Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition *

Jim Pitman

Technical Report No. 597

Department of Statistics
University of California
367 Evans Hall # 3860
Berkeley, CA 94720-3860

To appear in *Combinatorics, Probability and Computing*

June 6 2001. Revised January 11, 2002

Abstract

This paper introduces a split-and-merge transformation of interval partitions which combines some features of one model studied by Gnedin and Kerov [12, 11] and another studied by Tsilevich [30, 31] and Mayer-Wolf, Zeitouni and Zerner [21]. The invariance under this split-and-merge transformation of the interval partition generated by a suitable Poisson process yields a simple proof of the recent result of [21] that a Poisson-Dirichlet distribution is invariant for a closely related fragmentation-coagulation process. Uniqueness and convergence

*Research supported in part by N.S.F. Grant DMS-0071448

to the invariant measure are established for the split-and-merge transformation of interval partitions, but the corresponding problems for the fragmentation-coagulation process remain open.

1 Introduction

There has been recent interest in models for the evolution of a random process whose state may be regarded as an *interval partition*, that is a partition of some fixed interval I into collection of disjoint subintervals I_j, whose lengths sum to the length of I. Section 2 introduces a split-and-merge transformation of interval partitions which combines some features of one model studied by Gnedin and Kerov [12, 11] and another studied by Tsilevich [30, 31] and Mayer-Wolf, Zeitouni and Zerner [21]. It is easily seen that the interval partition generated by a suitable Poisson process is invariant for the model of Section 2. As shown in Section 4, this yields a simple proof of the result of [21] that a Poisson-Dirichlet distribution is invariant for a closely related fragmentation-coagulation process. The transfer of ideas from Section 2 to Section 4 relies on some basic results about size-biased random permutations, which are recalled in Section 3. Uniqueness and convergence to the invariant measure are established in Section 2 for the split-and-merge transformation of interval partitions. The corresponding problems for the fragmentation-coagulation process of Section 4 remain open, though for this model Tsilevich [31] has proved convergence to the Poisson-Dirichlet law of binomial means along the trajectory of the process started with only one part. Section 5 points out a connection between the uniqueness problem of Section 4 and Kingman's theory of exchangeable random partitions. Finally, Section 6 presents a discrete analog of the fragmentation-coagulation process of Section 4, in terms of the cycle structure of a random permutation defined by a suitable sequence of random transpositions.

See Durrett and Limic [7] for study of another model which may be regarded as a split-and-merge transformation of an interval partition. Their model has the same invariant Poisson process as the model described in Section 2 for suitably matched parameters, and their model behaves similarly with respect to continuous monotonic transformations. But the equilibrium state of the model decribed in Section 2 is reversible, whereas the equilibrium state of Durrett-Limic model is not. See Aldous-Diaconis [2] and Seppäläinen [29] for study of a more sophisticated model of a similar kind.

2 The split-and-merge transformation

For $W \geq 0$ and indicator variables $1_s, 1_m \in \{0,1\}$, where $1_s = 1$ or 0 according to a decision to split or not, and $1_m = 1$ or 0 according to a decision to merge or not, define a transformation $T_{\mathcal{S}^\uparrow}$ on the space $\mathcal{S}^\uparrow$ of sequences (x_n) with

$$0 \leq x_1 \leq x_2 \leq \cdots \leq x_\infty := \lim_n x_n \leq \infty$$

as follows. Let $T_{\mathcal{S}^\uparrow}((x_n); W, 1_s, 1_m) := (x'_n)$ where

$$(x'_1, x'_2, x'_3, x'_4, \ldots) = \begin{cases} (W, x_1, x_2, x_3, \ldots) & \text{if } W < x_1 \text{ and } 1_s = 1 \\ (x_2, x_3, x_4, x_5 \ldots) & \text{if } x_1 \leq W \text{ and } 1_m = 1 \\ (x_1, x_2, x_3, x_4 \ldots) & \text{otherwise.} \end{cases} \tag{1}$$

Note that in any case $x'_\infty = x_\infty$. Let $x_0 := 0$ and regard the x_n for $n \geq 1$ as the endpoints of successive intervals $I_n := [x_{n-1}, x_n)$ in a partition of $[0, x_\infty)$. Then the x'_n are the endpoints of intervals in the partition (I'_n) derived from (I_n) by the operation of

- splitting I_1 at W if $W \in I_1$ and $1_s = 1$;
- merging I_1 and I_2 if $W \notin I_1$ and $1_m = 1$;
- otherwise leaving the partition unchanged.

Alternatively, the sequence (x_n) may be regarded as encoding the set of points $\{x_n : x_n < x_\infty\}$, which is either finite, or countably infinite. Ties among the x_n would mean multiple points, but this possibility can be ignored with little loss of interest. A split corresponds to appearance or birth of a new point in the set, strictly to the left of the current left-most point, while a merge corresponds to disappearance or death of the left-most point (or of one of them if there is more than one). With this interpretation, and the notation $x \wedge c$ for the minimum of x and c, for $0 \leq c < x_\infty$ the *truncated sequence* $(x_n \wedge c)$ encodes the restriction of the set of points to $[0, c)$, that is the necessarily finite set of points $\{x_n : x_n < c\}$.

Definition 1 Let W be a random variable with continuous probability distribution on $[0, \infty)$, and let β_m and β_s be two probability parameters in $(0, 1]$. The (W, β_m, β_s)*split-and-merge chain* is the Markov chain with state space $\mathcal{S}^\uparrow$ such that given the current state is (x_n), the next state is distributed as

$T_{\mathcal{S}\uparrow}((x_n); W, 1_s, 1_m)$ where 1_s and 1_m are indicator random variables independent of W, with

$$P(1_s = 1) = \beta_s \quad \text{and} \quad P(1_m = 1) = \beta_m.$$

Note that the definition of the Markov chain depends only on the distribution of W, and that repeated steps of the chain can be made using independent copies of $(W, 1_s, 1_m)$ at each step. The continuity assumption on the distribution of W could be relaxed, but at the expense of minor complications in the formulation of the following theorem. This model was introduced by Gnedin and Kerov [12, 11] for $\beta_s = \beta_m = 1$ and $0 \leq W \leq 1$. The extra randomization, with the split probability β_s and the merge probability β_m, was suggested by the model of [21] whose definition is recalled in Section 4. Letting W have values in $[0, \infty)$ instead of $[0, 1]$ allows W to have the *standard exponential distribution* $P(W > t) = e^{-t}$ for $t \geq 0$. This is the simplest choice of W, because the Poisson process appearing in the following theorem is then homogeneous. This theorem is a development of results of [12, 11] suggested by the model of [21] discussed in the next section.

