Electron. J. Probab. **22** (2017), no. 44, 1–26.
ISSN: 1083-6489 DOI: 10.1214/17-EJP59

Extremes and gaps in sampling from a GEM random discrete distribution

Jim Pitman* Yuri Yakubovich†

Abstract

We show that in a sample of size n from a GEM$(0,\theta)$ random discrete distribution, the gaps $G_{i:n} := X_{n-i+1:n} - X_{n-i:n}$ between order statistics $X_{1:n} \le \cdots \le X_{n:n}$ of the sample, with the convention $G_{n:n} := X_{1:n} - 1$, are distributed like the first n terms of an infinite sequence of independent geometric$(i/(i+\theta))$ variables G_i. This extends a known result for the minimum $X_{1:n}$ to other gaps in the range of the sample, and implies that the maximum $X_{n:n}$ has the distribution of $1+\sum_{i=1}^{n} G_i$, hence the known result that $X_{n:n}$ grows like $\theta\log(n)$ as $n \to \infty$, with an asymptotically normal distribution. Other consequences include most known formulas for the exact distributions of GEM$(0,\theta)$ sampling statistics, including the Ewens and Donnelly–Tavaré sampling formulas. For the two-parameter GEM(α,θ) distribution we show that the maximal value grows like a random multiple of $n^{\alpha/(1-\alpha)}$ and find the limit distribution of the multiplier.

Keywords: GEM samples; order statistics; sample maximum; random discrete distribution; size-biased permutation.
AMS MSC 2010: 60G70; 60G09; 60F05.
Submitted to EJP on February 10, 2017, final version accepted on April 21, 2017.

1 Introduction

Consider a sequence of real random variables $X_1, X_2, \ldots$ with order statistics

$$X_{1:n} := \min_{1\le i\le n} X_i \le X_{2:n} \le X_{3:n} \le \cdots \le X_{n:n} := \max_{1\le i\le n} X_i.$$

For an independent and identically distributed (i.i.d.) sequence (X_n) with a continuous distribution function $F(x) := \mathbb{P}(X_n \le x)$, the probabilistic structure of order statistics is well understood. Many exact distributional identities are obtained by reduction to the simplest i.i.d. cases:

*Statistics Department, 367 Evans Hall # 3860, University of California, Berkeley, CA 94720-3860, U.S.A. E-mail: pitman@berkeley.edu
†Saint Petersburg State University, St. Petersburg State University, 7/9 Universitetskaya nab., St. Petersburg, 199034 Russia. E-mail: y.yakubovich@spbu.ru

- $X_i = U_i$, signifying $F(x) = x$ for $0 \le x \le 1$, the *uniform distribution* on $(0, 1)$;
- $X_i = \varepsilon_i/\lambda$ for ε_i i.i.d. exponential(1) and a constant $\lambda > 0$, in which case $F(x) = 1 - e^{-\lambda x}$ for $x \ge 0$, the *exponential*(λ) *distribution* on $(0, \infty)$.

This reduction involves the identity of n-dimensional joint distributions

$$(F(X_{i:n}), 1 \le i \le n) \stackrel{d}{=} (U_{i:n}, 1 \le i \le n) \stackrel{d}{=} (1 - \exp(-\varepsilon_{i:n}), 1 \le i \le n) \tag{1.1}$$

which holds with almost sure identities if the U_i and ε_i are defined by $U_i := F(X_i)$ and $\varepsilon_i := -\log(1 - U_i)$. For discrete distributions the situation is complicated by possible ties but also well understood. See [14] [53] for further background on order statistics.

We are primarily interested here in the structure of the *gaps between sample values* which we list from the top of the sample down, as

$$G_{i:n} := X_{n+1-i:n} - X_{n-i:n} \qquad (1 \le i \le n)$$

for distributions of X_i whose support has a *minimal value* $m_0 \ge 0$, with $X_{0:n} := m_0$. So we interpret $G_{n:n} := X_{1:n} - m_0$ as the *gap below the minimum* of the sample, with $G_{n:n} = 0$ iff some sample value hits the minimum of the range. The order statistics are then encoded in the gaps as $X_{k:n} = m_0 + \sum_{i=0}^{k-1} G_{n-i:n}$. In particular, the minimal and maximal values of the sample are

$$X_{1:n} := m_0 + G_{n:n} \qquad \text{and} \qquad X_{n:n} = m_0 + \sum_{i=1}^{n} G_{i:n}\,. \tag{1.2}$$

According to a well known result of Sukhatme–Rényi [65], [53, Repr. 3.4], for i.i.d. sampling the structure of the gaps is simplest for exponential variables:

$$\text{for } X_i = \frac{\varepsilon_i}{\lambda} \text{ the } G_{i:n} \text{ are independent with } (G_{i:n}, 1 \le i \le n) \stackrel{d}{=} \left(\frac{X_i}{i}, 1 \le i \le n\right). \tag{1.3}$$

Among absolutely continuous distributions, the family of shifted exponential distributions is characterized by quite weak forms of this assertion, for instance that $G_{1:2}$ is independent of $G_{2:2}$. See Ferguson [27] and earlier work cited there, and [69] for more recent results in this vein. Formulas (1.2) and (1.3) explain the well known identity in distribution

$$M_n := \max_{1 \le i \le n} \varepsilon_i \stackrel{d}{=} T_n := \sum_{i=1}^{n} \frac{\varepsilon_i}{i} \tag{1.4}$$

which implies the convergence in distribution, with centering but no normalization

$$M_n - \log n \;\stackrel{d}{=}\; T_n - \log n \;\stackrel{d}{\to}\; -\gamma + \sum_{i=1}^{\infty} \frac{(\varepsilon_i - 1)}{i} \tag{1.5}$$

where $\gamma := \lim_{n\to\infty}(-\log n + \sum_{i=1}^{n} 1/i)$ is Euler's constant. The infinite sum converges both almost surely and in mean square, by Kolmogorov's theorem for sums of independent random variables with mean 0, and the limit has the *Gumbel distribution function* $F(x) = \exp(-e^{-x})$. More generally, the asymptotic behavior of the maximum M_n of an i.i.d. sample from a continuous distribution is well understood. If after proper rescaling, the distribution of M_n has a non-degenerate weak limit, that limit must have the distribution function $F_\rho(x) = \exp(-(1 + x\rho)^{-1/\rho})$, for some $\rho \in (-\infty, \infty)$ and x such that $1 + x\rho > 0$ (and $F_\rho(x)$ equals 0 or 1 for other x), see, e.g., [53]. Here ρ (and the scaling) depends on the behavior of the distribution near the supremum of its support. Limits can be also degenerate, and there exist distributions for which no non-degenerate limit is possible.

The situation is quite different for an infinite exchangeable sequence X_1, $X_2, \ldots$. In this case any distribution can appear as a limiting distribution of the finite sample maximum $M_n := X_{n:n}$, as shown by the following example, which seems to be folklore. Let $Z_1, Z_2, \ldots$ be a sequence of i.i.d. random variables and let M be a random variable, independent of this sequence, with some given distribution. Take $X_n = Z_n + M$ to obtain an exchangeable sequence $X_1, X_2, \ldots$. If the support of the distribution of Z_1 is bounded above then M_n converges a.s. to a shift of M, without any rescaling. So to obtain results of any interest about limit distributions of maxima from an exchangeable sequence, some further structure must be involved.

We are interested here in the distribution of sample gaps, sample extremes, and related statistics, for exchangeable samples from a random discrete distribution $P_\bullet := (P_1, P_2, \ldots)$ on the set $\mathbb{N} := \{1, 2, \ldots\}$ of positive integers, subject to

$$0 < P_j < 1 \text{ for every } j = 1, 2, \ldots \text{ and } \sum_{j=1}^{\infty} P_j = 1 \text{ almost surely.} \tag{1.6}$$

We specify $P_\bullet$ by the *residual allocation model* (RAM) or *stick-breaking scheme* [42], [68] [58, §5]

$$P_j := H_j \prod_{i=1}^{j-1}(1 - H_i), \quad \text{with} \tag{1.7}$$

$$0 < H_i < 1 \qquad \text{and} \qquad \prod_{i=1}^{\infty}(1 - H_i) = 0 \text{ almost surely.} \tag{1.8}$$

The random variables H_i may be called *residual fractions*, *random discrete hazards*, or *factors*. We use the term RAM to indicate that the H_i are independent, but not necessarily that they are identically distributed. But we also consider these models in the broader context of $H_\bullet$ subject to (1.8), corresponding to $P_\bullet$ subject to (1.6), without any further dependence assumptions, which we call a *generalized residual allocation model* (GRAM) [58, §5].

The case of i.i.d. factors H_i has been extensively studied by Gnedin and coauthors [36], [30], who call this model the *Bernoulli sieve*. Another RAM of particular interest, because of its *invariance under size-biased permutation* [59] is the GEM model with parameters (α, θ). In this model, H_i has the beta$(1 - \alpha, \theta + i\alpha)$ density on $(0, 1)$

$$\frac{\mathbb{P}[H_i \in dx]}{dx} = \frac{x^{-\alpha}(1 - x)^{\theta + i\alpha - 1}}{B(1 - \alpha, \theta + i\alpha)} \qquad (0 < x < 1, i \in \mathbb{N}), \tag{1.9}$$

where $0 \le \alpha < 1$ and $\theta > -\alpha$ are real parameters, and $B(\cdot, \cdot)$ is Euler's beta function. The GEM hazard variables are i.i.d. only in the important special case $\alpha = 0$ covered by the following theorem:

Theorem 1.1. *Let $X_1, X_2, \ldots$ be an exchangeable sequence obtained by i.i.d. sampling from the* GEM$(0, \theta)$ *distribution* (1.7) *for i.i.d. beta$(1, \theta)$ hazards H_i. For each fixed $n \ge 1$, the gaps $G_{i:n}$ between the order statistics of the sample, read from right to left, with $G_{n:n} + 1 := \min_{1 \le i \le n} X_i$, are independent geometric$(i/(i + \theta))$ variables, with means θ/i for $1 \le i \le n$.*

As detailed in Section 3, this theorem contains most known results about GEM$(0, \theta)$ samples, including the well known sampling formulas for GEM$(0, \theta)$, due to Ewens [23], Antoniak [2], and Donnelly and Tavaré [18]. It is also very close to recent studies of random compositions derived from RAMs with i.i.d. factors, as we acknowledge further below. Our simple description of GEM$(0, \theta)$ gaps is hidden in these studies by different

encodings of the values and their multiplicities in discrete random sampling, based on the *count sequence* $N^\circ_{\bullet:n} := (N^\circ_{1:n}, N^\circ_{2:n}, \ldots)$ defined by

$$N^\circ_{b:n} := \sum_{i=1}^{n} \mathbf{1}(X_i = b) \qquad (b = 1, 2, \ldots). \tag{1.10}$$

Here $\mathbf{1}(A)$ denotes the indicator of an event or set A. The notion of a random sample from a random discrete distribution admits a variety of interpretations, some of which are recalled in Section 3. But as our primary metaphor for sampling, we follow recent studies of the Bernoulli sieve [30] in regarding the sample $X_1, \ldots, X_n$ as an allocation of n balls labeled by $i = 1, 2, \ldots, n$ into an unlimited number of boxes labeled by $b \in \{1, 2, \ldots\}$. So X_i is the label of the box into which ball i is thrown. Given $P_\bullet$ the X_i are independent allocations with $\mathbb{P}(X_i = b \,|\, P_\bullet) = P_b$. The count $N^\circ_{b:n}$ is the number of balls thrown into box b, the sample maximum $X_{n:n} = \max\{b : N^\circ_{b:n} > 0\}$ is the label of the rightmost occupied box, and so on.

The key to Theorem 1.1 is the close parallel between the structure of gaps in sampling from GEM$(0, \theta)$, and from exponential(λ), as in (1.3). This parallel guided our choice to list the gaps from top down rather than bottom up, as well as the definition of the final gap $G_{n:n} := X_{1:n} - 1$ in the discrete case. We show in Section 2 how Theorem 1.1 follows easily from its exponential(λ) analog, using Ignatov's construction [44] of GEM$(0, \theta)$ from a Poisson point process.

