Theory of Stochastic Processes
Vol. 16 (32), no. 2, 2010, pp. 44–57

ALEXANDER GNEDIN, ALEXANDER IKSANOV, AND ALEXANDER MARYNYCH

LIMIT THEOREMS FOR THE NUMBER OF OCCUPIED BOXES IN THE BERNOULLI SIEVE

The Bernoulli sieve is a version of the classical 'balls-in-boxes' occupancy scheme, in which random frequencies of infinitely many boxes are produced by a multiplicative renewal process also known as the residual allocation model or stick-breaking. We focus on the number K_n of boxes occupied by at least one of n balls, as $n \to \infty$. A variety of limiting distributions for K_n is derived from the properties of associated perturbed random walks. A refined approach based on the standard renewal theory allows us to remove a moment constraint and to cover the cases left open in previous studies.

1. Introduction and main result

In a classical occupancy scheme, balls are thrown independently in an infinite series of boxes with probability p_k of hitting a box $k = 1, 2, \dots$ for each particular ball, where $(p_k)_{k\in\mathbb{N}}$ is a fixed collection of positive frequencies summing up to unity. A quantity of traditional interest is the number K_n of boxes occupied by at least one of n balls. In specific applications, 'boxes' correspond to distinguishable species or types, and K_n is the number of distinct species represented in a random sample of size n. Starting from Karlin's fundamental paper [19], the behavior of K_n and related functionals was studied by many authors [3, 8, 17, 20]. In particular, it is known that the limiting distribution of K_n is normal if the variance of K_n goes to infinity with n, a property which holds, when p_k's have a power-like decay, but does not hold when the frequencies decay exponentially as $k \to \infty$ [7]. See [5, 10] for the survey of recent results on the infinite occupancy.

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

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

and to compare fluctuations of K_n about R_n^* with fluctuations of R_n^* itself. By Karlin's law of large numbers [19], we always have $K_n \sim R_n^*$ a.s. (as $n \to \infty$), so the environment may be regarded as strong if the sampling variability is negligible to the extent that R_n^* and K_n, normalized by the same constants, have the same limiting distributions, see [13] for examples.

2000 *Mathematics Subject Classification.* Primary 60F05; Secondary 60C05.

Key words and phrases. Infinite occupancy scheme; perturbed random walk; random environment; weak convergence.

In this paper, we focus on the limiting distributions of K_n for the Bernoulli sieve [9, 11, 12] which is the infinite occupancy scheme with random frequencies

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

where $(W_k)_{k\in\mathbb{N}}$ are independent copies of a random variable W taking values in $(0, 1)$. From a viewpoint, K_n is the number of blocks of a regenerative composition structure [4, 13] induced by a compound Poisson process with jumps $|\log W_k|$. Discrete probability distributions with random masses (1) are sometimes called residual allocation models, the best known being the instance associated with Ewens' sampling formula, when $W \overset{d}{=} \text{beta}(c, 1)$ for $c > 0$. Following [9, 12], frequencies (1) can be considered as sizes of the component intervals obtained by splitting $[0, 1]$ at points of the multiplicative renewal process $(Q_k : k \in \mathbb{N}_0)$, where

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

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

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

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

which may be finite or infinite.

Under the assumptions $\nu < \infty$ and $\sigma^2 < \infty$, the central limit theorem for K_n was proved in [9] by using the analysis of random recursions. Under the sole assumption $\nu < \infty$, the criterion for weak convergence of K_n and the list of all possible limit distributions were obtained in [12] from the behavior of

$$\rho^*(x) := \inf\{k \in \mathbb{N} : W_1 \dots W_k < e^{-x}\}, \quad x \geq 0. \tag{2}$$

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

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

in the range of small frequencies (large x). This allows us to treat the cases of finite and infinite ν in a unified way and to see how the centering of K_n needs to be adjusted in the case $\nu = \infty$. Although the approach based on the asymptotics of small frequencies is familiar from [5, 13, 19], the application to the Bernoulli sieve is new. We emphasize here that the connection of K_n to $N^*(x)$ is more fundamental, since $N^*(x)$ is not sensitive to the arrangement of boxes in some order, as compared to $\rho^*(x)$ involving explicitly the ordered features of the environment. Thus, we believe that the present paper offers a more insightful way to study the occupancy problem and calls for generalizations. Our main result is the following theorem.