Theorem 2 *Let* $C := \sup\{x : P(W > x) > 0\}$ *be the supremum of the support of the distribution of* W*, and let* $X_1 < X_2 < \cdots$ *be the points of a Poisson point process on* $[0, C]$ *with intensity measure* $(\beta_s/\beta_m)\mu(\cdot)$ *where* $\mu(dx) := P(W \in dx)/P(W > x)$*, or equivalently*

$$\mu[0, x] := -\log P(W > x) \qquad (0 \leq x < C).$$

(i) *For each* $c \in [0, C]$ *the distribution of* $(X_n \wedge c)_{n=1,2,\ldots}$ *is a reversible equilibrium distribution for the* (W, β_m, β_s) *split-and-merge chain.*
(ii) *For each* $c \in [0, C]$ *the distribution of* $(X_n \wedge c)_{n=1,2,\ldots}$ *is the unique equilibrium distribution for the* (W, β_m, β_s) *split-and-merge chain that is concentrated on sequences* (x_n) *with* $x_\infty = c$.
(iii) *Let* $(X_n^{(N)})_{n=1,2,\ldots}$ *denote the state of the* (W, β_s, β_m) *chain after* N *steps, starting from an initial state* (x_n) *with* $x_\infty = c \leq C$*. Then for each* $b < c$ *the law of the infinite sequence* $(X_n^{(N)} \wedge b)_{n=1,2,\ldots}$ *converges in total variation norm to that of* $(X_n \wedge b)_{n=1,2,\ldots}$.
(iv) *If the initial state* (x_n) *has* $x_\infty = C$*, then for each* k *the distribution of the random vector* $(X_n^{(N)})_{1\leq n\leq k}$ *converges in total variation norm to that of* $(X_n)_{1\leq n\leq k}$.

Proof. As observed in [12, 11] in the special case $\beta_s = \beta_m = 1$, it follows immediately from Definition 1 that if $(X_n^{(N)})_{n=1,2,\ldots}$ is the state of a (W, β_s, β_m) chain after N steps, and g an increasing function such that $g(W)$ has a continuous distribution, then $(g(X_n^{(N)}))_{n=1,2,\ldots}$ is the state of a $(g(W), \beta_s, \beta_m)$ chain after N steps. In view of this fact, and the well known transformation rule for Poisson processes, it suffices to prove the stated results for any particular continuous distribution of W. Now take W to be standard exponential, so μ is Lebesgue measure on $[0, \infty)$. For X_1 the first point of a homogeneous Poisson process on $[0, \infty)$ with intensity $\theta > 0$, and X_1 independent of W, introduce the events

$$\texttt{SPLIT} := (W < X_1, 1_s = 1) \text{ and } \texttt{MERGE} := (X_1 < W, 1_m = 1).$$

Since $P(W < X_1) = 1/(1+\theta)$ and the event $(X_1 < W)$ is independent of 1_s and 1_m,

$$P(\texttt{SPLIT}) = \frac{\beta_s}{1+\theta} \quad \text{and} \quad P(\texttt{MERGE}) = \frac{\theta\beta_m}{1+\theta}.$$

Thus

$$\theta = \beta_s/\beta_m \quad \Leftrightarrow \quad P(\texttt{SPLIT}) = P(\texttt{MERGE}) \tag{2}$$

as will now be assumed. To check that the homogeneous Poisson point process on $[0, \infty)$ with rate θ provides a reversible equilibrium, it must be shown that $(X'_n) := T_{\mathcal{S}\uparrow}((X_n), W, 1_s, 1_m)$ is such that

$$((X_n), (X'_n)) \stackrel{d}{=} ((X'_n), (X_n)). \tag{3}$$

The definition (1), the identity (2) and the fact that the $X_n - X_{n-1}$ are i.i.d. reduce (3) to the equality of conditional bivariate distributions

$$(W, X_1 - W) \text{ given } \texttt{SPLIT} \stackrel{d}{=} (X_1, X_2 - X_1) \text{ given } \texttt{MERGE}. \tag{4}$$

By independence of $X_1, W, 1_s$ and 1_m, the conditioning events SPLIT and MERGE in (4) can be replaced by $(W < X_1)$ and $(W \geq X_1)$ respectively. It is elementary and well known that given $(W < X_1)$ the random variables W and $X_1 - W$ are independent exponential variables with rates $1 + \theta$ and θ respectively, and the same applies to X_1 and $X_2 - X_1$ given $W \geq X_1$. This gives part (i) for $c = C$ and the case $c < C$ follows easily. Part (ii) obviously follows from (iii), which is obtained as follows. The key observation, already made in [12, 11] for $\beta_s = \beta_m = 1$, is that for $b < c$ the truncated sequence $(X_n^{(N)} \wedge b)_{n=1,2,\ldots}$ is a Markov chain, whose state when the sequence identically equal to b, meaning $[0,b)$ is empty of points, is a state which is a positive recurrent and aperiodic atom which is reached almost surely from any initial state (x_n) with $x_\infty = c \leq C$. These properties of this special state will be checked again below. It follows that the truncated chain is Harris recurrent [5], hence the convergence in total variation norm by a standard result for Harris chains. In the case $\beta_m = \beta_s = 1$, Gnedin and Kerov [11] established the required properties of the special state by explicit calculation of the distributions and expectations of various hitting times. These calculations could be generalized to incorporate β_s and β_m, but the required recurrence can be seen more quickly as follows. For small enough b the result is obtained by coupling the number of points in $[0,b)$ (starting with some non-zero number, up to the time when the number first reaches zero) below a random walk on the positive integers with negative drift. Then the same result for $2b < c$ is obtained similarly, watching the count of points in $[b,2b)$ only when $[0,b)$ is empty; and so on, for $2^m b < c$ for any $m = 1,2,\ldots$. Part (iv) follows easily from (iii), by first taking b so large that $P(X_k < b) > 1-\varepsilon$, then letting $N \to \infty$ and finally $\varepsilon \to 0$. □

