

# A Phase Transition for the Heights of a Fragmentation Tree

**Adrien Joseph**

*Laboratoire de Probabilités et Modèles Aléatoires, Université Pierre et Marie Curie,  
4 place Jussieu, 75005 Paris, France; e-mail: adrien.joseph@upmc.fr*

*Received 21 May 2009; accepted 8 January 2010; received in final form 1 April 2010*

*Published online 29 September 2010 in Wiley Online Library (wileyonlinelibrary.com).*

*DOI 10.1002/rsa.20340*

**ABSTRACT:** We provide information about the asymptotic regimes for a homogeneous fragmentation of a finite set. We establish a phase transition for the asymptotic behavior 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. © 2010 Wiley Periodicals, Inc. *Random Struct. Alg.*, 39, 247–274, 2011

*Keywords:* occupancy scheme; multiplicative cascade; random split tree; homogeneous fragmentation; branching random walk

## 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 [7, 14], analysis of algorithms [11], learning theory [6], etc. The books by Johnson and Kotz [17] and by Kolchin et al. [18] 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  $i$  ( $p_i$  should be seen as the size of that box). We may be interested in the number of boxes containing exactly  $j$  balls, or in the number of occupied boxes, etc.

---

Correspondence to: A. Joseph

© 2010 Wiley Periodicals, Inc.

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 all integers  $k \in \mathbb{Z}_+$ , 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 that 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 of size  $p_i$  (provided that  $p_i \neq 0$ ), such that the box  $ij$  is contained into the box  $i$  (if  $p_{ij} \neq 0$ ). 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 the first generation at which all the boxes have less than  $j \geq 2$  balls when  $n$  have been thrown by  $H_{n,j}$  and the first level at which there exists a box containing less than  $j \geq 1$  balls by  $G_{n,j}$ .  $H_{n,j}$  is called a *height* and  $G_{n,j}$  is called a *saturation level*. Our aim is to study the asymptotic behavior 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  $\rho$ . The random 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 split to its children. This model is called the occupancy scheme of multiplicative cascades. We put emphasis on the fact that the numbers  $p_i$  are random. Once we are given a realization of the numbers  $p_i(\omega), i \in \mathcal{U}$ , the balls are thrown in the boxes according to that realization. There are thus two levels of randomness in our model, namely the arrangement of boxes and the way one throws balls.

In the particular case when  $\rho$  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 [10]. 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 [3] for background and Section 3.2 in [4] for a description which is closely related to the present work. In a different direction, we mention the work of Haas et al. [13] who associate another random tree to homogeneous fragmentation processes.

We shall show the following result: there exist an integer  $\tau \in \{2, 3, \dots\} \cup \{\infty\}$  and a sequence of positive real numbers  $C_2 > \dots > C_\tau = C_{\tau+1} = \dots =: C^*$  such that for all integers  $j \geq 2$ ,

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

It should not be surprising that the heights  $H_{n,j}$  have logarithmic asymptotics, as a large class of trees have such a behavior, like random split trees. On the other hand, it is remarkable that there exists a critical parameter  $\tau \in \{2, 3, \dots\} \cup \{\infty\}$  at which a phase transition occurs (provided that  $2 < \tau < \infty$ ). It is a peculiar behavior for a random split tree. Devroye indeed proved in [10] that for a large class of random split trees such that their internal nodes contain at least one ball<sup>1</sup>, then all the heights, regardless of the value of the rupture threshold  $j$ , have the same asymptotic behavior (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

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

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

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 all positive integers  $j$ ,

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

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 behavior of  $G_{n,n^\alpha}$ .

Our approach essentially relies on the theory of branching random walks, and more precisely on their large deviations behaviors whose descriptions are due to Biggins [5]. 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  $\tau$  and the real numbers  $C_2, \dots, C_\tau$  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 all integers  $k$ , 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 lower bound of  $H_{n,j}$ , depending on whether the integer  $j$  is less than the critical parameter  $\tau$

---

<sup>1</sup>We stress that this assumption is crucial in the proof of Theorem 1 in [10] and seems to have been overlooked in the statement.

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  $H_{n,n^\alpha}$ ,  $G_{n,j}$  and  $G_{n,n^\alpha}$ .

## 2. FORMULATION OF THE MAIN RESULTS

Recall that  $\rho$  is a random probability measure on  $\mathbb{N}$ . We denote its law by  $\nu$ . If we denote by  $\text{Prob}_{\mathbb{N}}$  the space of probability measures on  $\mathbb{N}$ ,  $\nu$  is a probability measure on  $\text{Prob}_{\mathbb{N}}$ , called the splitting law. We exclude the degenerate case when  $\rho$  is a Dirac point mass a.s.

As explained in the introduction, we consider a family  $(\rho(i), i \in \mathcal{U})$  of independent copies of  $\rho$  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  $\mathbb{R}_+$

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

where  $\delta_z$  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 all 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  $\nu$ . 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