Theorem 1.1. *If there exist functions $f : \mathbb{R}_+ \to \mathbb{R}_+$ and $g : \mathbb{R}_+ \to \mathbb{R}$ such that $(\rho^*(x) - g(x))/f(x)$ converges weakly (as $x \to \infty$) to some non-degenerate and proper distribution, then also $(X_n - b_n)/a_n$ converges weakly (as $n \to \infty$) to the same distribution, where X_n can be either K_n or $N^*(\log n)$, and the constants are given by*

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

In more details, the possible limits for ρ^* and the convergence criteria, as summarized in Appendix to [12], lead to the following characterization.

Corollary 1.1. *The assumption of* Theorem 1.1 *holds iff either the distribution of* $|\log W|$ *belongs to the domain of attraction of a stable law or the function* $\mathbb{P}\{|\log W| > x\}$ *slowly varies at* ∞*. Accordingly, there are five possible types of convergence:*

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

$$b_n = \mu^{-1}\bigg(\log n - \int_0^{\log n} \mathbb{P}\{|\log(1-W)| > x\}\mathrm{d}x\bigg) \tag{3}$$

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

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

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

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

(c) *If*

$$\mathbb{P}\{|\log W| > x\} \ \sim\ x^{-\alpha}L(x), \quad x \to \infty, \tag{4}$$

for some L *slowly varying at* ∞ *and* $\alpha \in (1,2)$*, then, with* b_n *given in* (3) *and* $a_n = \mu^{-(\alpha+1)/\alpha} c_{[\log n]}$*, where* (c_n) *is any positive sequence satisfying*

$$\lim_{n\to\infty} nL(c_n)/c_n^\alpha = 1,$$

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

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

(d) *Assume that relation* (4) *holds with* $\alpha = 1$*. Let* $r : \mathbb{R} \to \mathbb{R}$ *be any nondecreasing function such that* $\lim_{x\to\infty} x\mathbb{P}\{|\log W| > r(x)\} = 1$*. We set*

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

Then, with

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

and

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

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

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

(e) *If the relation* (4) *holds for* $\alpha \in [0,1)$*, then, with* $b_n \equiv 0$ *and*

$$a_n := \log^\alpha n / L(\log n),$$

the limiting distribution of K_n/a_n *is the Mittag-Leffler law* θ_α *with moments*

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

To connect to the previous results, we consider the index I_n of the last occupied box, which is the value of k satisfying $Q_k < \min(U_1, \dots, U_n) < Q_{k-1}$. Let $L_n := I_n - K_n$ be the number of empty boxes with indices not exceeding I_n. From [12], we know that the number L_n of empty boxes is regulated by μ and ν via the relation $\lim_{n\to\infty} \mathbb{E}L_n = \nu/\mu$ (provided at least one of these is finite), and that the weak asymptotics of I_n coincides with that of $\rho^*(\log n)$, i.e. $(I_n - b_n)/a_n$ and $(\rho^*(\log n) - b_n)/a_n$ have the same proper and non-degenerate limiting distribution (if any). In [9, 12] it was shown that, under the condition $\nu < \infty$, the weak asymptotics of K_n coincides with that of I_n and, hence, with that of $\rho^*(\log n)$. That is to say, when $\nu < \infty$, the way L_n varies does not affect the asymptotics of K_n, meaning that L_n is dominated by I_n in the representation $K_n = I_n - L_n$. Clearly, this result is a particular case of Theorem 1.1 because, when $\nu < \infty$,

$$\lim_{x\to\infty} \frac{g(x) - \int_0^x g(x-y)\,\mathbb{P}\{|\log(1-W)| \in \mathrm{d}y\}}{f(x)} = 0 \tag{5}$$

