A reversible allelic partition process and Pitman sampling formula

A reversible allelic partition process and Pitman sampling formula

Matteo Giordano*, Pierpaolo De Blasi†,⋆ and Matteo Ruggiero†,⋆

*University of Cambridge**, *University of Torino*† *and Collegio Carlo Alberto*⋆

March 17, 2020

Abstract

We introduce a continuous-time Markov chain describing dynamic allelic partitions which extends the branching process construction of the Pitman sampling formula in Pitman (2006) and the classical birth-and-death process with immigration studied in Karlin and McGregor (1967), in turn related to the celebrated Ewens sampling formula. A biological basis for the scheme is provided in terms of a population of individuals grouped into families, that evolves according to a sequence of births, deaths and immigrations. We investigate the asymptotic behaviour of the chain and show that, as opposed to the birth-and-death process with immigration, this construction maintains in the temporal limit the mutual dependence among the multiplicities. When the death rate exceeds the birth rate the system is shown to have a reversible distribution, identified as a mixture of Pitman sampling formulae, with negative binomial mixing distribution on the population size. The population therefore converges to a stationary random configuration, characterised by a finite number of families and individuals.

Keywords: birth-and-death process, branching process, immigration, stationary process, Pólya urn, population dynamics.

MSC Primary: 60G10, 60J10. Secondary: 92D25, 60J80.

1 Allelic partitions and sampling formulae

A partition of $n \in \mathbb{N}$ is an unordered collection $\pi = \{n_1, \dots, n_k\}$ of $k \leq n$ positive integers whose sum equals n. A common equivalent way to describe π is by means of the so-called *allelic partition*, which groups the partition sets by size. We denote by $\mathbf{m} = (m_1, \dots, m_n) \in \mathbb{Z}_+^n$ the associated vector of multiplicities, where m_i counts

the number of repetitions of i in π, and let

$$\mathbb{A}_n = \left\{ \mathbf{m} \in \mathbb{Z}_+^n,\ \sum_{i=1}^{n} i m_i = n \right\}$$

be the finite set of all allelic partitions of n. It will be useful to embed $\mathbb{A}_n$ into $\mathbb{Z}_+^\infty$ by considering infinite vectors $\mathbf{m} = (m_1, \ldots, m_n, 0, 0, \ldots)$, defining also

$$s(\mathbf{m}) := \sum_{i \geq 1} i m_i, \quad k(\mathbf{m}) := \sum_{i \geq 1} m_i,$$

which are respectively the number of items and groups (or positive multiplicities) in $\mathbf{m}$, and

$$\mathbb{A} := \bigcup_{n \geq 1} \mathbb{A}_n = \{ \mathbf{m} \in \mathbb{Z}_+^\infty,\ s(\mathbf{m}) < \infty \},$$

the countable set of all allelic partitions. A random allelic partition $\mathbf{M} = (M_i)_{i \geq 1}$ can then be defined as a random variable taking values in $\mathbb{A}$.

The literature on the distributional properties of random allelic partition is rich and well established. Ewens (1972) derived the distribution of the random allelic partition induced by a sample of n genes drawn from a selectively neutral population at equilibrium (see also Karlin and McGregor, 1972). This is described by the celebrated *Ewens sampling formula* (ESF), which assigns probability

$$\mathrm{ESF}_n^\theta(\mathbf{m}) = \frac{n!}{\theta_{(n)}} \prod_{i \geq 1} \left(\frac{\theta}{i} \right)^{m_i} \frac{1}{m_i!} \mathbb{1}_{\{s(\mathbf{m}) = n\}}, \tag{1}$$

to the configuration $\mathbf{m} \in \mathbb{A}$ with m_i alleles appearing exactly i times for each $i \geq 1$. Here $\theta > 0$ and $\theta_{(n)} = \theta(\theta+1)\ldots(\theta+n-1)$ is the ascending factorial. The impact of the ESF has been significant in a number of fields beyond population genetics. For example, Antoniak (1974) derived it independently from a sample from a Dirichlet process, a cornerstone nonparametric prior distribution in Bayesian nonparametric statistics; Hoppe (1984) recovered the formula as the marginal distribution of a Markov chain generated by a Pólya-like urn model comprising an infinite number of colors, analogous to that introduced in Blackwell and MacQueen (1973) and described equivalently in Aldous (1985) with the famous metaphor of the *Chinese restaurant process*. See Johnson et al. (1997) and Crane (2016) for reviews and Feng (2010) for connections with population genetics.

Pitman (1995) introduced a two-parameter generalization of (1), often referred to as the *Pitman sampling formula* (PSF), whereby for parameters $0 \leq \alpha < 1$ and $\theta > -\alpha$, the probability assigned to a random allelic partition $\mathbf{m} \in \mathbb{A}$ is

$$\mathrm{PSF}_n^{\alpha,\theta}(\mathbf{m}) = \frac{n!}{\theta_{(n)}} \prod_{i=1}^{k(\mathbf{m})} (\theta + i\alpha) \prod_{j \geq 1} \left[\frac{(1-\alpha)^{j-1}}{j!} \right]^{m_j} \frac{1}{m_j!} \mathbb{1}_{\{s(\mathbf{m}) = n\}}. \tag{2}$$

