

# Diffusion Processes and the Ewens Sampling Formula

Shui Feng

*Abstract.* Crane [*Statist. Sci.* **31** (2016) 1–19] provides an excellent review of Ewens’ sampling formula (henceforth, ESF), its applications in and connections with various subjects. This note intends to extend the discussion a little bit. The focus will be on nonequilibrium ESF involving diffusion processes, ESF with symmetric selection and asymptotics of ESF. The references listed are by no means exhaustive.

*Key words and phrases:* Asymptotics, selection, nonequilibrium.

Let  $\mathcal{S}^\downarrow$  be the ranked simplex and  $\nu_\theta$  be Kingman’s Poisson–Dirichlet distribution on  $\mathcal{S}^\downarrow$  with parameter  $\theta > 0$ . For notational convenience, we denote the generic element of  $\mathcal{S}^\downarrow$  by  $\mathbf{x} = (x_1, x_2, \dots)$ ,  $\mathbf{y} = (y_1, y_2, \dots)$ , etc. Consider a population of individuals of various types. If the random proportions of types follow the law  $\nu_\theta$ , then ESF gives the distribution of the allelic partitions of random samples from the population. Given the sample size  $n$  and an allelic partition  $\mathbf{m} = (m_1, \dots, m_n)$ , set  $f_{\mathbf{m}}(\mathbf{x}) = 1$  for  $n = 1$  and for  $n > 1$ ,

$$f_{\mathbf{m}}(\mathbf{x}) = \frac{n!}{\prod_{j=1}^n (j!)^{m_j} m_j!} \cdot \sum_{\text{distinct } k_{ij}} x_{k_{11}} \cdots x_{k_{1m_1}} x_{k_{21}}^2 \cdots \cdot x_{k_{2m_2}}^2 \cdots x_{k_{n1}}^n \cdots x_{k_{nm_n}}^n.$$

The ESF can be written as  $ESF_n(\mathbf{m}; \theta) = \int_{\mathcal{S}^\downarrow} f_{\mathbf{m}}(\mathbf{x}) \nu_\theta(d\mathbf{x})$ .

The distribution  $\nu_\theta$  can be constructed from a sequence of Dirichlet distributions through a Poisson type limiting procedure. Ethier and Kurtz (1981) generalized this construction to a dynamical setting and constructed an infinite dimensional diffusion process with reversible measure  $\nu_\theta$ . The process is an infinite dimensional limit of a sequence of finite-dimensional Wright–Fisher diffusions. It describes the evolution of

random proportions of a population under the influence of parent independent mutation with mutation rate  $\theta$  and random sampling. The generator of the process on an appropriate domain has the form

$$\mathcal{A}_\theta = \frac{1}{2} \left[ \sum_{i,j=1}^{\infty} x_i (\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \partial x_j} - \theta \sum_{i=1}^{\infty} x_i \frac{\partial}{\partial x_i} \right],$$

$\delta_{ij}$  is the Kronecker delta.

Starting from any point  $\mathbf{x}$ , the distribution  $\nu_\theta(t)$  of the process at each fixed time  $t > 0$  is shown in Ethier (1992) to be absolutely continuous with respect to  $\nu_\theta$  and the density function is

$$q(t, \mathbf{x}, \mathbf{y}) = 1 + \sum_{k=2}^{\infty} e^{-\lambda_k t} \varphi_k(\mathbf{x}, \mathbf{y}),$$

where  $\lambda_k = \frac{k(k-1+\theta)}{2}$ , and

$$\varphi_k(\mathbf{x}, \mathbf{y}) = \frac{2k-1+\theta}{k!} \sum_{n=0}^k (-1)^{k-n} \binom{k}{n} \cdot \frac{\Gamma(n+k-1+\theta)}{\Gamma(n+\theta)} p_n(\mathbf{x}, \mathbf{y}).$$

Here  $\Gamma(\cdot)$  denotes the gamma function, and the function  $p_n(\mathbf{x}, \mathbf{y})$  has the form

$$p_n(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{m}} \frac{f_{\mathbf{m}}(\mathbf{x}) f_{\mathbf{m}}(\mathbf{y})}{ESF_n(\mathbf{m}; \theta)},$$

and the summation is over all allelic partitions of the size  $n$  sample. It is clear that  $q(t, \mathbf{x}, \mathbf{y})$  converges to 1 as  $t$  tends to infinity. The summation starting from 2 is a result of the ordering procedure.

Shui Feng is Professor, Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada L8S 4K1 (e-mail: shuifeng@mcmaster.ca).

