\beta \in [1/2, 1]$ (see Remark 25), without loss of generality we can assume that f is nondecreasing. Hence

$$\begin{aligned} |A(x)| &\le \int_0^{x-x_0} \frac{|g(x-y) - g_1(x-y)|}{f(x-y)}\mathrm{d}\, F(y) \\ &+ \int_{x-x_0}^{x} \frac{|g(x-y) - g_1(x-y)|}{f(x-y)}\mathrm{d}\, F(y) \\ &\le \varepsilon + \sup_{y\in[0,x_0]} \frac{|g(y) - g_1(y)|}{f(y)}(F(x) - F(x-x_0)). \end{aligned}$$

Sending $x \to \infty$ and then $\varepsilon \downarrow 0$ proves (2.34). □

If the law of $|\log W|$ belongs to the domain of attraction of an α-stable law, $\alpha \in (1, 2]$ then $(\rho(x) - g(x))/f(x)$ weakly converges with $g(x) = x/\mu$ and appropriate $f(x)$. Such a g trivially verifies (2.10) which, by Lemma 32, entails that every g_1 from the same f-equivalence class verifies (2.10).

If the law of $|\log W|$ belongs to the domain of attraction of a 1-stable law, then $(\rho(x) - g(x))/f(x)$ weakly converges for $g(x) = \frac{x}{m(r(x/m(x)))}$ and $f(x) = \frac{r(x/m(x))}{m(x)}$, with m and r as defined in part (d) of Corollary 24. Since r is regularly varying with index one, without loss of generality we can assume it and hence g are differentiable. Since $\frac{g(x)}{xf(x)}$ is regularly varying with index

(-1), it converges to 0, as $x \to \infty$. Besides, $\lim_{x\to\infty} x\mathbb{P}\{\zeta > x\} = 0$ in view of $\nu < \infty$, where we denoted $|\log(1-W)|$ by ζ. Hence,

$$\lim_{x\to\infty} \frac{g(x)\mathbb{P}\{\zeta > x\}}{f(x)} = 0.$$

Thus it suffices to check that

$$\lim_{x\to\infty} \frac{\mathbb{E}(g(x) - g(x-\zeta))1_{\{\zeta\le x\}}}{f(x)} = 0. \tag{2.35}$$

Now the subadditivity and differentiability of g can be exploited in order to show that

$$|g(x) - g(y)| \le c|x-y|, \quad x, y > 0,$$

where $c := 1/m(r(1))$. This immediately implies (2.35) and the whole claim by virtue of Lemma 32.

2.6 Number of empty boxes

Apart from K_n, other functionals which are the objects of intrinsic interest for infinite occupancy schemes and, in particular, for the Bernoulli sieve are

- M_n the index of the last occupied box;
- $L_n = M_n - K_n$ the number of empty intervals with indices not exceeding the M_n;
- $K_{n,r}$ the number of boxes occupied by exactly r balls;
- $Z_{n,k}$ the number of balls in the kth box in the left-to-right order of boxes.

Ultimate results for the weak convergence of M_n and $Z_{n,1}$ are known [56]. Under the assumption $\mu < \infty$ the weak convergence of $(K_{n,1}, \dots, K_{n,r})$ and $(Z_{n,1}, \dots, Z_{n,k})$ to some non-degenerate limiting random vectors is proved in [57]. The asymptotics of the $K_{n,r}$'s in the case $\mu = \infty$ will be discussed in Section 2.7 of this chapter.

The purpose of this section is a careful analysis of the asymptotics of $\mathbb{E}L_n$ (Theorem 33) which turns out to be an intricate analytical problem. Even though there is an explicit formula [56]

$$\mathbb{E}L_n = \sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k}\frac{1-\mathbb{E}(1-W)^k}{1-\mathbb{E}W^k}, \tag{2.36}$$

it does not seem possible to employ it in order to conclude on the asymptotic behavior of $\mathbb{E}L_n$ without restrictive additional assumptions. Also we prove that the distribution of L_n is geometric with parameter $1/2$ whenever $W \stackrel{d}{=} 1-W$ (Proposition 34).

2.6.1 Main results.

Theorem 33. *The expectation $\mathbb{E}L_n$ exhibits the following asymptotic behavior:*

(i) *If $\mu = \infty$ and $\nu = \infty$ then*

$$\liminf_{n\to\infty}\frac{\mathbb{E}W^n}{\mathbb{E}(1-W)^n} \le \liminf_{n\to\infty}\mathbb{E}L_n \le \limsup_{n\to\infty}\mathbb{E}L_n \le \limsup_{n\to\infty}\frac{\mathbb{E}W^n}{\mathbb{E}(1-W)^n}.$$

In particular,

$$\lim_{n\to\infty}\frac{\mathbb{E}W^n}{\mathbb{E}(1-W)^n} = \gamma_0 \in [0,\infty]$$

implies $\lim_{n\to\infty}\mathbb{E}L_n = \gamma_0$.

(ii) *If $\nu < \infty$ and $\mu \le \infty$ then*

$$\lim_{n\to\infty}\mathbb{E}L_n = \nu/\mu.$$

(iii) *If $\mu < \infty$ and $\nu = \infty$ then, as $n \to \infty$,*

$$\mathbb{E}L_n \sim \frac{1}{\mu}\int_1^n \frac{\mathbb{E}e^{-y(1-W)}}{y}\mathrm{d}y.$$

Proof. (i): Set $s_m = \frac{\mathbb{E}W^m}{\mathbb{E}(1-W)^m}$ and $Z_n := Z_{n,1}$. The following equality holds

$$\mathbb{P}\{Z_n = m\} = g_{n,m}\mathbb{P}\{Q_m(1) = 0\} = g_{n,m}\frac{\mathbb{E}(1-W)^m}{1-\mathbb{E}W^m}, \quad m \in \mathbb{N}, m \le n,$$

which implies the representation

$$\mathbb{E}L_n = \sum_{m=1}^{n} \frac{\mathbb{E}W^m}{\mathbb{E}(1-W)^m} g_{n,m} \frac{\mathbb{E}(1-W)^m}{1-\mathbb{E}W^m} = \mathbb{E}s_{Z_n}. \tag{2.37}$$

The array $c_{n,m} := \mathbb{P}\{Z_n = m\}$ verifies the conditions of Lemma 35. In particular, $\lim_{n\to\infty} c_{n,m} = 0$ in view of (2.6) and the assumption $\mu = \infty$. Hence, the first assertion of part (i) follows by applying that lemma with $t_n = \mathbb{E}L_n$. When γ_0 is well defined the proof is simpler, as in this case the statement follows from (2.37), divergence of Z_n (recall that $\lim_{n\to\infty} \mathbb{P}\{Z_n = m\} = 0$), by using the dominated convergence theorem in the case $\gamma_0 < \infty$ and Fatou's lemma in the case $\gamma_0 = \infty$.

See [56] and [57] for the proof of (ii).

(iii): We use the same poissonized version of the Bernoulli sieve and the same notation as on p. 46. In particular, the poissonized version of L_n is denoted by $L(t)$. Denote by $\pi_t^{(k)}$ the number of balls falling in box k, so that $\pi_t = \sum_{k\geq 1} \pi_t^{(k)}$. Then, conditionally on (W_j), $(\pi_t^{(k)})_{t\geq 0}$ is a Poisson process with rate $P_k = W_1 \dots W_{k-1}(1-W_k)$, and for different boxes these Poisson processes are independent. From the representation

$$L(t) = \sum_{k\geq 1} 1_{\{\pi_t^{(k)}=0,\ \sum_{j\geq k+1}\pi_t^{(j)}\geq 1\}}$$

we conclude that

$$\begin{aligned} \mathbb{E}(L(t)|(W_k)) &= \sum_{k\geq 1} e^{-tP_k}(1 - e^{-t(1-P_1-\dots-P_k)}) \\ &= \sum_{k=1}^{\infty} \left(e^{-tW_1\cdot\dots\cdot W_{k-1}(1-W_k)} - e^{-tW_1\cdot\dots\cdot W_{k-1}}\right). \end{aligned}$$

Recall the notation

$$\rho(x) := \inf\{k \in \mathbb{N} : S_k > x\}, \quad x \geq 0,$$

where $(S_k)_{k\in\mathbb{N}_0}$ is a zero-delayed random walk with generic step $|\log W|$. Set $\varphi(x) := \mathbb{E}e^{-x(1-W)}$ and

$$U(x) := \mathbb{E}\rho(x) = \sum_{k=0}^{\infty} \mathbb{P}\{S_k \leq x\}.$$

Then we have

$$\begin{aligned}\mathbb{E}L(t) &= \mathbb{E}\sum_{k=1}^{\infty}\Big(\varphi(te^{-S_{k-1}}) - \exp(-te^{-S_{k-1}})\Big)\\ &= \int_0^{\infty}\Big(\varphi(te^{-x}) - \exp(-te^{-x})\Big)U(\mathrm{d}x) \qquad (2.38)\\ &= \sum_{k=1}^{\infty}(-1)^{k+1}\frac{t^k}{k!}\frac{(1-\mathbb{E}(1-W)^k)}{1-\mathbb{E}W^k}, \qquad (2.39)\end{aligned}$$

where the familiar formula for Laplace transform of the renewal measure

$$\int_0^{\infty} e^{-sx}U(\mathrm{d}x) = \frac{1}{1-\mathbb{E}W^s}, \quad s>0,$$

has been utilised. Note that (2.39) is an obvious counterpart of (2.36).

Set $B(x) = \varphi(e^x) - \exp(-e^x)$, $x \in \mathbb{R}$. Since $\nu = \infty$ and

$$\int_0^{\infty}\frac{e^{-z(1-W)} - e^{-z}}{z}\mathrm{d}z = |\log(1-W)|,$$

we conclude that

$$\lim_{t\to\infty}\int_{-\infty}^{t} B(z)\mathrm{d}z = \infty. \qquad (2.40)$$

Applying a minor extension of [130, Theorem 5] to the equality

$$\mathbb{E}L(e^t) = \int_0^{\infty} B(t-x)U(\mathrm{d}x), \qquad (2.41)$$

which is equivalent to (2.38), yields

$$\mathbb{E}L(e^t) \sim \frac{1}{\mu}\int_0^t \varphi(e^x)\mathrm{d}x \sim \frac{1}{\mu}\int_1^{e^t}\frac{\varphi(x)}{x}\mathrm{d}x, \quad t\to\infty.$$

The asymptotics of $\mathbb{E}L_n$ is now obtained by the depoissonization Lemma 36. The lemma is applicable because $\mathbb{E}L(t)$ is slowly varying. Indeed, slow variation of $\int_1^t \varphi(u)\mathrm{d}u/u$ is checked straightforwardly from $\varphi(t) \downarrow 0$ and the divergence of the integral for $t = \infty$. □

In the case $\nu < \infty$ it is known that L_n converges in distribution to some proper random variable [56, Theorem 2.2]. However this result only gives

implicit specification of the limiting law through the distributions of L_n's, which are not easy to determine, with one remarkable exception. Obviously from the recursive construction of the Bernoulli sieve, the distribution of L_1 is geometric with parameter $\mathbb{E}W$. Curiously, the same is true for all $n \in \mathbb{N}$ provided the law of W is symmetric around the midpoint $1/2$.

Proposition 34. *If $W \stackrel{d}{=} 1 - W$ then L_n is geometrically distributed with parameter $1/2$, for all $n \in \mathbb{N}$.*

Proof. The argument is based on the recurrence (2.5) for marginal distributions of the L_n's. The symmetry $W =_d 1 - W$ yields $\mathbb{E}W^k = \mathbb{E}(1-W)^k$, for all $k \in \mathbb{N}$ and

$$\mathbb{P}\{Q_n^*(1) = n\} = \mathbb{P}\{Q_n^*(1) = 0\}, \tag{2.42}$$

for all $n \in \mathbb{N}$. We will show by induction on n that $\mathbb{P}\{L_n = k\} = 2^{-k-1}$, for all $k \in \mathbb{N}_0$. Using (2.5) and (2.42) we obtain

$$\begin{aligned}\mathbb{P}\{L_n = 0\} &= \mathbb{P}\{Q_n^*(1) = 0\} + \sum_{k=1}^{n-1} \mathbb{P}\{L_k = 0\}\mathbb{P}\{Q_n^*(1) = k\} \\ &= \mathbb{P}\{Q_n^*(1) = 0\} + \frac{1}{2}\Big(1 - 2\mathbb{P}\{Q_n^*(1) = 0\}\Big) = \frac{1}{2},\end{aligned}$$

by the induction hypothesis. Assuming now that $\mathbb{P}\{L_n = i\} = 2^{-i-1}$, for all $i < k$, we have

$$\begin{aligned}\mathbb{P}\{L_n = k\} &= \sum_{j=1}^{n-1} \mathbb{P}\{Q_n^*(1) = j\}\mathbb{P}\{L_j = k\} \\ &+ \mathbb{P}\{Q_n^*(1) = n\}\mathbb{P}\{L_n = k-1\} \\ &= 2^{-k-1}\Big(1 - 2\mathbb{P}\{Q_n^*(1) = 0\}\Big) \\ &+ \mathbb{P}\{Q_n^*(1) = 0\}2^{-k} = 2^{-k-1},\end{aligned}$$

and the proof is complete. □

Alternatively, one can use a representation of L_n through the sojourns of the Markov chain Q_n^* in positive states. Indeed, recall that L_1 has geometric distribution with parameter $\mathbb{E}W$. Then using (2.42) and induction it can be checked that the distribution of L_n does not depend on $n \in \mathbb{N}$.

2.6.2 Auxiliary results.

For ease of reference we include a result due to Toeplitz and Schur (see [69], Theorem 2 on p. 43 and Theorem 9 on p. 52). We rewrite it in a form which is suitable for our purposes.

Lemma 35. *Let $(s_n)_{n\in\mathbb{N}}$ be any sequence of real numbers and let $(c_{n,m})_{n,m\in\mathbb{N}}$ be a nonnegative array. Define another sequence $(t_n)_{n\in\mathbb{N}}$ by $t_n = \sum_{m=1}^{n} c_{n,m}s_m$. If*

(i) $\lim_{n\to\infty} c_{n,m} = 0$ *for all* m,

(ii) $\lim_{n\to\infty} \sum_{m=1}^{n} c_{n,m} = 1$,

then

$$\liminf_{n\to\infty} s_n \le \liminf_{n\to\infty} t_n \le \limsup_{n\to\infty} t_n \le \limsup_{n\to\infty} s_n \le +\infty.$$

Now we address the issue of depoissonization. Recall that the function $\mathbb{E}L(t)$ is slowly varying.

Lemma 36. *If* $\lim_{t\to\infty} \mathbb{E}L(t) = +\infty$ *then* $\mathbb{E}L_n \sim \mathbb{E}L(n)$, as $n \to \infty$.

Proof. For any fixed $\varepsilon \in (0,1)$,

$$\mathbb{E}L(t) \;=\; \mathbb{E}L(t)1_{\{|\pi_t - t| > \epsilon t\}} + \mathbb{E}L(t)1_{\{|\pi_t - t| \le \epsilon t\}} =: A_1(t) + A_2(t),$$

where (π_t) is the Poisson process such that $L(t) = L_{\pi_t}$ (see p. 46 for more details). The sublinearity of $\mathbb{E}L(t)$ and the elementary large deviation bound for Poisson processes [9],

$$\mathbb{P}\{|\pi_t - t| > \epsilon t\} \le c_1 e^{-c_2 t},\ t > 0,$$

with some $c_1, c_2 > 0$, yield $A_1(t) \to 0$, $t \to \infty$.

It remains to evaluate $A_2(t)$. Since both M_n and K_n are non-decreasing, we have

$$A_2(t) = \mathbb{E}(M(t) - K(t))1_{\{|\pi_t - t| \le \varepsilon t\}} \le \mathbb{E}L_{[(1-\varepsilon)t]} + \mathbb{E}(M_{[(1+\varepsilon)t]} - M_{[(1-\varepsilon)t]}).$$

Similarly, $A_2(t) \ge \mathbb{E}L_{[(1+\varepsilon)t]} - \mathbb{E}(M_{[(1+\varepsilon)t]} - M_{[(1-\varepsilon)t]})$. The first step is to prove that

$$\lim_{\varepsilon\downarrow 0} \lim_{n\to\infty} \mathbb{E}(M_{[(1+\varepsilon)n]} - M_n) = 0. \tag{2.43}$$

Setting $T_n = \max(E_1, \ldots, E_n)$ and using the subadditivity of the renewal function U we obtain

$$\begin{aligned} D(n) &:= \mathbb{E}\Big(M_{[(1+\varepsilon)n]} - M_n\Big) = \mathbb{E}\Big(U(T_{[(1+\varepsilon)n]}) - U(T_n)\Big) \\ &\leq \mathbb{E}U(T_{[(1+\varepsilon)n]} - T_n)1_{\{T_{[(1+\varepsilon)n]} > T_n\}}. \end{aligned}$$

An appeal to estimate $U(x) \leq ax + b$, with some $a, b > 0$, allows us to conclude that

$$\begin{aligned} D(n) &\leq \mathbb{E}\Big(a(T_{[(1+\varepsilon)n]} - T_n) + b\Big)1_{\{T_{[(1+\varepsilon)n]} > T_n\}} \\ &\leq a(H_{[(1+\varepsilon)n]} - H_n) + b\mathbb{P}\{T_{[(1+\varepsilon)n]} > T_n\}, \end{aligned}$$

where the equality $\mathbb{E}T_n = H_n := \sum_{k=1}^{n} \frac{1}{k}$ has been utilized. Since

$$\lim_{\varepsilon \downarrow 0} \lim_{n \to \infty} (H_{[(1+\varepsilon)n]} - H_n) = 0$$

and, by exchangeability,

$$\mathbb{P}\{T_{[(1+\varepsilon)n]} > T_n\} = 1 - \frac{n}{[(1+\varepsilon)n]} \to \varepsilon,$$

as $n \to \infty$, we arrive at (2.43).

We are ready to finish the proof. Divide the inequality

$$\mathbb{E}L(n/(1-\varepsilon)) \leq A_1(n/(1-\varepsilon)) + \mathbb{E}L_n + \mathbb{E}(M_{[\frac{1+\varepsilon}{1-\varepsilon}n]} - M_n),$$

by $\mathbb{E}L(n)$. Letting $n \to \infty$ then $\varepsilon \to 0$ and using the slow variation of $\mathbb{E}L(n)$ we obtain $\liminf_{n\to\infty} \frac{\mathbb{E}L_n}{\mathbb{E}L(n)} \geq 1$. The upper bound follows in the same way from the inequality

$$\mathbb{E}L(n/(1+\varepsilon)) \geq \mathbb{E}L_n - \mathbb{E}(M_n - M_{[\frac{1-\varepsilon}{1+\varepsilon}n]}).$$

□

2.7 Small parts

Recall that $K_{n,r}$ is the number of boxes occupied by exactly r balls in the Bernoulli sieve, $r \in \mathbb{N}$, $r \leq n$. Proposition 38 given below complements previously known results [57] about the weak convergence $K_{n,r}$.

Let $(p_k)_{k\in\mathbb{N}}$ be a probability mass function. Consider the classical multinomial occupancy scheme in which n balls are thrown independently in boxes, with probability p_k of hitting box $k = 1, 2, \ldots$ The expected number $\varkappa_{n,r}$ of boxes occupied by exactly r out of n balls is

$$\varkappa_{n,r} = \binom{n}{r} \sum_{j\geq 1} p_j^r (1-p_j)^{n-r}, \quad 1 \leq r \leq n. \tag{2.44}$$

Lemma 37. *For fixed $r < s$ there exists a constant c such that*

$$\varkappa_{n,r} \geq c\, \varkappa_{2n,s}, \quad n \in \mathbb{N}. \tag{2.45}$$

Proof. Using $(1-x)^{-1} \geq e^x$ for $x \in (0,1)$, we obtain

$$\begin{aligned} \frac{\binom{n}{r} x^r (1-x)^{n-r}}{\binom{2n}{s} x^s (1-x)^{2n-s}} &\geq c_1 \frac{s!}{2^s r!} (nx)^{r-s} (1-x)^{s-r-n} \\ &\geq c_2 (nx)^{r-s} e^{nx/2} \\ &\geq c_2 \min_{y>0} y^{r-s} e^{y/2} \\ &= c_2 \left(\frac{e}{2(s-r)}\right)^{s-r}. \end{aligned}$$

□

The result extends immediately to the case of random (P_k), particularly, to those of the form (2.1). This generalization is used in the proof of Proposition 38.

Proposition 38. (a) *If $\mu < \infty$ then, for every $r \in \mathbb{N}$, the vector $(K_{n,1}, K_{n,2}, \ldots, K_{n,r})$ converges weakly, as $n \to \infty$, to a proper multivariate distribution, and $\lim_{n\to\infty} \mathbb{E}K_{n,r} = (\mu r)^{-1}$.*

(b) *If $\mu = \infty$ then, for every $r \in \mathbb{N}$, $\lim_{n\to\infty} \mathbb{E}K_{n,r} = 0$, so $\lim_{n\to\infty} K_{n,r} = 0$ in probability.*

Proof. Part (a) was proved in [57, Theorem 3.3].