For $\alpha = 0$, equation (2) immediately recovers (1), so that $\mathrm{PSF}_n^{0,\theta} \equiv \mathrm{ESF}_n^{\theta}$. For $0 < \alpha < 1$, the distribution arises for example in the study of stable processes with index α, the case $\alpha = 1/2$ being related to the zeros of Brownian motion (see also Pitman, 1997), and in the study of the partition structures induced by a sample from a two-parameter Poisson–Dirichlet distribution (Perman et al., 1992; Perman, 1993; Pitman and Yor, 1997). The corresponding generalization of Hoppe's urn is defined in Pitman (1995, 1996*b*) as a Markov chain of allelic partitions with initial state $\mathbf{e}_0$ and transition probabilities

$$p(\mathbf{m}'|\mathbf{m}) \propto \begin{cases} \theta + \alpha k(\mathbf{m}), & \mathbf{m}' = \mathbf{m} + \mathbf{e}_1, \\ (i - \alpha) m_i, & \mathbf{m}' = \mathbf{m} - \mathbf{e}_i + \mathbf{e}_{i+1}, \ i \geq 1, \\ 0, & \text{else}, \end{cases} \tag{3}$$

where $\mathbf{e}_i = (\delta_{ij})_{j \geq 1}$, $i \geq 0$, and the normalising constant is given by $\theta + s(\mathbf{m})$. The first transition in (3) can be understood as the addition to the gene sample of a new gene of a previously unobserved allele, while the transition from $\mathbf{m}$ to $\mathbf{m} - \mathbf{e}_i + \mathbf{e}_{i+1}$ corresponds to adding a gene whose allele had been previously observed i times. For $\alpha = 0$, (3) coincides with Hoppe's urn, while for any $0 \leq \alpha < 1$ the marginal distribution of the state of the chain after n transitions is given by (2).

Among other relevant contributions to the related theory, the asymptotic behavior of the number $K_n^{\alpha,\theta}$ of different alleles was derived by Korwar and Hollander (1973) in the case $\alpha = 0$, for which

$$\lim_{n \to \infty} \frac{K_n^{0,\theta}}{\log n} = \theta \text{ a.s.}, \tag{4}$$

and by Pitman (1997) for $0 < \alpha < 1$, where we have that

$$\lim_{n \to \infty} \frac{K_n^{\alpha,\theta}}{\log n} = S^{\alpha,\theta} \text{ a.s.}, \tag{5}$$

$S^{\alpha,\theta}$ being an absolutely continuous random variable on $(0, \infty)$ whose law is related to the Mittag–Leffler distribution. Concerning the PSF, we further refer the reader to Kerov (2006); Pitman (1996*a*) for various characterizations thereof, James et al. (2008) for connections with Bayesian nonparametric statistics and again Feng (2010) for connections with population genetics. Note that similar asymptotic results to the above are available for Gibbs-type models (Gnedin and Pitman, 2006; De Blasi et al., 2015), which include the one- and two-parameter models recalled above, as well as a model in Gnedin (2010) which produces a finite but random number of families in the limit.

In this paper we are interested in the connection between Ewens–Pitman sampling formulae and the theory of stochastic processes for population growth. In

particular, we construct a birth-and-death process with immigration whose dynamics modifies the one arising as after (3) by allowing the removal of items from the system; and we show, in a particular regime, that the process is reversible with respect to a certain mixture of Pitman sampling formulae. The rest of the paper is organised as follows. In Section 2 we present our contribution and its connections with past related work. Section 3 describes in detail the birth-and-death population model with immigration, and how it can be constructed via branching processes. Finally, Section 4 analyses the reversible regime, identifying explicitly the reversible distribution and its connections with Pitman sampling formulae.

2 Main contribution and related work

The seminal work of Karlin and McGregor (1967) describes a population wherein new families are initiated at a sequence of random times generated by a stochastic process $I = \{I(t), t \geq 0\}$, which can be thought of as immigration events, and then evolve, independently of one another, according to the law of a common continuous-time Markov chain on $\mathbb{Z}_+$, whose infinitesimal rates are denoted q_{ij}. As pointed out by Tavaré (1989), if new families are interpreted as novel mutant alleles of a given gene, then the scheme may be regarded as a version of the infinitely-many-neutral-alleles model and can provide, under specific choices of its probabilistic components, a generating mechanism for the sampling formulae (1) and (2). In particular, let I be a pure-birth processes started at 0 with birth rates

$$\lim_{h\to 0} \frac{1}{h}\mathbb{P}(I(t+h) - I(t) = 1 | I(t) = k) = \theta + \alpha k, \quad k \geq 0,$$

and assume that the process describing the evolution of each family is started at 0 and has rates

$$q_{i,i+1} = i - \alpha, \ i \geq 1. \tag{6}$$

If now $M_i(t)$ is the number of families of size i at time t, and $\mathbf{M}(t) = (M_i(t))_{i\geq 1}$ is the induced allelic partition, then $\mathbf{M} = \{\mathbf{M}(t), t \geq 0\}$ defines a continuous-time Markov chain on $\mathbb{A}$ whose embedded jump chain has transition probabilities (3) and marginal distributions described by the PSF; see Feng and Hoppe (1998) and Pitman (2006). When $\alpha = 0$, this construction reduces to the birth process with immigration considered in Tavaré (1987), which provides the corresponding embedding of Hoppe's urn leading to the ESF (cf. Section 2.7.2 in Feng, 2010).

In the above schemes, each transition entails the addition of one individual to the population, and both the population size and the number of families diverge almost surely as time increases. This is a convenient mathematical simplification but may be undesirable or result in a lack of flexibility in different contexts, and

is arguably unrealistic when modelling the evolution of populations. Therefore, it can be of interest to account for the death of individuals and to study the related implications on the family structure.

Motivated by the above considerations, we consider here a modified version of the population model underlying the PSF, by assuming that the individuals are endowed with independent random exponential lifetimes of parameter $\mu > 0$, with $\mu = 0$ recovering the original construction, thereby introducing uniform death events in the scheme. See Definition 1 below for the details. By leaving unchanged the other rules governing the evolution of the family structure, we thus obtain more flexible dynamics whereby families are started, fluctuate in size and possibly become extinct with the passage of time, resulting in a population with varying size and number of families. Note that, because of exchangeability, the PSF is consistent with respect to uniform deletion, hence a death event induces a random partition still following the PSF. For consistency with respect to deletion of an entire family, a property known in the literature as regeneration, see Gnedin and Pitman (2005).