(see Remark 3.1 for the proof). Now, Theorem 1.1 says that, in the case where $\nu = \infty$, the asymptotics of L_n may affect the asymptotics of K_n, and this is indeed the case whenever (5) fails, hence a more sophisticated two-term centering of K_n is indispensable. The following example illustrates the phenomenon.

Example 1.1. Assume that, for some $\gamma \in (0, 1/2)$,

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

Then

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

and, in this case,

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

Thus, we see that the centering by $\mu^{-1}\log n$ is not enough, as the remainder $b_n - \mu^{-1}\log n$ is not killed by scaling. Moreover, one can check that, indeed,

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

which demonstrates the substantial contribution of L_n.

As in much of the previous work, we make use of the poissonized version of the occupancy model, in which balls are thrown in boxes in continuous time, at the epochs of a unit rate Poisson process. The variables associated with time $t \ge 0$ will be denoted $K(t), R^*(t)$, *etc.* For instance, the expected number of occupied boxes within a time interval $[0, t]$ conditionally given (P_k) is

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

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

The variable $N^*(x)$ is the number of sites on $[0, x]$ visited by a perturbed random walk with generic components $|\log W|, |\log(1-W)|$. Therefore, we shall develop some general renewal theory for perturbed random walks, which we believe might be of some independent interest. The approach based on perturbed random walks is more general than the one exploited in [12] and is well adapted to treat the cases $\nu < \infty$ and $\nu = \infty$ in a unified way.

2. Renewal theory for perturbed random walks

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

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

is called a *perturbed random walk* (see, e.g., [2], [14, Chapter 6], [18]). Since $\lim_{k\to\infty} T_k = \infty$ a.s., there is some finite number

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

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

$$R(x) := \sum_{k=0}^{\infty} \Big(1 - \exp\left(-xe^{-T_k}\right)\Big), \quad x \geq 0. \tag{6}$$

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

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

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

In the case $\mathbb{E}\eta < \infty$, the weak convergence of one of the variables $(\rho(x) - g(x))/f(x)$ and $(N(x) - g(x))/f(x)$ (with suitable f, g) implies the weak convergence of another one to the same distribution. Our main focus is thus on the cases where the contribution of η_k does affect the asymptotics of $N(x)$. To our knowledge, the questions about the asymptotics of perturbed random walks were circumvented in the literature by imposing an appropriate moment condition which allowed the reduction to (S_k) (see, e.g., [14, Chapter 6], [16], [21, Theorems 2.1 and 2.2]).

A well-known property of ρ is the subadditivity: for $x, y \geq 0$,

$$\rho(x + y) - \rho(x) \overset{a.s.}{\leq} \rho'(x, y) \overset{d}{=} \rho(y), \tag{7}$$

where $\rho'(x, y) := \inf\{k - N(x) \in \mathbb{N} : S_k - S_{\rho(x)} > y\}$, and $(\rho'(x, y) : y \geq 0)$ is independent of $\rho(x)$ and has the same distribution as $(\rho(y) : y \geq 0)$. Consider $U(x) := \mathbb{E}\rho(x) = \sum_{k=0}^{\infty} \mathbb{P}\{S_k \leq x\}$, the renewal function of (S_k). From (7) and Fekete's lemma, we have

$$U(x + y) - U(x) \leq C_1 y + C_2, \quad x, y \geq 0, \tag{8}$$

for some positive constants C_1 and C_2.

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

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

Throughout, we shall consider the functions involved in the centering of random variables up to this kind of equivalence. For instance, when we write $g = 0$, what is really meant is that g is equivalent to zero, with context-dependent f involved in the scaling of some random variable.

