Electronic Journal of Probability

Vol. 1 (1996) Paper no. 4, pages 1–28.

Journal URL
http://www.math.washington.edu/˜ejpecp/
Paper URL
http://www.math.washington.edu/˜ejpecp/EjpVol1/paper4.abs.html

RANDOM DISCRETE DISTRIBUTIONS DERIVED FROM SELF-SIMILAR RANDOM SETS

Jim Pitman

Department of Statistics, University of California, 367 Evans Hall # 3860, Berkeley, CA 94720-3860, pitman@stat.berkeley.edu

Marc Yor

Laboratoire de Probabilités, Tour 56, 4 Place Jussieu, 75252 Paris Cedex 05, France

Abstract: A model is proposed for a decreasing sequence of random variables $(V_1, V_2, \cdots)$ with $\sum_n V_n = 1$, which generalizes the Poisson-Dirichlet distribution and the distribution of ranked lengths of excursions of a Brownian motion or recurrent Bessel process. Let V_n be the length of the nth longest component interval of $[0,1]\backslash Z$, where Z is an a.s. non-empty random closed of $(0,\infty)$ of Lebesgue measure 0, and Z is self-similar, i.e. cZ has the same distribution as Z for every $c > 0$. Then for $0 \le a < b \le 1$ the expected number of n's such that $V_n \in (a,b)$ equals $\int_a^b v^{-1}F(dv)$ where the structural distribution F is identical to the distribution of $1 - \sup(Z \cap [0,1])$. Then

$F(dv) = f(v)dv$ where $(1 - v)f(v)$ is a decreasing function of v, and every such probability distribution F on $[0, 1]$ can arise from this construction.

Keywords: interval partition, zero set, excursion lengths, regenerative set, structural distribution.

AMS subject classification: 60G18, 60G57, 60K05.

Research supported in part by NSF grant MCS9404345.

Submitted to EJP on October 27, 1995. Final version accepted on February 20, 1996.

Contents

1 Introduction

Random discrete distributions have been widely studied, motivated by a variety of applications including the following:

- prior distributions in Bayesian statistics [12, 1],
- asymptotic distributions in combinatorics and number theory [42, 43, 41],
- models for gene frequencies in population genetics [20, 6, 11]
- models for species diversity in ecology [28, 10],
- the representation of partition structures [21],
- models for storage and search [4, 19, 7]
- analysis of the zero sets of stochastic processes such as Brownian motion and Brownian bridge [44, 35, 30, 32].

While the last of these applications was the main source of inspiration for the present study, the results described here admit interesting interpretations in some of the other applications as well.

Let $(V_n) = (V_1, V_2, \cdots)$ be a sequence of random variables such that $V_n \geq 0$ and $\sum_n V_n = 1$ a.s., where n always ranges over $\{1, 2, \cdots\}$. Call (V_n) a *random discrete distribution*, or RDD. Call a random variable V a *size-biased pick from* (V_n), if $V = V_N$ for a positive integer valued random variable N such that

$$P(N = n \,|\, V_1, V_2, \cdots) = V_n \qquad (n = 1, 2, \cdots) \tag{1}$$

This construction, and its iteration to define a *size-biased random permutation* of (V_n), play a key role in both theory and applications of random discrete distributions [14, 8, 33]. Denote by F the distribution on $(0, 1]$ of a size-biased pick V from (V_n). Following Engen [10], call F the *structural distribution* of (V_n). It is well known that many probabilities and expectations related to (V_n) can be expressed in terms of this one distribution F. For example, (1) implies that for any positive measurable function g

$$E\left[\sum_n g(V_n)\right] = E\left[\frac{g(V)}{V}\right] = \int_0^1 \frac{g(v)}{v} F(dv) \tag{2}$$

Taking $g(v) = 1(a < v < b)$ gives an expression in terms of F for the mean number of n's such that $a < V_n < b$. This shows in particular that if (V_n) is *ranked*, i.e. if $V_1 \geq V_2 \geq \cdots$ a.s., then the distribution of V_1 restricted to the interval $(1/2, 1]$ can be recovered from F:

$$P(V_1 \in dv) = v^{-1} F(dv) \qquad (1/2 < v \leq 1) \tag{3}$$

As noted in [33], the structural distribution F also appears in formulae related to Kingman's *partition structure* induced by (V_n), which is a natural construction of interest in several of the applications listed above.

Call a distribution F on $(0, 1]$ a *structural distribution* if F is the structural distribution of some RDD. Pitman [33] posed the problem of characterizing the set of all structural distributions, and gave a simple *necessary* condition for a distribution F to be structural, namely that for every $0 < a \leq 1$ (or, equivalently, for every $0 < a \leq 1/2$),

$$\int_0^a (1 - x) F(dx) \geq \int_{1-a}^1 (1 - x) F(dx) \tag{4}$$

This paper introduces a class of models for a RDD with the feature that the structural distribution can be identified explicitly. Analysis of these

RDD's shows that the following condition is *sufficient* for F to be a structural distribution. We note however that this condition is far from necessary, even assuming F has a density (See Example 19).

Condition 1 *F admits a density $f(u) = F(du)/du$ such that*

$$(1-u)f(u) \text{ is a decreasing function of } u \text{ for } 0 < u < 1. \tag{5}$$

From a mathematical point of view, it is natural to represent a RDD by the lengths of a random collection of disjoint open sub-intervals of $[0,1]$. The complement of such a random collection of intervals is then a *random closed subset* of $[0,1]$, as defined more formally in Section 2.

Definition 2 Let Z be a random closed subset of $(0,\infty)$ with Lebesgue$(Z) = 0$ a.s.. Say (V_n) *is derived from* Z, if V_n is the length of the nth longest component interval of $[0,1]\backslash Z$.

The assumption that Lebesgue$(Z) = 0$ ensures that $\sum_n V_n = 1$ a.s.. So (V_n) derived from Z is a ranked RDD. Think of each point of Z as the location of a *cut* in the line. Then (V_n) is defined by the ranked lengths of the intervals that remain after cutting $[0,1]$ at the points of Z. One natural construction of such a Z, corresponding to an arbitrary prescribed distribution for (V_n), is obtained from the *exchangeable interval partition* considered by Berbee [3] and Kallenberg [17]. Here we consider constructions with a different sort of symmetry:

Definition 3 SELF-SIM$_0$ *set.* Call Z *self-similar* if

$$Z \stackrel{d}{=} cZ \text{ for all } c > 0, \tag{6}$$

where $cZ = \{cz, z \in Z\}$, and $\stackrel{d}{=}$ denotes equality in distribution. (See Section 2 for the formal definition of the distribution of Z). Call Z a SELF-SIM$_0$ *set* if Z is an a.s. non-empty self-similar random closed subset of $\mathbb{R}^+$ with Lebesgue$(Z) = 0$ a.s..

That Condition 1 is sufficient for F to be a structural distribution is implied by the following theorem:

Theorem 4 *A distribution F on $(0,1]$ is the structural distribution of (V_n) derived from some* SELF-SIM$_0$ *set if and only if F satisfies Condition 1.*

This result is derived in Section 4 using the characterization of the structural distribution of a SELF-SIM$_0$ set provided by Theorem 7 below. Our formulation of this theorem was guided by the following two examples of SELF-SIM$_0$ sets which have been extensively studied. Both examples involve the beta(a,b) distribution on $(0,1)$, which is defined for $a>0, b>0$ by the probability density proportional to $u^{a-1}(1-u)^{b-1}, 0<u<1$.

