Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Computing likelihoods for coalescents with multiple collisions in the infinitely-many-sites model

Matthias Birkner[1,*], Jochen Blath[2,†]

submitted: 14th June 2007

[1] Weierstraß-Institut für Angewandte Analysis und Stochastik
Mohrenstraße 39
10117 Berlin
Germany

E-mail: birkner@wias-berlin.de

[2] Institut für Mathematik, Technische Universität Berlin
Straße des 17. Juni 136
10623 Berlin
Germany

E-mail: blath@math.tu-berlin.de

No. 1237
Berlin 2007

2000 *Mathematics Subject Classification.* Primary: 92D15; Secondary: 60G09, 60G52, 60J75, 60J85.

Key words and phrases. Λ-coalescent, likelihood-based inference, infinitely-many-sites, population genetics, Monte-Carlo method.

*Partially supported by EPSRC GR/R98563/01.

†Partially supported by EPSRC GR/R985603.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

One of the central problems in mathematical genetics is the inference of evolutionary parameters of a population (such as the mutation rate) based on the observed genetic types in a finite DNA sample. If the population model under consideration is in the domain of attraction of the classical Fleming-Viot process, such as the Wright-Fisher- or the Moran model, then the standard means to describe its genealogy is Kingman's coalescent. For this coalescent process, powerful inference methods are well-established.

An important feature of the above class of models is, roughly speaking, that the number of offspring of each individual is small when compared to the total population size, and hence all ancestral collisions are binary only. Recently, more general population models have been studied, in particular in the domain of attraction of so-called *generalised* Λ*-Fleming-Viot processes*, as well as their (dual) genealogies, given by the so-called Λ*-coalescents*, which allow multiple collisions. Moreover, Eldon and Wakeley (2006) provide evidence that such more general coalescents might actually be more adequate to describe real populations with extreme reproductive behaviour, in particular many marine species. In this paper, we extend methods of Ethier and Griffiths (1987) and Griffiths and Tavaré (1994, 1995) to obtain a likelihood based inference method for general Λ-coalescents. In particular, we obtain a method to compute (approximate) likelihood surfaces for the observed type probabilities of a given sample. We argue that within the (vast) family of Λ-coalescents, the parametrisable sub-family of Beta$(2-\alpha, \alpha)$-coalescents, where $\alpha \in (1, 2]$, are of particular relevance. We illustrate our method using simulated datasets, thus obtaining maximum-likelihood estimators of mutation and demographic parameters.

1 Introduction

Even though coalescents with multiple collisions have been studied quite extensively in the mathematical literature over the last decade (see, e.g. [P99], [S99], [S03], [MS01], [BLG03] and [BBS06]), and have been explicitly proposed as a model for genealogies in various biological scenarios, their use in biological studies has been rather limited up to now (see, however, [EW06]).

We suspect that this is at least in part due to a lack of statistical tools, which would allow to decide which among various multiple merger coalescents is most suitable for a given population, and which would furthermore allow to draw inference about parameters of interest, e.g. mutation rates, in such scenarios. Our aim is to contribute to remedying this lack by describing and implementing methods to compute likelihoods of observed sequence data in scenarios with multiple collisions. These in turn can form the basis of tests and estimation procedures.

In the present paper, we give particular attention to the so-called Beta-coalescents, which are a one-parameter subfamily of Λ-coalescents including Kingman's coalescent (see (1.4) below), and which exhibit interesting theoretical properties as well as practical advantages (see Section 8).

1.1 Coalescent processes

For neutral population models of fixed population size in the domain of attraction of the classical Fleming-Viot process, such as the Wright-Fisher- and the Moran model, the genealogy of a finite sample can be described by the now classical Kingman-coalescent, which we introduce briefly, followed by the more recently discovered and much more general Λ-coalescents. For background

on (classical and generalised) Fleming-Viot processes and variations of Kingman's coalescent, see e.g. [EK86], [D93], [EK93] and [DK99] as well as [H90], [N01], [HSW05] and [W06].