Consider now the particular case of Theorem 2 when $0 \leq W \leq 1$ and the initial sequence (x_n) has $x_\infty = 1$. It is convenient to recode (x_n) by its sequence of differences $p_n := x_n - x_{n-1}$ where $x_0 := 0$. So $(p_n) \in \mathcal{P}$ where

$$\mathcal{P} := \{(p_n) : p_n \geq 0, \Sigma_{n=1}^\infty p_n = 1\}$$

is the space of probability measures on the positive integers. Let $T_\mathcal{P}$ denote the action on $\mathcal{P}$ induced by the transformation $T_{\mathcal{S}\uparrow}$ in (1). That is $T_\mathcal{P}((p_n); W, 1_s, 1_m) := (p'_n)$ where

$$(p'_1, p'_2, p'_3, p'_4, \ldots) := \begin{cases} (W, p_1 - W, p_2, p_3, \ldots) & \text{if } W < p_1 \text{ and } 1_s = 1 \\ (p_1 + p_2, p_3, p_4, p_5 \ldots) & \text{if } p_1 \leq W \text{ and } 1_m = 1 \\ (p_1, p_2, p_3, p_4 \ldots\ldots) & \text{otherwise.} \end{cases}$$

Definition 3 For a random variable W with $0 \leq W \leq 1$ let the (W, β_m, β_s) *split-and-merge chain with state space* $\mathcal{P}$ be defined by the following transition mechanism. Given the current state is $(p_n) \in \mathcal{P}$, the next state is distributed as $T_\mathcal{P}((p_n); W, 1_s, 1_m)$ for Bernoulli variables 1_s and 1_m with parameters β_m and β_s, with W, 1_s and 1_m independent.

The following lemma is elementary and well-known:

Lemma 4 *An increasing sequence of positive random variables $0 < X_1 < X_2 < \cdots$ is the sequence of points of a Poisson process with intensity measure $\theta(1-x)^{-1}dx$ on $(0,1)$ for some $\theta > 0$ if and only if the corresponding differences $P_n := X_n - X_{n-1}$ with $X_0 := 0$ can be represented as*

$$P_n = V_n \prod_{i=1}^{n-1} (1 - V_i) \tag{5}$$

where the V_i are i.i.d. variables with the beta$(1,\theta)$ distribution $P(V_i \in dx) = \theta(1-x)^{\theta-1}dx$, so $V_i \overset{d}{=} U^{1/\theta} - 1$ for U with uniform distribution on $[0,1]$.

The law of (P_n) defined by (5), for independent beta$(1,\theta)$ variables V_i, is commonly known as GEM(θ), after Griffiths, Engen and McCloskey. See [9, 14, 22, 25] for background. Theorem 2 combined with Lemma 4 immediately yields the following result, part (i) of which is due to Gnedin-Kerov [12, 11] for $\beta_m = \beta_s = 1$.

Corollary 5 *For $\eta > 0$ let W_η have the beta$(1,\eta)$ distribution on $[0,1]$. Then*
(i) *The $(W_\eta, \beta_s, \beta_m)$ split-and-merge chain with state space $\mathcal{P}$ has a unique invariant probability distribution, namely GEM(θ) for $\theta = \eta\beta_s/\beta_m$.*
(ii) *The $(W_\eta, \beta_s, \beta_m)$ chain is reversible in its equilibrium state.*
(iii) *Let $(P_n^{(N)})_{n=1,2,\ldots}$ denote the state of the $(W_\eta, \beta_s, \beta_m)$ chain after N steps started with arbitrary initial state in $\mathcal{P}$. Then for each k the distribution of the random vector $(P_n^{(N)})_{1\le n\le k}$ converges in total variation norm to that of $(P_n)_{1\le n\le k}$ for (P_n) with GEM(θ) distribution.*

3 Size-biased permutations

In many applications of random discrete distributions, for instance to combinatorics [13], population genetics [9], species sampling [25], and models for coagulation and fragmentation [1, 6], the main feature of interest is the sizes of atoms of the distribution, rather than their *labels* or *locations* in some ambient space. For this reason, it is common to regard the state of some process of interest as an element of the set *partitions of* 1 or *ranked probability distributions*, say

$$\mathcal{P}^{\downarrow} := \{(p_n) : p_1 \ge p_2 \ge \cdots \ge 0 \text{ and } \Sigma p_n = 1\} \subseteq \mathcal{P}.$$

The analysis of the distribution of random states $(Q_n) \in \mathcal{P}^{\downarrow}$, and of mechanisms for the time evolution of such states, is often simplified by introduction of some other encoding of the state involving additional structure or randomization. In particular, the following construction has found numerous applications [24, 25].

Definition 6 For (Q_n) a random element of $\mathcal{P}$, a *size-biased permutation* (SBP) of (Q_n) is a random sequence $(P_n) \in \mathcal{P}$ whose joint distribution with (Q_n) is that created by the following construction: let (I_j) be some partition of $[0,1]$ into disjoint intervals with the length of I_j equal to Q_j, and independent of (I_j) let $U_1, U_2, \ldots$ be a sequence of independent uniform $[0,1]$ variables. Let $P_1 := Q_{J_1}$ be the length of the interval I_{J_1} say which contains U_1, and for $n \geq 2$ let $P_n := Q_{J_n}$ be the length of the interval I_{J_n} which contains the first U_j not in $\cup_{i=1}^{n-1} I_{J_i}$, with the convention $P_n := 0$ if there is no such interval.