We first prove (b) for $r = 1$, the result for $r \geq 2$ follows from Lemma 37. The frequencies P_k's given by (2.1) can be expressed as

$$P_k = (1 - W_k) \exp(-S_{k-1}), \tag{2.46}$$

where $(S_k)_{k\in\mathbb{N}_0}$ is a random walk with $S_0 = 0$ and the generic step $|\log W|$.

Formula (2.44) gives

$$\mathbb{E}(K_{n,1}|(W_k)_{k\in\mathbb{N}}) = n\sum_{j\geq 1} P_j(1-P_j)^{n-1}.$$

Using $1 - x \leq e^{-x}$ for $x \in [0,1]$, equality (2.46) and substituting e^z for n, we reduce estimating $\mathbb{E}K_{n,1}$ to estimating

$$\begin{aligned}\mathbb{E}\sum_{j\geq 1} e^z P_j e^{-e^z P_j} &= \mathbb{E}\sum_{j\geq 1}(1-W_j)\exp\{z - S_{j-1} - e^{z-S_{j-1}}(1-W_j)\} \\ &= \int_0^\infty f(z-y)\mathrm{d}U(y),\end{aligned}$$

where $f(y) := \mathbb{E}\{(1-W)\exp(y - e^y(1-W))\}$ and $U(y) := \sum_{j\geq 0}\mathbb{P}\{S_j \leq y\}$ is the renewal function of (S_j). The function f is nonnegative and integrable, since $\int_{-\infty}^{\infty} f(y)\mathrm{d}y = 1$. Furthermore, the function $y \to e^{-y}f(y)$ is nonincreasing. It is known that these properties together ensure that f is directly Riemann integrable (see, for instance, the proof of Corollary 2.17 in [43]). When $\mu = \infty$, application of the key renewal theorem yields

$$\lim_{z\to\infty}\int_0^\infty f(z-y)U(\mathrm{d}y) = 0,$$

whence $\lim_{n\to\infty}\mathbb{E}K_{n,1} = 0$. □

2.8 Bibliographic comments

The model of *stick-breaking partitions* called the *Bernoulli sieve* was introduced by Alexander Gnedin in [63]. The Bernoulli sieve is a generalization of at least three models of applied probability.

Leader election procedure. If the law of W is degenerate at some point $p \in (0,1)$, then the Bernoulli sieve reduces to the *leader election procedure*. Assume that n persons play the following game. In the first round of the game each of the players tosses once a coin with probability p for heads,

and while those who throw tails are eliminated, those who obtain heads play the second round with the same probability p of getting heads and so on until there are no players. If in some round all remaining players get heads, the round is deemed inconclusive and must be repeated, as many times as necessary until some players are eliminated. The leader election problem asks for the 'probability of a single winner' $\mathbb{P}\{Z_{n,1} = 1\}$. Also of interest are the number M_n of rounds until the leader(s) is (are) elected, the number L_n of inconclusive rounds and some other quantities. These characteristics were studied by many authors and in different (but equivalent) contexts: leader election algorithm [45, 46, 83, 90, 91, 92, 93, 98, 116], extremes in the geometric samples [6, 13, 27, 29, 30, 64, 70], etc.

Cycles in random permutations. Assume that the law of W is beta$(1, \theta)$ with some $\theta > 0$. In this case the Bernoulli sieve is closely related to the cycles of random permutations. Consider a random vector $C^{(n)} := (C_{n,1}, \dots, C_{n,n})$, where $C_{n,j}$ is the number of cycles of length j in the θ-biased random permutation of set $\{1, \dots, n\}$ (see [7] for exhaustive information on random permutations and related fields). Then

$$C^{(n)} \stackrel{d}{=} (K_{n,1}, \dots, K_{n,n}),$$

where $K_{n,r}$ is the number of boxes occupied by exactly r out of n balls in the Bernoulli sieve. The distribution of $C^{(n)}$ is given by the formula:

$$\mathbb{P}\{C^{(n)} = (c_1, \dots, c_n)\} = 1_{\{\sum_{j=1}^n j c_j = n\}} \frac{n!}{\theta^{(n)}} \prod_{j=1}^{n} \left(\frac{\theta}{j}\right)^{c_j} \frac{1}{c_j!}, \quad c_j = 0, 1, \dots, n,$$

which is widely known as *Ewens sampling formula*. Other classical results concerning the cycles of random permutations include the asymptotic independence of the number of small cycles, the weak convergence of small cycles to Poisson laws, the asymptotic normality of the total number of cycles.

Infinite occupancy schemes. An occupancy scheme in which n balls are thrown in an infinite array of boxes with *deterministic* probabilities p_k of

hitting box $k = 1, 2, \ldots$ is known as *Karlin's occupancy scheme*. Although such a scheme was partly investigated in [8, 34], the first systematic treatment is due to S. Karlin [84]. In particular, in the last cited work the central limit theorem for the number of occupied boxes was proved. Later on a more general result in that direction was obtained in [44]. A comprehensive survey of the results related to the infinite occupancy scheme can be found in [51]. The Bernoulli sieve is the Karlin's occupancy scheme with random frequencies (2.1).

Several papers were published which investigated the weak convergence of K_n, the number of occupied boxes in the Bernoulli sieve. Under the assumptions $\sigma^2 < \infty$ and $\nu < \infty$, the central limit theorem was proved in [63]. Under the sole assumption $\nu < \infty$, this result was subsequently generalized in [56] where the whole set of possible limiting laws for K_n was obtained. Unlike the original Gnedin's [63] proof which was analytic in nature and based on the careful analysis of random recurrences, the approach in [56] was mainly probabilistic and relied on the connection of K_n and $\rho^*(x)$ defined in (2.7). Yet another probabilistic argument exploited in this thesis is based on studying the asymptotics of small frequencies. Even though such an approach is familiar from [12, 60, 84], the application to the Bernoulli sieve is new. We emphasize that the connection of K_n and $N^*(x)$ defined in (2.3) is more fundamental that that of K_n and $\rho^*(x)$ exploited in [56], since $N^*(x)$ is not sensitive to the arrangement of boxes in some order, as compared to $\rho^*(x)$ involving explicitly the ordered features of the environment.

The small parts $K_{n,r}$ have received some attention in the literature, too. See [57] for results concerning the Bernoulli sieve and [12, 128] and references therein, for general occupancy schemes.

The *perturbed random walks*, as defined in this thesis, arise in diverse areas of applied probability. In particular, these are closely related to perpetuities, shot noise processes, regenerative processes, GI/G/∞ queues, etc. The perturbed random walks were studied in different contexts (see, for example, [5], [67, Chapter 6], [81], [110]), but the general theory of these objects has not

been developed so far (a work in progress [2] intends to contribute to such a theory, at least to some extent). As a rule, questions about the asymptotics of perturbed random walks were circumvented in the literature by imposing an appropriate moment condition which allowed reduction to standard random walks (see, for instance, [67, Chapter 6], [71], [118, Theorems 2.1 and 2.2]). We are unaware of any results in the spirit of those presented in Section 2.4.

While Sections 2.1-2.5 are based on [53], Section 2.6 is based on [52]. The results of Section 2.7 were proved in [54]

Chapter 3

Exchangeable coalescents

3.1 Lambda-coalescents with dust component

3.1.1 Preliminaries. The lambda-coalescent with values in partitions of n integers is a Markovian process $\Pi_n = (\Pi_n(t))_{t\geq 0}$ which starts at $t = 0$ with n singletons and evolves according to the rule: for each $t \geq 0$ when the number of clusters is m, each k tuple of them is merging in one cluster at probability rate

$$\lambda_{m,k} = \int_0^1 x^k (1-x)^{m-k} \theta(\mathrm{d}x), \qquad 2 \leq k \leq m, \tag{3.1}$$

where θ is a measure on the unit interval with finite second moment.

The integral representation of rates (3.1) ensures that the processes Π_n can be defined consistently for all n, as restrictions of a coalescent process Π_∞ which starts with infinitely many clusters and assumes values in the set of partitions of $\mathbb{N}$, see [113]. The infinite coalescent Π_∞ may be regarded as a limiting form of Π_n as $n \to \infty$, and uniquely connected to a process with values in the infinite-dimensional space of partitions of a unit mass. The lambda-coalescents were introduced in the papers by Pitman [113] and Sagitov [124], where the parameterization by finite measure $\Lambda(\mathrm{d}x) = x^2\theta(\mathrm{d}x)$ was

used. The reader is referred to the recent lecture notes [16, 18] for accessible introduction to the theory of lambda-coalescents and a survey.

After some number of collisions (merging events) Π_n enters the absorbing state with a sole cluster. Two basic characteristics of the speed of the coalescence are *the absorption time* τ_n and *the number of collisions* X_n. The large-n properties of τ_n and X_n are strongly determined by the concentration of measure θ on the unit interval near the endpoints of $[0, 1]$.

We suppose that θ has no mass at 1, which excludes forced termination of Π_∞ at an independent exponential time. The coalescent is said to come down from infinity if $\Pi_\infty(t)$ has finitely many clusters, for each $t > 0$, almost surely; then τ_n converge to a finite random variable τ_∞ which is the absorption time of Π_∞. Otherwise, $\Pi_\infty(t)$ almost surely stays with infinitely many clusters, for all t. There is a delicate criterion in terms of the rates $\lambda_{m,k}$ to distinguish between the two alternatives [127].

In this section we study τ_n and X_n under the assumption that Π_∞ stays infinite due to infinitely many original clusters which do not engage in collisions before any given time $t > 0$. This family of lambda-coalescents can be characterized by the moment condition

$$\int_0^1 x\,\theta(\mathrm{d}x) < \infty. \tag{3.2}$$

We call the collection of singleton clusters of $\Pi_\infty(t)$ *the dust component*. The dust component has a positive total frequency, meaning that the number of singletons within $\Pi_n(t)$ grows approximately linearly in n, as $n \to \infty$.

The coalescents with dust component do not exhaust all coalescents which stay infinite. One distinguished example is the Bolthausen-Sznitman coalescent with $\theta(\mathrm{d}x) = x^{-2}\mathrm{d}x$ which stays infinite although (3.2) fails. Such examples on the border between 'coming down from infinity' and 'possessing dust component' are more of an exception if one considers e.g. measures θ satisfying a condition of regular variation near zero.

Under (3.2) every transition of Π_∞ will involve infinitely many singletons. This suggests that most of the collision events of Π_n will involve some of the

original n clusters, for large n. Another way to express this idea is to say that in a tree representing the complete merging history of Π_n, most of the internal nodes are linked directly to one of n leaves. We will show that this intuition is indeed correct, to the extent that the behaviour of τ_n and X_n can be derived from that of analogous quantities associated with the evolution of the dust component. In turn, the total frequency of the dust component of Π_∞ undergoes a relatively simple process, which may be represented as $\exp(-S_t)$, where $S = (S_t)_{t\geq 0}$ is a subordinator. Similarly for Π_n, the engagement of original n clusters in their first collisions follows a Markovian process which has been studied in the context of regenerative composition structures derived from subordinators [58]. A coupling of Π_∞ with S will enable us to apply known results about the level-passage for subordinators, and about the asymptotics of regenerative composition structures.

The connection between Π_∞ and S was first explored in [55] in the special case when θ is a finite measure, hence subordinator S is a compound Poisson process. While in the present paper we are mainly interested in infinite θ, the case of finite θ is not excluded. Moreover, we will be able to extend the results of [55] by removing a condition on θ imposed in that paper.

In a recent paper by Haas and Miermont [68] results on counting collisions in the coalescent and counting blocks in the regenerative composition were derived separately in the context of absorption times of decreasing Markov chains. Our approach adds some insight to the connection between these two models, and it entails some delicate features like differentiating between collisions which involve some original clusters of Π_n and the collisions which do not.

3.1.2 The coalescent and singleton clusters.

In the role of the state space of the coalescent Π_n with initially n clusters we take the set of partitions of $[n] := \{1, \ldots, n\}$, in which every singleton cluster is classified as either *primary* or *secondary*. Under the *dust component of* $\Pi_n(t)$ we shall understand the collection of primary clusters. Every nonsingleton cluster of $\Pi_n(t)$ is regarded as secondary. For the notational convenience the clusters

are written by increase of their minimal elements, the elements within the clusters are written in increasing order, and the secondary clusters are written in brackets. For instance, 1 (2) (3 5 6) 4 7, a partition of the set [7], has three primary clusters and two secondary: $1, 4, 7$ and $(2), (3\ 5\ 6)$, respectively.

Introduce $\lambda_{m,1}$ as in (3.1) with $k = 1$. We have $\lambda_{m,1} < \infty$ by assumption (3.2).

We define the lambda-coalescent Π_n as a cádlág Markov process with values in such partitions of $[n]$ and the initial state $1\ 2\ \cdots\ n$ with n primary clusters. Each admissible transition is either merging some clusters in one cluster, or turning a primary singleton cluster into secondary. From partition with m clusters, the transition rate for merging each particular k-tuple of m clusters in one is $\lambda_{m,k}$ $(2 \le k \le m)$, and the transition rate for turning each particular primary singleton cluster into secondary singleton cluster is $\lambda_{m,1}$. For instance, the sequence of distinct states visited by Π_7 could be

$$1\ 2\ 3\ 4\ 5\ 6\ 7 \to 1\ 2\ (3\ 5\ 6)\ 4\ 7 \to 1\ (2)\ (3\ 5\ 6)\ 4\ 7 \to$$
$$1\ (2\ 4)\ (3\ 5\ 6)\ 7 \to 1\ (2\ 3\ 4\ 5\ 6\ 7) \to (1\ 2\ 3\ 4\ 5\ 6\ 7).$$

Let $N_n(t)$ be the number of clusters in $\Pi_n(t)$. Then $N_n = (N_n(t))_{t\ge 0}$ is a nonincreasing Markov process, with the transition rate

$$\varphi_{m,k} := \binom{m}{k} \lambda_{m,k} \tag{3.3}$$

for jumping from m to $m-k+1$, for $2 \le k \le m$. Turning a primary singleton cluster into a secondary singleton cluster does not cause a jump of N_n. The absorption time of Π_n can be recast as $\tau_n = \inf\{t : N_n(t) = 1\}$, and the number of collisions X_n is equal to the number of jumps the process N_n needs to proceed from n to 1 (which is 4 in the above example where the second transition does not alter the number of clusters).

Removing element n transforms partition of $[n]$ into partition of $[n-1]$. For example, partitions 1 (2 4) (3), 1 (2) 3 4 and 1 (2) 3 (4) all become 1 (2) (3). Restricting in this way Π_n to $[n-1]$, pointwise in $t \ge 0$, yields a stochastic copy of Π_{n-1}. This follows as in [113] since the rates satisfy the

consistency relation

$$\lambda_{m,k} = \lambda_{m+1,k} + \lambda_{m+1,k+1}, \quad 1 \le k \le m.$$

Therefore we may define Π_n on the same probability space consistently for all n. Explicit realization will appear in the sequel.

The projective limit of the processes $\Pi_n,\ n \in \mathbb{N}$, is a Markov process Π_∞ starting at $t = 0$ with the infinite configuration of primary clusters $1\ 2\ \cdots$, and assuming values in the space of partitions of the infinite set $\mathbb{N}$. Each partition $\Pi_\infty(t)$ has only primary singletons, namely those original clusters which do not engage in collisions up to time t. For a generic singleton, e.g. labelled 1, the time before its first collision has exponential distribution with parameter $\lambda_{1,1}$, and when such a collision occurs infinitely many other clusters are engaged.

The differentiation of singletons of $\Pi_n(t)$ into primary and secondary becomes transparent by considering Π_n as restriction of Π_∞ on $[n]$. The secondary singletons of $\Pi_n(t)$ are the unique representatives in $[n]$ of some infinite clusters of $\Pi_\infty(t)$. The primary singletons of $\Pi_n(t)$ are also singletons in the partition $\Pi_\infty(t)$.

There is a construction of Π_∞ based on a planar Poisson point process in the strip $[0,1] \times [0,\infty)$ with intensity measure $\theta(\mathrm{d}x) \times \mathrm{d}t$, see [16, 18, 113]. With each atom (t,x) one associates a transition of Π_∞ performed by tossing a coin with probability x for heads. To pass from $\Pi_\infty(t-)$ to $\Pi_\infty(t)$, the coin is tossed for each cluster of $\Pi_\infty(t-)$, then those clusters marked heads are merged in one, while the clusters marked tails remain unaltered. Although there are infinitely many transitions within any time interval if θ is an infinite measure, condition (3.2) ensures that Π_∞ does not terminate. In the case of finite θ transitions of Π_∞ occur at the epochs of Poisson process with rate $\theta([0,1])$.

Let $N_n^*(t)$ be the number of primary clusters in $\Pi_n(t)$. By homogeneity properties of Π_n, the process $N_n^* = (N_n^*(t))_{t\ge 0}$ is a nonincreasing Markov

process, jumping at rate $\varphi_{m,k}$ from m to $m-k$ for $1 \le k \le m$. Let

$$\tau_n^* := \inf\{t : N_n^*(t) = 0\}$$

be the random time when the last of n primary clusters disappears. For $1 \le r \le n$, let $K_{n,r}$ be the number of decrements of size r of (N_n^*) on the way from n to 0, let $K_n := \sum_{r=1}^n K_{n,r}$ be the total number of decrements of (N_n^*), and let X_n^* be the number of non-unit decrements of (N_n^*). Obviously,

$$X_n^* = K_n - K_{n,1}. \tag{3.4}$$

We call the clusters of partition $\Pi_n(\tau_n^*)$ that remain at time τ_n^* *residual*, and we denote R_n the number of residual clusters.

Processes N_n and N_n^* look very similar, thus at a first glance it might seem surprising that N_n^* is much easier to handle. The simplification comes from the identification of the sequence of decrements of N_n^* with the nth level of a regenerative composition structure [58], and further connection to the range of a subordinator. We show that N_n^* yields a good approximation for N_n for large n, thus X_n^* and τ_n^* are close to their counterparts X_n and τ_n. In one direction, the connection is quite obvious:

$$X_n^* \le X_n, \quad N_n^*(t) \le N_n(t), \quad \tau_n^* \le \tau_n.$$

For instance, the first inequality holds since every collision taking at least two primary clusters contributes to X_n, and since with positive probability some $R_n \ge 2$ clusters remain at time τ_n^* when the last primary clusters disappears.

3.1.3 Coupling with a subordinator.

Condition (3.2) implies that there exists a pure-jump subordinator $S = (S_t)_{t\ge 0}$ with the Laplace transform

$$\mathbb{E}(e^{-zS_t}) = e^{-t\Phi(z)}, \quad z \ge 0, \tag{3.5}$$

where the Laplace exponent is given by

$$\Phi(z) := \int_0^1 (1-(1-x)^z)\theta(\mathrm{d}x).$$

The coalescent process will be represented in terms of passage of S through multiple exponentially distributed levels. We describe first the evolution of the dust component.

Let $\epsilon_1, \ldots, \epsilon_n$ be independent of S i.i.d. standard exponential random variables, and let $\epsilon_{n:n} < \cdots < \epsilon_{n:1}$ be their order statistics. It is not difficult to see that $\Phi(n)$ coincides with the probability rate at which the subordinator passes through the level $\epsilon_{n:n}$ from any state $S_t = s < \epsilon_{n:n}$. The following lemma extends this observation.

Lemma 39. *For $t \geq 0$, conditionally given $S_t = s$ with $s \in (\epsilon_{n:m+1}, \epsilon_{n:m})$ the subordinator is passing through $\epsilon_{n:m}$ at rate $\Phi(m)$, and is hitting at this passage each of the intervals $(\epsilon_{n:m-k+1}, \epsilon_{n:m-k})$ at rate $\varphi_{m,k}$, for $1 \leq k \leq m \leq n$.*

Proof. The proof exploits the Lévy-Khintchine formula (3.5) and the memoryless property of the exponential distribution. See computations around [58, Theorem 5.2] for details. □

Now suppose that each of the primary clusters $1\ 2 \ldots n$ is given an exponential mark $\epsilon_1, \ldots, \epsilon_n$, and that for every $t \geq 0$ the marks $\epsilon_j > S_t$ are associated with primary clusters j existing at time t. If t is a jump-time of S and the interval $(S_{t-}, S_t]$ covers exactly one mark ϵ_j, we interpret the event of passage through ϵ_j as turning the primary cluster j into secondary. If $(S_{t-}, S_t]$ covers at least two of the ϵ_j's, we interpret this event as a collision which takes the corresponding primary clusters. Setting $N_n^*(t) := \#\{j \in [n] : \epsilon_j > S_t\}$ we obtain a process with desired rates $\varphi_{m,k}$ for transition from m to $m-k$, as it follows from the lemma. In particular, $\Phi(n) = \sum_{k=1}^n \varphi_{n,k}$ coincides with the total transition rate of the coalescent Π_n from the initial state $1\ 2 \cdots\ n$.