Kingman's coalescent. Let $\mathcal{P}_n$ be the set of partitions of $\{1,\dots,n\}$ and let $\mathcal{P}$ denote the set of partitions of $\mathbb{N}$. For each $n \in \mathbb{N}$, Kingman [K82] introduced the so-called *n-coalescent*, which is a $\mathcal{P}_n$-valued continuous time Markov process $\{\Pi_n(t), t \geq 0\}$, such that $\Pi_n(0)$ is the partition of $\{1,\dots,n\}$ into singleton blocks, and then each pair of blocks merges at rate one. Given that there are b blocks at present, this means that the overall rate to see a merger between blocks is $\binom{b}{2}$. Note that only *binary mergers* are allowed. Kingman [K82] also showed that there exists a $\mathcal{P}$-valued Markov process $\{\Pi(t), t \geq 0\}$, which is now called the (standard) *Kingman-coalescent*, and whose restriction to the first n positive integers is the n-coalescent. To see this, note that the restriction of any n-coalescent to $\{1,\dots,m\}$, where $1 \leq m \leq n$, is an m-coalescent. Hence the process can be constructed by an application of the standard extension theorem.

Λ-coalescents. Pitman [P99] and Sagitov [S99] introduced and discussed coalescents which allow *multiple collisions*, i.e. more than just two blocks may merge at a time. Again, such a coalescent with multiple collisions (which will be later called a *Λ-coalescent*) is a $\mathcal{P}$-valued Markov-process $\{\Pi(t), t \geq 0\}$, such that for each n, its restriction to the first n positive integers is a $\mathcal{P}_n$-valued Markov process (the "n-Λ-coalescent") with the following transition rates. Whenever there are b blocks in the partition at present, each k-tuple of blocks (where $2 \leq k \leq b \leq n$) is merging to form a single block at rate $\lambda_{b,k}$, and no other transitions are possible. The rates $\lambda_{b,k}$ do not depend on either n or on the structure of the blocks. Pitman showed that in order to be consistent, which means that for all $2 \leq k \leq b$,

$$\lambda_{b,k} = \lambda_{b+1,k} + \lambda_{b+1,k+1},$$

such transition rates must necessarily satisfy

$$\lambda_{b,k} = \int_0^1 x^k (1-x)^{b-k} \frac{1}{x^2} \Lambda(dx), \tag{1.1}$$

for some finite measure Λ on the unit interval. Note that (1.1) sets up a one-to-one correspondence between coalescents with multiple collisions and finite measures Λ. Indeed, it is easy to see that the $\lambda_{b,k}$ determine Λ by an application of Hausdorff's moment problem, which has a unique solution in this case.

Due to the restriction property, the Λ-coalescent on $\mathcal{P}$ (with rates obtained from the measure Λ as described above) can be constructed from the corresponding n-Λ-coalescents via extension.

Note that the family of Λ-coalescents is rather large, and in particular it cannot be parametrised by a few real variables. Important examples include $\Lambda = \delta_0$ (Kingman's coalescent) and $\Lambda = \delta_1$ (leading to star-shaped genealogies, i.e. one huge merger into one single block). Later, we will be concerned with two important parametric subclasses of Λ-coalescents, namely the so-called *Beta-coalescents*, where Λ has a Beta$(2-\alpha, \alpha)$-density for some $\alpha \in (1,2]$, and simple linear combinations of atomic measures of the type $\Lambda = c_1\delta_0 + c_2\delta_y$ for some constants $c_1, c_2 > 0$ and $y \in (0,1]$. To avoid trivialities, we will henceforth assume that $\Lambda \neq 0$.

Remarks (Multiple collisions and reproduction events).

1. An important difference between the classical Kingman-coalescent and coalescents which allow multiple collisions should be pointed out here. Roughly speaking, a Kingman coalescent arises as the limiting genealogy of a so-called Cannings population model ([C74], [C75]), if the individuals produce a number of offspring that is negligible when compared to the total population size (in particular, this requires that the variance of the number of offspring per individual converges to a finite limit). More general coalescents with multiple mergers arise, once the offspring distribution is such that a non-negligible proportion, say $x \in (0,1]$, of the population alive in the next generation

goes back to a single reproduction event from a single ancestor in the present generation. In this case, $x^{-2}\Lambda(dx)$ can be interpreted as the intensity at which we see such proportions x. Precise conditions on the underlying Cannings-models and the classification of the corresponding limiting genealogies can be found in [MS01].

