Random forests and the additive coalescent

This chapter reviews how various representations of additive coalescent processes, whose state space may be either finite or infinite partitions, can be constructed from random trees and forests. These constructions establish deep connections between the asymptotic behaviour of additive coalescent processes and the theory of Brownian trees and excursions. There are some close parallels with the theory of multiplicative coalescents and the asymptotics of critical random graphs, described in Section 6.4.

10.1. Random $p$-forests and Cayley’s multinomial expansion For each probability distribution $p$ on a set $S$ of $n$ elements, Cayley’s multinomial expansion allows the definition of a random $p$-forest $F_{n,k}$ of $k$ trees labeled by $S$. There is a natural way to realize these forests as a forest-valued fragmentation process $(F_{n,k}, 1 \leq k \leq n)$ where one edge of the forest is lost at each step by uniform random selection from all remaining edges. Time-reversal of this process yields a forest-valued coalescent process. The corresponding sequence of random partitions of $S$ coalesces in such a way that two blocks $A_i$ and $A_j$ merge at each step with probability proportional to $p(A_i) + p(A_j)$.

10.2. The additive coalescent A continuous time variant of this construction yields a partition-valued additive coalescent process in which blocks $A_i$ and $A_j$ merge at rate $p(A_i) + p(A_j)$. This is compared with other constructions of additive coalescent processes with various state spaces.

10.3. The standard additive coalescent This continuous time process, parameterized by $\mathbb{R}$, with state space the set $\mathcal{P}_{\downarrow 1}$ of sequences $x = (x_i)_{i \geq 1}$ with $x_1 \geq x_2 \geq \cdots \geq 0$ and $\sum_i x_i = 1$, is obtained as the limit in distribution as $n \to \infty$ of a time-shifted sequence of ranked additive coalescent processes, starting with $n$ equal masses of size $1/n$ at time $-\frac{1}{2} \log n$. It is known that there are many other such “eternal” additive coalescents.

10.4. Poisson cutting of the Brownian tree An explicit construction of the standard additive coalescent is obtained by cutting the branches of a Brownian tree by a Poisson point process of cuts along the skeleton of the tree at rate one per unit time per unit length of skeleton. This yields
the Brownian fragmentation process, from which the standard additive coalescent is recovered by a non-linear time reversal.

10.1. Random \( p \)-forests and Cayley’s multinomial expansion

It is hard to overemphasize the importance of Cayley’s discovery that in the expansion of \( (x_1 + \cdots + x_n)^{n-2} \) the multinomial coefficient of \( \prod_i x_i^{n_i} \) is the number of unrooted trees labeled by \([n]\) in which each vertex \( i \) has degree \( n_i - 1 \).

Many variations of Cayley’s expansion are known. One of the most useful can be presented as follows. For a finite set \( S \) and \( R \subseteq S \), let for \((S, R)\) be the set of all forests labeled by \( S \), whose set of roots is \( R \). And for \( F \in \text{for}(S, R) \) and \( s \in S \) let \( F_s \) denote the set of children of \( s \) in \( F \). So \(|F_s|\) is the number of children, or in-degree of \( s \) in \( F \). Recall that edges of \( F \) are assumed to be directed towards the roots, and note that for each \( F \in \text{for}(S, R) \) the sets \( F_s \) as \( s \) ranges over \( S \) are disjoint, possibly empty sets, whose union is \( S - R \). Then there is the forest volume formula

\[
\sum_{F \in \text{for}(S, R)} \prod_{s \in S} x_s^{\mid F_s \mid} = \left( \sum_{r \in R} x_r \right) \left( \sum_{s \in S} x_s \right)^{|S| - |R| - 1}. \tag{10.1}
\]

For \( |R| = 1 \) this amounts to Cayley’s expansion of \( (\sum_{s \in S} x_s)^{|S| - 2} \), and for \( x_s \equiv 1 \) it yields Cayley’s formula

\[
\mid \text{for}(S, R) \mid = |R| |S|^{|S| - |R| - 1} . \tag{10.2}
\]

See [360] for various proofs of the forest volume formula (10.1), and [359] for a number of probabilistic applications. Taking \( S = [n] \) and summing (10.1) over all subsets \( R \) of \([n]\) with \( |R| = k \) gives the cruder identity

\[
\sum_{F \in \text{for}[n, k]} \prod_{s=1}^n x_s^{\mid F_s \mid} = \binom{n-1}{k-1} \left( \sum_{s=1}^n x_s \right)^{n-k} \tag{10.3}
\]

where \( \text{for}[n, k] \) is the set of all forests of \( k \) trees labeled by \([n]\). This was obtained earlier in formula (6.22) as one of several enumerations equivalent to the Lagrange inversion formula Section 6.1.

Take \( x_s = p_s \) for a probability distribution \( p := (p_s) \) on \([n]\), or any other set \( S \) with \( |S| = n \), to see that for each \( k \in [n] \) the formula

\[
\mathbb{P}(\mathcal{F}_{n,k} = F) = \binom{n-1}{k-1}^{-1} \prod_{s \in S} p_s^{\mid F_s \mid} \tag{10.4}
\]

defines the distribution for a random forest \( \mathcal{F}_{n,k} \) of \( k \) trees labeled by \( S \) with \( |S| = n \), call it a \( p \)-forest of \( k \) trees labeled by \( S \). In particular, call \( \mathcal{F}_{n,1} \) a \( p \)-tree. Several natural constructions of \( p \)-trees from a sequence of independent and identically distributed random variables with distribution \( p \) are recalled in the exercises. The next theorem is fundamental to everything which follows.