This construction, and the change of variables (1.1), which maps sampling by independent uniforms in $(0, 1)$ to sampling by independent exponentials in $(0, \infty)$, was developed and applied in a number of previous works [44], [30], to deduce results for sampling from RAMs and related regenerative composition structures from corresponding results in renewal theory. The method yields also the following corollary of Theorem 1.1:

Corollary 1.2. *The* GEM$(0, \theta)$ *models for* $0 < \theta < \infty$ *are the only* RAM*s with i.i.d. factors such that for all sufficiently large* n *the gaps between order statistics in a sample of size* n *are independent.*

We conjecture that the GEM$(0, \theta)$ models are the only random discrete distributions of any kind subject to (1.6) with this property of independence of sample gaps for all n, or for all large n. But resolving this question seems beyond the reach of our current methods.

We discovered these properties of gaps in GEM$(0, \theta)$ samples by seeking an adequate explanation of the identity in distribution for the maximum M_n of a GEM$(0, \theta)$ sample, presented in the next corollary of Theorem 1.1. We first found this identity by a different method indicated in Section 4, without consideration of gaps. But the gaps explain it much better:

Corollary 1.3. *For the maximum* M_n *of a sample of size* n *from* GEM$(0, \theta)$

$$M_n := \max_{1 \le i \le n} X_i = \max\{b : N^\circ_{b:n} > 0\} = 1 + \sum_{i=1}^{n} G_{i:n} \stackrel{d}{=} 1 + \sum_{i=1}^{n} G_i \tag{1.11}$$

for independent G_i *with the geometric*$(i/(i+\theta))$ *distribution.*

Since G_i has mean $\mathbb{E}G_i = \theta/i$ and variance $\operatorname{Var} G_i = \theta(i+\theta)/i^2 \sim \theta/i$ as $i \to \infty$, Lindeberg's central limit theorem implies the limit in distribution

$$\frac{M_n - \theta \log n}{\sqrt{\theta \log n}} \stackrel{d}{\to} Z \qquad \text{as } n \to \infty, \tag{1.12}$$

where Z has the standard normal distribution. This limit theorem for M_n is known as an instance of a more general normal limit theorem for M_n in sampling from a large class of

RAMs with i.i.d. factors [37, Theorem 2.1 (b)]. Also known [4, (5.22)] and [37, Theorem 2.3] is the fact that (1.12) holds also with M_n replaced by the number K_n of distinct values in the sample, which may expressed in various ways parallel to the expressions for M_n in (1.11):

$$K_n := \sum_{j=1}^{n} \mathbf{1}\big(X_j \notin \{X_1, \ldots, X_{j-1}\}\big) = \sum_{b=1}^{\infty} \mathbf{1}(N^\circ_{b:n} > 0) = 1 + \sum_{i=1}^{n-1} \mathbf{1}(G_{i:n} > 0). \tag{1.13}$$

In Section 3 we discuss further these different representations of K_n, their interpretations in various applications, and related representations of the counts

$$K_{j:n} := \sum_{b=1}^{M_n} \mathbf{1}(N^\circ_{b:n} = j)$$

which for $1 \le j \le n$ gives the numbers of clusters of size j in the sample, and for $j = 0$ gives

$$K_{0:n} := M_n - K_n = \sum_{j=1}^{M_n} \prod_{i=1}^{n} \mathbf{1}(X_i \neq j) = G_{n:n} + \sum_{i=1}^{n-1} (G_{i:n} - 1)_+ \tag{1.14}$$

which is the total count of all values between 1 and M_n that are missing in the sample of size n. (Here and below $(x)_+ = \frac{1}{2}(x + |x|)$ is the positive part of x.) There is by now a substantial theory of asymptotics for these and related statistics of samples from RAMs with i.i.d. factors, developed by Gnedin and coauthors by various techniques. In particular, the work of [37] shows that the central limit theorems (1.12) for M_n and the same result for K_n hold jointly with the same limit variable Z, for the simple reason that for a large class of RAMs with i.i.d. factors, including GEM$(0, \theta)$, the difference $M_n - K_n$ in (1.14) has a limit in distribution as $n \to \infty$, without any centering or scaling. See [37] for further discussion, especially [37, Proposition 5.1] for a pretty formula involving the gamma function for the probability generating function of the large n limit in distribution of $K_{0:n}$ for GEM$(0, \theta)$, and [29] for generalizations to other RAMs. Other recent articles about refined limit theorems for various counting processes derived from RAMs with i.i.d. factors are [47] [46] [45] [1].

We find the normal limit law for M_n derived by sampling from a RAM with i.i.d. factors interesting, because normal limits cannot occur for the maximum of an i.i.d. sequence. Compare with the quite different conclusion of (1.5) for i.i.d. sampling from an exponential distribution, where M_n has a limit in law with just centering and no normalization. So, even though we regard Theorem 1.1 as a discrete analog of the more familiar description of gaps in i.i.d. sampling from an exponential distribution, the limit theorems for M_n implied by these results are quite different. Naively, it might be expected that the discrete analog of the simple structure of gaps in an exponential(λ) sample should be the gaps in sampling from a geometric(p) distribution, which is the RAM (1.7) with deterministic factors $H_i \equiv p$. But apart from some simple results for a sample of size $n = 2$, which are easily seen to hold for any RAM with H_1 independent of $H_2, H_3, \ldots$, the possibilities of various configurations of ties for $n \ge 3$ makes the description of gaps and related statistics for geometric(p) sampling more complicated than might at first be expected. The distribution of the count of missing values $K_{0:n}$ in (1.14), as well as the number L_n of *ties with the maximal value* M_n

$$L_n := n - \max\{j : X_{j:n} < X_{n:n}\} = N^\circ_{M_n:n} = \min\{i : G_{i:n} > 0\} \tag{1.15}$$

have been the subject of many studies of i.i.d. sampling from fixed discrete distributions, especially from the geometric(p) distribution. See for instance [10] [41] and earlier

literature cited there. In sampling from the geometric(p) distribution it is known the distributions of $K_{0:n}$ and L_n remain tight as $n \to \infty$, but that they do not converge, due to a periodic phenomenon which has been extensively studied in this literature. As shown in [37] however, for a large class of RAMs including GEM$(0, \theta)$, the periodic phenomena which arise from i.i.d. geometric(p) sampling get smoothed out very nicely: the count L_n and related statistics such as the $K_{j:n}$ have limits in distribution without any centering or normalization as $n \to \infty$. The idea that such results should be informed by analysis of gaps as well as counts leads to some developments of those limit theorems explored in a sequel [61] of this article.

For RAMs with independent but non-identically distributed factors, our results are more limited. For the two parameter GEM(α, θ) distribution the representation (1.11) of M_n as a sum of independent random variables is no longer valid. Nevertheless we show in Section 6 that in this case the maximum of a size n sample behaves as a random multiple of $n^{\alpha/(1-\alpha)}$ as $n \to \infty$, and in Section 7 we indicate some companion limit theorems for L_n in this case. See also a sequel to this article [62], where for sampling from GEM(α, θ) we derive a result presented here without proof as Theorem 3.4, which gives the distribution of the value-ranked frequencies, meaning the sequence of non-zero components of $(N^\circ_{b:n}, b = 1, 2, \ldots)$.

2 Point processes associated with random discrete distributions

For a random discrete distribution $P_\bullet := (P_j)$ on the positive integers governed by the GRAM (1.7) let

$$F_0 := 0 \text{ and } F_j := \sum_{i=1}^{j} P_i = 1 - \prod_{i=1}^{j} (1 - H_i) \text{ for } k = 1, 2, \ldots .$$

Thinking of $P_\bullet$ as a random discrete distribution on the real line, which happens to be concentrated on positive integers, $F_j = P_\bullet(-\infty, j]$ is the random cumulative distribution function evaluated at positive integers j. In the stick-breaking interpretation, the $F_j \in [0, 1]$ are the *break points*. But we prefer the language of the *stars and bars model*, discussed further in [61]. Following the method developed by Ignatov [44] for GEM$(0, \theta)$, and further developed in [31] [33] [34] for various other models of random discrete distributions, we treat the *bars* F_j as the points of a simple point process N_F on $(0, 1)$, which counts bars not including endpoints of $[0, 1]$. So

$$N_F(a, b] := \sum_{k=1}^{\infty} \mathbf{1}(a < F_k \le b) \qquad (0 \le a < b < 1)$$

is the number of bars in $(a, b]$. The *stars* are the i.i.d. uniform on $[0, 1]$ points $U_1, U_2, \ldots$ which fall between bars and define a random sample $X_1, X_2, \ldots$ from $P_\bullet$ as $X_i = N_F(0, U_i] + 1$. Then

$$\mathbb{P}(X_i \le j \,|\, P_\bullet) = F_j \qquad (i = 1, 2, \ldots, j = 1, 2, \ldots).$$

In terms of Gnedin's *balls in boxes model* of [36], the bars F_j in $(0, 1)$ are the dividers between boxes $[F_{j-1}, F_j)$ which collect the stars (balls) U_i into clusters falling in the same box. For the F_j, the prevailing assumption (1.6) becomes

$$0 < F_1 < F_2 < \cdots \uparrow 1 \qquad \text{almost surely.}$$

It is convenient to make the change of variable from $[0, 1)$ to $[0, \infty)$ by the map $u \mapsto x = -\log(1 - u)$. This is the inverse of the cumulative distribution function

$x \mapsto 1 - e^{-x}$ of a standard exponential variable $\varepsilon := -\log(1-U)$ for U uniform on $(0,1)$. Let $S_0 := 0$ and

$$S_j := -\log(1-F_j) = \sum_{i=1}^{j} -\log(1-H_i) \qquad (j = 1, 2, \ldots).$$

We regard these images S_j of bars F_j as the points of an associated point process N_S on $(0,\infty)$:

$$N_S(s,t] := \sum_{k=1}^{\infty} \mathbf{1}(s < S_k \le t) = N_F(1-\mathrm{e}^{-s}, 1-\mathrm{e}^{-t}] \qquad (0 \le s < t < \infty).$$

Note that the assumption (1.6) translates into $0 < S_1 < S_2 < \cdots \uparrow \infty$ a.s. For convenience we recall the Ignatov's construction of GEM$(0,\theta)$.

Lemma 2.1 ([44])**.** *With above notation, the following conditions are equivalent:*

(i) $P_\bullet$ *has the* GEM$(0,\theta)$ *distribution, meaning the* H_i *are i.i.d.* beta$(1,\theta)$.
(ii) N_F *is an inhomogeneous Poisson process on* $(0,1)$ *with intensity* $\theta\, du/(1-u)$ *at* $u \in (0,1)$.
(iii) N_S *is a homogeneous Poisson process on* $(0,\infty)$ *with intensity* $\theta\, dt$ *at* $t > 0$.
(iv) *The scaled spacings* $(S_k - S_{k-1})/\theta$, $k = 1, 2, \ldots$ *are i.i.d.* exponential(1).

Proof of Theorem 1.1. Whatever the random discrete distribution $P_\bullet$ subject to (1.6), for the sample $(X_1, \ldots, X_n)$ constructed as above, the order statistics of the discrete sample are obtained by counting bars to the left of the order statistics of the corresponding uniform and exponential samples, according to the formula

$$X_{i:n} - 1 = N_F(0, U_{i:n}] = N_S(0, \varepsilon_{i:n}], \tag{2.1}$$

while the gaps between order statistics of the discrete sample are obtained by counting bars between corresponding order statistics of the uniform and exponential samples:

$$G_{i:n} = N_F(U_{n-i:n}, U_{n-i+1:n}] = N_S(\varepsilon_{n-i:n}, \varepsilon_{n-i+1:n}]. \tag{2.2}$$

According to Lemma 2.1 the point process N_S is homogeneous Poisson with rate θ. By construction it is independent of the gaps between exponential order statistics appearing in (2.2), which due to (1.3) are independent and distributed like ε_i/i for $1 \le i \le n$. It follows that the $G_{i:n}$ are independent, with $G_{i:n} \stackrel{d}{=} N_S(0, \varepsilon/i]$ for ε standard exponential independent of N_S. So the distribution of $G_{i:n}$ is the mixture of Poisson(θt) distributions with t assigned the exponential(i) distribution of ε/i. It is well known that such a mixture is geometric(p) for p determined by equating means: $(1-p)/p = \theta/i$. This gives $p = i/(i+\theta)$, and the conclusion follows. □

Proof of Corollary 1.2. For a RAM with i.i.d. factors the point process N_S is a renewal process. For each fixed δ with $0 < \delta < 1$, the number of points F_j such that $F_j \le 1-\delta$ is $N_S(0, -\log(\delta)]$, which is well known to have finite exponential moments. On the other hand, by the law of large numbers for the sampling process, provided $-\log(\delta)$ is chosen to be a continuity point of the renewal measure, with probability one, for sufficiently large n, both the sum of gaps $\sum_{j=\delta n}^{n} G_{j:n}$ and the sum of indicators of non-zero gaps $\sum_{j=\delta n}^{n} \mathbf{1}(G_{j:n} > 0)$ will eventually equal the number of renewals $N_S(0, -\log(\delta)]$, because every relevant box will be occupied. Assuming the gaps are independent, it then follows from the Poisson approximation to sums of independent Bernoulli(p_i) random variables with small parameters p_i that $N_S(0, -\log(\delta)]$ is Poisson distributed. A variation of this

argument shows that N_S has independent, Poisson distributed increments. In other words, N_S is a possibly inhomogeneous Poisson process. But the RAM assumption makes N_S a renewal process. By comparison of the Poisson and renewal decompositions of N_S at its first point S_1, it is easily shown that the point processes that are both Poisson processes and renewal processes are the Poisson processes of constant rate θ for some $\theta > 0$. So the conclusion follows from Ignatov's construction of GEM$(0, \theta)$ in Lemma 2.1. □

As shown in [31] [29] [37] an asymptotic analysis of sampling statistics for a RAM with i.i.d. factors can be made by exploiting properties of the renewal counting process N_S in this case. But for GEM(α, θ) with $0 < \alpha < 1$, the spacings between the S_j are independent but not identically distributed, with $S_j - S_{j-1}$ converging almost surely to 0 as $j \to \infty$, by a simple Borel–Cantelli argument. It follows that in sampling from GEM(α, θ) for $0 < \alpha < 1$ the behavior of the order statistics and gaps is very different from the case $\alpha = 0$, and not approachable by methods of renewal theory.