For a random closed subset Z of $\mathbb{R}$, define

$$\begin{aligned} D_t &= \inf\{Z \cap (t,\infty)\} &(7)\\ G_t &= \sup\{Z \cap (-\infty,t]\} &(8)\\ A_t &= t - G_t &(9)\end{aligned}$$

Following terminology from renewal theory, when Z is a random discrete set of renewal times, we call (A_t) the *age process* derived from Z. If (V_n) is derived from Z, and $A_1>0$, then A_1 is one of the lengths in the sequence (V_n), say $A_1 = V_N$, where N is the *rank* of A_1 in (V_n). That is, $N=n$ if A_1 is the nth longest component interval of $[0,1]\backslash Z$.

Example 5 POISSON-DIRICHLET(θ). Suppose Z is the set of points of a Poisson process on $(0,\infty)$ with intensity measure $\theta x^{-1}dx$ for some $\theta>0$. Then the points of $Z\cap(0,1]$ can be ranked in decreasing order, say

$$Z \cap (0,1] = \{Z_1 > Z_2 > \cdots\} \tag{10}$$

It is known that Z_n may be represented as

$$Z_n = (1-X_1)\cdots(1-X_n) \quad (n \geq 1) \tag{11}$$

where $X_1 = A_1$ and the X_i are i.i.d. beta$(1,\theta)$ variables [16]. In terms of the X_i the sequence (V_n) is obtained by ranking the terms $\tilde{V}_n$ defined by

$$\tilde{V}_1 = X_1; \quad \tilde{V}_n = (1-X_1)\cdots(1-X_{n-1})X_n \qquad (n = 2,3,\cdots) \tag{12}$$

The distribution of (V_n) derived from this Z is known as the *Poisson-Dirichlet distribution* with parameter θ [19, 16]. It is known that $(\tilde{V}_n)$ is a size-biased permutation of (V_n) [27, 28, 8, 33]. In particular, $\tilde{V}_1 = A_1$ is a size-biased pick from (V_n), so the structural distribution of (V_n) is identical to the beta$(1,\theta)$ distribution of A_1.

Example 6 STABLE(α). Let Z be the closure of the set of zeros of a self-similar strong-Markov process B, such as a Brownian motion or a recurrent Bessel process, started at $B_0 = 0$. It is well known that Z is then the closure of the range of a stable subordinator of index α for some $0 < \alpha < 1$. For example, $\alpha = 1/2$ for Brownian motion, and $\alpha = (2 - \delta)/2$ for a Bessel process of dimension δ. The distribution of (V_n) in this case is an analog of the Poisson-Dirichlet distribution that has been studied by several authors [44, 29, 35]. It is well known that this Z is a.s. *perfect*, i.e. Z contains no isolated points. Consequently, Z is uncountable, and its points cannot be simply ranked as in the previous example. Still, it was shown in [35] that A_1 is a size-biased pick from (V_n), as in the previous example. So the structural distribution of (V_n) is again identical to the distribution of A_1, in this case beta$(1 - \alpha, \alpha)$, also known as generalized arcsine [9]. It was shown further in [30] that in this example a size-biased random permutation $(\tilde{V}_n)$ of (V_n), constructed with extra randomization, admits the representation (12) for independent beta$(1 - \alpha, n\alpha)$ random variables X_n.

Theorem 7 *Let (V_n) be the sequence of ranked lengths of component intervals of $[0,1]\backslash Z$, for Z a* SELF-SIM$_0$ *set. Let $A_1 := 1 - \sup\{Z \cap [0,1]\}$. Then $A_1 = V_N$ where (N, V_N) has the same joint distribution as if N were a size-biased pick from (V_n). Consequently:*

$$\text{the structural distribution of } (V_n) \text{ equals the distribution of } A_1, \tag{13}$$

and the distribution of N, the rank of A_1 in (V_n), is given by

$$P(N = n) = E(V_n) \qquad (n \in \mathbb{N}) \tag{14}$$

Theorem 7 is proved in Section 2. Note the subtle phrasing of the conclusion of Theorem 7. It is not claimed, nor is it true for every SELF-SIM$_0$ set Z, that A_1 is a size-biased pick from (V_n), as was observed in Examples 5 and 6. Spelled out in detail, the conclusion of Theorem 7 is that the rank N of A_1 in (V_n) has the following property, call it the *weak sampling property*:

$$\text{(weak sampling):} \quad \begin{array}{l} \text{for all positive measurable } f \\ E[f(V_n)1(N = n)] = E[f(V_n)V_n] \text{ for all } n \in \mathbb{N} \end{array} \tag{15}$$

Equivalently, by definition of conditional probabilities,

$$\text{(weak sampling):} \qquad P(N = n \,|\, V_n) = V_n \qquad \text{for all } n \in \mathbb{N}. \tag{16}$$

Compare with the *strong sampling property* which was observed in Examples 5 and 6:

$$\text{(strong sampling):} \qquad P(N = n \mid V_1, V_2, \cdots) = V_n \qquad \text{for all } n \in \mathbb{N} \tag{17}$$

To paraphrase Theorem 7,

$$\textit{every } \text{SELF-SIM}_0 \textit{ set has the weak sampling property.} \tag{18}$$

Example 20 in Section 5 shows that

$$\textit{not every } \text{SELF-SIM}_0 \textit{ set has the strong sampling property.} \tag{19}$$

Proposition 23 can be used to generate a large class of SELF-SIM_0 sets with the strong sampling property. But we do not know a nice sufficient condition for a SELF-SIM_0 set to have this property.

The most important conclusion of Theorem 7 is the identification of the structural distribution of (V_n) with the distribution of A_1. We provide another approach to this result in Section 4. The idea is to exploit the fact that Z is self-similar iff $\log Z$ is stationary, and make use of the generalizations to stationary random sets [31] of some standard formulae for stationary renewal processes. An advantage of this approach is that it gives an explicit description of all possible joint distributions of (G_t, D_t) derived from a SELF-SIM_0 set, which leads to Theorem 4.

2 Self-Similar Random Sets

Call Z a *random closed subset of* $\mathbb{R}$ if $\omega \to Z(\omega)$ is a map from a probability space $(\Omega, \mathcal{F}, P)$ to closed subsets of $\mathbb{R}$, and A_t is $\mathcal{F}$-measurable for every $t > 0$, where we define D_t, G_t and A_t in terms of Z as below Definition 2. To emphasize the Z underlying these definitions, we may write e.g. $A_t(Z)$ instead of A_t. Define the *distribution of* Z to be the distribution of the age process $(A_t(Z), t \geq 0)$ on the usual path space of cadlag paths. We refer to Azéma [2] for a general treatment of random closed subsets of $\mathbb{R}$.

A real or vector-valued process $(X_t, t \geq 0)$ is called *β-self-similar* where $\beta \in \mathbb{R}$ if for every $c > 0$

$$(X_{ct}, t \geq 0) \stackrel{d}{=} (c^\beta X_t, t \geq 0) \tag{20}$$

Such processes were studied by Lamperti [24, 25], who called them *semi-stable*. See [40] for a survey of the literature of these processes. A random closed subset Z of $\mathbb{R}$ is self-similar in the sense (6) iff its age process is 1-self-similar. A natural example of a self-similar random closed subset of $(0, \infty)$ is provided by the closure of the zero set of a β-self-similar process for any β.