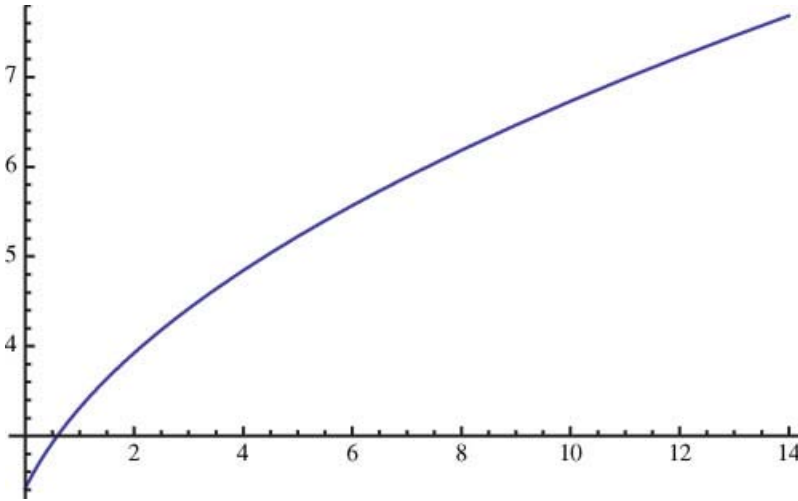
$$L(\theta) = \mathbb{E} \left[ \sum_{j \in \mathbb{N}} \rho_j^\theta \right] = \int_{\text{Prob}_{\mathbb{N}}} \left( \sum_{j \in \mathbb{N}} p_j^\theta \right) \nu(\text{d}p)$$

with the convention that  $p^\theta = 0$  when  $p = 0$  even when  $\theta \leq 0$ . Because  $\rho$  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

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

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

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



**Fig. 1.** graph of the function  $a \mapsto \theta^*(a, a)$ . [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

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  $(\theta_*, \theta^*)$ :

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

Note that  $\theta^* > 1$  and  $\theta_* < 0$  if  $\underline{\theta} < 0$ . Let us consider the example of beta-splitting laws.

**Example.** Let us describe the situation when  $\rho = (B, 1 - B, 0, \dots)$ , where  $B$  has a Beta( $a, b$ ) distribution. Easy calculations yield:

$$L(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)} \frac{\Gamma(a+\theta)}{\Gamma(a+b+\theta)} + \frac{\Gamma(a+b)}{\Gamma(b)} \frac{\Gamma(b+\theta)}{\Gamma(a+b+\theta)}.$$

Let  $\theta^*(a, b)$  denote the parameter  $\theta^*$ . Because this critical parameter is only defined implicitly as the solution of an equation involving the derivative of the gamma function, there is no simple explicit expression in terms of the parameter  $(a, b)$ . However numerical estimates can easily be obtained, and are represented in Figure 1 above in the special case  $a = b$ .

The convexity of  $\ln L$  implies that  $\frac{L}{-L'}$  is an increasing function. We define

$$C_* := \lim_{\substack{\theta \rightarrow \theta_* \\ \theta > \theta_*}} \frac{L(\theta)}{-L'(\theta)} \quad \text{and} \quad C^* := \lim_{\substack{\theta \rightarrow \theta^* \\ \theta < \theta^*}} \frac{L(\theta)}{-L'(\theta)}.$$

Inspired by the article of Hu and Shi in [16] (see Lemma 4 below), we shall sometimes need the following assumption: there exists  $\delta > 0$  such that

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

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

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

We can now state the results that we shall prove. Concerning the asymptotic behavior of the heights  $H_{n,j}$ , we have the following results which complete and improve Proposition 2 in [4].

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

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

Then

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

Furthermore, if  $j < \theta^*$  or if (1) holds, then

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

In fact, the following result always holds:

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

We see that there is a phase transition in the asymptotic behavior of  $H_{n,j}$  at the integer  $\lceil \theta^* \rceil$  when  $2 < \theta^* < \infty$  ( $\lceil \theta^* \rceil$  is the cut-off parameter denoted by  $\tau$  in the introduction). 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 all  $\alpha \in [\frac{1}{2}, 1)$ , if  $\rho$  is equal to  $(\frac{1}{2}, \frac{1}{2}, 0, \dots)$  with probability  $\alpha$  and to  $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, \dots)$  with probability  $1 - \alpha$ , then  $\theta^* = \infty$ .

**Remark 1.** *The critical parameter  $\theta^*$  is finite whenever*

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

where  $\rho = (\rho_j, j \in \mathbb{N})$  denotes a random probability measure on  $\mathbb{N}$  with law  $\nu$  (recall that  $\| \cdot \|_\infty$  is the essential supremum of the r.v.  $\max_{j \in \mathbb{N}} \rho_j$ ). Indeed, it is easily seen that

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

If the right-hand side equals 1, then  $g : \theta \mapsto \frac{-\ln L(\theta)}{\theta}$  has limit 0 at infinity. Now  $g$  is differentiable 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  $\theta_0 = \theta^* < \infty$ .

**Remark 2.** *The parameter  $\theta^*$  is infinite whenever  $\rho$  is not random, i.e.,  $\rho = (\rho_1, \rho_2, \dots)$  where  $\rho_j, j \in \mathbb{N}$ , are deterministic real numbers (note that we excluded the case when only*

one  $\rho_j$  is nonzero). Indeed, in that setting, as for all  $\theta > 1$ ,  $L(\theta) = \sum_{j=1}^{\infty} \rho_j^\theta$ , one can show that

$$\varphi(\theta) = \ln \left( \sum_{j=1}^{\infty} \rho_j^\theta \right) + \theta \frac{\sum_{j=1}^{\infty} -\ln \rho_j \rho_j^\theta}{\sum_{j=1}^{\infty} \rho_j^\theta}.$$

Let  $f$  denote the function  $x \mapsto -x \ln x$ . It is easily seen that  $f(\sum_{j=1}^{\infty} \rho_j^\theta) < \sum_{j=1}^{\infty} f(\rho_j^\theta)$ , which proves that  $\varphi(\theta) > 0$  for all  $\theta > 1$ . As a result,  $\theta^* = \infty$ .

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

**Proposition 1.** *Let  $\alpha \in (0, 1)$ . Then*

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

Moreover, if  $\theta^* < \infty$  and (1) holds, then

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

We shall see that if  $\theta^* < \infty$  then

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

Concerning the asymptotic behavior of the saturation levels, we shall prove the following theorem.