The next lemma will be used in the proof of Theorem 2.2.

Lemma 2.1. *If $\frac{\rho(x) - g(x)}{f(x)}$ weakly converges, then*

$$\lim_{x\to\infty} \frac{g(x) - g(x - y)}{f(x)} = 0 \quad \textit{locally uniformly in } y, \tag{9}$$

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

$$\lim_{x\to\infty} \frac{\int_0^x g(x-y)\,\mathrm{d}G(y) - \int_0^{x+\lambda} g(x+\lambda-y)\,\mathrm{d}G(y)}{f(x)} = 0, \tag{10}$$

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

Proof. Clearly, (9) is a property of the class of f-equivalent functions g.

We refer to the list of possible limiting laws and corresponding normalizations for $\rho(x)$ [12, Proposition A.1]. Relation (9) trivially holds, when $g(x) \equiv 0$. It is known that $g(x)$ cannot be chosen as zero if the law of ξ belongs to the domain of attraction of the α-stable law for $\alpha \in [1,2]$. It is known that, for ξ in the domain of attraction of a stable law with $\alpha \in (1,2]$, one can take $g(x) = x/\mathbb{E}\xi$ which satisfies (9).

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

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

where $m(x) := \int_0^x \mathbb{P}\{\xi > y\}\mathrm{d}y$, and $r(x)$ is any nondecreasing function such that $\lim_{x\to\infty} x\mathbb{P}\{\xi > r(x)\} = 1$. The concavity of $m(x)$ implies that $x \mapsto x/m(x)$ is nondecreasing. Thus, $x \mapsto m(r(x/m(x)))$ is nondecreasing too as a superposition of three nondecreasing functions. Hence, for every $\gamma \in (0,1)$,

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

which readily implies the subadditivity of g via

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

Thus,

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

For the converse inequality for lim inf, it is enough to choose a non-increasing g from the f-equivalence class. By [1, Theorem 2], this can be done, indeed, by taking inverse function to $x \mapsto xm(r(x))$.

The stated uniformity of convergence is checked along the same lines, and (10) follows from the subadditivity of b and easy estimates. □

2.2. **The case without centering.** We start with criteria for the weak convergence of $\rho(x)$ and $R(x)$ in the case where no centering is needed.

Theorem 2.1. *For* $Y(x)$ *and any of the variables* $\rho(x)$, $N(x)$, *or* $R(e^x)$, *the following conditions are equivalent:*

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

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

$$\mathbb{P}\{\xi > x\} \sim x^{-\alpha} L(x), \quad x \to \infty. \tag{11}$$

Furthermore, if (11) *holds, then the limiting law is the Mittag-Leffler distribution* θ_α, *and one can take* $f(x) = x^\alpha / L(x)$.

The assertion of Theorem 2.1 regarding $\rho(x)$ follows from [12, Appendix]. For the other two variables, the result is a consequence of the following lemma.

Lemma 2.2. *We have*

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

and

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

Proof. By definition of the perturbed random walk,

$$\rho(x-y)-\sum_{j=1}^{\rho(x)} 1_{\{\eta_j>y\}}\leq N(x)\leq \rho(x) \tag{12}$$

for $0<y<x$.

Clearly, $\rho(x)\uparrow\infty$ a.s. and

$$\rho(x-y)\geq \rho(x)-\rho'(x-y,y) \quad \text{a.s.} \tag{13}$$

with ρ' as in (7), from which

$$\frac{\rho(x-y)}{\rho(x)}\overset{\mathbb{P}}{\to}1, \quad x\to\infty. \tag{14}$$

Finally, by the strong law of large numbers, we have

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

Therefore, dividing (12) by $\rho(x)$ and letting firstly $x\to\infty$ and then $y\to\infty$, we obtain the first part of the lemma.

