

Asymptotics of the occupancy scheme in a random environment and its applications to tries

Silvia Businger

Universität Zürich, Switzerland

received 19th Sep. 2016, accepted 26th July 2017.

Consider m copies of an irreducible, aperiodic Markov chain Y taking values in a finite state space. The asymptotics as m tends to infinity, of the first time from which on the trajectories of the m copies differ, have been studied by Szpankowski (1991) in the setting of tries. We use a different approach and model the m trajectories by a variant of the occupancy scheme, where we consider a nested sequence of boxes. This approach will enable us to extend the result to the case when the transition probabilities are random. We moreover use the same techniques to study the asymptotics as m tends to infinity of the time up to which we have observed all the possible trajectories of Y in random and nonrandom scenery.

Keywords: Occupancy scheme, Coupon collector Problem, random environment, tries

1 Introduction

Let $Y = (Y_k)_{k \in \mathbb{N}}$ be an irreducible, aperiodic Markov chain taking values in a finite state space, say $\Sigma = \{1, \dots, K\}$, denote its transition probabilities by $(p_{ij}, i, j \in \Sigma)$, and consider m independent copies of Y . We define $H_m \in \mathbb{N}$, the first time from which on the trajectories of the m copies differ and $G_m \in \mathbb{N}$, the maximal time up to which we have observed all the possible trajectories of Y . A natural question which has generated a lot of interest in the literature concerns the asymptotic behavior of H_m and G_m as m tends to infinity.

Typically in information theory one considers the Markov source model (see for example Ash (2012)), that is an infinite data string is modeled by the trajectory of a Markov chain taking values in a finite alphabet Σ . We call the first n letters of the data string a word of length n . If we consider m independent data strings modeled by a Markov source, then H_m is the minimal length n such that the m words with length n are all distinct. In the same spirit G_m describes up to which length we observe all the possible words. More generally we will consider $H_{m,j}$ the minimal length such that out of the m words with length n at most $j - 1$ words are non-distinct, and $G_{m,j}$ describing up to which length we observe all the possible words at least j times.

The length H_m can be described in terms of a class of combinatorial trees, the so called K -ary tries that can be constructed with an iterative procedure. Consider m words with letters in the alphabet Σ . We

start with the root of a tree. We then look at the first letter of the m words. For each different letter we add a node to the root, in increasing order. If for a letter there is only one word starting with this letter, then the corresponding node will be a leaf of the tree and we store the word in that leaf. We then look at the second letters of the words not yet being stored in a leaf. If there are at least two words starting with the letter $i \in \Sigma$, we then look at the second letters of all the words starting with i . For each different letter we add a node to the node i in the same way as in the first step. We then repeat this procedure until all the words are stored in a leaf. For more details see for example Drmota (2009). When the data strings have been generated by a Markov source then the height of this K -ary trie is equal to H_m .

The height of tries have been a subject in research for many years, see Flajolet and Steyaert (1982) and Devroye (1984). The limit law of the height has finally been derived by Pittel (1985), see also Pittel (1986). The Markov source model, was studied by Szpankowski (1991). He established that

$$H_m \sim \frac{2}{-\ln(\rho(2))} \ln(m), \quad a.s.,$$

as m tends to infinity, where $\rho(2)$ denotes the maximum modulus eigenvalue of the matrix $A(2) = (p_{ik}^2)_{i,k \in \Sigma}$, by analyzing the longest common prefix of each possible pair of two data strings.

This result can be extended to the case when each external node is allowed to store up to $j - 1$ data strings. The tree defined by this modification is called a $(j - 1)$ -trie. Szpankowski (see Szpankowski (1991) and Szpankowski (2011)) studied the asymptotics of the height of a $(j - 1)$ -trie, he established that

$$H_{m,j} \sim \frac{j}{-\ln(\rho(j))} \ln(m), \quad a.s.,$$

as m tends to infinity, where $\rho(j)$ denotes the maximum modulus eigenvalue of the matrix given by $A(j) = (p_{ik}^j)_{i,k \in \Sigma}$.

Our purpose in this work is to investigate a similar problem in random environment. This is a natural model if for instance one wants to take into account transmission errors in the setting of information theory.

In order to explain how, in our model, the random environment acts on the trajectories of chains, let us first recall a simple construction of i.i.d. copies of the Markov chain Y with transition matrix $(p_{ik})_{i,k \in \Sigma}$ using the balls-in-bins setting (see, e.g. Devroye (2005)). For a general introduction to balls-in-bins in deterministic environment and the classical occupancy scheme with finitely many boxes see for example Kolchin et al. (1978) or Johnson and Kotz (1977). There is a broad literature on the occupancy scheme in deterministic environment and comparatively few investigations of the occupancy scheme in random environment. An infinite occupancy scheme in a random environment called the Bernoulli sieve has been introduced in Gnedin (2004) and then studied in depth, a survey can be found for example in Gnedin et al. (2010). A scheme of a similar type can also be found in Robert and Simatos (2009).

The set of words with length n in the alphabet Σ is Σ^n , with the convention that for $n = 0$, $\Sigma^0 = \{\emptyset\}$ is the empty word. For the sake of simplicity, we assume that the initial state of the Markov chain Y is always 1, and construct a nested family of boxes (or bins) indexed by the regular K -ary tree $\mathcal{U} = \bigcup_{n \in \mathbb{Z}_+} \Sigma^n$ as follows. At generation 0, there is a single box b_\emptyset with unit size and type 1. At the first generation, we divide the box b_\emptyset into $(b_i : i \in \Sigma)$ where b_i has type i and size $|b_i| = p_{1i}$. We iterate for the next generations in an obvious way. For each word $u = (i_1, \dots, i_n) \in \Sigma^n$, the box b_u has type i_n and is split at the next generation into sub-boxes $(b_{uk} : k \in \Sigma)$, where b_{uk} has type k and size $|b_{uk}| = |b_u|p_{i_n k}$. Now imagine

that we throw a ball into the initial box, distribute it uniformly at random to the next generation of boxes, and observe the sequence of the types of the sub-boxes it passes through, generation after generation. We then clearly obtain a version of the Markov chain Y , and more generally, throwing m balls independently yields the trajectories of m i.i.d. copies of Y . In this setting, the height $H_{m,j}$ of a $(j-1)$ -trie corresponds to the first generation at which all boxes contain strictly less than j balls. For $j \geq 1$ the first generation when there is a box containing strictly less than j balls when m balls have been thrown corresponds to $G_{m,j}$, and is referred to as saturation level.

The random environment that we shall consider corresponds to splitting boxes randomly rather than deterministically, in a Markovian manner. More precisely, we now consider for each word $u = (i_1, \dots, i_n) \in \Sigma^n$, an independent copy $A_u = (p_{ik}(u))_{i,k \in \Sigma}$ of a *random* transition matrix $A = (p_{ik})_{i,k \in \Sigma}$. The box b_u with type i_n is split at the next generation into $(b_{uk} : k \in \Sigma)$, where b_{uk} has type k and size $|b_{uk}| = |b_u|p_{i_n k}(u)$. Then throwing m balls uniformly at random yields the m data strings in random environment that we are interested in. Note that knowing the family of box sizes, the trajectory of a ball through boxes is not described by a Markov chain. It is thus not possible to reduce the study to applying the results due to Szpankowski by conditioning on the environment.

In order to investigate the behavior of the heights $H_{m,j}$ and saturation levels $G_{m,j}$ as m tends to infinity, we shall first show how the results by Szpankowski in deterministic environment can be recovered using an occupancy scheme analysis. Even though the main result has already been established by different arguments, we shall provide a detailed account as the same approach can then be adapted to the random environment setting. Specifically, we shall investigate the distribution of the sizes of the boxes at a large generation n , being especially interested in large deviation type estimates. We will moreover use the same techniques to study the asymptotics of the height $H_{m,j}$ when j depends on m , more precisely $j = j(m) = m^\alpha$ for $\alpha \in (0, 1)$.

We shall then show that this approach can be adapted in the random environment setting. In this direction, we shall first observe that taking the logarithm of the sizes of boxes yields a multitype branching random walk, which then enables us to apply large deviations estimates due to Biggins and Rahimzadeh Sani (2005). We will see that different from the nonrandom case there is a phase transition in the asymptotic behavior of $H_{m,j}$ as m tends to infinity and that

$$H_{m,j} \sim C(j) \cdot \ln(m) \quad a.s.$$

where $C(j)$ is a constant depending on j for j smaller than a critical parameter and $C(j) = \zeta^*$ is a constant arising in the limit behavior of the largest box for j larger than the critical parameter.

The study of the limiting behavior of the saturation level is closely related to the coupon collector's problem. (See for example Rosen (1970)). We will consider each box of generation n as a coupon, each of a different sort and suppose that a collector wants to have at least one of each sort. We say that a collector buys a certain coupon if a ball lands in the corresponding box. Then $G_{m,j}$ is the first generation n when the collector fails to have at least j coupons of each sort. We will see that:

$$G_{m,j} \sim \zeta_* \cdot \ln(m) \quad a.s.$$

where ζ_* is a constant appearing in the limit behavior of the smallest box.

Some of our arguments are inspired by the works by Bertoin (2008) and Joseph (2010), who used the fundamental results of Biggins (1992) on the asymptotic behaviors of branching random walks to investigate limits of occupancy scheme in the setting of random multiplicative (monotype) cascades.

1.1 Large deviation behavior of the box sizes

In this section we give some results on the asymptotic behavior of the box sizes as the generation n tends to infinity. We will for simplicity assume that we start from a box of type 1.

Recall that $(p_{ij}, i, j \in \Sigma)$ denotes the transition probabilities of the irreducible, aperiodic Markov chain Y . For each $\theta \in \mathbb{R}$ we define the $K \times K$ -matrix $A(\theta) := (p_{ij}^\theta)_{i,j \in \Sigma}$, with the convention that if $p_{ij} = 0$, then $p_{ij}^\theta = 0$, even for $\theta \leq 0$. The matrix $A(\theta)$ is connected to the box sizes in the following way:

Let $(l_{i,k}^{(n)})_k$ denote the sequence of the sizes of boxes with type i at generation n , lexicographically ordered. Let us further introduce the point measure

$$Z_j^{(n)} = \sum_k \delta_{-\ln(l_{j,k}^{(n)})},$$

and the Laplace transform of $Z_j^{(n)}$, that is:

$$\mathcal{L}_j^{(n)}(\theta) = \sum_k (l_{j,k}^{(n)})^\theta.$$

A simple iteration argument then shows, that

$$(\mathcal{L}_1^{(n)}(\theta), \dots, \mathcal{L}_K^{(n)}(\theta)) = (1, 0, \dots, 0)(A(\theta))^n. \quad (1)$$

Recall that we call an eigenvalue ρ of a matrix A a maximum modulus eigenvalue, if it is a simple root of the characteristic polynomial and its modulus is strictly larger than the modulus of the other roots.