**Theorem 2.** *Let  $j \geq 1$  be an integer and  $\alpha \in [0, 1)$ . Suppose that  $\theta_* = -\infty$  or (2) holds. Then*

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

Furthermore, if (2) and (1) hold, then

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

In fact, if (2) holds, then

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

**Remark 3.** *A sufficient condition to guarantee that  $-\infty < \theta_* < 0$  and  $\varphi(\theta_*) = 0$  is:*

$$-\infty < \underline{\theta} < 0 \quad \text{and} \quad \lim_{\theta \rightarrow \underline{\theta}} L(\theta) = \infty.$$

Indeed, suppose this condition fulfilled. Imagine that for all  $\theta < 0$ ,  $\varphi(\theta) > 0$ . Let  $\theta_0 \in (\underline{\theta}, 0)$ . As  $\theta \mapsto \frac{\ln L(\theta)}{\theta}$  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  $\lim_{\theta \rightarrow \underline{\theta}} L(\theta) = \infty$ .

**Remark 4.** *The final assertions in Theorem 1, Proposition 1 and Theorem 2 rely on the work of Hu and Shi [16]. McDiarmid’s setting in [19] can however be considered; we can prove that the results stated in Theorem 1, Proposition 1 and Theorem 2 still hold if the assumption (1) is replaced by the following:*

$$\int_{\text{Prob}\mathbb{N}} \left( \sum_{i \in \mathbb{N}} \mathbf{1}_{p_i > 0} \right)^2 \nu(d\mathbf{p}) < \infty. \tag{3}$$

*For instance, suppose that  $\rho = (\rho_1, 1 - \rho_1, 0, 0, \dots)$ , where  $\rho_1$  is a random variable with density  $\mathbf{1}_{0 < x < e^{-1}} x^{-1} \ln^{-2} x dx$ . Then  $\theta = 0$ , so (1) does not hold. Nonetheless,  $\rho$  is supported by two integers a.s., so (3) holds. As a result, for all  $j \geq 2$ ,  $H_{n,j} = C_j \ln n + O(\ln \ln n)$  a.s. Furthermore, as  $\|1 - \rho_1\|_\infty = 1$ , Remark 1 ensures that  $\theta^*$  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, Proposition 1 and Theorem 2 only in the setting of [16]. The general proof can however be easily carried out.*

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  $L(\theta) = \frac{2}{\theta+1}$ , so  $\theta = -1$  and (1) holds. According to Remark 1 and Remark 3,  $\theta^* < \infty$  and (2) hold. Hence, all our results may be applied. Easy calculations yield  $\varphi(\theta) = \ln 2 - \ln(\theta + 1) + \frac{\theta}{\theta+1}$  and  $[\theta^*] = 4$ , so that there is a phase transition. We can show that  $C_2 = \frac{2}{\ln 3 - \ln 2} \approx 4.93260 \dots$ ,  $C_3 = \frac{3}{\ln 2} \approx 4.32808 \dots$  and that  $C^* > C_*$  are the solutions of the equation

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

Note that  $C^* \approx 4.31107 \dots$  and  $C_* \approx 0.37336 \dots$

As  $\rho$  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. Let  $T_{n,j}$  denote 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 in such a way: each ball choses the left child of  $i$  with probability  $U_i$ , the right child of  $i$  with probability  $1 - U_i$ , where the r.v.  $U_i$ ,  $i \in \mathcal{U}$ , are independent and uniformly distributed on  $[0, 1]$ . 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  $i$  according to the probability vector  $(U_i, 1 - U_i)$ , and one has to consider whether the ball stays at that new box or not. Let  $\tilde{T}_{n,j}$  denote 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 [22] was the first to be interested in the height of the random binary search tree. In [20], Pittel studied its height and its saturation level. Devroye proved in [8, 9] 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 [10] 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 [5]. For all  $\theta > \underline{\theta}$ , 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 all integers  $k \in \mathbb{Z}_+ \cup \{\infty\}$ , we denote by  $\mathcal{F}_k$  the  $\sigma$ -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.

In some proofs, we shall need a precise estimate of the number of boxes at generation  $k$  with size of order  $\exp(k \frac{L'(\theta)}{L(\theta)})$ . If the splitting law  $\nu$  is not geometric, 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  $\rho$  belong to  $\{r^n, n \in \mathbb{Z}_+\}$ , we may apply Corollary 4 in [5]:

**Lemma 2.** *Suppose that  $\nu$  is not geometric. Then for all real numbers  $a$  and  $b$  such that  $a < b$  and for all  $\theta \in (\theta_*, \theta^*)$ , we have with probability one that*

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

Note that Lemma 2 has a version in the geometric case (technically, replace Stone’s local limit theorem by Gnedenko’s local limit theorem). In our proofs, we shall always work with a non-geometric splitting law.

We next turn our interest to the asymptotic behavior 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*

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

*Likewise, if (2) holds, then*

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

To have precise estimates of the behavior 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 [16], will be very useful.

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

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

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

- *If (2) holds, then*

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

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 [12] or Holst in [15]) 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  $\mathcal{F}_\infty$ . For all individuals  $i \in \mathcal{U}$  and for all  $x \in (0, \infty)$ , we denote the number of balls at  $i$  when the first  $\mathcal{P}_x$  balls have been thrown by  $\mathcal{C}(i; x)$ . For all  $x, y \in (0, \infty)$  and  $k \in \mathbb{Z}_+$ , we denote the number of boxes at generation  $k$  containing at least  $y$  balls when the first  $\mathcal{P}_x$  balls have been thrown by  $\mathcal{N}_{x,y}(k)$ :

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

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, let  $\mathcal{M}_{x,y}(k)$  denote 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 let  $M_{n,j}$  denote 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*

$$\sup_{\substack{j \geq p \\ j \in \mathbb{N}}} \sup_{x > 0} \mathbb{P}(\mathcal{P}_x \geq j) j^p x^{-p} \leq c(p), \tag{4}$$

and

$$\sup_{j \in \mathbb{N}} \sup_{x > 0} \mathbb{P}(\mathcal{P}_x < j) j^{-p} x^p \leq d(p). \tag{5}$$

*Proof.* We begin by showing (4). 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  $\frac{\Gamma(k-p+1)k^p}{k!} \rightarrow 1$  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 (4).