2. In [EW06], Eldon and Wakeley assume that there are extreme reproductive events, which account for non-negligible proportions of the population in a single reproduction event, in the population dynamics of the Pacific Oyster. In fact, many marine species seem to exhibit such behaviour (see also [A04] and [BBB94]). □

Remarks ("Coming down from infinity").

1. Not all Λ-coalescents seem to be reasonable as models for biological populations, since some do not allow for a finite "time to the most recent common ancestor" of the entire population (T_{MRCA}). This is equivalent to "coming down from infinity in finite time": it means that, given any initial partition in $\mathcal{P}$, and for all $\varepsilon > 0$, the partition $\Pi(\varepsilon)$ a.s. consists of finitely many blocks only. Letting

$$\lambda_b = \sum_{k=2}^{b} (k-1) \binom{b}{k} \lambda_{b,k},$$

Schweinsberg [S00] proves that if either Λ has an atom at 0 or Λ has no atom at zero and

$$\lambda^* := \sum_{b=2}^{\infty} \lambda_b^{-1} < \infty, \tag{1.2}$$

then the corresponding coalescent does come down from infinity (and if so, the time to come down to only one block has finite expectation).

2. An important example for a coalescent, which (only just) does not come down from infinity is the Bolthausen-Sznitman coalescent, where $\Lambda(dx) = dx$ is the uniform distribution on $[0,1]$. This is the Beta$(2-\alpha, \alpha)$-coalescent with $\alpha = 1$, and it plays an important role in statistical mechanics models for disordered systems (see e.g. [Bo06] for an introduction).

3. However, it should be observed that all n-Λ-coalescents (for finite n) do have an a.s. finite T_{MRCA}. □

Examples for coalescents which satisfy (1.2) are the process considered in [EW06], corresponding to

$$\Lambda = c_1 \delta_0 + c_2 \delta_y, \quad c_1 > 0,\ c_2 \geq 0 \tag{1.3}$$

for $y \in (0,1)$ (in particular Kingman's coalescent if $c_1 = 1, c_2 = 0$; but note that [EW06] also consider a scenario where $c_1 = 0$), the so-called Beta$(2-\alpha, \alpha)$-coalescents with $\alpha \in (1,2)$, where

$$\Lambda(dx) = \frac{\Gamma(2)}{\Gamma(2-\alpha)\Gamma(\alpha)} x^{1-\alpha}(1-x)^{\alpha-1}\, dx, \tag{1.4}$$

(even though the right-hand side of (1.4) makes no sense for $\alpha = 2$, Kingman's coalescent can be included as the weak limit Beta$(2-\alpha, \alpha) \to \delta_0$ as $\alpha \to 2$), and a coalescent discussed in Durrett and Schweinsberg [DS05],

$$\Lambda(dx) = c_1 \delta_0 + c_2 x\, dx, \quad c_1, c_2 \geq 0, c_1 + c_2 > 0, \tag{1.5}$$

which they propose to describe the genealogy at a neutral locus which is suitably linked to selected loci undergoing recurrent selective sweeps.

It is easy to interpret the behaviour of the population corresponding to the coalescent associated with (1.3). The first atom stands for a Kingman-component, i.e. essentially reproduction with

finite variance. The second atom means that with rate c_2, a single individual can produce $100 \times y\%$ of the population currently alive in a single reproduction event.

Populations with extreme reproductive behaviour. Recently, biologists have studied the genetic variation of certain marine species with rather extreme reproductive behaviour, see, e.g., Árnason [A04] (Atlantic Cod) and [BBB94] (Pacific Oyster). Eldon and Wakeley [EW06] analysed the sample described in [BBB94] and proposed a one-parameter family of Λ-coalescents, which comprises Kingman's coalescent as a boundary case, namely those described by (1.3), as models for their genealogy. Inference is then based on a simple *summary statistic*, the number of *segregating sites* and *singleton polymorphisms*. They conclude that ([EW06], p. 2622):

> *For many species, the coalescent with multiple mergers might be a better null model than Kingman's coalescent.*

In this paper, we obtain a method to compute the full likelihood of sequence observations under the infinitely-many sites model for general Λ-coalescents. This method can then be used to obtain maximum-likelihood estimators for demographic and mutational parameters.