A *regenerative* ordered partition of the set $[n]$ is defined by sending i, j to the same block iff $T_{\epsilon_i} = T_{\epsilon_j}$, see [58], where

$$T_s := \inf\{t \geq 0 : S_t > s\}$$

is the first passage time through level $s \geq 0$. The number of blocks of the

partition is equal to the number of jumps of N_n^* prior to the absorption at state 0.

These evolutions of primary clusters are consistent in n. Assigning the exponential marks $\epsilon_1, \epsilon_2, \ldots$ to infinitely many primary clusters 1 2 ... defines the initial state of the dust component. The frequency of the dust component of Π_∞ as time passes is the decaying process $(\exp(-S_t))_{t\geq 0}$.

One straightforward application of the representation by S concerns τ_n^*, the maximal lifetime of primary clusters in Π_n. We can identify τ_n^* with $T_{\epsilon_{n:1}}$, hence connect the limit behaviour of τ_n^* to that of T_s for high levels s. Indeed, from the extreme-value theory it is known that $\epsilon_{n:1} - \log n$ converges in distribution, as $n \to \infty$, to a random variable with the Gumbel distribution. It is also known that the scaled and centered random variables $(T_s - g(s))/f(s)$ can converge in distribution only if the normalizing constant $f(s)$ goes to ∞ with s. Thus, $T_{\epsilon_{n:1}}$ and $T_{\log n}$ have the same limit law, if any. Moreover, it can be shown that $(T_s - g(s))/f(s)$ converges weakly to a given proper and non-degenerate probability law if and only if the same holds for $(T'_s - g(s))/f(s)$, where T'_s is the number of points within $[0, s]$ of a random walk which starts at 0 and has the generic step distributed like S_1. See [22] or [103, Proposition 27] for a complete list of limit distributions of T'_s and the conditions of convergence. Summarizing the above, we have

Proposition 40. *For constants $a_n > 0$ and $b_n \in \mathbb{R}$, if one of the random variables $(\tau_n^* - b_n)/a_n$ and $(T_{\log n} - b_n)/a_n$ converges weakly, as $n \to \infty$, to a nondegenerate proper distribution, then the other random variable converges weakly to this distribution too.*

To realize the full dynamics of Π_n in terms of the level-passage, a mark is assigned to each cluster according to the following rule. At time 0 the marks $\epsilon_1, \ldots, \epsilon_n$ represent the primary clusters $1\ 2 \cdots n$. At time $t > 0$ there is some collection of marks on $[S_t, \infty)$ representing the clusters existing at this time. If at time $t > 0$ the subordinator passes through exactly k marks corresponding to some clusters $I_1, \ldots, I_k \subset [n]$, then a new cluster $I_1 \cup \cdots \cup I_k$ is born and assigned a mark $S_t + \epsilon$, where ϵ is a copy of the unit exponential random

variable, independent of S and all other marks assigned before t. For instance, if at the first passage time $t = T_{\epsilon_{n:n}}$ the subordinator jumps through exactly k levels $\epsilon_{j_1}, \ldots, \epsilon_{j_k}$ out of $\epsilon_1, \ldots, \epsilon_n$, then the secondary cluster $J = \{j_1, \ldots, j_k\}$ is born (which is a singleton if $k = 1$) and assigned a mark exponentially distributed on $[S_t, \infty)$.

In particular, when S passes at some time t through only one mark, there is no change in $\Pi_n(t)$, and the mark of the corresponding singleton cluster is just re-assigned.

3.1.4 The absorption time.

We wish to exploit the lifetime τ_n^* of primary clusters as approximation to the absorption time τ_n. At time τ_n^* the coalescent process is left with R_n residual clusters, whence the distributional identity

$$\tau_0 := 0, \quad \tau_n \stackrel{d}{=} \tau_n^* + \tilde{\tau}_{R_n}, \quad n \in \mathbb{N}, \tag{3.6}$$

where $\tilde{\tau}_m$ is assumed independent of (τ_n^*, R_n) and distributed like τ_m, for each $m \in \mathbb{N}_0$. To address the quality of approximation we need to estimate R_n.

We begin with some preparatory work. By the first transition the Markov chain N_n^* goes from n to a random state with distribution $p_{n,k} := \varphi_{n,n-k}/\Phi(n)$, $0 \le k \le n-1$. Let $g_{n,k}$ be the probability that N_n^* ever visits state k, so $g_{n,n} = 1$ and, for $1 \le k \le n-1$, in terms of the realization via subordinator, $g_{n,k} = \mathbb{P}(T_{\epsilon_{n:k+1}} < T_{\epsilon_{n:k}})$ is the probability that the interval $[\epsilon_{n:k+1}, \epsilon_{n:k}]$ intersects the range of S. An explicit formula for $g_{n,k}$ in terms of Φ is available (see [58, Eq. (50)]), but it is complicated and inconvenient for computations.

Lemma 41. *Suppose $(r_k)_{k\in\mathbb{N}}$ is a nonnegative sequence such that the sequence $\left(\frac{\Phi(k) r_k}{k}\right)_{k\in\mathbb{N}}$ is nonincreasing. Then the sequence $(a_n)_{n\in\mathbb{N}_0}$ defined by*

$$a_0 = 0, \quad a_n := \sum_{k=1}^{n} g_{n,k} r_k, \qquad n \ge 1$$

satisfies the relation

$$a_n = O\Big(\sum_{k=1}^{n} \frac{r_k \Phi(k)}{k}\Big), \quad n \to \infty.$$

Proof. The assertion follows from Lemma 15. Indeed, conditioning on the size of the first jump of N_n^* we see that the sequence (a_n) satisfies the recurrence

$$a_0 = 0, \quad a_n = r_n + \sum_{k=0}^{n-1} p_{n,k} a_k, \quad n \in \mathbb{N}.$$

To apply Lemma 15 we take $\psi_n = \Phi(n)$. Condition (C2) holds by the assumptions and condition (C1) follows from

$$\begin{aligned} \Phi(n) \sum_{k=0}^{n} (1 - k/n) p_{n,k} &= \frac{1}{n} \sum_{k=0}^{n-1} (n-k) \varphi_{n,n-k} \\ &= \frac{1}{n} \sum_{k=1}^{n} k \varphi_{n,k} = \int_0^1 x \theta(\mathrm{d}x) > 0. \end{aligned}$$

□

Note that, since the function $s \mapsto \Phi(s)/s$ is nonincreasing, the sequence $\left(\frac{\Phi(k) r_k}{k}\right)$ is nonincreasing whenever (r_k) is itself nonincreasing.

Denote $\vec{\theta}(x) := \theta([x, 1])$, $x \in (0, 1)$.

Lemma 42. *If either of two equivalent conditions*

$$\int_0^1 x^{-1} \mathrm{d}x \int_0^x \vec{\theta}(y) \mathrm{d}y < \infty, \tag{3.7}$$

$$\sum_{k=1}^{\infty} \frac{\Phi(k)}{k^2} < \infty \tag{3.8}$$

holds then

$$\mathbb{E} R_n = O(1), \quad n \to \infty,$$

in which case the sequence of distributions of the R_n's is tight.

Proof. The equivalence of (3.7) and (3.8) is established by repeated integration by parts.

In the genealogical history of each residual cluster there is the last secondary cluster appearing as a result of collision or switch involving some primary clusters. If secondary cluster b is born at some time $t \le \tau_n^*$ of such

an event, and if at this time some $j \geq 0$ other primary clusters co-exist, then b corresponds to a residual cluster provided that b and its followers do not collide with these j primary clusters or their followers before time τ_n^*. That is to say, b and the j primary clusters belong to distinct branches if the coalescent tree is cut at time τ_n^*. Let q_j be the probability that such cluster b corresponds to a residual cluster; restricting the coalescent to $j+1$ clusters it is seen that q_j indeed depends only on j. The consistency property of the coalescent with respect to the restrictions entails that q_j is decreasing in j. Averaging over the times when some primary clusters engaged we find the expected number of residual clusters

$$\mathbb{E}R_n = \sum_{j=0}^{n-1} g_{n,j} q_j. \tag{3.9}$$

Furthermore, given $S_t = s$, we have exactly j exponential marks of the primary clusters larger than s. The cluster b is assigned a new exponential mark $u = s + \epsilon$ which lies within each of the spacings in (s, ∞) generated by $\epsilon_{n:j}, \ldots, \epsilon_{n:1}$ with the same probability $1/(j+1)$. If this spacing is $(\epsilon_{n:k+1}, \epsilon_{n:k})$ then b *may* correspond to a residual cluster only if (i) $T_{\epsilon_{n:k+1}} < T_u < T_{\epsilon_{n:k}}$ and (ii) b does not collide further with k primary clusters and their followers before time τ_n^*. If (i) occurs, condition (ii) is not sufficient for the correspondence since possible collisions with some of j primary clusters or their followers are ignored. This leads to the inequality

$$q_j \leq \frac{1}{j+1} \sum_{k=0}^{j} g_{j+1,k+1} p_{k+1,k} q_k, \quad 1 \leq j \leq n-1,$$

and $q_0 = 1$. Substituting $\varphi_{k,1} = k(\Phi(k) - \Phi(k-1))$ we obtain

$$\begin{aligned} q_j &\leq \frac{1}{j+1} \sum_{k=1}^{j+1} g_{j+1,k} \frac{k(\Phi(k) - \Phi(k-1))}{\Phi(k)} q_{k-1} \\ &\leq \frac{c}{j+1} \sum_{k=1}^{j+1} (\Phi(k) - \Phi(k-1)) q_{k-1}, \end{aligned}$$

where Lemma 41 was applied with

$$r_k = \frac{k(\Phi(k) - \Phi(k-1)) q_{k-1}}{\Phi(k)}.$$

The required monotonicity condition holds since both q_k and $\Phi(k)-\Phi(k-1)$ are decreasing in k, the latter by concavity of Φ. Here and throughout c will denote a positive constant whose value is not important and may change from line to line.

Setting $a_j = (j+1)q_j$ and $b_j = c(\Phi(j+1)-\Phi(j))/(j+1)$, we obtain from the above

$$a_j \le \sum_{k=0}^{j} b_k a_k, \quad j \in \mathbb{N}_0.$$

We want to show that the sequence (a_j) is bounded. To that end, let $M_j := \max_{i=0,\dots,j} a_i$, then also

$$M_j \le \sum_{k=0}^{j} b_k M_k.$$

Since $\Phi(j)/j$ decreases we have $\Phi(j+1)-\Phi(j) \le \Phi(j+1)/(j+1)$, which taken together with (3.8) implies that the series $\sum_{k=0}^{\infty} b_k$ converges, so we can choose

$$n_0 := \inf\{k \ge 0 : \sum_{i=k}^{\infty} b_i < 1/2\}.$$

If $\lim_{n\to\infty} M_n = \infty$ then

$$1 \le \liminf_{n\to\infty} \frac{\sum_{k=0}^{n} b_k M_k}{M_n} = \liminf_{n\to\infty} \frac{\sum_{k=n_0}^{n} b_k M_k}{M_n} \le \sum_{k=n_0}^{\infty} b_k \le 1/2,$$

which is an obvious contradiction. Therefore (a_n) is bounded. From this

$$q_j \le M_j/(j+1) \le c/j.$$

Substituting this bound into (3.9) and applying Lemma 41 leads to the conclusion that $\mathbb{E}R_n$ remains bounded, as $n \to \infty$, by the virtue of (3.8). $\square$

Recall that the convergence of T_s in distribution always requires a scaling constant going to ∞ as $s \to \infty$. Under conditions of Lemma 42 the sequence of laws of τ_{R_n} is tight. Now from Proposition 40 and the decomposition (3.6) the following main result of this subsection emerges.

Theorem 43. *Suppose* (3.7) *holds. For some constants* $a_n > 0$ *and* $b_n \in \mathbb{R}$*, if one of the variables* $(T_{\log n} - b_n)/a_n$ *and* $(\tau_n - b_n)/a_n$ *converges weakly, as* $n \to \infty$*, to a nondegenerate proper distribution then the other variable converges weakly to the same distribution.*

The value of this result lies in the fact that the limit laws for T_s and the conditions of convergence are immediately translated into the convergence of τ_n. Normalizing and centering constants are known explicitly, see [103, Proposition 27] or [22]. It follows that only stable laws and the Mittag-Leffler laws can appear as the limit distributions of τ_n.

If measure θ is finite the condition (3.7) obviously holds. In this case S is a compound Poisson process. Theorem 43 has been proved in [55] under the assumptions that θ is not supported by a geometric sequence $(1 - x^k)_{k\in\mathbb{N}}$ (meaning that the law of S_1 is nonlattice) and that

$$\nu := \int_0^1 |\log x|\, \theta(\mathrm{d}x) < \infty. \tag{3.10}$$

Theorem 43 shows that the result of [55] is still true without requiring (3.10).

Assumption (3.7) is not very restrictive since $\Phi(k) = o(k)$, $k \to \infty$, always holds. Concretely, suppose the right tail of θ has the property of regular variation at 0, that is

$$\vec{\theta}(x) \sim x^{-\gamma}\ell_1(1/x), \quad x \downarrow 0, \tag{3.11}$$

for some function ℓ_1 of slow variation at ∞, and $\gamma \in [0, 1]$. Then condition (3.7) is satisfied for $\gamma \in [0, 1)$. In the edge case $\gamma = 1$ the behaviour of ℓ_1 is important, for instance (3.7) holds for $\ell_1(y) = (\log y)^{-\delta}$ if $\delta > 2$ and does not hold if $\delta \in (1, 2]$.

We use condition (3.7) to bound R_n, although we perceive that (3.7) can be omitted and the equivalence in Theorem 43 holds in full generality for the coalescents with dust component. Note that (3.7) is the local property of $\vec{\theta}$ near 0. More substantially, the limit law is affected by the decay at ∞ of the right tail of the distribution of S_1, for which the behaviour of $\vec{\theta}$ near 1 is

responsible. We illustrate this by two examples.

Example: normal limits. Assume in addition to (3.7) that

$$\mathtt{s}^2 := \mathrm{Var}(S_1) = \int_0^1 |\log(1-x)|^2 \theta(\mathrm{d}x) < \infty.$$

Then, as $n \to \infty$,

$$\frac{\tau_n - \mathtt{m}^{-1}\log n}{(\mathtt{m}^{-3}\mathtt{s}^2 \log n)^{1/2}} \stackrel{d}{\to} \mathcal{N}(0,1), \tag{3.12}$$

where $\mathtt{m} := \mathbb{E}S_1 = \int_0^1 |\log(1-x)|\theta(\mathrm{d}x)$. In view of [103, Proposition 27] and the remark in the paragraph preceding Proposition 40 the latter asymptotic result holds with τ_n replaced by $T_{\log n}$. By Theorem 43, (3.12) holds. The same argument also applies to relation (3.16).

This setting applies to beta (a,b) coalescents, $a>1$, $b>0$, which are the lambda-coalescents driven by

$$\theta(\mathrm{d}x) = (1/\mathrm{B}(a,b))x^{a-3}(1-x)^{b-1}1_{(0,1)}(x)\mathrm{d}x,$$

where $B(u,v) := \int_0^1 z^{u-1}(1-z)^{v-1}\mathrm{d}z$, $u,v>0$, denotes the Euler beta function.

The case $a>2$ was settled in [55]. We focus on the previously open case $1<a\le 2$.

For $a=2$ we compute the constants as

$$\mathtt{m} = b(b+1)\zeta(2,b), \quad \mathtt{s}^2 = 2b(b+1)\zeta(3,b),$$

where $\zeta(u,v) := \sum_{i=0}^{\infty}(i+v)^{-u}$, $u>1$, $v>0$, is the Hurwitz zeta function.

For $a\in(1,2)$ we have

$$\mathtt{m} = \frac{a+b-1}{(a-1)(2-a)}\Big(1-(a+b-2)\big(\Psi(a+b-1)-\Psi(b)\big)\Big), \tag{3.13}$$

$$\begin{aligned}\mathtt{s}^2 &= \frac{a+b-1}{(a-1)(2-a)} \times \Big(2\big(\Psi(a+b-1)-\Psi(b)\big) \\ &- (a+b-2)\big((\Psi(a+b-1)-\Psi(b))^2 + \Psi'(b) - \Psi'(a+b-1)\big)\Big)\end{aligned} \tag{3.14}$$

where Ψ is the logarithmic derivative of the gamma function. Finally, condition (3.7) holds since (3.11) is satisfied with $\gamma = 2 - a \in [0, 1)$ and constant function ℓ_1. Therefore, convergence (3.12) holds with the computed $\mathtt{m}$ and $\mathtt{s}^2$.

Example: stable limits. Assume (3.7) and

$$\vec{\theta}(1 - e^{-y}) \sim y^{-\beta}\ell(y),\ y \to \infty, \tag{3.15}$$

for some function ℓ slowly varying at ∞ and $\beta \in (1, 2)$. Then

$$\frac{\tau_n - \mathtt{m}^{-1}\log n}{\mathtt{m}^{-(\beta+1)/\beta} c_{\lfloor \log n \rfloor}} \xrightarrow{d} \mathcal{S}(\beta), \quad n \to \infty, \tag{3.16}$$

where c_n is any sequence satisfying $\lim_{n\to\infty} n\ell(c_n)/c_n^\beta = 1$, and $\mathcal{S}(\beta)$ is a random variable with the β-stable distribution with characteristic function

$$z \mapsto \exp\{-|z|^\beta \Gamma(1-\beta)(\cos(\pi\beta/2) + i\sin(\pi\beta/2)\operatorname{sgn}(z))\}, \quad z \in \mathbb{R}.$$

To illustrate, consider

$$\theta(\mathrm{d}x) = \frac{x^{a-2}\mathrm{d}x}{(1-x)|\log(1-x)|^d},$$

where $d \in (2, 3)$ and $a \in (d, d+1)$. Then (3.2) is satisfied, and condition (3.11) holds with $\gamma = d + 1 - a \in (0, 1)$ which implies (3.7). Condition (3.15) is fulfilled with $\beta = d - 1 \in (1, 2)$. Therefore, the absorption time τ_n of such coalescent has limiting law (3.16).

3.1.5 The number of collisions.

As an approximation to the number of collisions X_n we shall consider X_n^*, the number of jumps of N_n^* of size at least two. We will not be able to derive a complete result comparable with Theorems 45 or 43 because the universal criterion for convergence of X_n^* is not available. The cases when we know the behaviour of X_n^* (Corollary 24 and Proposition 38 of the present work and [11, 60, 61]) are all covered by the assumption that θ satisfies (3.11). We shall also proceed in this direction but exclude the case $\gamma = 1$ when $K_{n,1}$ is the term of dominating growth in the sum $K_n = \sum_{r=1}^n K_{n,r}$. By Karamata's Tauberian theorem [24] condition

(3.11) with $\gamma < 1$ is equivalent to the analogous asymptotics of the Laplace exponent

$$\Phi(z) \sim \Gamma(1-\gamma) z^{\gamma} \ell_1(z), \quad z \to \infty.$$

The case of finite θ appears when $\gamma = 0$ and Φ is an increasing bounded function.

The sequence (X_n) is nondecreasing and satisfies a distributional recurrence

$$X_1 = 0, \quad X_n \stackrel{d}{=} \tilde{X}_{n-J_n+1} + 1_{\{J_n \geq 2\}}, \quad n \geq 2, \tag{3.17}$$

where in the right-hand side $\tilde{X}_i$ is assumed independent of J_n and distributed like X_i, for each $i \in \mathbb{N}$, and J_n is distributed like the first decrement of N_n^*, that is $\mathbb{P}\{J_n = k\} = p_{n,n-k}$ for $1 \leq k \leq n$. Similarly, the number X_n^* of collisions which involve at least two primary clusters satisfies

$$X_1^* = 0, \quad X_n^* \stackrel{d}{=} \tilde{X}^*_{n-J_n} + 1_{\{J_n \geq 2\}}, \quad n \geq 2, \tag{3.18}$$

with the convention $X_0^* = 0$. We may decompose X_n as

$$X_n = X_n^* + D_n \stackrel{(3.4)}{=} K_n - K_{n,1} + D_n \ \text{ a.s.} \tag{3.19}$$

where D_n is the number of collisions which take at most one primary cluster. Thus a collision contributes to D_n if either exactly one primary cluster merges with at least one secondary cluster, or at least two secondary and no primary clusters are merged.

Lemma 44. *We have*

$$\mathbb{E}D_n \leq c \sum_{k=1}^{n} \left(\Phi(k)/k\right)^2, \quad n \in \mathbb{N}. \tag{3.20}$$

In particular, if either of two equivalent conditions

$$\int_0^1 x^{-2} \left(\int_0^x \vec{\theta}(y) \mathrm{d}y \right)^2 \mathrm{d}x < \infty, \tag{3.21}$$

$$\sum_{k=1}^{\infty} \left(\Phi(k)/k\right)^2 < \infty \tag{3.22}$$

holds then the sequence of distributions of the D_n's is tight.

Proof. The equivalence of (3.21) and (3.22) follows from [17, Proposition 1.4].