Thus (P_n) describes the sequence of lengths (Q_n) of intervals in a random partition of $[0,1]$, in the length-biased random order in which they are discovered by sampling with independent uniform variables. The operation of size-biased permutation defines a regular conditional distribution for (P_n) given (Q_n), say

$$P((P_n) \in \cdot \mid (Q_n)) = \mathtt{SBP}((Q_n), \cdot)$$

where $\mathtt{SBP}$ is a Markov transition kernel from $\mathcal{P}$ to $\mathcal{P}$. The distribution $\mathtt{SBP}((Q_n), \cdot)$ on $\mathcal{P}$ depends only the decreasing rearrangement of (Q_n), denoted $\mathtt{RANK}(Q_n) \in \mathcal{P}^{\downarrow}$. Some elementary and well known properties of the Markov kernel $\mathtt{SBP}$ on $\mathcal{P}$ are mentioned now for ease of reference in later discussion.

- The kernel $\mathtt{SBP}$ is idempotent : $(\mathtt{SBP})^2 = \mathtt{SBP}$.
- The formula $\mu\, \mathtt{SBP} = \nu$ sets up a bijection between probability distributions μ on $\mathcal{P}^{\downarrow}$ and probability distributions ν on $\mathcal{P}$ which are *invariant under size-biased permutation*, meaning $\nu\, \mathtt{SBP} = \nu$.

The family of distributions on $\mathcal{P}$ that are invariant under $\mathtt{SBP}$ has been characterized in a number of ways [24].

Following Kingman [16], the *Poisson-Dirichlet distribution with parameter* $\theta > 0$, denoted PD(θ), is defined as the distribution on $\mathcal{P}^{\downarrow}$ of $(Q_n) :=$

(Y_n/Σ) where $Y_1 > Y_2 > \cdots$ are the points of a Poisson process on $(0,\infty)$ with intensity measure $\theta y^{-1}e^{-y}dy$, and $\Sigma := \sum_n Y_n$. Two well known features of this construction are that Σ has gamma(θ) distribution, and that Σ is independent of (Q_n). Each of these features is known to characterize the special form of the intensity measure of the Poisson process [22]. It is well known that PD(θ) is characterized by the following relation with the GEM(θ) distribution (5) on $\mathcal{P}$, involving the ranking map from $\mathcal{P}$ to $\mathcal{P}^{\downarrow}$ and its inversion by the size-biased permutation kernel SBP:

$$\begin{array}{c} PD(\theta) \\ \mathtt{SBP}\downarrow \quad \uparrow \mathtt{RANK} \\ GEM(\theta) \end{array} \tag{6}$$

To spell out the meaning of the diagram: if (P_n) has GEM(θ) distribution and $(Q_n) := \mathtt{RANK}(P_n)$ then (Q_n) has PD(θ) distribution and (P_n) is a size-biased permutation of (Q_n). Hence of course, if (Q_n) has PD(θ) distribution and (P_n) is a size-biased permutation of (Q_n), then (P_n) has GEM(θ) distribution. Note the consequence, which is not obvious from the definition (5), that GEM(θ) is invariant under SBP. See [25] for further discussion and references to earlier sources. See also [27] regarding the larger two-parameter Poisson-Dirichlet family of distributions on $\mathcal{P}^{\downarrow}$, which shares some but not all of the remarkable properties of PD(θ).

Suppose now that a GEM(θ) distributed sequence (P_n) has been constructed in accordance with (6) and Definition 6 by size-biased sampling of some interval partition whose ranked lengths have PD(θ) distribution. For positive integers $n_1,\ldots,n_k$ with $\sum_{i=1}^k n_i = n$, let $(n_1,\ldots,n_k)$ denote the event that in the sampling process with independent uniform variables U_j, for $1 \le j \le n$ there are exactly n_i values U_j in the ith interval discovered by the sampling process, whose length is P_i. As a consequence of Ferguson's well known updating rule for sampling from a Dirichlet prior [10],

$$\left(P_1,\ldots,P_k, 1-\Sigma_{i-1}^k P_i\right) \text{ given } (n_1,\ldots,n_k) \stackrel{d}{=} \text{Dirichlet } (n_1,\ldots,n_k,\theta) \tag{7}$$

meaning that the joint density of $(P_1,\ldots,P_k)$ at $(x_1,\ldots,x_k)$ given $(n_1,\ldots,n_k)$ is proportional to

$$\left(\prod_{i=1}^k x_i^{n_i-1}\right)\left(1-\sum_{i=1}^k x_i\right)^{\theta-1} \quad \text{for } x_i \ge 0 \text{ and } \sum_{i=1}^k x_i < 1. \tag{8}$$

The probability of the event $(n_1,\ldots n_k)$ is given by a variant of the Ewens sampling formula given in [4, (7.2)].

4 A fragmentation-coagulation process

Following Mayer-Wolf, Zeitouni and Zerner [21], for a random variable W with values in $(0,1)$ and $\beta_s, \beta_m \in (0,1]$ as before, consider the (W, β_s, β_m) *fragmentation-coagulation chain with state-space* $\mathcal{P}^{\downarrow}$ defined as follows. Given that the current state is $(p_n) \in \mathcal{P}^{\downarrow}$, let J_1, J_2, 1_s, 1_m and W be independent random variables, with J_1 and J_2 distributed according to (p_n), with 1_s and 1_m Bernoulli variables with parameters β_s and β_m

- if $J_1 = J_2 = j$ say and $1_s = 1$, let the new state be obtained by replacing p_j by Wp_j and $(1-W)p_j$ and re-ranking;
- if $J_1 = i$ and $J_2 = j$ say with $i \neq j$, and $1_m = 1$, let the new state be obtained by replacing the two atoms p_i and p_j by a single atom of size $p_i + p_j$, and re-ranking;
- else no change in state.

