



A phase transition for the heights of a fragmentation tree

Adrien Joseph

► **To cite this version:**

Adrien Joseph. A phase transition for the heights of a fragmentation tree. 2009. <hal-00386468>

HAL Id: hal-00386468

<https://hal.archives-ouvertes.fr/hal-00386468>

Submitted on 21 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A phase transition for the heights of a fragmentation tree

Adrien Joseph

*Laboratoire de Probabilités et Modèles Aléatoires
Université Pierre et Marie Curie
175, rue du Chevaleret
F-75013 Paris, France*

Summary. We provide information about the asymptotic regimes for a homogeneous fragmentation of a finite set. We establish a phase transition for the asymptotic behaviours of the shattering times, defined as the first instants when all the blocks of the partition process have cardinality less than a fixed integer. Our results may be applied to the study of certain random split trees.

Key words. Occupancy scheme, multiplicative cascade, random split tree, homogeneous fragmentation, branching random walk.

A.M.S. Classification. 60 F 15, 60 J 80.

e-mail. `adrien.joseph@etu.upmc.fr`

1 Introduction

The occupancy scheme is a simple urn model which often occurs in probability, statistics, combinatorics and computer science. We can cite for example species sampling [9, 19], analysis of algorithms [16], learning theory [8], etc. The books by Johnson and Kotz [22] and by Kolchin *et al.* [23] are standard references. The present work is partly motivated by the study of fragmentation trees.

Let us recall the occupancy scheme. Let \mathcal{I} be a countable set and $\mathbf{p} = (p_i : i \in \mathcal{I})$ be a probability measure on \mathcal{I} . The occupancy scheme on $(\mathcal{I}, \mathbf{p})$ is described as follows. For all $i \in \mathcal{I}$ such that $p_i \neq 0$, one places a box at i . One then throws successively and independently n balls in the boxes by assuming that each ball has probability p_i of falling into the box located at i . We may be interested in the number of boxes containing exactly j balls, or in the number of occupied boxes, etc.

We consider here a variant of the occupancy scheme which corresponds to a nested family

of boxes. We introduce the infinite genealogical tree \mathcal{U} :

$$\mathcal{U} := \bigcup_{k=0}^{\infty} \mathbb{N}^k,$$

with $\mathbb{N} = \{1, 2, \dots\}$ and the convention $\mathbb{N}^0 = \{\emptyset\}$. The elements of \mathcal{U} are called individuals, and for every integer $k \in \mathbb{Z}_+$, the k -th generation of \mathcal{U} is formed by the individuals $i \in \mathbb{N}^k$ (we write $|i| = k$). For every individual $i = (i_1, \dots, i_k)$ of \mathcal{U} and $j \in \mathbb{N}$, the individual $ij = (i_1, \dots, i_k, j)$ is called the j -th child of i and i is the parent of ij . We suppose a real number $p_i \in [0, 1]$ is assigned to each individual i , such that $p_{\emptyset} = 1$ and $\sum_{j \in \mathbb{N}} p_{ij} = p_i$ for all $i \in \mathcal{U}$. Note that for all k , $\mathbf{p}(k) := (p_i : |i| = k)$ is a probability measure on \mathbb{N}^k . We can couple the occupancy schemes on $(\mathbb{N}^k, \mathbf{p}(k))$ as follows. Each individual i should be viewed as a box, such that the box ij is contained into the box i for every $i \in \mathcal{U}$ and $j \in \mathbb{N}$ such that $p_{ij} \neq 0$. Initially the n balls are thrown in the box located at \emptyset . Then one places successively and independently the n balls in the boxes of the first generation by assuming that each ball has probability p_i of falling into the box situated at $i \in \mathbb{N}$. Likewise, by iteration, a ball located in the box i is placed independently of the others into the sub-box ij with probability p_{ij}/p_i .

We denote by $H_{n,j}$ the first generation at which all the boxes have less than $j \geq 2$ balls when n have been thrown ($H_{n,j}$ is called a *height*) and by $G_{n,j}$ the first level at which there exists a box containing less than $j \geq 1$ balls ($G_{n,j}$ is called a *saturation level*). Our aim is to study the asymptotic behaviours of $H_{n,j}$ and $G_{n,j}$ as n tends to infinity when we consider a certain randomized version of $(p_i, i \in \mathcal{U})$.

More precisely, in the present work, we shall assume that we are given a random probability measure $\rho = (\rho_1, \rho_2, \dots)$ on \mathbb{N} . We assign to each individual i an independent copy $\rho(i)$ of ρ . The real numbers $p_i, i \in \mathcal{U}$, are defined by induction : $p_{\emptyset} := 1$ and $p_{ij} := p_i \rho_j(i)$, for all $i \in \mathcal{U}$ and $j \in \mathbb{N}$. It means that $\rho(i)$ describes how the mass of the individual i is splitted to its children. This model is called the *occupancy scheme of multiplicative cascades*. We put emphasis on the fact that there are two levels of randomness in our model, namely the arrangement of boxes and the way one throws balls.

In the particular case when ρ is supported by a finite number of integers, in the sense that $\#\{j : \rho_j > 0\} \leq b$ a.s. for some integer $b \geq 2$, the height $H_{n,j}$ has a natural interpretation in terms of a special class of random split trees which have been considered *e.g.* by Devoye [14]. Specifically, imagine that each box has a rupture threshold of j , in the sense that when a ball falls into some box i already containing $j - 1$ balls, then this box is removed and the j balls are shared out amongst the children of i according to the random probability $\rho(i)$ (*i.e.* conditionally on $\rho(i)$, each ball is put in the box ij with probability $\rho_j(i)$, independently of the other balls). This procedure yields a random tree where all balls are stored at leaves, and $H_{n,j}$ is the height of this tree when n balls have been thrown.

We further point out that the height $H_{n,j}$ also arises as a natural shattering time in homogeneous fragmentation chains, a class of partition-valued Markov chains. More precisely, the shattering time is defined as the first instant when all the blocks of the partition process have cardinality less than j . See [4] for background and Section 3.2 in [5] for a description which is closely related to the present work. In a different direction, we mention the work of Haas *et al.* [18] who associate another random tree to homogeneous fragmentation processes.

We shall show the following result : there exist an integer $j^* \in \{2, 3, \dots\} \cup \{\infty\}$ and a

sequence of positive real numbers $C_2 > \dots > C_{j^*} = C_{j^*+1} = \dots =: C^*$ such that for every integer $j \geq 2$,

$$H_{n,j} \underset{n \rightarrow \infty}{\sim} C_j \ln n \quad \text{a.s.}$$

It should not be surprising that the heights $H_{n,j}$ have logarithmic asymptotics, as a large class of trees have such a behaviour, like random split trees. On the other hand, it is remarkable that there exists a critical parameter $j^* \in \{2, 3, \dots\} \cup \{\infty\}$ at which a phase transition occurs (provided that $2 < j^* < \infty$). It is a peculiar behaviour for a random split tree. Devroye indeed proved in [14] that for a large class of random split trees such that their internal nodes contain at least one ball¹, then all the heights, regardless of the value of the rupture threshold j , have the same asymptotic behaviour (in probability). In particular, no phase transition occurs.

It may also be interesting to consider the situation where the parameter j depends on the number of balls n . In the case of power functions, we shall prove that under some technical conditions that will be detailed below,

$$H_{n,n^\alpha} \underset{n \rightarrow \infty}{\sim} (1 - \alpha)C^* \ln n \quad \text{a.s.},$$

for all $\alpha \in (0, 1)$, so that there is no phase transition in the asymptotics of H_{n,n^α} .

Concerning the asymptotics of the saturation level $G_{n,j}$, we shall prove the following result : under some technical conditions that will be detailed below, there exists a constant $C_* < C^*$ such that for every positive integer j ,

$$G_{n,j} \underset{n \rightarrow \infty}{\sim} C_* \ln n \quad \text{a.s.}$$

In particular, there is no phase transition in the asymptotics of $G_{n,j}$. We shall also show that under some further technical conditions, for all $\alpha \in (0, 1)$,

$$G_{n,n^\alpha} \underset{n \rightarrow \infty}{\sim} (1 - \alpha)C_* \ln n \quad \text{a.s.},$$

so that there is also no phase transition in the asymptotic behaviours of G_{n,n^α} .

Our approach essentially relies on the theory of branching random walks, and more precisely on their large deviations behaviours whose descriptions are due to Biggins [7]. The construction of the boxes given by the multiplicative cascades indeed enables us to define a branching random walk giving the sizes of the boxes at each generation. We shall see that the critical parameter j^* and the real numbers C_2, \dots, C_{j^*} and C_* are described by that branching random walk. Another key technique is Poissonization; instead of throwing exactly n balls, one throws \mathcal{P}_n balls, where \mathcal{P}_n is a Poisson variable with parameter n which is independent of $(\rho(i), i \in \mathcal{U})$. For every integer k , conditionally on the sizes of the boxes of the k -th generation, the numbers of balls per box of the k -th generation are thus independent Poisson variables.

The results will be stated in Section 2. We shall see the main techniques in Section 3. Section 4 will be devoted to the study of the heights. We shall first turn our attention to the upper bound. Due to the phase transition, we shall give two different proofs to show the

¹We stress that this assumption is crucial in the proof of Theorem 1 in [14] and seems to have been overlooked in the statement.

lower bound of $H_{n,j}$, depending on whether the integer j is less than the critical parameter j^* or not. The results on the saturation levels will be proved in Section 5. Finally, in Section 6, we shall explain heuristically why a phase transition may occur in the asymptotics of $H_{n,j}$, but not in those of H_{n,n^α} , $G_{n,j}$ and G_{n,n^α} .

2 Formulation of the main results

Recall that ρ is a random probability measure on \mathbb{N} . We denote its law by ν . If we denote by $\text{Prob}_{\mathbb{N}}$ the space of probability measures on \mathbb{N} , ν is a probability measure on $\text{Prob}_{\mathbb{N}}$, called the splitting law. We assume that ν is not geometric², in the sense that there is no real number $r \in (0, 1)$ such that with probability one, all the masses of the atoms of ρ belong to $\{r^n, n \in \mathbb{Z}_+\}$. In particular, the degenerate case when ρ is a Dirac point mass a.s. is therefore excluded.

As explained in the introduction, we consider a family $(\rho(i), i \in \mathcal{U})$ of independent copies of ρ labeled by the individuals of the genealogical tree \mathcal{U} . The multiplicative cascade construction defines for each generation k a probability measure $(p_i, |i| = k)$ on \mathbb{N}^k . Taking logarithm of masses, we may encode the latter by the following random point measure on \mathbb{R}_+

$$Z^{(k)}(\mathrm{d}y) := \sum_{i:|i|=k} \delta_{-\ln p_i}(\mathrm{d}y),$$

where δ_z stands for the Dirac point mass at z . Note that if $p_i = 0$, *i.e.* if there is no box at i , the individual i is omitted in the sum defining $Z^{(k)}$. Likewise, in the sequel, we shall always consider individuals which have a positive mass. It follows immediately from the structure of the multiplicative cascades that $(Z^{(k)}, k \in \mathbb{Z}_+)$ is a branching random walk, in the sense that for every integers $k, k' \geq 0$, $Z^{(k+k')}$ is obtained from $Z^{(k)}$ by replacing each atom z of $Z^{(k)}$ by a family $\{z + y, y \in \mathcal{Y}_z\}$, where each \mathcal{Y}_z is an independent copy of the family of atoms of $Z^{(k')}$.

