ORDERED AND SIZE-BIASED FREQUENCIES IN GEM AND GIBBS' MODELS FOR SPECIES SAMPLING

BY JIM PITMAN AND YURI YAKUBOVICH

University of California, Berkeley and St. Petersburg University

We describe the distribution of frequencies ordered by sample values in a random sample of size n from the two parameter $\mathsf{GEM}(\alpha, \theta)$ random discrete distribution on the positive integers. These frequencies are a (size-α)-biased random permutation of the sample frequencies in either ranked order, or in the order of appearance of values in the sampling process. This generalizes a well-known identity in distribution due to Donnelly and Tavaré [*Adv. in Appl. Probab.* **18** (1986) 1–19] for $\alpha = 0$ to the case $0 \leq \alpha < 1$. This description extends to sampling from $\mathsf{Gibbs}(\alpha)$ frequencies obtained by suitable conditioning of the $\mathsf{GEM}(\alpha, \theta)$ model, and yields a value-ordered version of the Chinese restaurant construction of $\mathsf{GEM}(\alpha, \theta)$ and $\mathsf{Gibbs}(\alpha)$ frequencies in the more usual size-biased order of their appearance. The proofs are based on a general construction of a finite sample $(X_1, \ldots, X_n)$ from any random frequencies in size-biased order from the associated exchangeable random partition Π_∞ of $\mathbb{N}$ which they generate.

1. Introduction. We are interested in various natural orderings of clusters of common values, in sampling from random discrete distributions. The recent review by Crane [9, 10] presents some of the widespread applications of these models of random partitions, in the contexts of population genetics, ecology, Bayesian nonparametrics, combinatorial stochastic processes and inductive inference. Let $P_\bullet := (P_1, P_2, \ldots)$ denote a random discrete distribution on the positive integers, to be thought of as a model for population frequencies of various species in a large population of individuals classified by species, or otherwise partitioned by type in some way. We take these population frequencies to be represented in the *stick-breaking form* [25, 39],

$$(1) \quad P_1 := H_1, \qquad P_2 = (1 - H_1)H_2, \qquad P_3 := (1 - H_1)(1 - H_2)H_3, \qquad \ldots$$

for some sequence of random variables H_i with values in $(0, 1)$ such that $P_i > 0$ for all i and $\sum_{i=1}^\infty P_i = 1$ almost surely. We call this model for population frequencies a *residual allocation model* (RAM) to indicate the H_i are independent, though not necessarily identically distributed. Let $X_1, \ldots, X_n$ denote a *sample of size n* from population frequencies $P_\bullet$, that is, the first n terms of a sequence $X_1, X_2, \ldots$

which conditionally given $P_\bullet$ is i.i.d. according to $P_\bullet$. We are most interested in settings where the possible sample values $1, 2, \ldots$ have a clear meaning in the context of some larger population model, such as the age-ranks of alleles in the infinitely-many-neutral-alleles diffusion model [15]. This model involves $P_\bullet$ with GEM$(0, \theta)$ distribution, while other models of current interest ([8, 33]) involve ranked frequencies derived from the GEM(α, θ) distribution, in which H_i has the beta$(1 - \alpha, \theta + i\alpha)$ density on $(0, 1)$. Here, $0 \leq \alpha < 1$ and $\theta > -\alpha$ are real parameters, and beta(a, b) is the probability distribution on $(0, 1)$ with probability density proportional to $u^{a-1}(1 - u)^{b-1}$ at $0 < u < 1$; see [17, 37] for further background on GEM distributions. As discussed further in the following sections, a special property of GEM(α, θ), important in many contexts, is that the GEM(α, θ) frequencies $P_\bullet$ are in a size-biased random order. It is also known ([18, 35]) that

(2)   the only RAMs with frequencies in size-biased order are the GEM(α, θ) models.

Combined with the fact that the large n asymptotics of GEM(α, θ) samples exhibit a variety of logarithmic and power law behaviors as (α, θ) varies [37], this draws attention to the GEM(α, θ) family as a tractable and versatile family of models for use in applications.

A basic problem is to describe the distribution of the partition of n determined by the *size-ordered* or *ranked sample frequencies*, meaning the list of sizes of clusters of equal values in a sample from a random discrete distribution $P_\bullet$, in decreasing order of size. As recalled in Section 2, that problem has been solved for both GEM(α, θ) and for RAMs with i.i.d. factors, hand-in-hand with a description of the distribution of the *appearance-ordered* sample frequencies, that is, the list of sizes of clusters of equal values in the order in which these values appeared in the random sampling process. It is well known that in a sample from any random discrete distribution $P_\bullet$, the appearance-ordered sample frequencies are a size-biased permutation of the partition of n. A more difficult problem is to provide a corresponding description of the *value-ordered sample frequencies* which can be obtained by ordering the sample in (weakly) increasing order and then reading counts of equal values, so that the number of times the minimal value in the sample is attained comes first, and the frequency of the maximal sample value comes last. See a discussion and an example in Section 2.1 below that should clarify these notions.

For the GEM$(0, \theta)$, with i.i.d. beta$(1, \theta)$ factors H_i, a remarkably simple description of the value-ordered sample frequencies was provided by Donnelly and Tavaré [14]:

> in sampling from GEM$(0, \theta)$, there is no difference in distribution between the value-ordered frequencies and the appearance-ordered frequencies: they are both in a size-biased random order.

For sampling from a RAM with i.i.d. factors, a more complicated description of the distribution of value-ordered frequencies was found in Gnedin and Pitman [20],

Section 11. But there seems to be a gap in the literature regarding the distribution of value-ordered frequencies for a $\mathsf{GEM}(\alpha, \theta)$ sample for $0 < \alpha < 1$. Our main result is that this problem has a surprisingly simple solution, almost as simple as the Donnelly and Tavaré result for $\alpha = 0$. Compared to the case $\alpha = 0$, the only difference is that the usual notion of size-biased permutation of a composition $(n_1, \ldots, n_k)$ of n needs to be replaced by the notion of a *(size-α)-biased random permutation*, defined as follows:

- for each $1 \leq i \leq k$, the first component is set equal to n_i with probability $\frac{n_i - \alpha}{n - k\alpha}$,
- given $k > 1$ and that the ith component was chosen to be the first, for each $j \neq i$ the second component is set equal to n_j with probability $\frac{n_j - \alpha}{n - n_i - (k-1)\alpha}$

and so on, as discussed more carefully in Section 2.2 and also in Appendix A.

THEOREM 1. *For each $n \geq 1$, in a random sample of size n from $\mathsf{GEM}(\alpha, \theta)$, the value-ordered sample frequencies are (size-α)-biased.*

Because sample frequencies in appearance order are size-biased in the usual way, this theorem shows that the Donnelly–Tavaré identity in distribution between value-ordered and appearance-ordered frequencies is very special to $\mathsf{GEM}(0, \theta)$. It does not extend to $\mathsf{GEM}(\alpha, \theta)$ for $0 < \alpha < 1$ without extending the notion of size-biasing to (size-α)-biasing. Hence the theorem dispels the tempting but false idea that value-ordered and appearance-ordered sample frequencies might be identically distributed in any model with value-ordered population frequencies in size-biased order. For except in trivial cases of equality between counts, for $0 < \alpha < 1$ a (size-α)-biased permutation is not the same in distribution as simple size-biased permutation.

Our proof of Theorem 1 in Section 3 shows much more: according to Theorem 5, the conclusion of Theorem 1 holds also for sampling from the size-biased presentation of frequencies of any Gibbs(α) *partition*, that is, for $P_\bullet$ derived as the limits of relative frequencies in order of appearance of any random partition (Π_n) of positive integers with the conditional distribution of Π_n given $K_n = k$ that is shared by all $\mathsf{GEM}(\alpha, \theta)$ partitions [37], Theorem 4.6, [21], described in more detail in Section 2.3 below. This leads us to speculate that there is a converse to Theorem 5: if in sampling from $P_\bullet$ the value-ordered clusters are (size-α)-biased, then $P_\bullet$ is the size-biased presentation of some Gibbs(α) frequencies. But we do not have any proof of this.

The case $\alpha = 0$ of Theorem 1, due to Donnelly and Tavaré [14], was a culmination of earlier work by Watterson and others ([40–42]) on random sampling from models of limit populations in genetics with random frequencies governed by $\mathsf{GEM}(0, \theta)$ when listed in order of *age-rank*, meaning the frequencies of the oldest, second-oldest, third-oldest, ... alleles in the population. The age-ranked

frequencies in these models are in size-biased random order, and the ith sample value X_i in this setting is then the age rank in the large population of the allelic types of the ith individual to be sampled. Thus it reasonable to study samples from more general frequencies in size-biased order thinking of the sample value as of the age-rank in an infinite idealized population. A natural question in this setting, is given that the allelic composition of a sample is $(n_1, \ldots, n_k)$, what is the probability that a particular allele with n_i representatives is the oldest in the sample? According a result of Watterson and Guess [42], Theorem 3, under assumptions that are known [13] to imply $\mathsf{GEM}(0, \theta)$ frequencies by age-rank in the limit population, the allele with n_i representatives is the oldest in the sample with probability n_i/n. Theorem 5 extends this result as follows.

COROLLARY 2. *In sampling from a limit population with frequencies by age-rank which are in a size-biased random order, and distributed according either to* $\mathsf{GEM}(\alpha, \theta)$, *or to the size-biased presentation of frequencies in a* Gibbs(α) *model, the allele with n_i representatives in the sample composition $(n_1, \ldots, n_k)$ is the oldest in the sample with probability $(n_i - \alpha)/(n - k\alpha)$.*

There is a combinatorial construction of the Gibbs(α) sample frequencies in size-biased appearance order known as the Chinese restaurant process (CRP) [37], Section 3.1. Our proof of Theorem 5 and its corollary Theorem 1 involves the Ordered Chinese restaurant process (OCRP), which produces value-ordered sample frequencies, as considered in [27] and [20], Section 11, and discussed further in Section 3. In comparing the prescriptions for appearance-ordered and value-ordered frequencies in sampling from a Gibbs(α) model, while there are obvious similarities between the two schemes, there are also subtle differences. Observe first that if you are given *both* the appearance-ordered and the value-ordered frequencies, by exchangeability the appearance-ordering is just a size-biased random ordering of the value-ordered frequencies. So the value-ordered frequencies serve as a sufficient statistic for predictions about the next sample value X_{n+1}. It is a subtle point of the prediction rule given the value-ordered frequencies, spelled out in Corollary 3 below, that the value-ordered frequencies provide no more information than the appearance-ordered frequencies, so far as predicting whether X_{n+1} will be a new value or not: all that matters is the number of existing clusters k and the sample size n: the probability that the next observation is a new value depends only on n and k, no matter what the appearance-ordered or value-ordered frequencies of the k clusters. This is a very special property of Gibbs(α) models, closely associated with the Markov property of K_n for these models. What is even more interesting, considering that the probability of a new value is unaffected by the value-ordered frequencies, is that the value-ordered frequencies *do* affect the probabilities that the new observation equals one of the clusters of previous observations, as is plain from comparison of the two formulas (19) and (25)

below. The sequential scheme for selecting a value is the same in both cases, except that the scheme given value-ordered frequencies puts weight $n_1 + 1 - \alpha$ on the lowest-valued cluster and weight $n - n_1 - (k - 1)\alpha$ on the rest, whereas the scheme given appearance-ordered frequencies puts lesser weight $n_1 - \alpha$ on the first cluster to appear, and the same weight $n - n_1 - (k - 1)\alpha$ on the rest. So the value-ordered frequency data changes the prediction of the next observation given it is one of those previously observed, always pushing it to be more likely to be the lowest previous value observed n_1 times, no matter what the previously observed frequencies in value-order $n_1, \ldots, n_k$.