To show (5), 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  $\frac{\Gamma(k+p+1)}{(k+1)^p k!} \rightarrow 1$  as  $k$  tends to infinity, there exists a finite constant  $d(p) \geq 1$  such that for all  $k \in \mathbb{Z}_+$ ,  $\frac{\Gamma(k+p+1)}{(k+1)^p k!} \leq d(p)$ . 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. ■

**Remark 5.** We stress that in (4), 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 2 below, and more precisely Lemma 6). On the other hand, there is no restriction in (5); no phase transition appears in the asymptotics of  $G_{n,j}$  (see Proposition 6 below).

#### 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 6 below). It is however proved in the paragraph dealing with the lower bound.

##### 4.1. Upper Bound

Equation (4) 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 upper bounds stated in Theorem 1 and in Proposition 1.

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

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

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

**Remark 6.** Suppose that  $\alpha = 0$ . We deduce from Proposition 2 that:

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

where  $h$  is the function  $\theta \mapsto \frac{\theta}{-\ln L(\theta)}$ . Now,  $h$  is differentiable 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

- $\frac{j}{-\ln L(j)}$  if  $j < \theta^*$ ,
- $\frac{\theta^*}{-\ln L(\theta^*)} = C^*$  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  $jn^\alpha$  balls when  $n$  balls have been thrown. Because we want to show that  $H_{n,jn^\alpha}$  is bounded from above by  $((1 - \alpha)\frac{\theta}{-\ln L(\theta)} \ln(n) + r \ln \ln n)$  for some  $r$  sufficiently large, we take  $k = ((1 - \alpha)\frac{\theta}{-\ln L(\theta)} \ln(n) + r \ln \ln n)$ . In other words, one initially throws  $n = \exp(k\frac{1}{1-\alpha}\frac{-\ln L(\theta)}{\theta})k^{-s}$  balls and we show that every box of the  $k$ -th generation contains less than  $jn^\alpha$  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 all  $k \geq 1$

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

Informally,  $x_k$  corresponds to the number of balls thrown when we consider boxes at generation  $k$ . 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  $\omega$ , there exists  $k_0(\omega)$  such that

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

*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  $\mathcal{F}_k$ ,  $(\mathcal{C}(i;x))_{i:|i|=k}$  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  $\lceil y \rceil \geq \theta$ , (4) ensures that

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

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:

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

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

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

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  $k, j\phi_k^\alpha \geq \theta$ ):

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

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  $\mathcal{N}_{x_k, j\phi_k^\alpha}(k) \geq 1$ . ■

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

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

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

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  $\omega$ , 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  $\Omega_0$  by

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

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  $\phi_{k_5}$  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. ■

**Remark 7.** Note that to deduce from Proposition 2 that  $H_{n,n^\alpha} \leq (1 - \alpha)C^*$  a.s. in the case  $\theta^* = \infty$ , one has to show that  $\lim_{\theta \rightarrow \infty} \frac{\theta}{-\ln L(\theta)} = C^*$ . Let us now prove that result.

Let  $\psi = \ln L$ . The condition  $\theta^* = \infty$  implies that  $C^* = \lim_{\theta \rightarrow \infty} -\frac{1}{\psi'(\theta)}$ . As  $\psi$  is convex, it is known that  $\frac{\psi(\theta)}{\theta}$  tends to  $l \in ]-\infty, \infty]$  as  $\theta$  tends to  $\infty$ . Let  $f$  be the function  $\theta \mapsto \psi(\theta) - l\theta$ .  $f$  is convex and  $\frac{f(\theta)}{\theta}$  tends to 0 as  $\theta$  tends to  $\infty$ . One can easily deduce that  $f$  is non-increasing. As a result,  $\psi' \leq l$ . If  $l \neq -\frac{1}{C^*}$ , then  $\psi'(\theta) \rightarrow l$  as  $\theta$  tends to  $\infty$ . Now,  $\psi'$  is non-decreasing. So there exists  $\varepsilon > 0$  such that  $\psi' \leq l - \varepsilon$ . Let  $\theta > 1$ . Then  $\int_1^\theta \psi' \leq (l - \varepsilon)(\theta - 1)$ , i.e.,  $\psi(\theta) - l\theta \leq -\varepsilon\theta + \varepsilon - l$ . Dividing by  $\theta$  and taking the limit, we get  $\varepsilon \leq 0$ , which is absurd. We deduce that  $l = -\frac{1}{C^*}$ , which means that  $\frac{\theta}{-\ln L(\theta)} \rightarrow C^*$  as  $\theta$  tends to  $\infty$ .

### 4.2. Lower Bound

We now study the lower bound. Combined with the upper bound supplied by Proposition 2, 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 3.** 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  $jn^\alpha$  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  $H_{n,jn^\alpha}$  is bounded from below by  $((1 - \alpha)C^*) \ln n - r \ln \ln n$ , we take  $k = ((1 - \alpha)C^*) \ln n - r \ln \ln n$ , i.e., one initially throws  $n = \exp(k \frac{1}{(1-\alpha)C^*}) k^s$  balls and we show that the largest box of the  $k$ -th generation contains at least  $jn^\alpha$  balls.