Suppose now that $W = U$ say has uniform distribution on $[0,1]$. It is easily seen that for an arbitrary random initial state $(Q_n) \in \mathcal{P}^{\downarrow}$, the state (Q'_n) of the (U, β_s, β_m) fragmentation-coagulation chain after one step from state (Q_n) can be constructed as $(Q'_n) = \mathtt{RANK}(P'_n)$, where (P'_n) is the state after one step of the (U, β_s, β_m) split-and-merge chain with state space $\mathcal{P}$, as constructed in Section 2, with initial state (P_n) which is a size-biased permutation of (Q_n).

Thus the transition mechanisms of these two processes are related according to the following diagram:

$$\begin{array}{ccc} (Q_n) & \overset{(U,\beta_s,\beta_m)\text{ frag-coag}}{\longrightarrow} & (Q'_n) \\ \mathtt{SBP}\downarrow \quad \uparrow \mathtt{RANK} & & \uparrow \mathtt{RANK} \\ (P_n) & \overset{(U,\beta_s,\beta_m)\text{ split-merge}}{\longrightarrow} & (P'_n) \end{array} \tag{9}$$

It is obvious from the diagram how a reversible invariant measure for the lower transition mechanism transfers to give a reversible invariant measure for the upper transition mechanism. Thus Corollary 5 and (6) imply the following result:

Theorem 7 [30, 21] *For U with uniform distribution on $(0,1)$, the PD(θ) distribution for $\theta = \beta_s/\beta_m$ gives a reversible equilibrium distribution for the (U, β_s, β_m) fragmentation-coagulation chain with state space $\mathcal{P}^{\downarrow}$.*

This result may be compared to the characterization of PD(θ) as the unique stationary distribution of the infinitely many alleles diffusion model [8], and the characterization of PD(1) in terms of virtual permutations provided in [30].

To discuss the issue of uniqueness of the invariant measure for the (U, β_s, β_m) fragmentation-coagulation chain with state space $\mathcal{P}^{\downarrow}$, suppose as above that (Q_n) is an arbitrary random initial state $(Q_n) \in \mathcal{P}^{\downarrow}$, and (P_n) is a size-biased random permutation of (Q_n). As remarked in [24], by exchangeability of the sampling process in the construction of (P_n) from an interval partition whose ranked lengths are given by (Q_n), if $(1, \ldots, 1)_n$ denotes the event that each of the n independent uniform variables falls in a different interval of the partition, then

$$\textit{the law of } (P_1, \ldots, P_n) \textit{ restricted to the event } (1, \ldots, 1)_n \textit{ is exchangeable.} \tag{10}$$

In particular, if (Q_n) has PD(θ) distribution, then from (7) the law in (10) has density at $(x_1, \ldots, x_n)$ proportional to

$$\left(1 - \sum_{i=1}^{n} x_i\right)^{\theta - 1} \quad \text{on } \mathcal{S}_n := \{(x_1, \ldots, x_n) : x_i \geq 0 \text{ and } \Sigma_{i=1}^n x_i < 1\}.$$

Let $\mathcal{A}_n$ be the set of functions on $\mathcal{S}_n$ which are the restrictions to $\mathcal{S}_n$ of some function which is real analytic in a neighbourhood of $\mathcal{S}_n$. It was shown in [21] that PD(θ) for $\theta = \beta_s/\beta_m$ is the only equilibrium distribution for the (U, β_s, β_m) fragmentation-coagulation chain such that

$$\textit{for each } n \textit{ the restricted law (10) admits a density in } \mathcal{A}_n. \tag{11}$$

It is known [24] that whenever (P_n) is the size-biased permutation of some (Q_n), the restricted law in (10) has density $\prod_{i=1}^{n-1}(1 - \sum_{j=1}^{i} x_j)$ relative to the unconditional law of $(P_1, \ldots, P_n)$ at $(x_1, \ldots, x_n)$. Therefore, the side condition (11) can be reformulated as

$$\textit{the law of } (P_1, \ldots, P_n) \textit{ admits a density in } \mathcal{A}_n. \tag{12}$$

The uniqueness result of Corollary 5 and diagram (9) show that PD(θ) for $\theta = \beta_s/\beta_m$ is the only equilibrium distribution of (Q_n) for the (U, β_s, β_m) fragmentation-coagulation chain on $\mathcal{P}^{\downarrow}$ such that

$$(P'_n) \textit{ is invariant under size-biased permutation} \tag{13}$$

where (P'_n) is the state after one step of the (U, β_s, β_m) split-and-merge chain with state space $\mathcal{P}$, started with initial state which is a size-biased permutation (P_n) of (Q_n). It might be that (13) must hold for any equilibrium distribution (Q_n) of the (U, β_s, β_m) fragmentation-coagulation chain on $\mathcal{P}^{\downarrow}$. But this does not seem to be at all obvious. Note that (13) may fail if the assumption that (Q_n) is an equilibrium is dropped. For instance, take $(Q_n) = (1, 0, 0, \ldots)$ and $\beta_s = \beta_m = 1$. Then $(P_n) = (Q_n)$, and P'_1 and $P'_2 = 1 - P'_1$ are both uniform on $(0, 1)$, so obviously not in size-biased order.