Both to illustrate applications of Theorem 1.1 in the case $\alpha = 0$, and to motivate study of the GEM order statistics for $0 < \alpha < 1$, before proceeding with that study we first show how Theorem 1.1 leads to some known results about the GEM model, and recall some of its diverse applications.

3 Applications

The GEM acronym was assigned by Warren Ewens [24, p. 321] to acknowledge the work of Griffiths [40], Engen [21] [20] and McCloskey [52] in developing the GEM model for random frequencies in genetics and ecology. The GEM$(0, \theta)$ with i.i.d. beta$(1, \theta)$ factors was studied first, following which the two parametric extension proposed by Engen [20] has also been extensively studied [56, 60, 26], motivated by its appearance in the structure of interval partitions generated by the zeros of various stochastic processes. New models of stationary reversible dynamics for population frequencies consistent with GEM(α, θ) for $0 < \alpha < 1$ have recently been developed and are of continuing interest [57] [11]. The GEM model has also been applied in Bayesian non-parametric statistics as a building block for Bayesian analysis and machine learning algorithms for inferences about clustered data. See the recent review by Crane [12] [13].

3.1 Species sampling

The GEM distributions were first studied in the setting of ecology, where a random discrete distribution $P_\bullet$ may be regarded as a *species abundance model*, meaning an idealized listing of frequencies of an unlimited number of distinct species in a population of unlimited size. To justify the use of infinite models in this setting, some preliminary remarks may be in order. All actual populations are finite, with only a finite number of species. However ecologists discovered that while models with a large finite number of species are quite intractable, remarkable simplifications occur in some particular infinite models. In a sample of n individuals from a large population of size say m, one can suppose that size m population itself is a sample from some ideal infinite population. Finite samples from this population may be assumed to exchangeable, and de Finetti's theorem leads to a representation of consistent models of species sampling from a large population in terms of limiting random frequencies, as shown by Kingman.

In the context of species sampling, in a sample of size n from some population, the data acquired is most naturally regarded as just the partition of n, that is a collection of positive integers that sum to n, obtained by classifying individuals by species, and ignoring any species labels. Such a partition of n can be described in two different ways.

The first is to list the number $N_{i:n}$ of individuals of type i, for $1 \le i \le k$, where $k = K_n$ is the number of distinct species in the sample, and there is some convention for the ordering of types. The second is to provide for each $j \in [n]$ the count

$$K_{j:n} := \sum_{i=1}^{n} \mathbf{1}(N_{i:n} = j) \qquad (j = 1, 2, \ldots) \tag{3.1}$$

of species with j representatives in the size n sample. The total number of distinct species is then

$$K_n = \sum_{j=1}^{n} K_{j:n} \, .$$

The frequencies P_i of species in the ideal infinite population then arise as almost sure limits of $N_{i:n}/n$ as $n \to \infty$, for some consistently chosen ordering of $N_{\bullet:n}$.

It is an awkward aspect of random frequency models that most labeling of frequencies P_i by integers or other countable sets are somewhat arbitrary. There are several possible workarounds for this problem. The easiest way is to list the *ranked sample frequencies*, meaning the numbers of representatives of various species in descending order as

$$N^{\downarrow}_{\bullet:n} = (N^{\downarrow}_{1:n}, \ldots, N^{\downarrow}_{k:n}), \qquad N^{\downarrow}_{1:n} \ge \cdots \ge N^{\downarrow}_{k:n} > 0,$$

which is the decreasing rearrangement of any other listing of the sample frequencies $(N_{1:n}, \ldots, N_{k:n})$. This is also a common way to list parts of partitions in combinatorics. Another way to arrange species is to obtain a size n sample by sampling the population one by one and listing the species in order of their appearance in the sampling process. Given that any particular species in the whole population has m representatives in a sample of size n, the probability of that species being listed first in order of appearance is m/n, by the assumed exchangeability of the sampling process. This leads to the general notion of a *size-biased random permutation* of a finite or countably infinite index set I, or of a collection of components of some kind $C_i, i \in I$ that is indexed by I, for some notion of sizes $V_i := V(C_i)$ of the components being permuted [16] [59]. Typically $V(C_i)$ is the number of elements for a finite set C_i, or some measure such as length for infinite sets C_i like intervals. The size function V of components is subject to the requirement that $V_i > 0$ and that $\Sigma := \sum_i V_i < \infty$, which needs to hold almost surely for a collection of random components $(C_i, i \in I)$. Given such random components $(C_i, i \in I)$, their *size-biased permutation* is a random indexing of these components $(C_{\sigma(i)}, i = 1, 2, \ldots)$, indexed either by the set $[k] := \{1, 2, \ldots, k\}$ if there are a finite number k of components, or by $\mathbb{N}$ if there are an infinite number of them. It is defined by a random bijection σ from either $[k]$ or $\mathbb{N}$ to I, such that $\mathbb{P}[\sigma(1) = j \,|\, C_i, i \in I) = V_j/\Sigma$ for $j \in I$, when $C_{\sigma(1)}$ is called an *size-biased pick* from the components, and for each $m \ge 1$ and $j_1, \ldots, j_m$ with $\Sigma_m := \Sigma - V_{j_1} - \cdots - V_{j_m} > 0$, the next component $C_{\sigma(m+1)}$ is an size-biased pick from the remaining components indexed by $I \setminus \{j_1, \ldots, j_m\}$:

$$\mathbb{P}[\sigma(m+1) = j | C_i, i \in I; \sigma(1) = j_1, \ldots, \sigma(m) = j_m] = V_j/\Sigma_m \text{ for all } j \in \ I \setminus \{j_1, \ldots, j_m\}.$$

With component C_i being a non-empty set of representatives of some species type i in an exchangeable random sample of size n, for some arbitrary labeling of species by $i \in [k]$, the sequence of sample frequencies *in order of appearance*

$$N^{*}_{\bullet:n} = (N^{*}_{1:n}, \ldots, N^{*}_{k:n})$$

is found to be the sequence of sizes $N^{*}_{i:n} = V(C_{\sigma(i)})$ of a size-biased random permutation of the clusters of different species found in the sample, with size $V(C_i) = N_{i:n}$ being

the number of each species present in the sample. The notation $N^*_{\bullet:n}$ is a mnemonic for this size-biasing which is involved in any ordering of species by appearance in an exchangeable process of random sampling. This size-biased listing of sample cluster sizes in order of appearance turns out to be very convenient to work with, especially when the frequencies $P_\bullet$ are in a size-biased random order themselves, an assumption which is quite natural in this context. This assumption holds for GEM(α, θ) model, as shown in [56], which is one of the reasons for use and study of this model.

In the context of species sampling, the values X_i in a size n sample from a random discrete distribution such as GEM(α, θ) seem to be of little interest besides the way that clusters of these values define a partition or a composition of n. However a natural meaning for these sample values X_i and their order statistics $X_{\bullet:n}$ can be provided using the fact that GEM(α, θ) frequencies $P_\bullet$ are already in size-biased random order, hence invariant in distribution under size-biased permutation (ISBP) [59], meaning that

$$(P^*_1, P^*_2, \ldots) \stackrel{d}{=} (P_1, P_2, \ldots)$$

where $(P^*_1, P^*_2, \ldots)$ is a size-biased permutation of $(P_1, P_2, \ldots)$.

Proposition 3.1. *Consider an exchangeable process of species sampling $(Y_1, Y_2, \ldots)$ from random frequencies $P_\bullet$. Split $(Y_1, Y_2, \ldots)$ into an* initial sample $(Y_1, \ldots, Y_n)$ *of size n, followed by a* secondary sample $(Y_{n+1}, Y_{n+2}, \ldots)$ *of unlimited size. For each $1 \le i \le n$ let X_i be the number of species discovered by the secondary sample up to and including discovery of Y_i in the secondary sample. Then $(X_1, \ldots, X_n)$ is a sample of size n from $P^*_\bullet$, a size-biased random permutation of $P_\bullet$. In particular if $P_\bullet$ is* ISBP*, as is the case for* GEM(α, θ)*, then $(X_1, \ldots, X_n)$ is distributed like a sample of size n from $P_\bullet$. Moreover, for a sample $(X_1, \ldots, X_n)$ from $P^*_\bullet$ so constructed, as well as the usual interpretation of $K_{j:n}$ as numbers of species with j representatives in the initial sample, and $K_n = \sum_{j=1}^n K_{j:n}$ the total number of species found by the initial sample, various features of the order statistics in the sample can be interpreted as follows:*

- *the minimum sample value $X_{1:n}$ is the number of species encountered in the secondary sample up to and including when the first species in the primary sample is encountered.*
- *the maximum sample value $M_n := X_{n:n}$ is the number of species encountered in the secondary sample up to and including the first time τ_n that a member of each of the initial K_n species has been encountered in the secondary sample.*
- *the number $K_{0:n} := M_n - K_n$ of missing values below the maximum of the sample, is the number of new species, not present in the initial sample, which are encountered in the secondary sample before this stopping time τ_n.*

Proof. More formally, $X_i = x$ iff $i \in C(x)$, where $C(1), C(2), \ldots$ is the list of clusters of individuals by species in the combined process, in their order of appearance in the secondary sample. So for instance $C(1) := \{i \ge 1 : Y_i = Y_{N(1)}\}$ with $N(1) := n + 1$, $C(2) := \{i \ge 1 : Y_i = Y_{N(2)}\}$, where $N(2) > N(1)$ is the index of the first individual of the second species to appear in the secondary sample, and so on. The relabeling of species by their order of appearance in the secondary sample yields a size-biased permutation $P^*_\bullet$ of $P_\bullet$ with P^*_x the almost sure limiting relative frequency of $C(x)$ for each $x = 1, 2, \ldots$. It then follows by exchangeability that $X_1, \ldots, X_n$ is a sample of size n from $P^*_\bullet$. The interpretations of the various statistics are then straightforward. □

One can also give similar interpretations of the gaps $G_{\bullet:n}$, but this is a bit trickier. For a sample $X_1, \ldots, X_n$ let $N^{X\uparrow}_{\bullet:n}$ denote the list of sample frequencies in *increasing order of X values*, that is the subsequence of all strictly positive counts in the complete list

of counts $N^{\circ}_{\bullet:n}$ derived from the sample $X_1, \ldots, X_n$ as in (1.10). So the count of values equal to $X_{1:n}$ comes first, and the count for $X_{n:n}$ last. In any random sample $X_1, \ldots, X_n$, it is clear from the definition that for any given composition $(n_1, \ldots, n_k)$ of n,

$$N^{X\uparrow}_{\bullet:n} = (n_1, \ldots, n_k)$$

if and only if the sequence of gaps $G_{\bullet:n}$ between order statistics is such that

$$\{1 \le i \le n-1 : G_{i:n} > 0\} \cup \{n\} = \{n_k, n_k + n_{k-1}, n_k + n_{k-1} + n_{k-2}, \ldots, n\}. \tag{3.2}$$

In the species sampling setting of Proposition 3.1, X_i is the number of species discovered by the secondary sample up to and including discovery of the species of the ith initial individual, and $N^{X\uparrow}_{\bullet:n}$ is the listing of cluster sizes in the primary sample in *order of their discovery* by the secondary sample. In this setting the gaps can be described as follows:

- For $\ell \in \{2, \ldots, k\}$, $G_{n_\ell + \cdots + n_k:n}$ is the number of new species encountered in the secondary sample after $\ell - 1$ species of the primary sample are found up to and including the time when ℓ-th species from the primary sample is found.
- $G_{n:n} := X_{1:n} - 1$ is the number of species encountered in the secondary sample before some species present in the primary sample is found.
- All other gaps are zero according to (3.2).