COROLLARY 3. *In sampling from the limit frequencies of any* $\mathrm{Gibbs}(\alpha)$ *model in size-biased order, conditionally given the number* K_n *of distinct values in the sample, the event* $X_{n+1} \notin \{X_1, \ldots, X_n\}$ *of a new value at time* $n + 1$ *is independent of the value-ordered frequencies of the sample* $X_1, \ldots, X_n$. *In other words, the conditional probability of this event given value-ordered frequencies* $(n_1, \ldots, n_k)$ *in a sample with* $K_n = k$ *depends only on* n *and* k *and does not depend otherwise on* $(n_1, \ldots, n_k)$.

The rest of this article is organized as follows. In the next section, we introduce the notation and recall some notions we use. In Section 3, we formulate and prove our main result, Theorem 5, and also discuss the OCRP which produces value-ordered sample frequencies step by step. In Section 4, we present an alternative computational proof of Theorem 1 and also derive some consequences from the OCRP description of Corollary 7. This allows us to reproduce well-known results for the $\mathrm{GEM}(0, \theta)$ model with this new approach, thus providing an additional check for it. Finally, in the Appendices we collect some basic facts about a generalization of the (size-α)-biasing procedure, and compare the value-ordering used in this paper with the regenerative ordering of [20].

2. Background and notation.

2.1. *Partitions generated by random samples.* Let Π_n denote the random partition of n generated by the sample values $X_1, \ldots, X_n$, that is, if there are say $K_n = k$ distinct sample values $X_1, \ldots, X_n$, the partition of the set $[n] := \{1, \ldots, n\}$ is

$$(3) \qquad\qquad \Pi_n := \{C_1, \ldots, C_k\}$$

with $C_1 := \{i \leq n : X_i = X_{M(1)}\}$, where $M(1) = 1$ is the least element of both $[n]$ and C_1, and if $k \geq 2$ then $C_2 := \{i \leq n : X_i = X_{M(2)}\}$, where $M(2)$ is the least element of both $[n] \setminus C_1$ and of C_2, and so on. These *clusters* C_i generated by the sample, are listed here in their *order of appearance*. We are interested in various orderings of these clusters. Each ordering of clusters induces a list of their sizes in that order, which is a random sequence of strictly positive integers with sum n,

called a *random composition of n*. The *value* of cluster C_i is the common value of X_j for every $j \in C_i$. We are particularly concerned with:

- The *cluster sizes in appearance-order*: $N_{\bullet:n}^* := (n_1, \ldots, n_k)$ if cluster C_j as above has n_j members for each $1 \leq j \leq k$.
- The *clusters in value-order* define a *random ordered partition of* $[n]$ ([20], Section 11, [27], Section 5.2.1)

$$(4) \qquad \widetilde{\Pi}_n := (\widetilde{C}_1, \ldots, \widetilde{C}_k) := (C_{\pi(1:n)}, C_{\pi(2:n)}, \ldots, C_{\pi(k:n)})$$

for some permutation $\pi(\bullet : n)$ of $[k]$, which encodes the additional value-order structure. Explicitly:

$$(5) \qquad \widetilde{C}_1 := \left\{ i \in [n] : X_i = \min_{j \in [n]} X_j \right\}$$

and if $\widetilde{C}_1 \neq [n]$, then

$$(6) \qquad \widetilde{C}_2 := \left\{ i \in [n] : i \notin \widetilde{C}_1, X_i = \min_{j \in [n] \setminus \widetilde{C}_1} X_j \right\}$$

and so on. Notice that for $\widetilde{\Pi}_n$ with K_n clusters the permutation $\pi(\bullet : n)$ of $[K_n]$ is encoded in the state $\widetilde{\Pi}_n$: the inverse of $\pi(\bullet : n)$ is obtained by rearranging the clusters $(\widetilde{C}_1, \ldots, \widetilde{C}_k)$ in order of their least elements.

- The *cluster sizes in value-order*: $N_{\bullet:n}^{X\uparrow}$ with $N_{i:n}^{X\uparrow} = \#\widetilde{C}_i = n_{\pi(i:n)}$ for $1 \leq i \leq K_n$. This is the sequence of sizes of clusters in increasing order of their common X-values. For instance,

$$N_{1:n}^{X\uparrow} \text{ is the number of } j \text{ such that } X_j = \min_{1 \leq i \leq n} X_i,$$

$$N_{K_n:n}^{X\uparrow} \text{ is the number of } j \text{ such that } X_j = \max_{1 \leq i \leq n} X_i.$$

- The *partition of n generated by the sample*: $N_{\bullet:n}^{\downarrow}$ is the weakly decreasing rearrangement of either $N_{\bullet:n}^*$ or $N_{\bullet:n}^{X\uparrow}$.

We illustrate these definitions by an adaptation of Kingman's paintbox model [29] for generating random partitions. Let (I_j) be the interval partition of $[0, 1]$ defined by

$$(7) \quad I_1 = (0, P_1), \qquad I_2 = (P_1, P_1 + P_2), \qquad I_3 = (P_1 + P_2, P_1 + P_2 + P_3)$$

and so on. Define the sample values by $X_i = j$ if $U_i \in I_j$ where U_i is a sequence of i.i.d. uniform$[0, 1]$ variables. In the following, display a particular realization of the successive partial sums of probabilities $P_1, P_1 + P_2, \ldots$ is marked by a series of vertical bars $|$ in a unit interval $[0, 1]$. Then $n = 6$ sample points U_i picked from $[0, 1]$ landed between the bars as indicated:

$$(8) \quad [0 \underbrace{\qquad}_{1} | \underbrace{\qquad}_{2} | \underbrace{U_6 \qquad U_5}_{3} | \underbrace{\qquad}_{4} | \underbrace{U_3}_{5} | \underbrace{\qquad}_{6} | \underbrace{\qquad}_{7} | \underbrace{U_2 \, U_1 \qquad U_4}_{8} | \underbrace{\qquad}_{9} | | | | | | \ldots 1]$$

Regarding the bars as separators between interval boxes with labels $1, 2, 3, \ldots$ shown under the braces, various quantities under consideration are in this in-

stance:

- the sample from $P_\bullet$ is $(X_1, \ldots, X_6) = (8, 8, 5, 8, 3, 3)$;
- the partition of $[6]$ is $\Pi_6 = \{\{1, 2, 4\}, \{3\}, \{5, 6\}\}$;
- the clusters in appearance-order are $(C_1, C_2, C_3) = (\{1, 2, 4\}, \{3\}, \{5, 6\})$;
- the cluster sizes in appearance-order are $N^*_{\bullet:6} = (3, 1, 2)$;
- the clusters in value-order are $(\widetilde{C}_1, \widetilde{C}_2, \widetilde{C}_3) = (C_3, C_2, C_1) = (\{5, 6\}, \{3\}, \{1, 2, 4\})$;
- the cluster sizes in value-order are $N^{X\uparrow}_{\bullet:6} = (2, 1, 3)$, corresponding to numbers of repeated values in the increasing rearrangement $(3, 3, 5, 8, 8, 8)$ of the sample;
- the partition of 6 defined by the cluster sizes is $N^{\downarrow}_{\bullet:6} = (3, 2, 1)$.

For any partition $\{C_1, \ldots, C_k\}$ of $[n]$, the probability of the basic event (3) that this particular partition is generated by an exchangeable sample $X_1, \ldots, X_n$ depends just on cluster sizes $n_i = \#C_i$, so defines a function $p(n_1, \ldots, n_k)$ of compositions $(n_1, \ldots, n_k)$ of n. Following [37], Section 2.2, this function of compositions of n is called the *exchangeable partition probability function* (EPPF) of Π_n. For each fixed k, the EPPF is a symmetric function of k positive integer arguments. As n varies, the EPPF satisfies an addition rule [37], (2.9), reflecting the consistency property of the random partitions, that Π_m is the restriction to $[m]$ of Π_n for each $m < n$. However, one can also consider the EPPF for a fixed n, as we do in Lemmas 10 and 11 in Appendix A.

Similarly, for an ordered partition $(C_1, \ldots, C_k)$ of $[n]$ in some order, with $n_i = \#C_i$, the probability $\widetilde{p}(n_1, \ldots, n_k)$ of the event that this particular ordered partition is obtained by some ordered partition construction from an exchangeable sample is called an *ordered exchangeable partition probability function* (OEPPF). This function may no longer be symmetric in $(n_1, \ldots, n_k)$. The term exchangeable means only that the probability of achieving the ordered partition $(C_1, \ldots, C_k)$ depends only on sizes $(n_1, \ldots, n_k)$ of clusters of the partition. As n varies, the OEPPF will also satisfy some consistency relations; see [20], equations (2), (3).

It is a well-known consequence of exchangeability of Π_n, that no matter what $P_\bullet$

(9) $N^*_{\bullet:n}$ is a size-biased random permutation of $N^{\downarrow}_{\bullet:n}$, as well as of $N^{X\uparrow}_{\bullet:n}$.

As the sample size $n \to \infty$, it follows easily from the strong law of large numbers and (9) that no matter what the distribution of $P_\bullet$, there is the almost sure convergence of relative cluster sizes

(10) $n^{-1} N^{X\uparrow}_{\bullet:n} \to P_\bullet$ and $n^{-1} N^*_{\bullet:n} \to P^*_\bullet$ almost surely,

where $P^*_\bullet$ is a size-biased random permutation of $P_\bullet$. Consequently,

(11)   if $P_\bullet$ is in a size-biased random order, then $P_\bullet \stackrel{d}{=} P^*_\bullet$.

Such a random discrete distribution $P_\bullet$ is said to be *invariant under size-biased permutation* (ISBP). This condition plays a central role in the theory of partitions

generated by random sampling, for a number of reasons. One reason is that if $P_\bullet$ is ISBP, there is a simple general formula [34], [37], (3.4), for the probability of the basic event (3) for any particular partition $\{C_1, \ldots, C_k\}$ of $[n]$, in terms of multivariate moments of $P_\bullet$. Another reason is that distributions that are ISBP are precisely the distributions of frequencies of species in appearance-order in any exchangeable species sampling model with proper frequencies [37]. Especially in contexts where there is no a priori natural ordering of frequencies by positive integers, for instance in the setting of population genetics where different alleles might be identified only by some biochemical tag, or in the theory of interval partitions generated by the zeros of stochastic processes [37], one may as well use the size-biased ordering whenever that is tractable, because of its close connection to partition probabilities.