The previous discussion suggests the following construction. Define a new Markov transition kernel on $\mathcal{P}$, say (U, β_s, β_m) split-merge-`SBP`, by composition of the (U, β_s, β_m) split-merge and the `SBP` kernels on $\mathcal{P}$. In terms of interval partitions, this means first performing the (U, β_s, β_m) split-merge operation on the first two intervals, then rearranging all the intervals according to the order in which they are discovered by a process of uniform random sampling of $[0, 1]$. If (P''_n) denotes the output of this process, starting from some arbitrary initial $(Q_n) \in \mathcal{P}^{\downarrow}$ as before, then the previous diagram (9) implies

$$\begin{array}{ccc} (Q_n) & \overset{(U,\beta_s,\beta_m)\ \text{frag-coag}}{\longrightarrow} & (Q'_n) \\ \texttt{SBP}\downarrow\ \ \uparrow\texttt{RANK} & & \texttt{SBP}\downarrow\ \ \uparrow\texttt{RANK} \\ (P_n) & \overset{(U,\beta_s,\beta_m)\ \text{split-merge-}\texttt{SBP}}{\longrightarrow} & (P''_n). \end{array} \tag{14}$$

Thus

$$(Q_n) \textit{ is an equilibrium for the } (U, \beta_s, \beta_m) \textit{ fragmentation-coagulation process} \tag{15}$$

if and only if

$$(P_n) \textit{ is an equilibrium for the } (U, \beta_s, \beta_m) \textit{ split-merge-}\texttt{SBP} \textit{ process.} \tag{16}$$

By application of the criterion of [28] for a function of a Markov chain to be Markov, using the `SBP` kernel to invert the function `RANK`, the diagram (14) implies that if the $\mathcal{P}$-valued (U, β_s, β_m) split-merge-SBP chain is started in state (P_n) which is a SBP of (Q_n), then the sequence in $\mathcal{P}^{\downarrow}$ obtained by ranking the state of this chain at each step is the $\mathcal{P}^{\downarrow}$-valued (U, β_s, β_m) fragmentation-coagulation chain. Unfortunately, however, the (U, β_s, β_m) split-merge-SBP chain does not seem to be nearly as easy to analyze as the (U, β_s, β_m) split-merge chain.

5 Exchangeable random partitions

Kingman's theory of *exchangeable random partitions* [19] provides one approach to the analysis of Markov kernels on $\mathcal{P}^{\downarrow}$. See for instance [26]. For an arbitrary random element (Q_n) of $\mathcal{P}^{\downarrow}$, consider the corresponding *exchangeable partition probability function* (EPPF) [23] which is the symmetric function of sequences of positive integers $(n_1, \ldots, n_k)$ of arbitrary length k defined by

$$P(n_1, \ldots, n_k) := E\left[\sum_{(i_1,\ldots,i_k)} \prod_{j=1}^{k} Q_{i_j}^{n_j}\right] \tag{17}$$

where the sum is over all permutations of k positive integers $(i_1, \ldots, i_k)$. If (I_j) is a random interval partition such that the length of I_j is Q_j, then $P(n_1, \ldots, n_k)$ can be interpreted as follows. If $U_1, U_2, \ldots$ are independent uniform $(0,1)$ variables, independent also of (Q_j), then for each particular partition of the first n integers into sets $\{A_1, \ldots, A_k\}$ of sizes $n_1, \ldots, n_k$, with the A_i in order of their least elements, $P(n_1, \ldots, n_k)$ is the probability that for each $1 \leq i \leq k$, each of the U_j for $j \in A_i$ falls in the interval of length P_i which is the ith interval discovered by the sampling process, as discussed earlier around (7). See also [23, (5) and (21)]. The work of Kingman [18] shows that the EPPF and the law of (Q_n) on $\mathcal{P}^{\downarrow}$ determine each other uniquely. If (P_n) is any random element of $\mathcal{P}$ such that $\mathtt{RANK}(P_n) = (Q_n)$, then formula (17) holds just as well with (Q_n) replaced by (P_n), by an obvious symmetry argument. If (P_n) is a size-biased permutation of (Q_n), there is the alternative formula [23]

$$P(n_1, \ldots, n_k) := E\left[\left(\prod_{i=1}^{k} P_i^{n_i - 1}\right) \prod_{i=1}^{k-1} \left(1 - \sum_{j=1}^{i} P_j\right)\right]. \tag{18}$$

which sets up a bijective correpondence between all distributions of (P_n) that are invariant under size-biased permutation, and the EPPF associated with the corresponding random element $\mathtt{RANK}(P_n)$ of $\mathcal{P}^{\downarrow}$.

According to a variant of the Ewens sampling formula [25, (30) and (36)] the EPPF corresponding to PD(θ) is given by the formula

$$P_\theta(n_1,\ldots,n_k) := \frac{\theta^k}{[\theta]_n}\prod_{i=1}^{k}(n_i-1)! \tag{19}$$

where $[\theta]_n := \theta(\theta+1)\cdots(\theta+n-1)$. Thus a natural approach to uniqueness problem considered in the previous section is to try to use the definition of the (U,β_s,β_m) fragmentation-coagulation process to characterize the EPPF of any equilibrium for this process. Apart from issues of notation and coding of partitions, this is essentially the same approach suggested in Section 6 of [21]. With present notation, and $\theta := \beta_s/\beta_m$, the equations discussed in Section 6 of [21] can be recast as follows. Let $P'(\cdots)$ denote the EPPF of the state of the (U,β_s,β_m) fragmentation-coagulation chain after one step started in a state with EPPF $P(\cdots)$. Then the equilibrium condition $P(n)=P'(n)$ holds if and only if

$$\theta\frac{(n-1)}{(n+1)}P(n+2) = \sum_{k=1}^{n-1}\binom{n}{k}P(k+1,n-k+1). \tag{20}$$

For instance

$$\theta P(4) = 6P(2,2);\quad \theta P(5) = 12P(3,2);\quad \theta P(6) = \frac{3}{5}(8P(2,4)+6P(3,3))$$

and so on. Note that with notation as in diagram (14), condition (20) for all n is necessary and sufficient for $P_1 \stackrel{d}{=} P_1''$, or again for $E\sum_n f(Q_n) = E\sum_n f(Q_n')$ for arbitrary non-negative measurable f. Similarly, one can check that the condition $P(n,m)=P'(n,m)$ holds iff

$$\theta\left[-\frac{2n!m!}{(n+m+1)!}P(2+n+m)+\frac{(n-1)}{(n+1)}P(2+n,m)+\frac{(m-1)}{(m+1)}P(2+m,n)\right]$$

$$= -2P(n{+}1,m{+}1)+\sum_{a=1}^{n-1}\binom{n}{a}P(a{+}1,n{-}a{+}1,m)+\sum_{a=1}^{m-1}\binom{m}{a}P(a{+}1,m{-}a{+}1,n)$$

and this for all n and m is equivalent to $(P_1,P_2) \stackrel{d}{=} (P_1'',P_2'')$.