For $\alpha = 0$, our construction recovers the detailed version of the birth-and-death process with immigration (BDI) in Karlin and McGregor (1967), object of thorough study in Kendall (1975) and Tavaré (1989). In this case, the new families are started at the times of a Poisson point process of intensity θ, and each family subsequently evolves according to an independent linear birth-and-death process. Theorem 2.1 in Karlin and McGregor (1967) derives the marginal distribution of the induced allelic partition $\mathbf{M}(t)$, which is that of a sequence of independent Poisson random variables given by

$$\mathbf{M}(t) \sim \prod_{i \geq 1} \mathrm{Po}(\theta b_t^i / i), \;\; b_t = \begin{cases} [e^{(1-\mu)t} - 1]/[e^{(1-\mu)t} - \mu], & \mu \neq 1 \\ t/(1+t), & \mu = 1. \end{cases} \tag{7}$$

Anticipated by earlier results in Watterson (1974) and Kendall (1975), Tavaré (1989) reformulated the above distribution as a mixture of Ewens sampling formulae in (1) with a negative binomial mixing distribution on n. Specifically, letting $S(t)$ being the number of alive individuals at time t, the population size process $S = \{S(t), t \geq 0\}$ defines by construction a BDI process, with negative binomial marginal distribution

$$S(t) \sim \text{N-Bin}(\theta, b_t), \;\; \mathbb{P}(S(t) = n) = \frac{\theta_{(n)}}{n!}(1 - b_t)^{\theta} b_t^n, \;\; n = 0, 1, \ldots \tag{8}$$

Then, by combining (7) and (8), the ESF is recovered as the conditional distribution of the actual allelic partition given the population size, i.e.,

$$\mathbb{P}(\mathbf{M}(t) = \mathbf{m} | S(t) = n) = \mathrm{ESF}_n^{\theta}(\mathbf{m}), \;\; \mathbf{m} \in \mathbb{A}, \;\; n \geq 1. \tag{9}$$

Within the study of the BDI process, a great attention has been dedicated to investigating the long-run behavior of the model. By taking the limit as $t \to \infty$ in

(7), we have the convergence in law

$$\mathbf{M}(t) \xrightarrow{d} (X_1, X_2, \dots), \ t \to \infty, \tag{10}$$

where, for $0 < \mu \leq 1$, $X_i \overset{\text{ind}}{\sim} \text{Po}(\theta/i)$, while for $\mu > 1$, $X_i \overset{\text{ind}}{\sim} \text{Po}(\theta\mu^{-i}/i)$. In the first case the death rate does not exceed the birth rate, and the asymptotic regime corresponds to the weak limit as $n \to \infty$ of the Ewens partition structure (1) derived in Arratia et al. (1992), characterized by an infinite sample size and a logarithmic growth of the underlying number of groups. Instead, for $\mu > 1$, the death events occur at a faster rate than births, causing each family to eventually become extinct almost surely (see Kendall, 1975) and preventing the indefinite growth observed previously. In fact, in the second regime the population size process S is easily seen in Lemma 2 below to have N-Bin(θ, μ^{-1}) reversible distribution. In turn, Kendall (1975) showed the reversibility of the induced allelic partition process, and shed light on the representation of the reversible distribution as an analogous mixture as that arising from (9), characterized by a N-Bin(θ, μ^{-1}) mixing on the population size.

In this paper, we encode the description of the evolving family structure directly in terms of the induced allelic partition process, and then focus on investigating the long-run behavior of the model. Because of the more involved rules governing the formation of new families, the sharp distributional result derived in Karlin and McGregor (1967) is not accessible in our construction. Nonetheless, by observing that the overall dynamics of the total population size is left unchanged, we show that for $\mu > 1$ the model is reversible, and identify the reversible distribution as a mixture of Pitman sampling formulae with respect to the same negative binomial mixing on the population size appearing for $\alpha = 0$. For $0 < \alpha < 1$, the mixture can be written in closed form as

$$\pi(\mathbf{m}) = C(\theta/\alpha)_{(k(\mathbf{m}))} \prod_{i \geq 1} \text{Po}(m_i; \alpha_i \mu^{-i}), \ \alpha_i = \frac{\alpha(1-\alpha)_{(i-1)}}{i!}, \ \mathbf{m} \in \mathbb{A}, \tag{11}$$

where

$$C = e^{1-(1-1/\mu)^\alpha}(1 - 1/\mu)^\theta. \tag{12}$$

Thus, in the reversible regime, the system will converge to a stationary random configuration that features almost surely finitely many families and individuals. However, as opposed to the previous model, the mutual dependence among the multiplicities is seen in (11) to be preserved in the limit.