2.2. *Pseudo-size biased random order.* It was mentioned in the Introduction that the value-ordering of clusters of Π_n involves the procedure of (size-α)-biased random permutation of component sizes. We treat the (size-α)-biased permutation an instance of a more general notion of an s-biased permutation [32], where s is some strictly positive function of a cluster size called *pseudo-size*. For a partition $\Pi_n = \{C_1, \ldots, C_k\}$ of $[n]$, with $\#C_i = n_i$ for $i \in [k]$, and such a function s, an *s-biased pick* is a randomly chosen cluster of Π_n, which given Π_n equals cluster C_i with probability $s(n_i)/(s(n_1) + \cdots + s(n_k))$. An *$s$-biased permutation* is a random permutation of clusters in order of an exhaustive process of sampling without replacement by a sequence of s-biased picks. The usual *size-biased permutation* of Π_n is just its s-biased permutation for the choice of the pseudo-size $s(m) := m$ to be just the ordinary size.

With the pseudo-size function s, we associate the probability

$$
(12) \qquad \widetilde{s}(n_1, \ldots, n_k) := \prod_{i=1}^{k} \frac{s(n_i)}{s(n_i) + \cdots + s(n_k)}
$$

that an ordered collection of clusters $(C_1, \ldots, C_k)$ of sizes $(n_1, \ldots, n_k)$ stays exactly in the same order after s-biased permutation. Notice that this is *not* the probability that after s-biased permutation of clusters their sizes will be $(n_1, \ldots, n_k)$. That probability is $\widetilde{s}(n_1, \ldots, n_k)$ multiplied by an appropriate combinatorial coefficient. We shall need the following result which we prove in Appendix A.

LEMMA 4. *Let Π_n be an exchangeable random partition of $[n]$ with* EPPF *p. Consider a strictly positive size function $s(m)$ of positive integers $m \leq n$, and associate with it a function $\widetilde{s}$ of compositions of n as in* (12).

(i) *If $N_{\bullet:n}$ is a listing of sizes of clusters of an s-biased permutation of Π_n, then the probability function of $N_{\bullet:n}$ on compositions of n is*

$$
(13) \qquad \mathbb{P}[N_{\bullet:n} = (n_1, \ldots, n_k)] = \binom{n}{n_1, \ldots, n_k} \widetilde{s}(n_1, \ldots, n_k) p(n_1, \ldots, n_k).
$$

(ii) *Conversely, if the probability function of a random composition $N_{\bullet:n}$ is given by* (13) *for all compositions of n, for some symmetric function p of compositions, then p is an* EPPF, *and $N_{\bullet:n}$ has the same distribution as an s-biased random ordering of component sizes of Π_n with* EPPF *p.*

2.3. *Gibbs' partitions.* We are particularly interested in the EPPFs which can be represented in the Gibbs' form

$$(14) \qquad p(n_1, \ldots, n_k) = V_{k:n} \prod_{i=1}^{k} w(n_i), \qquad \text{where } n = \sum_{i=1}^{k} n_i,$$

for some positive weights $V_{k:n}$, $1 \le k \le n$ and $w(1), w(2), \ldots$. It is known [37], Theorem 4.6, [21] that the only EPPFs of this form which are produced by sampling from some frequencies P_j that are strictly positive for all j are those obtained by taking

$$(15) \qquad w(n) = (1 - \alpha)_{n-1} \qquad \text{for some } 0 \le \alpha < 1,$$

with $(x)_n := \Gamma(x + n)/\Gamma(x)$ the rising factorial. For such weights $w(\cdot)$, it is easy to see [21] that $V_{k:n}$ must satisfy the consistency relation

$$(16) \qquad V_{k:n} = (n - k\alpha)V_{k:n+1} + V_{k+1:n+1} \qquad (1 \le k \le n < \infty).$$

Following [21], Definition 3, we call an exchangeable partition of positive integers with EPPF of form (14) with w weights given by (15) and V weights subject to (16) a *Gibbs (α) partition*. For a given $\alpha \in [0, 1)$, the collection of all arrays of nonnegative weights $V_{k:n}$ satisfying (16) is a convex set [37], Theorem 4.6, [21]. For each α, there is a one-parameter family of extreme weights. These are indexed by $\theta \ge 0$ for $\alpha = 0$, and $0 < \alpha < 1$ by another parameter $\ell \ge 0$, called the *α-diversity* of the associated random partition in [37]. In both cases, by general convex analysis, every consistent family of weights $V_{k:n}$ admits a unique integral representation over this one-parameter family of extreme weight arrays.

For each $\alpha \in [0, 1)$ and $\theta > -\alpha$, there is the distinguished family of weights

$$(17) \qquad V_{k:n}(\alpha, \theta) := \frac{1}{(1 + \theta)_{n-1}} \prod_{i=1}^{k-1} (\theta + i\alpha).$$

It is easily checked that (16) holds for these special weights $V_{k:n} = V_{k:n}(\alpha, \theta)$, so (17) provides an instance of Gibbs(α) exchangeable partition. It is known [28] that such $V_{k:n}$ are the only weights of form V_k/c_n for some positive sequences $V_\bullet$ and $c_\bullet$, and that the weights (17) produce the EPPF of a random partition of positive integers whose frequencies in order of appearance have the GEM(α, θ) distribution described in the Introduction [37].

Gibbs partitions were introduced in [36] and further developed in [21], and received much attention in recent probabilistic and statistical literature. A wide range

of Gibbs EPPF in terms of special functions can be found in [26]. Not trying to provide a full review of the literature, we mention papers [3–7, 11, 16, 30] which deal with various statistical applications of Gibbs-type priors. Interpretations of Gibbs' partitions in terms of records were developed in [24]. Some features of the construction we use in Section 4 are interpreted in terms of Bayesian inference for species sampling in [31] and [2].

2.4. *The Chinese restaurant process.* The *Chinese restaurant process* (CRP) [37], Section 3.1, provides an intuitive expression of various notions in sampling from random discrete distributions, with its metaphorical customers arriving to be seated at tables in the restaurant, with interpretations in various contexts, of

- *customers* corresponding to individuals/tokens/elements;
- *tables* corresponding to values/alleles/species/types/cycles/clusters/blocks/ intervals.

In the basic CRP as described in [37], Section 3.1, an exchangeable random partition of positive integers is constructed sequentially. Starting from a first customer assigned to table 1, after n customers have been assigned to some k tables labeled by $1, 2, \ldots, k$ in appearance-order, with say n_i customers seated at table i, for $1 \le i \le k$, there is a probabilistic rule for assigning customer $n + 1$ either to one of the k tables already occupied or to a new table. In the ecological context of species sampling, the customers are individuals and assigning a new customer to one of previously occupied tables corresponds to observing an individual of some previously seen species, while introducing a new table corresponds to sampling an individual of some species previously unseen. In this basic CRP, the relative frequencies of customers occupying the tables in order of appearance converge back to $P_\bullet^*$, the size-biased permutation of population frequencies, as in (10).

In the context of Gibbs(α) partitions with EPPF (14), given appearance-ordered frequencies $(n_1, \ldots, n_k)$ in a sample of size n, the appearance-ordered frequencies in a sample of size $n + 1$ are obtained by:

- either adding the frequency $n_{k+1} = 1$ (discovering a new species) with probability

$$(18) \qquad p_{k:n} := \frac{p(n_1, \ldots, n_k, 1)}{p(n_1, \ldots, n_k)} = \frac{V_{k+1:n+1}}{V_{k:n}} \qquad (1 \le k \le n);$$

- or, for some $i \in [k]$, incrementing n_i by 1 (new observation is the ith species in order of appearance) with probability

$$\frac{p(n_1, \ldots, n_i + 1, \ldots, n_k)}{p(n_1, \ldots, n_k)} = (1 - p_{k:n})\frac{n_i - \alpha}{n - k\alpha}$$

$$(19)$$

$$= (1 - p_{k:n})h_\alpha(n_i, \ldots, n_k) \prod_{j=1}^{i-1}[1 - h_\alpha(n_j, \ldots, n_k)],$$

where for a composition $(n_1, \ldots, n_k)$ of n

$$(20) \qquad h_\alpha(n_1, \ldots, n_k) := \frac{n_1 - \alpha}{n_1 - \alpha + \cdots + n_k - \alpha} = \frac{n_1 - \alpha}{n - k\alpha}$$

is the probability of choosing the first cluster in a (size-α)-biased pick from k distinct clusters of sizes $n_1, \ldots, n_k$.

The consistency relation (16) ensures that these probabilities sum to 1. The second, product form in (19) emphasizes the idea that a single cluster can be chosen from the existing clusters of sizes $(n_1, \ldots, n_k)$, in a (size-α)-biased fashion, by a succession of (size-α)-biased choices, the first to decide if it is the first cluster or not, if not whether it is the second cluster, and so on. In particular, for $\mathsf{GEM}(\alpha, \theta)$ frequencies defined by (17) one has

$$(21) \qquad p_{k:n} = \frac{\theta + k\alpha}{\theta + n} \quad \text{and} \quad \frac{p(n_1, \ldots, n_i + 1, \ldots, n_k)}{p(n_1, \ldots, n_k)} = \frac{n_i - \alpha}{\theta + n}.$$

Applying this procedure step by step leads to appearance-ordered tables, with relative frequencies converging to size-biased permutation of the Gibbs(α) probabilities, as in (10), which have the same distribution as the original frequencies just in the ISBP $\mathsf{GEM}(\alpha, \theta)$ case.

3. Main results. Our central result is the following more refined version of Theorem 1.

THEOREM 5. *Fix $0 \le \alpha < 1$. Suppose that $X_1, \ldots, X_n$ is a random sample from $P_\bullet$ which is the size-biased presentation of limit frequencies of a* Gibbs(α) *partition of positive integers. Let p be the* EPPF, *as in* (14)–(15), *corresponding to* (17) *for $P_\bullet$ with* $\mathsf{GEM}(\alpha, \theta)$ *distribution. Then the composition probability function of $N_{\bullet:n}^{X\uparrow}$, the sequence of sizes of clusters of X-values in increasing order of those values, is given by the formula*

$$(22) \qquad \mathbb{P}\big[N_{\bullet:n}^{X\uparrow} = (n_1, \ldots, n_k)\big] = \binom{n}{n_1, \ldots, n_k}\left(\prod_{i=1}^{k} \frac{n_i - \alpha}{n_i - \alpha + \cdots + n_k - \alpha}\right) \\ \times p(n_1, \ldots, n_k)$$

for each composition $(n_1, \ldots, n_k)$ of n. Moreover,

$$(23) \qquad N_{\bullet:n}^{X\uparrow} \text{ is a (size-}\alpha\text{)-biased permutation of } N_{\bullet:n}^*,$$

where the composition probability function of $N_{\bullet:n}^*$ is given by the right-hand side of (22) *with the (size-α)-biasing product replaced by the ordinary size-biasing product with every α replaced by* 0, *as in the known formula* (48).

Our proof of this result makes use of the following key lemma.