For the second assertion, we use the representation

$$\begin{aligned} R(x) &= \int_1^\infty (1-e^{-x/y})\mathrm{d}\,N(\log y) \\ &= \int_0^x N(\log x-\log y)e^{-y}\mathrm{d}y-(1-e^{-x})N(0). \end{aligned} \tag{15}$$

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

$$\int_0^x N(\log x-\log y)e^{-y}\mathrm{d}y\geq \int_0^a N(\log x-\log y)e^{-y}\mathrm{d}y\geq N(\log x-\log a)(1-e^{-a}).$$

Dividing this inequality by $\rho(\log x)$, sending $x\to\infty$, using (14) and the already established part of the lemma, and finally letting $a\to\infty$, we obtain a half of the desired conclusion.

To get the other half, we write

$$\begin{aligned} \int_0^x N(\log x-\log y)e^{-y}\mathrm{d}y \ &\overset{a.s.}{\leq}\ \rho(\log x)(1-e^{-x}) \\ &+\int_0^1(\rho(\log x-\log y)-\rho(\log x))e^{-y}\mathrm{d}y, \end{aligned} \tag{16}$$

where (7), the inequality $N(x)\leq\rho(x)$ a.s., and the fact that $\rho(y)$ is a.s. non-decreasing in y have been used. From (8), we have

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

thus, to complete the proof, it remains to divide (16) by $\rho(\log x)$ and send $x\to\infty$. □

2.3. **The case with nonzero centering.** Now we turn to a more intricate case where some centering is needed. We denote the distribution function of η by $F(x)$ and the renewal function of (S_k) by $U(x)$.

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

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

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

Lemma 2.3. *We have*

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

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

$$\mathbb{E}\Big(N(x) - M(x)\Big)^2 \ = \ O\Big(\int_0^x (1-F(y))\mathrm{d}y\Big) = o(x). \tag{17}$$

Proof. For integer $i < j$,

$$\mathbb{E}\bigg(1_{\{S_i \leq x\}}(1_{\{S_i+\eta_{i+1}\leq x\}} - F(x-S_i))1_{\{S_j\leq x\}}(1_{\{S_j+\eta_{j+1}\leq x\}} - F(x-S_j))\bigg|(\xi_k,\eta_k)_{k=1}^j\bigg)$$
$$= 1_{\{S_i\leq x\}}(1_{\{S_i+\eta_{i+1}\leq x\}} - F(x-S_i))1_{\{S_j\leq x\}}\Big(F(x-S_j) - F(x-S_j)\Big) = 0.$$

Hence,

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

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

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

where $a := \mathbb{E}\xi$ may be finite or infinite. If $\mathbb{E}\eta = \infty$ and $a < \infty$, a generalization of the key renewal theorem due to Sgibnev [22, Theorem 4] yields

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

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

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

Thus, (17) follows in any case.

□

Theorem 2.2. *If, for some random variable Z,*

$$\frac{\rho(x)-g(x)}{f(x)} \overset{d}{\to} Z,\ x\to\infty, \tag{18}$$

then also

$$\frac{M(x)-\int_0^x g(x-y)\mathrm{d}F(y)}{f(x)} \overset{d}{\to} Z,\ x\to\infty, \tag{19}$$

$$\frac{N(x)-\int_0^x g(x-y)\mathrm{d}F(y)}{f(x)} \overset{d}{\to} Z,\ x\to\infty, \tag{20}$$

and

$$\frac{R(x)-\int_0^{\log x} g(\log x-y)\mathrm{d}F(y)}{f(\log x)} \overset{d}{\to} Z,\ x\to\infty. \tag{21}$$

Proof. Integrating by parts yields

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

So to prove (19), it is enough to show that, as $x\to\infty$,

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

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

$$T_1(x)+T_2(x):=\int_0^\delta \frac{\rho(x-y)-g(x-y)}{f(x)}\mathrm{d}F(y)+\int_\delta^x \frac{\rho(x-y)-g(x-y)}{f(x)}\mathrm{d}F(y). \tag{22}$$

From the proof of Lemma 2.1, we know that it can be assumed, without loss of generality, that $g(x)$ is nondecreasing. Thus, almost surely,

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

In view of (7) and (9), we have the convergence $\lim_{\delta\to\infty}\lim_{x\to\infty} T_1(x)=Z$ in distribution.