Note that since Y_n is aperiodic and irreducible, $A(\theta)$ is positive regular (that is its entries are finite and there exists some positive integer r such that all entries of the matrix $A(\theta)^r$ are strictly positive.) Recall that we then have from the Perron-Frobenius theorem:

1. $A(\theta)$ possesses a unique maximum modulus eigenvalue $\rho(\theta) \in \mathbb{R}$,
2. there exists a strictly positive left-eigenvector $w(\theta) = (w_1(\theta), \dots, w_K(\theta))$ and a strictly positive right-eigenvector $v(\theta) = (v_1(\theta), \dots, v_K(\theta))$ with eigenvalue $\rho(\theta)$, normalized such that we have $(w(\theta))^t v(\theta) = 1$,
3. $\min_i \sum_j A(\theta)_{ij} \leq \rho(\theta) \leq \max_i \sum_j A(\theta)_{ij}$.

We moreover have that:

Lemma 1 (Biggins and Rahimzadeh Sani (2004)) *The maximum modulus eigenvalue $\rho(\theta)$ is analytic in θ .*

We deduce:

Lemma 2 *For each $\theta \in \mathbb{R}$ we have:*

$$\lim_{n \rightarrow \infty} \mathcal{L}_i^{(n)}(\theta) \rho(\theta)^{-n} = v_1(\theta) w_i(\theta).$$

We shall also need the following:

Proposition 1 (Kingman (1961)) *The logarithm of the maximum modulus eigenvalue $\ln(\rho(\theta))$ is a convex function of θ .*

Remark 1 *Note that this also entails the convexity of $\rho(\theta)$.*

Now, introduce the constants

$$C_* := \lim_{\theta \rightarrow -\infty} \frac{\rho(\theta)}{-\rho'(\theta)} \quad \text{and} \quad C^* := \lim_{\theta \rightarrow \infty} \frac{\rho(\theta)}{-\rho'(\theta)},$$

that will play a crucial role in the asymptotic behavior of the smallest and largest box.

Lemma 3 *We have $0 < C_* \leq C^* < \infty$.*

Proof: First note that the fact that $\ln(\rho(\theta))$ is convex entails that $\frac{\rho(\theta)}{-\rho'(\theta)}$ is an increasing function. Assume first $\theta > 0$ and let $p_* := \inf_{(i,j) \in S} p_{ij}$, where $S := \{(i, j) : p_{ij} > 0\}$. We then have

$$0 < p_*^\theta \leq \min_i \sum_j p_{ij}^\theta \leq \rho(\theta).$$

Now, let r be such that the matrix A^r has only positive entries and define the supremum of its entries $p^* := \sup_{(i,j)} A_{i,j}^r$. Note that $p^* < 1$, that $A(\theta)^r$ has only positive entries, and that its maximum modulus eigenvalue is $\rho(\theta)^r$. Let $(l_{i,j,k}^{(r)})_k$ denote the sequence of the sizes of boxes with type j at generation r when the first box was of type i . Then:

$$\rho(\theta)^r \leq \max_i \sum_j A(\theta)_{ij}^r = \max_i \sum_j \sum_k \left(l_{i,j,k}^{(r)} \right)^\theta \leq K \cdot K^r (p^*)^\theta.$$

Similarly if $\theta < 0$ we have $\rho(\theta) \leq K \cdot p_*^\theta$ and $\rho(\theta)^r \geq (p^*)^\theta$, and thus:

$$-\infty < \ln(p_*) \leq \lim_{\theta \rightarrow \pm\infty} \frac{\ln(\rho(\theta))}{\theta} \leq \frac{\ln(p^*)}{r} < 0.$$

We get by l'Hôpital's rule that $-\infty < -\frac{1}{C_*} < 0$ and $-\infty < -\frac{1}{C^*} < 0$. □

Now, let us define the size-biased pick of a box of generation n . That is define a random variable L_n by $\mathbb{P}(L_n = l_{i,k}^{(n)}) = l_{i,k}^{(n)}$. (Recall that $\sum_{i=1}^K \sum_k l_{i,k}^{(n)} = 1$.) We then have:

Lemma 4 *Let $\theta \in \mathbb{R}$, $z_n(\theta) := e^{n \frac{\rho'(\theta)}{\rho(\theta)}}$ and $\phi(\theta) := \ln(\rho(\theta)) - (\theta - 1) \frac{\rho'(\theta)}{\rho(\theta)}$, then $-\infty < \phi(\theta) \leq 0$ for all $\theta \in \mathbb{R}$, and:*

$$\lim_{\varepsilon \downarrow 0} \lim_{n \rightarrow \infty} \frac{1}{n} \ln \left(\mathbb{P} \left(L_n \in (z_n(\theta)e^{-n\varepsilon}, z_n(\theta)e^{n\varepsilon}) \right) \right) = \phi(\theta).$$

Proof: First note that $\phi(1) = 0$, and $\phi'(\theta) = (1 - \theta) \ln(\rho(\theta))''$, thus $\phi(\theta)$ is decreasing for $\theta > 1$ and increasing for $\theta < 1$. We then have $\phi(\theta) \leq 0$, and the fact that $-\infty < \phi(\theta)$ follows from the fact that ρ is analytic. Now, let $X_n := \frac{1}{n} \ln(L_n)$. We want to apply the Gärtner-Ellis theorem (see for

example Dembo and Zeitouni (2010), Theorem 2.3.6), to the random variable X_n . Let $\lambda \in \mathbb{R}$ and define $\Lambda_n(\lambda) := \ln \mathbb{E}[e^{\lambda X_n}]$. We first need to check that the limit

$$\Lambda(\lambda) := \lim_{n \rightarrow \infty} \frac{1}{n} \Lambda_n(n\lambda)$$

exists as an extended real number. Recall that by Lemma 1 we have

$$\frac{K}{2} v(\lambda) (w(\lambda))^t \rho(\lambda)^n \leq \sum_{i=1}^K \mathcal{L}_i^{(n)}(\lambda) \leq 2K v(\lambda) w^t(\lambda) \rho(\lambda)^n,$$

for sufficiently large n , and thus

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \Lambda_n(n\lambda) &= \lim_{n \rightarrow \infty} \frac{1}{n} \ln \left(\sum_{i=1}^K \sum_k (l_{i,k}^{(n)})^{(\lambda+1)} \right) \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \ln \left(\sum_{i=1}^K \mathcal{L}_i^{(n)}(\lambda+1) \right) \\ &= \ln(\rho(\lambda+1)) \\ &= \Lambda(\lambda). \end{aligned}$$

We further need to check that Λ is lower semicontinuous and essentially smooth, that is for

$$\mathcal{D}_\Lambda := \{\lambda \in \mathbb{R} : \Lambda(\lambda) < \infty\} = \mathbb{R},$$

the interior $\mathcal{D}_\Lambda^\circ$ is non empty, Λ is differentiable on $\mathcal{D}_\Lambda^\circ$, and $\lim_{n \rightarrow \infty} |\Lambda'(\lambda_n)| = \infty$ for every sequence in $\mathcal{D}_\Lambda^\circ$ converging to a boundary point of $\mathcal{D}_\Lambda^\circ$. The essential smoothness and lower semicontinuity then follow readily from the fact that Λ is differentiable on \mathbb{R} . Thus, by the Gärtner-Ellis theorem the large deviation principle holds with good rate function

$$\Lambda^*(z) = \sup_{\mu \in \mathbb{R}} (\mu z - \ln(\rho(\mu+1))).$$

Note that $\Lambda^*(z) \geq 0$ as $\rho(1) = 1$. For $\varepsilon > 0$ let $B_\varepsilon(\theta) := \left(\frac{\rho'(\theta)}{\rho(\theta)} - \varepsilon, \frac{\rho'(\theta)}{\rho(\theta)} + \varepsilon \right)$ and let $\overline{B}_\varepsilon(\theta)$ denote its closure. By the large deviation principle:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln(\mathbb{P}(X_n \in B_\varepsilon(\theta))) \geq - \inf_{x \in B_\varepsilon(\theta)} \Lambda^*(x)$$

and

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \ln(\mathbb{P}(X_n \in B_\varepsilon(\theta))) \leq - \inf_{x \in \overline{B}_\varepsilon(\theta)} \Lambda^*(x).$$

Now, note that $-\Lambda^*\left(\frac{\rho'(\theta)}{\rho(\theta)}\right) = \phi(\theta)$ and thus:

$$\begin{aligned} \phi(\theta) &\leq \liminf_{\varepsilon \downarrow 0} \liminf_{n \rightarrow \infty} \frac{1}{n} \ln(\mathbb{P}(X_n \in B_\varepsilon(\theta))) \\ &\leq \limsup_{\varepsilon \downarrow 0} \limsup_{n \rightarrow \infty} \frac{1}{n} \ln(\mathbb{P}(X_n \in B_\varepsilon(\theta))) \leq \phi(\theta). \end{aligned}$$

This proves our claim. \square

We next provide another approach to the previous lemma. Imagine that we follow the trajectory of a ball through the nested sequence of boxes and recall that the sequence of types of the boxes the ball passes through corresponds to a trajectory of the Markov chain Y . Now, let T_n denote the type of the box of the n -th generation the ball is passed through and note that L_n defined in the previous Lemma corresponds to its size. Then (T_n) is an irreducible Markov chain with transition probabilities p_{ij} and (T_n, T_{n+1}) is an irreducible Markov chain with state space $\{(i, j) \in \Sigma \times \Sigma : p_{ij} \neq 0\}$ and transition probabilities $\Pi((i, j), (j, k)) = p_{ij}$ and $\Pi((i, j), (m, k)) = 0$ if $m \neq j$. Further we have that

$$L_n = \prod_{k=0}^{n-1} p_{T_k, T_{k+1}}.$$

We prove again Lemma 4 now using the large deviation principle for additive functionals of Markov chains.