Let us introduce quantities defined via the splitting law ν . First, we define the Laplace transform of the intensity measure $Z^{(1)}$ by

$$L(\theta) := \mathbb{E} [\langle Z^{(1)}, e^{-\theta \cdot} \rangle]$$

for $\theta \in \mathbb{R}$. We can also write

$$L(\theta) = \mathbb{E} \left[\sum_{j \in \mathbb{N}} \rho_j^\theta \right] = \int_{\text{Prob}_{\mathbb{N}}} \left(\sum_{j \in \mathbb{N}} p_j^\theta \right) \nu(\mathrm{d}\mathbf{p})$$

with the convention that $p^\theta = 0$ when $p = 0$ even when $\theta \leq 0$. Because ρ is not a Dirac point mass a.s., $L(0) > 1$. The function $L : \mathbb{R} \rightarrow (0, \infty]$ is decreasing with $L(1) = 1$. We define

$$\underline{\theta} := \inf\{\theta \in \mathbb{R} : L(\theta) < \infty\},$$

²Working with a geometric splitting law would induce a phenomenon of periodicity which we shall not discuss here for simplicity. However results similar to those proven in this work can be established by the same techniques for geometric splitting laws.

so that $L(\theta) < \infty$ when $\theta > \underline{\theta}$. Note that $\underline{\theta}$ may be negative. We can show from Hölder's inequality that $\ln L$ is a convex function, which implies the convexity of L and that the function

$$\varphi : \theta \mapsto \ln L(\theta) - \theta \frac{L'(\theta)}{L(\theta)}$$

is increasing on $(\underline{\theta}, 0)$ and decreasing on $(0, \infty)$. As $L(1) = 1$ and L decreases, we have $\varphi(1) = -L'(1) > 0$, and thus the set of $\theta \in (\underline{\theta}, \infty)$ such that $\varphi(\theta) > 0$ is a non-empty open interval (θ_*, θ^*) :

$$\theta_* := \inf\{\theta > \underline{\theta} : \varphi(\theta) > 0\} \quad \text{and} \quad \theta^* := \sup\{\theta > \underline{\theta} : \varphi(\theta) > 0\}.$$

Note that $\theta^* > 1$ and $\theta_* < 0$ if $\underline{\theta} < 0$. The convexity of $\ln L$ implies that $-L/L'$ is an increasing function. We define

$$C_* := \lim_{\theta \downarrow \theta_*} -\frac{L(\theta)}{L'(\theta)} \quad \text{and} \quad C^* := \lim_{\theta \uparrow \theta^*} -\frac{L(\theta)}{L'(\theta)}.$$

Inspired by the article of Hu and Shi in [21] (see Lemma 4 below), we shall sometimes need the following assumption : there exists $\delta > 0$ such that

$$L(-\delta) < \infty \quad \text{and} \quad \int_{\text{Prob}_{\mathbb{N}}} \left(\sum_{i \in \mathbb{N}} \mathbf{1}_{p_i > 0} \right)^{1+\delta} \nu(d\mathbf{p}) < \infty. \quad (1)$$

To study the asymptotics of the saturation levels, we shall sometimes need the following hypothesis :

$$-\infty < \theta_* < 0 \quad \text{and} \quad \varphi(\theta_*) = 0. \quad (2)$$

We can now state the results that we shall prove. Concerning the asymptotic behaviours of the heights $H_{n,j}$, we have the following results which complete and improve Proposition 2 in [5].

Theorem 1 *Let $j \geq 2$ be an integer, set*

$$C_j := \begin{cases} -j / \ln L(j) & \text{if } j < \theta^*, \\ C^* & \text{if } j \geq \theta^*. \end{cases}$$

Then

$$H_{n,j} \sim C_j \ln n \quad \text{a.s.}$$

More precisely,

$$H_{n,j} \leq C_j \ln n + O(\ln \ln n) \quad \text{a.s.}, \quad (3)$$

and (3) is an equality if $j < \theta^$ or if (1) holds.*

We see that there is a phase transition in the asymptotic behaviour of $H_{n,j}$ at the integer $\lceil \theta^* \rceil$ when $2 < \theta^* < \infty$. We point out that it may happen that $\theta^* \leq 2$ or $\theta^* = \infty$, in which case there is no phase transition. For instance, one can show that if $\rho = (\rho_1, \rho_2, \dots)$ with $\rho_1 = 1 - 0.75U$, $\rho_j = 0.05U$ for all $j \in 2, \dots, 16$ and $\rho_j = 0$ for all $j \geq 17$, where U is uniformly distributed on $(0, 1)$, then $\theta^* < 1.99$. On the other hand, for every $\alpha \in [1/2, 1)$, if

ρ is equal to $(1/2, 1/2, 0, \dots)$ with probability α and to $(1/3, 1/3, 1/3, 0, \dots)$ with probability $1 - \alpha$, then $\theta^* = \infty$.

In this direction, we also point out that the critical parameter θ^* is finite whenever

$$\left\| \max_{j \in \mathbb{N}} \rho_j \right\|_{\infty} = 1,$$

where $\rho = (\rho_j, j \in \mathbb{N})$ denotes a random probability measure on \mathbb{N} with law ν . Indeed, it is easily seen that

$$\lim_{\theta \rightarrow \infty} L(\theta)^{1/\theta} = \left\| \max_{j \in \mathbb{N}} \rho_j \right\|_{\infty}.$$

If the right-hand side equals 1, then $g : \theta \mapsto -\ln L(\theta)/\theta$ has limit 0 at infinity. Now g is derivable and $g(1) = 0$, so there exists $\theta_0 \in (1, \infty)$ such that $g'(\theta_0) = 0$. As $g'(\theta) = \theta^{-2}\varphi(\theta)$, we conclude that $\theta_0 = \theta^* < \infty$.

We may also be interested in the asymptotics of H_{n,n^α} , where $\alpha \in (0, 1)$. We shall prove the following result.

Proposition 1 *Suppose $\theta^* < \infty$. Let $\alpha \in (0, 1)$. Then*

$$H_{n,n^\alpha} \sim (1 - \alpha)C^* \ln n \quad a.s.$$

Furthermore,

$$H_{n,n^\alpha} \leq (1 - \alpha)C^* \ln n + O(\ln \ln n) \quad a.s., \quad (4)$$

and (4) is an equality whenever (1) holds.

Remark 1 The final assertions in Theorem 1 and Proposition 1 rely on the work of Hu and Shi [21]. McDiarmid's setting in [24] can however be considered; we can prove that the results stated in Theorem 1 and in Proposition 1 still hold if the assumption (1) is replaced by the following :

$$\int_{\text{Prob}_{\mathbb{N}}} \left(\sum_{i \in \mathbb{N}} \mathbf{1}_{p_i > 0} \right)^2 \nu(d\mathbf{p}) < \infty. \quad (5)$$

For instance, suppose that $\rho = (\rho_1, 1 - \rho_1, 0, 0, \dots)$, where ρ_1 is a random variable with density $\mathbf{1}_{0 < x < e^{-1}} x^{-1} \ln^{-2} x dx$. Then $\underline{\theta} = 0$, so (1) does not hold. Nonetheless, ρ is supported by two integers a.s., so (5) holds. As a result, for all $j \geq 2$, $H_{n,j} = C_j \ln n + O(\ln \ln n)$ a.s. Furthermore, as $\left\| \max_{j \in \mathbb{N}} \rho_j \right\|_{\infty} = 1$, the discussion below Theorem 1 ensures that θ^* is finite, so for all $\alpha \in (0, 1)$, $H_{n,n^\alpha} = (1 - \alpha)C^* \ln n + O(\ln \ln n)$ a.s.

For the sake of simplicity, we shall show Theorem 1 and Proposition 1 only in the setting of [21] (see Proposition 4 below). The general proof can however be easily carried out.

Concerning the asymptotic behaviours of the saturation levels $G_{n,j}$, we shall prove the following theorem.

Theorem 2 *Let $j \geq 1$ be an integer.*

- *If $\theta_* = -\infty$, then*

$$G_{n,j} \sim C_* \ln n \quad a.s.$$

- Suppose that (2) holds. Then

$$G_{n,j} \sim C_* \ln n \quad a.s.$$

More precisely,

$$G_{n,j} \geq C_* \ln n + O(\ln \ln n) \quad a.s., \quad (6)$$

and (6) is an equality if (1) holds.

A sufficient condition to guarantee that $-\infty < \theta_* < 0$ and $\varphi(\theta_*) = 0$ is :

$$\underline{\theta} > -\infty \quad \text{and} \quad \lim_{\theta \downarrow \underline{\theta}} \uparrow L(\theta) = \infty.$$

Indeed, suppose this condition fulfilled. Imagine that for all $\theta < 0$, $\varphi(\theta) > 0$. Let $\theta_0 \in (\underline{\theta}, 0)$. As $\theta \mapsto \ln L(\theta)/\theta$ is decreasing (its derivative is $-\theta^{-2}\varphi(\theta) < 0$), we have for all $\theta < \theta_0$:

$$\ln L(\theta) \leq \theta \frac{\ln L(\theta_0)}{\theta_0} \leq \underline{\theta} \frac{\ln L(\theta_0)}{\theta_0},$$

which contradicts $\lim_{\theta \rightarrow \underline{\theta}} L(\theta) = \infty$.

We shall also show the following proposition.

Proposition 2 *Suppose that (2) holds. Let $\alpha \in (0, 1)$. Then*

$$G_{n,n^\alpha} \sim (1 - \alpha)C_* \ln n \quad a.s.$$

Moreover,

$$G_{n,n^\alpha} \geq (1 - \alpha)C_* \ln n + O(\ln \ln n) \quad a.s., \quad (7)$$

and (7) is an equality if (1) holds.

We now conclude this section by discussing an illustrative example. Consider the case $\rho = (U, 1 - U, 0, 0, \dots)$, where U is uniformly distributed on $[0, 1]$. Then $L(\theta) = 2/(\theta + 1)$, so $\underline{\theta} = -1$ and (1) holds. The discussions below the theorems yield : $\theta^* < \infty$ and (2) hold. Hence, all our results may be applied. Easy calculations yield $\varphi(\theta) = \ln 2 - \ln(\theta + 1) + \theta/(\theta + 1)$ and $\lceil \theta^* \rceil = 4$, so that there is a phase transition. We can show that $C_2 = 2/\ln(3/2) \approx 4, 93260\dots$, $C_3 = 3/\ln 2 \approx 4, 32808\dots$ and that $C^* > C_*$ are the solutions of the equation

$$\ln \left(\frac{2}{c} \right) + \frac{c-1}{c} = 0,$$

i.e. $C^* \approx 4, 31107\dots$ and $C_* \approx 0, 37336\dots$

As ρ is supported by two integers, our model may be interpreted in terms of random split trees; the procedure described in the introduction yields a random tree where all balls are stored at leaves. We denote by $T_{n,j}$ the tree obtained when n balls have been thrown and when the boxes have a rupture threshold of j . We now define another random split tree $\tilde{T}_{n,j}$ also related to our model. We imagine that each box has a rupture threshold of j , but when a ball falls into some box i already containing $j - 1$ balls, it remains in that box and the $j - 1$ other balls are shared out amongst the two children of i independently and with probability $(1/2, 1/2)$. No other ball is then allowed to be stored at the box i : when a

ball falls into i , it is placed into one of the two children of i with probability $(1/2, 1/2)$, and one has to consider whether the ball stays at that box or not. We denote by $\tilde{T}_{n,j}$ the tree obtained when n balls have been thrown. Note that each internal node of $\tilde{T}_{n,j}$ has exactly one ball, whereas its leaves have less than j balls. We see that the height $\tilde{H}_{n,j}$ of the tree $\tilde{T}_{n,j}$ is less than or equal to the height $H_{n,j}$ of the tree $T_{n,j}$.