For $x>0$, we set

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

and

$$\mathcal{Z}_x:=(Z_x(t):t\ge 0).$$

We will establish next that there exists a cadlag process $\mathcal{Z}=(Z(t),\ t\ge 0)$ such that

$$\frac{\sup_{y\in[0,x]}(\rho(y)-g(y))}{f(x)}=\sup_{t\in[0,1]} Z_x(t) \overset{d}{\to} \sup_{t\in[0,1]} Z(t),\quad x\to\infty, \tag{23}$$

and, similarly,

$$\frac{\inf_{y\in[0,x]}(\rho(y)-g(y))}{f(x)}=\inf_{t\in[0,1]} Z_x(t) \overset{d}{\to} \inf_{t\in[0,1]} Z(t),\quad x\to\infty. \tag{24}$$

CASE 1: Suppose that we can choose $g(x) = x/\mathbb{E}\xi$. Then Z is an α-stable random variable for some $\alpha \in (1, 2]$. We denote, by $\mathcal{Z} = (Z(t) : t \geq 0)$, a stable Lévy process such that $Z(1) \overset{d}{=} Z$. Regard $\mathcal{Z}_x$ and $\mathcal{Z}$ as random elements of Skorokhod's space $D[0, \infty)$ endowed with the M_1-topology.

By [6, Theorem 1b],

$$\mathcal{Z}_x \Rightarrow \mathcal{Z}, \quad x \to \infty. \tag{25}$$

Since the sup and inf functionals are M_1-continuous, we obtain (23) and (24), using the continuous mapping theorem.

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

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

and $(\hat{Z}(t) : t \geq 0)$ is a stable Lévy process such that $\hat{Z}(1) \overset{d}{=} Z$. With this notation, we derive (25) from [15, Theorem 2], from which (23) and (24) follow along the above lines.

Now it remains to estimate

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

with T_2 from (22). Using (23) and (24), we conclude that $\lim\limits_{\delta\to\infty}\lim\limits_{x\to\infty} T_2(x) = 0$ in probability. The proof of (19) is complete.

In view of (17), $\mathbb{E}(M(x) - N(x))^2 = o(x)$. Since $f^2(x)$ grows not slower than x (see [12, Proposition A.1]), Chebyshev's inequality yields

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

Now (20) follows from (19).

It remains to establish (21). To this end, we introduce, for $x > 1$,

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

Using

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

and (8), we conclude that, for $y \in (1, x)$,

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

Therefore, $\mathbb{E}Q_1(x) = O(1)$, as $x \to \infty$, whence $\frac{Q_1(x)}{f(\log x)} \xrightarrow{\mathbb{P}} 0$. Similarly, $\frac{Q_2(x)}{f(\log x)} \xrightarrow{\mathbb{P}} 0$. Thence, recalling (15)

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

As $N(\log x)$ grows in probability not faster than $\log x$, we conclude that

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

Now an appeal to (20) completes the proof. □

3. Proof of Theorem 1.1

The results on perturbed random walks can be applied now. Set

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

and

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

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

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

the appeal to Theorem 2.1 in the case $g=0$ and to Theorem 2.2 in the case $g\neq 0$ proves the result for $N^*(\log n)$. To prove the statement for K_n, we use the poissonization.