3.2 Some corollaries of Theorem 1.1

We now explain how Theorem 1.1 implies a number of known results about $\mathsf{GEM}(0, \theta)$ samples. Most of these were first discovered in the context of population genetics, where the age-ordering of alleles in a large population provides a natural indexing of allelic types, and the age-ordering of clusters of alleles found in a sample is of interest. The first result is an easy corollary of Theorem 1.1:

Corollary 3.2. (Gnedin-Pitman [34, (3.1)]) *In sampling from* $\mathsf{GEM}(0, \theta)$*, the sequence of indicators of strictly positive gaps* $\mathbf{1}(G_{i:n} > 0)$ *for* $1 \le i \le n$ *has the same distribution as the initial segment of* n *trials in an unlimited sequence* $(B_1, B_2, \ldots)$ *of independent Bernoulli trials with* $\mathbb{P}(B_i = 1) = \theta/(i + \theta)$*:*

$$(\mathbf{1}(G_{i:n} > 0), 1 \le i \le n) \stackrel{d}{=} (B_1, B_2, \ldots, B_n). \tag{3.3}$$

See also the work of Arratia, Barbour and Tavaré [3] [4] [5] [6] for further study of this process of independent Bernoulli trials $(B_1, B_2, \ldots)$ and its relation to the Ewens sampling formula (3.5) below.

Consider now the three random compositions of n defined above in terms of a sample $X_1, \ldots, X_n$ from a random discrete distribution $P_\bullet$:

- the value-ordered sample frequencies $N^{X\uparrow}_{\bullet:n}$,
- the appearance-ordered sample frequencies $N^{*}_{\bullet:n}$, and
- the ranked sample frequencies$N^{\downarrow}_{\bullet:n}$, which are the decreasing rearrangement of either $N^{X\uparrow}_{\bullet:n}$ or of $N^{*}_{\bullet:n}$;

For $P_\bullet$ with $\mathsf{GEM}(0, \theta)$ distribution, from (3.2) and (3.3) it is quite easy to deduce the *Donnelly–Tavaré sampling formula* [18] for the value-ordered frequencies:

$$\mathbb{P}_{0,\theta}(N^{X\uparrow}_{\bullet:n} = (n_1, \ldots, n_k)) = \frac{n!\,\theta^k}{(\theta)_n} \prod_{i=1}^{k} \frac{1}{n_i + \cdots + n_k} \tag{3.4}$$

for every composition $(n_1, \ldots, n_k)$ of n into $k \le n$ parts, with $(x)_n := \Gamma(x+n)/\Gamma(x)$ the Pochhammer symbol. The subscript notation $\mathbb{P}_{\alpha,\theta}$ or $\mathbb{E}_{\alpha,\theta}$ signals that probabilities

or expectations are governed by the GEM(α,θ) model. As indicated by Donnelly and Tavaré, summing (3.4) over all compositions of n with a prescribed weakly decreasing rearrangement yields the celebrated *Ewens sampling formula* [2] [23] for the distribution of the partition of n generated by sampling from GEM$(0,\theta)$. That is, for $K_{j:n}$ the count of clusters of size j as in (3.1), for each weak composition $(m_1,\dots,m_n)$ of k, meaning $m_j \geq 0$ with $\sum_{j=1}^n m_j = k$, with $\sum_{j=1}^n jm_j = n$

$$\mathbb{P}_{0,\theta}(K_{j:n} = m_j, 1 \leq j \leq n) = \frac{n!\,\theta^k}{(\theta)_n} \prod_{j=1}^n \frac{1}{m_j! j^{m_j}}\,. \tag{3.5}$$

It is also easily seen from (3.4) and (3.5) that the composition probability function displayed in (3.4) is also the composition probability function of the appearance-ordered frequencies $N^*_{\bullet:n}$. Recalling from earlier discussion, that in sampling from any $P_\bullet$,

$$N^*_{\bullet:n} \text{ is a size-biased permutation of } N^{X\uparrow}_{\bullet:n}.$$

the Donnelly–Tavaré formula implies the following very special property of GEM$(0,\theta)$, in which the roles of $N^*_{\bullet:n}$ and $N^{X\uparrow}_{\bullet:n}$ can be reversed.

Corollary 3.3 (Donnelly and Tavaré [18, (4.4)])**.** *In a sampling from* GEM$(0,\theta)$*, the sample frequencies in value-order $N^{X\uparrow}_{\bullet:n}$ and in appearance-order $N^*_{\bullet:n}$ are identically distributed. Their common distribution is described by formula* (3.4)*. Consequently, in sampling from* GEM$(0,\theta)$*,*

$$N^{X\uparrow}_{\bullet:n} \textit{ is a size-biased permutation of } N^*_{\bullet:n}.$$

The question of how to extend this result to GEM(α,θ) for $0<\alpha<1$ led us combine the representation of sampling from GEM(α,θ) provided by Proposition 3.1 with the known description of the distribution of $N^*_{\bullet:n}$ for GEM(α,θ) [58] [60, §3.2], to obtain the following theorem, whose proof will be detailed elsewhere [62].

Theorem 3.4. *In sampling from* GEM(α,θ)*, for all* $0 \leq \alpha < 1$

$$N^{X\uparrow}_{\bullet:n} \textit{ is a (size}-\alpha\textit{)-biased permutation of } N^*_{\bullet:n}. \tag{3.6}$$

The meaning of (3.6) is that given $N^*_{\bullet:n} = (n_1,\dots,n_k)$, the frequency $N^{X\uparrow}_{1:n}$ of the minimal value is distributed like a random choice of $(n_1,\dots,n_k)$, with n_i chosen with probability $(n_i-\alpha)/(n-k\alpha)$, and so on, as in the general definition of a size-biased permutation, just with the usual size n_i of a cluster replaced by $n_i-\alpha$. In particular, for $0<\alpha<1$ and $n\geq 3$ the two random compositions $N^{X\uparrow}_{\bullet:n}$ and $N^*_{\bullet:n}$ are not identically distributed. Remarkably, the conclusion (3.6) holds not only for GEM(α,θ), but also for the sampling from $P_\bullet$ the size-biased presentation of frequencies in any of the Gibbs(α) models introduced in [60, Theorem 4.6] and studied further in [32].

We note that Theorem 1.1 for GEM$(0,\theta)$ yields also the following further corollary, which identifies the known distribution of the minimal order statistic $X_{1:n}$. This can be read from a result for the infinitely many alleles model, due to Saunders, Tavaré, and Watterson [67, Theorem 8] and the fact that this model generates age-ordered GEM$(0,\theta)$ frequencies. See Donnelly and Tavaré [18] and Donnelly [15, Proposition 3.5] for further discussion. The independence assertion in the corollary is easily shown to hold in sampling from any RAM with i.i.d. factors, due to the regenerative property of these models discussed in [31].

Corollary 3.5. *In sampling from* GEM$(0,\theta)$*, the minimum of the sample, $X_{1:n} = 1 + G_{n:n}$ has a shifted geometric$(n/(n+\theta))$ distribution, and $X_{1:n}$ is independent of the pair of random compositions $N^{X\uparrow}_{\bullet:n}$ and $N^*_{\bullet:n}$, hence also independent of the Ewens(θ) distributed partition of n generated by the sample.*

Proof. The independence of $G_{n:n}$ and $N^{X\uparrow}_{\bullet:n}$ is clear, because $N^{X\uparrow}_{\bullet:n}$ is a function of the $G_{i:n}$ $1 \le i \le n-1$. But by exchangeability, conditionally given the entire collection of order statistics, $N^{*}_{\bullet:n}$ is just a size-biased permutation of $N^{X\uparrow}_{\bullet:n}$. So the order statistics and $N^{*}_{\bullet:n}$ are conditionally independent given $N^{X\uparrow}_{\bullet:n}$, from which the conclusion follows easily. □

3.3 Combinatorial limit theorems

Combinatorial models often involve exchangeable random partitions of $[n]$ into a collection of subsets of various sizes, typically connected components of a graph associated with the model, such as the cycles of a permutation, trees in a forest, or connected components of a mapping digraph. It is known [4] [60] that in many models for such a combinatorial structure picked uniformly at random, the sequence $N^{\downarrow}_{\bullet:n}$ of ranked component sizes converges in law after scaling by n:

$$n^{-1}(N^{\downarrow}_{1:n}, N^{\downarrow}_{2:n}, \ldots) \xrightarrow{d} (P^{\downarrow}_1, P^{\downarrow}_2, \ldots) \sim \mathrm{PD}(\alpha, \theta), \qquad n \to \infty, \tag{3.7}$$

for some (α, θ), where $\mathrm{PD}(\alpha, \theta)$, the *Poisson–Dirichlet distribution with parameters* (α, θ) is the distribution of the decreasing rearrangement of $\mathsf{GEM}(\alpha, \theta)$ [63]. According to the general theory of such limit distributions [17] [35], this is equivalent to the corresponding convergence

$$n^{-1}(N^{*}_{1:n}, N^{*}_{2:n}, \ldots) \xrightarrow{d} (P^{*}_1, P^{*}_2, \ldots) \sim \mathsf{GEM}(\alpha, \theta), \qquad n \to \infty, \tag{3.8}$$

for the size-biased reordering $N^{*}_{\bullet:n}$ of the component sizes, where the limit has the ISBP $\mathsf{GEM}(\alpha, \theta)$ distribution. The treatment of [4] presents a large number of such examples with $\alpha = 0$. An example with $\alpha = \theta = 1/2$ is provided by the tree components of a uniform random mapping digraph [60, (9.7)]. There are many similar examples, with ranked and size-biased compositions of n derived from other constructions. For instance, if the partition of n is the decreasing arrangement of lengths of excursions of an aperiodic Markov chain away from some recurrent state 0 run for n steps, and the return time of the state is in the domain of attraction of the stable law of index $\alpha \in (0, 1)$, then it is known [63] that (3.7) holds for this α with $\theta = 0$, for the Markov chain started in state 0, and with the same α with $\theta = \alpha$ for the Markovian bridge obtained by further conditioning to return to state 0 at a late time n. In this setting, the lengths of excursions are most naturally listed in their order of creation by the Markov chain, which is neither ranked nor size-biased. Still, the analysis of such limit laws for excursions is assisted by the device of deliberately size-biasing the order of excursions, to create a $\mathsf{GEM}(\alpha, \theta)$ limit as in (3.8) which is much easier to deal with than the $\mathrm{PD}(\alpha, \theta)$ limit of ranked lengths.