Proof: Define the deterministic function

$$f : \Sigma \times \Sigma \rightarrow \mathbb{R}, \quad (i, j) \mapsto p_{ij},$$

and the empirical means

$$X_n := \frac{1}{n} \sum_{k=0}^{n-1} f(T_k, T_{k+1}) = \frac{1}{n} \sum_{k=0}^{n-1} \ln(p_{T_k, T_{k+1}}) = \frac{1}{n} \ln(L_n).$$

By Theorem 3.1.2 in Dembo and Zeitouni (2010), X_n fulfills the large deviation principle with good rate function

$$\Lambda^*(z) = \sup_{\mu \in \mathbb{R}} (\mu z - \ln(\rho(\mu + 1))).$$

The claim then follows in the same spirit as before. \square

Let us define a function that will play a crucial role in our analysis:

$$\psi(\theta) := \ln(\rho(\theta)) - \frac{\rho'(\theta)}{\rho(\theta)} \theta \quad (2)$$

We then have:

Corollary 1 *Let $\theta \in \mathbb{R}$ and define $z_n(\theta) := e^{n \frac{\rho'(\theta)}{\rho(\theta)}}$. We then have:*

$$\lim_{\epsilon \downarrow 0} \lim_{n \rightarrow \infty} \frac{1}{n} \ln \left(\sum_{i=1}^K \sum_k 1_{\{l_{i,k}^{(n)} \in (z_n(\theta)e^{-n\epsilon}, z_n(\theta)e^{n\epsilon})\}} \right) = \psi(\theta).$$

Proof: First note that

$$\mathbb{P}(L_n \in (z_n(\theta)e^{-n\epsilon}, z_n(\theta)e^{n\epsilon})) = \sum_{i=1}^K \sum_k 1_{\{l_{i,k}^{(n)} \in (z_n(\theta)e^{-n\epsilon}, z_n(\theta)e^{n\epsilon})\}} l_{i,k}^{(n)}.$$

Thus

$$\begin{aligned} \sum_{i=1}^K \sum_k 1_{\{l_{i,k}^{(n)} \in (z_n(\theta)e^{-n\epsilon}, z_n(\theta)e^{n\epsilon})\}} e^{-n\epsilon} &\leq z_n(\theta)^{-1} \cdot \mathbb{P}(L_n \in (z_n(\theta)e^{-n\epsilon}, z_n(\theta)e^{n\epsilon})) \\ &\leq \sum_{i=1}^K \sum_k 1_{\{l_{i,k}^{(n)} \in (z_n(\theta)e^{-n\epsilon}, z_n(\theta)e^{n\epsilon})\}} e^{n\epsilon}. \end{aligned}$$

By Lemma 4 we derive that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \ln \left(\sum_{i=1}^K \sum_k 1_{\{l_{i,k}^{(n)} \in (z_n(\theta)e^{-n\epsilon}, z_n(\theta)e^{n\epsilon})\}} \right) \leq \psi(\theta) + \epsilon,$$

and

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \ln \left(\sum_{i=1}^K \sum_k 1_{\{l_{i,k}^{(n)} \in (z_n(\theta)e^{-n\epsilon}, z_n(\theta)e^{n\epsilon})\}} \right) \geq \psi(\theta) - \epsilon$$

and we easily conclude. \square

We moreover have that:

Lemma 5 *We have that $\psi(\theta) > 0$ for all $\theta \in \mathbb{R}$.*

Proof: First note that Corollary 1 implies that $\psi(\theta) \geq 0$ for all $\theta \in \mathbb{R}$. Further we have that $\psi(0) > 0$ since $\rho(0) > 1$. Moreover $\psi'(\theta) = -\theta \cdot \ln(\rho(\theta))''$ and thus, by convexity of $\ln(\rho(\theta))$, we have that ψ is increasing on the interval $(-\infty, 0)$ and decreasing on the interval $(0, \infty)$. Together with the fact that ψ is analytic and thus cannot be zero on any interval, we arrive at $\psi(\theta) > 0$ for all $\theta \in \mathbb{R}$. \square

We can use the previous results to gain information about the asymptotic behavior of the sizes of the smallest box and the largest box.

Lemma 6 *Let $\underline{l}^{(n)}$ denote the size of the smallest box at generation n . We then have:*

$$\lim_{n \rightarrow \infty} \frac{\ln(\underline{l}^{(n)})}{n} = -\frac{1}{C_*}$$

Proof: Let $\theta < 0$. By Lemma 1 there exists a n_0 , such that for all $n \geq n_0$:

$$(\underline{l}^{(n)})^\theta \leq \mathcal{L}_i^{(n)}(\theta) \leq 2v_1(\theta)(w_i(\theta))^t \rho(\theta)^n.$$

By some rearrangement we thus get:

$$\liminf_{n \rightarrow \infty} \frac{\ln(\underline{l}^{(n)})}{n} \geq \frac{\ln(\rho(\theta))}{\theta},$$

and we conclude by letting θ tend to $-\infty$. Now let $\varepsilon > 0$. By Lemma 4 there exists a natural number $n_0(\varepsilon)$ such that $\underline{l}^{(n)} \leq e^{n \frac{\rho'(\theta)}{\rho(\theta)}} e^{\varepsilon n}$ for all $n \geq n_0(\varepsilon)$, thus

$$\limsup_{n \rightarrow \infty} \frac{\ln(\underline{l}^{(n)})}{n} \leq \frac{\rho'(\theta)}{\rho(\theta)} + \varepsilon,$$

and the result follows by letting ε tend to zero. □

In the same way one shows for the largest box that:

Lemma 7 *Let $\bar{l}^{(n)}$ denote the size of the largest box at generation n . We then have:*

$$\lim_{n \rightarrow \infty} \frac{\ln(\bar{l}^{(n)})}{n} = -\frac{1}{C^*}$$

2 Height and saturation level of Markovian tries

2.1 Height of Markovian tries

Recall that $H_{m,j}$ for $j \geq 2$, denotes the first generation of boxes at which all the boxes contain strictly less than j balls when m balls have been thrown independently. We shall show how the large deviation estimates of the preceding section enable us to recover the following result.

Theorem 1 (Szpankowski (1991)) *For every $j \geq 2$ we have*

$$\lim_{m \rightarrow \infty} \frac{1}{\ln(m)} H_{m,j} = \frac{j}{-\ln(\rho(j))} \quad a.s.$$

Note that we now continue to successively throw balls forever, and that the randomness is coming from the way we distribute the balls into the boxes.

We will first tackle the upper bound. Let $N_{m,j}^{(n)}$ denote the number of boxes at generation n containing j or more balls when m balls have been thrown. Note that $H_{m,j} < n$ when $N_{m,j}^{(n)} = 0$. The idea will be to analyze the asymptotic behavior of $N_{m,j}^{(n)}$ as n and m tend to infinity. We will show that

$$H_{m,j} \leq \frac{j}{-\ln(\rho(j))} \ln(m) + O(\ln \ln(m)) \quad a.s.,$$

as m tends to infinity. Let $a > \frac{1}{j}$ and define the sequence

$$x_n := \rho(j)^{-\frac{n}{j}} n^{-a}.$$

We then have:

Lemma 8 *For almost all ω , there exists a natural number $n_0(\omega)$ s.t for all $n \geq n_0(\omega)$, $N_{[x_n],j}^{(n)} = 0$.*

Proof: Let $B(m, p)$ denote a generic Binomial variable with parameter $p \in [0, 1]$ and $m \in \mathbb{N}$ and note that

$$\mathbb{P}(B(m, p) \geq j) \leq \binom{m}{j} p^j \leq m^j p^j. \quad (3)$$

Further note that the number of balls in a box of size l when m balls have been thrown is $B(m, l)$ distributed. We thus have

$$\begin{aligned}\mathbb{E}[N_{\lfloor x_n \rfloor, j}^{(n)}] &= \sum_{i=1}^K \sum_k \mathbb{P}\left(B\left(\lfloor x_n \rfloor, l_{i,k}^{(n)}\right) \geq j\right) \\ &\leq \sum_{i=1}^K \sum_k \left(\lfloor x_n \rfloor l_{i,k}^{(n)}\right)^j \\ &\leq \sum_{i=1}^K x_n^j \mathcal{L}_i^n(j).\end{aligned}$$

By Lemma 2 there exists a natural number n_1 such that for all $n \geq n_1$

$$\mathcal{L}_i^{(n)}(j) \rho(j)^{-n} \leq 2v_1(j)w_i(j)^t.$$

Taking $c_1(j) := 2v_1(j)w_i(j)^t$ we get that

$$\mathbb{E}[N_{\lfloor x_n \rfloor, j}^{(n)}] \leq \sum_{i=1}^K c_1(j) x_n^j \rho(j)^n \leq \sum_{i=1}^K c_1(j) n^{-aj},$$

for all $n \geq n_1$. We finally arrive at

$$\mathbb{E}\left[\sum_{n \geq n_1} \mathbf{1}_{\{N_{\lfloor x_n \rfloor, j}^{(n)} \geq 1\}}\right] \leq \mathbb{E}\left[\sum_{n \geq n_1} N_{\lfloor x_n \rfloor, j}^{(n)}\right] \leq \sum_{n \geq n_1} c_1(j) n^{-aj} < \infty,$$

and we conclude by the Borel-Cantelli lemma. \square

This lemma will be enough to show that:

Proposition 2 *For every integer $j \geq 2$ we have*

$$\frac{1}{\ln(m)} H_{m,j} \leq \frac{j}{-\ln(\rho(j))} + O\left(\frac{\ln \ln(m)}{\ln(m)}\right) \quad a.s.$$

as m tends to infinity.

Proof: First note that there exists a natural number n_1 such that $\forall n \geq n_1$ we have $n^{-a} \geq \exp\left(-n \frac{-\ln(\rho(j))}{2j}\right)$ and there exists a natural number n_2 such that for all $n \geq n_2$ the sequence x_n is increasing. Then using Lemma 8 choose an $\omega \in \Omega$ for which there exists a natural number $n_0(\omega)$ such that for all $n \geq n_0(\omega)$ we have $N_{\lfloor x_n \rfloor, j}^{(n)} = 0$. Let $n_3 \geq \max(n_0(\omega), n_1 + 1, n_2 + 1)$ and note that for each $m \geq x_{n_3}$ there exists a unique $n \geq n_3$ such that $x_{n-1} \leq m < x_n$. From $m \leq \lfloor x_n \rfloor$ we have $H_{m,j} \leq n$. Moreover taking logarithm on both sides of the inequality $x_{n-1} \leq m$, we derive that

$$n \leq \frac{j}{-\ln(\rho(j))} \ln(m) + \frac{j}{-\ln(\rho(j))} a \ln(n-1) + 1.$$

Since $n \geq n_1 + 1$, we have that

$$m \geq x_{n-1} = \exp\left((n-1)\frac{-\ln(\rho(j))}{j}\right) (n-1)^{-a} \geq \exp\left((n-1)\frac{-\ln(\rho(j))}{2j}\right).$$

Taking logarithm, we derive that

$$(n-1) \leq \frac{2j}{-\ln(\rho(j))} \ln(m)$$

and finally arrive at:

$$H_{m,j} \leq n \leq \frac{j}{-\ln(\rho(j))} \ln(m) + \frac{j}{-\ln(\rho(j))} a \ln\left(2 \frac{j}{-\ln(\rho(j))} \ln(m)\right) + 1,$$

and we easily conclude. \square

Remark 2 One could also have derived Proposition 2, by applying Theorem 6.B in Barbour et al. (1992).