Rearranging the terms, one obtains the following representation for  $q(t, \mathbf{x}, \mathbf{y})$ :

$$q(t, \mathbf{x}, \mathbf{y}) = d_0(t) + \sum_{n=1}^{\infty} d_n(t) p_n(\mathbf{x}, \mathbf{y}),$$

where

$$d_0(t) = 1 - \sum_{k=1}^{\infty} e^{-\lambda_k t} \frac{2k + \theta - 1}{k!} (-1)^{k-1} \frac{\Gamma(\theta + k - 1)}{\Gamma(\theta)}$$

and for  $n \geq 1$ ,

$$d_n(t) = \sum_{k=n}^{\infty} e^{-\lambda_k t} \frac{2k + \theta - 1}{k!} (-1)^{k-n} \binom{k}{n} \cdot \frac{\Gamma(n + \theta + k - 1)}{\Gamma(n + \theta)}.$$

Here  $d_n(t)$  is the probability of having  $n$  ancestors at time  $t$  in Kingman's coalescent and the representation of  $d_n(t)$  is derived in [Tavaré \(1984\)](#).

This representation of  $q(t, \mathbf{x}, \mathbf{y})$  gives a clear picture about the population structure at each positive time  $t$ . Initially individuals in the population have types (old types) with proportions  $\mathbf{x}$ . The population evolves under the influence of random sampling and mutation. Random sampling changes proportions of each type while each mutation results in a new type not seen before. At each positive time, the population is a mixture of individuals of old types and new types. The number of old types is always finite and the distribution is given by Kingman's coalescent. For each  $n \geq 1$ , the function  $p_n(\mathbf{x}, \mathbf{y})$  reflects the details of the mixture when the number of old types is  $n$  and the type of proportions in the population is  $\mathbf{y}$ . An unordered model, a particular Fleming–Viot process, is studied in [Ethier and Griffiths \(1993\)](#) where the distribution at each positive time is represented as a mixture of posteriors of the Dirichlet process. The mixing factor is given by Kingman's coalescent.

For each fixed  $t > 0$ , the nonequilibrium ESF gives the distribution of the allelic partitions of random samples from the population when the random proportions follow the law  $\nu_\theta(t)$ . For any  $n \geq 1$  and allelic partition  $\mathbf{m}$ , the nonequilibrium ESF is

$$ESF_n(\mathbf{m}; t, \theta) = ESF_n(\mathbf{m}; \theta) + F_n(t),$$

where

$$F_n(t) = \sum_{k=2}^{\infty} e^{-\lambda_k t} \int_{S^\downarrow} \varphi_k(\mathbf{x}, \mathbf{y}) f_{\mathbf{m}}(\mathbf{y}) \nu_\theta(d\mathbf{y}).$$

The nonequilibrium factor  $F_n(t)$  describes the impact of finite time and diminishes as  $t$  tends to infinity.

[Griffiths \(1979a\)](#) is the first to obtain the nonequilibrium or transient ESF. The integral on the right-hand of the above equation can be calculated explicitly.

For  $0 < \alpha < 1, \theta + \alpha > 0$ , let  $\nu_{\alpha, \theta}$  denote the two-parameter Poisson–Dirichlet distribution. [Petrov \(2009\)](#) constructed an infinite dimensional diffusion process that has  $\nu_{\alpha, \theta}$  as the reversible measure. Alternate constructions were obtained later in [Feng and Sun \(2010\)](#), [Ruggiero and Walker \(2009\)](#). The generator of the diffusion process has the form

$$\mathcal{A}_{\alpha, \theta} = \frac{1}{2} \left[ \sum_{i, j=1}^{\infty} x_i (\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \partial x_j} - \theta \sum_{i=1}^{\infty} (x_i + \alpha) \frac{\partial}{\partial x_i} \right]$$

on an appropriate domain and the transition density function is obtained in [Feng et al. \(2011\)](#). Given the sample size  $n$  and the allelic partition  $\mathbf{m}$ , the Pitman's sampling formula or the two-parameter ESF,  $PSF_n(\mathbf{m}; \alpha, \theta)$ , also has a nonequilibrium version:

$$PSF_n(\mathbf{m}; t, \alpha, \theta) = PSF_n(\mathbf{m}; \alpha, \theta) + G_n(t),$$

where  $G_n(t)$  is obtained by replacing  $\nu_\theta$  with  $\nu_{\alpha, \theta}$  in the expression of  $F_n(t)$ . More details are found in [Xu \(2011\)](#), [Zhou \(2015\)](#).

*ESF with selection.*

For any real numbers  $s$  and  $r \geq 1$ , set

$$h_r(\mathbf{x}) = \sum_{i=1}^{\infty} x_i^r, \quad \phi_r(\mathbf{x}) = \exp\{s h_r(\mathbf{x})\}.$$

The function  $h_2$  is the homozygosity and the parameter  $s$  is the selection intensity. The probability

$$\nu_\theta^{\phi_r}(d\mathbf{x}) = \phi_r(\mathbf{x}) \nu_\theta(d\mathbf{x})$$

is called the Poisson–Dirichlet distribution with symmetric selection. [Grote and Speed \(2002\)](#) studied the sampling formula under  $\nu_\theta^{\phi_2}$  when  $s < 0$  and obtained a useful approximation. [Handa \(2005\)](#) studied the general case. For given sample size  $n$  and allelic partition  $\mathbf{m}$ , the sampling formula is

$$\begin{aligned} & \int_{S^\downarrow} f_{\mathbf{m}}(\mathbf{x}) \nu_\theta^{\phi_r}(d\mathbf{x}) \\ &= ESF_n(\mathbf{m}; \theta) + n! \prod_{i=1}^n \frac{\theta^{m_i}}{(j!)^{m_i} m_i!} \sum_{l=1}^{\infty} \frac{\theta^l}{l!} I_l(\mathbf{m}), \end{aligned}$$

where  $I_l(\cdot)$  has explicit integral form depending on  $s$  and  $r$ . The structure of this formula is very similar

to the nonequilibrium ESF. An alternate derivation is found in Huillet (2007).