We apply our method to the special case of the one-parameter family of $\mathrm{Beta}(2-\alpha, \alpha)$-coalescents from (1.4), where $\alpha \in (1, 2]$, and illustrate its use on simulated datasets. These coalescents arise as limits of genealogies of a class of neutral models, where the probability that the individual litter size exceeds $k \in \mathbb{N}$ decreases like $C \times k^{-\alpha}$ for some $C > 0$ ([S03], [BBC05]). See Section 8 for further details. Still, it appears an open problem to determine which Λ-coalescent is most suitable in which biological scenario.

For an application of our method to real sequence data and a more thorough discussion of underlying biological assumptions, we refer to a forthcoming article.

Inference for Kingman's coalescent. Efficient likelihood-based inference methods for Kingman's coalescent, some solving recursion (4.4) approximately via Monte Carlo methods, others using MCMC, have been developed since the beginning of the 1990ies, see [EG87], [GT94a], [GT94b], [GT94c], [GT96a], [GT96b], [GT97], [FKY99], [DIG04a], [SD00]. In [SD00], Stephens and Donnelly provide proposal distributions for importance sampling, which are optimal in some sense, and compare them to various other methods. Their importance sampling scheme seems, at present, to be the most efficient tool for inference for relatively large datasets.

1.2 Outline of the paper

In Section 2, we discuss some combinatorial properties of observations complying with the infinitely-many-sites model which we will require subsequently.

In Section 3, we present the probabilistic neutral coalescent model that gives rise to our data.

Section 4 contains recursions for the type probabilities assuming a given underlying Λ-coalescent.

In Section 5, we briefly state recursions of the above kind in the finite- and infinite-alleles cases. A detailed derivation of the finite-alleles recursions can be found in [BB07]. For completeness, we recall the recursion obtained by Möhle in [M06] for the infinite-alleles model.

In Section 6, we derive proposal transitions for a Markov chain that we then use to obtain a Monte Carlo scheme for the type probabilities resp. likelihoods obtained in Section 4 under the Λ-coalescent in the infinite-sites model.

Section 7 contains an urn-like algorithm for convenient generation of datasets under the general coalescent model.

Figure 2.1: Forbidden sub-patterns in the IMS

			0	0
0	1		0	1
1	0		1	0
1	1		1	1
(a)			(b)	

(a) Known ancestral types, (b) Unknown ancestral types

In Section 8, we discuss population models whose genealogies are naturally approximated by Beta-coalescents, and present some likelihood-surfaces, obtained by applying our Monte Carlo method to several simulated datasets.

Finally, in Section 9 (the Appendix), we include the original genetrees corresponding to our samples that lead to the likelihood-surfaces in Section 8.

2 Combinatorics of the infinitely-many-sites model

The infinitely-many-sites (IMS) model ([K69], [W75]) is a popular model in population genetics to describe the variability in DNA samples. It assumes that the locus under consideration consists of a(n in theory infinitely long) sequence of completely linked sites and that each site is hit at most once by a mutation in the entire history of the sample. It may e.g. be justified by considering a suitable limit of diverging sequence lengths and small mutation rates. In this section, we discuss some combinatorial properties of observations complying with the IMS model which we require later. See e.g. [EG87], [H90], [GT95] or [T01] for an overview.

Observations consist of n aligned sequences, where at most two different bases are visible at each site, and say s sites are segregating. To fix notation, we think of numbering the samples and the segregating sites in some (arbitrary) fashion.