In any of these settings where $\mathrm{PD}(\alpha, \theta)$ and $\mathsf{GEM}(\alpha, \theta)$ arise hand-in-hand as limit laws for ranked and size-biased counts of some kind, it was shown by Kingman [50] that the structure of the limit theorems extends to corresponding limit theorems for sampling components of the structure of size n. To set this up, consider a limited sample of n elements from a much larger set of size m supporting a combinatorial structure whose ranked relative component sizes $m^{-1}N^{\downarrow}_{\bullet:m}$ are well approximated in distribution by $\mathrm{PD}(\alpha, \theta)$ for some α, θ. Let each component in the structure of size m be labeled by its index of appearance in a size-biased listing of components. If the components of the combinatorial structure generate an exchangeable partition of $[m]$, these labels can be assigned to components in order of their least elements. If the components of the combinatorial structure do not generate an exchangeable partition of $[m]$, as in the case of excursion lengths of a Markov chain run for m steps, let the component labels be assigned by a size-biased random permutation of component sizes, as in Section 3.1. The following proposition is an easy consequence of Kingman's theory of partition structures [50] [17] [35]:

Proposition 3.6. *Let $U_i^{[m]}$ for $1 \le i \le n < m$ be a simple random sample of size n from $[m]$, either with or without replacement, and let $X_i^{[m]}$ be the label of the component of the combinatorial structure that contains $U_i^{[m]}$, for a size-biased labeling of components, that is independent of the random sample $U_i^{[m]}, 1 \le i \le n$. Suppose there is the convergence in distribution* (3.8) *of size-biased relative component sizes to* GEM(α, θ) *with n replaced by $m \to \infty$. Then for each fixed n there is convergence of joint distributions*

$$(X_i^{[m]}, 1 \le i \le n) \xrightarrow{d} (X_i, 1 \le i \le n) \text{ a sample from GEM}(\alpha, \theta) \text{ as } m \to \infty,$$

which implies also convergence in distribution of corresponding order statistics, counts and gaps to those derived from the GEM *sample.*

This proposition shows how any exact result for a sample of size n from a GEM can be turned into the conclusion of a limit theorem for sampling from various random combinatorial structures. Just that a double sampling process is involved, much as in Proposition 3.1, which acquires further interpretations in this context. The sample of size n may be regarded as an initial sample of size n, as in Proposition 3.1. Then there needs to be a secondary sample, to generate a size-biased labeling of components, run at least long enough to allocate a label to every component that intersects the initial sample. The limit in distribution as $m \to \infty$ of the list of secondary labels found in the primary sample of size n is then a size n sample from GEM(α, θ). The case of exchangeable partitions is particularly natural, as the initial sample of size n can be taken to be the set $[n]$ instead of a random subset of size n, and the secondary labeling of components can be taken to be the order of least elements of components, starting the labeling afresh after the initial sample of size n, as in Proposition 3.1. As $m \to \infty$ there is a negligible difference between this construction and a completely independent size-biased listing of components, so the conclusions of the above Proposition are valid in either setup.

For application to the excursions of a Markov chain, given the path of the Markov chain of length m, due to lack of exchangeability, two random samples are required, one to choose n sample times from $[m]$, and the other to assign secondary labels to excursions in their order of discovery by a random permutation of $[m]$. As in the exchangeable case, it makes no difference if the second random sample is just a continuation of the first, avoiding the first n elements drawn and continuing without replacement until all m time points have been covered, and all the excursions found. In this scenario there is a tiny probability that the first n time points sampled might find excursions which were not part of the subsequent sample, and hence to which no label can be assigned, but the probability of this event and all other differences in the distribution of the first n sample labels is negligible in the large m limit, because the assumption of convergence to GEM(α, θ) proper frequencies means that with overwhelming probability these first n sample points will fall in some collection of K_n excursions each of which acquires a significant fraction of the rest of the sample points in $[m]$ and $m \to \infty$.

The interpretation in this setting of large n limit theorems for sampling a GEM distribution is not immediate. But such interpretations might still be made with adequate provision of a limit regime with both n and m tending to infinity. To adequately justify such a double limit theorem, some estimate of the adequacy of the approximation of $m^{-1} N^*_{\bullet:m}$ by GEM(α, θ) would be required, such as that provided in [5] for the random permutation statistics approaching GEM$(0, 1)$.

4 The maximum of a sample from a random discrete distribution

This section develops some general formulas for the distribution of the maximum of a sample from a random discrete distribution on positive integers. These formulas allow

us to check Corollary 1.3 without appeal to Ignatov's representation of GEM$(0,\theta)$. Our interest in this approach is that it at least gives us an explicit if difficult formula for the distribution of the maximum of a sample from GEM(α,θ).

We begin with a well known representation of the probability generating function of a discrete random variable in terms of its tail probabilities.

Lemma 4.1 ([25, p. 265, Theorem 1])**.** *For X a non-negative integer valued random variable, the probability generating function*

$$\mathbb{E}z^X := \sum_{n=0}^{\infty} \mathbb{P}[X=n]z^n$$

may be represented for $|z|<1$ as

$$\mathbb{E}z^X = 1-(1-z)\sum_{m=0}^{\infty}\mathbb{P}[X>m]z^m \tag{4.1}$$

$$= (1-z)\sum_{m=0}^{\infty}\mathbb{P}[X\le m]z^m. \tag{4.2}$$

This allows us to provide the following general expression for the distribution of the maximum of a sample from a random discrete distribution:

Lemma 4.2. *Let $M_n = \max_{1\le k\le n} X_k$ for a sequence of exchangeable positive integer valued random variables $X_1,\dots,X_n$ which are conditionally i.i.d. $P_\bullet$ given some random discrete distribution $P_\bullet$ with $R_k := 1-\sum_{j=1}^{k} P_j \downarrow 0$ a.s. Then the probability generating function of M_n-1 admits the representation*

$$\mathbb{E}z^{M_n-1} = (1-z)\sum_{j=0}^{n}\binom{n}{j}(-1)^j\sum_{k=1}^{\infty}\mathbb{E}R_k^j\, z^{k-1}. \tag{4.3}$$

Proof. We apply (4.2) to $X=M_n-1$. For $k=1,2,\dots$ the term for $m=k-1$ is evaluated by taking expectations in the following identity:

$$\mathbb{P}[M_n-1\le k-1|P_\bullet] = \mathbb{P}[M_n\le k|P_\bullet] = (1-R_k)^n = \sum_{j=0}^{n}\binom{n}{j}(-1)^j R_k^j.$$

Now (4.3) follows easily from (4.2). □

Since the GEM(α,θ) model makes the $1-H_i$ independent with beta$(\theta+i\alpha,1-\alpha)$ distributions, for $j=0,1,\dots$

$$\mathbb{E}_{\alpha,\theta}(1-H_i)^j = \frac{B(\theta+i\alpha+j,1-\alpha)}{B(\theta+i\alpha,1-\alpha)} = \frac{(\theta+i\alpha)_j}{(\theta+(i-1)\alpha+1)_j}$$

hence the GEM(α,θ) tail moment formula

$$\mathbb{E}_{\alpha,\theta}R_k^j = \prod_{i=1}^{k}\frac{(\theta+i\alpha)_j}{(\theta+(i-1)\alpha+1)_j}. \tag{4.4}$$

Thus we obtain:

Proposition 4.3. *The probability generating function of M_n-1 for the maximum M_n of a sample of size n from* GEM(α,θ) *is given by formula* (4.3) *for the* GEM(α,θ) *tail moments* (4.4)*.*

For $j = 1$ the product in (4.4) gives the tail probability formula

$$\mathbb{P}_{\alpha,\theta}[X_1 > k] = \prod_{i=1}^{k} \frac{\theta + i\alpha}{1 + \theta + (i-1)\alpha} \tag{4.5}$$

for X_1 a sample of size 1 from $\mathsf{GEM}(\alpha, \theta)$. For $\alpha = 0$ the product reduces to $(\theta/(1+\theta))^k$. This geometric distribution of X_1 with parameter $1/(1+\theta)$ was indicated by Engen [21] in the context of ecological models. Formula (4.3) in this case reduces easily to the familiar formula for the probability generating function of the geometric distribution of $X_1 - 1$ on non-negative integers,

$$\mathbb{E}_{0,\theta} z^{X_1 - 1} = \frac{1}{1 - \theta(z-1)} = 1 + \theta(z-1) + \theta^2(z-1)^2 + \cdots$$

whose binomial moments can be read from the expansion in powers of $(z-1)$:

$$\mathbb{E}_{0,\theta} \binom{X_1 - 1}{k} = \theta^k \qquad (k = 0, 1, \ldots). \tag{4.6}$$

For $0 < \alpha < 1$ the tail probabilities (4.5) may be recognized as the terms of a hypergeometric series. This allows the following evaluation in terms of the Gaussian hypergeometric function ${}_2F_1$:

$$\mathbb{E}_{\alpha,\theta} z^{X_1 - 1} = 1 - \frac{(\alpha + \theta)}{(1 + \theta)} (1 - z) \, {}_2F_1 \left(\begin{matrix} 1, \, 2 + \theta/\alpha \\ 1 + (1+\theta)/\alpha \end{matrix} ; z \right)$$

with associated binomial moments

$$\mathbb{E}_{\alpha,\theta} \binom{X_1 - 1}{k} = \prod_{i=1}^{k} \frac{\theta + i\alpha}{1 - (i+1)\alpha} \qquad \text{if } 0 \le \alpha < \frac{1}{k+1} \tag{4.7}$$

and ∞ otherwise. Note that (4.7) reduces correctly to (4.6) for $\alpha = 0$. The case $k = 1$ of (4.7) is due to Kingman [49, (18)] for $\alpha = 0$ and [49, (58)] for $\theta = 0$. The case of (4.7) for $\theta = 0$ and general $k = 1, 2, \ldots$ was recently derived by Leisen, Lijoi, and Paroissin [51, Proposition 1] using a much more difficult approach.

For $j > 1$ only in the case $\alpha = 0$ does there seem to be much simplification in (4.3). Then we can proceed as follows:

Computational proof of Corollary 1.3. For $\alpha = 0$ in (4.4) we find that

$$\mathbb{E}_{0,\theta} R_k^j = \left(\frac{\theta}{\theta + j} \right)^k$$

and the series in (4.3) becomes

$$\sum_{k=1}^{\infty} \mathbb{E}_{0,\theta} R_k^j z^{k-1} = z^{-1} \sum_{k=1}^{\infty} \left(\frac{\theta z}{\theta + j} \right)^k = \frac{\theta}{j + \theta(1-z)}$$

hence

$$\mathbb{E}_{0,\theta} z^{M_n - 1} = (1-z) \sum_{j=0}^{n} \binom{n}{j} \frac{(-1)^j \theta}{j + \theta(1-z)} = \prod_{i=1}^{n} \frac{i}{i + \theta(1-z)} \, . \tag{4.8}$$

The last equality is the well-known partial fraction decomposition (see, e.g. [39, Eq. 5.41])

$$\frac{1}{(x)_{n+1}} = \frac{1}{n!} \sum_{j=0}^{n} \binom{n}{j} \frac{(-1)^j}{x + j} \tag{4.9}$$

which can be verified, for instance, by multiplying (4.9) by $x+k$ and plugging in $x=-k$ for $k=0,1,\dots,n$. Since the factors in the right-hand side of (4.8) are the probability generating functions of geometric variables G_i with parameters $i/(i+\theta)$, the conclusion of Corollary 1.3 follows. □

Remark 4.4. Looking on the form of (1.11) it is tempting to suppose that G_n is the difference $M_n - M_{n-1}$ and is independent of M_{n-1}. However this is not the case, because the new sample U_{n+1} gets into an arbitrary position ℓ in the order statistics and hence changes a value of $G_{n-\ell}$. Moreover, unlike the independent case, the successive maxima do not form a Markov chain. Heuristically, this happens because knowledge of the history provides some information about the realization of $\mathbf{Y}$. It can be shown, for instance, that $\mathbb{P}\big[M_1=j, M_2=M_3=\ell | M_1=j, M_2=\ell\big]$ for $j<\ell$ depends on j, but we omit this calculation. A similar issue arises in the identity in distribution (1.4) relating the distribution of the maximum M_n of n i.i.d. exponential variables to the sum T_n of scaled exponentials. But that identity fails to hold jointly as n varies for a more obvious reason: $\mathbb{P}(M_n = M_{n-1}) > 0$ while $\mathbb{P}(T_n = T_{n-1}) = 0$.

5 A generalization in the GEM$(0,\theta)$ case

The result of Corollary 1.3 can be generalized as follows. According to (2.1), the GEM$(0,\theta)$ model makes $M_n - 1 = N_F(0, U_{n:n}]$ a sum of independent geometrics, where $N_F := (N_F(0,u], 0 \le u < 1)$ is the GEM$(0,\theta)$ barrier process, which is Poisson with intensity $\theta(1-u)^{-1}du$ at $u \in (0,1)$, and $U_{n:n}$ is independent of N_F. Instead of $N_F(0, U_{n:n}]$, consider $N_F(0,\beta]$ for β with a suitable beta distribution, independent of N_F.