Assume now that Z is a SELF-SIM$_0$ set as in Definition 3. Let $V_n(t)$ be the length of the nth longest component interval of $[0,t]\backslash Z$. Then the sequence valued process $((V_n(t), n \in \mathbb{N}), t \geq 0)$ is 1-self-similar, and

$$\sum_n V_n(t) = t \text{ for all } t \geq 0 \text{ a.s.} \tag{21}$$

The random sequence $(V_n(t)/t, n \in \mathbb{N})$ then defines a ranked RDD which has the same distribution for every $t > 0$.

Proof of Theorem 7. Let N_t denote the rank of A_t in the sequence of ranked lengths $(V_n(t), n = 1, 2, \cdots)$ of $[0,t]\backslash Z$:

$$N_t = \sup\{n : A_t = V_n(t)\}, \tag{22}$$

with the convention $\sup \emptyset = 0$, so that

$$\{t : N_t = 0\} = \{t : A_t = 0\} \subseteq Z \tag{23}$$

It is a key observation that

$$V_n(t) = \int_0^t ds 1(N_s = n) \quad (n \in \mathbb{N}). \tag{24}$$

To check (24), start from the identity (21). Fix an $m \in \mathbb{N}$ and integrate $1(A_t = V_m(t))$ with respect to both sides of (21). Since for each n, $dV_n(t)$ is carried by the set $\{t : A_t = V_n(t)\}$, and this set differs from $\{t : N_t = n\}$ by at most the discrete set of times $\{t$ such that $A_t = V_k(t)$ for more than one $k\}$, we obtain (24) with m instead of n.

It is clear that $(N_t, t \geq 0)$ satisfies the assumptions of the following Lemma. Theorem 7 follows immediately from the conclusion of the Lemma. $\square$

Lemma 8 *Suppose that a process* $(N_t, t \geq 0)$ *with values in* $\mathbb{N}_0 := \{0, 1, 2, \cdots\}$ *is 0-self-similar, i.e. for every* $c > 0$

$$(N_{ct}, t \geq 0) \stackrel{d}{=} (N_t, t \geq 0) \tag{25}$$

Let $V_n(t)$ be defined by (24) for all $n \in \mathbb{N}_0$. Then for every $n \in \mathbb{N}_0$ and every $t > 0$

$$P[N_t = n \mid V_n(t)] = \frac{V_n(t)}{t} \tag{26}$$

In particular, if (21) holds, then for every $t > 0$, $(N_t, V_{N_t}(t)/t) \stackrel{d}{=} (N, V_N(1))$ where N is a size-biased pick from $(V_n(1), n \in \mathbb{N})$

Proof. Apply the next Lemma to the 0-self-similar process $X_t = 1(N_t = n)$. □

Lemma 9 *Let $\bar{X}_t = \frac{1}{t}\int_0^t X_s ds$ where $(X_s, s \geq 0)$ is a jointly measurable real-valued 0-self-similar process such that $E(|X_1|) < \infty$. Then for every $t > 0$*

$$E[X_t \mid \bar{X}_t] = \bar{X}_t \tag{27}$$

Proof. Because (X_t) is 0-self-similar,

$$(X_t, \bar{X}_t) \stackrel{d}{=} (X_1, \bar{X}_1) \text{ for all } t. \tag{28}$$

It will be shown that (27) follows from this identity. As a first consequence of (28), it suffices to prove (27) for $t = 1$. Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a C^1 function with $f(0) = 0$. The chain rule for Lebesgue integrals (see e.g. [39], Chapter 0, Prop. (4.6)) gives for all $t \geq 0$

$$f(t\bar{X}_t) = \int_0^t f'(s\bar{X}_s)X_s ds \tag{29}$$

Take $t = 1$ and use (28) to obtain

$$E[\,f(\bar{X}_1)\,] = \int_0^1 E[\,f'(s\bar{X}_s)X_s]ds = E\left[\,\int_0^1 f'(s\bar{X}_1)X_1 ds\,\right] \tag{30}$$

Now assume further that $f'(0) = 0$, so there is a vanishing contribution from the event $(\bar{X}_1 = 0)$ in the rightmost expectation above. Then we see that for all C^1 functions f with $f(0) = f'(0) = 0$

$$E[\,f(\bar{X}_1)\,] = E\left[\,f(\bar{X}_1)\bar{X}_1^{-1}X_1; \bar{X}_1 \neq 0\right] \tag{31}$$

By standard measure theory, (31) must hold also for every Borel f such that both expectations in (31) are well-defined and finite. Taking $f(x) = x1(x \in B)$ shows that for every Borel subset B of $\mathbb{R}$ that does not contain 0

$$E[\bar{X}_1 1(\bar{X}_1 \in B)] = E\left[X_1 1(\bar{X}_1 \in B)\right] \tag{32}$$

But (32) holds also for $B = \mathbb{R}$ because, using (28) again

$$E(\bar{X}_1) = E\left[\int_0^1 X_s ds\right] = \int_0^1 E(X_s)ds = \int_0^1 E(X_1)ds = E(X_1) \tag{33}$$

Subtracting (32) for $B = \mathbb{R} - \{0\}$ from (33) gives (32) for $B = \{0\}$, hence (32) for all Borel B. This is (27). □

3 Stationary random closed subsets of $\mathbb{R}$

Call a random closed subset Z of $\mathbb{R}$ *stationary* if Z is shift invariant: i.e. $Z \stackrel{d}{=} Z + c$ for any $c > 0$, where $\stackrel{d}{=}$ denotes equality in distribution, and $Z + c = \{z + c, z \in Z\}$. That is to say, Z is stationary iff its age process (A_t) is a stationary process in the usual strict sense. Then the $\mathbb{R}^2$-valued process $((A_t, D_t - t), t \geq 0)$ is also stationary. In particular $(A_t, D_t - t) \stackrel{d}{=} (A_0, D_0)$ for all $t \geq 0$.

Definition 10 Call Z a STATIONARY$_0$ set if Z is an a.s. non-empty stationary random closed subset of $\mathbb{R}$ with Lebesgue$(Z) = 0$ a.s..

Clearly, Z is a SELF-SIM$_0$ set iff $\log Z$ is a STATIONARY$_0$ set. The following proposition is the restriction to STATIONARY$_0$ sets of the result stated in Section VI of [31].

Proposition 11 [31] *Let Z be a* STATIONARY$_0$ *set. Define $L = A_0 + D_0$, so L is the length of the component interval of Z^c that contains 0. Then $P(0 < L < \infty) = 1$ and*

$$(A_0, D_0) = (LV, L(1 - V)) \tag{34}$$

where V has uniform distribution on $(0, 1)$, and V is independent of L.

Given a probability distribution F_0 on $(0, \infty)$, it is easy to construct a STATIONARY$_0$ set such that L has distribution F_0.

Example 12 STATIONARY-LATTICE(F_0). Call Z a *stationary random lattice with spacing distribution* F_0 if $Z = \{L(\mathbb{Z} - V)\}$ where $\mathbb{Z}$ is the set of integers, L is a random variable with distribution F_0, and V is uniform$(0, 1)$ independent of L.