Step 1. We firstly check that

$$\lim_{t\to\infty} \mathbb{E}\,\mathrm{Var}(K(t)|(P_k)) = \frac{\log 2}{\mu}, \tag{26}$$

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

$$\frac{K(t)-\mathbb{E}(K(t)|(P_k))}{q(t)} \overset{\mathbb{P}}{\to} 0, \tag{27}$$

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

According to [19, formula (25)],

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

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

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

which is the same as

$$\mathbb{E}\,\mathrm{Var}(K(e^x)|(W_k)) = \int_0^{\infty} A(x-y)\mathrm{d}U^*(y). \tag{28}$$

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

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

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

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

Furthermore, arguing in the same way as in [12, Section 5], we can prove that $A(t)$ is directly Riemann-integrable. Therefore, the application of the key renewal theorem on $\mathbb{R}$ to (28) yields (26).

Chebyshev's inequality together with (26) imply that, for every $\varepsilon > 0$,

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

which proves (27) upon taking expectation and invoking the Lebesgue bounded convergence theorem.

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

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

weakly converges to the same law.

Using this observation and exploiting Theorem 2.1 (in the case $g = 0$) or formula (21) of Theorem 2.2 (in the case $g \neq 0$), we conclude that the weak convergence of $\frac{\rho^*(x)-g(x)}{f(x)}$ to some distribution θ implies the weak convergence of both

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

and

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

to θ.

Step 3. It remains to pass from the poissonized occupancy model to the fixed-n model. In view of (10) and the fact that $f(\log t)$ is slowly varying,

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

satisfies

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

for every $0 < \varepsilon < 1$. Thus, we have

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

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

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

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

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

for all $x \geq 0$. To prove the converse inequality for $\liminf$, one has to note that

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

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

Remark 3.1. Here is the promised verification of (5). Below, we use the terminology introduced in the proof of Lemma 2.1.

Lemma 3.1. *Relation* (5) *is a property of the class of f-equivalent functions g.*

Proof. Assume that g satisfies (5). We have to show that any g_1 such that

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

satisfies (5) as well.

Plainly, it is enough to check that

$$A(x):=\frac{\int_0^x(g(x-y)-g_1(x-y))\mathrm{d}\,F(y)}{f(x)}\ \to 0,\quad x\to\infty. \tag{29}$$

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

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

Sending $x\to\infty$ and then $\varepsilon\downarrow 0$ proves (29). □

If the law of $|\log W|$ belongs to the domain of attraction of an α-stable law, $\alpha\in(1,2]$, then $(\rho(x)-g(x))/f(x)$ weakly converges with $g(x)=x/\mu$ and appropriate $f(x)$. Such a g trivially verifies (5) which, by Lemma 3.1, entails that every g_1 from the same f-equivalence class verifies (5).

If the law of $|\log W|$ belongs to the domain of attraction of a 1-stable law, then $(\rho(x)-g(x))/f(x)$ weakly converges for $g(x)=\frac{x}{m(r(x/m(x)))}$ and $f(x)=\frac{r(x/m(x))}{m(x)}$, with m and r as defined in part (d) of Corollary 1.1. Since r is regularly varying with index one, we can assume it without loss of generality, and, hence, g are differentiable. Since $\frac{g(x)}{xf(x)}$ is regularly varying with index (-1), it converges to 0 as $x\to\infty$. In addition, $\lim\limits_{x\to\infty}x\mathbb{P}\{\zeta>x\}=0$ in view of $\nu<\infty$, where we denoted $|\log(1-W)|$ by ζ. Hence,

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

Thus, it suffices to check that

$$\lim_{x\to\infty}\frac{\mathbb{E}(g(x)-g(x-\zeta))1_{\{\zeta\le x\}}}{f(x)}=0. \tag{30}$$

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

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

where $K:=1/m(r(1))$. This immediately implies (30) and the whole claim by virtue of Lemma 3.1.

Acknowledgement A. Iksanov gratefully acknowledges the support by a grant from the Utrecht University.