Choose some primary cluster b, to be definite let it be the cluster labelled 1, and suppose $\tilde{X}_{n-1}$ is realised as the number of collisions among $n-1$ primary clusters $[n] \setminus \{b\}$ and their followers. Then $X_n = \tilde{X}_{n-1} + Z_n$, where Z_n is the indicator of the event that the first collision of b involves exactly one other cluster a. At the time of the merge of b with a the Markov chain N_n^* decrements by two or one, depending on whether a is primary or secondary. Let Y_n be the indicator of the event that the first involvement of b is either turning b into secondary cluster, or a collision taking at most one other primary cluster and arbitrary number of secondary clusters. Clearly, $Y_n \geq Z_n$, therefore from (3.17)

$$X_n \overset{d}{\leq} \tilde{X}_{n-J_n} + Y_{n-J_n+1} + 1_{\{J_n \geq 2\}}, \tag{3.23}$$

where $\overset{d}{\leq}$ stands for 'stochastically smaller'. Passing to expectations in (3.23), (3.18) and (3.19) we see that, for $d_n := \mathbb{E}D_n$, $y_n := \mathbb{E}Y_n$,

$$d_1 = 0, \quad d_n \leq \sum_{k=1}^{n-1} p_{n,k}(d_k + y_{k+1}), \quad n = 2, 3, \ldots,$$

and iterating this inequality yields

$$d_1 = 0, \quad d_n \leq \sum_{j=1}^{n-1} g_{n,j} y_{j+1}, \quad n = 2, 3, \ldots$$

By exchangeability, we have $y_n = (\mathbb{E}K_{n,1} + 2\mathbb{E}K_{n,2})/n$. Since

$$\mathbb{E}K_{n,1} = \sum_{k=1}^{n} g_{n,k} p_{k,k-1} = \sum_{k=1}^{n} g_{n,k} \frac{k(\Phi(k) - \Phi(k-1))}{\Phi(k)},$$

using Lemma 41 with $r_k = k(\Phi(k) - \Phi(k-1))/\Phi(k)$ yields

$$\mathbb{E}K_{n,1} \leq c\,\Phi(n), \quad n \in \mathbb{N}. \tag{3.24}$$

Using this, the inequality from Lemma 37 and the monotonicity of Φ,

$$\mathbb{E}K_{n,2} \overset{(2.45)}{\leq} c_1\,\mathbb{E}K_{\lceil n/2 \rceil,1} \overset{(3.24)}{\leq} c_2\,\Phi(\lceil n/2 \rceil) \leq c_2\,\Phi(n).$$

Thus

$$d_n \leq c \sum_{k=1}^{n} g_{n,k}\Phi(k)/k,$$

and using Lemma 41 with $r_k = c\Phi(k)/k$ results in (3.20). □

The compound Poisson case. Assume that θ is a finite measure on $(0,1)$, not supported by a geometric sequence of the form $(1-x^k)_{k\in\mathbb{N}}$, for some $x \in (0,1)$. Since a linear time change of the coalescent does not affect the distribution of X_n we will not lose generality by assuming that θ is a probability measure on $(0,1)$. Let $(W_k)_{k\in\mathbb{N}}$ be independent copies of a random variable W such that the law of $1-W$ is θ. The subordinator S is then a unit rate compound Poisson process with the generic jump $|\log W|$ having some nonlattice law.

The next theorem improves upon a result from [55] by removing condition (3.10).

Theorem 45. *For constants $a_n > 0$ such that $\lim_{n\to\infty} a_n = \infty$, and $b_n \in \mathbb{R}$, whenever any of the variables*

$$\frac{K_n - b_n}{a_n}, \quad \frac{X_n^* - b_n}{a_n} \quad \text{or} \quad \frac{X_n - b_n}{a_n}$$

converges weakly, as $n \to \infty$, to a nondegenerate proper distribution then all three variables converge weakly to this distribution.

Proof. Recall representation (3.19). Since ν is a probability measure we have $\Phi(k) < 1$, hence condition (3.22) is satisfied, and the sequence of laws of the D_n's is tight by Lemma 44. By Proposition 38, the sequence of laws of the $K_{n,1}$'s is tight as well. By the assumption $a_n \to \infty$ the result follows. □

From Corollary 24 it is known that, depending on the behaviour of $\vec{\theta}(x)$ near $x = 1$ there are five different modes of the weak convergence of, suitably normalized and centered, K_n (hence X_n). Below we provide an example which partially coincides with Example on p. 45 to demonstrate a substantial role of the parameter $\nu = \int_0^1 |\log x|\theta(\mathrm{d}x)$.

Example. Suppose θ has the right tail of the form

$$\vec{\theta}(x) = \frac{|\log x|^\rho}{1 + |\log x|^\rho}, \qquad x \in (0, 1]$$

with $\rho > 0$. In the case $\rho \in (0, 1/2)$ we have $\nu = \infty$, and

$$\frac{X_n - \mathtt{m}^{-1} \log n + (\mathtt{m}(1-\rho))^{-1} \log^{1-\rho} n}{c \log^{1/2} n} \stackrel{d}{\to} \mathcal{N}(0, 1), \quad n \to \infty,$$

where $\mathtt{m} = \int_0^1 |\log(1-x)|\, \theta(\mathrm{d}x)$.

In the other case, when $\rho > 1/2$ (then $\nu < \infty$ for $\rho > 1$), the centering simplifies, so that

$$\frac{X_n - \mathtt{m}^{-1} \log n}{c \log^{1/2} n} \stackrel{d}{\to} \mathcal{N}(0, 1), \quad n \to \infty.$$

Evolution of secondary particles. In the compound Poisson case the number V_t of secondary clusters of $\Pi_\infty(t)$ is finite, for each $t \geq 0$. The process $V = (V_t)_{t \geq 0}$ starts with $V_0 = 0$ and is a Markov chain with the transition rate $\varphi_{m,k} = \binom{m}{k} \lambda_{m,k}$ for jumping from m to $m - k + 1$, $0 \leq k \leq m$, $k \neq 1$. The rate for $k = 0$ is given by the same formula (3.1), and $\varphi_{m,0} < \infty$ because θ is finite. The $k = 0$ transition, resulting in increase of the number of secondary clusters by one, occurs when some (in fact, infinitely many) primary clusters merge without engagement of secondary clusters. The stationarity of V is a consequence of the existence of the dust component with infinitely many clusters.

It can be shown that the Markov chain V is positively recurrent and has a unique stationary distribution (z_m) found from the balance equation

$$z_m = \sum_{k=0}^{\infty} z_{m+k-1} \varphi_{m+k-1,k} \tag{3.25}$$

supplemented by the conditions $z_0 = 0$ and $\sum_{m=1}^{\infty} z_m = 1$.

Suppose for example that $\theta(\mathrm{d}x) = \mathrm{d}x$ is the Lebesgue measure on $[0, 1]$. In this case $\varphi_{m,k} = (m+1)^{-1}$. Equation (3.25) becomes $z_m =$

$\sum_{j=m-1}^{\infty} z_j/(j+1)$. Differencing yields $z_m - z_{m+1} = z_{m-1}/m$, which is readily solved as

$$z_m = \frac{e^{-1}}{(m-1)!}, \qquad m \in \mathbb{N},$$

so in this case the stationary distribution is shifted Poisson.

In contrast, the number of secondary clusters in the finite coalescent Π_n is not a Markov process, because the transition rates depend on the number of remaining primary particles.

The case of slow variation. Suppose (3.11) holds with $\gamma = 0$ and slowly varying $\ell_1(z) \to \infty, \quad z \to \infty$. The Laplace exponent satisfies then $\Phi(z) \sim \ell_1(z)$. Suppose also that the subordinator has finite moments

$$\mathtt{m} = \mathbb{E}S_1 = \int_0^1 |\log(1-x)|\theta(\mathrm{d}x), \quad \mathtt{s}^2 = \operatorname{Var} S_1 = \int_0^1 |\log(1-x)|^2\theta(\mathrm{d}x).$$

Choose the centering/scaling constants as

$$b_n = \frac{1}{\mathtt{m}}\int_0^n \frac{\Phi(z)}{z}\mathrm{d}z\,, \quad a_n = \sqrt{\frac{\mathtt{s}^2}{\mathtt{m}^3}\int_0^n \frac{\Phi^2(z)}{z}\mathrm{d}z}.$$

In [11] it was shown that, as $n \to \infty$,

$$\mathbb{E}K_n \sim b_n, \quad \sqrt{\operatorname{Var}K_n} \sim a_n,$$

and that the normal limit $(K_n - b_n)/a_n \xrightarrow{d} \mathcal{N}(0,1)$ holds for various classes of functions ℓ_1. In particular, these include functions of slow variation at infinity with asymptotics as diverse as

$$\ell_1(z) = \log(\log(\ldots(\log(z))\ldots)), \quad \ell_1(z) = \log^{\beta} z\,, \quad \ell_1(z) = \exp(\log^{\beta} z),$$

where $\beta > 0$.

The series (3.22) converges for arbitrary ℓ_1, hence, by Lemma 44, $\mathbb{E}D_n = O(1)$. On the other hand, from (3.24) and by the properties of slowly varying functions [24]

$$\mathbb{E}K_{n,1} = O(\Phi(n)) = o(a_n).$$

It is immediate now that $(K_n - b_n)/a_n \xrightarrow{d} \mathcal{N}(0,1)$ implies both $(X_n^* - b_n)/a_n \xrightarrow{d} \mathcal{N}(0,1)$ and $(X_n - b_n)/a_n \xrightarrow{d} \mathcal{N}(0,1)$.

Example: gamma subordinators. Consider the classical gamma subordinator with Laplace exponent $\Phi(z) = \alpha \log(1 + z/\beta)$, where $\alpha, \beta > 0$. The corresponding θ driving the coalescent has density

$$\theta(\mathrm{d}x) = \frac{\alpha(1-x)^{\beta-1}}{|\log(1-x)|}\,\mathrm{d}x.$$

The central limit theorem for K_n was proved by different methods in [61] and [11]. From this we conclude that the number of collisions also satisfies $(X_n - b_n)/a_n \xrightarrow{d} \mathcal{N}(0,1)$, where the constants can be chosen as

$$a_n = \sqrt{\frac{\beta \log^3 n}{3}}\,, \qquad b_n = \frac{\beta \log^2 n}{2}.$$

Example: beta$(2,b)$-coalescents. For this family $\theta(\mathrm{d}x) = x^{-1}(1-x)^{b-1}\mathrm{d}x$. The convergence of X_n to the standard normal distribution holds with scaling/centering constants

$$a_n = \sqrt{\frac{\mathtt{s}^2}{3\mathtt{m}^3}\log^3 n}, \quad b_n = \frac{\log^2 n}{2\mathtt{m}},$$

where $\mathtt{m} = \zeta(2,b),\ \mathtt{s}^2 = 2\zeta(3,b)$.

In Section 3.2 this result will be proved by a different technique based on the asymptotic analysis of moments.

Regular variation with index $0 < \gamma < 1$. A key distribution in this case is the law of the random variable

$$I = \int_0^\infty \exp(-\gamma S_t)\mathrm{d}t,$$

known as the exponential functional of the subordinator γS. The distribution of I is uniquely determined by the moments

$$\mathbb{E}I^k = \frac{k!}{\prod_{i=1}^k \Phi(\gamma i)}.$$

From [60, Theorem 4.1 and Corollary 5.2] $X_n^*/a_n \xrightarrow{d} I$, where $a_n = \Gamma(2-\gamma)n^\gamma \ell_1(n)$, and no centering is required. In fact, K_n/a_n and $K_{n,r}/a_n$ $(r \geq 1)$ converge almost surely and in the mean.

To justify the convergence of X_n using (3.19) we need to estimate $\mathbb{E}D_n$. For $0 < \gamma < 1/2$ we have $\mathbb{E}D_n = O(1)$ since $\Phi(z) \sim c\,\ell_1(z)z^\gamma$, hence the series (3.22) converges. For $1/2 < \gamma < 1$ we have

$$\sum_{k=1}^{n} (\Phi(k)/k)^2 \sim c\, n^{2\gamma-1}\ell_1^2(n),$$

and for $\gamma = 1/2$ the latter sum, as a function of n, has the property of slow variation at infinity (see [24, Proposition 1.5.8]). Thus in any case $D_n/a_n \to 0$ in probability. It follows that $X_n/a_n \xrightarrow{d} I$.

Example: beta (a, b)-coalescents with $1 < a < 2$. In this case

$$\frac{X_n}{n^{2-a}} \xrightarrow{d} \frac{\Gamma(2-a)}{2-a}\int_0^\infty \exp\{-(2-a)S_t\}\,\mathrm{d}t, \quad n \to \infty,$$

where the Laplace exponent of S is given by

$$\Phi(z) = \int_0^1 (1-(1-x)^z)x^{a-3}(1-x)^{b-1}\mathrm{d}x.$$

This result was obtained in [68] by another method, and with a change of variables the equivalence with [78, Theorem 7.1] in the case $b = 1$ can be established.

The subfamily of beta-coalescents with parameters $b = 2 - a$ was intensively studied. In the literature sometimes $\alpha := 2 - a$ is taken as parameter, so that θ in this notation becomes

$$\theta(\mathrm{d}x) = x^{-\alpha-1}(1-x)^{\alpha-1}.$$

In this case N_n^* decrements like a random walk conditioned to hit 0 and, moreover, there is an explicit formula (see [58, p. 471])

$$g_{n,k} = \frac{(\alpha)_k(\alpha)_{n-k}}{(\alpha)_n}\binom{n}{k},$$

where $(\alpha)_k$ denotes the rising factorial. The variable K_n is then the number of blocks in Pitman's (α, α)-partition (or in the regenerative composition induced by excursions of a Bessel bridge [58]).

3.2 Number of collisions in beta $(2, b)$-coalescents

Most intensively studied lambda-coalescents are those driven by a beta measure

$$\theta(\mathrm{d}x) = cx^{a-3}(1-x)^{b-1}1_{(0,1)}(x)\mathrm{d}x, \quad a, b, c > 0, \tag{3.26}$$

and called the *beta-coalescents*. The weak convergence results surveyed in Section 3.4 (see Table 1 there) indicate that the two parameter values $a = 1$ and $a = 2$ play a kind of threshold-role when studying the limiting behavior of the number of collisions X_n of beta (a, b)-coalescents. In this section we focus on the asymptotics of X_n for beta $(2, b)$-coalescents, $b > 0$ and thus assume throughout this section that

$$\theta(\mathrm{d}x) = x^{-1}(1-x)^{b-1}1_{(0,1)}(x)\mathrm{d}x, \quad b > 0.$$

Although the result of forthcoming Theorem 49 has already been obtained on p. 92 we think that its alternative derivation is worth having.

3.2.1 Preliminaries. Throughout the section we often use the notation introduced in the previous section without special mention.

From the structure of the coalescent process it follows that $(X_n)_{n\in\mathbb{N}}$ satisfies the distributional recurrence (6) with I_n having distribution

$$\begin{aligned} \mathbb{P}\{I_n = k\} &= \varphi_{n,n-k+1} / \sum_{k=2}^{n} \varphi_{n,k} \qquad (3.27) \\ &= \frac{\Gamma(k+b-1)\Gamma(n+1)}{(n-k+1)\Gamma(k)\Gamma(n+b)H(n,b)}, \quad k \in \mathbb{N}, k \le n-1, \end{aligned}$$

where rates $\varphi_{n,k}$ were defined in (3.3),

$$H(n,b) := \frac{b}{b+n-1} + \Psi(b+n-1) - \Psi(b) - 1, \qquad n \in \mathbb{N},\ b > 0,$$

and $\Psi(\cdot)$ denotes the logarithmic derivative of the gamma function. The random variable I_n is the (random) state of the process N_n after its first

jump. Note that $\Psi(b+n-1) = \log n + O(1/n)$, $n \to \infty$ (see [1, formula (6.3.18)]), and, therefore,

$$H(n,b) \;=\; \log n - \Psi(b) - 1 + O\Big(\frac{1}{n}\Big), \quad n \to \infty. \tag{3.28}$$

Recall that

$$\sum_{k=2}^{n} \varphi_{n,k} = \Phi(n) = \int_0^1 (1-(1-x)^n)\theta(\mathrm{d}x).$$

In the proofs we will need the asymptotics of the total rates

$$\Phi_n \;=\; \frac{H(n,b)}{B(2,b)} \;\sim\; \frac{\log n}{B(2,b)}, \quad n \to \infty, \tag{3.29}$$

and explicit expressions of the moments

$$\begin{aligned} m_r^{(b)} \;&:=\; \int_0^1 |\log(1-x)|^r \theta(\mathrm{d}x) = \int_0^1 |\log(1-x)|^r \frac{(1-x)^{b-1}}{x}\,\mathrm{d}x \\ &=\; \Gamma(r+1)\,\zeta(r+1,b), \quad r>0, \end{aligned} \tag{3.30}$$

where $\zeta(\cdot,\cdot)$ is the Hurwitz zeta function. The last formula follows from a Hurwitz identity (see, for example, [1, formula (23.2.7)]). Set

$$\mathtt{m} := m_1^{(b)} = \zeta(2,b) \;\text{ and }\; \mathtt{s}^2 := m_2^{(b)} = 2\zeta(3,b).$$

3.2.2 Main results.

Our first result presents the asymptotic expansions of the moments of X_n. For convenience, we use the notation $\log^k n := (\log(n))^k$, $k, n \in \mathbb{N}$.

Theorem 46. *(Expansion of moments)*
As $n \to \infty$, for $k \in \mathbb{N}$,

$$\mathbb{E}X_n^k \;=\; \frac{1}{(2\mathtt{m})^k}\log^{2k} n + \frac{2k((2k+1)\mathtt{s}^2 + 6c\mathtt{m})}{3(2\mathtt{m})^{k+1}}\log^{2k-1} n + O(\log^{2k-2} n),$$

where $c := -\Psi(b) - 1$. In particular, the variance $\operatorname{Var} X_n$ has the asymptotic expansion

$$\operatorname{Var} X_n \;=\; \frac{\mathtt{s}^2}{3\mathtt{m}^3}\log^3 n + O(\log^2 n) \;=\; \frac{2\zeta(3,b)}{3\zeta^3(2,b)}\log^3 n + O(\log^2 n).$$

Remark 47. For $t \geq 0$, let $(f_i(t))_{i\in\mathbb{N}}$ be a sequence (in some order) of the asymptotic frequencies of the random exchangeable partition $\Pi_\infty(t)$ (corresponding to the beta $(2, b)$ - coalescent). By [113, Proposition 26], $(\widehat{S}_t := -\log(1 - \sum_{i=1}^\infty f_i(t)))_{t\geq 0}$ is a version of S a pure-jump subordinator with the Lévy exponent Φ. We will come back to this remark later in the proofs.

Corollary 48. (Strong law of large numbers)
As $n \to \infty$, $X_n/\log^2 n \to 1/(2\mathtt{m})$ almost surely.

Finally, given next is a central limit theorem for X_n.

Theorem 49. *(Central limit theorem)*
As $n \to \infty$, the sequence

$$\frac{X_n - \frac{1}{2\mathtt{m}}\log^2 n}{\sqrt{\frac{\mathtt{s}^2}{3\mathtt{m}^3}\log^3 n}}$$

weakly converges to the standard normal law.

Remark 50. The proof of Theorem 49 presented in Section 3.2.4 draws heavily from the coalescent theory and results on random exchangeable partitions. We leave open the question whether it is possible to deduce the asymptotic normality of X_n from recurrence (6) with I_n satisfying (3.27), i.e., without using pathwise results available in the coalescent setting.

Recently there appeared a preprint [31] whose authors readdressed the issue of finding the weak asymptotics of X_n in beta $(2, b)$-coalescent using the recurrence alone. Their proof is incomplete since it is based on an unproved conjecture which, however, seems to be correct and, in the first approximation, follows from our Lemma 54.