Theorem 5.1. *For $n \in \mathbb{N}$ and $\theta, b > 0$, let $\beta_{n,b}$ with the* beta(n,b) *density at u proportional to $u^{n-1}(1-u)^{b-1}$ be independent of the* GEM$(0,\theta)$ *barrier process N_F. Then*

$$N_F(0,\beta_{n,b}] \stackrel{d}{=} \sum_{i=1}^{n} G_i(b,\theta) \tag{5.1}$$

where the $G_i(b,\theta)$ are independent with geometric$(p_i(b,\theta))$ distributions, for

$$p_i(b,\theta) := \frac{b+i-1}{b+i-1+\theta}. \tag{5.2}$$

Proof. Consider first $N_F(0,W]$ where W is a random variable with some arbitrary distribution on $[0,1]$, independent of N_F. For $W=u$ fixed, the distribution of $N_F(0,u]$ is Poisson$(-\theta\log(1-u))$ with the probability generating function

$$\mathbb{E}z^{N_F(0,u]} = \exp\left[-(1-z)(-\theta\log(1-u)\right] = (1-u)^{\theta(1-z)}\,.$$

For general W the distribution of $N_F(0,W]$ ranges over all mixed Poisson distributions. Explicitly, the probability generating function of W is

$$\mathbb{E}z^{N_F(0,W]} = \mathbb{E}(1-W)^{\theta(1-z)}.$$

In particular, if $W=\beta_{a,b}$ has the beta(a,b) distribution then

$$\begin{aligned}\mathbb{E}z^{N_F(0,\beta_{a,b}]} &= \mathbb{E}(1-\beta_{a,b})^{\theta(1-z)} = \mathbb{E}\beta_{b,a}^{\theta(1-z)} \\ &= \frac{\Gamma(b+\theta(1-z))}{\Gamma(a+b+\theta(1-z))}\frac{\Gamma(a+b)}{\Gamma(b)}.\end{aligned}$$

If $a=n$ is a positive integer, then $\frac{\Gamma(n+b)}{\Gamma(b)} = (b)_n := \prod_{i=1}^{n}(b+i-1)$ so

$$\mathbb{E}z^{N_F(0,\beta_{n,b}]} = \prod_{i=1}^{n}\frac{b+i-1}{b+i-1+\theta(1-z)} = \prod_{i=1}^{n}\frac{p_i(b,\theta)}{1-(1-p_i(b,\theta))z}$$

for $p_i(n,\theta)$ as in (5.2). Since the i-th factor is the probability generating function for the geometric($p_i(n,\theta)$) distribution, the claim (5.1) follows. □

Remark 5.2. Notice that $U_{n,n} \stackrel{d}{=} \beta_{n,1}$, so (5.1) is a generalization of (1.11).

6 The maximum of a GEM(α,θ) sample for $0<\alpha<1$

The technique of the previous sections does not seem to work for the case $0<\alpha<1$. However the asymptotics of the GEM distribution in this case are known sufficiently well to find the asymptotic behavior of M_n as $n\to\infty$. In particular, it is known that GEM(α,θ) frequencies P_i almost surely decay as random factors of $i^{-1/\alpha}$. Similar behavior is also known for the sampling from a random branching process model introduced by Robert and Simatos [66] where different but similar aspects of samples are studied, such as the limit behavior of the first unoccupied box, as the sample size grows.

A key role in the study of the GEM(α,θ) distribution for the case $0<\alpha<1$ is played by the notion of the α-diversity of the exchangeable sample. It is known [60, Th. 3.8] that for K_n defined by (1.13) from a GEM(α,θ) sample with $0<\alpha<1$ and $\theta>-\alpha$ there exists a limit

$$\lim_{n\to\infty}\frac{K_n}{n^\alpha}=D_\alpha>0 \text{ almost surely } (\mathbb{P}_{\alpha,\theta})$$

and also in p-th mean for every $p>0$. The distribution of the limiting random variable D_α, which depends on θ, is known as the *α-diversity* and is determined by its moments

$$\mathbb{E}_{\alpha,\theta}D_\alpha^p=\frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)}\,\frac{\Gamma(p+\frac{\theta}{\alpha}+1)}{\Gamma(p\alpha+\theta+1)}\,. \tag{6.1}$$

Moreover, the α-diversity D_α is a.s. determined by $P_\bullet$ and

$$\mathbb{P}_{\alpha,\theta}[X_1>k|P_\bullet]\sim\alpha D_{\alpha,\theta}^{1/\alpha}\,k^{1-1/\alpha} \text{ almost surely } (\mathbb{P}_{\alpha,\theta}) \text{ as } k\to\infty,$$

see [28, Sec. 10] or [60, Lemma 3.11]. For such a power law it is well known that the maximum of an independent sample of size n converges in distribution to the Fréchet distribution. Namely, writing for short $\gamma=1/\alpha-1$, for any fixed $x>0$

$$\begin{aligned}\mathbb{P}_{\alpha,\theta}\big[M_n\le xn^{1/\gamma}\;\big|\;P_\bullet\big]&=\big(1-\mathbb{P}_{\alpha,\theta}\big[X_1>xn^{1/\gamma}\;\big|\;P_\bullet\big]\big)^n\\&\sim\left(1-\alpha D_{\alpha,\theta}^{1/\alpha}\frac{x^{-\gamma}}{n}\right)^n \text{ almost surely } (\mathbb{P}_{\alpha,\theta})\\&\to\exp\big(-\alpha D_{\alpha,\theta}^{1/\alpha}x^{-\gamma}\big),\qquad n\to\infty.\end{aligned}$$

Hence, by integration with respect to the distribution of the α-diversity, we have the following result.

Theorem 6.1. *Let M_n be the maximum of a size n GEM(α,θ) exchangeable sample with $0<\alpha<1$ and $\theta>-\alpha$. Then for each $x>0$*

$$\mathbb{P}_{\alpha,\theta}\big[M_n\le xn^{\alpha/(1-\alpha)}\big]\to\mathbb{E}_{\alpha,\theta}\exp\big(-\alpha D_\alpha^{1/\alpha}x^{-(1-\alpha)/\alpha}\big) \text{ as } n\to\infty. \tag{6.2}$$

Remark 6.2. For the case $\alpha=0$ the asymptotic result $K_n\sim M_n\sim\theta\log n$ almost surely $(\mathbb{P}_{0,\theta})$ of [37] shows that asymptotically K_n and M_n have the same behavior. For $\alpha>0$ the situation is different: K_n should be divided by n^α to get a proper limit, and M_n grows much faster as a random factor of $n^{\alpha/(1-\alpha)}$.

Note that (6.2) expresses the cumulative distribution function of $\lim n^{-\alpha/(1-\alpha)}M_n$ evaluated at x as the Laplace transform $\mathbb{E}_{\alpha,\theta}\big[\mathrm{e}^{-yD_\alpha^{1/\alpha}}\big]$ evaluated at $y=\alpha x^{-(1-\alpha)/\alpha}$. Since the $\mathbb{P}_{\alpha,\theta}$ moments of D_α given by (6.1) determine its distribution, we can obtain an explicit but clumsy expression for the limiting distribution function (6.2).

Theorem 6.3. *For the $\mathbb{P}_{\alpha,\theta}$ distribution of D_α determined by the moment function* (6.1)*,*

$$\mathbb{E}_{\alpha,\theta}\exp\bigl(-\alpha D_{\alpha,\theta}^{1/\alpha}x^{-(1-\alpha)/\alpha}\bigr) = \frac{2\alpha^{1-\theta-\alpha}\,\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)}x^{(1-\alpha)(\theta/\alpha+1)}\int_0^\infty v^{\theta+2\alpha-1}\mathrm{e}^{-(v^2/\alpha)^\alpha x^{1-\alpha}}J_\theta(2v)\,dv, \qquad (6.3)$$

where J_θ is the Bessel function.

Proof. Writing for short $y=\alpha x^{-(1-\alpha)/\alpha}$ we have, for any $c>0$,

$$\mathrm{e}^{-yD_\alpha^{1/\alpha}} = \frac{1}{2\pi\mathrm{i}}\int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty}\Gamma(s)\bigl(yD_\alpha^{1/\alpha}\bigr)^{-s}ds$$

because e^{-y} and $\Gamma(s)$ form the Mellin pair. We refer to [54] for the necessary information about Mellin's transform. By analyticity the expression (6.1) for $\mathbb{P}_{\alpha,\theta}$ moments of D_α is valid also for complex p at least with $\mathrm{Re}\,p>-1-\frac{\theta}{\alpha}$. Hence taking expectation and applying Fubini's theorem yields

$$\mathbb{E}_{\alpha,\theta}\bigl[\mathrm{e}^{-yD_{\alpha,\theta}^{1/\alpha}}\bigr] = \frac{1}{2\pi\mathrm{i}}\frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\alpha}+1)}\int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty}\Gamma(s)\,\frac{\Gamma(\frac{\theta-s}{\alpha}+1)}{\Gamma(\theta-s+1)}y^{-s}ds \qquad (6.4)$$

for $0<c<\alpha+\theta$. Now, $\Gamma(s)/\Gamma(\theta-s+1)$ is the Mellin transform of $y^{-\theta/2}J_\theta(2\sqrt{y})$ in the fundamental strip $0<\mathrm{Re}\,s<\frac{\theta}{2}+\frac{3}{4}$ ([54, II.5.38], where there is a misprint in the right bound) and $\Gamma(\frac{\theta-s}{\alpha}+1)$ is the Mellin transform of $\alpha y^{-\alpha-\theta}\mathrm{e}^{-y^{-\alpha}}$ for $\mathrm{Re}\,s<\alpha+\theta$, by the standard transformations of the Mellin pair e^{-y} and $\Gamma(s)$. Hence their product in the intersection of fundamental strips is the Mellin transform of the multiplicative convolution and for $0<c<\min\{\frac{\theta}{2}+\frac{3}{4},\alpha+\theta\}$ by the inversion formula

$$\frac{1}{2\pi\mathrm{i}}\int_{c-\mathrm{i}\infty}^{c+\mathrm{i}\infty}\Gamma(s)\,\frac{\Gamma(\frac{\theta-s}{\alpha}+1)}{\Gamma(\theta-s+1)}y^{-s}ds = \alpha\int_0^\infty (y/u)^{-\alpha-\theta}\mathrm{e}^{-(y/u)^{-\alpha}}u^{-\theta/2}J_\theta(2\sqrt{u})\,\frac{du}{u}\,.$$

Plugging this into (6.4), changing the variable $v=\sqrt{u}$ and returning to the variable x yields the result. □

The right-hand side of (6.3) does not seem to allow much simplification for general α. For some rational α Mathematica evaluates this integral in terms of the hypergeometric function. However for $\alpha=1/2$ the integral can be taken explicitly and leads to a simple expression. In this case the integral is the Mellin transform of $f(v)=\mathrm{e}^{-v\sqrt{2x}}J_\theta(2v)$ evaluated at $\theta+1$. According to [54, I.10.7]

$$\int_0^\infty v^{s-1}\mathrm{e}^{-v\sqrt{2x}}J_\theta(2v)\,dv = (2x)^{-(s+\theta)/2}\frac{\Gamma(\theta+s)}{\Gamma(\theta+1)}\,{}_2F_1\left(\begin{matrix}\frac{\theta+s}{2},\,\frac{\theta+s+1}{2}\\ \theta+1\end{matrix};-\tfrac{2}{x}\right)$$

and the last expression simplifies for $s=\theta+1$ because

$${}_2F_1\left(\begin{matrix}a,\,b\\ b\end{matrix};z\right) = \sum_{k=0}^\infty\frac{(a)_k}{k!}z^k = \frac{1}{(1-z)^a}$$

for $|z|<1$ and by analyticity also for all z with $\mathrm{Re}\,z<1$. Hence

$$\int_0^\infty v^\theta\mathrm{e}^{-v\sqrt{2x}}J_\theta(2v)\,dv = \frac{\Gamma(2\theta+1)}{\Gamma(\theta+1)}\,\frac{1}{(2x+4)^{\theta+1/2}}$$

and plugging it into (6.3) gives the following result.

Corollary 6.4. *Let M_n be the maximum of a size n $\mathsf{GEM}(\frac{1}{2},\theta)$ exchangeable sample with $\theta > -\frac{1}{2}$. Then M_n/n converges in distribution as $n \to \infty$ to a random variable with the cumulative distribution function $\big(x/(x+2)\big)^{\theta+1/2}$.*

Some simplification is also possible for rational $\alpha \neq 1/2$ using the representation of the α-diversity in terms of product of random variables with beta and gamma distributions given in [48, Sec. 8].

7 Limit laws for the number of missing values and number of ties at the maximum

This section offers some complements to the analysis of limit laws for M_n in $\mathsf{GEM}(0,\theta)$ settings, following the work of Gnedin et al. [37].