If F_0 is degenerate at $c > 0$, so $L = c$ is constant, it is obvious that STATIONARY-LATTICE(F_0) is the only possible distribution of a STATIONARY$_0$ set such that L has distribution F_0. But there are many other possible distributions of Z corresponding to each non-degenerate F_0. Among these is the following:

Example 13 STATIONARY-REGEN$_0$ (F_0). This is the unique distribution of a STATIONARY$_0$ set Z that is *regenerative* subset of $\mathbb{R}$ in the sense of [26], and such that L has distribution F_0. See Fristedt [13], who gives the following construction of Z, and further references. Let

$$Z = \{-A_0 - Z_1\} \cup \{D_0 + Z_2\} \tag{35}$$

where A_0 and D_0 are defined by (34) in terms of L with distribution F_0 and an independent uniform V, and, independent of these variables, Z_1 and Z_2 are two independent copies of the closed range of a pure jump subordinator with Lévy measure $\Lambda(dx) = cx^{-1}F_0(dx)$, for an arbitrary constant $c > 0$. It is easily seen that this yields the same distribution as various other constructions of Z that can be found in the literature. It is immediate from the above construction that this Z is reversible: $Z \stackrel{d}{=} -Z$. If Λ has total mass 1, then the STATIONARY-REGEN$_0$ set Z defined by (35) is just the stationary point process derived as in renewal theory from i.i.d. spacings with distribution Λ. Then $F_0(dx) = \mu^{-1}x\Lambda(dx)$ where μ is the mean of Λ. If $\Lambda(0, \infty) = \infty$, then Z is a.s. uncountable.

Another method of constructing a STATIONARY-REGEN$_0$ (F_0) is to let Z be the closure of the zero set of a suitable stationary strong-Markov process X. One can always take X to be the stationary version of the age process derived from the subordinator with Lévy measure Λ described above. This is the method of Horowitz [15]. But zero sets of other Markov processes X may be considered. For example, the zero set Z of a stationary diffusion process X on the line, for which 0 is recurrent, gives a $\text{STATIONARY-REGEN}_0$ set with $\Lambda(0,\infty) = \infty$. See Knight [22] and Kotani-Watanabe [23] regarding which Lévy measures Λ can be obtained by this construction from a diffusion.

Example 14 Suppose $W = \log Z$ where Z is the stable(α) regenerative SELF-SIM_0 set. If we represent Z as the zero set of $(B(t), t \geq 0)$ for a Brownian motion B (in case $\alpha = 1/2$) or Bessel process B of dimension $\delta = 2 - 2\alpha$, then W is the zero set of the process $(X_s, s \geq 0)$ defined by $X_s = e^{-s/2}B(e^s)$. It is well known that for B a Brownian motion, X is a one-dimensional Ornstein-Uhlenbeck process. See Section 6 of [35] regarding the Bessel case.

4 The joint law of (G_1, D_1) for a SELF-SIM_0 set

We start this section by presenting an alternative derivation of the key identity (13) that is part of the conclusion of Theorem 7.

Another proof of (13). Let Z be a SELF-SIM_0 set. Let $V_n := V_n(1)$, the length of the nth longest component interval of $[0,1]\backslash Z$. Let U be uniform$(0,1)$ independent of Z. A size-biased pick from (V_n) is provided by the length of the component interval covering U, that is $(D_U \wedge 1) - G_U$. So an equivalent of (13) is

$$(D_U \wedge 1) - G_U \stackrel{d}{=} 1 - G_1 \tag{36}$$

which, by scaling, amounts to

$$((U\,D_1) \wedge 1) - UG_1 \stackrel{d}{=} 1 - G_1 \tag{37}$$

Ignoring null sets, the event $(D_U \geq 1)$ is identical to $(U > G_1)$. On this event $G_U = G_1$, so the left side of (36) reduces to $1 - G_1$, and (36) can be rewritten as

$$(D_U - G_U)1(D_U < 1) \stackrel{d}{=} (1 - G_1)1(U \leq G_1) \tag{38}$$

That is to say

$$P(D_U - G_U \in dx; D_U < 1) = (1 - x)P(1 - G_1 \in dx) \qquad (0 < x < 1) \tag{39}$$

or equivalently, by scaling,

$$P(U(D_1-G_1)\in dx; UD_1<1)=(1-x)P(1-G_1\in dx) \qquad (0<x<1) \quad (40)$$

Consider now the random subset $\log Z$ of $\mathbb{R}$. Since Z is a SELF-SIM$_0$ set, $\log Z$ is a STATIONARY$_0$ set. Since $z\to\log z$ is increasing with $\log 1=0$,

$$\begin{aligned}\log D_1 &= D_0(\log Z)=LV &(41)\\ \log G_1 &= G_0(\log Z)=-L(1-V) &(42)\end{aligned}$$

where $D_0(\log Z)=\inf\{\log Z\cap(0,\infty)\}$, $G_0(\log Z)=\sup\{\log Z\cap(-\infty,0]\}$.

$$L:=\log(D_1/G_1)=D_0(\log Z)-G_0(\log Z)>0 \text{ a.s.} \qquad (43)$$

and V is uniform on $(0,1)$, and independent of L and U. Thus the identity (37) reduces to the following: for such L, V and U,

$$[(Ue^{LV})\wedge 1]-Ue^{L(V-1)}\stackrel{d}{=}1-e^{L(V-1)} \qquad (44)$$

As noted in Section 3, L can have an arbitrary distribution F_0 on $(0,\infty)$. By conditioning on L, (44) holds no matter what the distribution of L iff (44) holds for L an arbitrary positive constant, say $L=\log C$. Now (44) reduces to this:

for any constant $C>1$, and independent uniform $(0,1)$ variables U and V,

$$[(UC^V)\wedge 1]-UC^{V-1}\stackrel{d}{=}1-C^{V-1} \qquad (45)$$

Put $W=1-V$, so U and W are i.i.d. uniform $(0,1)$ too. By the same reduction made earlier in (40), it is enough to check that for $0<x<1$,

$$P[(C-1)UC^{-W}\in dx; UC^{-W}<C^{-1}]=(1-x)P(1-C^{-W}\in dx) \qquad (46)$$

Since $1-C^{-W}<1-C^{-1}$, the distribution on the right side of (46) vanishes for $x>1-C^{-1}$. So it does on the left, since the condition $UC^{-W}<C^{-1}$ makes

$$(C-1)UC^{-W}<(C-1)C^{-1}=1-C^{-1}$$

as well. So it is enough to compute the densities of both sides in (46) relative to dx for $0<x\le 1-C^{-1}$, and show they are equal. On the one hand,

conditioning on W shows that the density on the left side of (46) is constant over this range and equal to $E[C^W/(C-1)]$. On the other hand, the change of variables

$$x = 1 - C^{-w}; \quad \frac{dw}{dx} = \frac{1}{(1-x)\log C}$$

shows that the density on the right side of (46) is constant and equal to $1/\log C$. An easy integration confirms that the two constants are equal. □

Proof of Theorem 4. Consider the distribution of A_1 for a SELF-SIM$_0$ set. By application of Proposition 11 as in the preceding argument,

$$A_1 = 1 - G_1 = 1 - \exp(-Y) \tag{47}$$

where $Y := LW$ for a uniform$(0,1)$ variable W independent of L, and, from the discussion below Proposition 11, the random variable $L := \log(D_1/G_1)$ can have an arbitrary distribution F_0 on $(0,\infty)$. The distribution of Y is given by the density

$$P(Y \in dy) = h(y)dy \qquad (y > 0) \tag{48}$$

where

$$h(y) := \int_{(y,\infty)} x^{-1} F_0(dx). \tag{49}$$

¿From (47) and (48), the change of variables

$$a = 1 - e^{-y}, \quad y = -\log(1-a), \quad dy = da/(1-a) \tag{50}$$

shows that

$$f(a) := \frac{h(-\log(1-a))}{1-a} \qquad (0 < a < 1) \tag{51}$$

serves as a probability density for A_1. Formula (48) sets up a 1-1 correspondence between probability distributions F_0 on $(0,\infty)$, and probability densities h on $(0,\infty)$ satisfying

$$y \to h(y) \text{ is decreasing and right continuous for } 0 < y < \infty \tag{52}$$