3.2.3 Proof of Theorem 46 and Corollary 48.

Proof of Theorem 46. For $k \in \mathbb{N}$, set $a_n^{(k)} := \mathbb{E}X_n^k$. By induction on k we will prove the asymptotic expansion

$$a_n^{(k)} \;=\; \alpha^k \log^{2k} n + r_k \log^{2k-1} n + O(\log^{2k-2} n), \quad k \in \mathbb{N}, \tag{3.31}$$

where $\alpha := (2\mathtt{m})^{-1}$ and

$$r_k \;:=\; \frac{2}{3}k\alpha^{k+1}((2k+1)\mathtt{s}^2 + 6c\mathtt{m}). \tag{3.32}$$

For $k = 1$, write a_n instead of $a_n^{(1)}$ for simplicity. In view of (6), we have

$$a_1 \;=\; 0, \quad a_n \;=\; 1 + \sum_{i=1}^{n-1} a_i \mathbb{P}\{I_n = i\}, \quad n \geq 2. \tag{3.33}$$

Put $b_n := a_n - \alpha \log^2 n$, $n \in \mathbb{N}$. From (3.33) it follows that $b_1 = 0$ and

$$\begin{aligned} b_n &= 1 + \alpha \sum_{i=1}^{n-1} (\log^2(n-i) - \log^2 n)\mathbb{P}\{I_n = n-i\} + \sum_{i=1}^{n-1} b_i \mathbb{P}\{I_n = i\} \\ &=: c_n + \sum_{i=1}^{n-1} b_i \mathbb{P}\{I_n = i\}, \quad n \geq 2. \end{aligned} \tag{3.34}$$

Using Lemma 53 (with $k = 1$ and $k = 2$), we get

$$\begin{aligned} c_n &= 1 + \alpha \sum_{i=1}^{n-1} \left(\log^2(1 - i/n) + 2\log n \log(1 - i/n)\right) \mathbb{P}\{I_n = n - i\} \\ &= 1 + \frac{\alpha}{H(n,b)} \Big(\mathtt{s}^2 + O\Big(\frac{\log^2 n}{n^{b \wedge 1}}\Big) + 2\log n \Big(-\mathtt{m} + O\Big(\frac{\log n}{n^{b\wedge 1}}\Big)\Big)\Big) \\ &= 1 - \frac{\log n}{H(n,b)} + \frac{\mathtt{s}^2}{2\mathtt{m}H(n,b)} + O\Big(\frac{\log n}{n^{b\wedge 1}}\Big), \end{aligned}$$

and, by (3.28),

$$\begin{aligned} c_n &= 1 - \frac{H(n,b) + \Psi(b) + 1 + O(1/n)}{H(n,b)} + \frac{\mathtt{s}^2}{2\mathtt{m}H(n,b)} + O\Big(\frac{\log n}{n^{b\wedge 1}}\Big) \\ &= \frac{\frac{\mathtt{s}^2}{2\mathtt{m}} - \Psi(b) - 1}{H(n,b)} + O\Big(\frac{\log n}{n^{b\wedge 1}}\Big) \;=:\; \frac{C_1}{H(n,b)} + O\Big(\frac{\log n}{n^{b\wedge 1}}\Big). \end{aligned}$$

Substituting this relation into (3.34) yields

$$b_n \;=\; \frac{C_1}{H(n,b)} + O\Big(\frac{\log n}{n^{b\wedge 1}}\Big) + \sum_{i=1}^{n-1} b_i \mathbb{P}\{I_n = i\}.$$

Set $d_n := b_n - (C_1/\mathtt{m}) \log n$, $n \in \mathbb{N}$. Then, $d_1 = 0$ and

$$\begin{aligned} d_n &= \frac{C_1}{H(n,b)} + \frac{C_1}{\mathtt{m}} \sum_{i=1}^{n-1} \log(i/n)\mathbb{P}\{I_n = i\} \\ &\qquad + O\Big(\frac{\log n}{n^{b\wedge 1}}\Big) + \sum_{i=1}^{n-1} d_i \mathbb{P}\{I_n = i\}, \quad n \geq 2. \end{aligned}$$

Another appeal to Lemma 53 leads to

$$\begin{aligned} d_n &= \frac{C_1}{H(n,b)} + \frac{C_1}{\mathtt{m}H(n,b)}\Big(- m_1 + O\Big(\frac{\log n}{n^{b\wedge 1}}\Big)\Big) \\ &\qquad + O\Big(\frac{\log n}{n^{b\wedge 1}}\Big) + \sum_{i=1}^{n-1} d_i \mathbb{P}\{I_n = i\} \\ &= O\Big(\frac{\log n}{n^{b\wedge 1}}\Big) + \sum_{i=1}^{n-1} d_i \mathbb{P}\{I_n = i\}. \end{aligned}$$

By Lemma 54, $d_n = O(1)$. Therefore, $a_n = \alpha \log^2 n + r_1 \log n + O(1)$, and we have already proved (3.31) for $k = 1$.

The induction step from $\{1, \dots, k\}$ to $k+1$ works as follows. Using (6) and dropping terms of lower orders (which is possible due to the assumption of induction) we get $a_1^{(k+1)} = 0$ and

$$\begin{aligned} a_n^{(k+1)} &= (k+1)\alpha^k \log^{2k} n + (k+1) r_k \log^{2k-1} n + \\ &\quad + O(\log^{2k-2} n) + \sum_{j=1}^{n-1} a_j^{(k+1)} \mathbb{P}\{I_n = j\}, \quad n \geq 2. \end{aligned}$$

Put $b_n^{(k+1)} := a_n^{(k+1)} - \alpha^{k+1} \log^{2k+2} n$, $n \in \mathbb{N}$. Then, we have $b_1^{(k+1)} = 0$ and

$$b_n^{(k+1)} = c_n^{(k+1)} + \sum_{j=1}^{n-1} b_j^{(k+1)} \mathbb{P}\{I_n = j\}, \quad n \geq 2, \tag{3.35}$$

where

$$\begin{aligned} c_n^{(k+1)} &:= \alpha^{k+1} \sum_{j=1}^{n-1} \big(\log^{2k+2}(n-j) - \log^{2k+2} n\big)\, \mathbb{P}\{I_n = n-j\} \\ &\quad + (k+1)\alpha^k \log^{2k} n + (k+1) r_k \log^{2k-1} n + O(\log^{2k-2} n). \end{aligned}$$

Binomial expansion of $\log^{2k+2}(n-j) = (\log(1 - j/n) + \log n)^{2k+2}$ leads to

$$\begin{aligned} c_n^{(k+1)} &= (k+1)\alpha^k \log^{2k} n + (k+1) r_k \log^{2k-1} n + O(\log^{2k-2} n) \\ &\quad + \alpha^{k+1} \sum_{j=1}^{n-1} \mathbb{P}\{I_n = j\} \sum_{i=0}^{2k+1} \binom{2k+2}{i} \log^{2k+2-i}(j/n) \log^i n \\ &= (k+1)\alpha^k \log^{2k} n + (k+1) r_k \log^{2k-1} n + O(\log^{2k-2} n) \\ &\quad + \alpha^{k+1} \sum_{i=0}^{2k+1} \binom{2k+2}{i} \log^i n \sum_{j=1}^{n-1} \mathbb{P}\{I_n = j\} \log^{2k+2-i}(j/n). \end{aligned}$$

Using Lemma 53 gives

$$
\begin{aligned}
& c_n^{(k+1)} \\
&= (k+1)\alpha^k \log^{2k} n + (k+1) r_k \log^{2k-1} n + O(\log^{2k-2} n) \\
&\quad + \frac{\alpha^{k+1}}{H(n,b)} \sum_{i=0}^{2k+1} \binom{2k+2}{i} \log^i n \Big((-1)^i m_{2k+2-i} + O\Big(\frac{\log^{2k+2-i} n}{n^{b \wedge 1}} \Big) \Big) \\
&= (k+1)\alpha^k \log^{2k} n + (k+1) r_k \log^{2k-1} n + O(\log^{2k-2} n) \\
&\quad + \frac{\alpha^{k+1}}{H(n,b)} \left(-\mathtt{m} \binom{2k+2}{2k+1} \log^{2k+1} n + \mathtt{s}^2 \binom{2k+2}{2k} \log^{2k} n \right) \\
&= (k+1)\alpha^k \log^{2k} n \Big(1 - \frac{\log n}{H(n,b)} \Big) \\
&\quad + \Big((k+1) r_k + \alpha^{k+1} (2k+1)(k+1) \mathtt{s}^2 \frac{\log n}{H(n,b)} \Big) \log^{2k-1} n \\
&+ O(\log^{2k-2} n) \\
&= (k+1) \Big(r_k + (2k+1)\alpha^{k+1} \mathtt{s}^2 - (\Psi(b)+1)\alpha^k \Big) \log^{2k-1} n \\
&+ O(\log^{2k-2} n) \\
&=: c_k \log^{2k-1} n + O(\log^{2k-2} n).
\end{aligned}
$$

Plugging the last expression into (3.35) gives $b_1^{(k+1)} = 0$ and

$$
b_n^{(k+1)} = c_k \log^{2k-1} n + O(\log^{2k-2} n) + \sum_{j=1}^{n-1} b_j^{(k+1)} \mathbb{P}\{I_n = j\}, \ n \geq 2.
$$

Set $e_n^{(k+1)} := b_n^{(k+1)} - C_k \log^{2k+1} n$, $n \in \mathbb{N}$, where $C_k := c_k/((2k+1)\mathtt{m})$. The

so defined sequence is given by the recurrence

$$\begin{aligned} e_n^{(k+1)} &= c_k \log^{2k-1} n + O(\log^{2k-2} n) \\ &\quad + C_k \sum_{i=1}^{n-1} \Big(\log^{2k+1}(n-i) - \log^{2k+1} n \Big) \mathbb{P}\{I_n = n-i\} \\ &\quad + \sum_{j=1}^{n-1} e_j^{(k+1)} \mathbb{P}\{I_n = j\} \\ &= c_k \log^{2k-1} n + O(\log^{2k-2} n) \\ &\quad + C_k \sum_{i=1}^{n-1} \mathbb{P}\{I_n = i\} \sum_{j=0}^{2k} \binom{2k+1}{j} \log^j n \log^{2k+1-j}(i/n) \\ &\quad + \sum_{j=1}^{n-1} e_j^{(k+1)} \mathbb{P}\{I_n = j\}. \end{aligned}$$

Using again Lemma 53 yields

$$\begin{aligned} e_n^{(k+1)} &= c_k \log^{2k-1} n + O(\log^{2k-2} n) \\ &\quad + C_k \frac{\log^{2k} n}{H(n,b)} (2k+1) \Big(-m_1 + O\Big(\frac{\log n}{n^{b \wedge 1}} \Big) \Big) \\ &\quad + \sum_{j=1}^{n-1} e_{n-j}^{(k+1)} \mathbb{P}\{I_n = j\} \\ &= O(\log^{2k-2} n) + \sum_{j=1}^{n-1} e_{n-j}^{(k+1)} \mathbb{P}\{I_n = j\}, \end{aligned}$$

by the choice of C_k. For sufficiently large n, we can choose $M_k > 0$ such that the $O(\log^{2k-2} n)$ term is dominated by

$$M_k(k\alpha^{k-1} \log^{2k-2} n + k r_{k-1} \log^{2k-3} n + O(\log^{2k-4} n)).$$

Therefore, for large n, $e_n^{(k+1)} \leq M_k a_n^{(k)}$. By the assumption of induction, $a_n^{(k)} = O(\log^{2k} n)$. Therefore, $e_n^{(k+1)} = O(\log^{2k} n)$. Setting $r_{k+1} := C_k = c_k/((2k+1)\mathtt{m})$, we obtain

$$a_n^{(k+1)} = \alpha^{k+1} \log^{2k+2} n + r_{k+1} \log^{2k+1} n + O(\log^{2k} n).$$

The sequence $(r_k)_{k \in \mathbb{N}}$ satisfies the recurrence

$$r_{k+1} = \frac{k+1}{(2k+1)\mathtt{m}} (r_k + (2k+1)\alpha^{k+1}\mathtt{s}^2 - (\Psi(b)+1)\alpha^k),$$

with initial condition

$$r_1 \;=\; \frac{\frac{\mathtt{s}^2}{2\mathtt{m}} - \Psi(b) - 1}{m_1}.$$

The unique solution of this recurrence is given by (3.32). The proof of Theorem 46 is complete. □

Proof of Corollary 48. For $n \in \mathbb{N}$ and $\varepsilon > 0$, set $A_n(\varepsilon) := \{|X_n - \mathbb{E}X_n| \geq \varepsilon \mathbb{E}X_n\}$. By Chebyshev's inequality, $\mathbb{P}\{A_n(\varepsilon)\} \leq \operatorname{Var} X_n/(\varepsilon \mathbb{E}X_n)^2$. From Theorem 46 it follows that

$$\frac{\operatorname{Var} X_n}{(\mathbb{E}X_n)^2} \;=\; \frac{4\mathtt{s}^2}{3\mathtt{m}} \frac{1}{\log n} + O\Big(\frac{1}{\log^2 n}\Big).$$

Therefore, replacing n by $n_k := \lfloor \exp(k^2) \rfloor$, it follows that $\sum_{k=1}^{\infty} \mathbb{P}\{A_{n_k}(\varepsilon)\} < \infty$ and, hence, $X_{n_k}/\mathbb{E}X_{n_k} \to 1$ almost surely as $k \to \infty$ by the Borel-Cantelli lemma. Thus we have already verified the result along the subsequence $(n_k)_{k\in\mathbb{N}}$. For each integer $n \geq n_1$, there exists a unique index $k = k(n) \in \mathbb{N}$ such that $n_k \leq n < n_{k+1}$. It is clear that the sequence $(X_n)_{n\in\mathbb{N}}$ is almost surely non-decreasing. Therefore, the corollary follows from the standard sandwich argument

$$\frac{X_{n_k}}{\mathbb{E}X_{n_k}} \frac{\mathbb{E}X_{n_k}}{\mathbb{E}X_{n_{k+1}}} \;\leq\; \frac{X_n}{\mathbb{E}X_n} \;\leq\; \frac{X_{n_{k+1}}}{\mathbb{E}X_{n_{k+1}}} \frac{\mathbb{E}X_{n_{k+1}}}{\mathbb{E}X_{n_k}} \quad \text{almost surely}$$

and from $\mathbb{E}X_{n_k}/\mathbb{E}X_{n_{k+1}} \sim \log^2 n_k / \log^2 n_{k+1} \sim k^4/(k+1)^4 \to 1$. □

3.2.4 Proof of Theorem 49.

We will use [109, Theorem 2.1], which is written down below in a modified form as suggested by Gnedin, Pitman and Yor [61, Theorem 10]. In the following, for random variables X we use the notation $\|X\|_3 := (\mathbb{E}(|X|^3))^{1/3}$.

Proposition 51. *Assume that a random sequence $(U_n)_{n\in\mathbb{N}}$ of real-valued random variables satisfies the distributional recurrence*

$$U_n \;\stackrel{d}{=}\; U'_{J_n} + V_n, \quad n \geq n_0, \tag{3.36}$$

for some $n_0 \in \mathbb{N}$, where U'_k is assumed independent of (J_n, V_n) and distributed like U_k, for each integer $k \geq n_0$; J_n takes values in $\{0, 1, \ldots, n\}$ and $\mathbb{P}\{J_n = n\} < 1$ for each integer $n \geq n_0$. Suppose further that $\|U_n\|_3 < \infty$ and that, for some constants $C > 0$ and $\alpha > 0$, the following three conditions hold.

(i) $\limsup_{n\to\infty} \mathbb{E}\log\left(\frac{J_n \vee 1}{n}\right) < 0$ *and* $\sup_{n\geq 2} \left\|\log\left(\frac{J_n \vee 1}{n}\right)\right\|_3 < \infty$.

(ii) For some $\lambda \in [0, 2\alpha)$ *and some* $\kappa > 0$, *as* $n \to \infty$,

$$\|V_n - \mu_n + \mu_{J_n}\|_3 \;=\; O(\log^{\kappa} n), \qquad \operatorname{Var} U_n = C\log^{2\alpha} n + O(\log^{\lambda} n),$$

where $\mu_n := \mathbb{E}U_n$.

(iii) $\alpha > 1/3 + \max(\kappa, \lambda/2)$.

Then, as $n \to \infty$, *the sequence* $(U_n - \mu_n)/(\sqrt{C}\log^{\alpha} n)$ *weakly converges to the standard normal law.*

Remark 52. The recurrence (6) is of the form (3.36) with random indices $J_n := I_n$, where I_n has distribution (3.27). By Lemma 53 and (3.28),

$$\mathbb{E}\log\left(\frac{J_n}{n}\right) \;=\; \sum_{i=1}^{n-1} \log\Big(1 - \frac{i}{n}\Big)\mathbb{P}\{I_n = i\} \;\sim\; -\frac{\mathtt{m}}{\log n}.$$

Therefore, $\lim_{n\to\infty} \mathbb{E}\log(J_n/n) = 0$. In particular, the first part of condition (i) in Proposition 51 is not satisfied. Hence, Proposition 51 is not applicable to (6) with I_n having distribution (3.27).

Fix any $T > 0$. The total number X_n of collisions is the sum of the numbers of collisions occurring during the time intervals $[0, T)$ (denote it by $X_n(T)$) and $[T, \infty)$ (denote it by $\widehat{X}_n(T)$). Since the coalescent is a Markov process, $\widehat{X}_n(T) \stackrel{d}{=} X'_{N_n(T)}$, where X'_k is independent of $(J_n, V_n) := (N_n(T), X_n(T))$ and distributed like X_k, for each $k \in \mathbb{N}$. Thus we have proved that (X_n) satisfies another recurrence of the form (3.36), namely

$$X_n \;\stackrel{d}{=}\; X'_{N_n(T)} + X_n(T). \tag{3.37}$$

Proof of Theorem 49. Let us prove that recurrence (3.37) satisfies all the conditions of Proposition 51.

Since $X_n \leq n - 1$ almost surely, $\|X_n\|_3 < \infty$.

Recall that $X_n(T)$ is the number of jumps of the process $\Pi_n(t))_{t\in[0,T)}$. From the construction of the coalescent based on a planar Poisson point

process (see p. 76) it follows that with probability one $X_n(T)$ is bounded from above by a random variable with Poisson distribution with parameter $T\Phi(n)$. By (3.29), $T\Phi(n) \sim (T/B(2,b)) \log n$. Therefore,

$$\|X_n(T)\|_3 \;=\; O(\log n), \quad n \to \infty. \tag{3.38}$$

Let $Q(T) := (\widehat{f}_i(T))_{i\in\mathbb{N}}$ be the decreasing rearrangement of the asymptotic frequencies of the random exchangeable partition $\Pi_\infty(T)$ (corresponding to the beta $(2, b)$ - coalescent). According to Remark 47, $1-\sum_{i=1}^\infty \widehat{f}_i(T) = e^{-\widehat{S}_T}$. The elements of the set $Q(T) \cup \{1 - \sum_{i=1}^\infty \widehat{f}_i(T)\}$ are the lengths of the intervals (from left to right) comprising the partition of $[0, 1]$. Let $U_1, \ldots, U_n$ be independent random variables (points), uniformly distributed on $[0, 1]$ and independent of the lengths of the intervals. Let $W_{n,i}(T)$ be the number of points falling in the interval of length $\widehat{f}_i(T)$. Set

$$\eta_n(T) \;:=\; |\{i \in \{1, \ldots, n\} : U_i > 1 - e^{-\widehat{S}_T}\}|,$$

$$\zeta_n(T) \;:=\; |\{i \geq 1 : W_{n,i}(T) > 0\}|, \quad \theta_n(T) \;:=\; \zeta_n(T) + 1_{\{\eta_n(T)>0\}}.$$

From the paintbox construction [87] of a random exchangeable partition it follows that

$$N_n(T) \;\stackrel{d}{=}\; \zeta_n(T) + \eta_n(T).$$

Arguing in the same way as in [61, p. 592] we conclude that, as $n \to \infty$, $\eta_n(T)/n \to e^{-\widehat{S}_T}$ almost surely, which easily implies that

$$\lim_{n\to\infty} \left(-\log \left(\frac{\eta_n(T) \vee 1}{n} \right) \right) \;=\; \widehat{S}_T \tag{3.39}$$

almost surely, and that, for each $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathbb{E} \left| \left(\log \left(\frac{\eta_n(T) \vee 1}{n} \right) \right)^k \right| \;=\; \mathbb{E}\widehat{S}_T^k. \tag{3.40}$$

Note that in view of (3.30) the right-hand side is finite, for each $k \in \mathbb{N}$. Interpreting the intervals as "boxes" and the points as "balls", the $\theta_n(T)$ is just the number of occupied boxes in the classical multinomial occupancy scheme.

From the results in [51, p. 152] it follows that $\lim_{n\to\infty} n^{-1}\mathbb{E}(\theta_n(T)|(\widehat{f}_i(T))) = 0$ almost surely. This fact together with Proposition 2 of the same reference (see also [84, Theorem 8]) leads to $\lim_{n\to\infty} \theta_n(T)/n = 0$ almost surely conditionally on $(\widehat{f}_i(T))_{i\in\mathbb{N}}$, and, hence, unconditionally. The latter implies that $\lim_{n\to\infty} N_n(T)/n = e^{-\widehat{S}_T}$ almost surely and, hence,

$$\lim_{n\to\infty}\left(-\log\left(\frac{N_n(T)}{n}\right)\right) \;=\; \widehat{S}_T \tag{3.41}$$

almost surely. Since

$$-\log\left(\frac{N_n(T)}{n}\right) \;\le\; -\log\left(\frac{\eta_n(T)\vee 1}{n}\right)$$

almost surely, (3.39), (3.40), and (3.41) together imply that, for each $k\in\mathbb{N}$,

$$\lim_{n\to\infty}\mathbb{E}\left|\left(\log\left(\frac{N_n(T)}{n}\right)\right)^k\right| \;=\; \mathbb{E}\widehat{S}_T^k, \tag{3.42}$$

by a variant of Fatou's lemma, sometimes called Pratt's lemma [115].