2.1 Known ancestral types and rooted genealogical trees

Assuming that ancestral types are known, e.g. by comparing with a sequence from a suitable outgroup, the data is represented by an $n \times s$ matrix $S = (s_{ij})$, where $s_{ij} = 0$ if sample i has the ancestral type at segregating site j, and $s_{ij} = 1$ if it has the mutant type. It is natural to condense this matrix by removing identical rows (corresponding to types which were observed more than once in the sample). Enumerate the, say $d \leq n$, different types in some (arbitrary) way. Then the data can be equivalently described by a $d \times s$-matrix S together with an ordered partition $\mathbf{a} = (a_1, \ldots, a_d)$ of $\{1, \ldots, n\}$, where a_i are the (numbers of the) samples of type i. The data are compatible with the IMS model if no sub-pattern as in Figure 2.1 (a) or any of its row permutations appears in S; equivalently, if O_j denotes the set of types which carry mutation j, we must have for any pair k, j that $O_j \cap O_k \neq \emptyset \Rightarrow O_j \subset O_k$ or $O_k \subset O_j$. Violations of the IMS assumption can be caused by parallel or back mutations or by recombination. A data matrix S compatible with IMS can be equivalently described by (the partition $\mathbf{a}$ and) a rooted genealogical tree $\mathbf{t}$, where the leaves correspond to observed sequences and internal nodes to mutations. A possible way to encode such trees is via

$$\mathbf{t} = \big(\mathbf{x}_1, \ldots, \mathbf{x}_d\big), \tag{2.1}$$

where $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ij(i)})$ is the sequence of mutations observed when tracing from type i backwards to the root. The fact that $\mathbf{t}$ is a rooted tree is equivalent to the following conditions:

1) Coordinates within each sequence $\mathbf{x}_i$ are distinct.

2) If for some $i, i' \in \{1, \ldots, d\}$ and j, j' we have $x_{ij} = x_{i'j'}$, then

$$x_{i,j+k} = x_{i',j'+k}, \quad \text{for all } k.$$

3) There is at least one coordinate common to all sequences.

It is customary to number mutations by $\{1, \ldots, s\}$ and take $x_{ij(i)} = 0$ for the "root mutation". In order to recover S from $\mathbf{t}$, simply put 1s in row i at all columns x_{ik}, $0 \leq k < j(i)$. A constructive way of obtaining $\mathbf{t}$ from the matrix S is Gusfield's algorithm ([G91]): Interpret the columns of S as binary numbers (with the first row as the most significant bit) and re-order them according to decreasing size (with the largest in the leftmost column, and ties resolved arbitrarily). The entries of $\mathbf{x}_i$ are found by "reading off" from right to left the columns j with $s_{ij} = 1$. Note that this implicitly puts a temporal order on the observed mutations, and orders mutations according to this "age", which is not necessarily completely determined by the actual sequence data. This is harmless because we will later "factor out" the mutation labels by considering appropriate equivalence classes:

Introduce equivalence relations on the set of types by writing

$$(\mathbf{x}_1, \ldots, \mathbf{x}_d) \sim (\mathbf{y}_1, \ldots, \mathbf{y}_d), \tag{2.2}$$

if there is a bijection $\xi : \mathbb{N}_0 \to \mathbb{N}_0$ with $y_{ij} = \xi(x_{ij}), i \in 1, \ldots, d;\ j = 0, 1, \ldots$. Furthermore, write

$$(\mathbf{x}_1, \ldots, \mathbf{x}_d) \approx (\mathbf{y}_1, \ldots, \mathbf{y}_d), \tag{2.3}$$

if there is a bijection $\zeta : \mathbb{N}_0 \to \mathbb{N}_0$ and a permutation σ on $\{1, \ldots, d\}$, such that $\mathbf{y}_{\sigma(i),j} = \zeta(x_{ij}), i = 1, \ldots, d; j = 0, 1, \ldots$

Under "$\sim$", the concrete labels of mutations are irrelevant. Note that in what follows, we suppress the distinction between such an equivalence class, denoted by $[\mathbf{t}]$, and a representative, denoted by $\mathbf{t}$. Under "$\approx$", tags of types become irrelevant, too.

Example: A dataset of eight alleles, which is consistent with the above rules.