It is remarkable that $\tilde{T}_{n,2}$ has the law of the random binary search tree. Robson [27] was the first to be interested in the height of the random binary search tree. In [25], Pittel studied its height and its saturation level. Devroye proved in [11, 12] that the saturation level is asymptotically equivalent to $C_* \ln n$ in probability as in our model, but that the height $\tilde{H}_{n,2}$ is equivalent to $C^* \ln n$ in probability, so that $\tilde{H}_{n,2}$ is not equivalent to $H_{n,2}$. Regarding the random binary search tree as a random split tree, he proved in [14] that all the heights $\tilde{H}_{n,j}$ (regardless of the value of j) of the random binary search tree are asymptotically equivalent to $C^* \ln n$ in probability. In particular, there is no phase transition, contrary to our case. There is of course no contradiction, as the random binary search tree does not correspond to a model treated in this work. Indeed, in terms of random split trees, internal nodes of the binary search tree retain exactly one ball whereas all the balls are stored at leaves in our case.

3 Preliminaries

3.1 Some results on branching random walks

In this section, we recall some results on branching random walks that will be used in the proofs.

We begin by stating a key result obtained by Biggins in [7]. For every $\theta > \underline{\theta}$, we introduce

$$W^{(k)}(\theta) := L(\theta)^{-k} \langle Z^{(k)}, e^{-\theta \cdot} \rangle = L(\theta)^{-k} \sum_{i:|i|=k} p_i^\theta.$$

For every $k \in \mathbb{Z}_+ \cup \{\infty\}$, we denote by \mathcal{F}_k the σ -algebra generated by $(p_i, |i| \leq k)$.

Lemma 1 *For every $\theta > \underline{\theta}$, $(W^{(k)}(\theta), k \in \mathbb{Z}_+)$ is a martingale with respect to the filtration $(\mathcal{F}_k)_{k \in \mathbb{Z}_+}$. Moreover, if $\theta \in (\theta_*, \theta^*)$, it is bounded in $L^\gamma(\mathbb{P})$ for some $\gamma > 1$ and therefore uniformly integrable, and its terminal value*

$$W(\theta) := \lim_{k \rightarrow \infty} W^{(k)}(\theta)$$

is positive a.s.

Applying Corollary 4 in [7], we can obtain a precise estimate of the number of boxes at generation k with size of order $\exp(k L'(\theta)/L(\theta))$. Specifically

Lemma 2 *For all real numbers a and b such that $a < b$ and for all $\theta \in (\theta_*, \theta^*)$, we have with probability one that*

$$\begin{aligned} & \lim_{k \rightarrow \infty} \sqrt{k} e^{-k\varphi(\theta)} \# \{i \in \mathbb{N}^k : \exp(-b + kL'(\theta)/L(\theta)) \leq p_i \leq \exp(-a + kL'(\theta)/L(\theta))\} \\ &= \frac{1}{\sqrt{2\pi}} \frac{e^{\theta b} - e^{\theta a}}{\theta} \left(\frac{L''(\theta)}{L(\theta)} - \left(\frac{L'(\theta)}{L(\theta)} \right)^2 \right)^{-1/2} W(\theta). \end{aligned}$$

We next turn our interest to the asymptotic behaviours of extreme sizes of boxes at generation k :

$$\underline{p}(k) := \inf \{p_i : p_i > 0, |i| = k\} \quad \text{and} \quad \bar{p}(k) := \sup \{p_i : |i| = k\}.$$

We have the general following results :

Lemma 3 *If $\theta^* < \infty$, then*

$$\lim_{k \rightarrow \infty} (-\ln \bar{p}(k) - k/C^*) = \infty \quad \text{and} \quad \lim_{k \rightarrow \infty} -\frac{\ln \bar{p}(k)}{k} = \frac{1}{C^*} \quad a.s.$$

Likewise, if (2) holds, then

$$\lim_{k \rightarrow \infty} (\ln \underline{p}(k) + k/C_*) = \infty \quad \text{and} \quad \lim_{k \rightarrow \infty} -\frac{\ln \underline{p}(k)}{k} = \frac{1}{C_*} \quad a.s.$$

To have precise estimates of the behaviours of the heights and of the saturation levels, we shall sometimes need sharper results of the asymptotics of $\bar{p}(k)$ and $\underline{p}(k)$. The following result, proved by Hu and Shi in [21], will be very useful.

Lemma 4 *We assume that (1) holds.*

- *If $\theta^* < \infty$, then*

$$\limsup_{k \rightarrow \infty} \frac{-\ln \bar{p}(k) - k/C^*}{\ln k} = \frac{3}{2\theta^*} \quad a.s.$$

- *If (2) holds, then*

$$\limsup_{k \rightarrow \infty} \frac{\ln \underline{p}(k) + k/C_*}{\ln k} = -\frac{3}{2\theta_*} \quad a.s.$$

Addario-Berry and Reed in [2] also studied the minima in branching random walks, but they require a stronger condition than (1). Their assumption is however fulfilled for random split trees.

3.2 Poissonization

Let us present the methods used in the proofs. We have to consider the number of balls belonging to each box at each generation. Now, conditionally on \mathcal{F}_k , when n balls have been thrown, the number of balls in the box i , $|i| = k$, follows a binomial law of parameter np_i . Furthermore, these random variables are not independent. A classical idea to circumvent those difficulties (see for instance Gnedin *et al.* in [17] or Holst in [20]) is to consider a randomized version of the total number of balls : instead of throwing initially n balls, one throws \mathcal{P}_n balls, where \mathcal{P}_n is a Poisson variable with parameter n which is independent of $(\rho(i), i \in \mathcal{U})$.

More precisely, we suppose we are given a standard Poisson process $(\mathcal{P}_x)_{x \geq 0}$ independent of \mathcal{F}_∞ . For every individual $i \in \mathcal{U}$ and for every $x \in (0, \infty)$, we denote by $\mathcal{C}(i; x)$ the number of balls at i when the first \mathcal{P}_x balls have been thrown. For all $x, y \in (0, \infty)$ and $k \in \mathbb{Z}_+$, we

denote by $\mathcal{N}_{x,y}(k)$ the number of boxes at generation k containing at least y balls when the first \mathcal{P}_x balls have been thrown :

$$\mathcal{N}_{x,y}(k) := \#\{i \in \mathbb{N}^k : \mathcal{C}(i; x) \geq y\}.$$

Conditionally on \mathcal{F}_k , the random variables $(\mathcal{C}(i; n))_{|i|=k}$ are independent Poisson variables with parameters np_i . That is why in our proofs, we shall first focus on $\mathcal{N}_{n,j}(k)$. We shall then show that $\mathcal{N}_{n,j}(k)$ is close to the number $N_{n,j}(k)$ of boxes at generation k containing at least j balls when exactly n balls have been thrown. Similarly, we denote by $\mathcal{M}_{x,y}(k)$ the number of boxes at generation k containing less than y balls when the first \mathcal{P}_x balls have been thrown :

$$\mathcal{M}_{x,y}(k) := \#\{i \in \mathbb{N}^k : \mathcal{C}(i; x) < y\}$$

and by $M_{n,j}$ the number of boxes at generation k containing less than j balls when n balls have been thrown.

We now prove two estimates using the technique of Poissonization that will then be used in the sequel.

Lemma 5 *Let $p \in (0, \infty)$. There exist two finite constants $c(p)$ and $d(p)$ such that*

$$\sup_{\substack{j \geq p \\ j \in \mathbb{N}}} \sup_{x > 0} \mathbb{P}(\mathcal{P}_x \geq j) j^p x^{-p} \leq c(p), \quad (8)$$

and

$$\sup_{j \in \mathbb{N}} \sup_{x > 0} \mathbb{P}(\mathcal{P}_x < j) j^{-p} x^p \leq d(p). \quad (9)$$

Proof : We begin by showing (8). Let $j \geq p$ be an integer. Let $x > 0$. By Markov's inequality,

$$\mathbb{P}(\mathcal{P}_x \geq j) = \mathbb{P}(\mathcal{P}_x^p \geq j^p) \leq j^{-p} \mathbb{E}[\mathcal{P}_x^p \mathbf{1}_{\mathcal{P}_x \geq j}].$$

Therefore we only have to bound from above

$$x^{-p} \mathbb{E}[\mathcal{P}_x^p \mathbf{1}_{\mathcal{P}_x \geq j}] = x^{-p} \sum_{k=j}^{\infty} k^p e^{-x} \frac{x^k}{k!}.$$

As $\Gamma(k-p+1)k^p/k! \rightarrow 1$ as k tends to infinity, there exists a finite constant $c(p) \geq 1$ such that for all $k > p-1$,

$$\frac{k^p}{k!} \leq \frac{c(p)}{\Gamma(k-p+1)}.$$

As $j \geq p > p-1$, we thus have

$$x^{-p} \sum_{k=j}^{\infty} k^p e^{-x} \frac{x^k}{k!} \leq c(p) e^{-x} \sum_{k=j}^{\infty} \frac{x^{k-p}}{\Gamma(k-p+1)} = c(p) e^{-x} \sum_{k=0}^{\infty} \frac{x^{k+u}}{\Gamma(k+u+1)},$$

where $u := j-p \geq 0$. Applying the formulae 6.5.1, 6.5.4 and 6.5.29 in [1], we get that

$$\sum_{k=0}^{\infty} \frac{x^{k+u}}{\Gamma(k+u+1)} \leq e^x,$$

which proves (8).

To show (9), we write by Markov's inequality :

$$\mathbb{P}(\mathcal{P}_x < j) = \mathbb{P}((\mathcal{P}_x + 1)^{-p} \geq j^{-p}) \leq j^p \mathbb{E}[(\mathcal{P}_x + 1)^{-p}] = j^p \sum_{k=0}^{\infty} (k+1)^{-p} e^{-x} \frac{x^k}{k!}.$$

As $\Gamma(k+p+1)(k+1)^{-p}/k! \rightarrow 1$ as k tends to infinity, there exists a finite constant $d(p) \geq 1$ such that for all $k \in \mathbb{Z}_+$, $\Gamma(k+p+1)(k+1)^{-p}/k! \leq d(p)$. We thus have

$$\mathbb{P}(\mathcal{P}_x < j) \leq d(p) j^p x^{-p} e^{-x} \sum_{k=0}^{\infty} \frac{x^{k+p}}{\Gamma(k+p+1)}.$$

We conclude as before. \square

Remark 2 We stress that in (8), the integer j cannot be less than p . This restriction lies in the heart of the existence of the phase transition in the asymptotics of $H_{n,j}$ (see Proposition 3 below, and more precisely Lemma 6). On the other hand, there is no restriction in (9); no phase transition appears in the asymptotics of $G_{n,j}$.

4 Study of the heights

In this section, we prove Theorem 1 and Proposition 1. We are first interested in the upper bound. We shall then focus on the lower bound. We stress that the phase transition is glimpsed at the beginning of the study of the upper bound (see Remark 3 below). It is however proved in the paragraph dealing with the lower bound.

4.1 Upper bound

Equation (8) will enable us to have a uniform upper bound of $N_{n,j}(k)$ independent of the sizes of the boxes that will eventually lead to the inequalities (3) and (4).

Proposition 3 *Let $j \geq 1$ be an integer and $\alpha \in [0, 1)$ such that $(j, \alpha) \neq (1, 0)$. Then*

$$\limsup_{n \rightarrow \infty} \frac{H_{n, jn^\alpha} - (1 - \alpha) \frac{\theta}{-\ln L(\theta)} \ln n}{\ln \ln n} \leq \frac{1}{-\ln L(\theta)} \quad a.s.$$

for every $\theta > 1$ if $\alpha > 0$, and for every $\theta \in (1, j]$ if $\alpha = 0$.

Remark 3 Suppose that $\alpha = 0$. We deduce from Proposition 3 that :

$$H_{n,j} \leq \min_{\theta \in (1, j]} h(\theta) \ln n + O(\ln \ln n) \quad a.s.,$$

where h is the function $\theta \mapsto -\theta/\ln L(\theta)$. Now, h is derivable and $h'(\theta) = -\varphi(\theta) \ln^{-2} L(\theta)$, so h is decreasing on $(1, \theta^*]$ and increasing on $[\theta^*, \infty)$. The minimum of h is therefore

- $-j/\ln L(j)$ if $j < \theta^*$,

- $-\theta^*/\ln L(\theta^*) = C^*$ if $j \geq \theta^*$.