The condition (i) of Proposition 51 follows from (3.42). The estimate $\|\mu_n - \mu_{J_n}\|_3 = O(\log n)$ follows from Theorem 46 and (3.42). In view of this observation, (3.38) and Theorem 46, (ii) holds with $\kappa = 1$, $\alpha = 3/2$ and $\lambda = 2$. Therefore, (iii) holds too. □

3.2.5 Auxiliary results.

The proof of Theorem 46 relies upon the following two technical results.

Lemma 53. *For all $k\in\mathbb{N}$ and $b>0$, as $n\to\infty$,*

$$\Big|H(n,b)\sum_{i=1}^{n-1}\mathbb{P}\{I_n=i\}\Big(-\log\Big(\frac{i}{n}\Big)\Big)^k - m_k^{(b)}\Big| \;=\; O\Big(\frac{\log^k n}{n^{b\wedge 1}}\Big), \tag{3.43}$$

where $H(n,b)$ is the function defined after (3.27) and $m_k^{(b)} = k!\zeta(k+1,b)$.

Proof. We first prove that

$$\begin{aligned} J_n(b,k) \;&:=\; \Big|\sum_{i=1}^{n-1}\Big(1-\frac{i}{n}\Big)^{b-1}\frac{1}{i}\Big(-\log\Big(1-\frac{i}{n}\Big)\Big)^k - m_k^{(b)}\Big| \\ &=\; O\Big(\frac{\log^k n}{n^{b\wedge 1}}\Big) \end{aligned} \tag{3.44}$$

and that

$$\begin{aligned} L_n(b,k) &:= \Big|\sum_{i=1}^{n-1}\Big(1-\frac{i}{n}\Big)^{b-1}\frac{1}{i+1}\Big(-\log\Big(1-\frac{i}{n}\Big)\Big)^k - m_k^{(b)}\Big| \\ &= O\Big(\frac{\log^k n}{n^{b\wedge 1}}\Big). \end{aligned} \tag{3.45}$$

Fix $k \in \mathbb{N}$. For $b > 1$, introduce the continuous non-negative function $f_b : [0,1] \to \mathbb{R}$ via $f_b(x) := x^{-1}(1-x)^{b-1}(-\log(1-x))^k$ for $x \in (0,1)$, $f_b(0) := 1_{\{k=1\}}$, and $f_b(1) := 0$. Pick some $\delta \in (0,1)$ such that f_b is non-increasing on $[\delta, 1]$. Then,

$$\begin{aligned} &\Big|\frac{1}{n}\sum_{i=[n\delta]+1}^{n-1} f_b\Big(\frac{i}{n}\Big) - \int_\delta^1 f_b(x)\,\mathrm{d}x\Big| \\ &= \Big|\sum_{i=[n\delta]+1}^{n-1}\int_{i/n}^{(i+1)/n}\Big(f_b\Big(\frac{i}{n}\Big) - f_b(x)\Big)\,\mathrm{d}x - \int_\delta^{([n\delta]+1)/n} f_b(x)\,\mathrm{d}x\Big| \\ &\le \sum_{i=[n\delta]+1}^{n-1}\int_{i/n}^{(i+1)/n}\Big(f_b\Big(\frac{i}{n}\Big) - f_b\Big(\frac{i+1}{n}\Big)\Big)\,\mathrm{d}x + \int_\delta^{([n\delta]+1)/n} f_b(x)\,\mathrm{d}x \\ &= O\Big(\frac{1}{n}\Big). \end{aligned}$$

It is easily checked that f_b is continuously differentiable on $(0,\delta)$ with $\sup_{0<x<\delta}|f_b'(x)| < \infty$. Therefore, exploiting the mean value theorem for differentiable functions, we have

$$\Big|\frac{1}{n}\sum_{i=1}^{[n\delta]} f_b\Big(\frac{i}{n}\Big) - \int_0^\delta f_b(x)\,\mathrm{d}x\Big| = O\Big(\frac{1}{n}\Big).$$

Combining these two pieces together and using the equality $m_k^{(b)} = \int_0^1 f_b(x)\,\mathrm{d}x$, we get $J_n(b,k) = O(1/n)$, which is more than we need.

Assuming that $b \in (0,1]$, an application of the previous result to the function f_{b+1}, which satisfies

$$f_{b+1}(x) = \frac{(1-x)^{b-1}(-\log(1-x))^k}{x} - (1-x)^{b-1}(-\log(1-x))^k$$

for $x \in (0,1)$, yields

$$\Big| \sum_{i=1}^{n-1} \Big(\frac{(1-\frac{i}{n})^{b-1}(-\log(1-\frac{i}{n}))^k}{i} - \frac{(1-\frac{i}{n})^{b-1}(-\log(1-\frac{i}{n}))^k}{n} \Big) - \int_0^1 f_{b+1}(x)\,\mathrm{d}x \Big| \;=\; O\Big(\frac{1}{n}\Big). \qquad (3.46)$$

Note that $\int_0^1 f_{b+1}(x)\,\mathrm{d}x = m_k^{(b)} - k!/b^{k+1}$.

For all $n \in \mathbb{N}$ with $b \log n \geq 1$, we now use the inequalities

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n-1} \Big(\frac{i}{n}\Big)^{b-1} \Big(-\log\Big(\frac{i}{n}\Big)\Big)^k \\ &\qquad \geq \int_{\frac{1}{n}}^1 x^{b-1}(-\log x)^k\,\mathrm{d}x \;=\; \frac{k!}{b^{k+1}}\Big(1 - n^{-b}\sum_{i=0}^{k}\frac{(b\log n)^i}{i!}\Big) \\ &\qquad \geq \frac{k!}{b^{k+1}} - k!\frac{\log^k n}{bn^b}\sum_{i=0}^{k}\frac{1}{i!} \;\geq\; \frac{k!}{b^{k+1}} - k!e\frac{\log^k n}{bn^b} \end{aligned}$$

to conclude that, as $n \to \infty$,

$$\Big|\frac{1}{n}\sum_{i=1}^{n-1}\Big(1-\frac{i}{n}\Big)^{b-1}\Big(-\log\Big(1-\frac{i}{n}\Big)\Big)^k - \frac{k!}{b^{k+1}}\Big| \;=\; O\Big(\frac{\log^k n}{n^b}\Big).$$

Combining this estimate with (3.46) yields (3.44).

Let us now prove (3.45). If $k \in \mathbb{N} \setminus \{1\}$, then

$$\begin{aligned} 0 \;&\leq\; M_n(b,k) \\ &:=\; \sum_{i=1}^{n-1}\frac{(1-\frac{i}{n})^{b-1}(-\log(1-\frac{i}{n}))^k}{i} - \sum_{i=1}^{n-1}\frac{(1-\frac{i}{n})^{b-1}(-\log(1-\frac{i}{n}))^k}{i+1} \\ &\leq\; \sum_{i=1}^{n-1}\frac{(1-\frac{i}{n})^{b-1}(-\log(1-\frac{i}{n}))^k}{i^2} \\ &\sim\; \frac{1}{n}\int_0^1 \frac{(1-x)^{b-1}(-\log(1-x))^k}{x^2}\,\mathrm{d}x, \end{aligned}$$

and the last integral is finite. Therefore, $M_n(b,k) = O(1/n)$, which, together with (3.44), proves (3.45) under the current assumptions.

If $k = 1$, then

$$\begin{aligned} 0 &\leq M_n(b,k) \\ &\leq n^{(1-b)\vee 0} \sum_{i=1}^{n-1} \frac{-\log(1-\frac{i}{n})}{i^2} = n^{(1-b)\vee 0} \sum_{i=1}^{n-1} \frac{1}{i^2} \sum_{j=1}^{\infty} \frac{(\frac{i}{n})^j}{j} \\ &\leq n^{(1-b)\vee 0} \sum_{i=1}^{n-1} \frac{1}{i^2} \sum_{j=1}^{\infty} \Big(\frac{i}{n}\Big)^j = n^{(1-b)\vee 0} \sum_{i=1}^{n-1} \frac{1}{i^2} \frac{\frac{i}{n}}{1-\frac{i}{n}} \\ &= n^{(1-b)\vee 0} \sum_{i=1}^{n-1} \frac{1}{i(n-i)} = n^{(1-b)\vee 0} \frac{1}{n} \sum_{i=1}^{n-1} \Big(\frac{1}{i} + \frac{1}{n-i}\Big) \sim \frac{2\log n}{n^{b\wedge 1}}. \end{aligned}$$

This relation, together with (3.44), proves (3.45).

For $b = 1$, the left-hand side of (3.45) coincides with that of (3.43). Thus, we only have to check (3.43) for $b \neq 1$. To this end, keeping in mind (3.44) and (3.45), it suffices to show that

$$\begin{aligned} &\Big| \sum_{i=1}^{n-1} \Big(\frac{\Gamma(n-i+b-1)\Gamma(n+1)}{\Gamma(n-i)\Gamma(n+b)} - \Big(1-\frac{i}{n}\Big)^{b-1} \Big) \frac{1}{i+1} \Big(-\log\Big(1-\frac{i}{n}\Big)\Big)^k \Big| \\ &\quad = O\Big(\frac{\log^k n}{n^{b\wedge 1}}\Big). \end{aligned} \tag{3.47}$$

First, we will prove that for any $b > 0$, there exists a constant $M > 0$ such that for all $n \in \mathbb{N}$ and all $j \in \mathbb{N}$, $j \leq n-1$

$$\Big| \frac{\Gamma(n-j+b-1)\Gamma(n+1)}{\Gamma(n-j)\Gamma(n+b)} - \Big(1-\frac{j}{n}\Big)^{b-1} \Big| \leq \frac{M}{n} \Big(1-\frac{j}{n}\Big)^{b-2}, \tag{3.48}$$

or, equivalently,

$$\Big| \frac{\Gamma(j+b-1)\Gamma(n+1)}{\Gamma(j)\Gamma(n+b)} - \Big(\frac{j}{n}\Big)^{b-1} \Big| \leq \frac{M}{n} \Big(\frac{j}{n}\Big)^{b-2}. \tag{3.49}$$

The subsequent argument relies on the following inequality (see [1, formula (6.1.47)]). For $c, d > -1$, there exists $M_{c,d} > 0$ such that for all $n \in \mathbb{N}$,

$$\Big| \frac{\Gamma(n+c)}{\Gamma(n+d)} - n^{c-d} \Big| \leq M_{c,d}\, n^{c-d-1}.$$

(3.49) now follows from the chain of inequalities

$$\begin{aligned}
&\Big|\frac{\Gamma(j+b-1)\Gamma(n+1)}{\Gamma(j)\Gamma(n+b)} - \Big(\frac{j}{n}\Big)^{b-1}\Big| \\
&= \Big|\Big(\frac{\Gamma(j+b-1)}{\Gamma(j)} - j^{b-1}\Big)\frac{\Gamma(n+1)}{\Gamma(n+b)} + \frac{\Gamma(n+1)}{\Gamma(n+b)}j^{b-1} - \Big(\frac{j}{n}\Big)^{b-1}\Big| \\
&\le \frac{\Gamma(n+1)}{\Gamma(n+b)}\Big|\frac{\Gamma(j+b-1)}{\Gamma(j)} - j^{b-1}\Big| + j^{b-1}\Big|\frac{\Gamma(n+1)}{\Gamma(n+b)} - n^{1-b}\Big| \\
&\le \frac{\Gamma(n+1)}{\Gamma(n+b)}M_{b-1,0}\,j^{b-2} + j^{b-1}M_{1,b}\,n^{-b} \\
&\le \Big|\frac{\Gamma(n+1)}{\Gamma(n+b)} - n^{1-b}\Big|M_{b-1,0}\,j^{b-2} + n^{1-b}\,M_{b-1,0}\,j^{b-2} + j^{b-1}M_{1,b}\,n^{-b} \\
&\le \frac{M_{1,b}M_{b-1,0}}{n^2}\Big(\frac{j}{n}\Big)^{b-2} + \frac{M_{b-1,0}}{n}\Big(\frac{j}{n}\Big)^{b-2} + \frac{M_{1,b}}{n}\Big(\frac{j}{n}\Big)^{b-1} \\
&\le \frac{M}{n}\Big(\frac{j}{n}\Big)^{b-2},
\end{aligned}$$

where $M := M_{1,b}M_{b-1,0} + M_{b-1,0} + M_{1,b}$. Plugging (3.48) into the left-hand side of (3.47) gives

$$\begin{aligned}
&\Big|\sum_{i=1}^{n-1}\Big(\frac{\Gamma(n-i+b-1)\Gamma(n+1)}{\Gamma(n-i)\Gamma(n+b)} - \Big(1-\frac{i}{n}\Big)^{b-1}\Big)\frac{1}{i+1}\Big(-\log\Big(1-\frac{i}{n}\Big)\Big)^k\Big| \\
&\quad\le \frac{M}{n}\sum_{i=1}^{n-1}\Big(1-\frac{i}{n}\Big)^{b-2}\frac{1}{i+1}\Big(-\log\Big(1-\frac{i}{n}\Big)\Big)^k =: Q_n(b,k).
\end{aligned}$$

For $b > 1$, the function $x \mapsto x^{-1}(1-x)^{b-2}\log^k(1-x)$ is integrable on $[0,1]$, which implies that the latter sum is bounded and the right-hand side in (3.47) is $O(1/n)$. If $b \in (0,1)$, then noting that the function $x \mapsto x^{-1}(-\log(1-x))^k$ is non-decreasing on $(0,1)$, we conclude that, for $n \ge 2$,

$$\begin{aligned}
Q_n(b,k) &= \frac{M}{n^b}\sum_{i=1}^{n-1}(n-i)^{b-2}\frac{1}{(i+1)/n}\Big(-\log\Big(1-\frac{i}{n}\Big)\Big)^k \\
&\le \frac{M}{n^b}\sum_{i=1}^{n-1}(n-i)^{b-2}\frac{1}{i/n}\Big(-\log\Big(1-\frac{i}{n}\Big)\Big)^k \\
&\le \frac{2M\log^k n}{n^b}\sum_{i=1}^{n-1}(n-i)^{b-2} = O\Big(\frac{\log^k n}{n^b}\Big).
\end{aligned}$$

Thus, (3.47) is established and the proof is complete. □

Lemma 54. *Fix $k \in \mathbb{N}$ and $b > 0$, and suppose that $(b_n)_{n\in\mathbb{N}}$ is some sequence satisfying $b_n = O(n^{-b}\log^k n)$. If the sequence $(a_n)_{n\in\mathbb{N}}$ is defined recursively by*

$$a_1 \;:=\; 0, \quad a_n \;:=\; b_n + \sum_{i=1}^{n-1} a_i \mathbb{P}\{I_n = i\}, \quad n \geq 2,$$

where $\mathbb{P}\{I_n = k\}$ is defined in (3.27), then $a_n = O(1)$.

Proof. Since $\mathbb{E}(n - I_n) \sim n/(b\log n)$, there exists an $M > 0$ such that for all $n = 2, 3, \ldots,$

$$\frac{b}{2n^{1+b/2}}\mathbb{E}(n - I_n) \;\geq\; \frac{M\log^k n}{n^b}. \tag{3.50}$$

It suffices to prove the following. If

$$c_1 \;:=\; 0, \quad c_n \;=\; \frac{M\log^k n}{n^b} + \sum_{i=1}^{n-1} c_i \mathbb{P}\{I_n = i\}, \quad n \geq 2,$$

with M defined in (3.50), then

$$c_n \;\leq\; 2 - n^{-b/2} \quad \text{for all } n \in \mathbb{N}. \tag{3.51}$$

We will use induction. For $n = 1$, (3.51) is obviously satisfied as $c_1 = 0$. Assume (3.51) holds for all $n = 1, \ldots, m-1$. Then,

$$c_m \;\leq\; \frac{M\log^k m}{m^b} + \sum_{i=1}^{m-1}(2 - i^{-b/2})\mathbb{P}\{I_m = i\}.$$

We will now verify that the right hand side of the latter inequality is less than or equal to $2 - m^{-b/2}$ or, equivalently, that

$$\sum_{i=1}^{m-1}((m-i)^{-b/2} - m^{-b/2})\mathbb{P}\{I_m = m - i\} \;\geq\; \frac{M\log^k m}{m^b}.$$

The inequality $(1-x)^{-a} \geq 1 + ax$, $x \in (0,1)$, $a > 0$ yields

$$\begin{aligned}
&\sum_{i=1}^{m-1}\big(i^{-b/2} - m^{-b/2}\big)\mathbb{P}\{I_m = i\} \\
&\qquad = \; m^{-b/2}\sum_{i=1}^{m-1}\big((i/m)^{-b/2} - 1\big)\mathbb{P}\{I_m = i\} \\
&\qquad \geq \; \frac{b}{2m^{1+b/2}}\mathbb{E}(m - I_m) \;\geq\; \frac{M\log^k m}{m^b},
\end{aligned}$$

by (3.50). □

3.3 Functionals on the Poisson-Dirichlet coalescent

3.3.1 Main results. The Poisson-Dirichlet (PD) coalescent $\Pi_\infty^{(\gamma)}$ with parameter $\gamma > 0$ introduced by Sagitov in [125] is a particular example of the coalescent with simultaneous multiple collisions.

Denote by $\Pi_n^{(\gamma)}$ the PD coalescent restricted to the set $\{1, 2, \dots, n\}$. Let X_n be the number of collisions, τ_n the absorption time and L_n the total branch length of $\Pi_n^{(\gamma)}$. It was shown in [101] that (X_n), (τ_n) and (L_n) satisfy distributional recurrences (6), (7) and (8), respectively, with T_n having the exponential distribution with parameter $1 - \frac{\gamma^n}{[\gamma]_n}$ and I_n having distribution

$$\mathbb{P}\{I_n = k\} = \frac{\gamma^k}{[\gamma]_n - \gamma^n} s(n,k),\ k \in \mathbb{N}, k \le n-1, \tag{3.52}$$

where $s(n,k)$ are the absolute Stirling numbers of the first kind and $[\gamma]_n := \gamma(\gamma+1)\cdots(\gamma+n-1)$.

The moments of random sequences (X_n) and (τ_n) have the following asymptotics

Theorem 55. *For each $k \in \mathbb{N}$,*

$$\lim_{n\to\infty} \frac{\mathbb{E}Y_n^k}{(\log_\gamma^*(n))^k} = 1,$$

where Y_n denotes either X_n or τ_n, and the function $x \mapsto \log_\gamma^*(x)$ is defined by the functional equation

$$\log_\gamma^*(x) = (1 + \log_\gamma^*(\gamma \log x)) 1_{(e^{2\gamma\vee 1},\infty)}(x).$$

The function $x \mapsto \log_\gamma^*(x)$ is non-decreasing and unbounded; it grows slower than any iteration of the logarithm, i.e., extremely slowly.

From Theorem 55 and Chebyshev's inequality the weak law of large numbers emerges.

Corollary 56. As $n \to \infty$,

$$\frac{X_n}{\log_\gamma^*(n)} \xrightarrow{P} 1 \quad \text{and} \quad \frac{\tau_n}{\log_\gamma^*(n)} \xrightarrow{P} 1.$$

Proposition 57 given next is a weak convergence result for the total branch length L_n of the Poisson-Dirichlet coalescent.

Proposition 57. *As $n \to \infty$,*

$$\frac{L_n}{n} \xrightarrow{d} \eta,$$

where η has the standard exponential distribution.

We will give an elementary proof of this result which is completely different from that suggested by Martin Möhle in [101]. Indeed, (L_n) satisfies

$$L_1 = 0, \quad L_n \stackrel{d}{=} nT_n + L'_{I_n}, \quad n \geq 2.$$

It remains to note that while $L'_{I_n}/n \xrightarrow{P} 0$ in view of the estimates $\mathbb{E}L_n = O(n^2)$ and $\mathbb{E}I_n = O(\log n)$, the relation $T_n \xrightarrow{d} \eta$ holds trivially.

3.3.2 Proof of Theorem 55.

We use the method of iterative functions from Chapter 1. We begin with the proof for X_n.

From (3.52) it follows that

$$\mathbb{E}I_n = \gamma \log n + O(1) \quad \text{and} \quad \operatorname{Var} I_n = \gamma \log n + O(1), \quad n \to \infty.$$

Set

$$g(x) := \gamma \log x, \quad h(x) = 1 \quad \text{and} \quad g^*(x) = \operatorname{Iter}(h, g, x_0),$$

for some fixed $x_0 > \exp(2\gamma \vee 1)$. Such a choice of x_0 guarantees that condition (1.1) holds. Indeed, we have $\delta := x_0 - \gamma \log x_0 > 0$, and $x - \gamma \log x > \delta$, for all $x > x_0$. The function g^* satisfies the functional equation

$$g^*(x) = 1 + g^*(\gamma \log x), \ x > x_0.$$

Let F be the twice differentiable modification of g^* of the form

$$F(x) = \begin{cases} 1 + F(\gamma \log x), & x > x_0, \\ \alpha_1 x^3 + \alpha_2 x^2 + \alpha_3 x, x \in [0, x_0], \end{cases}$$

for some $\alpha_1, \alpha_2, \alpha_3$ (see Lemma 17). For fixed $j \in \mathbb{N}$, we have

$$\begin{aligned} F'(x) &= o\bigg(\frac{1}{x \log x \cdots \log^{\circ(j)}(x)}\bigg) \\ F''(x) &= o\bigg(\frac{1}{x^2 (\log x)^2 \cdots (\log^{\circ(j)}(x))^2}\bigg). \end{aligned}$$

An application of Theorem 12 yields, $\mathbb{E}X_n \sim g^*(n) \sim F(n)$.