References

1. Anderson, K. K. and Athreya, K. B. (1988). A note on conjugate Φ-variation and a weak limit theorem for the number of renewals. *Stat. Prob. Letters.* **6**, 151–154.
2. Araman, V. F. and Glynn, P. W. (2006). Tail asymptotics for the maximum of perturbed random walk. *Ann. Appl. Probab.* **16**, 1411–1431.
3. Barbour, A. D. (2009). Univariate approximations in the infinite occupancy scheme. *Alea.* **6**, 415–433.
4. Barbour, A.D. and Gnedin, A.V. (2006) Regenerative compositions in the case of slow variation, *Stoch. Proc. Appl.* **116**, 1012–1047.
5. Barbour, A. D. and Gnedin, A. V. (2009). Small counts in the infinite occupancy scheme. *Electron. J. Probab.* **14**, 365–384.
6. Bingham, N. H. (1973). Maxima of sums of random variables and suprema of stable processes. *Z. Wahrsch. verw. Geb.* **26**, 273– 296.
7. Bogachev, L. V., Gnedin, A. V. and Yakubovich, Yu. V. (2008). On the variance of the number of occupied boxes. *Adv. Appl. Math.* **40**, 401–432.
8. Dutko, M. (1989). Central limit theorems for infinite urn models. *Ann. Probab.* **17**, 1255–1263.
9. Gnedin, A. V. (2004). The Bernoulli sieve. *Bernoulli* **10**, 79–96.
10. Gnedin, A., Hansen, B. and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys* **4**, 146–171.
11. Gnedin, A., Iksanov, A. and Roesler, U. (2008). Small parts in the Bernoulli sieve. *Discrete Mathematics and Theoretical Computer Science*, Proceedings Series, Volume **AI**, 235–242.
12. Gnedin, A., Iksanov, A., Negadajlov, P., and Roesler, U. (2009). The Bernoulli sieve revisited. *Ann. Appl. Prob.* **19**, 1634–1655.
13. Gnedin, A., Pitman, J. and Yor, M. (2006) Asymptotic laws for compositions derived from transformed subordinators, *Ann. Probab.* **34**, 468–492.
14. Gut, A. (2009). Stopped Random Walks: Limit Theorems and Applications, Springer: New York.
15. de Haan, L. and Resnick, S. I. (1979). Conjugate Π-variation and process inversion. *Ann. Probab.* **7**, 1028–1035.
16. Hitczenko, P. and Wesolowski, J. (2010+). Renorming divergent perpetuities. *Bernoulli*, to appear.
17. Hwang, H. K. and Janson, S. (2008). Local limit theorems for finite and infinite urn models. *Ann. Probab.* **36**, 992–1022.
18. Iksanov, O. M. (2007). Perpetuities, Branching Random Walks and Self-Decomposability, Zirka: Kiev (in Ukrainian).
19. Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *J. Math. Mech.* **17,** 373–401.
20. Mirakhmedov, Sh. A. (1989). Randomized decomposable statistics in a generalized allocation scheme over a countable set of cells. *Diskretnaya Matematika.* **1**, 46–62.
21. Rachev, S. T. and Samorodnitsky, G. (1995). Limit laws for a stochastic process and random recursion arising in probabilistic modelling. *Adv. Appl. Prob.* **27**, 185–202.
22. Sgibnev, M. S. (1981). Renewal theorem in the case of an infinite variance. *Siberian Math. J.* **22**, 787–796.

Address of Alexander Gnedin

Current address: Department of Mathematics, Utrecht University, Postbus 80010, 3508 TA Utrecht, The Netherlands

E-mail address: A.V.Gnedin@uu.nl

Address of Alexander Iksanov

Current address: Faculty of Cybernetics, Taras Shevchenko National University of Kiev, Kiev 01033, Ukraine

E-mail address: iksan@unicyb.kiev.ua

Address of Alexander Marynych

Current address: Faculty of Cybernetics, Taras Shevchenko National University of Kiev, Kiev 01033, Ukraine

E-mail address: marynych@unicyb.kiev.ua