We now turn to the proof of the lower bound. Recall that $H_{m,j} \geq n$, if at generation n there is at least one box containing j or more balls. We thus have that $H_{m,j} > n$ if $N_{m,j}^{(n)} \geq 1$. As in the proof of the upper bound we want to analyze the asymptotic behavior of $N_{m,j}^{(n)}$ as n and m tend to infinity. We want to take $n = \frac{j}{-\ln(\rho(j))} \ln(m) + o(\ln(m))$. Recall that $\psi(\theta) = \ln(\rho(\theta)) - \frac{\rho'(\theta)}{\rho(\theta)}\theta$ and that $\psi(\theta) > 0$. Let $\psi(j) > \varepsilon' > 0$ and define the sequence:

$$x_n := \rho(j)^{\frac{-n}{j}} e^{n\varepsilon'}.$$

In this section we make use of the classical Poissonization trick. Instead of throwing x_n balls initially we will throw the random number of balls $\text{Poisson}(x_n)$. The advantage of this procedure is that if we consider two different boxes b and \tilde{b} with size l respectively \tilde{l} , then the number of balls in b and \tilde{b} are independent Poisson random variables with parameter $x_n l$ respectively $x_n \tilde{l}$.

Lemma 9 For almost all ω there exists a natural number $n_0(\omega)$, such that there exists at least one box at generation n containing j or more balls when $\text{Poisson}(x_n)$ balls have been thrown.

Proof: From Corollary 1, we know that for all $\varepsilon > 0$ there exists a natural number $n_1(\varepsilon)$ such that

$$\frac{1}{n} \ln \left(\sum_{i=1}^K \sum_k 1_{\left\{ l_{i,k}^{(n)} \in \left(e^{n \frac{\rho'(j)}{\rho(j)}} e^{-n\varepsilon}, e^{n \frac{\rho'(j)}{\rho(j)}} e^{n\varepsilon} \right) \right\}} \right) \geq \psi(j) - \varepsilon, \quad (4)$$

for all $n \geq n_1(\varepsilon)$. Now, let $\varepsilon < \frac{j}{j+1} \varepsilon'$ and let M_n denote the set containing all the boxes with size larger than $z_n := e^{n \left(\frac{\rho'(j)}{\rho(j)} - \varepsilon \right)}$ at generation n . From (4) we deduce that

$$|M_n| \geq v_n(j) := e^{n(\psi(j) - \varepsilon)},$$

for all $n \geq n_1(\varepsilon)$. For n large enough we can thus consider the first v_n boxes in M_n , say $b_1(n), \dots, b_{v_n(j)}(n)$ and denote their size with $l_1(n), \dots, l_{v_n(j)}(n)$. We then place an imaginary box $\mathbf{b}_i(n)$ in $b_i(n)$ for

$1 \leq i \leq v_n(j)$, each of size exactly z_n . When a ball falls into the box $b_i(n)$ it arrives in the imaginary box $\mathbf{b}_i(n)$ with probability $\frac{z_n}{l_i}$. We want to show that there exists a natural number n_0 such that for all $n \geq n_0$ at least one of the boxes $\mathbf{b}_i(n)$, $1 \leq i \leq v_n(j)$ contains more than j balls when $\text{Poisson}(x_n)$ balls have been thrown.

Let A_n denote the event that all boxes $\mathbf{b}_i(n)$, $1 \leq i \leq v_n(j)$ contain strictly less than j balls when $\text{Poisson}(x_n)$ balls have been thrown. Since the number of balls in each box are independent Poisson random variables with parameter $x_n z_n$, we have:

$$\mathbb{P}(A_n) \leq \mathbb{P}(\text{Poisson}(x_n z_n) < j)^{v_n(j)}$$

and thus

$$\ln(\mathbb{P}(A_n)) \leq v_n(j) \cdot \ln(\mathbb{P}(\text{Poisson}(x_n z_n) < j)).$$

Note that since $\psi(j) > 0$, the sequence $x_n z_n$ tends to zero as n tends to infinity and that $\ln(1+x) \sim x$ for small x . We get that for n large enough

$$\ln(\mathbb{P}(\text{Poisson}(x_n z_n) < j)) \sim -\frac{(x_n z_n)^j}{j!},$$

and thus for large enough n there exists a constant c such that:

$$\ln(\mathbb{P}(A_n)) \leq -c v_n(j) x_n^j z_n^j = -c e^{n(j\varepsilon' - (j+1)\varepsilon)},$$

and we finally arrive at

$$\mathbb{P}(A_n) \leq e^{-c e^{n(j\varepsilon' - (j+1)\varepsilon)}}.$$

Applying the Borel-Cantelli lemma, we get the result. \square

This lemma will be enough to show the following proposition:

Proposition 3 *For every integer $j \geq 2$ we have*

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

Proof: Define the sequence $y_n = \rho(j)^{-\frac{n}{j}} e^{n\varepsilon}$ with $\varepsilon > \varepsilon'$ and notice that the sequence $\frac{y_{n-1}}{x_n}$ tends to infinity. Thus there exists a natural number n_1 such that, for all $n \geq n_1$, $\frac{y_{n-1}}{x_n} > 3$. We have:

$$\begin{aligned} \mathbb{P}(\text{Poisson}(x_n) \geq \lceil y_{n-1} \rceil) &\leq \mathbb{P}(\text{Poisson}(x_n) \geq 3x_n) \\ &= \mathbb{P}(\text{Poisson}(x_n) - x_n \geq 2x_n) \\ &\leq \mathbb{P}(|\text{Poisson}(x_n) - x_n| \geq 2x_n) \leq \frac{1}{4x_n} \end{aligned}$$

by Chebyshev's inequality. By Borel-Cantelli's lemma we thus derive that for almost all ω there exists a natural number $n_2(\omega)$ such that for all $n \geq n_2(\omega)$ we have that $\text{Poisson}(x_n) < \lceil y_{n-1} \rceil$. Now by Lemma 9 we derive that for almost all ω there exists a natural number $n_3(\omega)$ such that for all $n \geq n_3(\omega)$, we have:

$$N_{\lceil y_{n-1} \rceil, j} \geq 1.$$

Now fix such an ω . Note that there exists a natural number n_4 such that $\forall n \geq n_4$ the sequence y_n is increasing. Let $n_5 \geq \max(n_3(\omega), n_4 + 1)$ and note that for each $m \geq y_{n_4}$ there exists a unique $n \geq n_4$ such that $y_{n-1} < m < y_n$. Now, since $\lceil y_{n-1} \rceil \leq m$, we have $H_{m,j} > n$. Further since $m < y_n$, taking logarithm on both sides, we have that:

$$\ln(m) < \left(\frac{-\ln(\rho(j))}{j} + \varepsilon \right) n.$$

We thus have that:

$$\left(\frac{-\ln(\rho(j))}{j} + \varepsilon \right)^{-1} \ln(m) < n < H_{m,j},$$

and

$$\frac{1}{m} H_{m,j} \geq \left(\frac{-\ln(\rho(j))}{j} + \varepsilon \right).$$

We conclude by letting ε tend to zero. □

2.2 Power regimes

In this section we study the case when $j = j(m) = m^\alpha$ for $\alpha \in (0, 1)$. In the setting of K -ary tries this corresponds to the case when not only the number of words that have to be stored, but also the storage capacity of the nodes tends to infinity. We aim to show that:

Theorem 2 *We have:*

$$\lim_{m \rightarrow \infty} \frac{1}{\ln(m)} H_{m,j} = (1 - \alpha)C^* \quad a.s.$$

We first establish the upper bound. Let $\theta > 0$, $a > \frac{1}{\theta(1-\alpha)}$, and define the sequence

$$x_n := \rho(\theta)^{-\frac{1}{1-\alpha} \frac{n}{\theta}} n^{-a}.$$

In the notation of the last section, we then have:

Lemma 10 *For almost all ω , there exists a natural number $n_0(\omega)$ s.t for all $n \geq n_0(\omega)$ we have $N_{\lfloor x_n \rfloor, x_{n-1}^\alpha}^{(n)} = 0$.*

Proof: Let $k \in \mathbb{N}$ and let $j \geq 2(k+1)$. A straight forward computation shows that then:

$$\mathbb{P}(B(m, p) \geq j) \leq 2^k \cdot \left(\frac{mp}{j} \right)^k.$$

Thus for n large enough such that $x_{n-1}^\alpha \geq 2(\lceil \theta \rceil + 1)$, we have::

$$\begin{aligned} \mathbb{E}[N_{\lfloor x_n \rfloor, x_{n-1}^\alpha}^{(n)}] &\leq \sum_{i=1}^K \sum_k \mathbb{P}\left(B(\lfloor x_n \rfloor, l_{i,k}^{(n)}) \geq \lceil \theta \rceil\right) \\ &\leq 2^{\lceil \theta \rceil} \sum_{i=1}^K \sum_k \left(\frac{\lfloor x_n \rfloor l_{i,k}^{(n)}}{x_{n-1}^\alpha} \right)^\theta. \end{aligned}$$

Let $c_1(\theta) = 2^{\lceil \theta \rceil} \cdot 2v_1(\theta)(w_i(\theta))^t$. By Lemma 2 there exists a natural number n_1 such that for all $n \geq n_1$:

$$\mathbb{E}[N_{\lfloor x_n \rfloor, x_{n-1}^\alpha}^{(n)}] \leq \sum_{i=1}^K c_1(\theta) \left(\frac{x_n}{x_{n-1}^\alpha} \right)^\theta \rho(\theta)^n \leq \sum_{i=1}^K c_2(\theta) n^{-a\theta(1-\alpha)}, \quad (5)$$

where $c_2(\theta) = c_1(\theta)\rho(\theta)^{-\frac{\alpha}{1-\alpha}}$, and we conclude by Borel-Cantelli's lemma. \square

The upper bound then follows in the same way as Proposition 2. We now turn to the proof of the lower bound. Let $\varepsilon' > 0$ and define the sequence:

$$y_n := \exp \left(n \left(\frac{1}{C^*} \frac{1}{(1-\alpha)} + \varepsilon' \right) \right).$$

We then have:

Lemma 11 *For almost all $\omega \in \Omega$ there exists a natural number $n_0(\omega)$, s.t. $N_{\lceil y_{n-1} \rceil, y_n^\alpha} \geq 1$ for all $n \geq n_0(\omega)$.*

Proof: Let $\varepsilon < (1-\alpha)\varepsilon'$. Define A_n , the event that the largest box at generation n contains strictly less than y_n^α balls when $\lceil y_{n-1} \rceil$ balls have been thrown and recall that the largest box at generation n has size larger or equal to $z_n := e^{n(-\frac{1}{C^*} - \varepsilon)}$. We then have:

$$\begin{aligned} \mathbb{P}(A_n) &\leq \mathbb{P}((B(\lceil y_{n-1} \rceil, z_n) > y_n^\alpha) \\ &\leq \mathbb{P}((B(\lceil y_{n-1} \rceil, z_n) + 1)^{-1} \leq y_n^{-\alpha}). \end{aligned}$$

A straight forward computation shows that:

$$\mathbb{E}[(B(m, p) + 1)^{-1}] = \frac{1 - (1-p)^{m+1}}{(m+1)p} \leq \frac{1}{mp},$$

and by Markov's inequality we thus get that

$$\begin{aligned} \mathbb{P}((B(\lceil y_{n-1} \rceil, z_n) + 1)^{-1} \leq y_n^{-\alpha}) &\leq y_n^\alpha \cdot \mathbb{E}[(B(\lceil y_{n-1} \rceil, z_n) + 1)^{-1}] \\ &\leq \frac{y_n^\alpha}{\lceil y_{n-1} \rceil \cdot z_n}. \end{aligned}$$

Taking $c(\alpha) := y_1^\alpha$ we arrive at:

$$\mathbb{P}(A_n) \leq c(\alpha) \cdot e^{(n-1)(\varepsilon + (\alpha-1)\varepsilon')}.$$

and we conclude by Borel-Cantelli's Lemma. \square

The lower bound follows by the usual computations.