Formula (51) in turn sets up a 1-1 correspondence between such probability densities h on $(0,\infty)$ and probability densities f on $(0,1)$ satisfying

$$a \to (1-a)f(a) \text{ is decreasing and right continuous for } 0 < a < 1 \tag{53}$$

The conclusion of Theorem 4 is now clear. □

As a complement to Theorem 4, the following corollary summarizes the collection of distributional identities implicit in the above argument:

Corollary 15 *The distribution of D_1/G_1 derived from a* SELF-SIM$_0$ *set can be any distribution on $(1,\infty)$. Let F_0 denote the corresponding distribution of $L := \log(D_1/G_1)$, which can be any distribution on $(0,\infty)$. Then*

(i) *The joint distribution of (G_1, D_1) is determined by F_0 via the formula*

$$(G_1, D_1) = (e^{-LW}, e^{-L(1-W)}) \tag{54}$$

where $W{:=} -(\log G_1)/L$ has uniform$(0,1)$ distribution, independent of L with distribution F_0;

(ii) *the random variables $-\log G_1$ and $\log D_1$ are identically distributed with a decreasing density $h(y)$ on $(0,\infty)$*

(iii) *the random variables G_1 and $1/D_1$ are identically distributed, with density $g(u)$ on $(0,1)$ such that $ug(u)$ is an increasing function of u;*

(iv) *the random variables $1/G_1$ and D_1 are identically distributed, with density $k(x)$ on $(1,\infty)$ such that $xk(x)$ is a decreasing function of x;*

(v) *the right-continous version of the density h is related to F_0 by (49); each of the densities h,g and k is arbitrary subject to the constraints stated above, and each of these densities can be recovered from any of the others via the formulae*

$$g(u) = u^{-2}k(u^{-1}) = u^{-1}h(-\log u) \quad (0 < u < 1) \tag{55}$$

$$h(y) = e^{-y}g(e^{-y}) = e^{y}k(e^{y}) \quad (0 < y < \infty) \tag{56}$$

$$k(x) = x^{-2}g(x^{-1}) = x^{-1}h(\log x) \quad (1 < x < \infty); \tag{57}$$

Remark 16 *Inversion.* The identity in distribution $D_1 \stackrel{d}{=} 1/G_1$ for any SELF-SIM$_0$ set, implied by (ii) above, can be seen immediately by scaling, using the relation $(G_t < u) = (t < D_u)$. In case Z is the stable(α) SELF-SIM$_0$ set, as in Example 6, the joint distribution of (G_1, D_1) is well known [9]. In particular, the distribution of G_1 is beta$(\alpha, 1-\alpha)$. The identity in distribution $D_1 \stackrel{d}{=} 1/G_1$ can be strengthened to $(G_t, t \geq 0) \stackrel{d}{=} (1/D_{1/t}, t \geq 0)$ in this case, and more generally whenever Z is *invertible*, that is $Z \stackrel{d}{=} 1/Z$ where $1/Z = \{1/z, z \in Z\}$. This amounts to reversibility of $\log Z$, which holds whenever $\log Z$ is regenerative [26], and also if $\log Z$ is a stationary random lattice. However, as shown by an example in [31], not all STATIONARY$_0$ sets are reversible, so not all SELF-SIM$_0$ sets are invertible.

Remark 17 *Distribution of $D_1 - G_1$.* For the stable(α) set Z, the distribution of the length $D_1 - G_1$ of the complementary interval covering 1 is found by integration from the joint law of (G_1, D_1) to be

$$P(D_1 - G_1 \in dx)/dx = [\Gamma(\alpha)\Gamma(1-\alpha)]^{-1} x^{-\alpha-1}(1-x)_+^{\alpha} \quad (x > 0) \tag{58}$$

where the last factor equals $(1-x)^\alpha$ for $0 < x < 1$ and 1 for $x \geq 1$. It can also be shown that this distribution of $D_1 - G_1$ is the distribution of $Z_{1-\alpha,\alpha}/Z_{\alpha,1}$ for independent $Z_{1-\alpha,\alpha}$ and $Z_{\alpha,1}$, where $Z_{a,b}$ has beta(a,b) distribution.

The distribution of $D_1 - G_1$ is also easily described for the set of points Z of a Poisson point process with intensity $\theta x^{-1} dx$ as considered in Example 5. In that case $D_1 - G_1$ is the sum of independent variables $D_1 - 1$ and $1 - G_1$, where $1 - G_1$ has beta$(1, \theta)$ distribution, and $D_1 \stackrel{d}{=} 1/G_1$. So the density of $D_1 - G_1$ can be expressed as a convolution.

For a general SELF-SIM$_0$ set, it is clear from (54) that the joint law of (G_1, D_1) has a density relative to Lebesgue measure in the plane iff F_0 is absolutely continuous. Still, it can be seen as follows that $D_1 - G_1$ always has a density, no matter what F_0. Use (54) to write

$$D_1 - G_1 = e^{-LU}(e^L - 1) \tag{59}$$

Conditioning on L gives

$$P(D_1 - G_1 \in dy | L = \ell) = \frac{1(1 - e^{-\ell} < y < e^\ell - 1)dy}{y\ell} \tag{60}$$

Integrating out with respect to $F_0(d\ell)$ gives a general if unwieldy formula for the density of $D_1 - G_1$. We do not know whether F_0 can be recovered from the density of $D_1 - G_1$, or if there is any nice description of all possible densities for $D_1 - G_1$ as Z ranges over all SELF-SIM$_0$ sets.

Example 18 *The zero set of a perturbed Brownian motion.* The above formulae can be applied to the SELF-SIM$_0$ set Z defined by the zero set of the perturbed Brownian motion $(|B_t| - \mu L_t, t \geq 0)$ studied in [5]. Here B is a standard Brownian motion, $(L_t, t \geq 0)$ is its local time process at zero, and $\mu > 0$ is a parameter. The law of G_1, found explicitly in [5] turns out to be fairly complicated. Still, without further calculation, the above results show how this law determines the structural distribution of (V_n) derived from this Z, the law of D_1, and joint law of (G_1, D_1). It seems intuitively clear that this Z is not invertible, but we do not see a proof.

5 Examples

Example 19 *A (V_n) with $V_n > 0$ for all n such that the structural distribution has a density f not satisfying Condition* 1. Let V_1 have a density with support $[q, 1]$ for some $1/2 < q < 1$. Let $V_{n+1} = (1 - V_1)W_n$ for $n \geq 1$ where (W_n) is any RDD with $W_n > 0$ for all n, whose structural distribution has a density that does not vanish on $(0, 1)$. Then (V_n) is a RDD whose structural distribution F has a density f that is strictly positive on $(0, 1-q)$ and $(q, 1)$, but which vanishes on $(1-q, q)$. Obviously this f does not satisfy Condition 1.