LEMMA 6 ([38], Proposition 3.1). *Consider an exchangeable random partition* $\Pi_\infty := (\Pi_n)$ *of the set* $\mathbb{N}$ *of positive integers, with proper frequencies, meaning that all clusters of* Π_∞ *are infinite almost surely. Fix* $n \geq 1$. *Let* $\mathcal{C}_1$ *be the cluster of* Π_∞ *containing* $n + 1$, *and for* $k \geq 1$, *given that* $\bigcup_{i=1}^k \mathcal{C}_i \neq \mathbb{N}$, *let* $\mathcal{C}_{k+1}$ *be the cluster of* Π_∞ *containing the least* $m > n + 1$ *with* $m \notin \bigcup_{i=1}^k \mathcal{C}_i$. *Define* $P_\bullet$ *by setting* P_k *equal to the almost sure limiting relative frequency of* $\mathcal{C}_k$, *and let* $X_i := j$ *iff* $i \in \mathcal{C}_j$. *Then:*

- *$P_\bullet$ is a size-biased ordering of frequencies of Π_∞;*
- *$(X_1, \ldots, X_n)$ is a sample of size n from $P_\bullet$;*
- *the partition of $[n]$ generated by $(X_1, \ldots, X_n)$ is Π_n.*

This key lemma has a "now you see it, now you don't" quality, depending on what metaphor is used for the intuitive description of Π_∞. Here is one way to see that the construction works. Regard the exchangeable partition Π_∞ as being generated by Kingman's paintbox construction, using random frequencies to create an interval partition of $[0, 1]$ into component intervals whose lengths in size-biased random order have the required distribution of $P_\bullet$. Let U be a sequence of i.i.d. uniform $[0, 1]$ random variables independent of the interval partition, and let Π_∞ be the random partition of $\mathbb{N}$ whose clusters are the equivalence classes for the random equivalence relation $i \sim j$ iff either $i = j$ or U_i and U_j fall in the same component of the interval partition. Use the order in which $U_{n+1}, U_{n+2}, \ldots$ discover the component intervals to label them by $1, 2, \ldots$, and define $P_\bullet$ by this labeling of interval lengths. Finally, let X_i for $i \in [n]$ be the numerical label of the interval component containing U_i. Then the conclusions of the lemma should be intuitively clear.

Our proof of Theorem 5 will be expressed in terms of another metaphor for exchangeable random partitions, the Chinese restaurant construction of Π_∞. Call the first n customers to enter the restaurant the *primary customers* and customers $n + 1, n + 2, \ldots$ the *secondary* customers. Then $\mathcal{C}_1, \mathcal{C}_2, \ldots$ is the list of clusters of Π_∞ *in order of their discovery by secondary customers*, and $P_\bullet$ is the listing of frequencies of these clusters of Π_∞ in that order of discovery by secondary customers. Compared to the usual listing of clusters in order of appearance, this is just a relabeling of tables. Each table in the restaurant is assigned a new label, with label 1 for the table at which customer $n + 1$ is seated, label 2 for the next table discovered by one of the secondary customers, and so on. By some almost surely finite random time, the first K_n tables at which the primary customers were seated will all have been discovered by secondary customers. At that random time, the values $X_1, \ldots, X_n$ are assigned to the primary customers, with $X_i = j$ if customer i is seated at the jth table in order of discovery by the secondary customers. With this metaphor, the fact that $X_1, \ldots, X_n$ generates Π_n, the partition of $[n]$ defined by the original seating plan in the Chinese restaurant, is completely obvious. That $X_1, \ldots, X_n$ is a sample of size n from $P_\bullet$ is less obvious, but nonetheless true.

PROOF OF THEOREM 5. It is enough to show (23) for any particular representation of a sample $X_1, \ldots, X_n$ from $P_\bullet$. For this purpose, we take $X_1, \ldots, X_n$ to be constructed as in Lemma 6, and use the Chinese restaurant metaphor. According to (19), in a Gibbs(α) CRP at any stage $m > n$ in the process of rediscovery of the initial tables by secondary customers, given that $N^*_{\bullet:n} = (n_1, \ldots, n_k)$ say, and given that up to stage m some nonempty subset of tables $S \subseteq [k]$ remains undiscovered, and given also that individual $m+1$ sits at one of these tables, that table is table $i \in S$ with probability $(n_i - \alpha)/\sum_{s \in S}(n_s - \alpha)$. That is just a (size-$\alpha$)-biased assignment of customer $m+1$ to one of the remaining tables. The conditional distribution of $N^{X\uparrow}_{\bullet:n}$ given $N^*_{\bullet:n} = (n_1, \ldots, n_k)$ is therefore that of a (size-α)-biased random permutation of $(n_1, \ldots, n_k)$. This proves (23), and the sampling formula (22) is read from Lemma 4. $\square$

The simplicity of these descriptions of value-ordered frequencies in sampling from GEM(α, θ), and Gibbs(α) models in general, suggests there should be some embellishment of the CRP generating appearance-ordered frequencies as described in Section 2.4, in which both the appearance-order and value-order of the sample are generated sequentially, in an entirely combinatorial way, that is distributionally equivalent to the model of sampling from an infinite list of frequencies. Such additional structure of sampling in an environment with totally ordered clusters, treated in [13, 20, 22], and developed here in Corollary 7, is well accommodated by an *ordered Chinese restaurant process* (OCRP). This is the usual CRP, with a sequentially developing total order of tables, as proposed in [27]. Here, the order of tables is taken to be the value-order, although any other order of tables can be treated in a similar fashion. It is assumed inductively that after n customers have arrived they are seated at some k tables which are placed from left to right by order of values in the sample, and a new customer is seated either to some previously occupied table or to a new table which is placed in one of $k+1$ possible places relative to the old tables. Technically, the *state* of the restaurant after n customers have been seated represents an *ordered partition of the set of n customers labeled by* $[n]$. Customer $n+1$ arrives with a value X_{n+1} and occupies the table where previous customers with the same value were seated, if any, or a new table if this value appears for the first time, and this new table is placed between tables with lower and higher values than X_{n+1}, or at the appropriate end of the line of tables if the value X_{n+1} is extreme compared to the values of previous customers. Implicit then in the state is the ordering of tables by appearance which can be restored by sorting the tables in order of the least customer number. If just the value-ordered frequencies of the occupied tables are given instead of the ordered partition of the set $[n]$, then this information is lost. But due to exchangeability, the conditional distribution of the appearance order given value-ordered frequencies is a size-biased permutation of these frequencies.

It turns out that the above procedure specialized to Gibbs(α) partitions with value-ordered frequencies can be described in a way quite similar to the basic CRP explained in Section 2.4. We summarize it in the following corollary of Theorem 5.

COROLLARY 7.   *In sampling from the limit frequencies of any* Gibbs(α) *model in size-biased order, with associated discovery probabilities $p_{k:n}$ as in* (18), *the sequential development of value-ordered sampling frequencies is as follows. Given frequencies $(n_1, \ldots, n_k)$ in value-order in a sample of size n, the value-ordered frequencies in the sample of size $n + 1$ are obtained by*:

• *either, for some $j \in [k + 1]$, putting a 1 into $(n_1, \ldots, n_k)$ at the jth of $k + 1$ possible places* (*new value not present previously and of rank j in the updated value order*), *to create frequencies $(1, n_1, \ldots, n_k)$, $(n_1, 1, \ldots, n_k)$, $\ldots$, $(n_1, \ldots, n_k, 1)$ as the case may be, with probabilities*

$$(24) \qquad p_{k:n} h_\alpha(1, n_j, \ldots, n_k) \prod_{i=1}^{j-1} \left[1 - h_\alpha(1, n_i, \ldots, n_k)\right];$$

• *or, for some $j \in [k]$, incrementing n_j by 1* (*new value of rank j both in the previous and in the updated value ordering*) *with probabilities*

$$(25) \qquad (1 - p_{k:n}) h_\alpha(n_j + 1, n_{j+1}, \ldots, n_k) \prod_{i=1}^{j-1} \left[1 - h_\alpha(n_i + 1, n_{i+1}, \ldots, n_k)\right]$$

for $h_\alpha(n_1, \ldots, n_m) := (n_1 - \alpha)/(n_1 - \alpha + \cdots + n_m - \alpha)$ as in (20).

This corollary is much simpler than similar descriptions of the development of value-ordered sampling frequencies for a RAM with i.i.d. factors provided by Gnedin and Pitman [20] and James [27], even in the case of sampling from GEM$(0, \theta)$, when it can be checked that Corollary 7 is consistent with results in these sources. What is remarkable and unexpected about these results is that it seems extremely difficult to provide any comparably simple descriptions of the value-ordered frequencies in sampling from a more general RAM with independent but not identically distributed factors. Our arguments make essential use of both the assumed size-biased order of the Gibbs(α) frequencies, and the sequential description of Gibbs(α) sampling frequencies in appearance order, discussed above.

PROOF OF COROLLARY 7.   Suppose that after $n \geq 1$ steps of the OCRP the value-ordered frequencies are $(n_1, \ldots, n_k)$, with $n = \sum_{i=1}^k n_i$. Given that event, according to (22) and Lemmas 11 and 4 the probability that a new customer occupies some new table which is placed in jth of $k + 1$ possible positions is

$$\frac{\widetilde{p}(n_1, \ldots, n_{j-1}, 1, n_j, \ldots, n_k)}{\widetilde{p}(n_1, \ldots, n_k)} = \frac{\widetilde{s}(n_1, \ldots, n_{j-1}, 1, n_j, \ldots, n_k)}{\widetilde{s}(n_1, \ldots, n_k)} \frac{V_{k+1:n+1}}{V_{k:n}}$$

for $\widetilde{s}$ associated with the pseudo-size function $s(n) = n - \alpha$ as in (12), where the second fraction on the right is the ratio of EPPFs (14), because products of w weights (15) cancel. By (18), this second fraction is exactly $p_{k:n}$ for the Gibbs(α)

partitions. Hence it remains to notice that the ratio of $\widetilde{s}$ functions can be written down in the stick-breaking form (24).

Similarly, given the frequencies $(n_1, \ldots, n_k)$ in value-order after n steps, a new customer is seated at the existing table j with probability

$$\frac{\widehat{p}(n_1, \ldots, n_j + 1, \ldots, n_k)}{\widehat{p}(n_1, \ldots, n_k)} = \frac{\widetilde{s}(n_1, \ldots, n_j + 1, \ldots, n_k)}{\widetilde{s}(n_1, \ldots, n_k)} \frac{V_{k:n+1} w(n_j + 1)}{V_{k:n} w(n_j)}.$$

From (15), (16), (18) and (12), it follows that

$$\frac{V_{k:n+1} w(n_j + 1)}{V_{k:n} w(n_j)} = \frac{n_j - \alpha}{n - k\alpha} (1 - p_{k:n}),$$

$$\frac{\widetilde{s}(n_1, \ldots, n_j + 1, \ldots, n_k)}{\widetilde{s}(n_1, \ldots, n_k)} = \frac{n_j + 1 - \alpha}{n_j - \alpha} \prod_{i=1}^{j} \frac{n_i - \alpha + \cdots + n_k - \alpha}{n_i - \alpha + \cdots + n_k - \alpha + 1}$$

and the stick-breaking representation (25) is just a rearrangement of the product of the right-hand sides above. $\square$

Two comments on the above argument include:

• Once the Bayes ratios have been calculated as indicated, the conditional independence asserted in Corollary 3 is obvious by inspection of the formulas. But this conditional independence does not seem at all obvious otherwise. Especially because the Bayes calculations show that the value-ordered frequencies do affect the probabilities of adding to old clusters, it does not seem at all clear why they might not also affect the probability of discovering a new species, in some way more complex than just through the total number of clusters. Even for sampling from $\mathsf{GEM}(0, \theta)$ this does not seem obvious.