Using the induction it is easy to check that, for $k \geq 2$, the sequence $(\mathbb{E}X_n^k)$ satisfies the recurrence

$$\mathbb{E}X_n^k = e_n(k) + \sum_{i=1}^{n-1} \mathbb{P}\{I_n = k\} \mathbb{E}X_i^k, \ n \geq 2,$$

where $e_n(k) = k(g^*(n))^{k-1} + o(k(g^*(n))^{k-1})$ and, for $k \geq 2$, the statement follows from Theorems 13 and 18. Since $g^*(x) \sim \log^* x$, as $x \to \infty$, the proof for X_n is complete.

As to τ_n, recall that a random variable T_n has an exponential distribution with parameter $1 - \frac{\gamma^n}{[\gamma]_n}$. Therefore,

$$\mathbb{E}T_n^k = k! \Big(1 - \frac{\gamma^n}{[\gamma]_n}\Big)^{-k} \to k!, \quad n \to \infty.$$

This implies that $(\mathbb{E}\tau_n^k)$ satisfies the recurrence

$$\mathbb{E}\tau_n^k = e'_n(k) + \sum_{i=1}^{n-1} \mathbb{P}\{I_n = k\} \mathbb{E}\tau_i^k, \ n \geq 2,$$

where $e'_n(k) = k\mathbb{E}T_n \mathbb{E}\tau_n^{k-1} + o(\mathbb{E}T_n \mathbb{E}\tau_n^{k-1})$. From the equivalence $e'_n(k) \sim e_n(k)$ and Theorem 13, we obtain that $\mathbb{E}\tau_n^k \sim \mathbb{E}X_n^k \sim (g^*(n))^k$. The proof is complete.

3.4 Bibliographic comments

The theory of exchangeable coalescents originates from the population genetics. The foundations of the theory were laid down in the pioneering works of Kingman [88, 89]. Based on the classical Wright-Fisher model, Kingman

introduced one of the simplest patterns of the coalescent which is commonly known as *Kingman coalescent* nowadays. Further developments of the theory were made in [39, 73].

In 1999, Pitman [113] and Sagitov [124] independently introduced *coalescents with multiple collisions* also known as *lambda-coalescents*. In particular, their results imply that the coalescents with multiple collisions are uniquely determined by certain finite measures 'lambda' on $[0, 1]$ which justifies the second term.

The last five years have seen an outbreak of activity around the lambda-coalescents and particularly their large sample properties. Important contributions were made by J. Berestycki, N. Berestycki, A. Gnedin, A. Iksanov, M. Möhle, J. Schweinsberg, Y. Yakubovich and others [14, 15, 36, 40, 42, 50, 53, 62, 68, 76, 77, 78, 100, 127].

As it has already been mentioned, the beta (a, b) - coalescents, i.e., the lambda-coalescents driven by measures θ defined in (3.26), had been receiving a lot of attention. This class covers many interesting cases. For instance, in the case $a = 2 - b,\ a \in (0, 1)$ the corresponding coalescent is closely related to b-stable branching [25]. We refer to p. 93 and to [114] and [16] for further multiple connections of these beta-coalescents to various random processes. The case $a = b = 1$ corresponds to the well-known Bolthausen-Sznitman coalescent which was introduced in [26]. The process has connections to stable subordinators [20] and the genealogy of continuous-state branching processes [19]. In the recent years it was an object of intensive research [40, 42, 65, 77, 113].

Now we would like to give an up-to-date overview of the weak convergence results for the three functionals acting on the beta (a, b) - coalescents: the number of collisions X_n, the absorption time τ_n and the total branch length L_n.

Table 1. Weak convergence of $(X_n - a_n)/b_n$ for beta (a, b) - coalescents.

a	b	a_n	b_n	Limit law	Source
$0<a<1$	$b>0$	$n(\alpha-1)$	$(\alpha-1)n^{1/\alpha}$	α-stable	[62]
$a=1$	$b=1$	(see comments)	$\frac{n}{(\log n)^2}$	1-stable	[42, 77]
$1<a<2$	$b>0$	0	κn^{α}	$\int_0^\infty e^{-\alpha S_t}\mathrm{d}t$	[54, 68]
$a=2$	$b>0$	$(2r_1)^{-1}(\log n)^2$	$(3^{-1}r_1^{-3}r_2\log^3 n)^{1/2}$	st. normal	[54, 76]
$a>2$	$b>0$	$m_1^{-1}\log n$	$(m_1^{-3}m_2\log n)^{1/2}$	st. normal	[54, 55]

Notation: $\alpha = 2-a$, $\kappa = \Gamma(\alpha)/\alpha$,

$$r_1 = \zeta(2,b), \quad r_2 = 2\zeta(3,b),$$

where $\zeta(\cdot,\cdot)$ is the Hurwitz zeta function; when $a>2$

$$m_1 = \Psi(a-2+b) - \Psi(b), \quad m_2 = \Psi'(b) - \Psi'(a-2+b),$$

where $\Psi(\cdot)$ is the logarithmic derivative of the gamma function. The Laplace exponent of a pure-jump subordinator $(S_t)_{t\geq 0}$ is given by

$$\Phi(z) = \int_0^1 (1-(1-x)^z)x^{a-3}(1-x)^{b-1}\mathrm{d}x.$$

Comments: In Table 1 the only open case is $a=1$ and $b\neq 1$. In the case $a=b=1$ centering constants are given by

$$a_n = n(\log n)^{-1} + n\log\log n(\log n)^{-2}. \tag{3.53}$$

For the Kingman coalsescent $X_n = n-1$, for all $n\in\mathbb{N}$.

In what follows we assume that the value $a=0$ corresponds to the Kingman coalescent and that c in (3.26) equals $1/B(a,b)$, where B is the beta function.

Table 2. Weak convergence of $(\tau_n - c_n)/d_n$ for beta (a,b) - coalescents.

a	b	c_n	d_n	Limit law	Source
$a=0$		0	1	ρ	[88]
$a=1$	$b=1$	$\log\log n$	1	st. Gumbel	[50, 65]
$1<a<2$	$b>0$	$\mathtt{m}^{-1}\log n$	$(\mathtt{m}^{-3}\mathtt{s}^2\log n)^{1/2}$	st. normal	[54]
$a=2$	$b>0$	$c_1^{-1}\log n$	$(c_1^{-3}c_2\log n)^{1/2}$	st. normal	[54]
$a>2$	$b>0$	$(\gamma m_1)^{-1}\log n$	$\gamma^{-1}(m_1^{-3}(m_2+m_1^2)\log n)^{1/2}$	st. normal	[55]

Notation: The law ρ is an infinite convolution of the exponential laws with parameters $i(i-1)/2$, $i \geq 2$. The distribution function of the standard Gumbel law is $x \mapsto e^{-e^{-x}}$, $x \in \mathbb{R}$. The constants $\mathtt{m}$ and $\mathtt{s}^2$ are defined in (3.13) and (3.14);

$$c_1 = b(b+1)\zeta(2,b), \quad c_2 = 2b(b+1)\zeta(3,b),$$

where ζ is the Hurwitz zeta function. The constants m_1 and m_2 are the same as in Table 1, and when $a > 2$

$$\gamma = \frac{a-1+b}{a-1}\frac{a-2+b}{a-2}.$$

Comments: In the case $a \in (0,1)$, $b > 0$ the beta (a,b) - coalescent comes down from infinity which implies that τ_n weakly converges without any normalization. However, the limiting law is unknown.

The case $a = 1$ and $b \neq 1$ is completely open.

Table 3. Weak convergence of $(L_n - e_n)/f_n$ for beta (a,b) - coalescents.

a	b	e_n	f_n	Limit law	Source
$a=0$		$2\log n$	2	st. Gumbel	[40, 131]
$a=1$	$b=1$	(see comments)	$\frac{n}{(\log n)^2}$	1-stable	[40]
$a>1$	$b>0$	0	n	$\int_0^\infty e^{-S_t}\mathrm{d}t$	[100, 101]

Notation: The Laplace exponent of a pure-jump subordinator $(S_t)_{t\geq 0}$ is given by

$$\Phi(z) = (1/B(a,b)) \int_0^1 (1-(1-x)^z)x^{a-3}(1-x)^{b-1}\mathrm{d}x.$$

Comments: In the case $a = b = 1$ centering constants e_n coincide with a_n given by (3.53). To our knowledge, in the case $0 < a < 1$ and $b = 2 - a$ only the law of large numbers has been proved so far [14]. The authors of [36] made an attempt to settle the case $a \in (0,1)$ and $b > 0$, however their results are incomplete. Summarizing we conclude that the cases $a \in (0,1)$, $b > 0$ and $a = 1$, $b \neq 1$ are open.

The weak convergence result for the number of collisions of the Bolthausen-Sznitman coalescent was first obtained in [42] with the aid of the singular analysis of generating functions. Soon thereafter an alternative,

purely probabilistic proof of this result was worked out in [77]. That proof was based on a coupling with *random walks with barrier*. Later on such an approach was further developed in [78] which allowed one to derive the weak convergence results for the number of collisions in the beta $(a, 1)$ - coalescents with $a \in (0, 2)$.

We would like to stress that several entries of the tables were completed by specializing results obtained for more general lambda-coalescent processes (see [55, 62, 68, 100, 101] and Section 3.1 of the present work).

Coalescents with simultaneous multiple collisions were introduced in [102]. Further results can be found in [48, 49, 99, 101, 125, 126].

Sections 3.1, 3.2 and 3.3 are based on [54], [76] and [95], respectively.

BIBLIOGRAPHY

[1] Abramowitz, M., and Stegun, I. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, New York, 1964.

[2] Alsmeyer, G., Iksanov, A., and Meiners, M. Power and exponential moments of the number of visits and related quantities for perturbed random walks. (2011+), work in progress.

[3] Anderson, K. K., and Athreya, K. B. A note on conjugate Φ-variation and a weak limit theorem for the number of renewals. *Stat. Probab. Letters. 6* (1988), 151–154.

[4] Angel, O., Berestycki, N., and Limic, V. Global divergence of spatial coalescents. *Probab. Theory Relat. Fields.* (2011+), to appear.

[5] Araman, V. F., and Glynn, P. W. Tail asymptotics for the maximum of perturbed random walk. *Ann. Appl. Probab. 16* (2006), 1411–1431.

[6] Archibald, M., Knopfmacher, A., and Prodinger, H. The number of distinct values in a geometrically distributed sample. *Europ. J. Combinat. 27(7)* (2006), 1059–1081.

[7] Arratia, R., Barbour, A. D., and Tavare, S. *Logarithmic combinatorial structures.* European Mathematical Society, 2003.

[8] Bahadur, R. R. On the number of distinct values in a large sample from an infinite discrete distribution. *Proc. Nat. Inst. Sci. India. 26A* (1960), 66–75.

[9] Bahadur, R. R. Some limit theorems in statistics. *CBMS Regional conference series in applied mathematics. Philadelphia: SIAM. 4* (1971).

[10] Bai, Z. D., Hwang, H. K., and Liang, W. Q. Normal approximations of the number of records in geometrically distributed random variables. *Random Struct. Algor. 13* (1998), 319–334.

[11] Barbour, A. D., and Gnedin, A. V. Regenerative compositions in the case of slow variation. *Stoch. Proc. Appl. 116* (2006), 1012–1047.

[12] Barbour, A. D., and Gnedin, A. V. Small counts in the infinite occupancy scheme. *Electron. J. Probab. 14* (2009), 365–384.

[13] Baryshnikov, Y., Eisenberg, B., and Stengle, G. A necessary and sufficient condition for the existence of the limiting probability of a tie for first place. *Stat. Probab. Letters. 23(3)* (1995), 203–209.

[14] Berestycki, J., Berestycki, N., and Schweinsberg, J. Small-time behavior of beta coalescents. *Annales de l'Institut Henri Poincaré 44* (2008), 214–238.

[15] Berestycki, J., Berestycki, N., and V.Limic. The Λ-coalescent speed of coming down from infinity. *Ann. Probab. 38* (2010), 207–233.

[16] Berestycki, N. Recent progress in coalescent theory. *Ensaios Matemáticos. 16* (2009), 1–193.

[17] Bertoin, J. Subordinators: Examples and applications. *Springer Lecture Notes in Math. 1727* (1996).

[18] Bertoin, J. Exchangeable coalescents. lecture notes. *ETH Zürich.* (2010).

[19] Bertoin, J., and Gall, J. F. L. The bolthausen–sznitman coalescent and the genealogy of continuous-state branching processes. *Probab. Theory Related Fields 117* (2000), 249–266.

[20] Bertoin, J., and Pitman, J. Two coalescents derived from the ranges of stable subordinators. *Electron. J. Probab. 5* (2000), 1–17.

[21] Bertoin, J., and Yor, M. Exponential functionals of Lévy processes. *Probability Surveys 2* (2005), 191–212.

[22] Bingham, N. H. Limit theorems for regenerative phenomena, recurrent events and renewal theory. *Z. Wahrsch. Verw. Gebiete. 21* (1972), 20–44.

[23] Bingham, N. H. Maxima of sums of random variables and suprema of stable processes. *Z. Wahrsch. Verw. Gebiete. 26* (1973), 273– 296.

[24] Bingham, N. H., Goldie, C. M., and Teugels, J. L. *Regular variation.* Cambridge, Cambridge University Press, 1989.

[25] Birkner, M., Blath, J., Capaldo, M., Etheridge, A., Möhle, M., Schweinsberg, J., and Wakolbinger, A. Alpha-stable branching and beta-coalescents. *Electron. J. Probab. 10* (2005), 303–325.

[26] Bolthausen, E., and Sznitman, A.-S. On Ruelle's probability cascades and an abstract cavity method. *Comm. Math. Phys. 197* (1998), 247–276.

[27] Brands, J., Steutel, F., and Wilms, R. On the number of maxima in a discrete sample. *Stat. Probab. Letters. 20* (1994), 209–217.

[28] Bruhn, V. *Eine Methode zur asymptotischen Behandlungeiner Klasse von Rekursionsgleichungen mit einer Anwendung in der stochastischen Analyse des Quicksort-Algorithmus.* PhD thesis, Christian-Albrechts-Universität zu Kiel, 1996.

[29] Bruss, F. T., and Grübel, R. On the multiplicity of the maximum in a discrete random sample. *Ann. Appl. Probab. 13(4)* (2003), 1252–1263.

[30] Bruss, F. T., and O'Cinneide, C. A. On the maximum and its uniqueness for geometric random samples. *J. Appl. Probab. 27* (1990), 598–610.

[31] Chen, C.-H., and Fuchs, M. On the moment-transfer approach for random variables satisfying a one-sided distributional recurrence. *preprint* (2010).

[32] Chern, H. H., and Hwang, H. K. Partial match queries in random quadtrees. *SIAM J. Comput. 32(4)* (2003), 904–915.

[33] Chern, H. H., Hwang, H. K., and Tsai, T. H. An asymptotic theory for Cauchy-Euler differential equations with applications to the analysis of algorithms. *Journal of Algorithms 44(1)* (2001), 177–225.

[34] Darling, D. A. Some limit theorems assiciated with multinomial trials. *Proc. Fifth Berkeley Symp. on Math. Statist. and Probab. 2* (1967), 345–350.

[35] de Haan, L., and Resnick, S. I. Conjugate Π-variation and process inversion. *Ann. Probab. 7* (1979), 1028–1035.

[36] Delmas, J. F., Dhersin, J. S., and Siri-Jegousse, A. Asymptotic results on the length of coalescent trees. *Ann. Appl. Probab. 18* (2008), 997—-1025.

[37] Devillers, O. Randomization yields simple $O(n \log^* n)$ algorithms for difficult $\Omega(n)$ problems. *Internat. J. Comput. Geom. Appl. 2* (1992), 621–635.

[38] Devroye, L. Limit laws for sums of functions of subtrees of random binary search trees. *SIAM J. Comput. 32* (2003), 152–171.

[39] Donnely, P., and Tavare, S. Coalescents and genealogical structure under neutrality. *Annual Review of Genetics. 29* (1995), 401–421.

[40] Drmota, M. *Random trees: An interplay between combinatorics and probability.* Springer, 2009.

[41] Drmota, M., Iksanov, A., Moehle, M., and Roesler, U. Asymptotic results about the total branch length of the bolthausen-sznitman coalescent. *Stoch. Proc. Appl. 117* (2007), 1404–1421.

[42] Drmota, M., Iksanov, A., Möhle, M., and Rösler, U. A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree. *Random Struct. Algor. 34* (2009), 319–336.

[43] Durrett, R., and Liggett, T. M. Fixed points of the smoothing transformation. *Z. Wahrsch. Verw. Gebiete. 64* (1983), 275–301.

[44] Dutko, M. Central limit theorems for infinite urn models. *Ann. Probab. 17* (1989), 1255–1263.

[45] Eisenberg, B., Stengle, G., and Strang, G. The asymptotic probability of a tie for first place. *Ann. Appl. Probab. 3* (1993), 731–745.

[46] Fill, J., Mahmoud, H., and Szpankowski, W. On the distribution for the duration of a randomized leader election algorithm. *Ann. Appl. Probab. 6* (1996), 1260–1283.

[47] Flajolet, P., and Sedgewick, R. *Analytic combinatorics.* Cambridge University Press, 2008.

[48] Freund, F. Almost sure asymptotics for mutated external branches of coalescent processes and applications to the number of types. *Electron. Comm. Probab.* (2010+), to appear.

[49] Freund, F., and Möhle, M. On the number of allelic types for samples taken from exchangeable coalescents with mutation. *Adv. Appl. Probab. 41* (2009), 1082–1101.

[50] Freund, F., and Möhle, M. On the time back to the most recent common ancestor and the external branch length of the bolthausen-sznitman coalescent. *Markov Process. Related Fields. 15* (2009), 387–416.

[51] Gnedin, A., Hansen, A., and Pitman, J. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys. 4* (2007), 146–171.

[52] Gnedin, A., Iksanov, A., and Marynych, A. The Bernoulli sieve: an overview. *Discrete Math. Theor. Comput. Sci. Proceedings Series Vol. AM* (2010), 329–342.

[53] Gnedin, A., Iksanov, A., and Marynych, A. Limit theorems fot the number of occupied boxes in the Bernoulli sieve. *Theory Stoch. Proc. 16(32)* (2010), 44–57.

[54] Gnedin, A., Iksanov, A., and Marynych, A. Lambda-coalescents with dust component. *submitted* (2011).

[55] Gnedin, A., Iksanov, A., and Möhle, M. On asymptotics of exchangeable coalescents with multiple collisions. *J. Appl. Probab. 45* (2008), 1186–1195.

[56] Gnedin, A., Iksanov, A., Negadajlov, P., and Roesler, U. The Bernoulli sieve revisited. *Ann. Appl. Probab. 19* (2009), 1634–1655.

[57] Gnedin, A., Iksanov, A., and Roesler, U. Small parts in the Bernoulli sieve. *Discrete Math. Theor. Comput. Sci. Proceedings Series Volume AI* (2008), 235–242.

[58] Gnedin, A., and Pitman, J. Regenerative composition structures. *Ann. Probab. 33* (2005), 445–479.

[59] Gnedin, A., and Pitman, J. Self-similar and markov composition structures. *Zapiski nauchnyh seminarov POMI. 326* (2005), 59–84.

[60] Gnedin, A., Pitman, J., and Yor, M. Asymptotic laws for compositions derived from transformed subordinators. *Ann. Probab. 34* (2006), 468–492.

[61] Gnedin, A., Pitman, J., and Yor, M. Asymptotic laws for regenerative compositions: gamma subordinators and the like. *Probab. Theory Relat. Fields. 135* (2006), 576–602.

[62] Gnedin, A., and Yakubovich, Y. On the number of collisions in Λ - coalescents. *Electron. J. Probab. 12* (2007), 1547–1567.

[63] Gnedin, A. V. The Bernoulli sieve. *Bernoulli. 10* (2004), 79–96.

[64] Goh, W. M. Y., and Hitczenko, P. Gaps in samples of geometric random variables. *Discrete Math. 22* (2007), 2871–2890.

[65] Goldschmidt, C., and Martin, J. Random recursive trees and the Bolthausen-Sznitman coalescent. *Electron. J. Probab. 10* (2005), 718–745.

[66] Greene, D. H., and Knuth, D. E. *Mathematics for the analysis of algorithms, 3d Edition.* Burkhauser, 1990.