3 A birth-and-death process with immigration

3.1 Definition

Definition 1. *For $0 \leq \alpha < 1$, $\theta > -\alpha$ and $\mu > 0$, let $\mathbf{M} = \{\mathbf{M}(t),\ t \geq 0\}$, $\mathbf{M}(t) = (M_i(t))_{i\geq 1}$, be a continuous-time Markov chain with state space $\mathbb{A}$, initial state $\mathbf{M}(0) = \mathbf{e}_0$ and temporally homogeneous transition rates*

$$q(\mathbf{m}'|\mathbf{m}) = \begin{cases} \theta + \alpha k(\mathbf{m}), & \mathbf{m}' = \mathbf{m} + \mathbf{e}_1, \\ (i-\alpha)m_i, & \mathbf{m}' = \mathbf{m} - \mathbf{e}_i + \mathbf{e}_{i+1},\ i \geq 1, \\ \mu i m_i, & \mathbf{m}' = \mathbf{m} - \mathbf{e}_i + \mathbf{e}_{i-1},\ i \geq 1, \\ -\theta - (1+\mu)s(\mathbf{m}), & \mathbf{m}' = \mathbf{m}, \\ 0, & \text{else.} \end{cases} \tag{13}$$

We notice that the first two lines in (13) coincide precisely with the transitions in the sequential construction by Pitman (1995) displayed in (3). These describe the initiation of a new family of size one and the addition of one individual to a pre-existing family of size i respectively. The third transition instead represents, for $i \geq 2$, the death of a member of a family of size i, while for $i = 1$ it corresponds to the extinction of a family comprising a single individual. The choice of the initial state $\mathbf{M}(0) = (0, 0, \dots)$ is rather common across the literature, and can be interpreted as the population being initially empty, as in a territory about to be colonised.

The infinitesimal generator corresponding to the transition rates in (13) are easily seen to be *stable*, as $q(\mathbf{m}|\mathbf{m})$ is finite for all $\mathbf{m} \in \mathbb{A}$, and *conservative*, since each row has zero sum (cf. Norris, 1997). Furthermore, by adapting the argument in Section 4 in Kendall (1975), we can deduce from the properties of the associated population size process that $\mathbf{M}$ is also *regular*, i.e. it performs almost surely only a finite number of jumps in every finite time interval. In particular, for any $t \geq 0$, let

$$S(t) := s(\mathbf{M}(t)) = \sum_{i\geq 1} i M_i(t) \tag{14}$$

be the random number of items underlying the allelic partition $\mathbf{M}(t)$. Then we have from Definition 1 that $S = \{S(t), t \geq 0\}$ defines a continuous-time Markov chain on $\mathbb{Z}_+$, with initial state $S(0) = 0$ and transition rates

$$r(n'|n) = \begin{cases} \theta + n, & n' = n + 1, \\ \mu n, & n' = n - 1, \\ -\theta - (1+\mu)n, & n' = n, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Indeed, by properly aggregating the rates in (13), we see that conditionally given the population size $S(t) = n$ (that is to say $\mathbf{M}(t) = \mathbf{m}$ for any $\mathbf{m} \in \mathbb{A}$ with $s(\mathbf{m}) = n$)

one of the following events occur. A new individual is introduced in the population (either initiating a new family or joining a pre-existing one) increasing its size by one unit, and this happens with overall rate

$$\theta + \alpha k(\mathbf{m}) + \sum_{i \geq 1} (i - \alpha) m_i = \theta + n.$$

Alternatively, the death of one of the n individuals occurs, decreasing the population size by one unit, which happens at the total rate

$$\sum_{i \geq 1} \mu i m_i = \mu n.$$

Thus, irrespective of the parameter α, S defines a classical BDI process, with unit birth rate, death rate μ and immigration rate θ. The law of S is seen in (15) to be insensitive to the choice of α, so that we conclude that the more refined rules governing the formation of new families do not impact the dynamics of the total population size, but rather influence the detailed probabilistic fluctuations in the family configuration.

Finally, we notice that $\mathbf{M}$ and S share the same random jump time, as each transition of the allelic partition process involves either the introduction or the death of an individual. Therefore, as the generator of the birth-and-death process with immigration is known to be regular (see, e.g., Section 2 in Kendall (1975)), we can directly conclude about the regularity of $\mathbf{M}$, which accordingly identifies (jointly with the initial state) a well-defined chain on $\mathbb{A}$.

3.2 Branching process construction

A biological basis for the three-parameter process in Definition 1 can be provided in terms of a population of individuals grouped into families evolving according to a sequence of births, deaths and immigrations, by introducing uniform death events in the models studied by Feng and Hoppe (1998) and Pitman (2006).

First assume that $\theta > 0$, and consider a population whose members have independent exponential lifetimes with parameter μ, and autonomously reproduce according to independent unit-rate Poisson processes (taken without loss of generality, upon appropriate time rescaling). At a birth, any offspring produced by the oldest member of each family will either join its parent's family, with probability $1-\alpha$, or start a new family, with probability α. All other family members give birth to individuals that will remain within their parent's family. Finally, an independent Poisson process of rate θ is assumed to bring immigrants into the population that, at the time of entry, will initiate a new family each.

For $-\alpha < \theta \leq 0$, the above rules can be modified by assuming that no immigrant can enter the population, but rather that the overall oldest member in the

population produces at rate $1-\alpha$ offspring that will join its family, and at rate $\alpha+\theta$ individuals that will initiate new families.

Thus, by exploiting basic properties of independent Poisson processes (e.g., a *Poissonization* argument as in Athreya and Karlin, 1968), it follows that, if we encode the family structure into the corresponding sequence of multiplicities, the dynamics of the above population generates a continuous-time Markov chain of allelic partitions, whose transition rates recover precisely those of the process $\mathbf{M}$ in (13). In particular, each family evolves according to an independent birth-and-death process with initial state 1 and transition rates

$$q_{i,i+1} = i-\alpha, \quad q_{i,i-1} = \mu i, \quad i \geq 1$$

directly generalizing (6). On the other hand, the introduction of deaths changes substantially the probabilistic laws that govern the input process $I = \{I(t),\ t \geq 0\}$ (counting the number of new families formed up to time t). In particular, as opposed to Karlin and McGregor (1967), here I does not define a Markov process, since the probability of introducing a new family at each transition depends on the actual number of families $k(\mathbf{M}(t))$ which, due to families possibly becoming extinct, is in general different from $I(t)$. This in turn represents a serious obstruction in deriving distributional results for quantities of interest related to the model, in particular for the total number of families. Still, as resulting from (15), the dynamics of the total population size is tractable, and this fact will be exploited in the next section to study a reversibility regime for the process.