The phase transition is shown. It will be proved in the study of the lower bound.

Recall that $H_{n,jn^\alpha} \leq k$ if and only if at generation k , every box contains less than jn^α balls when n balls have been thrown. Because we want to show that H_{n,jn^α} is bounded from above by $\left((1-\alpha)\frac{\theta}{-\ln L(\theta)} + \varepsilon\right) \ln(n)$, we take $k \approx \left((1-\alpha)\frac{\theta}{-\ln L(\theta)} + \varepsilon\right) \ln(n)$. In other words, one initially throws $n \approx \exp\left(k\left(\frac{1}{1-\alpha}\frac{-\ln L(\theta)}{\theta} - \varepsilon\right)\right)$ balls and we show that every box of the k -th generation contains less than jn^α balls.

Let u and u' be two real numbers such that

$$u > \frac{1}{(1-\alpha)\theta} \quad \text{and} \quad u < u' < \frac{1}{\alpha} \left(u - \frac{1}{\theta}\right),$$

where for $\alpha = 0$, the second condition reduces to $u' > u$. Define for every $k \geq 1$

$$x_k := k^{-u} \exp\left(k \frac{1}{1-\alpha} \frac{-\ln L(\theta)}{\theta}\right) \quad \text{and} \quad \phi_k := k^{-u'} \exp\left(k \frac{1}{1-\alpha} \frac{-\ln L(\theta)}{\theta}\right).$$

Informally, x_k corresponds to the number of balls thrown when we consider boxes at generation k . Note that $x_k \approx \exp\left(k\left(\frac{1}{1-\alpha}\frac{-\ln L(\theta)}{\theta} - \varepsilon\right)\right)$.

As mentioned in the preliminaries, the argument relies on Poissonization. We are first interested in $\mathcal{N}_{x,y}(k)$ and we shall see how to dePoissonize.

Lemma 6 *For almost all ω , there exists $k_0(\omega)$ such that*

$$\mathcal{N}_{x_k, j\phi_k^\alpha}(k) = 0, \quad \text{for all } k \geq k_0(\omega).$$

Proof : Let $x > 0$ and $y \geq \theta$. We calculate $\mathbb{E}[\mathcal{N}_{x,y}(k)|\mathcal{F}_k]$. We write

$$\mathcal{N}_{x,y}(k) = \sum_{i:|i|=k} \mathbf{1}_{\mathcal{C}(i;x) \geq y} = \sum_{i:|i|=k} \mathbf{1}_{\mathcal{C}(i;x) \geq \lceil y \rceil}.$$

Conditionally on \mathcal{F}_k , $(\mathcal{C}(i;x))_{|i|=k}$ are Poisson variables with parameters xp_i , so

$$\mathbb{E}[\mathcal{N}_{x,y}(k)|\mathcal{F}_k] = \sum_{i:|i|=k} \mathbb{P}(\mathcal{P}_{xp_i} \geq \lceil y \rceil).$$

As $\lceil y \rceil \geq \theta$, (8) ensures that

$$\mathbb{E}[\mathcal{N}_{x,y}(k)|\mathcal{F}_k] \leq \sum_{i:|i|=k} c(\theta) \lceil y \rceil^{-\theta} (xp_i)^\theta$$

(see Remark 2). Now, as $y \geq \theta$,

$$\lceil y \rceil \geq \frac{y-1}{y}y \geq \frac{\theta-1}{\theta}y.$$

In the notations of Lemma 1, we have :

$$\mathbb{E} [\mathcal{N}_{x,y}(k) | \mathcal{F}_k] \leq c'(\theta) y^{-\theta} x^\theta L(\theta)^k W^{(k)}(\theta),$$

where $c'(\theta) := c(\theta)(1 - 1/\theta)^{-\theta}$. We finally get

$$\mathbb{E} [\mathcal{N}_{x,y}(k)] \leq c'(\theta) y^{-\theta} x^\theta L(\theta)^k.$$

Taking $x = x_k$ and $y = j\phi_k^\alpha$, we get for all k sufficiently large so that $j\phi_k^\alpha \geq \theta$ (note that if $\alpha = 0$, then for all k , $j\phi_k^\alpha \geq \theta$) :

$$\mathbb{E} [\mathcal{N}_{x_k, j\phi_k^\alpha}(k)] \leq c'(\theta) j^{-\theta} k^{\theta(u'\alpha - u)}.$$

Now $\theta(u'\alpha - u) < -1$. As a consequence

$$\mathbb{E} \left[\sum_{k \in \mathbb{N}} \mathbf{1}_{\mathcal{N}_{x_k, j\phi_k^\alpha}(k) \geq 1} \right] \leq \mathbb{E} \left[\sum_{k \in \mathbb{N}} \mathcal{N}_{x_k, j\phi_k^\alpha}(k) \right] < \infty.$$

In particular, there is an a.s. finite number of integers k such that $\mathcal{N}_{x_k, j\phi_k^\alpha}(k) \geq 1$. \square

We now show how to have information on $N_{n, jn^\alpha}(k)$ itself.

Proof of Proposition 3 : Let E_k denote the event $\{\mathcal{P}_{x_k} \leq \phi_{k+1}\}$. Because $u' > u$, ϕ_{k+1}/x_k tends to 0 as k tends to infinity, so there exists an integer k_1 such that for all $k \geq k_1$, $\phi_{k+1} \leq x_k/2$. Consequently,

$$\mathbb{P}(E_k) \leq \mathbb{P}(\mathcal{P}_{x_k} \leq x_k/2) = \mathbb{P}(\mathcal{P}_{x_k} - x_k \leq -x_k/2) \leq \mathbb{P}(|\mathcal{P}_{x_k} - x_k| \geq x_k/2).$$

The variance of the Poisson variable \mathcal{P}_{x_k} being x_k , we get by Chebichev's inequality: $\mathbb{P}(E_k) \leq 4x_k^{-1}$ for all $k \geq k_1$. As $\sum_{k \in \mathbb{N}} x_k^{-1} < \infty$, the Borel-Cantelli lemma ensures that for almost all ω , there exists $k_2(\omega)$ such that $\mathcal{P}_{x_k} > \lfloor \phi_{k+1} \rfloor$ for all $k \geq k_2(\omega)$.

Applying Lemma 6, we deduce that, if we define the event Ω_0 by

$$\Omega_0 := \left\{ \omega : \text{there exists } k_3(\omega) \text{ such that } N_{\lfloor \phi_{k+1} \rfloor, j\phi_k^\alpha}(k) = 0 \text{ for all } k \geq k_3(\omega) \right\},$$

then $\mathbb{P}(\Omega_0) = 1$. Notice that there exists an integer k_4 such that for all $k \geq k_4$,

$$k^{-u'} \geq \exp \left(-k \frac{1}{2(1-\alpha)} \frac{-\ln L(\theta)}{\theta} \right).$$

There exists a rank k_5 from which the sequence $(\phi_k)_{k \geq k_5}$ is increasing. Furthermore, that sequence tends to infinity. Let $\omega \in \Omega_0$. Let n be an integer greater than ϕ_{k_5} large enough so that the unique $k \geq k_5$ satisfying $\phi_k \leq n < \phi_{k+1}$ is greater than $k_3(\omega)$ and k_4 . Because $N_{\lfloor \phi_{k+1} \rfloor, j\phi_k^\alpha}(k) = 0$ and $n \leq \lfloor \phi_{k+1} \rfloor$, $N_{n, j\phi_k^\alpha}(k) = 0$. Now, $jn^\alpha \geq j\phi_k^\alpha$, so $N_{n, jn^\alpha}(k) = 0$ and $H_{n, jn^\alpha} \leq k$. Moreover, as $n \geq \phi_k$,

$$k \leq (1-\alpha) \frac{\theta}{-\ln L(\theta)} \ln n + u'(1-\alpha) \frac{\theta}{-\ln L(\theta)} \ln k.$$

Further, as $k \geq k_4$, we have :

$$\exp\left(k \frac{1}{2(1-\alpha)} \frac{-\ln L(\theta)}{\theta}\right) \leq k^{-u'} \exp\left(k \frac{1}{1-\alpha} \frac{-\ln L(\theta)}{\theta}\right) \leq n,$$

so $k \leq 2(1-\alpha) \frac{\theta}{-\ln L(\theta)} \ln n$. Therefore

$$H_{n,jn^\alpha} \leq k \leq (1-\alpha) \frac{\theta}{-\ln L(\theta)} \ln n + u'(1-\alpha) \frac{\theta}{-\ln L(\theta)} \ln\left(2(1-\alpha) \frac{\theta}{-\ln L(\theta)} \ln n\right).$$

We conclude that

$$\limsup_{n \rightarrow \infty} \frac{H_{n,jn^\alpha} - (1-\alpha) \frac{\theta}{-\ln L(\theta)} \ln n}{\ln \ln n} \leq u'(1-\alpha) \frac{\theta}{-\ln L(\theta)} \quad \text{a.s.}$$

The fact that u' can be chosen arbitrarily close to $\frac{1}{(1-\alpha)\theta}$ completes the proof. \square

4.2 Lower bound

We now study the lower bound. Combined with the upper bound supplied by Proposition 3, Theorem 1 and Proposition 1 will be proved.

4.2.1 The case $j \geq \theta^*$ or $\alpha \in (0, 1)$ and $\theta^* < \infty$

In this section, we prove that $H_{n,j} \geq C^* \ln n + O(\ln \ln n)$ a.s. if $j \geq \theta^*$ and (1) holds. We also prove that $H_{n,n^\alpha} \geq (1-\alpha)C^* \ln n + O(\ln \ln n)$ a.s. if $\theta^* < \infty$ and (1) holds. We shall see that the largest box plays a key role.

Proposition 4 *We suppose that $\theta^* < \infty$. Let $j \geq 1$ be an integer and $\alpha \in [0, 1)$ such that $(j, \alpha) \neq (1, 0)$. Under the assumption (1), we have :*

$$\liminf_{n \rightarrow \infty} \frac{H_{n,jn^\alpha} - (1-\alpha)C^* \ln n}{\ln \ln n} \geq \frac{3}{2 \ln L(\theta^*)} \quad \text{a.s.}$$

By definition, $H_{n,jn^\alpha} > k$ if and only if at generation k , there exists a box containing at least jn^α balls when n balls have been thrown. We shall see that in our setting, it suffices to consider the largest box. As we intend to show that H_{n,jn^α} is bounded from below by $((1-\alpha)C^* - \varepsilon) \ln n$, we take $k \approx ((1-\alpha)C^* - \varepsilon) \ln n$, *i.e.* one initially throws $n \approx \exp\left(k \left(\frac{1}{(1-\alpha)C^*} + \varepsilon\right)\right)$ balls and we show that the largest box of the k -th generation contains at least jn^α balls.

Let γ , γ' and γ'' be three real numbers such that

$$\gamma > \frac{3}{2\theta^*}, \quad \gamma' > \frac{\gamma}{1-\alpha} \quad \text{and} \quad \gamma' < \gamma'' < \frac{\gamma' - \gamma}{\alpha}.$$

Note that if $\alpha = 0$, the third condition is simply $\gamma'' > \gamma'$. Define for all $k \in \mathbb{Z}_+$

$$x_k := k^{\gamma'} \exp\left(k \frac{1}{(1-\alpha)C^*}\right) \quad \text{and} \quad \phi_k := k^{\gamma''} \exp\left(k \frac{1}{(1-\alpha)C^*}\right).$$

We first show that a.s. $\mathcal{N}_{x_k, j\phi_k^\alpha}(k) \geq 1$ for all integers k sufficiently large. To do so, we simply consider the largest box.