It was observed in other notation in [37, (19)] that in sampling from $\mathsf{GEM}(0,\theta)$, for the number of missing values in the range $K_{0:n} := M_n - K_n$ there is the representation in distribution (1.14) in terms of independent geometric(p_i) random variables $G_{i:n}$ with $p_i := i/(\theta+i)$. Writing now $G(p_i)$ instead of $G_{i:n}$ to emphasize the lack of dependence on n in this representation, apart from the trivial term $G_{n:n}$ which obviously converges almost surely to 0 as $n \to \infty$, we deduce easily that

$$K_{0:n} \stackrel{d}{\to} K_{0:\infty} := \sum_{i=1}^{\infty} (G(p_i) - 1)_+ \tag{7.1}$$

This is just an explicit presentation of a random variable with the limit distribution of $K_{0:n}$ as $n \to \infty$ which was described in [37, Proposition 5.1] by the probability generating function

$$g_\theta(z) := \mathbb{E} z^{K_{0:\infty}} = \frac{\Gamma(1+\theta)\Gamma(1+(1-z)\theta)}{\Gamma(1+(2-z)\theta)} \qquad (|z| \leq 1) \tag{7.2}$$

which can be read from (7.1) as an infinite product of modified geometric generating functions. This product simplifies to (7.2) due to the Weierstrass product formula for the gamma function [22, Eq. (1.1.3), p. 1]. As observed in [37, Proposition 5.1], this distribution of $K_{0:\infty}$ is a mixed Poisson distribution with random parameter distributed as $\theta|\log \beta_{1,\theta}|$ for $\beta_{1,\theta}$ with the beta$(1,\theta)$ distribution of P_1, the first $\mathsf{GEM}(1,\theta)$ frequency.

That result may by understood as a refinement of (7.1) in which each term $(G(p_i)-1)_+$ is replaced by the distributionally equivalent random variable $N_i(\theta B_{1-p_i}\varepsilon_i/i)$ where the ε_i are independent standard exponential variables, the B_{1-p_i} are independent Bernoulli variables with the indicated parameters, independent also of the ε_i, and the N_i are independent rate one Poisson processes independent of both the ε_i and the B_{1-p_i}. Then there is the identity in distribution

$$\sum_{i=1}^{\infty} B_{1-p_i} \frac{\varepsilon_i}{i} \stackrel{d}{=} -\log \beta_{1,\theta} \text{ where } 1 - p_i = \theta/(i+\theta) \tag{7.3}$$

which can be checked by computing the Laplace transform of both sides at $\lambda > 0$. This identity (7.3) is the instance $a = 1, b = \theta$ of the identity (7.4) presented in the following proposition, which is the simpler variant for $\log$ beta variables of a representation of $\log$ gamma variables due to Gordon [38]. This identity can be found in a not easily accessible text by Pakes [55, (32.17)].

Proposition 7.1. *For each $a, b > 0$, for $0 < \beta_{a,b} < 1$ with density proportional to $u^{a-1}(1-u)^{b-1}$ at $0 < u < 1$ there is the identity in distribution*

$$\sum_{j=0}^{\infty} B_{b/(a+b+j)} \frac{\varepsilon_j}{a+j} \stackrel{d}{=} -\log \beta_{a,b}, \tag{7.4}$$

where the ε_j are i.i.d. standard exponential variables, independent of a sequence of independent Bernoulli variables B_{p_j} with parameters $p_j = b/(a+b+j)$.

Proof. The well known identity in distribution $\beta_{a,b}\gamma_{a+b} \stackrel{d}{=} \gamma_a$ for independent beta and gamma variables with the indicated parameters, and known representations of $\log$ gamma variables, show that the distribution of the non-negative random variable $-\log \beta_{a,b}$ is infinitely divisible with Lévy density at $x > 0$ which is given by the formula [8, p. 769]

$$\frac{x^{-1}}{1-e^{-x}}(e^{-ax} - e^{-(a+b)x}) = \sum_{j=0}^{\infty} x^{-1}(e^{-(a+j)x} - e^{-(a+b+j)x}). \tag{7.5}$$

But it is also well known and easily checked that the jth term on the right side of (7.5) is the Lévy density of the infinitely divisible law of the jth term in (7.4). So the conclusion follows easily from the additivity of Lévy measures. □

For $z = 0$ the generating function (7.2) gives the limiting probability of what is called in [43] the event of a *complete sample* with no gaps:

$$\lim_{n\to\infty} \mathbb{P}(K_{0:n} = 0) = \lim_{n\to\infty} \mathbb{P}(K_n = M_n) = g_\theta(0) = \frac{\Gamma(1+\theta)^2}{\Gamma(1+2\theta)}. \tag{7.6}$$

If $\theta = m$ say is a positive integer, these formulas simplify by the gamma recursion $\Gamma(1+x) = x\Gamma(x)$. The generating function (7.2) reduces to rational function of z, with m linear factors in the denominator. This implies that $K_{0:\infty}$ is distributed as the sum of just m independent geometrics G_i, with the by now familiar parameters $i/(i+\theta)$ for $1 \le i \le m$. This yields the remarkable chain of identities in law

$$K_{0:\infty} \stackrel{d}{=} \sum_{i=1}^{m} G_i \stackrel{d}{=} M_m - 1 \stackrel{d}{=} X_{n:n} - X_{n-m:n} \text{ for all } n \ge m \text{ if } \theta = m \in \mathbb{N}.$$

where the first $\stackrel{d}{=}$ holds only if $\theta = m \in \mathbb{N}$, but the next two $\stackrel{d}{=}$ hold for all $\theta > 0$ by Theorem 1.1. Also for $\theta = m \in \mathbb{N}$, the right side of (7.6) is the inverse of the central binomial coefficient $\binom{2m}{m}^{-1} \sim 2^{-2m}\sqrt{\pi m}$ as $m \to \infty$, and the approach of this probability to 0 is similarly rapid for $\theta \to \infty$ through real values, due to Stirling's approximation to the gamma function. In particular, for $\theta = 1$, the probability of a complete sample is simply $1/2$. This is also known [43] to be the common value of $\mathbb{P}(K_n = M_n)$ for every n in the case of i.i.d. sampling from geometric$(1/2)$. Some extensions of these results to more general RAMs with i.i.d. factors will be given in [61].

Also either from (7.1) or from the probability generating function g_θ given in (7.2) it is easy to find the generating function for the Lévy measure of $K_{0:\infty}$ which has atoms λ_k in $k = 1, 2, \ldots$:

$$\sum_{k=1}^{\infty} \lambda_k z^k = -\log g_\theta(0) + \log g_\theta(z) = \log \frac{\Gamma(1+2\theta)}{\Gamma(1+\theta)} \frac{\Gamma(1+(1-z)\theta)}{\Gamma(1+(2-z)\theta)}.$$

Its Taylor's expansion gives an expression for the atoms λ_k in terms of the polygamma function.

The right tail probability function of L_∞, the limit in distribution of L_n can be read from its definition (1.15): For $k = 0, 1, 2, \ldots$

$$\mathbb{P}(L_\infty > k) = \mathbb{P}(G_i = 0, 1 \le i \le k) = \frac{(1)_k}{(1+\theta)_k} \sim \frac{\Gamma(1+\theta)}{k^\theta} \text{ as } k \to \infty.$$

The tail probability generating function for L_∞ is the Gaussian hypergeometric function

$$\sum_{k=0}^{\infty} \mathbb{P}(L_\infty > k) z^k = \sum_{k=0}^{\infty} \frac{(1)_k}{(1+\theta)_k} z^k = {}_2F_1\left(\begin{matrix}1, 1\\ 1+\theta\end{matrix}; z\right)$$

from which the limiting mean is found to be

$$\lim_{n\to\infty} \mathbb{E}(L_n) = \mathbb{E}(L_\infty) = \sum_{k=0}^{\infty} \mathbb{P}(L_\infty > k) = \frac{\theta}{(\theta-1)_+}$$

where the last expression should be read as $\theta/(\theta-1) < \infty$ for $\theta > 1$, and $\theta/0 = \infty$ for $\theta \le 1$. Similarly, the limiting second moment is finite only if $\theta > 2$ with the simple limit formula for the second binomial moment

$$\lim_{n\to\infty} \mathbb{E}\binom{L_n}{2} = \sum_{k=0}^{\infty} k\, \mathbb{P}(L_\infty > k) = \frac{\theta}{(\theta-1)_+(\theta-2)_+}\,.$$

It appears that this pattern continues, with finite third binomial moment $2!\theta/(\theta-3)_3$ for $\theta > 3$, and so on.

Since the number of ties at the maximum L_n can be defined on a common probability space, one can also be interested in some stronger types of convergence for these random variables than the convergence in distribution. The answer to this question for independent random variables is well known: Brands et al. [9] conjectured and Baryshnikov et al. [7] confirmed that the number L_n of maxima in a sample of n independent discrete random variables can exhibit just three types of behavior as $n \to \infty$: either it converges to 1 or to ∞ in probability, or it does not have a limit. These three cases can be distinguished in terms of the discrete hazards (1.7) of the distribution of X_1, which are nonrandom in this case, say, $H_j = h_j$. If $h_j \to 0$ as $j \to \infty$, then the number of maxima converges in probability to 1, and this is the only possibility for convergence to a proper distribution. This result was extended to an almost sure convergence by Qi [64], who showed that a.s. convergence holds if and only if the series $\sum_j h_j^2$ converges. Later, a probabilistic proof of this result was given by Eisenberg [19] along with some extensions. His results are also formulated in terms of the discrete hazards:

Lemma 7.2 ([19])**.** *Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with values in $\mathbb{N}$ and infinitely supported distribution. Then, for any $\ell \in \mathbb{N}$, $\mathbb{P}[\limsup_n L_n = \ell] = 1$ if and only if $\sum_{j=1}^{\infty} h_j^{\ell} = \infty$ and $\sum_{j=1}^{\infty} h_j^{\ell+1} < \infty$. If the above series diverge for all $\ell \in \mathbb{N}$ then $\mathbb{P}[\limsup_n L_n = \infty] = 1$.*

This result can be immediately translated to the exchangeable GEM case, because in this case hazards are independent random variables. Heuristically, the next Theorem means that for $\alpha \in (0,1)$, as k becomes large, the GEM(α,θ) probabilities P_k a.s. tend to zero regularly and the maximum is unlikely to be hit twice before a new maximal value is reached, while for $\alpha = 0$ the situation is opposite, there exist k such that P_k is arbitrary large compared to the tail $1 - F_k$ and such values k are repeated as maxima of the sample many times.

Theorem 7.3. *Let $X_1, X_2, \dots$ have the* GEM(α,θ) *exchangeable distribution. Then*

$$\begin{aligned} &\mathbb{P}\big[\limsup_n L_n = 1\big] = 1, && \alpha > 0;\\ &\mathbb{P}\big[\limsup_n L_n = \infty\big] = 1, && \alpha = 0. \end{aligned}$$

Proof. If the distribution of H_i is defined by (1.9) then

$$\mathbb{E}[H_i^k] = \frac{(1-\alpha)_k}{(1+(i-1)\alpha+\theta)_k}\,.$$

Hence for $\alpha > 0$

$$\mathbb{E}[H_i^k] \sim \frac{(1-\alpha)_k}{(i\alpha)^k}, \qquad i \to \infty,$$

and since $H_i \in (0,1)$ by Kolmogorov's three series theorem the series $\sum H_i^2$ converges a.s. So $\mathbb{P}[\limsup_n L_n = 1|(H_i)] = 1$ by Lemma 7.2, and also unconditionally. On the other hand $\mathbb{E}[H_i^k]$ does not depend on i for $\alpha = 0$, so the series H_i^k diverges by the same theorem and again Lemma 7.2 implies $\mathbb{P}[\limsup_n L_n = \infty|(H_i)] = 1$ and hence unconditionally. □

References

[1] Gerold Alsmeyer, Alexander Iksanov, and Alexander Marynych, *Functional limit theorems for the number of occupied boxes in the Bernoulli sieve*, Stochastic Process. Appl. **127** (2017), no. 3, 995–1017. MR-3605718

[2] Charles E. Antoniak, *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems*, Ann. Statist. **2** (1974), 1152–1174. MR-0365969

[3] Richard Arratia, A. D. Barbour, and Simon Tavaré, *Poisson process approximations for the Ewens sampling formula*, Ann. Appl. Probab. **2** (1992), no. 3, 519–535. MR-1177897

[4] Richard Arratia, A. D. Barbour, and Simon Tavaré, *Logarithmic combinatorial structures: a probabilistic approach*, EMS Monographs in Mathematics, European Mathematical Society (EMS), Zürich, 2003. MR-2032426

[5] Richard Arratia, A. D. Barbour, and Simon Tavaré, *A tale of three couplings: Poisson–Dirichlet and GEM approximations for random permutations*, Combin. Probab. Comput. **15** (2006), nos 1–2, 31–62. MR-2195574

[6] Richard Arratia, A. D. Barbour, and Simon Tavaré, *Exploiting the Feller coupling for the Ewens sampling formula [comment on MR3458585]*, Statist. Sci. **31** (2016), no. 1, 27–29. MR-3458588

[7] Yuliy Baryshnikov, Bennett Eisenberg, and Gilbert Stengle, *A necessary and sufficient condition for the existence of the limiting probability of a tie for first place*, Statist. Probab. Lett. **23** (1995), no. 3, 203–209. MR-1340152

[8] Arup Bose, Anirban Dasgupta, and Herman Rubin, *A contemporary review and bibliography of infinitely divisible distributions and processes*, Sankhyā Ser. A **64** (2002), no. 3, part 2, 763–819, Special issue in memory of D. Basu. MR-1981512

[9] Jos J. A. M. Brands, Frederik W. Steutel, and Roeland J. G. Wilms, *On the number of maxima in a discrete sample*, Statist. Probab. Lett. **20** (1994), no. 3, 209–217. MR-1294106

[10] Franz Thomas Bruss and Rudolf Grübel, *On the multiplicity of the maximum in a discrete random sample*, Ann. Appl. Probab. **13** (2003), no. 4, 1252–1263. MR-2023876

[11] Cristina Costantini, Pierpaolo De Blasi, Stewart N. Ethier, Matteo Ruggiero, and Dario Spano, *Wright–Fisher construction of the two-parameter Poisson–Dirichlet diffusion*, arXiv preprint arXiv:1601.06064 (2016).