As remarked in [21], it is possible to check the PD(θ) equilibrium by use of these equations and the fact that a distribution on $\mathcal{P}^{\downarrow}$ is determined by the values of $P(n_1, \ldots, n_k)$ for (n_i) with $n_i \geq 2$ for all i. But it seems difficult to establish uniqueness with this system of equations, as they are ill-determined at every finite stage k.

6 Random transpositions

This section makes a connection between the fragmentation-coagulation model of Section 4 and a simple model for the evolution of a sequence of random permutations. See also Tsilevich [30, 31] for closely related studies.

Let S_n be the symmetric group of permutations of $[n] := \{1, \ldots, n\}$. For $x, y \in [n]$ let $\sigma_{x,y}$ be the transposition of x and y. It is elementary and well known that if π has k cycles then $\pi\sigma_{x,y}$ for $x \neq y$ has either $k-1$ or $k+1$ cycles according to whether x and y fall in the same or different cycles of π.

For $\beta_s, \beta_m \in (0,1]$ consider Markov chain with state-space S_n defined as follows. Given that the current state is $\pi \in S_n$, let $X, Y, 1_s$ and 1_m be independent random variables, with X and Y uniformly distributed on $[n]$, and 1_s and 1_m Bernoulli variables with parameters β_s and β_m respectively; if either

- X and Y fall in the same cycle of π and $1_s = 1$, or
- X and Y fall in different cycles of π and $1_m = 1$,

then let the next state be the product $\pi\sigma_{X,Y}$; else let the next state be π.

This Markov chain describes a process of random transpositions whereby at each step, either a cycle splits in two, or two cycles merge, or the permutation remains unchanged, with probabilities determined in a simple way by the cycle structure of the permutation and the split and merge probabilities β_s and β_m. For $\beta_s = \beta_m = 1$ this is the process of random transpositions studied by Diaconis-Shashahani [3] and Matthews [20]. In this simplest case, the stationary distribution is obviously the uniform distribution on S_n.

Proposition 8 *The process of random transpositions with state space S_n defined above for $\beta_s, \beta_m \in (0,1]$ has a unique equilibrium probability distribution, namely*

$$P_\theta(\pi) = \theta^{|\pi|}/[\theta]_n \qquad (\pi \in S_n)$$

where $\theta = \beta_s/\beta_m$, where $|\pi|$ is the number of cycles of π, and

$$[\theta]_n := \theta(\theta + 1)\cdots(\theta + n - 1).$$

Moreover, the process in equilibrium is reversible.

Proof. The fact that $P_\theta(\cdot)$ is a probability distribution on S_n is well known, and it is obvious that all states communicate, so it suffices to check the usual condition for a reversible equilibrium, that is

$$P_\theta(\pi)P(\pi, \pi') = P_\theta(\pi')P(\pi', \pi) \tag{21}$$

where $P(\cdot, \cdot)$ is the transition matrix on S_n determined by (β_m, β_s). In view of how the cycles of $\pi' = \pi\sigma_{x,y}$ are related to those of π, both sides of (21) are 0 unless $|\pi'| = |\pi| \pm 1$, so it is enough to consider π and π' with $|\pi| = k$ and $|\pi'| = k + 1$ for some $1 \leq k < n$. But then (21) reduces to

$$\theta^k \frac{2}{n^2}\beta_s = \theta^{k+1}\frac{2}{n^2}\beta_m$$

which holds if and only if $\theta = \beta_s/\beta_m$. □

For a permutation $\pi \in S_n$ let $\lambda(\pi)$ denote the partition of n defined by the sizes of the cycles of π. So $\lambda(\pi) \in \mathcal{P}_n^\downarrow$, the set of non-increasing sequences of non-negative integers with sum n. It is easily seen that if $(\Pi_N, N = 1, 2, \ldots)$ is the Markov chain with state space S_n just described, then $(\lambda(\Pi_N), N = 1, 2, \ldots)$ is a a Markov chain with state space $\mathcal{P}_n^\downarrow$, with the following transition mechanism. Given that the current state is $\lambda := (\lambda_i) \in \mathcal{P}_n^\downarrow$, where the λ_i are positive integers with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and $\sum_i \lambda_i = n$, let $\bar{\lambda} \in \mathcal{P}^\downarrow$ be the probability distribution defined by $\bar{\lambda}_i = \lambda_i/n$, and let $J_1, J_2, 1_s$ and 1_m be independent random variables, with J_1 and J_2 distributed on the positive integers according to on $\bar{\lambda}$, and 1_s and 1_m Bernoulli variables with parameters β_s and β_m respectively;

- if $J_1 = J_2 = j$ say and $1_s = 1$, let the new state be obtained by splitting the part λ_j into U_j and $\lambda_j - U_j$ and re-ranking, where U_j is uniformly distributed on $\{0, 1, \ldots, \lambda_j - 1\}$;
- if $J_1 = i$ and $J_2 = j$ say with $i \neq j$, and $1_m = 1$, let the new state be obtained by replacing the two parts λ_i and λ_j by a single part of size $\lambda_i + \lambda_j$, and re-ranking;

- else no change in state.