1 :	$(6,1,0)$	5 :	$(7,1,0)$
2 :	$(6,1,0)$	6 :	$(8,5,1,0)$
3 :	$(10,1,0)$	7 :	$(4,3,2,0)$
4 :	$(7,1,0)$	8 :	$(9,4,3,2,0)$

Note that the allelic types $(6,1,0)$ and $(7,1,0)$ appear twice in the example, i.e. have multiplicity two. For notational convenience, our sequences all end in 0, this reflects the existence of a common "root". The labels of the mutations and the root are by no means required to be decreasing, this is just suitable convention.

Given a sample of size n, we will now write $(\mathbf{t}, \mathbf{n})$ for the pair consisting of the set of *different* types $\mathbf{t} = (\mathbf{x}_1, \ldots, \mathbf{x}_d), d \leq n$, and the multiplicity vector $\mathbf{n}$. In the above example, we have $d = 6$ and

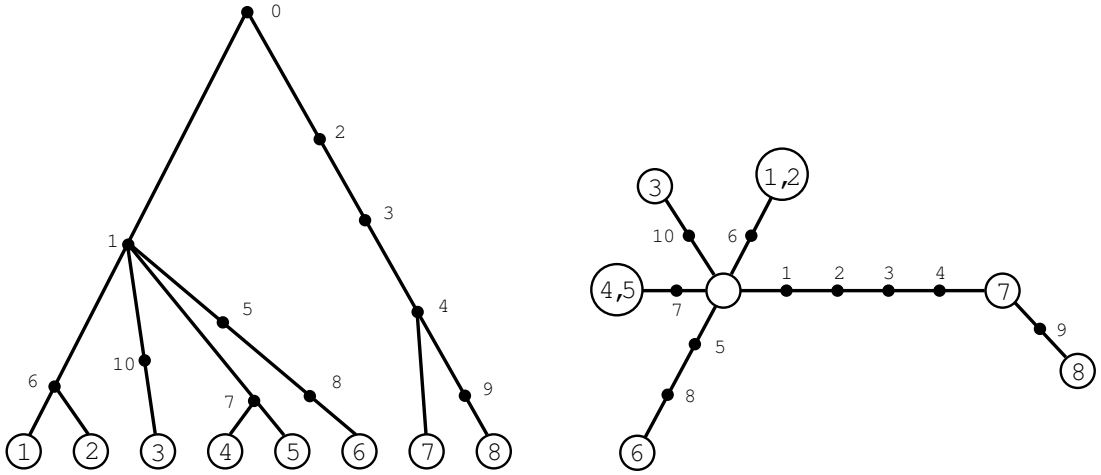
$$(\mathbf{t}, \mathbf{n}) = \Big(\big((6,1,0), (10,1,0), (7,1,0), (8,5,1,0), (4,3,2,0), (9,4,3,2,0)\big), (2,1,2,1,1,1)\Big).$$

If we take numbered samples into account, i.e. if we let $a_i \subset \{1, \ldots, n\}, i \in \{1, \ldots, d\}$ denote the set of the numbers of the sequences with type x_i, then one can also consider the set of types and ordered partitions $(t, \mathbf{a})$, where $\mathbf{a} = (a_1, \ldots, a_d)$, in the above example given by

$$(\mathbf{t}, \mathbf{a}) = \Big(\big((6,1,0), (10,1,0), (7,1,0), (8,5,1,0), (4,3,2,0), (9,4,3,2,0)\big), \big(\{1,2\}, \{3\}, \{4,5\}, \{6\}, \{7\}, \{8\}\big)\Big).$$

The probabilistic mechanism behind these data and the necessary equivalence relation will be discussed in detail in Section 3.

Figure 2.2: Rooted and unrooted tree corresponding to the example.



(a) Genetree (b) Unrooted tree

2.2 Unknown ancestral types and unrooted genealogical trees

If ancestral types are not known, the data matrix S is only specified up to flips of its columns. As above, it suffices to consider the condensed data matrix, which we again denote by S with d (pairwise different) rows together with the partition $\mathbf{a}$. The data are compatible with the IMS model in this case if and only if no sub-pattern as in Figure 2.1 (b) or any of its row permutations appears in S (the so-called "four gamete rule"). If they are compatible in this sense, they correspond to an unrooted genealogical tree, and a valid "polarised" data matrix (or equivalently, a rooted tree $\mathbf{t}$) can be obtained by flipping in such a way that in each column, 0 is the more frequent type. All other possible polarisations (resp. compatible rooted trees) can be obtained by passing to an unrooted tree, and subsequent re-rooting.