[12] Harry Crane, *The ubiquitous Ewens sampling formula*, Statist. Sci. **31** (2016), no. 1, 1–19. MR-3458585

[13] Harry Crane, *Rejoinder: The ubiquitous Ewens sampling formula [MR3458586; MR3458587; MR3458588; MR3458589; MR3458590; MR3458585]*, Statist. Sci. **31** (2016), no. 1, 37–39. MR-3458591

[14] Herbert A. David and Haikady N. Nagaraja, *Order statistics*, third ed., Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003. MR-1994955

[15] Peter Donnelly, *Partition structures, Pólya urns, the Ewens sampling formula, and the ages of alleles*, Theoret. Population Biol. **30** (1986), no. 2, 271–288. MR-865115

[16] Peter Donnelly, *The heaps process, libraries, and size-biased permutations*, J. Appl. Probab. **28** (1991), no. 2, 321–335. MR-1104569

[17] Peter Donnelly and Paul Joyce, *Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex*, Stochastic Process. Appl. **31** (1989), no. 1, 89–103. MR-996613

[18] Peter Donnelly and Simon Tavaré, *The ages of alleles and a coalescent*, Adv. in Appl. Probab. **18** (1986), no. 1, 1–19. MR-827330

[19] Bennett Eisenberg, *The number of players tied for the record*, Statist. Probab. Lett. **79** (2009), no. 3, 283–288. MR-2493010

[20] Steinar Engen, *Stochastic abundance models*, Chapman and Hall, London; Halsted Press [John Wiley & Sons], New York, 1978, With emphasis on biological communities and species diversity, Monographs on Applied Probability and Statistics. MR-515721

[21] Steiner Engen, *A note on the geometric series as a species frequency model*, Biometrika **62** (1975), no. 3, 697–699. MR-0411097

[22] Arthur Erdélyi, Wilhelm Magnus, Fritz Oberhettinger, and Francesco G. Tricomi, *Higher transcendental functions. Vols. I, II*, McGraw-Hill Book Company, Inc., New York-Toronto-London, 1953, Based, in part, on notes left by Harry Bateman. MR-0058756

[23] Warren J. Ewens, *The sampling theory of selectively neutral alleles*, Theoret. Population Biology **3** (1972), 87–112; erratum, ibid. **3** (1972), 240; erratum, ibid. **3** (1972), 376. MR-0325177

[24] Warren J. Ewens, *Mathematical population genetics. I*, second ed., Interdisciplinary Applied Mathematics, vol. 27, Springer-Verlag, New York, 2004, Theoretical introduction. MR-2026891

[25] William Feller, *An introduction to probability theory and its applications. Vol. I*, Third edition, John Wiley & Sons, Inc., New York-London-Sydney, 1968. MR-0228020

[26] Shui Feng, *The Poisson–Dirichlet distribution and related topics*, Probability and its Applications (New York), Springer, Heidelberg, 2010, Models and asymptotic behaviors. MR-2663265

[27] Thomas S. Ferguson, *On characterizing distributions by properties of order statistics*, Sankhyā Ser. A **29** (1967), 265–278. MR-0226804

[28] Alexander Gnedin, Ben Hansen, and Jim Pitman, *Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws*, Probab. Surv. **4** (2007), 146–171. MR-2318403

[29] Alexander Gnedin, Alex Iksanov, and Uwe Roesler, *Small parts in the Bernoulli sieve*, Fifth Colloquium on Mathematics and Computer Science, Discrete Math. Theor. Comput. Sci. Proc., AI, Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2008, pp. 235–242. MR-2508790

[30] Alexander Gnedin, Alexander Iksanov, and Alexander Marynych, *The Bernoulli sieve: an overview*, 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10), Discrete Math. Theor. Comput. Sci. Proc., AM, Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2010, pp. 329–341. MR-2735350

[31] Alexander Gnedin and Jim Pitman, *Regenerative composition structures*, Ann. Probab. **33** (2005), no. 2, 445–479. MR-2122798

[32] Alexander Gnedin and Jim Pitman, *Exchangeable Gibbs partitions and Stirling triangles*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **325** (2005), no. Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 12, 83–102, 244–245. MR-2160320

[33] Alexander Gnedin and Jim Pitman, *Self-similar and Markov composition structures*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **326** (2005), no. Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 13, 59–84, 280–281. MR-2183216

[34] Alexander Gnedin and Jim Pitman, *Poisson representation of a Ewens fragmentation process*, Combin. Probab. Comput. **16** (2007), no. 6, 819–827. MR-2351686

[35] Alexander V. Gnedin, *On convergence and extensions of size-biased permutations*, J. Appl. Probab. **35** (1998), no. 3, 642–650. MR-1659532

[36] Alexander V. Gnedin, *The Bernoulli sieve*, Bernoulli **10** (2004), no. 1, 79–96. MR-2044594

[37] Alexander V. Gnedin, Alexander M. Iksanov, Pavlo Negadajlov, and Uwe Rösler, *The Bernoulli sieve revisited*, Ann. Appl. Probab. **19** (2009), no. 4, 1634–1655. MR-2538083

[38] Louis Gordon, *A stochastic approach to the gamma function*, Amer. Math. Monthly **101** (1994), no. 9, 858–865. MR-1300491

[39] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik, *Concrete mathematics*, Addison-Wesley Publishing Company, Advanced Book Program, Reading, MA, 1989, A foundation for computer science. MR-1001562

[40] Robert C. Griffiths, *Unpublished notes*, Monash Univ., Melbourne, Australia, 1980.

[41] Rudolf Grübel and Paweł Hitczenko, *Gaps in discrete random samples*, J. Appl. Probab. **46** (2009), no. 4, 1038–1051. MR-2582705

[42] Paul R. Halmos, *Random alms*, Ann. Math. Statistics **15** (1944), 182–189. MR-0010342

[43] Paweł Hitczenko and Arnold Knopfmacher, *Gap-free compositions and gap-free samples of geometric random variables*, Discrete Math. **294** (2005), no. 3, 225–239. MR-2137565

[44] Tsvetan Ignatov, *A constant arising in the asymptotic theory of symmetric groups, and Poisson–Dirichlet measures*, Teor. Veroyatnost. i Primenen. **27** (1982), no. 1, 129–140. MR-645134

[45] Alexander M. Iksanov, Alexander V. Marynych, and Vladimir A. Vatutin, *Weak convergence of finite-dimensional distributions of the number of empty boxes in the Bernoulli sieve*, Theory Probab. Appl. **59** (2015), no. 1, 87–113. MR-3416065

[46] Alexander Iksanov, *On the number of empty boxes in the Bernoulli sieve II*, Stochastic Process. Appl. **122** (2012), no. 7, 2701–2729. MR-2926172

[47] Alexander Iksanov, *On the number of empty boxes in the Bernoulli sieve I*, Stochastics **85** (2013), no. 6, 946–959. MR-3176494

[48] Lancelot F. James, *Lamperti-type laws*, Ann. Appl. Probab. **20** (2010), no. 4, 1303–1340. MR-2676940

[49] John F. C. Kingman, *Random discrete distributions*, J. Roy. Statist. Soc. Ser. B **37** (1975), 1–22. MR-0368264

[50] John F. C. Kingman, *The representation of partition structures*, J. London Math. Soc. (2) **18** (1978), no. 2, 374–380. MR-509954

[51] Fabrizio Leisen, Antonio Lijoi, and Christian Paroissin, *Limiting behavior of the search cost distribution for the move-to-front rule in the stable case*, Statist. Probab. Lett. **81** (2011), no. 12, 1827–1832. MR-2845896

[52] John William McCloskey, *A model for distribution of individuals by species in an environment*, Annals of Mathematical Statistics **35** (1964), no. 4, 1839–1840, Abstract of Ph. D. Thesis, Michigan State University, 1965.

[53] Valery B. Nevzorov, *Records: Mathematical theory*, Translations of Mathematical Monographs, vol. 194, American Mathematical Society, Providence, RI, 2001, Translated from the Russian manuscript by D. M. Chibisov. MR-1791071

[54] Fritz Oberhettinger, *Tables of Mellin transforms*, Springer-Verlag, New York-Heidelberg, 1974. MR-0352890

[55] Anthony G. Pakes, *The laws of some random series of independent summands*, Advances in the theory and practice of statistics, Wiley Ser. Probab. Statist. Appl. Probab. Statist., Wiley, New York, 1997, pp. 499–516. MR-1481190

[56] Mihael Perman, Jim Pitman, and Marc Yor, *Size-biased sampling of Poisson point processes and excursions*, Probab. Theory Related Fields **92** (1992), no. 1, 21–39. MR-1156448

[57] Leonid A. Petrov, *A two-parameter family of infinite-dimensional diffusions on the Kingman simplex*, Funktsional. Anal. i Prilozhen. **43** (2009), no. 4, 45–66. MR-2596654

[58] Jim Pitman, *Exchangeable and partially exchangeable random partitions*, Probab. Theory Related Fields **102** (1995), no. 2, 145–158. MR-1337249

[59] Jim Pitman, *Random discrete distributions invariant under size-biased permutation*, Adv. in Appl. Probab. **28** (1996), no. 2, 525–539. MR-1387889

[60] Jim Pitman, *Combinatorial stochastic processes*, Lecture Notes in Mathematics, vol. 1875, Springer-Verlag, Berlin, 2006, Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard. MR-2245368

[61] Jim Pitman, *Extremes and gaps in sampling from a residual allocation model*, 2017, In preparation.

[62] Jim Pitman and Yuri Yakubovich, *Ordered and size-biased frequencies in GEM and Gibbs models for species sampling*, arXiv preprint, 2017, arXiv:1704.04732.

[63] Jim Pitman and Marc Yor, *The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator*, Ann. Probab. **25** (1997), no. 2, 855–900. MR-1434129

[64] Yongcheng Qi, *A note on the number of maxima in a discrete sample*, Statist. Probab. Lett. **33** (1997), no. 4, 373–377. MR-1458007

[65] Alfréd Rényi, *On the theory of order statistics*, Acta Math. Acad. Sci. Hungar. **4** (1953), 191–231. MR-0061792

[66] Philippe Robert and Florian Simatos, *Occupancy schemes associated to Yule processes*, Adv. in Appl. Probab. **41** (2009), no. 2, 600–622. MR-2541191

[67] Ian W. Saunders, Simon Tavaré, and G. A. Watterson, *On the genealogy of nested subsamples from a haploid population*, Adv. in Appl. Probab. **16** (1984), no. 3, 471–491. MR-753512

[68] Stanley Sawyer and Daniel Hartl, *A sampling theory for local selection*, Journal of Genetics **64** (1985), no. 1, 21–29.

[69] George P. Yanev and Santanu Chakraborty, *A characterization of exponential distribution and the Sukhatme–Rényi decomposition of exponential maxima*, Statist. Probab. Lett. **110** (2016), 94–102. MR-3474742

Acknowledgments. Thanks to Daniel Dufresne for references to the infinite divisibility of log beta distributions.