Proposition 8 implies that this chain with state space $\mathcal{P}_n^{\downarrow}$ has a reversible equilibrium which is the distribution of the partition of n induced by the cycles of a random permutation with distribution P_θ on S_n, as in (21). It is elementary and well known that this distribution on $\mathcal{P}_n^{\downarrow}$ is given by the Ewens sampling formula with parameter θ. That is to say, for $\lambda = (\lambda_i)$ such that the number of i with $\lambda_i = j$ is m_j for each $1 \leq j \leq n$, where $m_j \geq 0$ and $\sum_j jm_j = n$, the equilibrium probability of λ is

$$P_\theta\{\pi \in S_n : \lambda(\pi) = \lambda\} = \frac{n!}{[\theta]_n} \prod_{j=1}^{n} \frac{1}{m_j!} \left(\frac{\theta}{j}\right)^{m_j}$$

where $\theta = \beta_s/\beta_m$. This identification of the equilibrium distribution of the $\mathcal{P}_n^{\downarrow}$-valued process with parameters β_m and β_s is a particular case of a result of Whittle [32] for a more general model of $\mathcal{P}_n^{\downarrow}$-valued processes of coagulation and fragmentation. See also Kelly [15, Theorem 8.1] and [6]. Now regard $\mathcal{P}_n^{\downarrow}$ as a subset of $\mathcal{P}^{\downarrow}$ by use of the normalization $\lambda \to \bar{\lambda}$. As $n \to \infty$ the transition mechanism of $\mathcal{P}_n^{\downarrow}$-valued process with parameters (β_s, β_m) approaches that of the (U, β_s, β_m) fragmentation-coagulation chain described in Section 4, while the equilibrium distribution of the $\mathcal{P}_n^{\downarrow}$-valued process, determined by the Ewens sampling formula (21), approaches PD(θ) by a result of Kingman [17]. Thus Theorem 7 could be deduced by a weak-convergence argument, but this approach does not seem to help solve the uniqueness problem.

Acknowledgment

Thanks to the authors of [21] for showing me their work prior to publication, and for making helpful comments on a draft of this paper.

References

[1] D.J. Aldous. Deterministic and stochastic models for coalescence (aggregation and coagulation): a review of the mean-field theory for probabilists. *Bernoulli*, 5:3–48, 1999.

[2] D.J. Aldous and P. Diaconis. Hammersley's interacting particle process and longest increasing subsequences. *Probab. Th. Rel. Fields*, 103:199–213, 1995.

[3] P. Diaconis and M. Shahshahani. Generating a random permutation with random transpositions. *Z. Wahrsch. Verw. Gebiete*, 57(2):159–179, 1981.

[4] P. Donnelly and S. Tavaré. The ages of alleles and a coalescent. *Adv. Appl. Probab.*, 18:1–19 & 1023, 1986.

[5] R. Durrett. *Probability: theory and examples.* Duxbury Press, Belmont, CA, second edition, 1996.

[6] R. Durrett, B. L. Granovsky, and S. Gueron. The equilibrium behavior of reversible coagulation-fragmentation processes. *J. Theoret. Probab.*, 12(2):447–474, 1999.

[7] R. Durrett and V. Limic. A surprising Poisson process arising from a species competition model. Preprint, Dept. Mathematics, Cornell University, 2001.

[8] S. N. Ethier and T.G. Kurtz. The infinitely-many-neutral-alleles diffusion model. *Adv. in Appl. Probab.*, 13:429 – 452, 1981.

[9] W. J. Ewens. Population genetics theory—the past and the future. In S. Lessard, editor, *Mathematical and statistical developments of evolutionary theory (Montreal, PQ, 1987)*, pages 177–227. Kluwer Acad. Publ., Dordrecht, 1990.

[10] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1:209–230, 1973.

[11] A. Gnedin and S. Kerov. Fibonacci solitaire. Preprint, to appear in *Random Structures and Algorithms*, 2000.

[12] A. Gnedin and S. Kerov. A characterization of GEM distributions. *Combin. Probab. Comput.*, 10(3):213–217, 2001.

[13] J.C. Hansen. Order statistics for decomposable combinatorial structures. *Rand. Struct. Alg.*, 5:517–533, 1994.

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

[15] F.P. Kelly. *Reversibility and Stochastic Networks.* Wiley, 1979.

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

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

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

[19] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[20] P. Matthews. A strong uniform time for random transpositions. *J. Theoret. Probab.*, 1(4):411–423, 1988.

[21] E. Mayer-Wolf, O. Zeitouni, and M. Zerner. Asymptotics of certain coagulation-fragmentation processes and invariant Poisson-Dirichlet measures. Preprint. Technion, Haifa, Israel. Available via `http://tiger.technion.ac.il/~zeitouni/ps/cofra.ps`, 2001.

[22] M. Perman, J. Pitman, and M. Yor. Size-biased sampling of Poisson point processes and excursions. *Probab. Th. Rel. Fields*, 92:21–39, 1992.

[23] J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Th. Rel. Fields*, 102:145–158, 1995.

[24] J. Pitman. Random discrete distributions invariant under size-biased permutation. *Adv. Appl. Prob.*, 28:525–539, 1996.

[25] J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. In T.S. Ferguson et al., editor, *Statistics, Probability and Game Theory; Papers in honor of David Blackwell*, volume 30 of *Lecture Notes-Monograph Series*, pages 245–267. Institute of Mathematical Statistics, Hayward, California, 1996.

[26] J. Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27:1870–1902, 1999.

[27] J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.*, 25:855–900, 1997.

[28] L. C. G. Rogers and J. Pitman. Markov functions. *Annals of Probability*, 9:573–582, 1981.

[29] T. Seppäläinen. Hydrodynamic scaling, convex duality and asymptotic shapes of growth models. *Markov Process. Related Fields*, 4:1–26, 1998.

[30] N. V. Tsilevich. Stationary random partitions of a natural series. *Teor. Veroyatnost. i Primenen.*, 44(1):55–73, 1999.

[31] N. V. Tsilevich. On the simplest split-merge operator on the infinite-dimensional simplex. PDMI Preprint 03, 2001.

[32] P. Whittle. Statistical processes of aggregation and polymerisation. *Proc. Camb. Phil. Soc.*, 61:475–495, 1965.