Let  $\gamma, \gamma'$  and  $\gamma''$  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  $\omega$ , 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  $\mathfrak{b}(k)$  contained in the largest box  $\bar{\mathfrak{b}}(k)$  of size  $\bar{p}(k)$  such that a ball fallen in  $\bar{\mathfrak{b}}(k)$  is thrown in  $\mathfrak{b}(k)$  with probability

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

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

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

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

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

Applying (5), we get:

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

where  $p := \frac{2}{\gamma' - \gamma - \alpha\gamma''} > 0$ , 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,  $\mathfrak{p}(k) \geq \bar{p}(k)$  or the box  $\mathfrak{b}(k)$ , which is contained in the box  $\bar{\mathfrak{b}}(k)$ , 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,  $\mathfrak{p}(k) < \bar{p}(k)$ . Lemma 7 is therefore proved. ■

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

*Proof of Proposition 3.* The same calculations performed at the beginning of the proof of Proposition 2 show that for almost all  $\omega$ , 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  $\Omega_0$  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  $(\phi_k)$  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  $N_{n, j n^\alpha}(k) \geq 1$  and  $H_{n, j n^\alpha} > k$ . Further, as  $n \leq \phi_k$ ,

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



As  $\exp((k - 1)\frac{1}{(1-\alpha)C^*}) \leq \phi_{k-1} \leq n$ , 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  $\gamma''$  can be chosen arbitrarily close to  $\frac{3}{2(1-\alpha)\theta^*}$  completes the proof. ■

**Remark 8.** *In the case  $\theta^* < \infty$ , if the assumption (1) does not hold, one can still show that for all integers  $j \geq 1$  and for all real numbers  $\alpha \in [0, 1)$  such that  $(j, \alpha) \neq (1, 0)$ , with probability one,  $H_{n,jn^\alpha} \geq (1 - \alpha)C^* \ln n + o(\ln n)$ . To do so, recall that Lemma 3 ensures that  $\frac{-\ln \tilde{p}(k)}{k}$  tends to  $\frac{1}{C^*}$  a.s., and use an argument similar to that in Proposition 3.*

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

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

The main difference with the case  $j \geq \theta^*$  is that we cannot consider the largest box any more. 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  $\frac{j}{-\ln L(j)} \ln n - r \ln \ln n$ , we consider the situation at the  $k$ -th generation when one initially throws  $\exp(k \frac{-\ln L(j)}{j})k^s$  balls.

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

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

where

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

and

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

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

$$v(k) := \left\lceil \frac{Q(j)k^{-\frac{1}{2}} e^{k\varphi(j)}}{2} \right\rceil.$$

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

$$\#\{i \in \mathbb{N}^k : p_i \geq \mathfrak{s}(k)\} \geq \nu(k) \geq 1, \quad \text{for all } k \geq k_0(\omega). \tag{6}$$

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  $\omega$ , 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 decreasing order. Denote the family of boxes having a size at least  $\mathfrak{s}(k)$  by  $\mathcal{G}_k$ . 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  $\mathfrak{b}_l(k)$  inside the box  $b_l(k)$  such that a ball fallen in  $b_l(k)$  is thrown in  $\mathfrak{b}_l(k)$  with probability  $\frac{\mathfrak{s}(k)}{s_l(k)}$ , where  $s_l(k)$  is the size of  $b_l(k)$ . In particular, every imaginary box has size  $\mathfrak{s}(k)$ . If  $\#\mathcal{G}_k < \nu(k)$ , we denote the first  $\nu(k)$  boxes by  $\mathfrak{b}_1(k), \dots, \mathfrak{b}_{\nu(k)}(k)$ . Introduce the events  $A_k := \{\#\mathcal{G}_k \geq \nu(k)\}$  and

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

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

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

Thus

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

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

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

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

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

Finally, we get

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

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 (6), 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. ■

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

*Proof of Proposition 4.* Let  $u$  and  $u'$  be two real numbers such that  $\frac{1}{2j} < u < u'$ . Define  $x_k := k^u \exp(k \frac{-\ln L(j)}{j})$  and  $\phi_k := k^{u'} \exp(k \frac{-\ln L(j)}{j})$ . One can show that for almost all  $\omega$ , 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  $\Omega_0$  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  $(\phi_k)$  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  $n \geq \phi_{k-1} \geq \exp((k-1) \frac{-\ln L(j)}{j})$ , 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. ■

**4.2.3. The Case  $\theta^* = \infty$  and  $\alpha \in (0, 1)$ .** In this section, we prove that  $H_{n,n^\alpha} \geq (1 - \alpha)C^* \ln n + o(\ln n)$  a.s. if  $\theta^* = \infty$  and  $\alpha \in (0, 1)$ . We shall see that, like the case  $\theta^* < \infty$ , the largest box plays a key role.

**Proposition 5.** *We suppose that  $\theta^* = \infty$ . Let  $\alpha \in (0, 1)$ . With probability one, we have:*

$$\liminf_{n \rightarrow \infty} \frac{H_{n,n^\alpha}}{\ln n} \geq (1 - \alpha)C^*.$$

We are inspired by the proofs of Proposition 3 and Proposition 4.

*Proof.* Let  $\theta$ ,  $\varepsilon$  and  $\varepsilon'$  be three real numbers such that

$$\theta > 1 \quad \text{and} \quad 0 < \varepsilon < \varepsilon' < \frac{\varepsilon}{\alpha}.$$

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

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

We first show that a.s.,  $\mathcal{N}_{x_k, \phi_k^\alpha}(k) \geq 1$  for all integers  $k$  sufficiently large. To do so, we consider the largest box. Even though we cannot apply Lemma 3, Lemma 2 provide information about the size  $\bar{p}(k)$  of the largest box  $\bar{b}(k)$ .