Lemma 7 For almost all ω , there exists $k_0(\omega)$ such that

$$\mathcal{N}_{x_k, j\phi_k^\alpha}(k) \geq 1, \quad \text{for all } k \geq k_0(\omega).$$

Proof : For every generation $k \in \mathbb{N}$, we consider an imaginary box $\mathbf{b}(k)$ contained in the largest box $\bar{\mathbf{b}}(k)$ of size $\bar{p}(k)$ such that a ball fallen in $\bar{\mathbf{b}}(k)$ is thrown in $\mathbf{b}(k)$ with probability

$$\frac{\mathbf{p}(k)}{\bar{p}(k)} \wedge 1, \quad \text{where } \mathbf{p}(k) := k^{-\gamma} e^{-k/C^*}.$$

Informally, the box $\mathbf{b}(k)$ has size $\mathbf{p}(k) \wedge \bar{p}(k)$. We denote by A_k the event $\{\mathbf{p}(k) < \bar{p}(k)\}$ and by B_k the event defined by

$$B_k := \{\text{the box } \mathbf{b}(k) \text{ contains less than } j\phi_k^\alpha \text{ balls when the first } \mathcal{P}_{x_k} \text{ have been thrown}\}.$$

Conditionally on A_k , the box $\mathbf{b}(k)$ has size $\mathbf{p}(k)$ so the number of balls contained in $\mathbf{b}(k)$ when \mathcal{P}_{x_k} balls have been thrown is a Poisson variable with parameter $x_k \mathbf{p}(k)$. As a result

$$\mathbb{P}(A_k \cap B_k) \leq \mathbb{P}(B_k | A_k) = \mathbb{P}(\mathcal{P}_{x_k \mathbf{p}(k)} < j\phi_k^\alpha).$$

Applying (9), we get :

$$\mathbb{P}(A_k \cap B_k) \leq d(p) j^p k^{-2},$$

where $p := 2/(\gamma' - \gamma - \alpha\gamma'') > 0$, and hence $\sum \mathbb{P}(A_k \cap B_k) < \infty$. By the Borel-Cantelli lemma, we deduce that a.s., for all k sufficiently large, $\mathbf{p}(k) \geq \bar{p}(k)$ or the box $\mathbf{b}(k)$, which is contained in the box $\bar{\mathbf{b}}(k)$, has at least $j\phi_k^\alpha$ balls when the first \mathcal{P}_{x_k} balls have been thrown. Now, by Lemma 4, we know that a.s., for all integers k sufficiently large, $\mathbf{p}(k) < \bar{p}(k)$. Lemma 7 is therefore proved. \square

We now deduce from Lemma 7 the lower bound of H_{n, jn^α} .

Proof of Proposition 4 : The same calculations performed at the beginning of the proof of Proposition 3 show that for almost all ω , there exists $k_1(\omega)$ such that $\mathcal{P}_{x_k} < \lceil \phi_{k-1} \rceil$ for all $k \geq k_1(\omega)$. Applying Lemma 7, we deduce that, if we define the event Ω_0 by

$$\Omega_0 := \{\omega : \text{there exists } k_2(\omega) \text{ such that } N_{\lceil \phi_{k-1} \rceil, j\phi_k^\alpha}(k) \geq 1 \text{ for all } k \geq k_2(\omega)\},$$

then $\mathbb{P}(\Omega_0) = 1$. Let $\omega \in \Omega_0$. The sequence (ϕ_k) is increasing and tends to infinity. Let n be an integer large enough so that the unique integer k satisfying $\phi_{k-1} < n \leq \phi_k$ is greater than $k_2(\omega)$. Because $N_{\lceil \phi_{k-1} \rceil, j\phi_k^\alpha}(k) \geq 1$ and $\lceil \phi_{k-1} \rceil \leq n$, $N_{n, j\phi_k^\alpha}(k) \geq 1$. Now, $j n^\alpha \leq j\phi_k^\alpha$, so $N_{n, jn^\alpha}(k) \geq 1$ and $H_{n, jn^\alpha} > k$. Further, as $n \leq \phi_k$,

$$k \geq (1 - \alpha)C^* \ln n - \gamma''(1 - \alpha)C^* \ln k.$$

As $\exp\left((k-1)\frac{1}{(1-\alpha)C^*}\right) \leq \phi_{k-1} \leq n$, we have : $k \leq (1 - \alpha)C^* \ln n + 1$. Thus

$$H_{n, jn^\alpha} > k \geq (1 - \alpha)C^* \ln n - \gamma''(1 - \alpha)C^* \ln((1 - \alpha)C^* \ln n + 1)$$

and

$$\liminf_{n \rightarrow \infty} \frac{H_{n, jn^\alpha} - (1 - \alpha)C^* \ln n}{\ln \ln n} \geq -\gamma''(1 - \alpha)C^* \quad \text{a.s.}$$

The fact that γ'' can be chosen arbitrarily close to $\frac{3}{2(1-\alpha)\theta^*}$ completes the proof. \square

4.2.2 The case $2 \leq j < \theta^*$ and $\alpha = 0$

We now prove that $H_{n,j} \geq C_j \ln n + O(\ln \ln n)$ a.s. if $j < \theta^*$. We shall notice that contrary to the case $j \geq \theta^*$, the number of boxes matters more than their sizes.

Proposition 5 *Let $j < \theta^*$ be an integer greater than 1. Then*

$$\liminf_{n \rightarrow \infty} \frac{H_{n,j} - \frac{j}{-\ln L(j)} \ln n}{\ln \ln n} \geq \frac{1}{2 \ln L(j)} \quad a.s.$$

The main difference with the case $j \geq \theta^*$ is that we cannot consider the largest box any more. If $\theta^* < \infty$, one can even show by performing the usual calculations that with probability one, for all $k \geq C^* \ln n + o(\ln n)$, the largest box of the k -th generation contains no ball when n balls have been initially thrown. We shall rather focus on some other boxes which are smaller but sufficiently numerous so that it is very unlikely that all of them contain less than j balls when n balls have been thrown. As we want to prove that $H_{n,j}$ is bounded from below by $\left(\frac{j}{-\ln L(j)} - \varepsilon\right) \ln n$, we consider the situation at the k -th generation when one initially throws approximatively $\exp(k(-\ln L(j)/j + \varepsilon))$ balls.

The boxes that will play a key role are those appearing in Lemma 2 (recall that $j < \theta^*$) : with probability one,

$$\lim_{k \rightarrow \infty} \sqrt{k} e^{-k\varphi(j)} \# \{i \in \mathbb{N}^k : \mathbf{s}(k) \leq p_i \leq 2\mathbf{s}(k)\} = Q(j),$$

where

$$\mathbf{s}(k) := \exp\left(-k \frac{-L'(j)}{L(j)}\right)$$

and

$$Q(j) := \frac{1}{\sqrt{2\pi}} \frac{1 - 2^{-j}}{j} \left(\frac{L''(j)}{L(j)} - \left(\frac{L'(j)}{L(j)} \right)^2 \right)^{-1/2} W(j).$$

Notice that, by Lemma 1, as $j < \theta^*$, $Q(j) > 0$ a.s. Define

$$\nu(k) := \lceil Q(j) k^{-1/2} e^{k\varphi(j)} / 2 \rceil.$$

Then (recall that $\varphi(j) > 0$), for almost all ω , there exists $k_0(\omega)$ such that

$$\# \{i \in \mathbb{N}^k : p_i \geq \mathbf{s}(k)\} \geq \nu(k) \geq 1, \quad \text{for all } k \geq k_0(\omega). \quad (10)$$

We can now prove the following result.

Lemma 8 *Define for all $k \in \mathbb{Z}_+$*

$$x_k := k^u \exp\left(k \frac{-\ln L(j)}{j}\right), \quad \text{where } u > \frac{1}{2j}.$$

Then for almost all ω , there exists $k_1(\omega)$ such that

$$\mathcal{N}_{x_k, j}(k) \geq 1, \quad \text{for all } k \geq k_1(\omega).$$

Proof : As \mathbb{N}^k is countable, we can order the boxes of the k -th generation in the decreasing order. Denote by \mathcal{G}_k the family of boxes having a size at least $\mathbf{s}(k)$. We create $\nu(k)$ new boxes : if $\#\mathcal{G}_k \geq \nu(k)$, consider the first $\nu(k)$ boxes belonging to the family \mathcal{G}_k , denoted by $b_1(k), \dots, b_{\nu(k)}(k)$. For all $l \leq \nu(k)$, we place an imaginary box $\mathbf{b}_l(k)$ inside the box $b_l(k)$ such that a ball fallen in $b_l(k)$ is thrown in $\mathbf{b}_l(k)$ with probability $\mathbf{s}(k)/s_l(k)$, where $s_l(k)$ is the size of $b_l(k)$. In particular, every imaginary box has size $\mathbf{s}(k)$. If $\#\mathcal{G}_k < \nu(k)$, we denote by $\mathbf{b}_1(k), \dots, \mathbf{b}_{\nu(k)}(k)$ the first $\nu(k)$ boxes. Introduce the events $A_k := \{\#\mathcal{G}_k \geq \nu(k)\}$ and

$$B_k := \{\forall l \leq \nu(k), \mathbf{b}_l(k) \text{ contains less than } j \text{ balls when the first } \mathcal{P}_{x_k} \text{ have been thrown}\}.$$

Conditionally on \mathcal{F}_∞ and A_k , the boxes $\mathbf{b}_l(k)$ have size $\mathbf{s}(k)$ so the numbers of balls contained in $\mathbf{b}_l(k)$, $l \leq \nu(k)$, when \mathcal{P}_{x_k} balls have been thrown are independent Poisson variables with parameters $x_k \mathbf{s}(k)$. Therefore

$$\mathbb{P}(A_k \cap B_k | \mathcal{F}_\infty) \leq \mathbb{P}(\mathcal{P}_{x_k \mathbf{s}(k)} < j)^{\nu(k)}.$$

Thus

$$\ln \mathbb{P}(A_k \cap B_k | \mathcal{F}_\infty) \leq \frac{Q(j)}{2} k^{-1/2} e^{k\varphi(j)} \ln \mathbb{P}(\mathcal{P}_{x_k \mathbf{s}(k)} < j).$$

It can be easily seen that $x_k \mathbf{s}(k)$ tends to 0, and that, as a result,

$$\ln \mathbb{P}(\mathcal{P}_{x_k \mathbf{s}(k)} < j) \sim -x_k^j \mathbf{s}(k)^j / j!.$$

Thus there exists a constant $c > 0$ such that for all $k \in \mathbb{Z}_+$,

$$\ln \mathbb{P}(\mathcal{P}_{x_k \mathbf{s}(k)} < j) \leq -2c x_k^j \mathbf{s}(k)^j.$$

Finally, we get

$$\mathbb{P}(A_k \cap B_k | \mathcal{F}_\infty) \leq \exp(-cQ(j)k^{-1/2}e^{k\varphi(k)}x_k^j\mathbf{s}(k)^j) = \exp(-cQ(j)k^{ju-1/2}).$$

As $u > \frac{1}{2j}$ and $Q(j) > 0$ a.s., we get $\mathbb{E}[\sum \mathbf{1}_{A_k \cap B_k} | \mathcal{F}_\infty] < \infty$ a.s., so $\sum \mathbf{1}_{A_k \cap B_k} < \infty$ a.s. Combined with (10), this proves that a.s., for all integers k sufficiently large, there exists an imaginary box containing at least j balls when \mathcal{P}_{x_k} have been thrown. As every imaginary box is contained in a real box, Lemma 8 is proved. \square

We now deduce from Lemma 8 the lower bound of $H_{n,j}$.