Example 20 *A* SELF-SIM$_0$ *set Z that does not have the strong sampling property.* For every possible structural density f for (V_n) derived from a SELF-SIM$_0$ set, as described in Theorem 4, there is a SELF-SIM$_0$ set that generates a (V_n) with the given structural density f, and which does not have the strong sampling property (17). Such a SELF-SIM$_0$ set Z is obtained as $Z = \exp(W)$ where W is the STATIONARY-LATTICE(F_0) for F_0 the distribution in (49) derived from f as in (53). Let L be the length of the component interval of W^c that contains 0. So L has distribution F_0. And $Z := \{\exp(L(\mathbb{Z} - V))\}$ where V is uniform $(0, 1)$ independent of L, and $\mathbb{Z}$ is the set of integers. Consequently,

$$Z \cap (0, 1] = \{Z_1 > Z_2 > \cdots\} \text{ a.s.} \tag{61}$$

where $Z_n = e^{-L(n-1+V)}$ for $n = 1, 2, \cdots$, and the spacings $\tilde{V}_n := Z_{n-1} - Z_n$, where $Z_0 := 1$, are given by

$$\tilde{V}_1 = 1 - e^{-LV}; \qquad \tilde{V}_n = e^{-L(n-2+V)} - e^{-L(n-1+V)} \quad (n \geq 2)$$

The sequence (V_n) is obtained by ranking $(\tilde{V}_n)$. The expression for $\tilde{V}_n$ shows that $\tilde{V}_2, \tilde{V}_3, \cdots$ is a geometric progression with common ratio e^{-L}. Let N be the rank of $\tilde{V}_1$ in (V_n), so $\tilde{V}_1 = V_N$. Clearly $V_n = \tilde{V}_n$ for all $n > N$, so

$$P\left(\frac{V_{n+1}}{V_n} = e^{-L} \text{ for all sufficiently large } n\right) = 1$$

and N can be recovered from (V_n) as

$$N = \max\left\{n : \frac{V_{n+1}}{V_n} \neq e^{-L}\right\} = \max\left\{n : \frac{V_{n+1}}{V_n} \neq \frac{V_{n+2}}{V_{n+1}}\right\} \tag{62}$$

Thus both L and N are measurable functions of (V_n). In particular, N is not a size-biased pick from (V_n) in the sense of (17).

Example 21 [10, 30, 38] POISSON-DIRICHLET(α, θ). This distribution for a ranked RDD with two parameters, abbreviated PD(α, θ), and defined for

$$0 \leq \alpha < 1 \text{ and } \theta > -\alpha \tag{63}$$

generalizes the one parameter POISSON-DIRICHLET(θ) distribution of Example 5, which is the special case PD$(0, \theta)$. It was shown in [30] that the PD$(\alpha, 0)$ distribution of (V_n) is that derived from the stable(α) set Z, as in Example 6, while PD(α, α) is the distribution of (V_n) derived from this stable (α) set Z by conditioning on $0 \in Z$, an operation made precise in [44], [18]. A sequence (V_n) with PD(α, θ)distribution can be constructed by ranking $(\tilde{V}_n)$ defined by the residual allocation model (12) for independent X_n such that X_n has beta $(1-\alpha, \theta+n\alpha)$ distribution. Moreover $(\tilde{V}_n)$ is then the size-biased permutation of (V_n) [10, 30, 33], and consequently

$$\text{the structural distribution of PD}(\alpha, \theta)\text{is beta}(1-\alpha, \theta+\alpha) \tag{64}$$

As shown by Examples 5 and 6, for $\alpha = 0$ or $\theta = 0$ the following statement is true:

> the PD(α, θ)distribution is generated by the unique SELF-SIM$_0$ set Z such that $\log Z$ is a STATIONARY-REGEN$_0$ set and the distribution of $A_1(Z)$ is the beta$(1-\alpha, \theta+\alpha)$ distribution required by (64) and (13).

But we do not know if this holds for any other choices of (α, θ). It is easily checked that for (α, θ) in the range (63) this beta distribution on $(0, 1)$ satisfies the necessary Condition 1 for existence of a SELF-SIM$_0$ set generating a (V_n) with this structural distribution. In Corollary 26 we show how to derive PD$(\alpha, \theta)$from a SELF-SIM$_0$ set for $0 < \alpha < 1$ and $\theta > 0$, but we do not know whether this is possible for $0 < \alpha < 1$ and $-\alpha < \theta \leq 0$.

6 Operations

There are some natural operations related to both random discrete distributions and self-similar random sets, which allow examples to be combined in some interesting ways.

Define the *ranked product* of two RDD's (U_n) and (V_n) defined on the same probability space to be the RDD (W_n) obtained by ranking the collection of products $\{U_m V_n, m \in \mathbb{N}, n \in \mathbb{N}\}$. As noted in [33], if $\tilde{U}_1$ is a size-biased pick from (U_n) and $\tilde{V}_1$ is a size-biased pick from (U_n), and $(\tilde{U}_1, U_1, U_2, \cdots)$ and $(\tilde{V}_1, V_1, V_2, \cdots)$ are independent, then $\tilde{W}_1 := \tilde{U}_1 \tilde{V}_1$ is a size-biased pick from the ranked product (W_n) of (U_n) and (V_n). So the set of structural distributions on $(0,1]$ is closed under the multiplicative analog of convolution. In particular, if f and g are two structural densities then so is h defined by

$$h(u) := \int_0^1 y^{-1} f(u/y) g(y) dy \quad (0 < u < 1) \tag{65}$$

Let $P \bullet Q$ denote the distribution of the ranked product of a RDD (P), i.e. a RDD with distribution P, and an independent RDD (Q). Let STR(P) denote the structural distribution on $(0,1]$ of a RDD (P), and let $*$ denote the multiplicative convolution operation on distributions on $(0,1]$. Then the above remarks may be summarized as follows: $\mathrm{STR}(P \bullet Q) = \mathrm{STR}(P) * \mathrm{STR}(Q)$. Note that the operation $\bullet$ on distributions of RDD's is commutative: $P \bullet Q = Q \bullet P$. Note also that with mild non-degeneracy assumptions on P and Q,

if (W_n) is a RDD $(P \bullet Q)$ *then with probability 1 there are distinct positive integers (k, ℓ, m, n) with $W_k/W_\ell = W_m/W_n$.* (66)

So, for example, $P \bullet Q$ could not be $\mathrm{PD}(\alpha, \theta)$for any (α, θ) .

A more interesting operation on laws of RDD's is the *composition* operation $\otimes$ defined as follows. Given two laws P and Q for a RDD, let (U_n) be a RDD (P), and, independent of (U_n), let $(V_{mn}, n = 1, 2, \cdots,)$, $m = 1, 2 \cdots$ be a sequence of i.i.d. copies of a RDD (Q). Let $P \otimes Q$ be the law of the RDD obtained by ranking the collection of products $\{U_m V_{mn}, m \in \mathbb{N}, n \in \mathbb{N}\}$. It is easily seen that the operation $\otimes$, like $\bullet$, has the property $\mathrm{STR}(P \otimes Q) = \mathrm{STR}(P) * \mathrm{STR}(Q)$. However, except in trivial cases, $P \otimes Q \neq P \bullet Q$. This is clear because mild conditions on P and Q ensure that the probability considered in (66) becomes 0 for $P \otimes Q$ instead of $P \bullet Q$. Indeed, the composition operation $\otimes$ is not even commutative. This is easily seen as follows. Take one of the laws, say P, to be the degenerate distribution that assigns probability 1 to the sequence $(1/2, 1/2, 0, \cdots)$, and let Q be the law of any (V_n) such that V_n has a continuous distribution for each n and $V_1 > V_2 > \cdots$ a.s.. Then (W_n) governed by $P \otimes Q$ has $W_1 > W_2 > \cdots$ a.s. whereas (W_n) governed by

$Q \otimes P$ has $W_1 = W_2 > W_3 = W_4 > \cdots$ a.s. Typically then, $P \bullet Q, P \otimes Q$ and $Q \otimes P$ will be three distinct laws for a RDD with the same structural distribution $\text{STR}(P) * \text{STR}(Q)$.