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

$$\mathcal{N}_{x_k, \phi_k^\alpha}(k) \geq 1, \quad \text{for all } k \geq k_0(\omega). \tag{7}$$

We denote by  $A_k$  the event  $\{\bar{p}(k) \geq \exp(k \frac{L'(\theta)}{L(\theta)})\}$  and by  $B_k$  the event  $\{\text{the box } \bar{b}(k) \text{ contains less than } \phi_k^\alpha \text{ balls when the first } \mathcal{P}_{x_k} \text{ have been thrown}\}$ . Then:

$$\mathbb{P}(A_k \cap B_k) \leq \mathbb{P}(B_k|A_k) \leq \mathbb{P}(\mathcal{P}_{x_k \exp(k \frac{L'(\theta)}{L(\theta)})} < \phi_k^\alpha).$$

Applying (5), we get:

$$\mathbb{P}(A_k \cap B_k) \leq d(1)e^{-k(\varepsilon - \alpha\varepsilon')}.$$

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,  $\bar{p}(k) < \exp(k \frac{L'(\theta)}{L(\theta)})$  or the box  $\bar{b}(k)$  has at least  $\phi_k^\alpha$  balls when the first  $\mathcal{P}_{x_k}$  balls have been thrown. Now, by Lemma 2 (recall that  $\theta \in (\theta_*, \theta^*)$  since  $\theta^* = \infty$ ), we know that a.s., for all integers  $k$  sufficiently large,  $\bar{p}(k) \geq \exp(k \frac{L'(\theta)}{L(\theta)})$ . (7) is therefore proved.

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

$$\liminf_{n \rightarrow \infty} \frac{H_{n,n^\alpha}}{\ln n} \geq \frac{1}{\varepsilon' + \frac{1}{1-\alpha} \frac{-L'(\theta)}{L(\theta)}} \quad \text{a.s.}$$

The fact that  $\varepsilon'$  can be chose arbitrarily close to 0 ensures that:

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

Taking the limit as  $\theta$  goes to  $\infty = \theta^*$ , we get the result. ■

### 5. STUDY OF THE SATURATION LEVELS

In this section, we prove Theorem 2. We shall assume that  $\theta_* = -\infty$  or (2) holds. We first deal with the lower bound. We shall then focus on the upper bound.

#### 5.1. Lower Bound

We are inspired by the techniques developed in the previous section. Instead of applying (4) as in the study of the heights, we will apply (5). It may explain why there is no phase transition in the asymptotics of  $G_{n,j}$  (see Remark 5).

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

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

for every  $\theta \in [\theta^*, 0)$  if (2) holds, for every  $\theta \in (-\infty, 0)$  if  $\theta_* = -\infty$ .

Recall that  $G_{n,jn^\alpha} > k$  if and only if at generation  $k$ , every box contains at least  $jn^\alpha$  balls when  $n$  balls have been thrown. Because we want to show that  $G_{n,jn^\alpha}$  is bounded from below by  $(1 - \alpha) \frac{-\theta}{\ln L(\theta)} \ln n - r \ln \ln n$ , we take  $k = (1 - \alpha) \frac{-\theta}{\ln L(\theta)} \ln n - r \ln \ln n$ , i.e.,  $n = \exp(k \frac{1}{1-\alpha} \frac{\ln L(\theta)}{-\theta}) k^s$ .

*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  $k \in \mathbb{N}$ :

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

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

$$\mathcal{M}_{x_k, j\phi_k^\alpha}(k) = 0, \quad \text{for all } k \geq k_0(\omega). \tag{8}$$

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

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

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

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

Applying (5), we get

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

Now, as  $y \geq j$ ,

$$\lceil y \rceil \leq \frac{y+1}{y} y \leq \frac{j+1}{j} y.$$

In the notations of Lemma 1, we have:

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

where  $d' := d(-\theta)(1 + \frac{1}{j})^{-\theta}$ . 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 all integers  $k \in \mathbb{N}$ :

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

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 (8).

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

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

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

**Remark 9.** Note that to deduce from Proposition 6 that  $G_{n, jn^\alpha} \geq (1 - \alpha)C_* \ln n$  a.s. in the case  $\theta_* = -\infty$ , one has to show that

$$\lim_{\theta \rightarrow -\infty} \frac{-\theta}{\ln L(\theta)} = C_*.$$

Let us now prove that result. Let  $\psi = \ln L$ . The condition  $\theta_* = -\infty$  implies that  $C_* = \lim_{\theta \rightarrow -\infty} -\frac{1}{\psi'(\theta)}$ . As  $\psi$  is convex, it is known that  $\frac{\psi(\theta)}{-\theta}$  tends to  $l \in (-\infty, \infty]$  as  $\theta$  tends to  $-\infty$ . We distinguish two cases.

Either  $l$  is finite. Let  $g : \theta \mapsto \psi(\theta) + l\theta$ .  $g$  is a convex function and  $\frac{g(\theta)}{\theta}$  tends to 0 as  $\theta$  tends to  $-\infty$ . One can easily deduce that  $g$  is non-decreasing. As a result,  $\psi' \geq -l$ . If  $l \neq \frac{1}{C_*}$ , then  $\psi'(\theta) \rightarrow -l$  as  $\theta$  tends to  $-\infty$ . Now,  $\psi'$  is non-decreasing. 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  $l = \frac{1}{C_*}$ , which means that  $\frac{-\theta}{\ln L(\theta)} \rightarrow C_*$  as  $\theta$  tends to  $-\infty$ .