Proof of Proposition 5 : Let u and u' be two real numbers such that $\frac{1}{2j} < u < u'$. Define $x_k := k^u \exp(-k \ln L(j)/j)$ and $\phi_k := k^{u'} \exp(-k \ln L(j)/j)$. One can show that for almost all ω , there exists $k_2(\omega)$ such that $\mathcal{P}_{x_k} < \lceil \phi_{k-1} \rceil$ for all $k \geq k_2(\omega)$. Applying Lemma 8, we deduce that if we define the event Ω_0 by

$$\Omega_0 := \{\omega : \text{there exists } k_3(\omega) \text{ such that } N_{\lceil \phi_{k-1} \rceil, j}(k) \geq 1 \text{ for all } k \geq k_3(\omega)\},$$

then $\mathbb{P}(\Omega_0) = 1$. The sequence (ϕ_k) is increasing and tends to infinity. Let $\omega \in \Omega_0$. Let n be an integer large enough so that the unique integer k satisfying $\phi_{k-1} < n \leq \phi_k$ is greater

than $k_3(\omega)$. Because $N_{\lceil\phi_{k-1}\rceil,j}(k) \geq 1$ and $n \geq \lceil\phi_{k-1}\rceil$, $N_{n,j}(k) \geq 1$, so $H_{n,j} > k$. Moreover, as $n \leq \phi_k$,

$$k \geq \frac{j}{-\ln L(j)} \ln n - u' \frac{j}{-\ln L(j)} \ln k.$$

As $n \geq \phi_{k-1} \geq \exp(-(k-1) \ln L(j)/j)$, we have : $k \leq \frac{j}{-\ln L(j)} \ln n + 1$. Therefore

$$H_{n,j} > k \geq \frac{j}{-\ln L(j)} \ln n - u' \frac{j}{-\ln L(j)} \ln \left(\frac{j}{-\ln L(j)} \ln n + 1 \right).$$

We conclude that

$$\liminf_{n \rightarrow \infty} \frac{H_{n,j} - \frac{j}{-\ln L(j)} \ln n}{\ln \ln(n)} \geq -u' \frac{j}{-\ln L(j)} \quad \text{a.s.}$$

The fact that u' can be chosen arbitrarily close to $\frac{1}{2j}$ completes the proof. \square

Remark 4 Theorem 1 states that $H_{n,j} \sim C_j \ln n$ a.s. This asymptotic behaviour was proved in the case $\theta^* = \infty$ (we even showed that $H_{n,j} = C_j \ln n + O(\ln \ln n)$ a.s.). If $\theta^* < \infty$, then $-\ln \bar{p}(k)/k$ tends to $1/C^*$ a.s. (see Lemma 3). We deduce from an argument similar to that in Proposition 4 that $H_{n,j} \sim C_j \ln n$ a.s. Finally, Theorem 1 and Proposition 1 have been proved.

5 Study of the saturation levels

In this section, we prove Theorem 2 and Proposition 2. We shall first assume that (2) holds. We shall then study the case $\theta_* = -\infty$ to complete the proof of Theorem 2.

5.1 The case $-\infty < \theta_* < 0$ and $\varphi(\theta_*) = 0$

Throughout this section, we suppose that (2) holds. In particular, $C_* = -\theta_*/\ln L(\theta_*) = -L(\theta_*)/L'(\theta_*)$. We are first interested in the lower bound. We shall then focus on the upper bound. We are inspired by the techniques developed in the previous section.

5.1.1 Lower bound

We show the inequalities (6) and (7) when (2) holds. Equation (9) will be the key tool.

Proposition 6 *Suppose that (2) holds. Let $j \geq 1$ be an integer and $\alpha \in [0, 1)$. Then*

$$\liminf_{n \rightarrow \infty} \frac{G_{n,jn^\alpha} - (1-\alpha)C_* \ln(n)}{\ln \ln(n)} \geq -\frac{1}{\ln L(\theta_*)} \quad \text{a.s.}$$

Recall that $G_{n,jn^\alpha} > k$ if and only if at generation k , every box contains at least jn^α balls when n balls have been thrown. Because we want to show that G_{n,jn^α} is bounded from below by $((1-\alpha)C_* - \varepsilon) \ln(n)$, we take $k \approx ((1-\alpha)C_* - \varepsilon) \ln(n)$, i.e. $n \approx \exp\left(k \left(\frac{1}{(1-\alpha)C_*} + \varepsilon\right)\right)$.

Proof : Let u and u' be two real numbers such that

$$u > -\frac{1}{(1-\alpha)\theta_*} \quad \text{and} \quad u < u' < \frac{1}{\alpha} \left(u + \frac{1}{\theta_*} \right),$$

where the second condition reduces to $u' > u$ for $\alpha = 0$. Define for all $k \in \mathbb{N}$:

$$x_k := k^u \exp \left(k \frac{1}{(1-\alpha)C_*} \right) \quad \text{and} \quad \phi_k := k^{u'} \exp \left(k \frac{1}{(1-\alpha)C_*} \right).$$

Proceeding as usual, let us prove the following key result : for almost all ω , there exists $k_0(\omega)$ such that

$$\mathcal{M}_{x_k, j\phi_k^\alpha}(k) = 0, \quad \text{for all } k \geq k_0(\omega). \quad (11)$$

Let x and y be two real numbers such that $y \geq j$ and $k \in \mathbb{N}$. We calculate $\mathbb{E}[\mathcal{M}_{x,y}(k)|\mathcal{F}_k]$. We write

$$\mathcal{M}_{x,y}(k) = \sum_{i:|i|=k} \mathbf{1}_{\mathcal{C}(i;x) < y} = \sum_{i:|i|=k} \mathbf{1}_{\mathcal{C}(i;x) < [y]}$$

Now, conditionally on \mathcal{F}_k , $(\mathcal{C}(i;x))_{|i|=k}$ are Poisson variables with parameters xp_i , so

$$\mathbb{E}[\mathcal{M}_{x,y}(k)|\mathcal{F}_k] = \sum_{i:|i|=k} \mathbb{P}(\mathcal{P}_{xp_i} < [y]).$$

Applying (9), we get

$$\mathbb{E}[\mathcal{M}_{x,y}(k)|\mathcal{F}_k] \leq d(-\theta_*)[y]^{-\theta_*} \sum_{i:|i|=k} (xp_i)^{\theta_*}.$$

Now, as $y \geq j$,

$$[y] \leq \frac{y+1}{y}y \leq \frac{j+1}{j}y.$$

In the notations of Lemma 1, we have :

$$\mathbb{E}[\mathcal{M}_{x,y}(k)|\mathcal{F}_k] \leq d'y^{-\theta_*}x^{\theta_*}L(\theta_*)^kW^{(k)}(\theta_*),$$

where $d' := d(-\theta_*)(1+1/j)^{-\theta_*}$. We finally get

$$\mathbb{E}[\mathcal{M}_{x,y}(k)] \leq d'y^{-\theta_*}x^{\theta_*}L(\theta_*)^k.$$

For $x = x_k$ and $y = j\phi_k^\alpha \geq j$, we obtain for every integer $k \in \mathbb{N}$:

$$\mathbb{E}[\mathcal{M}_{x_k, j\phi_k^\alpha}(k)] \leq d'j^{-\theta_*}k^{\theta_*(u-u'\alpha)}.$$

Now $\theta_*(u-u'\alpha) < -1$. As a consequence

$$\mathbb{E} \left[\sum_{k \in \mathbb{Z}_+} \mathbf{1}_{\mathcal{M}_{x_k, j\phi_k^\alpha}(k) \geq 1} \right] \leq \mathbb{E} \left[\sum_{k \in \mathbb{Z}_+} \mathcal{M}_{x_k, j\phi_k^\alpha}(k) \right] < \infty.$$

In particular, there is an a.s. finite number of integers k such that $\mathcal{M}_{x_k, j\phi_k^\alpha}(k) \geq 1$, which proves (11).

Performing the same calculations as at the end of the proof of Proposition 4, one finally gets :

$$\liminf_{n \rightarrow \infty} \frac{G_{n, jn^\alpha} - (1-\alpha)C_* \ln n}{\ln n} \geq -u'(1-\alpha)C_* \quad \text{a.s.}$$

The fact that u' can be chose arbitrarily close to $-\frac{1}{(1-\alpha)\theta_*}$ completes the proof. \square

5.1.2 Upper bound

In this section, we are interested in the upper bound of the saturation levels $G_{n,j}$. We prove that the inequalities (6) and (7) are in fact equalities whenever (1) and (2) hold by studying the smallest box, regardless of the value of j ; there is no phase transition.

Proposition 7 *Suppose that (1) and (2) hold. Let $j \geq 1$ be an integer and $\alpha \in (0, 1)$. Then*

$$\limsup_{n \rightarrow \infty} \frac{G_{n,j} - C_* \ln n}{\ln \ln n} \leq \frac{3}{2 \ln L(\theta_*)} + \frac{C_*}{j} \quad a.s.$$

and

$$\limsup_{n \rightarrow \infty} \frac{G_{n,jn^\alpha} - (1 - \alpha)C_* \ln n}{\ln \ln n} \leq \frac{3}{2 \ln L(\theta_*)} \quad a.s.$$

In order to prove both inequalities simultaneously, we may suppose that $\alpha \in [0, 1)$, the case $G_{n,j}$ corresponding to G_{n,jn^α} with $\alpha = 0$.

By definition, $G_{n,jn^\alpha} \leq k$ if and only if at generation k , there exists a box containing less than jn^α balls when n balls have been thrown. We shall see that it suffices to consider the smallest box. As we intend to show that G_{n,jn^α} is bounded from above by $((1 - \alpha)C_* + \varepsilon) \ln n$, we take $k \approx ((1 - \alpha)C_* + \varepsilon) \ln n$, i.e. $n \approx \exp\left(k \left(\frac{1}{(1 - \alpha)C_*} - \varepsilon\right)\right)$.

Proof : Let γ, γ' and γ'' be three real numbers such that :

- $\gamma > -\frac{3}{2\theta_*}$, $\gamma' > \gamma + \frac{1}{j}$ and $\gamma'' > \gamma'$ if $\alpha = 0$,
- $\gamma > -\frac{3}{2\theta_*}$, $\gamma' > \frac{\gamma}{1 - \alpha}$ and $\gamma' < \gamma'' < \frac{\gamma' - \gamma}{\alpha}$ if $\alpha > 0$.

Define

$$x_k := k^{-\gamma'} \exp\left(k \frac{1}{(1 - \alpha)C_*}\right) \quad \text{and} \quad \phi_k := k^{-\gamma''} \exp\left(k \frac{1}{(1 - \alpha)C_*}\right).$$

Let us prove the following result : for almost all ω , there exists $k_0(\omega)$ such that

$$\mathcal{M}_{x_k, j\phi_k^\alpha}(k) \geq 1, \quad \text{for all } k \geq k_0(\omega). \quad (12)$$

Define $\mathbf{p}(k) := k^\gamma e^{-k/C_*}$. We consider an imaginary box $\mathbf{b}(k)$ at generation k such that

1. if $1 \leq \mathbf{p}(k)$: every ball thrown is placed in $\mathbf{b}(k)$.
2. if $\underline{\mathbf{p}}(k) < \mathbf{p}(k) < 1$, where $\underline{\mathbf{p}}(k)$ is the size of the smallest box $\underline{\mathbf{b}}(k)$: every ball fallen in $\underline{\mathbf{b}}(k)$ is also placed in the imaginary box and every other ball is placed in $\mathbf{b}(k)$ with probability $(\mathbf{p}(k) - \underline{\mathbf{p}}(k))/(1 - \underline{\mathbf{p}}(k))$. Hence, the imaginary box has size $\mathbf{p}(k)$.
3. if $\mathbf{p}(k) \leq \underline{\mathbf{p}}(k)$: every ball fallen in the smallest box is placed in the imaginary box with probability $\mathbf{p}(k)/\underline{\mathbf{p}}(k)$, and no other ball is placed in $\mathbf{b}(k)$.