• The elementary algebra of cancellation in the Bayes' calculations used to prove Corollary 7 can be easily used to show that the OCRP defined by that Corollary 7 gives an ordered exchangeable partition of positive integers, without any assumption that it is derived by the value-orders in successive sampling. In view of the general representation theorem for such an exchangeable OCRP due to Gnedin [22], it follows that this OCRP must in fact be derived from value-order generated by some exchangeable sequence $X_1, X_2, \ldots$, which is a sample from some random discrete distribution F on the line, whose size-biased atoms have the $\mathsf{Gibbs}(\alpha)$ distribution determined by the discovery probabilities, because the distribution of partitions of n is built into the dynamics of the restaurant. Even for $\mathsf{GEM}(\alpha, \theta)$ it seems far from obvious from this approach why the atoms of F are simply those of $\mathsf{Gibbs}(\alpha)$ in their usual order, which is at the heart of what Theorems 1 and 5 are saying. But this approach might be used in combination with some other means of identifying F to provide an alternate proof of Theorem 1.

4. Related calculations.

4.1. *Inductive proof of Theorem* 1. For a general RAM, a decomposition over the minimal value m of the sample, which is repeated n_1 times if the value-ordered

counts are $(n_1, \ldots, n_k)$, leads to the following recursive formula:[1] for $k \geq 2$,

$$
(26) \quad \mathbb{P}\big[N_{\bullet:n}^{X\uparrow} = (n_1, \ldots, n_k)\big] = \binom{n}{n_1} \sum_{m=1}^{\infty} \mathbb{E}\left[H_m^{n_1}(1 - H_m)^{n-n_1} \prod_{i=1}^{m-1} (1 - H_i)^n \right]
$$
$$
\times \mathbb{P}^{(m)}\big[N_{\bullet:n}^{X\uparrow} = (n_2, \ldots, n_k)\big],
$$

where $\mathbb{P}^{(m)}$ refers to the probability in a generally different RAM, that is one generated by $(H_{m+1}, H_{m+2}, \ldots)$ instead of $(H_1, H_2, \ldots)$ in (1). So only for RAMs with i.i.d. factors H_i, this is indeed a recursion, but for $\mathsf{GEM}(\alpha, \theta)$ (26) connects the probabilities of value-ordered counts for different parameters:

$$
(27) \quad \mathbb{P}_{\alpha,\theta}\big[N_{\bullet:n}^{X\uparrow} = (n_1, \ldots, n_k)\big] = \binom{n}{n_1} \sum_{m=1}^{\infty} \mathbb{E}\left[H_m^{n_1}(1 - H_m)^{n-n_1} \prod_{i=1}^{m-1} (1 - H_i)^n \right]
$$
$$
\times \mathbb{P}_{\alpha,\theta+m\alpha}\big[N_{\bullet:n}^{X\uparrow} = (n_2, \ldots, n_k)\big].
$$

This leads to a direct proof of Theorem 1 by induction on the number k of distinct values in the sample, which is outlined below.

We need to show that, with $n = n_1 + \cdots + n_k$,

$$
(28) \quad \begin{aligned}
&\mathbb{P}_{\alpha,\theta}\big[N_{\bullet:n}^{X\uparrow} = (n_1, \ldots, n_k)\big] \\
&\quad = \frac{n!(\frac{\theta}{\alpha} + 1)_{k-1}\alpha^{k-1}}{(\theta + 1)_{n-1}} \prod_{\ell=1}^{k} \frac{(1 - \alpha)_{n_\ell}}{n_\ell!(n_\ell + \cdots + n_k - (k - \ell + 1)\alpha)}
\end{aligned}
$$

which is (22) for the $\mathsf{GEM}(\alpha, \theta)$ EPPF p given by (14) with w weights (15) and V weights (17). Comparing it to the well-known formula [37], (3.6),

$$
(29) \quad \mathbb{P}_{\alpha,\theta}\big[N_{\bullet:n}^{*} = (n_1, \ldots, n_k)\big] = \frac{n!(\frac{\theta}{\alpha} + 1)_{k-1}\alpha^{k-1}}{(\theta + 1)_{n-1}} \prod_{\ell=1}^{k} \frac{(1 - \alpha)_{n_\ell-1}}{n_\ell!}
$$

shows that $N_{\bullet:n}^{X\uparrow}$ is a (size-α)-biased permutation of $N_{\bullet:n}^{*}$.

In order to evaluate (27), note that

$$
(30) \quad \mathbb{E}_{\alpha,\theta} H_i^r (1 - H_i)^s = \frac{B(1 - \alpha + r, \theta + i\alpha + s)}{B(1 - \alpha, \theta + i\alpha)} = \frac{(1 - \alpha)_r (\theta + i\alpha)_s}{(\theta + (i-1)\alpha + 1)_{r+s}}.
$$

For samples with just one value repeated n times, it is well known and easy to calculate using (30) that

$$
\begin{aligned}
\mathbb{P}_{\alpha,\theta}\big[N_{\bullet:n}^{X\uparrow} = (n)\big] &= \sum_{m=1}^{\infty} \mathbb{E}_{\alpha,\theta} P_m^n \\
&= \sum_{m=1}^{\infty} \mathbb{E}_{\alpha,\theta} H_m^n \prod_{i=1}^{m-1} \mathbb{E}_{\alpha,\theta}(1 - H_i)^n
\end{aligned}
$$

---

[1] We thank the anonymous referee for this observation.

$$= \sum_{m=1}^{\infty} \frac{(1-\alpha)_n}{(\theta + (m-1)\alpha + 1)_n} \prod_{i=1}^{m-1} \frac{(\theta + i\alpha)_n}{(\theta + (i-1)\alpha + 1)_n}$$

$$= \frac{(1-\alpha)_n}{(\theta+1)_n} \sum_{m=1}^{\infty} \prod_{i=1}^{m-1} \frac{\theta + i\alpha}{\theta + i\alpha + n}.$$

The sum in the last line is the evaluation of the hypergeometric function $_2F_1(1, \frac{\theta+\alpha}{\alpha}; \frac{\theta+\alpha+n}{\alpha}; 1)$, and since for $b > a + 1$ one has [43], Section 14.11,

$$\tag{31} \sum_{j=0}^{\infty} \frac{(a)_j}{(b)_j} = {}_2F_1(1, a; b; 1) = \frac{b-1}{b-a-1}$$

it finally gives

$$\tag{32} \mathbb{P}_{\alpha,\theta}\big[N_{\bullet:n}^{X\uparrow} = (n)\big] = \frac{(1-\alpha)_{n-1}}{(\theta+1)_{n-1}}$$

in accordance with (28). This is the induction base.

Now suppose that (28) is true for some k and check that it is also true with $k+1$ instead of k. Let $n = n_1 + \cdots + n_{k+1}$, then by (27), (30) and the induction assumption (28)

$$\mathbb{P}_{\alpha,\theta}\big[N_{\bullet:n}^{X\uparrow} = (n_1, \ldots, n_{k+1})\big]$$

$$= \binom{n}{n_1} \sum_{m=1}^{\infty} \mathbb{E}\Bigg[H_m^{n_1} (1 - H_m)^{n-n_1} \prod_{i=1}^{m-1} (1 - H_i)^n \Bigg]$$

$$\times \mathbb{P}_{\alpha,\theta+m\alpha}\big[N_{\bullet:n-n_1}^{X\uparrow} = (n_2, \ldots, n_k)\big]$$

$$= \frac{n!}{n_1!} \sum_{m=1}^{\infty} \frac{(1-\alpha)_{n_1}(\theta+m\alpha)_{n-n_1}}{(\theta + (m-1)\alpha + 1)_n} \prod_{i=1}^{m-1} \frac{(\theta + i\alpha)_n}{(\theta + (i-1)\alpha + 1)_n}$$

$$\times \frac{(\frac{\theta+m\alpha}{\alpha} + 1)_{k-1} \alpha^{k-1}}{(\theta + m\alpha + 1)_{n-n_1-1}} \prod_{\ell=2}^{k+1} \frac{(1-\alpha)_{n_\ell}}{n_\ell!(n_\ell + \cdots + n_{k+1} - (k - \ell + 2)\alpha)}$$

$$= \frac{n!(n - (k+1)\alpha)\alpha^k}{(\theta+1)_n} \prod_{\ell=1}^{k+1} \frac{(1-\alpha)_{n_\ell}}{n_\ell!(n_\ell + \cdots + n_{k+1} - (k - \ell + 2)\alpha)}$$

$$\times \sum_{m=1}^{\infty} \frac{(\frac{\theta}{\alpha} + 1)_{k+m-1}}{(\frac{\theta+n}{\alpha} + 1)_{m-1}}.$$

Writing $(\frac{\theta}{\alpha} + 1)_{k+m-1} = (\frac{\theta}{\alpha} + 1)_k (\frac{\theta}{\alpha} + k + 1)_{m-1}$ and using (31) allows to calculate

$$\sum_{m=1}^{\infty} \frac{(\frac{\theta}{\alpha} + 1)_{k+m-1}}{(\frac{\theta+n}{\alpha} + 1)_{m-1}} = \left(\frac{\theta}{\alpha} + 1\right)_k \frac{\theta + n}{n - (k+1)\alpha}$$

which gives (28) with k replaced by $k + 1$, as desired.

4.2. *Some checks on Corollary* 7.   As the result of Corollary 7 is a new and not obvious property of Gibbs(α) partitions, even in the heavily studied case $\alpha = 0$ of a GEM($0, \theta$) partition, this section offers some checks on the result by different approaches. We are able to do this for $\alpha = 0$, but providing any significant checks on the result for $0 < \alpha < 1$ remains a challenging problem.

We start with some general identity in distribution which is a consequence of Lemma 6. Let $(X_1, X_2, \dots)$ be a sample from random discrete distribution $P_\bullet$, and consider the sequence of indicators Δ_n and L_n where $\Delta_n := K_n - K_{n-1}$ is the indicator of discovery of a new value at step n, and L_n is the indicator of placement at the extreme left, that is, the new value is less than all previous values. Obviously, $0 \le L_n \le \Delta_n$.

Let the new values be discovered in the random times

$$(33) \qquad \{1 = M_1 < M_2 < \cdots\} := \{n \ge 1 : \Delta_n = 1\}.$$

Let X be the index in this sequence of the time value 1 first appears in the sample, that is,

$$(34) \qquad M_X = \min\{m : X_m = 1\}.$$

COROLLARY 8.   *Suppose that frequencies $P_\bullet$ are in size-biased random order. Then X has the same distribution as the first sample X_1.*

PROOF.   We can think of sampling from any realization of frequencies $P_\bullet$. Consider Kingman's paintbox construction of Section 2.1 and suppose that U_1 hits some interval $\mathcal{I}$. Take $P_\bullet$ as in Lemma 6 with $n = 1$, then they are in size-biased random order, and X_1 as defined in the lemma is the number of clusters in Π_∞ restricted to $\{2, \dots, L\}$, where L is the random time when the sequence $U_2, U_3, \dots$ hits $\mathcal{I}$. On the other hand, suppose that $M_1, M_2, \dots$ and X are produced from the sample from $P_\bullet^* \stackrel{d}{=} P_\bullet$, where P_1^* is the size-biased pick defined as the length of $\mathcal{I}$. Then the order of other frequencies of $P_\bullet^*$ is irrelevant to the definition of X, and $X = X_1$ almost surely, hence the result.   □

For the Gibbs(α) partition, Corollary 7 gives us a very different way to compute the law of X.