2.3 Saturation level of Markovian tries

Let $j \geq 1$ and recall that $G_{m,j}$ denotes the first generation at which there exists a box containing strictly less than j balls, when m balls have been thrown initially. We aim to show that:

Theorem 3 *Let $j \geq 1$, and recall that $C_* = \lim_{\theta \rightarrow -\infty} \frac{\rho(\theta)}{-\rho'(\theta)}$. We then have*

$$\lim_{m \rightarrow \infty} \frac{1}{\ln(m)} G_{m,j} = C_* \quad a.s.$$

We will first study the asymptotic behavior of $G_{m,1}$ and then extend this to the asymptotic behavior of $G_{m,j}$. We will shorthand write G_m for $G_{m,1}$.

In order to establish an upper and a lower bound for G_m it will be useful to study the asymptotic behavior of the number of balls T_n , one needs to throw initially to observe at least one ball in each box of generation n .

As already mentioned in the introduction, this can be interpreted in terms of the coupon collector's problem. We will see each box of generation n as a different sort of coupon, T_n then corresponds to the number of coupons one has to buy to get at least one of each coupon and the probability to buy a sort of coupon is given by the size of the corresponding box.

It will sometimes be convenient to use a Poissonization technique to model T_n . Suppose that the collector continues to buy coupons forever (rather than stopping when having a full collection). Moreover suppose that the coupons are bought at times distributed as the arrival times of a Poisson process with rate 1. The times a coupon of sort i is bought are then the arrival times of independent Poisson processes with rate p_i . Let \mathbf{T}_n denote the waiting time until the collector has completed his collection. We then have

$$\mathbf{T}_n = \max_{k \leq K^n} \mathbf{exp}(l_k^{(n)}),$$

where $\mathbf{exp}(l_k^{(n)})$ denote independent exponential random variables with parameter $l_k^{(n)}$, that is to say $\mathbb{E}[\mathbf{exp}(l_k^{(n)})] = \frac{1}{l_k^{(n)}}$. The connection between \mathbf{T}_n and T_n is then given by

$$\mathbf{T}_n = \sum_{k=1}^{T_n} \mathbf{exp}_k(1),$$

where $\mathbf{exp}_k(1)$ are i.i.d. exponential random variables with parameter 1, independent of T_n . In the same spirit, let T_n^j denote the number of balls one needs to throw initially to observe at least j ball in each box of generation n . Let \mathbf{T}_n^j denote the waiting time until the collector has completed j copies of his collection. We then have

$$\mathbf{T}_n^j = \max_{k \leq K^n} \Gamma(j, l_k^{(n)}),$$

where $\Gamma(j, l_k^{(n)})$ denote independent Gamma random variables with parameter $l_k^{(n)}$ and j . As in the case $j = 1$ we then have

$$\mathbf{T}_n^j = \sum_{k=1}^{T_n^j} \mathbf{exp}_k(1),$$

where $\mathbf{exp}_k(1)$ are i.i.d. exponential random variables with parameter 1, independent of T_n^j .

For the lower bound, we want to find an upper bound for the waiting time \mathbf{T}_n until the collector has completed his collection.

Lemma 12 *Let $\theta < 0$ and define $x_n := e^{n \frac{\ln(\rho(\theta))}{-\theta}} n^\mu$, $\mu > 2$. For almost all ω there exists a natural number $n_0(\omega)$, such that for all $n \geq n_0(\omega)$:*

$$\mathbf{T}_n(\omega) < x_n.$$

Proof: We use the Poissonization technique explained in the previous section. We have:

$$\mathbb{P}(\mathbf{T}_n \geq x_n) = \mathbb{P}(\max_{k \leq K^n} \mathbf{exp}_k^{(n)} \geq x_n) \leq \mathbb{P}(\max_{k \leq K^n} \mathbf{exp}_k(1) \geq \underline{l}^{(n)} x_n),$$

where $\mathbf{exp}_k(1)$ are i.i.d exponential r.v. with parameter 1. Now, for $\theta < 0$ we have:

$$\left(\underline{l}^{(n)}\right)^\theta \leq \sum_k \left(l_{j,k}^{(n)}\right)^\theta = \mathcal{L}_j^{(n)}(\theta).$$

By Lemma 1 there exists a n_0 , such that for all $n \geq n_0$:

$$\left(\underline{l}^{(n)}\right)^\theta \leq \mathcal{L}_j^{(n)}(\theta) \leq 2v_1(\theta)(w_i(\theta))^t \rho(\theta)^n.$$

Let $c_1(\theta) := \frac{1}{\theta} 2 \ln(v_1(\theta)(w_i(\theta))^t)$, by taking logarithm on both sides and some rearrangement, we get that there exists a natural number n_1 such that for all $n \geq n_1$:

$$\frac{n \ln(\rho(\theta))}{\theta} + c_1(\theta) \leq \ln(\underline{l}^{(n)}),$$

and thus

$$e^{n \frac{\ln(\rho(\theta))}{\theta} + c_1(\theta)} \leq \underline{l}^{(n)},$$

for all $n \geq n_1$. Let $c_2(\theta) = e^{c_1(\theta)}$, we deduce that for n large enough:

$$\mathbb{P}(\mathbf{T}_n \geq x_n) \leq \mathbb{P}(\max_{k \leq K^n} \mathbf{exp}_k(1) \geq c_2(\theta) n^\mu).$$

Now, recall that

$$\max_{k \leq K^n} \mathbf{exp}_k(1) - n \ln(K)$$

converges in distribution to the standard Gumbel distribution $\mathbf{G}(1)$ as n tends to infinity. Thus for large enough n , we have:

$$\begin{aligned} \mathbb{P}(\mathbf{T}_n \geq x_n) &\leq 2 \cdot \mathbb{P}(\mathbf{G}(1) \geq c_2(\theta) n^\mu - n \ln(K)) \\ &= 2 \cdot \left(1 - \exp\left(-e^{-(c_2(\theta) n^\mu - n \ln(K))}\right)\right) \\ &\leq 2 \cdot e^{-(c_2(\theta) n^\mu - n \ln(K))}. \end{aligned}$$

We then conclude by Borel-Cantelli's lemma. \square

Since the times at which a ball is thrown are the arrival times of independent Poisson processes with rate 1, the number of balls thrown up to time x_n is Poisson distributed with parameter x_n . Applying Lemma 12, we thus get, that:

Corollary 2 For almost all ω it exists an integer number $n_0(\omega)$ such that $\forall n \geq n_0(\omega)$ there exists no box at generation n containing no balls when Poisson(x_n) balls have been thrown.

We then arrive at:

Proposition 4 We have:

$$\frac{1}{\ln(m)} G_{m,j} \geq C_* + O\left(\frac{\ln \ln(m)}{\ln(m)}\right) \quad a.s.$$

as m tends to infinity.

Note that $T_n^j \leq T_n^{(1)} + \dots + T_n^{(j)}$ a.s., where $T_n^{(1)}, \dots, T_n^{(j)}$ denote independent copies of T_n . Further we have $G_{\lfloor \frac{m}{j} \rfloor} \geq n$ a.s. which implies that $\lfloor \frac{m}{j} \rfloor > T_n$ a.s. and thus $T_n^j \leq m$ a.s. But then $G_{m,j} \geq n$ a.s. and we conclude that $G_{m,j} \geq G_{\lfloor \frac{m}{j} \rfloor}$ a.s. It will be thus enough to show the statement for G_m .

Proof of Proposition 4: Define the sequence $y_n = e^{n \frac{\ln(\rho(\theta))}{-\theta}} n^{\mu'}$ with $\mu' < \mu$ and note that the sequence $\frac{y_n}{x_n}$ tends to zero. Thus there exists a natural number n_0 such that for all $n \geq n_0$, we have $\frac{y_n}{x_n} < \frac{1}{2}$. Thus:

$$\begin{aligned} \mathbb{P}(\text{Poisson}(x_n) \leq \lfloor y_n \rfloor) &\leq \mathbb{P}(\text{Poisson}(x_n) \leq \frac{x_n}{2}) \\ &= \mathbb{P}(\text{Poisson}(x_n) - x_n \leq -\frac{x_n}{2}) \\ &\leq \mathbb{P}(|\text{Poisson}(x_n) - x_n| \geq \frac{x_n}{2}) \leq \frac{4}{x_n} \end{aligned}$$

by Chebyshev's inequality. By Borel-Cantelli's lemma we thus derive that for almost all ω there exists a natural number $n_1(\omega)$ such that for all $n \geq n_1(\omega)$ we have that $\text{Poisson}(x_n) > \lfloor y_{n-1} \rfloor$.

Now, let $M_m^{(n)}$ denote the number of boxes at generation n containing zero balls when m balls have been thrown. By Corollary 2 we deduce that for almost all ω there exists a natural number $n_2(\omega)$, such that for all $n \geq n_2(\omega)$, we have $M_{y_{n-1}}^{(n)} = 0$. Fix such an ω and note that since the sequence (y_n) is increasing, for each $m \geq y_{n_2(\omega)}$ there exists a unique $n \geq n_2(\omega)$ such that $y_{n-1} < m \leq x_n$. Since $\lfloor y_{n-1} \rfloor \leq m$ we have $G_m > n - 1$. Now, by taking logarithm on both sides of the inequality $m \leq x_n$ and some rearrangement, we get that:

$$n \geq \frac{-\theta}{\ln(\rho(\theta))} \ln(m) - \mu \ln(n).$$

From the inequality $m > x_{n-1} \geq e^{(n-1) \frac{\ln(\rho(\theta))}{-\theta}}$, we get that $\frac{\theta}{\ln(\rho(\theta))} \ln(m) \geq n$ and thus:

$$G_m > n - 1 \geq \frac{-\theta}{\ln(\rho(\theta))} \ln(m) - \mu \ln\left(\frac{-\theta}{\ln(\rho(\theta))} \ln(m)\right) - 1, \quad (6)$$

and we are left to show that:

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

Indeed by l'Hôpital's rule, we have that:

$$\lim_{\theta \rightarrow -\infty} \frac{-\theta}{\ln(\rho(\theta))} = \lim_{\theta \rightarrow -\infty} \frac{\rho(\theta)}{-\rho'(\theta)} = C_*,$$

and we conclude. \square

Remark 3 *We could have gained the same result by applying Theorem 6.E in Barbour et al. (1992).*

For the proof of the upper bound, we want to find an lower bound for the number of balls T_n one needs to throw initially to observe at least one ball in each box of generation n .