4 The reversible regime

As seen above, the population size process S defines a classical BDI process. Hence, from the transition rates in (15) we immediately deduce that S is irreducible, as each pair of states are mutually accessible. Moreover, the process is known to be *transient* for $0<\mu<1$, *null recurrent* for $\mu=1$ and *positive recurrent* for $\mu>1$ (cf. Section 2 in Kendall, 1975). In the latter case, S is *ergodic*, implying the convergence in law

$$\lim_{t\to\infty} \mathbb{P}(S(t)=n|S(0)=0) = \lambda(n),\ \forall n \in \mathbb{Z}_+$$

for a unique *stationary distribution* $\lambda = (\lambda(n))_{n\geq 0}$ on $\mathbb{Z}_+$ satisfying

$$\sum_{n\in\mathbb{Z}_+} \lambda(n) r(n'|n) = 0,\ \forall n' \in \mathbb{Z}_+.$$

The above sequence of equations can be solved recursively to obtain

$$\lambda(n) = \frac{\theta_{(n)}}{n!}\mu^{-n}(1-1/\mu)^{\theta},\ n \geq 0, \tag{16}$$

so that λ results to be a negative binomial distribution N-bin(θ, μ^{-1}) with parameters θ and μ^{-1}. The following lemma shows that, under the condition $\mu > 1$, the population size process in fact enjoys the stronger property of being reversible with respect to λ.

Lemma 2. *Let S be the population size process associated to $\mathbf{M}$ in Definition 1 through (14). Then, if $\mu > 1$, S is reversible with respect to its stationary distribution λ in (16).*

Proof. As S is regular and irreducible, the reversibility of the process is equivalent to the detailed balance condition

$$\lambda(n)r(n'|n) = \lambda(n')r(n|n'), \ n \neq n'. \tag{17}$$

In view of (15), we only need to show (17) for the cases $n \geq 0$, $n' = n + 1$, and $n \geq 1$, $n' = n - 1$, as the transition rates corresponding to all other transitions are equal to 0. For $n \geq 0$, $n' = n + 1$, the left hand side of (17) becomes

$$\begin{aligned}
\lambda(n)r(n+1|n) &= \frac{\theta_{(n)}}{n!}\mu^{-n}(1-1/\mu)^{\theta}(\theta+n) \\
&= \frac{\theta_{(n+1)}}{(n+1)!}\mu^{-(n+1)}(1-1/\mu)^{\theta}\mu(n+1) \\
&= \lambda(n+1)r(n|n+1);
\end{aligned}$$

and similarly, for $n \geq 1$, $n' = n - 1$ we have

$$\begin{aligned}
\lambda(n)r(n-1|n) &= \frac{\theta_{(n)}}{n!}\mu^{-n}(1-1/\mu)^{\theta}\mu n \\
&= \frac{\theta_{(n-1)}}{(n-1)!}\mu^{-(n-1)}(1-1/\mu)^{\theta}(\theta+n) \\
&= \lambda(n-1)r(n|n-1).
\end{aligned}$$

□

We now move on to study the properties of $\mathbf{M}$. We first notice that $\mathbf{M}$ is *irreducible*, as we see from (13) that the state space $\mathbb{A}$ constitutes a single communication class, i.e., for any $\mathbf{m}, \mathbf{m}' \in \mathbb{A}$ we can find a finite path $\mathbf{m}^{(1)}, \ldots, \mathbf{m}^{(l)} \in \mathbb{A}$ for which

$$q(\mathbf{m}^{(1)}|\mathbf{m})q(\mathbf{m}^{(2)}|\mathbf{m}^{(1)}) \ldots q(\mathbf{m}'|\mathbf{m}^{(l)}) > 0.$$

Thus, we can argue as in Section 5 in Kendall (1975) to draw some preliminary conclusions concerning the influence of the death rate on the dynamics of the model,

based upon the properties of the sample size process S. In particular, in view of the one-to-one correspondence

$$S(t) = \sum_{i \geq 1} iM_i(t) = 0 \iff \mathbf{M}(t) = \mathbf{e}_0 = (0, 0, \dots), \ t \geq 0,$$

we deduce that the initial condition $\mathbf{e}_0$ and, by irreducibility, the entire state space $\mathbb{A}$ itself, are characterized by the same recurrence and transience properties of the state 0 for S. Accordingly, we conclude that $\mathbf{M}$ is transient for $0 < \mu < 1$, null recurrent for $\mu = 1$ and ergodic for $\mu > 1$.

We shall henceforth focus on the latter case, wherein the underlying population almost surely undergoes a sequence of extinctions, separated by intervals of times of finite expected length during which its size and family configuration fluctuate according to the description in Section 3. Such dynamics were shown in Lemma 2 to generate a reversible population size process. The following theorem, which states our main result, shows that the same property holds for $\mathbf{M}$.

Theorem 3. *Let* $\mathbf{M}$ *be as in Definition 1, with* $0 < \alpha < 1$ *and* $\mu > 1$. *Then,* $\mathbf{M}$ *is reversible with respect to the distribution* π *with weights given by (11).*

Proof. That π is a well-defined probability distribution on $\mathbb{A}$, as well as the explicit expression of the constant C in (12), will follow from the mixture representation below. We then proceed to show the detailed balance condition

$$\pi(\mathbf{m}')q(\mathbf{m}|\mathbf{m}') = \pi(\mathbf{m})q(\mathbf{m}'|\mathbf{m}), \ \mathbf{m} \neq \mathbf{m}'. \tag{18}$$

which, since $\mathbf{M}$ is regular and irreducible, is equivalent to the reversibility.