[67] Gut, A. *Stopped random walks: Limit theorems and applications, 2nd edition.* Springer: New York, 2009.

[68] Haas, B., and Miermont, G. Self-similar scaling limits of non-increasing Markov chains. *Bernoulli.* (2011+), to appear.

[69] Hardy, G. H. *Divergent Series.* AMS Bookstore, 2000.

[70] Hitczenko, P., and Knopfmacher, A. Gap-free compositions and gap-free samples of geometric random variables. *Discrete Math. 294(3)* (2005), 225–239.

[71] Hitczenko, P., and Wesolowski, J. Renorming divergent perpetuities. *Bernoulli* (2011+), to appear.

[72] Hoare, C. R. Quicksort. *The Computer Journal. 5(1)* (1962), 10–16.

[73] Hudson, R. R. Gene genealogies and the coalescent process. *Oxford Surv. Evol. Biol. 7* (1991), 1–44.

[74] Hwang, H. K. Profiles of random trees: plane-oriented recursive trees. *Random Struct. Algor. 30(3)* (2007), 380–413.

[75] Hwang, H. K., and Neininger, R. Phase change of limit laws in the quicksort recurrence under varying toll functions. *SIAM J. Comput. 31* (2002), 1687–1722.

[76] Iksanov, A., Marynych, A., and Möhle, M. On the number of collisions in beta(2, b)-coalescents. *Bernoulli. 15* (2009), 829–845.

[77] Iksanov, A., and Möhle, M. A probabilistic proof of a weak limit law for the number of cuts needed to isolate the root of a random recursive tree. *Electron. Commun. Probab. 12* (2007), 28–35.

[78] Iksanov, A., and Möhle, M. On the number of jumps of random walks with a barrier. *Adv. Appl. Probab. 40* (2008), 206–228.

[79] Iksanov, A., and Negadajlov, P. On the number of zero increments of random walks with a barrier. *Discrete Math. Theor. Comput. Sci. Proceedings Series Vol. AI* (2008), 243–250.

[80] Iksanov, A., and Terletsky, Y. On asymptotic behavior of certain recursions with random indices of linear growth. *ProbStat Forum 1* (2008), 62–67.

[81] Iksanov, O. M. *Perpetuities, branching random walk and self-decomposability, (in Ukrainian).* Kiev: Zirka, 2007.

[82] Janson, S., Lavault, C., and Louchard, G. Convergence of some leader election algorithm. *Discrete Math. Theor. Comput. Sci. 10(3)* (2008), 171–196.

[83] Janson, S., and Szpankowski, W. Analysis of an asymmetric leader election algorithm. *Electron. J. Combin. 4 #R17* (1997).

[84] Karlin, S. Central limit theorems for certain infinite urn schemes. *J. Math. Mech. 17* (1967), 373–401.

[85] Karp, R. M. Probabilistic recurrence relations. *J. Assoc. Comput. Mach. 41* (1994), 1136–1150.

[86] Kesten, H. The number of distinguishable alleles according to the Ohta-Kimura model of neutral mutation. *J. Math. Biol. 10* (1980), 167–187.

[87] Kingman, J. F. C. The representation of partition structures. *J. London Math. Soc. 18* (1978), 374–380.

[88] Kingman, J. F. C. The coalescent. *Stoch. Proc. Appl. 13* (1982), 235–248.

[89] Kingman, J. F. C. On the genealogy of large populations. *J. Appl. Probab. 19* (1982), 27–43.

[90] Kirschenhofer, P., and Prodinger, H. The number of winners in a discrete geometrically distributed sample. *Ann. Appl. Probab. 6* (1996), 687–694.

[91] Lavault, C., and Louchard, G. Asymptotic analysis of a leader election algorithm. *Theoretical Computer Science. 359(1)* (2006), 239–254.

[92] Louchard, G., and Prodinger, H. On gaps and unoccupied urns in sequence of geometrically distributed random variables. *Discrete Math. 308(9)* (2008), 1538–1562.

[93] Louchard, G., and Prodinger, H. The asymmetric leader election algorithm: another approach. *Ann. Combinat. 12(4)* (2009), 449–478.

[94] Mahmoud, H. *Evolution of random search trees.* Wiley, New York, 1992.

[95] Marynych, A. On the asymptotics of moments of linear random recurrences. *Theory Stoch. Proc. 16(32)* (2010), 106–119.

[96] Marynych, O. V. Asymptotic behaviour of absorption time in decreasing Markov chains (in Ukrainian). *Bulletin of Kiev University. 1* (2010), 118–121.

[97] Meir, A., and Moon, J. Cutting down recursive trees. *Math. Biosci. 21* (1974), 173–181.

[98] Mohamed, H. A probabilistic analysis of a leader election algorithm. *Discrete Math. Theor. Comput. Sci. Proceedings Series Vol. AD* (2006), 225–256.

[99] Möhle, M. On sampling distributions for coalescent processes with simultaneous multiple collisions. *Bernoulli. 12* (2006), 35–53.

[100] Möhle, M. On the number of segregating sites for populations with large family sizes. *Adv. Appl. Probab. 38* (2006), 750–767.

[101] Möhle, M. Asymptotic results for coalescent processes without proper frequencies and applications to the two-parameter Poisson - Dirichlet coalescent. *Stoch. Proc. Appl. 120* (2010), 2159–2173.

[102] Möhle, M., and Sagitov, S. A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab. 29* (2001), 1547–1562.

[103] Negadajlov, P. *Limit theorems for random recurrences and renewal-type processes.* PhD thesis, Utrecht University, 2010.

[104] Negadajlov, P. A. Asymptotic results for the absorption times of random walks with a barrier. *Theor. Probab. Math. Statist. 79* (2008), 127–138.

[105] Negadajlov, P. A. On the absorption times in random walk with barrier (in Ukrainian). *Bulletin of Kiev University. 4* (2008), 149–152.

[106] Neininger, R. On a multivariate contraction method for random recursive structures with applications to Quicksort. *Random Struct. Algor. 19* (2001), 498–524.

[107] Neininger, R. On binary search tree recursions with monomials as toll functions. *J. Comput. Appl. Math. 142* (2002), 185–196.

[108] Neininger, R., and Rüschendorf, L. A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Probab. 14* (2004), 378–418.

[109] Neininger, R., and Ruschendorf, L. On the contraction method with degenerate limit equation. *Ann. Probab. 32* (2004), 2838–2856.

[110] Palmowski, Z., and Zwart, B. On perturbed random walks. *J. Appl. Probab. 47* (2010), 1203–1204.

[111] Panholzer, A. Non-crossing trees revisited: cutting down and spanning subtrees. *Discrete Math. Theor. Comput. Sci. Proceedings Series Vol. AC* (2003), 265–276.

[112] Panholzer, A. Cutting down very simple trees. *Quaest. Math. 29* (2006), 211–227.

[113] Pitman, J. Coalescents with multiple collisions. *Ann. Probab. 27* (1999), 1870–1902.

[114] Pitman, J. Combinatorial stochastic processes. *Springer Lecture Notes in Math. 1875* (2006).

[115] Pratt, J. W. On interchanging limits and integrals. *Ann. Math. Stat. 31* (1960), 74–77.

[116] Prodinger, H. How to select a loser. *Discrete Math. 120* (1993), 149–159.

[117] Rachev, S. T., and Rüschendorf, L. Probability metrics and recursive algorithms. *Adv. Appl. Probab. 27* (1995), 770–799.

[118] Rachev, S. T., and Samorodnitsky, G. Limit laws for a stochastic process and random recursion arising in probabilistic modelling. *Adv. Appl. Probab. 27* (1995), 185–202.

[119] Rösler, U. A limit theorem for "Quicksort". *RAIRO, Inform. Theor. Appl. 25* (1991), 85–100.

[120] Rösler, U. A fixed point theorem for distributions. *Stoch. Proc. Appl. 42* (1992), 195–214.

[121] Rosler, U. On the analysis of stochastic divide and conquer algorithms. *Algorithmica 29* (2001), 238–261.

[122] Rösler, U., and Rüschendorf, L. The contraction method for recursive algorithms. *Algorithmica 29* (2001), 3–33.

[123] Ross, S. M. A simple heuristic approach to simplex efficiency. *Europ. J. Operat. Res. 9* (1982), 344–346.

[124] Sagitov, S. The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab. 36* (1999), 1116–1125.

[125] Sagitov, S. Convergence to the coalescent with simultaneous multiple mergers. *J. Appl. Probab. 40* (2003), 839–854.

[126] Schweinsberg, J. Coalescents with simultaneous multiple collisions. *Electron. J. Probab. 5* (2000), 1–50.

[127] Schweinsberg, J. A necessary and sufficient condition for the Λ-coalescent to come down from infinity. *Electron. Comm. Probab. 5* (2000), 1–11.

[128] Schweinsberg, J. The number of small blocks in exchangeable random partitions. *Alea 7* (2010), 217–242.

[129] Sedgewick, R. The analysis of quicksort programs. *Acta Informatika. 7(4)* (1977), 327–355.

[130] Sgibnev, M. On a renewal function when the second moment is infinite. *Stat. Probab. Letters. 79* (2009), 1242–1245.

[131] Tavaré, S. Ancestral inference in population genetics. *Springer Lecture Notes in Math. 1837* (2004).

[132] van Cutsem, B., and Ycart, B. Renewal-type behaviour of absorption times in markov chains. *Adv. Appl. Probab. 26* (1994), 988–1005.

Acknowledgement

I am grateful to my teachers and daily supervisors Alexander Gnedin and Alexander Iksanov for their help, both direct and indirect, in writing this thesis. The synergetic effect of being a part of the 3-A team was a great experience which tripled my energy. Every time I felt despair and lack of self-confidence in solving mathematical problems their guidance and encouragement helped to bring me back on the rail. I am thankful to Roberto Fernandez for his engagement as promotor.

One person played a special role in my life and must be mentioned here – my late grandfather Prof. Dr. Oleksandr M. Marynych. Although his scientific interests were far from mathematics, his lateral thinking and enthusiasm as researcher had become a role model for me. I owe my decision to become a mathematician to him.

I wish to express my appreciation to the colleagues from the Stochastics Group in Utrecht and the Operations Research Department in Kiev, for stimulating discussions on major subjects of my work. My thanks go to Eduard Belitser and Martin Bootsma for their help with preparation of the Dutch summary of the thesis.

Finally, I am indebted to the members of the reading committee (leescommissie) Jean-François Delmas, Michael Drmota, Michel Mandjes and Ralph Neininger for their careful reading and helpful comments, which have led to significant improvements in several parts of this thesis.

Samenvatting

Stochastische (toevallige) recursieve combinatorische structuren (SRCS) zijn de voornaamste onderzoeksobjecten van dit proefschrift. Onder de term "recursieve combinatorische structuur" verstaan we een zelf-ontbindbaar combinatorisch object dat zelf-gelijkend is: het object kan worden ontleed in kleinere objecten die gelijksoortig zijn aan het originele object. De toevalligheid zit in de manier van het construeren van het object en/of in de manier van het ontleden van het object in delen. Door de interne recursieve structuur van deze objecten kunnen we hun eigenschappen beschrijven via zogenaamde stochastische recursieve rijen (SRR). Door de stochastische functie van deze parameters voor een kleinere SRCS te berekenen met SRR kunnen we de parameters van het grotere SRCS bepalen.

Een van de best bestudeerde klassen van SSR is de klasse van lineaire SRR. Zulke stochastische rijen $(X_n)_{n\in\mathbb{N}}$ worden beschreven door een stochastische vergelijking voor kansverdelingen van de vorm:

$$X_1 = a \geq 0, \qquad X_n \stackrel{d}{=} V_n + \sum_{r=1}^{K} A_r(n) X^{(r)}_{I^n_{(r)}}, \quad n \geq 2.$$

In deze formule is X_n een parameter van een SRCS van grootte n. In plaats van X_n te onderzoeken, kunnen we ook $K \geq 1$ subproblemen van stochastische grootte $I^n_{(r)} \in \{1, \ldots, n\}$ onderzoeken. $V_n \geq 0$ is een stochastische niet-homogene term, de uitdrukkingen $A_r(n) > 0$ zijn de random gewichten van groep r en K is een vast natuurlijk getal. Er wordt verondersteld dat voor elke $r = 1, \ldots, K$ de stochastische variabele $X^{(r)}_k$, die correspondeert met de bijdrage van de subgroep r, onafhankelijk is van $((I^n_{(1)}, \ldots, I^n_{(K)}, A_1(n), \ldots, A_K(n), V_n))_{n\geq 2}$ en dat deze dezelfde verdeling heeft als X_k voor elke $k \in \mathbb{N}$. Deze rijen (vaak in een vereenvoudigde vorm met $K = 1$) beschrijven de eigenschappen van een groot aantal stochastische recursieve objecten zoals de absorptietijden van dalende Markov-ketens, de duur van stochastische algoritmen, functionalen op de coalescenten en random partities, de cyclische structuur van stochastische transposities etc.

Vanuit methodologisch oogpunt bestaat dit proefschrift uit twee delen: in Deel 1 (Part 1 in het proefschrift) wordt een nieuwe methode voor de analyse van de momenten van SRR voorgesteld en wordt een nieuw resultaat over de absorptietijd van dalende Markov-ketens bewezen; het tweede deel (Part 2 en 3) is gewijd aan twee bijzondere modellen voor SRCS: partities voortgebracht door het "stok-breken" en onderling verwisselbare coalescenten.

De eerste helft van het eerste deel van het proefschrift is gewijd aan de methode van iteratieve functies, een nieuwe methode voor de analyse van de momenten van SRR. Oorspronkelijk werd deze methode ontwikkeld voor het bepalen van het asymptotisch gedrag van de momenten van het aantal botsingen X_n en de absorptietijd T_n voor de zogenaamde Poisson-Dirichlet coalescent. Stelling 55, die bewezen wordt met behulp van de methode van iteratieve functies, zegt dat $\mathbb{E}X_n^k$ en $\mathbb{E}T_n^k$, $k \in \mathbb{N}$ asymptotisch equivalent zijn aan de machten van de log*-functie die als volgt is gedefinieerd:

$$\log^*(x) = \begin{cases} 1 + \log^*(\log x) & \text{als } x > 1, \\ 0 & \text{als } x \in [0, 1]. \end{cases}$$

Deze functie groeit langzamer dan elke iteratie van de logaritme. Dit niet-standaard asymptotische gedrag van de momenten maakt duidelijk waarom de bekende methoden van de asymptotische analyse geen resultaten hebben opgeleverd. Dit was het geval voor het aantal botsingen en de absorptietijd voor de Poisson-Dirichlet coalescent hetgeen ons gemotiveerd heeft om een nieuwe methode voor de analyse van SRR te ontwikkelen. Deze methode kan worden toegepast op een groot aantal SRR, zoals het hoofdstuk "Applications" van Part 1 laat zien. Het algoritme beschreven op pagina 21 geeft een stapsgewijs schema voor hoe de methode van iteratieve functies gebruikt kan worden om het asymptotische gedrag van lineaire SRR te bepalen. De hoofdresultaten van dit deel, die de precieze fundering voor dit algoritme geven, zijn Stellingen 11 en 12. In het tweede deel van Deel 1 wordt het zwakke asymptotische gedrag bestudeerd van de absorptietijd X_n voor een zekere klasse van dalende Markov-ketens. In het bijzonder, stelt Stelling 19 de zwakke convergentie van X_n/n naar een exponentiële functionaal van zekere

subordinator vast (met (eventueel) een niet-nul intensiteit van de absorptie en verschuiving). Het bewijs is gebaseerd op de momentenmethode, hetgeen de reden was om deze resultaten in Deel 1 te plaatsen.

In Deel 2 (Part 2) bestuderen we de asymptotiek van het aantal blokken in een random partitie, voortgebracht door het "stok-breken" proces. We beschouwen een partitie van het interval $[0, 1]$ door te kijken naar de punten van een multiplicatieve "random walk" $(Q_k)_{k\in\mathbb{N}_0}$, waarbij

$$Q_0 := 1, \quad Q_j := \prod_{i=1}^{j} W_i, \quad j \in \mathbb{N},$$

en $(W_k)_{k\in\mathbb{N}}$ onafhankelijke kopieën zijn van een stochastische variabele W die waarden aanneemt in het open interval $(0, 1)$. Zij $U_1, \ldots, U_n$ een steekproef uit een uniforme verdeling op $[0, 1]$, onafhankelijk van de multiplicatieve "random walk". We noemen de open intervallen (Q_k, Q_{k-1}) "urnen" en punten $U_1, \ldots, U_n$ "ballen". De gebeurtenis $U_i \in (Q_k, Q_{k-1})$ betekent dat bal i in urn k terecht is gekomen. Op deze manier krijgen we een bezettingsschema van de urnen dat gerelateerd is aan een random partitie van het getal n in niet-negatieve gehele termen:

$$n = \sum_{k=1}^{\infty} \#\{1 \le i \le n : U_i \in (Q_k, Q_{k-1})\}.$$

Een van de meest belangrijke karakteristieken van zulke partities is het aantal termen ongelijk aan nul K_n (ofwel het aantal bezette urnen). In Stelling 23 geven we het algemene resultaat over de partities, voortgebracht door het bovenbeschreven "stok-breken" proces; het gaat over de zwakke convergentie van de op een geschikte wijze gecentreerde en genormeerde rij K_n. Het bewijs van deze stelling is gebaseerd op een koppeling met de zogenaamde "verstoorde random walk". De resultaten voor dit proces zijn afgeleid in gedeelte 2.4 en zijn op zichzelf gezien interessant. Naast het aantal bezette urnen K_n, hebben we ook de functionalen L_n en $K_{n,r}$ onderzocht (Stelling 33, bewering 34 en bewering 38) die respectievelijk het aantal lege urnen tot de laatste bezette urn en het aantal urnen met precies r ballen weergeven. Alle genoemde functionalen zijn SRR.

Het laatste deel van het proefschrift, Deel 3 (Part 3), is gewijd aan onderling verwisselbare coalescenten waarbij meerdere partities tegelijkertijd kunnen samengaan. Het idee is als volgt: Neem n individuen uit een generatie van de populatie en bepaal hun genealogie. Voor elk paar individuen bestaat er een tijdstip waarop de twee stambomen samenkomen bij hun recentste gemeenschappelijke voorouder. Als we met dit proces terug in tijd doorgaan, krijgen we een coalescent-boom: de bladeren zijn de n individuen, de wortel is de recentste gemeenschappelijke voorouder voor de hele steekproef. De boomstructuur van deze coalescent beschrijft de genealogie van de populatie en de boomstructuur hangt af van de intensiteit waarmee stambomen samengevoegd worden. In het simpelste geval, als alleen binaire samenvoegingen zijn toegestaan, krijgen we Kingman's coalescent. Als meerdere samenvoegingen zijn toegestaan, dan heet deze coalescent lambda-coalescent. Op de lambda-coalescenten kan men functionalen invoeren die de boomstructuur beschrijven: het aantal samenvoegingen, de absorptietijden (de tijd tot de recentste gezamenlijke voorouder) enzovoort. Merk op dat deze functionalen SRR zijn. Deel 3 is onverdeeld in 3 gedeeltes. In het eerste gedeelte bestuderen we lambda-coalescenten met positieve intensiteit van eenlingen, i.e., in de limiet dat $n \to \infty$ zijn er voor elke tijd $t > 0$ oneindig veel clusters die nooit betrokken zijn geweest bij een samenvoeging. Voor deze klasse coalescenten leiden we de zwakke asymptotiek voor het aantal botsingen en de absorptietijden (Stellingen 45 en 43) af, door gebruik te maken van een koppeling met een toenemend Lévy-proces. Het tweede gedeelte van Deel 3 gaat over het aantal botsingen X_n in de beta$(2, b)$-coalescent. In het bijzonder, voor X_n bepalen we een ontbinding in de momenten (Stelling 46), leiden we de wet van de grote aantallen af (Gevolg 48) en bewijzen we de centrale limietstelling (Stelling 49). Het laatste gedeelte van Deel 3 is gewijd aan de momenten van de functionalen op de Poisson-Dirichlet coalescent. De in Deel 1 beschreven Stelling 55, die bewezen is met behulp van de methode van iteratieve functies, beschrijft de asymptotiek van de momenten van het aantal botsingen X_n en de absorptietijden T_n voor dit type coalescenten.

Curriculum vitae

Alexander Marynych was born on August the 4^{th}, 1986 in Kiev, Ukraine. He attended a secondary school in Kiev in 1993-2003. In the years from 2003 to 2008 he was studying applied mathematics at the National Taras Shevchenko University of Kiev. He graduated from the university with master's diploma cum laude.

After the graduation he occupied a part-time lectureship position at Faculty of Cybernetics of the University of Kiev.

In the Fall 2008 he engaged in the cooperation project 'Combinatorial stochastic processes' supported by the Utrecht University and the University of Kiev. His contribution to the project is summarised in the present thesis. The results of the work have been reported at a number of international conferences on probability theory, decision making and theory of algorithms.

He plans to do a postdoc at the Eindhoven University of Technology after the defence.