Lemma 13 *Let $\theta < 1$ and define $x_n := e^{n\left(\frac{-\rho'(\theta)}{\rho(\theta)} - \varepsilon'\right)}$, $\varepsilon' > 0$. For almost all ω there exists a natural number $n_0(\omega)$, such that for all $n \geq n_0(\omega)$:*

$$\mathbf{T}_n(\omega) > x_n.$$

Proof: We use the Poissonization technique explained in the last section. We have:

$$\mathbb{P}(\mathbf{T}_n \leq x_n) = \mathbb{P}\left(\max_{k \leq K^n} \mathbf{exp}(l_k^{(n)}) \leq x_n\right) \leq \mathbb{P}(\mathbf{exp}(1) \leq \underline{l}^{(n)} x_n).$$

Now, by Lemma 4 there exists for each $\varepsilon > 0$ a natural number $n_0(\varepsilon)$ such that $\underline{l}^{(n)} \leq e^{n\left(\frac{\rho'(\theta)}{\rho(\theta)} + \varepsilon\right)}$ for all $n \geq n_0(\varepsilon)$. Let $\varepsilon < \varepsilon'$. We then have for each $n \geq n_0(\varepsilon)$:

$$\mathbb{P}(\mathbf{T}_n \leq x_n) \leq \mathbb{P}(\mathbf{exp}(1) \leq 2e^{-n\varepsilon}) = 1 - \exp(-e^{-n(\varepsilon' - \varepsilon)}) \leq e^{-n(\varepsilon' - \varepsilon)}.$$

We conclude by Borel-Cantelli's lemma. \square

Since the times at which a ball is thrown are the arrival times of independent Poisson processes with rate 1, the number of balls thrown up to time x_n is Poisson distributed with parameter x_n . Applying Lemma 13, we thus get that:

Corollary 3 *For almost all ω it exists a natural number $n_0(\omega)$ such that $\forall n \geq n_0(\omega)$ there is at least one box at generation n containing zero balls when Poisson(x_n) balls have been thrown.*

We now tackle the proof of the main result in this section.

Proposition 5 *We have:*

$$\limsup_{m \rightarrow \infty} \frac{1}{\ln(m)} G_m \leq C_* \quad a.s.$$

Proof: Define the sequence $y_n = e^{n\left(\frac{-\rho'(\theta)}{\rho(\theta)} - \varepsilon\right)}$ with $\varepsilon' > \varepsilon$ and notice that the sequence $\frac{y_n}{x_n}$ tends to infinity. Thus there exists a natural number n_1 such that for all $n \geq n_1$ $\frac{y_n}{x_n} > 3$. We have:

$$\begin{aligned} \mathbb{P}(\text{Poisson}(x_n) \geq \lfloor y_n \rfloor) &\leq \mathbb{P}(\text{Poisson}(x_n) \geq 3x_n) \\ &= \mathbb{P}(\text{Poisson}(x_n) - x_n \geq 2x_n) \\ &\leq \mathbb{P}(|\text{Poisson}(x_n) - x_n| \geq 2x_n) \leq \frac{1}{4x_n}, \end{aligned}$$

by Chebyshev's inequality. By Borel-Cantelli we thus derive that for almost all ω there exists a natural number $n_2(\omega)$ such that for all $n \geq n_2(\omega)$ we have that $\text{Poisson}(x_n) < \lfloor y_n \rfloor$. Now by Corollary 3 we derive that for almost all ω there exists a natural number $n_3(\omega)$ such that for all $n \geq n_3(\omega)$, we have:

$$M_{\lfloor y_n \rfloor, j} = 0.$$

Now fix such an ω . Note that there exists a natural number n_4 such that $\forall n \geq n_4$ the sequence y_n is increasing. Let $n_5 \geq \max(n_3(\omega), n_4)$ and note that for each $m \geq y_{n_4}$ there exists a unique $n \geq n_4$ such that $y_n < m < y_{n+1}$. Now, since $\lfloor y_{n-1} \rfloor \leq m$, we have $G_m < n$. Further since $y_n < m$, taking logarithm on both sides, we have that:

$$\frac{1}{\left(\frac{-\rho'(\theta)}{\rho(\theta)} - \varepsilon'\right)} \ln(m) \geq n > G_m.$$

We thus derive that

$$\limsup_{m \rightarrow \infty} \frac{1}{\ln(m)} G_m \leq \frac{1}{\frac{-\rho'(\theta)}{\rho(\theta)} - \varepsilon},$$

and we conclude by letting ε tend to zero and θ tend to $-\infty$. \square

Now, let $j \geq 2$ and note that the first generation when there exists a box containing no ball when m balls have been thrown is larger than the first generation at which there exists a box containing less than j balls when m balls have been thrown. That is $G_{m,j} \leq G_m$ and we derive that:

Proposition 6 *Let $j \geq 1$. We have:*

$$\liminf_{m \rightarrow \infty} \frac{1}{\ln(m)} G_{m,j} \leq C_* \quad a.s.$$

3 Occupancy scheme in random environment

3.1 Random probability cascades

Consider a random transition matrix $A = (p_{ij})$. To the box with label u , of type i , we will associate an independent copy of $A_u = (p_{ij}(u))$ of A . The length of the box $u = (i_1, \dots, i_n)$ is then given by some multiplicative cascade, that is:

$$l_u^{(n)} = p_{i_1 i_2}(i_1) \times \dots \times p_{i_{n-1} i_n}((i_1, i_2, \dots, i_{n-1})). \quad (7)$$

Let $(l_{i,j,k}^{(n)})_k$ denote the sequence of length of the boxes of type j at generation n , when the first box was of type i . The process $(Z_i^n)_n = (Z_{i1}^n, \dots, Z_{iK}^n)$, with

$$Z_{ij}^n = \sum_k \delta_{-\ln(l_{i,j,k}^{(n)})},$$

is then a multitype branching random walk. Let $|Z_{ij}| := \int_{\mathbb{R}} Z_{ij}(dx)$ denote the total mass. We will assume that the embedded Galton-Watson process $((|Z_{i1}^n|, \dots, |Z_{iK}^n|))$ is positive regular, that is

$$\text{the matrix } (\mathbb{P}(|Z_{ij}| > 0))_{ij} \text{ is positive regular.} \quad (8)$$

Let $\theta \in \mathbb{R}$ and let us introduce the Laplace transform of the intensity measure of Z_{ij} :

$$m_{ij}(\theta) = \mathbb{E} \left[\int e^{-\theta x} Z_{ij}(dx) \right] = \mathbb{E} [p_{ij}^\theta].$$

Let

$$L = \bigcap_{i,j} \{\theta \in \mathbb{R} : m_{ij}(\theta) < \infty\}$$

and note that L is an interval since m_{ij} is decreasing in θ . Let us introduce $M(\theta)$, the matrix with entries $\mathbb{E}[p_{ij}^\theta]$, where we agree that if $p_{ij} = 0$, then $p_{ij}^\theta = 0$ even if $\theta \leq 0$. Note that the entries of $M^n(\theta)$ are given by

$$m_{ij}^n(\theta) = \mathbb{E} \left[\int e^{-\theta x} Z_{ij}^n(dx) \right] = \mathbb{E} \left[\sum_k (l_{ij,k}^{(n)})^\theta \right].$$

Condition (8) implies that $M(\theta)$ is positive regular for each $\theta \in L$, and thus the Perron-Frobenius theorem applies and $M(\theta)$ possesses a maximum modulus eigenvalue $\varrho(\theta)$, a strictly positive left-eigenvector $w(\theta)$ and a strictly positive right-eigenvector $v(\theta)$ with eigenvalue $\varrho(\theta)$, such that we have $(w(\theta))^t v(\theta) = 1$. Moreover $\ln(\varrho(\theta))$ is convex (see Kingman (1961)) and analytic on L (see Biggins and Rahimzadeh Sani (2004)). We will need to assume some even stronger condition on $\varrho(\theta)$. We will assume that:

$$\ln(\varrho(\theta)) \text{ is strictly convex.} \quad (9)$$

Similar to the function $\psi(\theta)$ in the last section, the function

$$f(\theta) := \ln(\varrho(\theta)) - \theta \frac{\varrho'(\theta)}{\varrho(\theta)},$$

will play a crucial role in our analysis. Note that $f'(\theta) = -\theta \cdot \ln(\varrho(\theta))''$. Thus f is strictly decreasing on the interval $(0, \infty) \cap L$ and strictly increasing on $(-\infty, 0) \cap L$. Since $f(0) > 0$ the function f is thus positive on some open interval. Let us define

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

The function f is then strictly positive on (θ_*, θ^*) .

Let (\mathcal{F}_n) denote the natural filtration of the the multitype branching random walk. Following Biggins and Rahimzadeh Sani (2005) one can define a remarkable martingale for each $\theta \in L$, with respect to the filtration \mathcal{F}_n :

$$W_i^n(\theta) := \sum_{j=1}^K \frac{v_j(\theta)}{v_i(\theta)} \varrho(\theta)^{-n} \cdot \sum_k (l_{ij,k}^{(n)})^\theta,$$

where $v_i(\theta)$ denotes the i -th entry of the right-eigenvector $v(\theta)$ with eigenvalue $\varrho(\theta)$. We then have:

Lemma 14 *For each $\theta \in (\theta_*, \theta^*)$ the martingale $W_i^n(\theta)$ is bounded in $L^\alpha(\mathbb{P})$ for some $\alpha > 1$. It converges almost surely and in mean and its terminal value*

$$W_i(\theta) := \lim_{n \rightarrow \infty} W_i^n(\theta)$$

is a.s. strictly positive.