COROLLARY 9.   *For the Gibbs(α) partition generated by a sample $X_1, X_2, \dots$ from Gibbs(α) frequencies $P_\bullet$ in size-biased order, let*

$$p_\alpha(n, k) := \frac{1 - \alpha}{n - k\alpha} = \mathbb{P}(L_n = 1 \mid M_k = n) = \mathbb{P}\Big(X_n = \min_{1 \le i \le n} X_i \mid M_k = n\Big)$$

which is the common conditional probability in every such Gibbs(α) model that a new minimal value is discovered at time n, given that the kth new value is discovered at time n. Then for each $k = 1, 2, \dots$

$$(35) \qquad \mathbb{E}(P_k) = \mathbb{P}(X = k) = \mathbb{E}\left[p_\alpha(M_k, k) \prod_{j=k+1}^{\infty} \big(1 - p_\alpha(M_j, j)\big) \right].$$

PROOF. The value of the common conditional probability declared above is read from Corollary 7, with n instead of $n+1$ and $k-1$ instead of k, as is the fact that given the entire sequence $M_1, M_2, \ldots$ the events of new minima at these times are conditionally independent with probabilities $p_\alpha(M_j, j)$ for $j \geq 1$. The event $(X = k)$ is the event that the kth new value to occur in the sample is minimal value of the whole sample, so all new values discovered later are not minimal. The second equality in (35) follows by conditioning on this sequence. The first equality in (35) is Corollary 8. □

It is hard to imagine how this formula (35) could be checked in any other way for a general Gibbs(α) partition, though Griffiths and Spanò [24] offer a deep study of the sequence $(M_1, M_2, \ldots)$ derived from a Gibbs(α) partition which might provide an alternate approach. For GEM(α, θ), there is at least a simple product formula for $\mathbb{E}_{\alpha,\theta} P_k$. But the expected product seems very difficult to check, because there is no independence to work with. For GEM($0, \theta$), the product is quite manageable, however. Then it is well known [1] that the indicators Δ_n are independent, with $\mathbb{P}_{0,\theta}(\Delta_n = 1) = \theta/(\theta + n - 1)$ for $n \geq 1$. It follows easily that for $k \geq 1$ the $\mathbb{P}_{0,\theta}$ distribution of M_k is given by the formula

$$(36) \qquad \mathbb{P}_{0,\theta}(M_k = n) = C_{n-1,k-1}\theta^k/(\theta)_n, \qquad n \geq 1,$$

where $C_{n,k} = (-1)^{n+k} S_{n,k}$ is the unsigned Stirling number of the first kind giving the number of permutations of $[n]$ with k cycles. We observe that the evaluation, with $(x)_r := \Gamma(x + r)/\Gamma(x)$,

$$(37) \qquad \mathbb{E}_{0,\theta}\left[\frac{1}{(M_k + \theta)_r}\right] = \frac{\theta^{k-1}}{(\theta + r)^k(\theta + 1)_{r-1}}, \qquad r > -\theta,$$

is an elementary consequence of the fact that these probabilities in (36) sum to 1 for each $\theta > 0$. This neat formula for inverse Pochhammer moments of M_k does not seem to be well known. We only noticed it after needing the case $r = 1$ to complete the check indicated below.

For $\alpha = 0$, the probability $p_0(n, k) = 1/n$ does not depend on k, and the identity (35) reduces easily to

$$(38) \qquad \frac{\theta^{k-1}}{(1+\theta)^k} = \mathbb{E}_{0,\theta}\left[\frac{1}{M_k} \prod_{m=M_k+1}^{\infty}\left(1 - \frac{\Delta_m}{m}\right)\right].$$

Using independence of the Δ_n, we can compute

$$(39) \qquad \mathbb{E}_{0,\theta}\left[\prod_{m=M_k+1}^{\infty}\left(1 - \frac{\Delta_m}{m}\right) \Big| M_k = n\right] = \mathbb{E}_{0,\theta}\left[\prod_{m=n+1}^{\infty}\left(1 - \frac{\Delta_m}{m}\right)\right]$$

$$= \prod_{m=n+1}^{\infty}\left(1 - \frac{\theta}{(\theta + m - 1)m}\right)$$

$$= \frac{n}{n + \theta}$$

by the factorization

$$1 - \frac{\theta}{(\theta + m - 1)m} = \frac{(m-1)}{m} \frac{(m+\theta)}{m-1+\theta}$$

which telescopes the product. Plugging this into (38) reduces it to (37) for $r = 1$.

In the same vein, conditioning on P_1 gives access to $M_X = \min\{n \geq 1 : X_n = 1\} = \max\{n : L_n = 1\}$:

$$\mathbb{P}_{\alpha,\theta}(M_X > n) = \mathbb{E}_{\alpha,\theta}(1 - P_1)^n = \frac{(\alpha + \theta)_n}{(1+\theta)_n}$$

which reduces to $\theta/(\theta + n)$ for $\alpha = 0$. In that case, differencing gives

$$(40) \qquad \mathbb{P}_{0,\theta}(M_X = n) = \frac{\theta}{(\theta + n)(\theta + n - 1)}.$$

On the other hand, from the description with L_n, the usual $(0, \theta)$ restaurant model, and the fact mentioned above that if the nth customer goes to a new table, this table is placed to the extreme left with probability $1/n$,

$$\mathbb{P}_{0,\theta}(M_X = n) = \mathbb{P}_{0,\theta}(\Delta_n = 1, L_n = 1, L_m = 0 \text{ for all } m > n)$$

$$(41) \qquad = \frac{\theta}{(n-1+\theta)} \frac{1}{n} \prod_{\ell=n+1}^{\infty} \left(1 - \frac{\theta}{(\ell - 1 + \theta)} \frac{1}{\ell}\right),$$

and this is again a telescoping product which reduces to (40). It is not obvious how to perform the same check for general α, because the Δ_n are no longer independent.

APPENDIX A: PSEUDO-SIZE-BIASED ORDERINGS

We need to extend a well-known notion of a size-biased permutation of a finite or countably infinite index set I, or of a collection of clusters or components of some kind $C_i, i \in I$ that is indexed by I, for some notion of sizes $s(C_i)$ of the clusters involved [12, 35]. Typically, $s(C_i)$ is some kind of measure of C_i. But other *pseudo-size* functions s may also be considered, subject to the requirements that $s(C_i) > 0$ for all i and that $\Sigma := \sum_i s(C_i) < \infty$, which need only be met almost surely. Given some collection of random components $(C_i, i \in I)$, and a pseudo-size function s, an *s-biased pick* from these components is $C_{\pi(1)}$, where $\pi(1) \in I$ is a random index with

$$(42) \qquad \mathbb{P}\big(\pi(1) = h \mid C_i, i \in I\big) = s(C_h)/\Sigma \qquad (h \in I).$$

An *s-biased random permutation* of $(C_i, i \in I)$ is an exhaustive random indexing of components $C_{\pi(j)}$ defined by a sequence of s-biased picks without replacement from these components, indexed by $j \in [k]$ if there are a finite number k of components, or by $j \in \mathbb{N}$ if there are an infinite number of them. So, conditionally given $(C_i, i \in I)$:

- $C_{\pi(1)}$ is an s-biased pick from $(C_i, i \in I)$;
- given also $\pi(1)$ and there is more than one component, $\pi(2)$ is an s-biased pick from $(C_i, i \in I \setminus \{\pi(1)\})$;
- given also $\pi(2)$ and there are more than two components, $\pi(3)$ is an s-biased pick from $(C_i, i \in I \setminus \{\pi(1), \pi(2)\})$, and so on.

By a *size-biased permutation*, one usually means the s-biased permutation as defined above, for the specific choice of the size function $s(C_i)$ equal to the number of elements for a finite set C_i, or some measure such as length for infinite sets C_i like intervals.

Intuitively, think in terms of a bag of balls C_i with pseudo-sizes $s(C_i)$ reflecting the ease with which they are drawn relative to other balls. Then a s-biased random permutation of the C_i is a listing of the balls in the order they are drawn in an exhaustive process of sampling without replacement.

We need this notion just for components which are blocks of a random set partition. If the pseudo-size function depends just on the size of a component, then the pseudo-size-biased permutation of an exchangeable random set partition will be an exchangeable ordered set partition. The following lemma presents some elementary facts about this construction for a general pseudo-size function depending just on the real size.

LEMMA 10. *Let s be a strictly positive function of positive integers $m \leq n$, for some fixed n. As in* (12), *define an associated function of compositions of n by*

$$\widetilde{s}(n_1, \ldots, n_k) := \prod_{i=1}^{k} \frac{s(n_i)}{s(n_i) + \cdots + s(n_k)}. \tag{43}$$

Suppose that $(n_1, \ldots, n_k)$ is the list of ordinary sizes of components $(C_1, \ldots, C_k)$ of a fixed ordered partition of $[n]$. Let $(C_{\pi(1)}, \ldots, C_{\pi(k)})$ be an s-biased random permutation of $(C_1, \ldots, C_k)$, for C_i assigned pseudo-size $s(\#C_i) = s(n_i)$. Then:

(i) $(n_{\pi(1)}, \ldots, n_{\pi(k)})$ *is an s-biased random permutation of $(n_1, \ldots, n_k)$.*

(ii) *For π the random permutation of $[k]$ so defined,* (43) *gives the probability that π is the identity, meaning the components are selected in their original order.*

(iii) *For each σ in the set $\mathcal{S}_k$ of all permutations of $[k]$,*

$$\mathbb{P}(\pi = \sigma) = \widetilde{s}(n_{\sigma(1)}, \ldots, n_{\sigma(k)}). \tag{44}$$

(iv) *For every composition $(n_1, \ldots, n_k)$ and every pseudo-size function s, there is the identity*

$$\sum_{\sigma \in \mathcal{S}_k} \widetilde{s}(n_{\sigma(1)}, \ldots, n_{\sigma(k)}) = 1. \tag{45}$$

PROOF.   Part (i) follows easily from the definition of an s-biased permutation, as does (44) and its special case stated in (ii), by multiplication of successive conditional probabilities. Finally, (iv) follows from (iii) by the law of total probability.
□

The operation of s-biasing is one easy way to turn an exchangeable random partition of $[n]$ into an ordered exchangeable random partition of $[n]$. After s-biasing we are dealing with an ordered exchangeable random partition of $[n]$. The notions of the EPPF of an exchangeable random partition and the OEPPF of an ordered exchangeable random partition were introduced in Section 2.1. The following lemma records a fundamental relation between the EPPF of Π_n and the OEPPF of its s-biased random permutation $\widetilde{\Pi}_n$.