Or  $l$  is infinite. As  $\theta_* = -\infty$ ,  $\varphi(\theta) > 0$  for all  $\theta < 0$ , so the function  $\theta \in (-\infty, 0) \mapsto \frac{\theta}{\psi(\theta)} - \frac{1}{\psi'(\theta)}$  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:  $1 \geq 1 - \frac{\theta\psi'(\theta)}{\psi(\theta)} \geq -\varepsilon\psi'(\theta)$ . Consequently  $\psi'(\theta) \geq -\frac{1}{\varepsilon}$  and  $\psi(0) - \psi(\theta) \geq \frac{\theta}{\varepsilon}$ . Dividing by  $\theta < 0$ , we obtain:  $-\frac{\psi(\theta)}{\theta} \leq \frac{1}{\varepsilon} - \frac{\psi(0)}{\theta}$ . Taking the limit, we have  $\infty \leq \frac{1}{\varepsilon}$ , which is absurd. Finally,  $\frac{\theta}{\psi(\theta)} - \frac{1}{\psi'(\theta)}$  tends to 0 as  $\theta$  tends to  $-\infty$ . Now, by definition of  $C_*$ ,  $\psi'(\theta)$  tends to  $-\frac{1}{C_*}$ . As a result,  $\frac{-\theta}{\psi(\theta)}$  tends to  $C_*$  as  $\theta$  tends to  $-\infty$ .

## 5.2. Upper Bound

In this section, we are interested in the upper bound of the saturation levels.

5.2.1. *The Case  $-\infty < \theta_* < 0$  and  $\varphi(\theta_*) = 0$ .* We prove that  $G_{n, jn^\alpha} \leq (1 - \alpha)C_* \ln n + O(\ln \ln n)$  a.s. if (2) and (1) hold by studying the smallest box, regardless of the value of  $j$  and  $\alpha$ ; there is no phase transition.

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

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

and

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

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

*Proof.* Let  $\gamma, \gamma'$  and  $\gamma''$  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  $\omega$ , there exists  $k_0(\omega)$  such that

$$\mathcal{M}_{x_k, j\phi_k^\alpha}(k) \geq 1, \quad \text{for all } k \geq k_0(\omega). \tag{9}$$

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

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

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 (9), 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 (4). To do so, let  $p := j$  if  $\alpha = 0$  and  $p := \frac{2}{\gamma' - \gamma - \alpha\gamma''} > 0$  if  $\alpha > 0$ . Note that for all integers  $k$  sufficiently large,  $j\phi_k^\alpha \geq p$ . (4) 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 (9).

Performing the same calculations as at the end of the proof of Proposition 2, 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  $\gamma''$  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. ■

**Remark 10.** *If the assumption (2) holds but (1) does not, one can still show that for all integers  $j \geq 1$  and for all real numbers  $\alpha \in [0, 1)$ , with probability one,  $G_{n,jn^\alpha} \leq (1 - \alpha)C_* \ln n + o(\ln n)$ . To do so, recall that Lemma 3 ensures that  $\frac{-\ln p(k)}{k}$  tends to  $\frac{1}{C_*}$  a.s., and use an argument similar to that in Proposition 7.*

5.2.2. *The Case  $\theta_* = -\infty$ .* In this section, we prove that  $G_{n,jn^\alpha} \leq (1 - \alpha)C_* \ln n + o(\ln n)$  a.s. if  $\theta_* = -\infty$ . We shall see that, once again, the smallest box plays a key role.

**Proposition 8.** *We suppose that  $\theta^* = -\infty$ . Let  $j$  be a positive integer and  $\alpha \in [0, 1)$ . Then*

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

We are inspired by the proof of Proposition 5.

*Proof.* Let  $\theta, \varepsilon$  and  $\varepsilon'$  be three real numbers such that

$$\theta < 0, \quad 0 < \varepsilon < \varepsilon' < \frac{\varepsilon}{\alpha} \quad \text{and} \quad \varepsilon' < \frac{1}{1 - \alpha} \frac{-L'(\theta)}{L(\theta)}.$$

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

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

Note that we suppose that  $\varepsilon' < \frac{1}{1 - \alpha} \frac{-L'(\theta)}{L(\theta)}$  so that  $x_k$  and  $\phi_k$  tend to  $\infty$  as  $k$  tends to  $\infty$ . We first show that a.s.,  $\mathcal{M}_{x_k, j\phi_k^\alpha}(k) \geq 1$  for all integers  $k$  sufficiently large. To do so, we consider the smallest box. Even though we cannot apply Lemma 3, Lemma 2 provide information about the size  $\underline{p}(k)$  of the smallest box  $\underline{b}(k)$ .

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

$$\mathcal{M}_{x_k, j\phi_k^\alpha}(k) \geq 1, \quad \text{for all } k \geq k_0(\omega). \tag{10}$$



We denote by  $A_k$  the event  $\{p(k) \leq \exp(k \frac{L'(\theta)}{L(\theta)})\}$  and by  $B_k$  the event  $\{\text{the box } \underline{b}(k) \text{ contains at least } j\phi_k^\alpha \text{ balls when the first } \mathcal{P}_{x_k} \text{ have been thrown}\}$ . Then:

$$\mathbb{P}(A_k \cap B_k) \leq \mathbb{P}(B_k | A_k) \leq \mathbb{P}(\mathcal{P}_{x_k \exp(k \frac{L'(\theta)}{L(\theta)})} \geq j\phi_k^\alpha).$$

By Markov's inequality, we get:

$$\mathbb{P}(A_k \cap B_k) \leq j^{-1} e^{-k(\varepsilon - \alpha\varepsilon')}.$$

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,  $p(k) > \exp(k \frac{L'(\theta)}{L(\theta)})$  or the box  $\underline{b}(k)$  has less than  $j\phi_k^\alpha$  balls when the first  $\mathcal{P}_{x_k}$  balls have been thrown. Now, by Lemma 2 (recall that  $\theta \in (\theta_*, \theta^*)$  since  $\theta_* = -\infty$ ), we know that a.s., for all integers  $k$  sufficiently large,  $\underline{p}(k) \leq \exp(k \frac{L'(\theta)}{L(\theta)})$ . Finally, (10) is proved.