In view of the expression for the transition rates (13) of $\mathbf{M}$, we only need to show (18) for the cases in which, for given $\mathbf{m} \in \mathbb{A}$, we have $\mathbf{m}' = \mathbf{m}+\mathbf{e}_1$, $\mathbf{m}' = \mathbf{m}-\mathbf{e}_i+\mathbf{e}_{i+1}$ for $i \geq 1$, or finally $\mathbf{m}' = \mathbf{m} + \mathbf{e}_i - \mathbf{e}_{i+1}$ for some $i \geq 0$. The rate associated to the other transitions are indeed null.

For $\mathbf{m} \in \mathbb{A}$ and $\mathbf{m}' = \mathbf{m}+\mathbf{e}_1$, denoting $k = k(\mathbf{m})$, $s = s(\mathbf{m})$, so that $k(\mathbf{m}+\mathbf{e}_1) = k+1$ and $s(\mathbf{m}+\mathbf{e}_1) = s+1$, the left hand side of (18) becomes

$$\begin{aligned}
&\pi(\mathbf{m}+\mathbf{e}_1)\, q(\mathbf{m}|\mathbf{m}+\mathbf{e}_1) \\
&\quad = C\left(\frac{\theta}{\alpha}\right)_{(k+1)} \mathrm{Po}(m_1+1; \alpha_1\mu^{-1}) \prod_{i \geq 2} \mathrm{Po}(m_i; \alpha_i\mu^{-i})\mu(m_1+1) \\
&\quad = C\left(\frac{\theta}{\alpha}\right)_{(k)} \left(\frac{\theta}{\alpha}+k\right) \mathrm{Po}(m_1; \alpha_1\mu^{-1}) \frac{\alpha_1\mu^{-1}}{m_1+1} \prod_{i \geq 2} \mathrm{Po}(m_i; \alpha_i\mu^{-i})\mu(m_1+1) \\
&\quad = C\left(\frac{\theta}{\alpha}\right)_{(k)} \prod_{i \geq 1} \mathrm{Po}(m_i; \alpha_i\mu^{-i})(\theta+\alpha k) \\
&\quad = \pi(\mathbf{m})q(\mathbf{m}+\mathbf{e}_1|\mathbf{m}).
\end{aligned}$$

Here we have used the identity

$$\mathrm{Po}(m_1+1;\alpha_1\mu^{-1}) = e^{-\alpha_1\mu^{-1}}\frac{(\alpha_1\mu^{-1})^{m_1+1}}{(m_1+1)!} = \mathrm{Po}(m_1;\alpha_1\mu^{-1})\frac{\alpha_1\mu^{-1}}{m_1+1}$$

and the fact that

$$\alpha_i = \frac{\alpha(1-\alpha)_{(i-1)}}{i!},\ i \geq 1$$

with the convention $(1-\alpha)_{(0)} = 1$, yielding $\alpha_1 = \alpha$.

Instead, for $\mathbf{m}' = \mathbf{m} - \mathbf{e}_i + \mathbf{e}_{i+1}$, denoting as before $k = k(\mathbf{m})$ and $s = s(\mathbf{m})$, we have $k(\mathbf{m}-\mathbf{e}_i+\mathbf{e}_{i+1}) = k$ and $s(\mathbf{m}-\mathbf{e}_i+\mathbf{e}_{i+1}) = s+1$. Thus, proceeding similarly

$$\begin{aligned}
&\pi(\mathbf{m}-\mathbf{e}_i+\mathbf{e}_{i+1})q(\mathbf{m}|\mathbf{m}-\mathbf{e}_i+\mathbf{e}_{i+1}) \\
&= C\left(\frac{\theta}{\alpha}\right)_{(k)}\prod_{i\geq 1}\mathrm{Po}(m_i;\alpha_i\mu^{-i})\frac{\alpha_{i+1}\mu^{-(i+1)}}{m_{i+1}+1}\frac{m_i}{\alpha_i\mu^{-i}}\mu(i+1)(m_{i+1}+1) \\
&= \pi(\mathbf{m})\frac{\alpha_{i+1}}{\alpha_i}(i+1)m_i = \pi(\mathbf{m})(i-\alpha)m_i \\
&= \pi(\mathbf{m})q(\mathbf{m}-\mathbf{e}_i+\mathbf{e}_{i+1}|\mathbf{m}),
\end{aligned}$$

where we have used

$$\frac{\alpha_{i+1}}{\alpha_i} = \frac{i-\alpha}{i+1}.$$

Finally, consider the case $\mathbf{m}' = \mathbf{m}+\mathbf{e}_i-\mathbf{e}_{i+1}$. Using the same notation as above, we have $k(\mathbf{m}+\mathbf{e}_i-\mathbf{e}_{i+1}) = k$ and $s(\mathbf{m}+\mathbf{e}_i-\mathbf{e}_{i+1}) = s-1$. A similar computation yields

$$\begin{aligned}
&\pi(\mathbf{m}+\mathbf{e}_i-\mathbf{e}_{i+1})q(\mathbf{m}|\mathbf{m}+\mathbf{e}_i-\mathbf{e}_{i+1}) \\
&= C\left(\frac{\theta}{\alpha}\right)_{(k)}\prod_{i\geq 1}\mathrm{Po}(m_i;\alpha_i\mu^{-i})\frac{m_{i+1}}{\alpha_{i+1}\mu^{-(i+1)}}\frac{\alpha_i\mu^{-i}}{m_i+1}(i-\alpha)(m_i+1) \\
&= \pi(\mathbf{m})\frac{\alpha_i}{\alpha_{i+1}}(i-\alpha)\mu m_{i+1} \\
&= \pi(\mathbf{m})\mu m_{i+1}(i+1) \\
&= \pi(\mathbf{m})q(\mathbf{m}+\mathbf{e}_i-\mathbf{e}_{i+1}|\mathbf{m}).
\end{aligned}$$

concluding the proof. □

The reversible distribution π in of Theorem 3 can be represented as a mixture of Pitman sampling formulae in (2) with respect to the reversible N-Bin(θ,μ^{-1}) distribution of the population size process S. Specifically, we have

$$\pi(\mathbf{m}) = \sum_{n\geq 0}\mathrm{PSF}_n^{\alpha,\theta}(\mathbf{m})\text{N-Bin}(n;\theta,\mu^{-1}),\ \mathbf{m}\in\mathbb{A}. \tag{19}$$