#### *Asymptotics.*

In the neutral evolution model, the parameter  $\theta$  is the scaled population mutation rate and is equal to  $4Nu$  with  $u$  being the individual mutation rate and  $N$  the effective population size. The Poisson type limit is to let  $N$  tend to infinity and  $u$  tend to zero while the product  $Nu$  is held constant. If  $N$  goes to infinity faster or slower than  $1/u$ , one would be dealing with limiting procedures of  $\theta$  tending to infinity or zero. Given sample size  $n$ , let  $\mathbf{m}_0 = (0, 0, \dots, 1)$  and  $\mathbf{m}_\infty = (n, 0, \dots, 0)$  be two allelic partitions. Then the ESFs corresponding to  $\theta = 0$  and  $\theta = \infty$  are Dirac measures at  $\mathbf{m}_0$  and  $\mathbf{m}_\infty$ , respectively. Asymptotic results for ESF such as central limit theorems and large deviations can be found in Griffiths (1979b), Joyce, Krone and Kurtz (2002) and Feng (2007).

#### ACKNOWLEDGMENTS

Supported in part by the Natural Science and Engineering Research Council of Canada.

#### REFERENCES

- CRANE, H. (2016). The ubiquitous Ewens sampling formula. *Statist. Sci.* **31** 1–19.
- ETHIER, S. N. (1992). Eigenstructure of the infinitely-many-neutral-alleles diffusion model. *J. Appl. Probab.* **29** 487–498. [MR1174426](#)
- ETHIER, S. N. and GRIFFITHS, R. C. (1993). The transition function of a Fleming–Viot process. *Ann. Probab.* **21** 1571–1590. [MR1235429](#)
- ETHIER, S. N. and KURTZ, T. G. (1981). The infinitely-many-neutral-alleles diffusion model. *Adv. in Appl. Probab.* **13** 429–452. [MR0615945](#)
- FENG, S. (2007). Large deviations associated with Poisson–Dirichlet distribution and Ewens sampling formula. *Ann. Appl. Probab.* **17** 1570–1595. [MR2358634](#)
- FENG, S. and SUN, W. (2010). Some diffusion processes associated with two parameter Poisson–Dirichlet distribution and Dirichlet process. *Probab. Theory Related Fields* **148** 501–525. [MR2678897](#)
- FENG, S., SUN, W., WANG, F.-Y. and XU, F. (2011). Functional inequalities for the two-parameter extension of the infinitely-many-neutral-alleles diffusion. *J. Funct. Anal.* **260** 399–413. [MR2737405](#)
- GRIFFITHS, R. C. (1979a). Exact sampling distributions from the infinite neutral alleles model. *Adv. in Appl. Probab.* **11** 326–354. [MR0526416](#)
- GRIFFITHS, R. C. (1979b). On the distribution of allele frequencies in a diffusion model. *Theoret. Population Biol.* **15** 140–158. [MR0528914](#)
- GROTE, M. N. and SPEED, T. P. (2002). Approximate Ewens formulae for symmetric overdominance selection. *Ann. Appl. Probab.* **12** 637–663. [MR1910643](#)
- HANDA, K. (2005). Sampling formulae for symmetric selection. *Electron. Commun. Probab.* **10** 223–234. [MR2182606](#)
- HUILLET, T. (2007). Ewens sampling formulae with and without selection. *J. Comput. Appl. Math.* **206** 755–773. [MR2333711](#)
- JOYCE, P., KRONE, S. M. and KURTZ, T. G. (2002). Gaussian limits associated with the Poisson–Dirichlet distribution and the Ewens sampling formula. *Ann. Appl. Probab.* **12** 101–124. [MR1890058](#)
- PETROV, L. A. (2009). A two-parameter family of infinite-dimensional diffusions on the Kingman simplex. *Funct. Anal. Appl.* **43** 279–296.
- RUGGIERO, M. and WALKER, S. G. (2009). Countable representation for infinite dimensional diffusions derived from the two-parameter Poisson–Dirichlet process. *Electron. Commun. Probab.* **14** 501–517. [MR2564485](#)
- TAVARÉ, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoret. Population Biol.* **26** 119–164. [MR0770050](#)
- XU, F. (2011). The sampling formula and Laplace transform associated with the two-parameter Poisson–Dirichlet distribution. *Adv. in Appl. Probab.* **43** 1066–1085. [MR2867946](#)
- ZHOU, Y. (2015). Ergodic inequality of a two-parameter infinitely-many-alleles diffusion model. *J. Appl. Probab.* **52** 238–246. [MR3336858](#)