LEMMA 11.   *Let $\widetilde{\Pi}_n$ be the ordered exchangeable random partition of $[n]$ obtained from an exchangeable random partition Π_n of $[n]$ by putting its components in an s-biased random order. Then the OEPPF $\widetilde{p}$ of $\widetilde{\Pi}_n$ and the EPPF p of Π_n are related by*

$$(46) \qquad \widetilde{p}(n_1, \ldots, n_k) = \widetilde{s}(n_1, \ldots, n_k)\, p(n_1, \ldots, n_k),$$

where $\widetilde{s}$ is defined by (43).

PROOF.   For any *particular* ordered partition $(C_1, \ldots, C_k)$ of $[n]$ with components of sizes $(n_1, \ldots, n_k)$, the s-biased permutation of components of Π_n equals $(C_1, \ldots, C_k)$ iff Π_n equals $\{C_1, \ldots, C_k\}$, which happens with probability $p(n_1, \ldots, n_k)$, and given that event the s-biased permutation puts these components in exactly the desired order, which according to (44) happens with probability $\widetilde{s}(n_1, \ldots, n_k)$.   □

PROOF OF LEMMA 4.   It does not change the distribution of $N_{\bullet:n}$ to assume that the s-biased random ordering is made at the level of clusters say $\{C_1, \ldots, C_k\}$ of Π_n. Formula (13) can then be understood as follows. According to the previous lemma, the right-hand side without the multinomial coefficient gives the probability that the s-biased permutation of clusters of Π_n results in any *particular* ordered partition $(C_1, \ldots, C_k)$ with clusters of these sizes. But the number of such ordered partitions of $[n]$ with the given cluster sizes is the multinomial coefficient, and the cases are equiprobable, so the conclusion follows.

As for the converse, for a general random composition $N_{\bullet:n}$ with probability function $q(n_1, \ldots, n_k) = \mathbb{P}[N_{\bullet:n} = (n_1, \ldots, n_k)]$, it is known [20], (4), by arguments much as above, that the EPPF say $\widehat{p}(n_1, \ldots, n_k)$ of the exchangeable random partition of $[n]$ with the same distribution of ranked component sizes as those of $N_{\bullet:n}$ is determined by a summation of $q(n_{\sigma(1)}, \ldots, n_{\sigma(k)})$ over all permutations σ of $[k]$, weighted by the inverse of the multinomial coefficient appearing in (13).

In the present context, assuming that $q(n_1, \ldots, n_k)$ is given by the right-hand side of (13), the multinomial coefficient cancels its inverse, and the general formula for $\widehat{p}$ becomes

$$(47) \qquad \widehat{p}(n_1, \ldots, n_k) = \sum_{\sigma \in \mathcal{S}_k} \widetilde{s}(n_{\sigma(1)}, \ldots, n_{\sigma(k)}) p(n_{\sigma(1)}, \ldots, n_{\sigma(k)}).$$

If p is symmetric, then $p(n_{\sigma(1)}, \ldots, n_{\sigma(k)}) \equiv p(n_1, \ldots, n_k)$ can be factored out of the sum, and the remaining sum is 1 by (45). So $\widehat{p} = p$. $\quad\square$

Some instances of Lemma 4 are as follows:

• The case $s(m) := m$ is ordinary size-biasing. Then the coefficient of $p(n_1, \ldots, n_k)$ on the right-hand side of (13) becomes

$$(48) \qquad \begin{aligned} &\binom{n}{n_1, \ldots, n_k}\left(\prod_{i=1}^k \frac{n_i}{n_i + \cdots + n_k}\right) \\ &= \frac{n!}{(n_1 - 1)! \cdots (n_k - 1)!}\left(\prod_{i=1}^k \frac{1}{n_i + \cdots + n_k}\right). \end{aligned}$$

This instance of formula (13) was given in [34] and [37], (2.6), for the ordinary size-biasing involved when $N_{\bullet:n} := N_{\bullet:n}^*$ is the sequence of cluster sizes of Π_n in order of appearance. In this case, the coefficient of $p(n_1, \ldots, n_k)$ displayed in (48) is a positive integer, the number of partitions of $[n]$ with k clusters of sizes $n_1, \ldots, n_k$ in order of appearance, as indicated by Donnelly and Tavaré in connection with their case $\alpha = 0$ of Theorem 1.

• If $s(m) \equiv 1$, then $\widetilde{s}(n_1, \ldots, n_k) = 1/k!$. Then (13) gives the probability function of the component sizes of Π_n presented in a random order which given $K_n = k$ is uniform on all permutations of $[k]$. This formula appears in [37], (2.7). It is of particular interest for sampling from $\mathsf{GEM}(\alpha, \alpha)$, when it gives the distribution of the composition of n derived by uniform sampling from the interval partition generated by excursions away from 0 of a standard Brownian bridge for $\alpha = 1/2$, and by a standard Bessel bridge of dimension $2 - 2\alpha$ for $0 < \alpha < 1$. See [37], Section 4.5.

• The pseudo-size function $s(m) := m - \alpha$ is involved in Theorems 1 and 5.

APPENDIX B: THE REGENERATIVE ORDERING OF A GEM(α, θ) SAMPLE

This appendix compares and contrasts:

• the value-ordered cluster sizes $N_{\bullet:n}^{X\uparrow}$ in a sample $X_1, \ldots, X_n$ from a GEM(α, θ) distribution on $\{1, 2, \ldots\}$,

which is the primary concern of this article, and

• the value-ordered cluster sizes $N_{\bullet:n}^{Y\uparrow}$ in a sample $Y_1, \ldots, Y_n$ from a particular random discrete distribution $F_{\alpha,\theta}$ on $(0, \infty)$ constructed in [20], with a regenerative property, whose atoms in size-biased order have GEM(α, θ) distribution.

See also [23] for a nice review of these and related concepts. Recall from [20] that a sequence of random compositions $(N_{\bullet:n}, n = 1, 2, \ldots)$ is called *regenerative* if deletion of the first component of $N_{\bullet:n}$ of some given size $n_1 < n$ produces a copy of $N_{\bullet:n-n_1}$:

(49)
$$\begin{aligned}(N_{2:n}, N_{3:n}, \ldots \mid N_{1:n} &= n_1) \\ &\stackrel{d}{=} (N_{1:n-n_1}, N_{2:n-n_1}, \ldots) \qquad \text{for each } 1 \le n_1 < n.\end{aligned}$$

It was shown in [19] that if in sampling from a random discrete distribution on the line, the value-ordered sample frequencies from various sample sizes n are regenerative in this sense, then the distribution of these value-ordered sample frequencies is uniquely determined by that of the size-ordered frequencies (Kingman's partition structure), or equally by that of the appearance-ordered frequencies, which are in distribution just a size-biased rearrangement of the size-ordered sample frequencies. So a random discrete distribution $P_\bullet$, or its associated partition structure, is called *regenerative* iff there exists such a regenerative rearrangement of its frequencies. The study of such regenerative composition structures was motivated by the appearance of these structures in the interval partitions generated by the excursions of a Brownian motion or other Markov process whose zero set is the range of a stable subordinator of index $\alpha \in (0, 1)$.

For any random interval partition of $[0, 1]$, defined by some sequence of interval components (I_j), say $I_j = (G_j, D_j)$ with lengths $P_j := D_j - G_j$, there is a canonical construction of a random discrete distribution F on $[0, 1]$ which puts mass P_j at the right end of I_j. The sample $Y_1, Y_2, \ldots$ from F is then constructed from an i.i.d. uniform $[0, 1]$ random sample $U_1, U_2, \ldots$ by setting $Y_i = D_j$ if $U_i \in (G_j, D_j)$. The value-ordered clusters in the sample $Y_1, \ldots, Y_n$ then reflect the order structure of the intervals (I_j) to the extent it is revealed by the intervals discovered by the uniform sampling points U_i.

First, we emphasize the similarity between these two models of value-ordered cluster sizes $N_{\bullet:n}^{X\uparrow}$ and $N_{\bullet:n}^{Y\uparrow}$ considered above. The cluster sizes in order of appearance in the two sampling schemes are identically distributed, as GEM(α, θ). If $\alpha = 0$, the $F_{0,\theta}$ mentioned above simply puts probability P_j at $1 - \prod_{i=1}^{j}(1 - H_i)$ where the H_i are the i.i.d. beta$(1, \theta)$ factors driving the stick-breaking construction (1) of the GEM$(0, \theta)$ frequencies. The order structure of these possible values is identical to that of their positive integer labels $j = 1, 2, \ldots$. So the value-ordered cluster sizes $N_{\bullet:n}^{X\uparrow}$ and $N_{\bullet:n}^{Y\uparrow}$ are identically distributed.

And now the big difference is the following. For $0 < \alpha < 1$, the random discrete distribution $F_{\alpha,\theta}$ involved in the regenerative ordering of GEM(α, θ) frequencies

cannot have its atoms listed in increasing order like this. Consequently, the value-ordered sample frequencies $N_{\bullet:n}^{X\uparrow}$ and $N_{\bullet:n}^{Y\uparrow}$ cannot be identically distributed for all n. There is some flexibility in the definition of $F_{\alpha,\theta}$, corresponding to change of variables from Y_n to $g(Y_n)$ by a continuous and strictly increasing function g. But that has no effect on the distribution of value-ordered sample frequencies $N_{\bullet:n}^{Y\uparrow}$. According to the results of [20] for $0 < \alpha < 1$, in any representation of the regenerative composition structure associated with $\mathsf{GEM}(\alpha, \theta)$ by value-ordered samples from a random discrete distribution $F_{\alpha,\theta}$ on the line, the atoms of $F_{\alpha,\theta}$ must with probability one accumulate at the left end of the support of $F_{\alpha,\theta}$, corresponding the fact that as the compositions of n grow, with probability one new singleton clusters are added infinitely often at the extreme left end of the sample values. For large n, the initial components of $N_{\bullet:n}^{Y\uparrow}$ are all small, with convergence in distribution to $(1, 1, \ldots)$ as $n \to \infty$, which is not very interesting. On the other hand, the initial component $N_{1:n}^{X\uparrow}$ has limiting relative frequency $n^{-1} N_{1:n}^{X\uparrow} \to P_1 > 0$ almost surely.

While the limiting behavior of these differently ordered sampling compositions derived from $\mathsf{GEM}(\alpha, \theta)$ is very different for $0 < \alpha < 1$, the stochastic mechanism by which they can be described turns out to be very similar. This involves just a slight extension of the notion of pseudo-size-biased random ordering as in Lemma 4.

The definition of an s-biased random permutation proposed in Section 2.2 and treated further in Appendix A admits an obvious generalization in which a strictly positive function $s(m)$ of a single integer variable m with $1 \le m \le n$ is replaced by strictly positive function $s(n', n'')$ of positive integer variables $1 \le n'' \le n' \le n$ which for each fixed n' gives the pseudo-size to be assigned to each component of size $n'' \le n'$ in making a pseudo-size-biased pick from clusters of sizes $n_1'', \ldots, n_j''$ with $\sum_{i=1}^{j} n_i'' = n'$. Then we can formulate the following straightforward extension of Lemmas 10 and 4.