Proof: We want to apply Theorem 2 in Biggins and Rahimzadeh Sani (2005). We need to check that for all $\theta \in (\theta_*, \theta^*)$ there is an $\alpha > 1$ such that we have $\varrho(\alpha\theta) < \varrho(\theta)^\alpha$ and $\max_i \mathbb{E}[(W_i^n(\theta))^\alpha] < \infty$. Consider the function $g(\theta) = \frac{1}{\theta} \ln(\varrho(\theta))$ and note that $g'(\theta) = -\frac{1}{\theta^2} f(\theta)$. Thus g is decreasing on (θ_*, θ^*) . For $\alpha > 1$ small enough we thus have

$$\frac{\ln(\varrho(\alpha\theta))}{\alpha\theta} < \frac{\ln(\varrho(\theta))}{\theta}$$

and thus $\varrho(\alpha\theta) < \varrho(\theta)^\alpha$. For the second criterion note that

$$\mathbb{E}[(W_i^n(\theta))^\alpha] \leq \sum_{j=1}^K \left(\frac{v_j(\theta)}{v_i(\theta)} \varrho(\theta)^{-1} \right)^\alpha \mathbb{E}[p_{ij}^{\alpha\theta}] < \infty \quad \forall i, j,$$

by Jensen's inequality and the fact that $\alpha\theta \in (\theta_*, \theta^*)$ for $\alpha > 1$ small enough. By Theorem 2 in Biggins and Rahimzadeh Sani (2005) we thus get the convergence of W_i^n almost surely and in mean. For the a.s. strictly positivity note that

$$\mathbb{E}[(W_i^n(\theta))^\alpha] = \sum_{j=1}^K \frac{v_j(\theta)}{v_i(\theta)} \varrho(\theta)^{-n} m_{ij}^n(\theta).$$

By the Perron-Frobenius theorem $\lim_{n \rightarrow \infty} \varrho(\theta)^{-n} m_{ij}^n(\theta) = v_i(\theta) w_j(\theta)$. Thus $\lim_{n \rightarrow \infty} \mathbb{E}[W_i^n(\theta)] = \mathbb{E}[W_i(\theta)] = 1$ and $\mathbb{P}(W_i(\theta) = 0) = \beta_i < 1$. Moreover $(\beta_1, \dots, \beta_K)$ is a fixed point of the multivariate generating function of the embedded Galton-Watson process and thus β_i is the extinction probability of the process started from type i and thus $\beta_i = 0$ a.s. \square

The following result (Corollary 2 in Biggins and Rahimzadeh Sani (2005)), will play the role of Corollary 1 in the previous section. It will help us to control the asymptotic behavior of the length of boxes at generation n , as n tends to infinity.

Lemma 15 *For all $a > b \in \mathbb{R}$ and θ in a compact subset of (θ_*, θ^*) we have:*

$$\sqrt{n} e^{-nf(\theta)} \# \left\{ k : l_{i,j,k}^{(n)} \in \left[e^{n \frac{\varrho'(\theta)}{\varrho(\theta)} - a}, e^{n \frac{\varrho'(\theta)}{\varrho(\theta)} - b} \right] \right\} \rightarrow \frac{v_i(\theta) w_j(\theta) W_i(\theta) e^{a\theta} - e^{b\theta}}{\sqrt{2\pi f''(\theta)} \theta}$$

almost surely as n tends to infinity.

As in the last section, we will further sometimes need to control the size of the smallest and the largest box at generation n , when n tends to infinity. Note that $\varrho(\theta)$ is decreasing in L since the entries $m_{ij}(\theta)$ are decreasing and that the strict convexity of $\varrho(\theta)$ thus entails that $\varrho'(\theta) < 0$ on L . Define the constants

$$\zeta^* := \lim_{\substack{\theta \rightarrow \theta^* \\ \theta \leq \theta^*}} \frac{\varrho(\theta)}{-\varrho'(\theta)}, \quad \text{and} \quad \zeta_* := \lim_{\substack{\theta \rightarrow \theta_* \\ \theta > \theta_*}} \frac{\varrho(\theta)}{-\varrho'(\theta)}.$$

The strict convexity of $\ln(\varrho(\theta))$ implies that $\frac{\varrho(\theta)}{-\varrho'(\theta)}$ is strictly increasing on L . Recall moreover that $\rho(\theta)$ is analytic on L . For $\theta^* < \infty$ we thus have $0 < \zeta^* < \infty$ since $0 \in L$ and $\frac{\varrho(0)}{-\varrho'(0)} > 0$. Moreover if

$$-\infty < \theta_* < 0 \quad \text{and} \quad \lim_{\substack{\theta \rightarrow \theta_* \\ \theta > \theta_*}} f(\theta) = 0, \quad (10)$$

then we have $0 < \zeta_* < \infty$ since then $\lim_{\substack{\theta \rightarrow \theta_* \\ \theta > \theta_*}} \frac{\varrho(\theta)}{-\varrho'(\theta)} = \lim_{\theta > \theta_*} \frac{\ln(\varrho(\theta))}{\theta}$.

Corollary 4 *Suppose that $\theta^* < \infty$ and let $\bar{l}_i^{(n)}$ denote the size of the largest box at generation n , when the first box was of type i . We then have:*

$$\lim_{n \rightarrow \infty} n^{-1} \cdot \ln(\bar{l}_i^{(n)}) = -\frac{1}{\zeta_*} \quad a.s.$$

Proof: We first show the upper bound. Let $\theta \in (0, \theta^*)$. By Lemma 14 we have for n large enough:

$$\varrho(\theta)^{-n} (\bar{l}_i^{(n)})^\theta \leq \sum_{j=1}^K \frac{v_j(\theta)}{v_i(\theta)} \varrho(\theta)^{-n} \cdot (\bar{l}_i^{(n)})^\theta \leq W_i^n(\theta) \leq 2W_i(\theta), \quad a.s.$$

By some rearrangement we thus get:

$$\limsup_{n \rightarrow \infty} n^{-1} \cdot \ln(\bar{l}_i^{(n)}) \leq \frac{\ln(\varrho(\theta))}{\theta} \quad a.s.$$

We conclude by letting θ tend to θ^* . For the lower bound note that by Lemma 15 we have $e^{n \frac{\varrho'(\theta)}{\varrho(\theta)}} \leq \bar{l}_i^{(n)}$ a.s., and we conclude by some rearrangement and letting θ tend to θ^* . \square

For the size of the smallest box one shows in the same way that:

Corollary 5 *Let (10) hold and let $l_i^{(n)}$ denote the smallest box at generation n , when the type of the first box was of type i . We then have*

$$\lim_{n \rightarrow \infty} n^{-1} \cdot \ln(l_i^{(n)}) = -\frac{1}{\zeta_*}, \quad a.s.$$

3.2 Study of the height

Recall that $H_{m,j}$ denotes the first generation of boxes at which all the boxes contain strictly less than j balls when m balls have been thrown independently. We will see that there is a phase transition in the limiting behavior of $H_{m,j}$ as m tends to infinity, depending on the values of j . We aim to show that:

Theorem 4 *Suppose that conditions (8) and (9) hold. We then have:*

1. *For every $j \in (\theta_*, \theta^*)$ we have*

$$\lim_{m \rightarrow \infty} \frac{1}{\ln(m)} H_{m,j} = \frac{j}{-\ln(\varrho(j))} + O\left(\frac{\ln \ln(m)}{\ln(m)}\right) \quad a.s.$$

2. *If $\theta^* < \infty$ we have for every $j \geq \theta^*$ that*

$$\lim_{m \rightarrow \infty} \frac{1}{\ln(m)} H_{m,j} = \zeta_* \quad a.s.$$

The phase transition in the limiting behavior of the height has first been observed by Joseph (2010).

Remark 4 As in the previous sections we assume that we start from a box of type 1. We will thus drop the subscript 1 and shorthand write $l_{i,k}^{(n)}$ for $l_{1i,k}^{(n)}$ and $\underline{l}^{(n)}$ for $\underline{l}_1^{(n)}$ and W for W_1 .

We first show the upper bound. For $\theta \in L$ with $\theta > 0$ such that $\lceil \theta \rceil \leq j$ and $a > \frac{1}{\theta}$, let us define the sequence

$$x_n = \varrho(\theta)^{-\frac{n}{\theta}} n^{-a}.$$

We then have:

Lemma 16 For almost all ω , there exists a natural number $n_0(\omega)$ s.t for all $n \geq n_0(\omega)$, there exists no box at generation n containing j or more balls when $\lfloor x_n \rfloor$ balls have been thrown.

Proof: Let $N_{m,j}^{(n)}$ denote the number of boxes at generation n containing j or more balls when m balls have been thrown. Conditionally on \mathcal{F}_n , the number of balls in a box at generation n of size l when m balls have been thrown has distribution $B(m, l)$. Similarly to the proof of Lemma 8, we thus have:

$$\begin{aligned} \mathbb{E}[N_{\lfloor x_n \rfloor, j}^{(n)} | \mathcal{F}_n] &= \sum_{i=1}^K \sum_k \mathbb{P}\left(B\left(\lfloor x_n \rfloor, l_{i,k}^{(n)}\right) \geq j\right) \\ &\leq \sum_{i=1}^K \sum_k \left(\lfloor x_n \rfloor l_{i,k}^{(n)}\right)^\theta \\ &\leq c_1(\theta) x_n^\theta \varrho(\theta)^n \sum_{i=1}^K \sum_k \varrho(\theta)^{-n} \frac{v_i(\theta)}{v_1(\theta)} (l_{i,k}^{(n)})^\theta \\ &= c_1(\theta) n^{-a\theta} W^{(n)}(\theta), \end{aligned}$$

where $c_1(\theta) := \left(\max_{1 \leq i \leq K} \left(\frac{v_i(\theta)}{v_1(\theta)}\right)\right)$. We thus derive that:

$$\mathbb{E}[N_{x_n, j}^{(n)}] \leq c_1(\theta) n^{-a\theta} \mathbb{E}[W^{(n)}(\theta)] = c_1(\theta) n^{-a\theta}.$$

We finally arrive at

$$\mathbb{E}\left[\sum 1_{\{N_{\lfloor x_n \rfloor, j}^{(n)} \geq 1\}}\right] \leq \mathbb{E}\left[\sum N_{\lfloor x_n \rfloor, j}^{(n)}\right] \leq \sum c_1(\theta) n^{-a\theta} < \infty,$$

and we conclude by Borel-Cantelli. □

In the same way as in the proof of Proposition 2, we derive that:

Proposition 7 For every integer $\theta \in L$ we have

$$\frac{1}{\ln(m)} H_{m,j} \leq \frac{\theta}{-\ln(\varrho(\theta))} + O\left(\frac{\ln \ln(m)}{\ln(m)}\right) \quad a.s.$$

as m tends to infinity.

Remark 5 The function $\theta \rightarrow \frac{\theta}{-\ln(\varrho(\theta))}$ is decreasing on (θ_*, θ^*) .

For the lower bound let first $j \in (\theta_*, \theta^*)$ and recall that $H_{m,j} \geq n$, if at generation n there is at least one box containing j or more balls. Let $a > \frac{1}{2j}$ and define the sequence

$$x_n = \varrho(\theta)^{-\frac{n}{j}} n^a.$$

As in the previous section we then have:

Lemma 17 *For almost all ω there exists a natural number $n_0(\omega)$, such that for all $n \geq n_0(\omega)$ there is at least one box containing j or more balls when $\text{Poisson}(x_n)$ balls have been thrown.*

Proof: Let

$$z_n := e^{n \frac{\varrho'(j)}{\varrho(j)}}.$$

and let M_n denote the number of boxes at generation n of type 1 with size in the interval $[z_n, 2z_n]$. Let

$$Z(j) = \frac{v_1(j)w_1(j)W(j)}{\sqrt{2\pi f''(j)}} \cdot \frac{1 - 2^{-j}}{j},$$

and $v_n(j) = \frac{1}{2}Z(j)e^{nf(j)}n^{-\frac{1}{2}}$. From Lemma 15 we know that a.s. there exists a natural number n_0 such that for all $n \geq n_0$, we have $M_n \geq v_n(j)$. We can thus a.s. consider the first $v_n(j)$ boxes in M_n , say $b_1(n), \dots, b_{v_n(j)}(n)$ and denote their size with $l_1(n), \dots, l_{v_n(j)}(n)$. We place an imaginary box $\mathbf{b}_i(n)$ in $b_i(n)$ for $1 \leq i \leq v_n(j)$, each of size exactly z_n . If a ball falls into the box $b_i(n)$ it is placed in the imaginary box $\mathbf{b}_i(n)$ with probability $\frac{z_n}{l_i}$.