A nice illustration of the composition operation is provided by the following result of [34]:

$$\text{for } \alpha > 0 \text{ and } \theta > 0, \ \text{PD}(0,\theta) \otimes \text{PD}(\alpha, 0) \ = \ \text{PD}(\alpha,\theta) \tag{67}$$

If, as in the above examples, both P and Q can be derived from a SELF-SIM$_0$ set, it is natural to ask whether $P \bullet Q$ and $P \otimes Q$ can be so derived. This is achieved for $\otimes$ by the following construction.

Construction 22 *Let X and Y be two random closed subsets of $\mathbb{R}$ of Lebesgue measure 0. Let $(\gamma_n, \delta_n), n \in \mathbb{N})$ be a list of all the component intervals of the complement of X. Let $Y_1, Y_2, \cdots$ be a sequence of independent copies of Y, independent also of X. Let*

$$Z = X \cup \bigcup_{n=1}^{\infty} [\{\gamma_n + Y_n\} \cap (\gamma_n, \delta_n)] \tag{68}$$

Informally, the new set Z contains all the points of X, and, within each component interval of X_c, the new set also contains points derived from a copy of Y shifted to start at the left end of the interval. Some basic properties of this construction are stated in the following Proposition, whose proof is straightforward and left to the reader:

Proposition 23 *Let P and Q denote the distributions of the RDD's derived from SELF-SIM$_0$ sets X and Y repectively. Let Z be constructed from X and Y as in Construction 22. Then Z is a SELF-SIM$_0$ set, the distribution of the RDD derived from Z is $P \otimes Q$, with structural distribution $\text{STR}(P \otimes Q) = \text{STR}(P) * \text{STR}(Q)$. Moreover, if both X and Y have the strong sampling property (17) then so does Z.*

A consequence of this proposition, which can also be checked directly, is the following:

Corollary 24 *The set of densities on $(0,1)$ satisfying Condition 1 is closed under multiplicative convolution.*

Remark 25 *Finite Unions.* If $Z_1, \cdots, Z_m$ are m independent SELF-SIM$_0$ sets, it is easily seen that their union Z is also a SELF-SIM$_0$ set. Since $A_1(Z) = \max_i A_1(Z_i)$, Theorem 7 identifies the structural distribution of the RDD derived from Z as the distribution of $\max_i \tilde{V}_i$ where $\tilde{V}_i$ is a size-biased pick from the RDD derived from Z_i. Let f_i denote the structural density of $\tilde{V}_i$ derived from Z_i. Then the structural density f derived from Z is

$$f(v) = \sum_i f_i(v) \prod_{j \neq i} \int_v^1 f_j(u) du$$

So the class of densities satisfying Condition 1 is closed under this operation too.

Note that if X is *discrete*, e.g. the Poisson process generating PD$(0, \theta)$ as in Example 21, and Y is *perfect*, like the stable(α) set generating PD$(\alpha, 0)$, then laying down shifted copies of Y in the component intervals of X, as in Construction 22, yields a perfect set Z. But if the roles of X and Y are switched, laying down shifted copies of X in the component intervals of Y yields a set that is a.s. neither discrete nor perfect. Certainly, the distributions of the sets Z so obtained are different, but whether or not the derived RDD's have the same distribution is not so obvious.

By combining the identity (67) with the above proposition, we obtain:

Corollary 26 *For every $\alpha > 0$ and $\theta > 0$, there exists an a.s. perfect* SELF-SIM$_0$ *set Z with the strong sampling property such that the* RDD *derived from Z has* PD(α, θ) *distribution.*

7 Open problems

In the setting of Lemma 8, fix t and write simply N for N_t and $V(B)$ for $\sum_{n \in B} V_n(t)/t$ for a subset B of $\mathbb{N}$. Applying Lemma 9 to $X_t = 1(N_t \in B)$ shows that for every subset B of $\mathbb{N}$, $P(N \in B | V(B)) = V(B)$. However, as in the discussion around (16) and (17), Example 20 shows that it does not necessarily follow that $P(N \in B | V(C), C \subseteq \mathbb{N})$ equals $V(B)$, as it does in Examples 5 and 6. See [36] for some applications of this property in the setting of Example 6. It is natural to ask what additional hypothesis is appropriate for this stronger conclusion to hold in a more general setting, but we do not have an answer to this question. In essence, the problem is the following:

Problem 27 Find a condition that implies the identity *(27)* for a vector-valued 0-self-similar process X.

See [37] for a number of reformulations of (27) and further discussion, including a simple example of an $\mathbb{R}^2$-valued 0-self-similar process X for which (27) fails to hold.

We do not know much about RDD's derived from SELF-SIM$_0$ sets besides the results presented in this paper. Some obvious questions seem very difficult to tackle. For instance:

Problem 28 Is it possible to characterize the set of all possible laws of RDD's that can be derived from a SELF-SIM$_0$ set Z?

Less abstractly, given some description of the distribution of a SELF-SIM$_0$ set or perhaps another random closed Z, there is the problem of how to describe the distribution of (V_n) derived from Z. Several papers in the literature can be viewed as treating instances of this problem for Z's of various special forms [16, 29, 38]. Problem 30 describes a SELF-SIM$_0$ set Z for which this question remains to be answered. For the random closed subset Z of $(0,1)$ associated with an *exchangeable interval partition* of $(0,1)$ derived from a RDD as in Berbee [3], Kallenberg [17], it is obvious that the law of Z is uniquely determined by that of the RDD. But if there exists a SELF-SIM$_0$ set Z that generates the RDD, uniqueness of the law of Z is not so obvious:

Problem 29 Given that a RDD with a particular distribution can be derived from some SELF-SIM$_0$ set Z, is the distribution of such a Z unique?

We do not even know if there is uniqueness in law for the two most basic examples 5 and 6. To conclude, we pose the following:

Problem 30 Suppose $Z = \exp W$ for W a STATIONARY-REGEN$_0$ (F_0) with $\int_0^\infty x^{-1} F_0(dx) < \infty$, so W is the set of points in a stationary renewal sequence. Let $\tilde{V}_n$ be the sequence of spacings between the points of the discrete set Z, as defined in (10) and (12). For which F_0 is it the case, as it is for W a homogeneous Poisson process, that $(\tilde{V}_n)$ is a size-biased random permutation of the ranked sequence (V_n)? Example 20 shows that there are discrete SELF-SIM_0 sets Z such that $(\tilde{V}_n)$ does not have the same distribution as a size-biased random permutation of (V_n), despite the identity in distribution of first terms implied by (13). It would be interesting to know if there were any other discrete SELF-SIM_0 sets besides $\exp(W)$ for homogeneous Poisson W which had this property. If there were, it would presumably be possible to explicitly describe the joint law of the size-biased sequence $(\tilde{V}_n)$, and then derive a sampling formula for the corresponding partition structure, as in [34].