Indeed, for $\alpha > 0$, the PSF (2) can be rewritten as

$$\mathrm{PSF}_n^{\alpha,\theta}(\mathbf{m}) = \frac{n!(\theta/\alpha)_{(k(\mathbf{m}))}}{\theta_{(n)}} \prod_{i\geq 1} \frac{\alpha_i^{m_i}}{m_i!} \mathbb{1}_{\{s(\mathbf{m})=n\}}, \ \mathbf{m} \in \mathbb{A}$$

where α_i is as in (11). Then, for $\mathbf{m} \in \mathbb{A}$, it follows that

$$\begin{aligned}
&\sum_{n\geq 0} \mathrm{PSF}_n^{\alpha,\theta}(\mathbf{m}) \text{N-Bin}(n;\theta,\mu^{-1}) \\
&= \sum_{n\geq 0} \frac{n!(\theta/\alpha)_{(k(\mathbf{m}))}}{\theta_{(n)}} \prod_{i\geq 1} \frac{\alpha_i^{m_i}}{m_i!} \mathbb{1}_{\{s(\mathbf{m})=n\}} \frac{\theta_{(n)}}{n!} \mu^{-n} (1-1/\mu)^{\theta} \\
&= (\theta/\alpha)_{(k(\mathbf{m}))} \prod_{i\geq 1} \frac{\alpha_i^{m_i}}{m_i!} \mu^{-s(\mathbf{m})} (1-1/\mu)^{\theta} \\
&= (1-1/\mu)^{\theta} (\theta/\alpha)_{(k(\mathbf{m}))} \prod_{i\geq 1} \frac{\alpha_i^{m_i}}{m_i!} \mu^{-\sum_{i\geq 1} i m_i} \\
&= (1-1/\mu)^{\theta} (\theta/\alpha)_{(k(\mathbf{m}))} \prod_{i\geq 1} \frac{(\alpha_i \mu^{-i})^{m_i}}{m_i!} e^{-\alpha_i \mu^{-i}} e^{\alpha_i \mu^{-i}} \\
&= (1-1/\mu)^{\theta} e^{\sum_{i\geq 1} \alpha_i \mu^{-i}} (\theta/\alpha)_{(k(\mathbf{m}))} \prod_{i\geq 1} \mathrm{Po}(m_i; \alpha_i \mu^{-i}).
\end{aligned}$$

The series appearing as the argument of the exponential can be simplified by noticing that

$$\begin{aligned}
1-(1-1/\mu)^{\alpha} &= 1 - \sum_{i\geq 0} \binom{\alpha}{i} (-\mu^{-i}) \\
&= 1 - 1 - \sum_{i\geq 1} (-1)^n \frac{\alpha(\alpha-1)\dots(\alpha-i+1)}{i!} \mu^{-i} \\
&= -\sum_{i\geq 1} -\frac{-\alpha(1-\alpha)\dots(1-\alpha+i-2)}{i!} \mu^{-i} \\
&= \sum_{i\geq 1} \frac{\alpha(1-\alpha)_{(i-1)}}{i!} \mu^{-i} = \sum_{i\geq 1} \alpha_i \mu^{-i}.
\end{aligned}$$

Replaced into the last line of the previous display, this recovers the expression for the reversible distribution π with the constant C as in (12), implying (19).

Theorem 3 yields interesting insights in the long-run behavior of the three-parameter process: in particular, in view of the ergodicity, $\mathbf{M}(t)$ will converge in distribution to a stationary random allelic partition whose distribution admits the mixture representation (19), and is characterized by a random, but almost surely

finite number of items and groups. Thus, the reversible regime of $\mathbf{M}$ reveals clear analogies with the results in Kendall (1975) for the BDI process, essentially due to the equivalence of the induced population size processes. However, as opposed to the case $\alpha = 0$, $\mathbf{M}$ in the limit to maintains the mutual dependence among the multiplicities, when $\alpha > 0$.

Concerning the range $0 < \mu \leq 1$, we deduce at once from the absence of positive recurrence that the process is necessarily non-stationary, and in fact, because of the slower death rate, the associated population size process S is seen in (8) to diverge as $t \to \infty$. On the other hand, for $\theta > 0$, we can deduce from the description of the population model in Section 3, that the number of families $K(t)$ alive at time t stochastically dominates the number of families in the corresponding model with $\alpha = 0$, as it can be rewritten as

$$K(t) = K_1(t) + K_2(t)$$

where $K_1(t)$ counts the family founded by immigrants, and $K_2(t)$ is the number of those started by the newborns leaving their family. Now $K_1(t)$ is evidently equal in distribution to the number of families comprised in a population governed by the BDI process, which, in turn, exhibits the logarithmic growth (4). Accordingly, we conclude that $\mathbf{M}$ will evolve towards infinite structures characterized by an infinite number of individuals and families. Whether a precise description of the limiting behavior similar to (10) can be achieved within this formulation, remains for the moment an open question.

