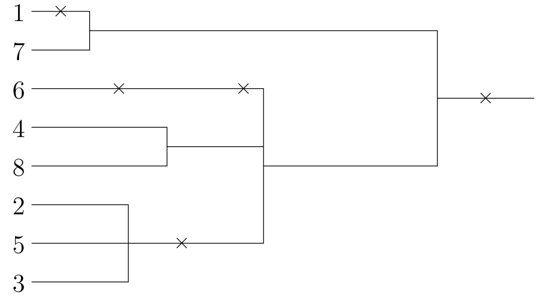
Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitman coalescent

Christina Goldschmidt
(joint work with Anne-Laure Basdevant)

We imagine a coalescent process as modelling the genealogy of a sample from a population which is subject to neutral mutation. We work under the assumptions of the infinitely many alleles model so that, in particular, every mutation gives rise to a completely new type in the population. Mutations occur as a Poisson process of rate ρ along the branches of the coalescent tree. The *allelic partition* groups together individuals of the same allelic type, and is obtained by tracing each individual's lineage back in time to the most recent mutation. An example of this construction is given below.



The allelic partition here is $\{1\}, \{2,3,5\}, \{4,7,8\}, \{6\}$. Let $N_k(n)$ be the number of blocks of size k in the allelic partition, when we start with a sample of n individuals and let $N(n) = \sum_{k=1}^{n} N_k(n)$. The *allele frequency spectrum* is the vector $(N_1(n), N_2(n), \ldots)$ of block counts.

We are interested in the distribution of the allele frequency spectrum associated with the Λ-coalescents, a class of exchangeable coalescent processes introduced by Pitman [10] and Sagitov [11]. Each such process corresponds to a finite measure Λ on $[0,1]$. Since the state of a Λ-coalescent is an exchangeable random partition of $\mathbb{N}$ for all times and the mutation occurs in a symmetric manner, the allelic partition is, itself, exchangeable. This entails that there exists a sequence of underlying random block frequencies $F_1 \geq F_2 \geq \ldots \geq 0$ such that $\sum_{i=1}^{\infty} F_i \leq 1$. The allelic partition could then be viewed as the partition created by sampling from these (unknown) frequencies in an i.i.d. manner, according to Kingman's paintbox process. For a general exchangeable random partition, the quantities $N_k(n)$, $k \geq 1$, and $N(n)$ (thought of the numbers of boxes discovered by the first n samples) have recently been of particular interest; see Gnedin, Hansen and Pitman [8] and the references therein.

For Kingman's coalescent, the distribution of the allele frequency spectrum is known completely and is given by the celebrated *Ewens sampling formula* [7]:

$$q(m_1, m_2, \ldots) := \mathbb{P}\left(N_1(n) = m_1, N_2(n) = m_2, \ldots\right) = \frac{n! \theta^{\sum_{i \geq 1} m_i}}{(\theta)_n \prod_{j \geq 1} j^{m_j} m_j!},$$

where $\theta = 2\rho$ and $(\theta)_n = \theta(\theta+1)\dots(\theta+n-1)$. For no other Λ-coalescent (apart from the degenerate star-shaped coalescent with $\Lambda = \delta_1$) is $q(m_1, m_2, \dots)$ known explicitly, although Möhle [9] has proved a recursion that it must satisfy. However, Berestycki, Berestycki and Schweinsberg [2, 3] have recently proved asymptotic results for the Beta coalescents with $\alpha \in (1,2)$; that is, the coalescents corresponding to

$$\Lambda(dx) = \frac{1}{\Gamma(\alpha)\Gamma(2-\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}dx.$$

In this case,

$$n^{\alpha-2}N(n) \xrightarrow{p} \frac{\rho\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}$$

and, for $k \geq 1$,

$$n^{\alpha-2}N_k(n) \xrightarrow{p} \frac{\rho\alpha(\alpha-1)^2\Gamma(k+\alpha-2)}{k!}$$

as $n \to \infty$.

The Bolthausen-Sznitman coalescent [4] is the $\alpha = 1$ Beta coalescent i.e. $\Lambda(dx) = dx$. It has several nice properties and seems to be more tractable than most other Λ-coalescents. A significant difference between it and the Beta coalescents with $\alpha \in (1,2)$ is that the Bolthausen-Sznitman coalescent does not come down from infinity; that is, it has infinitely many blocks for all time. The main result of [1] gives the corresponding (and rather different) asymptotics for the allele frequency spectrum of the Bolthausen-Sznitman coalescent:

Theorem. *As $n \to \infty$,*

$$\frac{\log n}{n} N_1(n) \xrightarrow{p} \rho$$

and, for $k \geq 2$,

$$\frac{(\log n)^2}{n} N_k(n) \xrightarrow{p} \frac{\rho}{k(k-1)}.$$

As a corollary, we obtain that $\frac{\log n}{n} N(n) \xrightarrow{p} \rho$.