The usual calculations yield:

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

The fact that  $\varepsilon'$  can be chose arbitrarily close to 0 ensures that:

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

Taking the limit as  $\theta$  goes to  $-\infty = \theta^*$ , we get the result. ■

## 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  $H_{n,n^\alpha}$ ,  $G_{n,j}$  and  $G_{n,n^\alpha}$ . 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 all integers  $k$ . 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 all interval components of  $F(k)$  contain 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  $\{\frac{j}{n+1} : 1 \leq j \leq n\}$  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  $H_{n,n^\alpha}$  would be equivalent to  $(1 - \alpha)C^* \ln n$  a.s. and the saturation levels  $G_{n,jn^\alpha}$  to  $(1 - \alpha)C_* \ln n$  a.s.

We first explain why the clusters of size  $n^\alpha$  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  $H_{n,n^\alpha}$  and  $G_{n,n^\alpha}$ . 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:

$$(\hat{U}_1, \dots, \hat{U}_n) \stackrel{(d)}{=} \left( \frac{\gamma_1}{\gamma_{n+1}}, \dots, \frac{\gamma_n}{\gamma_{n+1}} \right), \tag{11}$$

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  $p = \frac{3}{\varepsilon}$ , we have for all integers  $n$  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 for all integers  $n$  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  $\varepsilon$  tend to 0, we easily conclude that  $\underline{m}_{n,n^\alpha} \geq n^{-1+\alpha+o(1)}$  a.s. Similar calculations yield:  $\bar{m}_{n,n^\alpha} \leq n^{-1+\alpha+o(1)}$  a.s. We conclude that the clusters of size  $n^\alpha$  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  $H_{n,n^\alpha}$  and  $G_{n,n^\alpha}$ .

On the contrary, the clusters of size  $j$  behave very differently from the equidistribution case. For instance, thanks to the identification (11), we may apply well-known results on extreme values (see, e.g., [21]). One finds that  $\underline{m}_{n,j} \approx n^{-\frac{j}{j-1}}$ , so that  $\underline{m}_{n,j}$  is much smaller than  $\frac{j}{n}$ , 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  $\bar{m}_{n,j} \approx \frac{\ln n}{n}$ , 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  $G_{n,j}$ .

## ACKNOWLEDGMENTS

The author would like to express his gratitude to his thesis advisor Jean Bertoin who suggested this subject. The author thanks 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, New York, 1964.
- [2] L. Addario-Berry and B. Reed, Minima in branching random walks, *Ann Probab* 37 (2009), 1044–1079.
- [3] J. Bertoin, *Random Fragmentation and Coagulation Processes*, Vol. 102, Cambridge Studies in Advanced Mathematics, Cambridge, 2006.
- [4] J. Bertoin, Asymptotic regimes for the occupancy scheme of multiplicative cascades, *Stochastic Process Appl* 118 (2008), 1586–1605.
- [5] J. D. Biggins, Uniform convergence of martingales in the branching random walk, *Ann Probab* 20 (1992), 137–151.
- [6] S. Boucheron and D. Gardy, An urn model from learning theory, *Random Struct Algorithms* 10 (1997), 43–67.
- [7] J. Bunge and M. Fitzpatrick, Estimating the number of species: A review, *J Am Stat Assoc* 88 (1993), 364–373.
- [8] L. Devroye, A note on the height of binary search trees, *J Assoc Comput Mach* 33 (1986), 489–498.
- [9] L. Devroye, Branching processes in the analysis of the heights of trees, *Acta Inform* 24 (1987), 277–298.
- [10] L. Devroye, Universal limit laws for depths in random trees, *SIAM J Comput* 28 (1999), 409–432.
- [11] D. Gardy, Occupancy urn models in the analysis of algorithms, *J Statist Plann Inference* 101 (2002), 95–105.
- [12] A. Gnedin, B. Hansen, and J. Pitman, Notes on the occupancy problem with infinitely many boxes: General asymptotics and power laws, *Probab Surv* 4 (2007), 146–171.
- [13] B. Haas, G. Miermont, J. Pitman, and M. Winkel, Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models, *Ann Probab* 36 (2008), 1790–1837.
- [14] F. He and K. J. Gaston, Estimating species abundance from occurrence, *Am Naturalist* 156 (2000), 553–559.
- [15] L. Holst, On birthday, collectors', occupancy and other classical urn problems, *Int Statist Rev* 54 (1986), 15–27.
- [16] Y. Hu and Z. Shi, Minimal position and critical martingale convergence in branching random walks, and directed polymers on disordered trees, *Ann Probab* 37 (2009), 742–789.
- [17] N. L. Johnson and S. Kotz, *Urn models and their application. An approach to modern discrete probability theory*, Wiley, New York, 1977.
- [18] V. F. Kolchin, B. A. Sevast'yanov, and V. P. Chistyakov, *Random allocations*, Wiley, New York, 1978.

- [19] McDiarmid, Minimal positions in a branching random walk, *Ann Appl Probab* 5 (1995), 128–139.
- [20] B. Pittel, On growing random binary trees, *J Math Anal Appl* 103 (1984), 461–480.
- [21] S. Resnick, Extreme values, regular variation, and point processes, *Springer Series in Operations Research and Financial Engineering*, New York, 1987.
- [22] M. Robson, The height of binary search trees, *Austr Comput J* 11 (1979), 151–153.