Acknowledgements

The authors are grateful to the Associate Editor and two referees for helpful comments. The first author is supported by the ERC grant No. 647812 (UQMSI) and partially by the EPSRC grant EP/L016516/1 for the Cambridge Centre for Analysis. The second and third authors are partially supported by the Italian Ministry of Education, University and Research (MIUR), through PRIN 2015SNS29B and through "Dipartimenti di Eccellenza" grant 2018-2022.

References

Aldous, D. J. (1985), *Exchangeability and related topics*, number 1117 *in* 'Lecture Notes in Mathematics', Springer, Berlin, Heidelberg.

Antoniak, C. E. (1974), 'Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems', *Annals of Statistics* **2**(6), 1152–1174.

Arratia, R., Barbour, A. D. and Tavare, S. (1992), 'Poisson process approximations for the Ewens sampling formula', *Annals of Applied Probability* **3**(3), 519–535.

Athreya, K. B. and Karlin, S. (1968), 'Embedding of urn schemes into continuous time Markov branching processes and related limit theorems', *The Annals of Mathematical Statistics* **39**(6), 1801–1817.

Blackwell, D. and MacQueen, J. B. (1973), 'Ferguson distributions via Pólya urn schemes', *Annals of Statistics* **1**(2), 353–355.

Crane, H. (2016), 'The ubiquitous Ewens sampling formula', *Statistical Science* **31**(1), 1–19.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. and Ruggiero, M. (2015), 'Are Gibbs-type priors the most natural generalization of the Dirichlet process?', *IEEE transactions on pattern analysis and machine intelligence* **37**(2), 212–229.

Ewens, W. J. (1972), 'The sampling theory of selectively neutral alleles', *Theoretical Population Biology* **3**(1), 87–112.

Feng, S. (2010), *The Poisson-Dirichlet Distribution and Related Topics*, Probability and Its Applications, 1 edn, Springer-Verlag Berlin Heidelberg.

Feng, S. and Hoppe, F. M. (1998), 'Large deviation principles for some random combinatorial structures in population genetics and Brownian motion', *Annals of Applied Probability* **8**(4), 975–994.

Gnedin, A. (2010), 'A species sampling model with finitely many types', *Electron. Commun. Probab.* **15**, 79–88.

Gnedin, A. and Pitman, J. (2005), 'Regenerative partition structures', *Electronic Journal of Combinatorics* **11**(2), 1–21.

Gnedin, A. and Pitman, J. (2006), 'Exchangeable Gibbs partitions and Stirling triangles', *Journal of Mathematical sciences* **138**(3), 5674–5685.

Hoppe, F. M. (1984), 'Pólya-like urns and the Ewens' sampling formula', *Journal of Mathematical Biology* **20**(1), 91–94.

James, L. F., Lijoi, A. and Prünster, I. (2008), 'Distributions of linear functionals of two parameter poisson–dirichlet random measures', *Annals of Applied Probability* **18**(2), 521–551.

Johnson, N., Kotz, S. and Balakrishnan, N. (1997), *Multivariate discrete distributions*, Wiley, New York.

Karlin, S. and McGregor, J. (1967), 'The number of mutant forms maintained in a population', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* pp. 415–438.

Karlin, S. and McGregor, J. (1972), 'Addendum to a paper of W. Ewens', *Theoretical Population Biology* **3**(1), 113–116.

Kendall, D. G. (1975), 'Some problems in mathematical genealogy', *Journal of Applied Probability* **12**(1), 325–345.

Kerov, S. V. (2006), 'Coherent random allocations, and the Ewens-Pitman formula', *Journal of Mathematical Sciences* **138**(3), 5699–5710.

Korwar, R. M. and Hollander, M. (1973), 'Contributions to the theory of Dirichlet processes', *Annals of Probability* **1**(4), 705–711.

Norris, J. R. (1997), *Markov Chains*, number 2 *in* 'Cambridge Series in Statistical and Probabilistic Mathematics', Cambridge University Press.

Perman, M. (1993), 'The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator', *Stochastic Processes and their Applications* **46**(2), 267–281.

Perman, M., Pitman, J. and Yor, M. (1992), 'Size-biased sampling of Poisson point processes and excursions', *Probability Theory and Related Fields* **92**(1), 21–39.

Pitman, J. (1995), 'Exchangeable and partially exchangeable random partitions', *Probability Theory and Related Fields* **102**(2), 145–158.

Pitman, J. (1996*a*), 'Random discrete distributions invariant under size-biased permutation', *Advances in Applied Probability* **28**(2), 525–539.

Pitman, J. (1996*b*), Some developments of the Blackwell-MacQueen urn scheme, *in* T. F. et al., ed., 'Probability and Game Theory: Papers in Honor of David Blackwell', Vol. 30 of *Lecture Notes-Monograph Series*, Institute of Mathematical Statistics, pp. 245–267.

Pitman, J. (1997), 'Partition structures derived from Brownian motion and stable subordinators', *Bernoulli* **3**(1), 79–96.

Pitman, J. (2006), *Combinatorial Stochastic Processes*, number 1875 *in* 'École d'Été de Probabilités de Saint-Flour', 1 edn, Springer-Verlag Berlin Heidelberg.

Pitman, J. and Yor, M. (1997), 'The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator', *Annals of Probability* **25**(2), 855–900.

Tavaré, S. (1987), 'The birth process with immigration, and the genealogical structure of large populations', *Journal of Mathematical Biology* **25**(2), 161–168.

Tavaré, S. (1989), The genealogy of the birth, death and immigration process, *in* M. W. Feldman, ed., 'Mathematical Evolutionary Theory', Princenton University Press, pp. 41–56.

Watterson, G. A. (1974), 'The sampling theory of selectively neutral alleles', *Advances in Applied Probability* **6**(3), 463–488.