The proof of this theorem involves proving a fluid limit result for the path of the coalescent with mutations, using the method described in Darling and Norris [5]. We need to add some structure in order to follow the mutations and so (following Dong, Gnedin and Pitman [6]) we allow individuals to be in two possible states: *active* (unmutated) or *frozen* (mutated). Blocks may contain both active and frozen individuals, and the status of an individual is ignored by the operation of coalescence. At rate ρ, any block containing active individuals receives a mutation. A block can contain individuals which were frozen at different times, but all those frozen at the same time form a block in the final allelic partition. The process continues until all individuals are frozen. Now suppose that we start with n active individuals in singleton blocks. For $k \geq 1$, let $X_k^n(t)$ be the number of blocks containing k active individuals at time t. Let $Z_k^n(t)$ be the number of blocks of size k in the final allelic partition which have already been formed by time t. Let $T_n = \inf\{t \geq 0 : \sum_{k=1}^n X_k^n(t) = 0\}$. Then

$$(X_1^n(t), X_2^n(t), \dots, Z_1^n(t), Z_2^n(t), \dots)_{0 \leq t \leq T_n}$$

is a Markov jump process whose terminal value is of interest to us. We show that

$$\begin{aligned}&\bigg(\tfrac{1}{n}X_1^n\left(\tfrac{t}{\log n}\right), \tfrac{\log n}{n}X_2^n\left(\tfrac{t}{\log n}\right), \tfrac{\log n}{n}X_3^n\left(\tfrac{t}{\log n}\right), \ldots,\\ &\quad \tfrac{\log n}{n}Z_1^n\left(\tfrac{t}{\log n}\right), \tfrac{(\log n)^2}{n}Z_2^n\left(\tfrac{t}{\log n}\right), \tfrac{(\log n)^2}{n}Z_3^n\left(\tfrac{t}{\log n}\right), \ldots\bigg)_{0\le t\le (\log n)T_n}\end{aligned}$$

behaves asymptotically like the deterministic function

$$(x_1(t), x_2(t), x_3(t), \ldots, z_1(t), z_2(t), z_3(t), \ldots)_{t\ge 0},$$

where

$$\begin{aligned} x_1(t) &= e^{-t}, & x_k(t) &= \frac{te^{-t}}{k(k-1)}, \quad k\ge 2,\\ z_1(t) &= \rho(1-e^{-t}), & z_k(t) &= \frac{\rho}{k(k-1)}(1-e^{-t}-te^{-t}), \quad k\ge 2.\end{aligned}$$

Heuristically, then, $(\log n)T_n$ should not be too far from $\inf\{t\ge 0 : \sum_{k=1}^{\infty} x_k(t) = 0\} = \infty$. It is possible to show rigorously that we do, indeed, have

$$\frac{\log n}{n}Z_1^n(T_n) \sim z_1(\infty) = \rho \quad \text{ and } \quad \frac{(\log n)^2}{n}Z_k^n(T_n) \sim z_k(\infty) = \frac{\rho}{k(k-1)}, k\ge 2,$$

as stated in the Theorem. See [1] for a full proof.

We remark that this fluid limit is ususual in two respects. Firstly, we scale time *down* rather than up and, secondly, different co-ordinates of the process have different scalings.

References

[1] A.-L. Basdevant and C. Goldschmidt, *Asymptotics of the allele frequency spectrum associated with the Bolthausen-Sznitman coalescent*, `arXiv:0706.2808` (2007).

[2] J. Berestycki, N. Beresycki, and J. Schweinsberg, *Beta-coalescents and continuous stable random trees*, `arXiv:math/0602113`. To appear in Ann. Probab. (2007).

[3] J. Berestycki, N. Beresycki, and J. Schweinsberg, *Small-time behavior of Beta-coalescents*, `arXiv:math/0601032`. To appear in Ann. Inst. H. Poincaré Probab. Statist. (2007).

[4] E. Bolthausen and A.-S. Sznitman, *On Ruelle's probability cascades and an abstract cavity method*, Comm. Math. Phys. **197**(2) (1998), 247–276.

[5] R. W. R. Darling and J. R. Norris, *Differential equation approximations for Markov chains*, Preprint (2007).

[6] R. Dong, A. Gnedin, and J. Pitman, *Exchangeable partitions derived from Markovian coalescents*, `arXiv:math.PR/0603745` (2006).

[7] W. J. Ewens, *The sampling theory of selectively neutral alleles*, Theoret. Popul. Biol. **3** (1972), 87–112.

[8] A. Gnedin, B. Hansen, and J. Pitman, *Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws*, Probab. Surv. **4** (2007), 146–171 (electronic).

[9] M. Möhle, *On sampling distributions for coalescent processes with simultaneous multiple collisions*, Bernoulli **12**(1) (2006), 35–53.

[10] J. Pitman, *Coalescents with multiple collisions*, Ann. Probab. **27**(4) (1999), 1870–1902.

[11] S. Sagitov, *The general coalescent with asynchronous mergers of ancestral lines*, J. Appl. Probab. **36**(4) (1999), 1116–1125.