LEMMA 12. *Let $s = s(n', n'')$ be some arbitrary strictly positive pseudo-size function of positive integers $1 \le n'' \le n' \le n$ for some fixed positive integer n. Extend the definition* (12) *in Lemma* 10 *to*

$$\widetilde{s}(n_1, \ldots, n_k)$$

(50)
$$:= \prod_{i=1}^{k} \frac{s(v_i, n_i)}{s(v_i, n_i) + \cdots + s(v_i, n_k)} \qquad \textit{where } v_i := n_i + \cdots + n_k.$$

Then all four parts of Lemma 10 *remain valid, as does the sampling formula of Lemma* 4 *for the probability function of an s-biased random ordering of the cluster sizes of an exchangeable random partition Π_n with* EPPF *p.*

According to [20], Theorem 8.1, for $0 \le \alpha < 1$ and $\theta \ge 0$, in sampling from the regenerative arrangement of $\mathsf{GEM}(\alpha, \theta)$ frequencies, the probability function

of the value-ordered frequencies is given by a simple product formula, which in association with the simple product formula for the (α, θ) EPPF, which can be read from the formulas (14)–(15)–(17), is an expression of the fact [20], Corollary 8.2, that these value-ordered frequencies are in an s-biased random order for the pseudo-size function

$$(51) \qquad s_{\alpha,\theta}(n', n'') = \alpha(n' - n'') + \theta n''$$

which satisfies the strict positivity requirement only for $0 \leq \alpha < 1$ and $\theta \geq 0$. For $\alpha > 0$, this is quite a strange notion pseudo-size: a linear combination of the usual size n'' of a component, and the size $n' - n''$ of its complement in a universe of size n'. For $\alpha = 0$, $s_{0,\theta}(n', n'') = \theta n''$, the constant factor θ has no effect, the pseudo-size-biasing reduces to ordinary size-biasing, and we recover the case $\alpha = 0$ of Theorem 1 due to Donnelly and Tavaré. These results of [20] can now be seen in a broader context of descriptions of random compositions of n derived from each other, or from random partitions of n, by various schemes of pseudo-size-biased sampling. This operation of pseudo-size-biased sampling is a particularly tractable case of the more general notion of a *deletion kernel* for recursive sampling of parts of a partition of n, as treated further in [19]. The present approach of working with ordered partitions of the set $[n]$, as in the proof of Lemma 4, and in some passages of [20], seems to be technically easier than the formalism of unordered partitions of n adopted in [19].

REFERENCES

[1] ARRATIA, R., BARBOUR, A. D. and TAVARÉ, S. (2003). *Logarithmic Combinatorial Structures*: *A Probabilistic Approach*. European Mathematical Society (EMS), Zürich. MR2032426

[2] BACALLADO, S., FAVARO, S. and TRIPPA, L. (2015). Looking-backward probabilities for Gibbs-type exchangeable random partitions. *Bernoulli* **21** 1–37. MR3322311

[3] CERQUETTI, A. (2008). On a Gibbs characterization of normalized generalized Gamma processes. *Statist. Probab. Lett.* **78** 3123–3128. MR2479467

[4] CERQUETTI, A. (2009). A generalized sequential construction of exchangeable Gibbs partitions with application. In *S. Co.* 2009. *Sixth Conference. Complex Data Modeling and Computationally Intensive Statistical Methods for Estimation and Prediction* 115–120. Maggioli Editore, Santarcangelo di Romagna.

[5] CERQUETTI, A. (2013). Marginals of multivariate Gibbs distributions with applications in Bayesian species sampling. *Electron. J. Stat.* **7** 697–716. MR3035269

[6] CERQUETTI, A. (2013). Some contributions to the theory of conditional Gibbs partitions. In *Complex Models and Computational Methods in Statistics*. 77–89. Physica-Verlag/Springer, Milan. MR3051207

[7] CESARI, O., FAVARO, S. and NIPOTI, B. (2014). Posterior analysis of rare variants in Gibbs-type species sampling models. *J. Multivariate Anal.* **131** 79–98. MR3252637

[8] COSTANTINI, C., DE BLASI, P., ETHIER, S. N., RUGGIERO, M. and SPANÒ, D. (2017). Wright–Fisher construction of the two-parameter Poisson–Dirichlet diffusion. *Ann*. *Appl*. *Probab*. **27** 1923–1950. MR3678488

[9] CRANE, H. (2016). The ubiquitous Ewens sampling formula. *Statist*. *Sci*. **31** 1–19. MR3458585

[10] CRANE, H. (2016). Rejoinder: The ubiquitous Ewens sampling formula. *Statist*. *Sci*. **31** 37–39. MR3458591

[11] DE BLASI, P., FAVARO, S., LIJOI, A., MENA, R. H., PRÜNSTER, I. and RUGGIERO, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **37** 212–229.

[12] DONNELLY, P. (1991). The heaps process, libraries, and size-biased permutations. *J*. *Appl*. *Probab*. **28** 321–335. MR1104569

[13] DONNELLY, P. and JOYCE, P. (1991). Consistent ordered sampling distributions: Characterization and convergence. *Adv*. *in Appl*. *Probab*. **23** 229–258. MR1104078

[14] DONNELLY, P. and TAVARÉ, S. (1986). The ages of alleles and a coalescent. *Adv*. *in Appl*. *Probab*. **18** 1–19. MR827330

[15] ETHIER, S. N. (1990). The distribution of the frequencies of age-ordered alleles in a diffusion model. *Adv*. *in Appl*. *Probab*. **22** 519–532. MR1066961

[16] FAVARO, S. and JAMES, L. F. (2015). A note on nonparametric inference for species variety with Gibbs-type priors. *Electron*. *J*. *Stat*. **9** 2884–2902. MR3439188

[17] FENG, S. (2010). *The Poisson–Dirichlet Distribution and Related Topics*: *Models and Asymptotic Behaviors*. Springer, Heidelberg. MR2663265

[18] GNEDIN, A., HAULK, C. and PITMAN, J. (2010). Characterizations of exchangeable partitions and random discrete distributions by deletion properties. In *Probability and Mathematical Genetics*. *London Mathematical Society Lecture Note Series* **378** 264–298. Cambridge Univ. Press, Cambridge. MR2744243

[19] GNEDIN, A. and PITMAN, J. (2004). Regenerative partition structures. *Electron*. *J*. *Combin*. **11** Research Paper 12. MR2120107

[20] GNEDIN, A. and PITMAN, J. (2005). Regenerative composition structures. *Ann*. *Probab*. **33** 445–479. MR2122798

[21] GNEDIN, A. and PITMAN, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap*. *Nauchn*. *Sem*. *S*.*-Peterburg*. *Otdel*. *Mat*. *Inst*. *Steklov*. (*POMI*) **325** 83–102, 244–245. MR2160320

[22] GNEDIN, A. V. (1997). The representation of composition structures. *Ann*. *Probab*. **25** 1437–1450. MR1457625

[23] GNEDIN, A. V. (2010). Regeneration in random combinatorial structures. *Probab*. *Surv*. **7** 105–156. MR2684164

[24] GRIFFITHS, R. C. and SPANÒ, D. (2007). Record indices and age-ordered frequencies in exchangeable Gibbs partitions. *Electron*. *J*. *Probab*. **12** 1101–1130. MR2336601

[25] HALMOS, P. R. (1944). Random alms. *Ann*. *Math*. *Stat*. **15** 182–189. MR0010342

[26] HO, M.-W., JAMES, L. F. and LAU, J. W. (2007). Gibbs partitions (EPPF's) derived from a stable subordinator are Fox *H*- and Meijer *G*-transforms. Preprint. Available at arXiv:0708.0619v2.

[27] JAMES, L. F. (2006). Poisson calculus for spatial neutral to the right processes. *Ann*. *Statist*. **34** 416–440. MR2275248

[28] KEROV, S. (2005). Coherent random allocations, and the Ewens–Pitman formula. *Zap*. *Nauchn*. *Sem*. *S*.*-Peterburg*. *Otdel*. *Mat*. *Inst*. *Steklov*. (*POMI*) **325** 127–145, 246. PDMI Preprint, Steklov Math. Institute, St. Petersburg, 1995. MR2160323

[29] KINGMAN, J. F. C. (1978). The representation of partition structures. *J*. *Lond*. *Math*. *Soc*. (2) **18** 374–380. MR0509954

[30] LIJOI, A., PRÜNSTER, I. and WALKER, S. G. (2008). Investigating nonparametric priors with Gibbs structure. *Statist*. *Sinica* **18** 1653–1668. MR2469329

[31] LIJOI, A., PRÜNSTER, I. and WALKER, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18** 1519–1547. MR2434179

[32] PERMAN, M., PITMAN, J. and YOR, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92** 21–39. MR1156448

[33] PETROV, L. A. (2009). A two-parameter family of infinite-dimensional diffusions on the Kingman simplex. *Funktsional. Anal. i Prilozhen.* **43** 45–66. MR2596654

[34] PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102** 145–158. MR1337249

[35] PITMAN, J. (1996). Random discrete distributions invariant under size-biased permutation. *Adv. in Appl. Probab.* **28** 525–539. MR1387889

[36] PITMAN, J. (2003). Poisson–Kingman partitions. In *Statistics and Science*: *A Festschrift for Terry Speed. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **40** 1–34. IMS, Beachwood, OH. MR2004330

[37] PITMAN, J. (2006). *Combinatorial Stochastic Processes. Lecture Notes in Math.* **1875**. Springer, Berlin. MR2245368

[38] PITMAN, J. and YAKUBOVICH, Y. (2017). Extremes and gaps in sampling from a GEM random discrete distribution. *Electron. J. Probab.* **22** Paper No. 44. MR3646070

[39] SAWYER, S. and HARTL, D. (1985). A sampling theory for local selection. *J. Genet.* **64** 21–29.

[40] WATTERSON, G. A. (1976). Reversibility and the age of an allele. I. Moran's infinitely many neutral alleles model. *Theor. Popul. Biol.* **10** 239–253. MR0475994

[41] WATTERSON, G. A. (1977). Reversibility and the age of an allele. II. Two-allele models, with selection and mutation. *Theor. Popul. Biol.* **12** 179–196. MR0475995

[42] WATTERSON, G. A. and GUESS, H. A. (1977). Is the most frequent allele the oldest? *Theor. Popul. Biol.* **11** 141–160.

[43] WHITTAKER, E. T. and WATSON, G. N. (1996). *A Course of Modern Analysis*: *An Introduction to the General Theory of Infinite Processes and of Analytic Functions*; *with an Account of the Principal Transcendental Functions*. Cambridge Univ. Press, Cambridge. Reprint of the fourth (1927) edition. MR1424469

STATISTICS DEPARTMENT
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL # 3860
BERKELEY, CALIFORNIA 94720-3860
USA
E-MAIL: pitman@stat.berkeley.edu

ST. PETERSBURG STATE UNIVERSITY
7/9 UNIVERSITETSKAYA NAB.
ST. PETERSBURG, 199034
RUSSIA
E-MAIL: y.yakubovich@spbu.ru