References

[1] C. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.*, 2:1152–1174, 1974.

[2] J. Azéma. Sur les fermés aléatoires. In *Séminaire de Probabilités XIX*, pages 397–495. Springer, 1985. Lecture Notes in Math. 1123.

[3] H. Berbee. On covering single points by randomly ordered intervals. *Ann. Probability*, 9:520 – 528, 1981.

[4] P. J. Burville and J. F. C. Kingman. On a model for storage and search. *J. Appl. Probab.*, 10:697–701, 1973.

[5] P. Carmona, F. Petit, and M. Yor. Some extensions of the arc sine law as partial consequences of the scaling property of Brownian motion. *Probab. Th. Rel. Fields*, 100:1–29, 1994.

[6] P. Donnelly. Partition structures, Pólya urns, the Ewens sampling formula, and the ages of alleles. *Theoretical Population Biology*, 30:271 – 288, 1986.

[7] P. Donnelly. The heaps process, libraries and size biased permutations. *J. Appl. Prob.*, 28:322–335, 1991.

[8] P. Donnelly and P. Joyce. Continuity and weak convergence of ranked and size-biased permutations on the infinite simplex. *Stochastic Processes and their Applications*, 31:89 – 103, 1989.

[9] E. B. Dynkin. Some limit theorems for sums of independent random variables with infinite mathematical expectations. *IMS-AMS Selected Translations in Math. Stat. and Prob.*, 1:171–189, 1961.

[10] S. Engen. *Stochastic Abundance Models with Emphasis on Biological Communities and Species Diversity.* Chapman and Hall Ltd., 1978.

[11] W.J. Ewens. Population genetics theory - the past and the future. In S. Lessard, editor, *Mathematical and Statistical Problems in Evolution.* University of Montreal Press, Montreal, 1988.

[12] D. Freedman. On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.*, 34:1386–1403, 1963.

[13] B. Fristedt. Intersections and limits of regenerative sets. In D. Aldous and R. Pemantle, editors, *Random Discrete Structures*, volume 76 of *Mathematics and its Applications.* Springer-Verlag, 1995.

[14] F. M. Hoppe. Size-biased filtering of Poisson-Dirichlet samples with an application to partition structures in genetics. *Journal of Applied Probability*, 23:1008 – 1012, 1986.

[15] J. Horowitz. Semilinear Markov processes, subordinators and renewal theory. *Z. Wahrsch. Verw. Gebiete*, 24:167 – 193, 1972.

[16] T. Ignatov. On a constant arising in the theory of symmetric groups and on Poisson-Dirichlet measures. *Theory Probab. Appl.*, 27:136–147, 1982.

[17] O. Kallenberg. The local time intensity of an exchangeable interval partition. In A. Gut and L. Holst, editors, *Probability and Statistics, Essays in Honour of Carl-Gustav Esseen*, pages 85–94. Uppsala University, 1983.

[18] Olav Kallenberg. Splitting at backward times in regenerative sets. *Annals of Probability*, 9:781 – 799, 1981.

[19] J. F. C. Kingman. Random discrete distributions. *J. Roy. Statist. Soc. B*, 37:1–22, 1975.

[20] J. F. C. Kingman. The population structure associated with the Ewens sampling formula. *Theor. Popul. Biol.*, 11:274–283, 1977.

[21] J. F. C. Kingman. The representation of partition structures. *J. London Math. Soc.*, 18:374–380, 1978.

[22] F.B. Knight. Characterization of the Lévy measure of inverse local times of gap diffusions. In *Seminar on Stochastic Processes, 1981*, pages 53–78. Birkhäuser, Boston, 1981.

[23] S. Kotani and S. Watanabe. Krein's spectral theory of strings and generalized diffusion processes. In *Functional Analysis in Markov Processes*, pages 235–249. Springer, 1982. Lecture Notes in Math. 923.

[24] J. Lamperti. Semi-stable stochastic processes. *Trans. Amer. Math. Soc.*, 104:62–78, 1962.

[25] J. Lamperti. Semi-stable Markov processes.I. *Z. Wahrsch. Verw. Gebiete*, 22:205–225, 1972.

[26] B. Maisonneuve. Ensembles régénératifs de la droite. *Z. Wahrsch. Verw. Gebiete*, 63:501 – 510, 1983.

[27] J. W. McCloskey. A model for the distribution of individuals by species in an environment. Ph. D. thesis, Michigan State University, 1965.

[28] G. P. Patil and C. Taillie. Diversity as a concept and its implications for random communities. *Bull. Int. Stat. Inst.*, XLVII:497 – 515, 1977.

[29] M. Perman. Order statistics for jumps of normalized subordinators. *Stoch. Proc. Appl.*, 46:267–281, 1993.

[30] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*, 92:21–39, 1992.

[31] J. Pitman. Stationary excursions. In *Séminaire de Probabilités XXI*, pages 289–302. Springer, 1986. Lecture Notes in Math. 1247.

[32] J. Pitman. Partition structures derived from Brownian motion and stable subordinators. Technical Report 346, Dept. Statistics, U.C. Berkeley, 1992. To appear in *Bernoulli*.

[33] J. Pitman. Random discrete distributions invariant under size-biased permutation. Technical Report 344, Dept. Statistics, U.C. Berkeley, 1992. To appear in *Advances in Applied Probability*.

[34] J. Pitman. The two-parameter generalization of Ewens' random partition structure. Technical Report 345, Dept. Statistics, U.C. Berkeley, 1992.

[35] J. Pitman and M. Yor. Arcsine laws and interval partitions derived from a stable subordinator. *Proc. London Math. Soc. (3)*, 65:326–356, 1992.

[36] J. Pitman and M. Yor. On the relative lengths of excursions of some Markov processes. In preparation, 1995.

[37] J. Pitman and M. Yor. Some conditional expectation given an average of a stationary or self-similar random process. Technical Report 438, Dept. Statistics, U.C. Berkeley, 1995. In preparation.

[38] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Technical Report 433, Dept. Statistics, U.C. Berkeley, 1995. To appear in *The Annals of Probability*.

[39] D. Revuz and M. Yor. *Continuous martingales and Brownian motion.* Springer, Berlin-Heidelberg, 1994. 2nd edition.

[40] M. S. Taqqu. A bibliographical guide to self-similar processes and long-range dependence. In *Dependence in Probab. and Stat.: A Survey of Recent Results; Ernst Eberlein, Murad S. Taqqu (Ed.)*, pages 137–162. Birkhäuser (Basel, Boston), 1986.

[41] A. M. Vershik. The asymptotic distribution of factorizations of natural numbers into prime divisors. *Soviet Math. Dokl.*, 34:57–61, 1986.

[42] A.M. Vershik and A.A. Shmidt. Limit measures arising in the theory of groups, I. *Theor. Prob. Appl.*, 22:79–85, 1977.

[43] A.M. Vershik and A.A. Shmidt. Limit measures arising in the theory of symmetric groups, II. *Theor. Prob. Appl.*, 23:36–49, 1978.

[44] J.G. Wendel. Zero-free intervals of semi-stable Markov processes. *Math. Scand.*, 14:21 – 34, 1964.