The box $\mathfrak{b}(k)$ has thus size $\mathfrak{p}(k) \wedge 1$, and whenever $\mathfrak{p}(k) > \underline{p}(k)$, it contains the smallest box. From Lemma 4, we know that a.s., for all integers k sufficiently large, $\mathfrak{p}(k) > \underline{p}(k)$. To prove Lemma 12, all we have to do is therefore to prove that $\mathbb{P}(\limsup A_k) = 0$, where the event A_k is defined by

$$A_k := \{\mathfrak{b}(k) \text{ contains at least } j\phi_k^\alpha \text{ balls when the first } \mathcal{P}_{x_k} \text{ balls have been thrown}\}.$$

By the Borel-Cantelli lemma, it suffices to show that $\sum \mathbb{P}(A_k) < \infty$. Let k be an integer sufficiently large so that $\mathfrak{p}(k) < 1$. The imaginary box has then size $\mathfrak{p}(k)$, so we have :

$$\mathbb{P}(A_k) = \mathbb{P}(\mathcal{P}_{x_k \mathfrak{p}(k)} \geq j\phi_k^\alpha)$$

We would like to apply (8). To do so, let $p := j$ if $\alpha = 0$ and $p := 2/(\gamma' - \gamma - \alpha\gamma'') > 0$ if $\alpha > 0$. Note that for all integers k sufficiently large, $j\phi_k^\alpha \geq p$. Equation (8) then ensures that :

$$\mathbb{P}(A_k) \leq c(p)j^{-p}k^{p(\alpha\gamma'' + \gamma - \gamma')}.$$

As $p(\alpha\gamma'' + \gamma - \gamma') < -1$, $\sum \mathbb{P}(A_k) < \infty$, which proves (12).

Performing the same calculations as at the end of the proof of Proposition 3, one finally gets :

$$\limsup_{n \rightarrow \infty} \frac{G_{n, jn^\alpha} - (1 - \alpha)C_* \ln n}{\ln \ln n} \leq \gamma''(1 - \alpha)C_* \quad \text{a.s.}$$

The fact that γ'' can be chosen arbitrarily close either to $-\frac{3}{2\theta_*} + \frac{1}{j}$ if $\alpha = 0$ or to $-\frac{3}{2(1-\alpha)\theta_*}$ if $\alpha > 0$ completes the proof. \square

Proposition 2 and a part of Theorem 2 have been proved. We now turn our attention to the other part of Theorem 2.

5.2 The case $\theta_* = -\infty$

In this section, we prove that if $\theta_* = -\infty$, then $G_{n,j} \sim C_* \ln n$ a.s. We begin by showing the following lemma :

Lemma 9 *If $\theta_* = -\infty$, then $-\theta/\ln L(\theta)$ tends to C_* as θ tends to $-\infty$.*

Proof : The condition $\theta_* = -\infty$ means that $\varphi(\theta) > 0$ for all $\theta < 0$. Let $\psi = \ln L$. Recall that $C_* = \lim_{\theta \rightarrow -\infty} -1/\psi'(\theta)$. As ψ is convex decreasing, it is known that $-\psi(\theta)/\theta$ tends to $l \in (0, \infty]$ as θ tends to $-\infty$. We distinguish two cases.

Either l is finite, then one can easily show that $\theta \mapsto \psi(\theta) + l\theta$ is increasing. As a result, $\psi' \geq -l$. If $l \neq 1/C_*$, then $\psi'(\theta) \not\rightarrow -l$ as θ tends to $-\infty$, so there exists $\varepsilon > 0$ such that $\psi' \geq \varepsilon - l$. Let $\theta < 0$. Then $\int_\theta^0 \psi' \geq (\varepsilon - l)(-\theta)$, i.e. $\psi(\theta) + l\theta \leq \psi(0) + \varepsilon\theta$. Dividing by $-\theta$ and taking the limit, we get $\varepsilon \leq 0$, which is absurd. We deduce that $l = 1/C_*$, which means that $-\theta/\ln L(\theta) \rightarrow C_*$ as θ tends to $-\infty$.

Or l is infinite. As for all $\theta < 0$, $\varphi(\theta) > 0$, the function $\theta \in (-\infty, 0) \mapsto \theta/\psi(\theta) - 1/\psi'(\theta)$ is increasing. It is also positive. If it does not tend to 0 at $-\infty$, then it is bounded from below by some $\varepsilon > 0$. Multiplying by $-\psi'(\theta) > 0$, we get : $1 \geq 1 - \theta\psi'(\theta)/\psi(\theta) \geq -\varepsilon\psi'(\theta)$.

Consequently $\psi'(\theta) \geq -1/\varepsilon$ and $\psi(0) - \psi(\theta) \geq \theta/\varepsilon$. Dividing by $\theta < 0$, we obtain : $-\psi(\theta)/\theta \leq 1/\varepsilon - \psi(0)/\theta$. Taking the limit, we have $\infty \leq 1/\varepsilon$, which is absurd. Finally, $\theta/\psi(\theta) - 1/\psi'(\theta)$ tends to 0 as θ tends to $-\infty$. Now, by definition of C_* , $\psi'(\theta)$ tends to $-1/C_*$. As a result, $-\theta/\psi(\theta)$ tends to C_* as θ tends to $-\infty$. \square

5.2.1 Lower bound

We show that $G_{n,j} \geq C_* \ln n + o(\ln n)$ a.s. Equation (9) will be very useful.

Proposition 8 *Suppose that $\theta_* = -\infty$. Let $j \geq 1$ be an integer. Then*

$$\liminf_{n \rightarrow \infty} \frac{G_{n,j}}{\ln n} \geq C_* \quad a.s.$$

Proof : We follow the usual strategy. Let $\theta < 0$ and $u > -\ln L(\theta)/\theta$. Define for all $k \in \mathbb{Z}_+$, $x_k := e^{ku}$. Applying (9), we can show that

$$\mathbb{E}[\mathcal{M}_{x_k,j}(k)|\mathcal{F}_k] \leq d(-\theta)j^{-\theta} x_k^\theta L(\theta)^k W^{(k)}(\theta),$$

so

$$\mathbb{E}[\mathcal{M}_{x_k,j}(k)] \leq d(-\theta)j^{-\theta} \exp\{k\theta(u + \ln L(\theta)/\theta)\}.$$

As $\theta(u + \ln L(\theta)/\theta) < 0$, $\mathbb{E}[\sum \mathcal{M}_{x_k,j}(k)] < \infty$, so a.s., for all integers k sufficiently large, $\mathcal{M}_{x_k,j}(k) = 0$. The usual calculations yield :

$$\liminf_{n \rightarrow \infty} \frac{G_{n,j}}{\ln n} \geq \frac{-\theta}{\ln L(\theta)} \quad a.s.$$

Lemma 9 enables us to complete the proof. \square

5.2.2 Upper bound

We finally prove that $G_{n,j} \leq C_* \ln n + o(\ln n)$ a.s. As $\theta_* = -\infty$, Lemma 3 cannot be applied (we do not know how the size of the smallest box behaves). Surprisingly, we are inspired by the proof of Proposition 5.

Proposition 9 *Suppose that $\theta_* = -\infty$. Let $j \geq 1$ be an integer. Then*

$$\limsup_{n \rightarrow \infty} \frac{G_{n,j}}{\ln n} \leq C_* \quad a.s.$$

Proof : We may suppose that $j = 1$. Let $\theta < 0$. Here, the boxes of the k -th generation that will play a key role are those having a size approximatively $\mathbf{s}(k)$, where

$$\mathbf{s}(k) := \exp\left(-k \frac{-L'(\theta)}{L(\theta)}\right).$$

Applying Lemma 2, we have with probability one :

$$\lim_{k \rightarrow \infty} \sqrt{k} e^{-k\varphi(\theta)} \#\{i \in \mathbb{N}^k : \mathbf{s}(k)/2 \leq p_i \leq \mathbf{s}(k)\} = Q(\theta),$$

where

$$Q(\theta) := \frac{1}{\sqrt{2\pi}} \frac{1 - 2^\theta}{-\theta} \left(\frac{L''(\theta)}{L(\theta)} - \left(\frac{L'(\theta)}{L(\theta)} \right)^2 \right)^{-1/2} W(\theta).$$

Notice that, by Lemma 1, $Q(\theta) > 0$ a.s. Define

$$\nu(k) := \lceil Q(\theta) k^{-1/2} e^{k\varphi(\theta)} / 2 \rceil.$$

Then (recall that $\varphi(\theta) > 0$), for almost all ω , there exists $k_0(\omega)$ such that

$$\# \{i \in \mathbb{N}^k : p_i \leq \mathbf{s}(k)\} \geq \nu(k) \geq 1, \quad \text{for all } k \geq k_0(\omega). \quad (13)$$

Let $u < -L'(\theta)/L(\theta)$. Define for all $k \in \mathbb{Z}_+$, $x_k := e^{ku}$. Let us prove that for almost all ω , there exists $k_1(\omega)$ such that

$$\mathcal{M}_{x_k,1}(k) \geq 1, \quad \text{for all } k \geq k_1(\omega). \quad (14)$$

Denote by \mathcal{G}_k the family of boxes having a size at most $\mathbf{s}(k)$. We define $\nu(k)$ new boxes : if $\#\mathcal{G}_k \geq \nu(k)$, we create $\nu(k)$ imaginary boxes $\mathbf{b}_1(k), \dots, \mathbf{b}_{\nu(k)}(k)$ of size $\mathbf{s}(k)$ containing the first $\nu(k)$ real boxes belonging to the family \mathcal{G}_k . If $\#\mathcal{G}_k < \nu(k)$, we denote by $\mathbf{b}_1(k), \dots, \mathbf{b}_{\nu(k)}(k)$ the first $\nu(k)$ boxes of the k -th generation. Introduce the event $A_k := \{\#\mathcal{G}_k \geq \nu(k)\}$ and

$$B_k := \{\forall l \leq \nu(k), \mathbf{b}_l(k) \text{ contains at least one ball when the first } \mathcal{P}_{x_k} \text{ have been thrown}\}.$$

Conditionally on \mathcal{F}_∞ and A_k , the boxes $\mathbf{b}_l(k)$ have size $\mathbf{s}(k)$ so the numbers of balls contained in $\mathbf{b}_l(k)$, $l \leq \nu(k)$, when \mathcal{P}_{x_k} balls have been thrown are independent Poisson variables with parameters $x_k \mathbf{s}(k)$. Therefore

$$\mathbb{P}(A_k \cap B_k | \mathcal{F}_\infty) \leq \mathbb{P}(\mathcal{P}_{x_k \mathbf{s}(k)} \geq 1)^{\nu(k)}.$$

Thus

$$\ln \mathbb{P}(A_k \cap B_k | \mathcal{F}_\infty) \leq \frac{Q(\theta)}{2} k^{-1/2} e^{k\varphi(\theta)} \ln \mathbb{P}(\mathcal{P}_{x_k \mathbf{s}(k)} \geq 1).$$

Now $\mathbb{P}(\mathcal{P}_{x_k \mathbf{s}(k)} \geq 1) \leq x_k \mathbf{s}(k)$, so

$$\mathbb{P}(A_k \cap B_k | \mathcal{F}_\infty) \leq \exp \left\{ \left(u + \frac{L'(\theta)}{L(\theta)} \right) \frac{Q(\theta)}{2} k^{1/2} e^{k\varphi(\theta)} \right\}.$$

As $u + L'(\theta)/L(\theta) < 0$, $\varphi(\theta) > 0$ and $Q(\theta) > 0$ a.s., we get $\mathbb{E}[\sum \mathbf{1}_{A_k \cap B_k} | \mathcal{F}_\infty] < \infty$ a.s., so $\sum \mathbf{1}_{A_k \cap B_k} < \infty$ a.s. Combined with (13), this proves that a.s., for all integers k sufficiently large, there exists an empty imaginary box when \mathcal{P}_{x_k} have been thrown. As every imaginary box contains a real box, (14) is proved.