Let A_n denote the event that the boxes $\mathbf{b}_i(n)$ for $1 \leq i \leq v_n(j)$ contain strictly less than j balls when $\text{Poisson}(x_n)$ balls have been thrown. Since conditioned on \mathcal{F}_n the number of balls in each box of generation n are independent Poisson random variables with parameter $x_n z_n$, we have:

$$\mathbb{P}(A_n | \mathcal{F}_\infty) \leq \mathbb{P}(\text{Poisson}(x_n z_n) < j)^{v_n(j)}.$$

We then finish the proof in the same way as the proof of Lemma 9. □

Performing the same computations as in the proof of Proposition 3, we arrive at:

Proposition 8 *Suppose that (8) and (9) hold. For every integer $j \in (\theta_*, \theta^*)$ we have*

$$\frac{1}{\ln(m)} H_{m,j} \geq \frac{j}{-\ln(\varrho(j))} + O\left(\frac{\ln \ln(m)}{\ln(m)}\right) \quad a.s.$$

as m tends to infinity.

We now turn to the second case. Suppose that $\theta^* < \infty$ and $j \geq \theta^*$. Let $0 < \varepsilon' < \varepsilon$ and define the sequences

$$x_n := e^{n(\frac{1}{\zeta^*} + \varepsilon)} \quad \text{and} \quad y_n = e^{n(-\frac{1}{\zeta^*} - \varepsilon')}.$$

We then have:

Lemma 18 *For almost all ω there exists a natural number $n_0(\omega)$, s.t. $N_{\lceil x_n \rceil, j} \geq 1$ for all $n \geq n_0(\omega)$.*

Proof: Let \bar{b}_n denote the largest box at generation n and recall that $\bar{l}^{(n)}$ denotes its size. We place an imaginary box \mathbf{b}_n of size $\bar{l}^{(n)} \wedge y_n$ inside the largest box. When a ball falls into \bar{b}_n it is placed in \mathbf{b}_n with probability $\frac{y_n}{\bar{l}^{(n)}}$. Now, let

$$A_n := \{\bar{l}^{(n)} \geq y_n\}$$

and let B_n denote the event that the box \mathbf{b}_n contains strictly less than j balls when $\lceil x_n \rceil$ balls have been thrown. As in the proof of Lemma 11, we have:

$$\begin{aligned} \mathbb{P}(A_n \cap B_n) &\leq \mathbb{P}(B_n | A_n) \\ &\leq \mathbb{P}((B(\lceil x_n \rceil, y_n) + 1)^{-1} \leq j^{-1}) \\ &\leq \frac{j}{x_n \cdot y_n}, \end{aligned}$$

and we arrive at

$$\mathbb{P}(A_n \cap B_n) \leq j e^{-n(\varepsilon - \varepsilon')}.$$

We conclude by Borel-Canteli's lemma and the fact that $\mathbb{P}(A_n) = 1$ by Corollary 4. \square

Performing the usual computations, we arrive at:

Proposition 9 *Suppose that (8), (9) hold, that $\theta^* < \infty$ and $j \geq \theta^*$. We then have*

$$\liminf_{m \rightarrow \infty} \frac{1}{\ln(m)} H_{m,j} \geq \zeta^* \quad a.s.$$

3.3 Study of the saturation level

We aim to show that:

Theorem 5 *Suppose that (8), (9) and (10) hold. We then have*

$$\lim_{m \rightarrow \infty} \frac{1}{\ln(m)} G_{m,j} = \zeta_* \quad a.s.$$

We first tackle the proof of the lower bound. Let $\varepsilon' > 0$ and define the sequence

$$x_n = e^{n(\frac{1}{\zeta_*} + \varepsilon')},$$

where, ζ_* is the constant appearing in the limit behavior of the smallest box. In the same notation as in the previous sections we then have:

Lemma 19 *For almost all ω there exists a natural number $n_0(\omega)$, such that for all $n \geq n_0(\omega)$:*

$$\mathbf{T}_n(\omega) < x_n.$$

Proof: Let $0 < \varepsilon < \varepsilon'$. Define A_n , the event that the smallest box $\underline{l}^{(n)}$ of generation n is larger than $e^{n(-\frac{1}{\zeta_*} - \varepsilon)}$ and B_n , the event that $\mathbf{T}_n \geq x_n$. Let $\mathbf{exp}_k(1)$ for $k \leq K^n$ denote independent exponential

random variables with parameter 1. We then have:

$$\begin{aligned}
\mathbb{P}(B_n \cap A_n) &\leq \mathbb{P}(B_n|A_n) \\
&\leq \mathbb{P}(\max_{k \leq K^n} \mathbf{exp}_k(1) \geq e^{n(-\frac{1}{\zeta_*} - \varepsilon)} x_n) \\
&\leq 2 \cdot \mathbb{P}(\mathbf{G}(1) \geq e^{n(-\frac{1}{\zeta_*} - \varepsilon)} x_n - n \ln(K)) \\
&\leq 2 \cdot \exp(-(e^{n(\varepsilon' - \varepsilon)} - n \ln(K))),
\end{aligned}$$

and we conclude by the fact that $\mathbb{P}(A_n) = 1$ by Corollary 5 and Borel-Cantelli's lemma. \square

Performing the same calculations and generalizations as in the previous section, we arrive at:

Proposition 10 *Suppose that (8), (9) and (10) hold. We have:*

$$\liminf_{m \rightarrow \infty} \frac{1}{\ln(m)} G_{m,j} \geq \zeta_* \quad a.s.$$

For the upper bound let $\varepsilon' > 0$ and define the sequence

$$y_n = e^{n(\frac{1}{\zeta_*} - \varepsilon')}.$$

We then have:

Lemma 20 *For almost all ω there exists a natural number $n_0(\omega)$, such that for all $n \geq n_0(\omega)$:*

$$\mathbf{T}_n(\omega) > y_n.$$

Proof: Let $0 < \varepsilon' < \varepsilon$. Define A_n , the event that the smallest box $\underline{l}^{(n)}$ of generation n is smaller than $e^{n(-\frac{1}{\zeta_*} + \varepsilon)}$ and B_n , the event that $\mathbf{T}_n \leq y_n$. We then have:

$$\begin{aligned}
\mathbb{P}(B_n \cap A_n) &\leq \mathbb{P}(B_n|A_n) \\
&\leq \mathbb{P}(\mathbf{exp}_k(1) \leq e^{n(\frac{1}{\zeta_*} + \varepsilon)} y_n) \\
&\leq 1 - \exp(-e^{-n(\varepsilon - \varepsilon')}) \\
&\leq e^{-n(\varepsilon - \varepsilon')},
\end{aligned}$$

and we conclude by the fact that $\mathbb{P}(A_n) = 1$ by Corollary 5 and Borel-Cantelli's lemma. \square

The upper bound then easily follows.

Acknowledgements

I would like to thank Jean Bertoin for introducing me to this problem and for his advice and support. I would also like to thank two anonymous referees for carefully reading the first draft of this work and for their helpful comments.

References

- R. B. Ash. *Information Theory*. Dover Publications, New York, 2012.
- A. D. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Oxford University Press, Oxford, 1992.
- J. Bertoin. Asymptotic regimes for the occupancy scheme of multiplicative cascades. *Stochastic Processes and their Applications*, 118:1586–1605, 2008.
- J. D. Biggins. Uniform convergence of martingales in the branching random walk. *Annals of Probability*, 20(1):137–151, 1992.
- J. D. Biggins and A. Rahimzadeh Sani. Extended Perron-Frobenius results. 2004. available at <http://biggins.staff.shef.ac.uk/epfr.pdf>.
- J. D. Biggins and A. Rahimzadeh Sani. Convergence results on multitype multivariate branching random walks. *Advances in Applied Probability*, 37(3):681–705, 2005.
- A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, 2010.
- L. Devroye. A probabilistic analysis of the height of tries and of the complexity of triesort. *Acta Informatica*, 21:229–237, 1984.
- L. Devroye. Universal asymptotics for random tries and patricia trees. *Algorithmica*, 42:11–29, 2005.
- M. Drmota. *Random Trees*. Springer, Wien, New York, 2009.
- P. Flajolet and J.-M. Steyaert. A branching process arising in dynamic hashing, trie searching and polynomial factorization. In *Lecture Notes in Computer Science*, volume 140. 1982.
- A. Gnedin. The Bernoulli sieve. *Bernoulli*, 10(1):79–96, 2004.
- A. Gnedin, B. Hansen, and J. Pitman. Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys*, (4):146–171, 2007.
- A. Gnedin, A. Iksanov, and A. Marynych. The Bernoulli sieve: an overview. In *Proceedings of the 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA10), Discrete Math. Theor. comput. Sci. AM*, pages 329–341, 2010.
- L. Holst. On Birthday, Collector's, Occupancy and other Classical Urn Problems. *International Statistical Review*, 54(1):15–27, 1986.
- N. L. Johnson and S. Kotz. *Urn models and their application*. John Wiley and Sons, New York, London, Sydney, 1977.
- A. Joseph. A Phase Transition for the Heights of a Fragmentation Tree. *Random Structures and Algorithms*, 39:247–274, 2010.
- J. F. C. Kingman. A convexity property of positive matrices. *The Quarterly Journal of Mathematics*, 12(1):283–284, 1961.

- V. F. Kolchin, B. A. Sevast'yanov, and V. P. Chistyakov. *Random Allocations*. John Wiley and Sons, New York, London, Sydney, 1978.
- B. Pittel. Asymptotical growth of a class of random trees. *Annals of Probability*, 13(2):414–427, 1985.
- B. Pittel. Path in a Random Digital Tree: Limiting Distributions. *Advances in Applied Probability*, 18(1): 139–155, 1986.
- P. Robert and F. Simatos. Occupancy scheme associated to Yule processes. *Advances in Applied Probability*, 41(2):600–622, 2009.
- B. Rosen. On the coupon collector's waiting time. *The Annals of Mathematical Statistics*, 41(6):1952–1969, 1970.
- W. Szpankowski. On the Height of Digital Trees and related Problems. *Algorithmica*, 6:256–277, 1991.
- W. Szpankowski. *Average case analysis of algorithms on sequences*. John Wiley and Sons, New York, 2011.