Proceeding as usual, we deduce that for all $\theta < 0$,

$$\limsup_{n \rightarrow \infty} \frac{G_{n,1}}{\ln n} \leq \frac{L(\theta)}{-L'(\theta)} \quad \text{a.s.}$$

We get the result by letting θ tend to $-\infty$. □

Finally, Theorem 2 has been proved.

6 An explanation for the phase transition

In this section, we explain heuristically why a phase transition may occur in the asymptotics of $H_{n,j}$ but not in those of H_{n,n^α} , $G_{n,j}$ and G_{n,n^α} . To do so, we rephrase the setting in terms of interval-fragmentations. One can easily construct a family $(F(k), k \in \mathbb{Z}_+)$ of random open subsets of $(0, 1)$ with $F(0) = (0, 1)$, which is nested in the sense that $F(k') \subseteq F(k)$ for all $k \leq k'$, and such that the set of the lengths of the interval components of $F(k)$ is $\{p_i : |i| = k\}$ for every integer k . The boxes correspond to the interval components and the balls to a sequence $(U_i)_{i \in \mathbb{N}}$ of independent random variables uniformly distributed on $(0, 1)$ and independent of the fragmentation F . The height $H_{n,j}$ corresponds to the least integer k such that every interval component of $F(k)$ contains less than j elements of the set $\{U_1, \dots, U_n\}$, and the saturation level $G_{n,j}$ to the least integer k such that there exists an interval component of $F(k)$ containing less than j elements of the set $\{U_1, \dots, U_n\}$. Roughly speaking, the height $H_{n,j}$ depends crucially on the minimal length $\underline{m}_{n,j}$ of the intervals $[\hat{U}_i, \hat{U}_{i+j-1}]$ for $1 \leq i \leq n - j + 1$, where $0 < \hat{U}_1 < \dots < \hat{U}_n < 1$ are the ordered statistics of the family (U_1, \dots, U_n) , whereas the saturation level $G_{n,j}$ is related to the maximal length $\bar{m}_{n,j}$ of the intervals $[\hat{U}_i, \hat{U}_{i+j}]$ for $0 \leq i \leq n - j + 1$, where $\hat{U}_0 := 0$ and $\hat{U}_{n+1} := 1$. Indeed, the height $H_{n,j}$ is the first time k when no cluster $[\hat{U}_i, \hat{U}_{i+j-1}]$, $1 \leq i \leq n - j + 1$, of size j (in particular the smallest one) is included in an interval component of $F(k)$, and the saturation level $G_{n,j}$ is the first time k when there exists a cluster $[\hat{U}_i, \hat{U}_{i+j}]$, $0 \leq i \leq n - j + 1$, of size $j + 1$ (possibly the largest one) containing an interval component of $F(k)$.

It is easy to show from Lemma 3 that if we used equidistributed points $\{j/(n+1) : 1 \leq j \leq n\}$ instead of i.i.d. uniform points, then no phase transition would occur. More precisely, all the heights $H_{n,j}$ would be equivalent to $C^* \ln n$ a.s. Further, the heights H_{n,n^α} would be equivalent to $(1 - \alpha)C^* \ln n$ a.s. and the saturation levels G_{n,jn^α} to $(1 - \alpha)C_* \ln n$ a.s.

We first explain why the clusters of size n^α behave as if the points U_i , $1 \leq i \leq n$, were equidistributed on $(0, 1)$. It will follow that no phase transition occurs in the asymptotics of H_{n,n^α} and G_{n,n^α} . To do so, let us first prove that $\underline{m}_{n,n^\alpha} \geq n^{-1+\alpha+o(1)}$ a.s. Let $\varepsilon > 0$. We have :

$$\mathbb{P}(\underline{m}_{n,n^\alpha} \leq n^{-1+\alpha-2\varepsilon}) \leq n\mathbb{P}(\hat{U}_{n^\alpha} - \hat{U}_1 \leq n^{-1+\alpha-2\varepsilon}).$$

Let $(e_i)_{i \in \mathbb{N}}$ be a sequence of independent exponential variables with parameters 1. Define $\gamma_k := \sum_{1 \leq i \leq k} e_i$, for all $k \geq 1$. Then we have :

$$(\hat{U}_1, \dots, \hat{U}_n) \stackrel{(d)}{=} (\gamma_1/\gamma_{n+1}, \dots, \gamma_n/\gamma_{n+1}), \quad (15)$$

so

$$\begin{aligned} \mathbb{P}(\hat{U}_{n^\alpha} - \hat{U}_1 \leq n^{-1+\alpha-2\varepsilon}) &= \mathbb{P}\left(\frac{\gamma_{n^\alpha} - \gamma_1}{\gamma_{n+1}} \leq n^{-1+\alpha-2\varepsilon}\right) \\ &\leq \mathbb{P}(\gamma_{n^\alpha} - \gamma_1 < n^{1+\varepsilon}n^{-1+\alpha-2\varepsilon}) + \mathbb{P}(\gamma_{n+1} \geq n^{1+\varepsilon}) \\ &= \mathbb{P}(\mathcal{P}_{n^\alpha-\varepsilon} \geq n^\alpha) + \mathbb{P}(\mathcal{P}_{n^{1+\varepsilon}} \leq n+1) \end{aligned}$$

since the sequence $(\gamma_i)_{i \in \mathbb{N}}$ has the same law as the ordered sequence of points of a standard Poisson process. Applying Lemma 5 with $p = 3/\varepsilon$, we have for every integer n sufficiently large :

$$\mathbb{P}(\mathcal{P}_{n^\alpha-\varepsilon} \geq n^\alpha) + \mathbb{P}(\mathcal{P}_{n^{1+\varepsilon}} \leq n+1) \leq (c(p) + 2^p d(p)) n^{-3}.$$

We deduce that that for every integer n sufficiently large, we have :

$$\mathbb{P}(\underline{m}_{n,n^\alpha} \leq n^{-1+\alpha-2\varepsilon}) \leq (c(p) + 2^p d(p)) n^{-2}.$$

Applying the Borel-Cantelli lemma and letting ε tend to 0, we easily conclude that $\underline{m}_{n,n^\alpha} \geq n^{-1+\alpha+o(1)}$ a.s. Similar calculations yield : $\overline{m}_{n,n^\alpha} \leq n^{-1+\alpha+o(1)}$ a.s. We conclude that the clusters of size n^α essentially behave as if the points U_i , $1 \leq i \leq n$, were equidistributed on $(0, 1)$, which explains why no phase transition occur in the asymptotics of H_{n,n^α} and G_{n,n^α} .

On the contrary, the clusters of size j behave very differently from the equidistribution case. For instance, thanks to the identification (15), we may apply well-known results on extreme values (see *e.g.* [26]). One finds that $\underline{m}_{n,j} \approx n^{-j/(j-1)}$, so that $\underline{m}_{n,j}$ is much smaller than j/n , even at a logarithmic scale. Furthermore, the larger the integer j is, the closer these two quantities are, so that $H_{n,j}$ may eventually be equivalent to $C^* \ln n$ a.s. if j is sufficiently large. On the other hand, if the integer j is too small, then $H_{n,j}$ may be greater than $(1+\varepsilon)C^* \ln n + o(\ln n)$ a.s. for some $\varepsilon > 0$: one has to wait for a long time before the smallest clusters of size j are no longer included in any interval component of the fragmentation; in that case, there is a phase transition.

Concerning the saturation level $G_{n,j}$, one gets $\overline{m}_{n,j} \approx \ln n/n$, which differs from the equidistribution case only up to a logarithmic factor, and the impact of the latter is asymptotically negligible. It explains why no phase transition occurs in the asymptotics of $G_{n,j}$.

Acknowledgements. I would like to express my gratitude to my thesis director Jean Bertoin who suggested this subject to me. I sincerely thank him for his invaluable advice and for his careful reading.

References

- [1] M. Abramowitz and I. A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 1964.
- [2] L. Addario-Berry and B.A. Reed. Minima in branching random walks. To appear in *Ann. Probab.*
- [3] J. Bertoin. The asymptotic behavior of fragmentation processes. *J. Euro. Math. Soc.*, 5:395–416, 2003.
- [4] J. Bertoin. *Random Fragmentation and Coagulation Processes*, volume 102. Cambridge Studies in Advanced Mathematics, 2006.
- [5] J. Bertoin. Asymptotic regimes for the occupancy scheme of multiplicative cascades. *Stochastic Process. Appl.*, 118(9):1586–1605, 2008.
- [6] J. Bertoin and A. Rouault. Discretization methods for homogeneous fragmentations. *J. London Math. Soc.*, 72:91–109, 2005.
- [7] J. D. Biggins. Uniform convergence of martingales in the branching random walk. *Ann. Probab.*, 20(1):137–151, 1992.

- [8] S. Boucheron and D. Gardy. An urn model from learning theory. *Random Structures Algorithms*, 10(1-2):43–67, 1997.
- [9] J. Bunge and M. Fitzpatrick. Estimating the number of species : a review. *J. Am. Stat. Assoc.*, 88:364–373, 1993.
- [10] B. Chauvin, T. Klein, J.-F. Marckert, and A. Rouault. Martingales and profile of binary search trees. *Electron. J. Probab.*, 10(12):420–435, 2005.
- [11] L. Devroye. A note on the height of binary search trees. *J. Assoc. Comput. Mach.*, 33(3):489–498, 1986.
- [12] L. Devroye. Branching processes in the analysis of the heights of trees. *Acta Inform.*, 24(3):277–298, 1987.
- [13] L. Devroye. Branching processes and their applications in the analysis of tree structures and tree algorithms. *Probabilistic methods for algorithmic discrete mathematics*, Springer:249–314, 1998.
- [14] L. Devroye. Universal limit laws for depths in random trees. *SIAM J. Comput.*, 28(2):409–432, 1999.
- [15] M. Drmota. Stochastic analysis of tree-like data structures. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 460(2041):271–307, 2004.
- [16] D. Gardy. Occupancy urn models in the analysis of algorithms. *J. Statist. Plann. Inference*, 101(1-2):95–105, 2002.
- [17] A. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probab. Surv.*, 4:146–171, 2007.
- [18] B. Haas, G. Miermont, J. Pitman, and M. Winkel. Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models. *Ann. Probab.*, 36(5):1790–1837, 2008.
- [19] F. He and K. J. Gaston. Estimating species abundance from occurrence. *The American Naturalist*, 156(5):553–559, 2000.
- [20] L. Holst. On birthday, collectors’, occupancy and other classical urn problems. *Internat. Statist. Rev.*, 54(1):15–27, 1986.
- [21] Y. Hu and Z. Shi. Minimal position and critical martingale convergence in branching random walks, and directed polymers on disordered trees. To appear in *Ann. Probab.*
- [22] N. L. Johnson and S. Kotz. *Urn Models and Their Application. An Approach to Modern Discrete Probability Theory*. John Wiley & Sons, New York-London-Sydney, 1977.
- [23] V. F. Kolchin, B. A. Sevast’yanov, and V. P. Chistyakov. *Random Allocations*. John Wiley & Sons, New York-London-Sydney, 1978.

- [24] McDiarmid. Minimal positions in a branching random walk. *Ann. Appl. Probab.*, 5:128–139, 1995.
- [25] B. Pittel. On growing random binary trees. *J. Math. Anal. Appl.*, 103(2):461–480, 1984.
- [26] S. Resnick. *Extreme Values, Regular Variation, and Point Processes*. Springer Series in Operations Research and Financial Engineering, 1987.
- [27] M. Robson. The height of binary search trees. *Austral. Comput. J.*, 11:151–153, 1979.