To build an unrooted tree Q from a(n equivalence class of) rooted tree(s) $\mathbf{t}$ as encoded in (2.1), proceed as follows: Vertices correspond to observed and inferred sequences (types), where an inferred type represents an internal node of degree ≥ 3 in $\mathbf{t}$; edges of $\mathbf{t}$ are merged at internal nodes of degree 2 (which were "internal" mutations in $\mathbf{t}$), and the resulting edges of Q are marked by the number of mutations they carry. Thus, Q is described by

- its set of vertices V (together with an ordered "meta-partition" $\mathbf{a}$ describing which samples correspond to which vertex, where possibly some vertices, namely the inferred types, are marked by $\emptyset$), and
- a matrix (m_{ij}), where m_{ij} is the number of mutations between vertices i and j (with the stipulation that $m_{ij} = 0$ if there is no edge between i and j in Q).

Note that this tree need not be binary. Two (equivalence classes under $\sim$ of) rooted genealogical trees $\mathbf{t}$, $\mathbf{t}'$ (with the same enumerated types and the same set of mutation labels) are equivalent as unrooted trees, in symbols $\mathbf{t} \sim_u \mathbf{t}'$, if they lead to the same unrooted tree in the construction above.

Alternatively, given an unpolarised $d \times s$ observation matrix S one can compute the pairwise difference matrix with entries

$$D_{ij} := \#\big\{1 \leq k \leq s : S_{ik} = 0, S_{jk} = 1 \text{ or } S_{ik} = 1, S_{jk} = 0\big\}. \tag{2.4}$$

It is easy to see that the four-gamete rule for S implies that this metric D on the set of types satisfies the "four-point condition":

$$\begin{gathered}\text{any four elements can be named } x, y, u, v \text{ such that} \\ D_{xy} + D_{uv} \leq D_{xu} + D_{yv} = D_{xv} + D_{yu}.\end{gathered} \tag{2.5}$$

Thus, the pairwise distance (Hamming) metric D is additive, and corresponds to a unique unrooted tree Q with integer branch lengths (see e.g. [WSS77], or use neighbour-joining [SK88]).

These two methods of obtaining an unrooted tree from an unpolarised observation matrix S are equivalent. Since a rooted tree $\mathbf{t} = (\mathbf{x}_1, \ldots, \mathbf{x}_d)$ compatible with a polarisation of S also gives rise to (the same) pairwise distance matrix $C(\mathbf{t})$ on the d types with entries

$$c_{ij}(\mathbf{t}) := \#\{k : x_{ik} \notin \mathbf{x}_j\} + \#\{k : x_{jk} \notin \mathbf{x}_i\}, \qquad 1 \leq i, j \leq d$$

(with an obvious abuse of the "$\notin$"-notation), this follows from the uniqueness of the tree defining an additive metric. Thus we have

$$\mathbf{t} \sim_u \mathbf{t}' \quad \Longleftrightarrow \quad C(\mathbf{t}) = C(\mathbf{t}'). \tag{2.6}$$

For a given unrooted tree Q with γ sequences (including inferred sequences) with m_j mutations occuring on edge j ($j = 1, \ldots, |E|$) and s segregating sites altogether (i.e. $s = \sum_j m_j$), there are

$$\gamma + \sum_j \big(m_j - 1\big) = s + 1 \tag{2.7}$$

possible positions of the root (and thus this many different rooted trees corresponding to Q): the root could be at any of the γ sequences or between any two mutations on any edge.

3 Infinite sites data and Λ-coalescent trees

To obtain an n-sample under the infinite-sites model from a coalescent tree, we perform the following probabilistic experiment. Note that by duality, this describes the distribution of a sample of size n from the stationary distribution of a Λ-generalised Fleming-Viot process ([DK99], [BLG03]) with mutation process as in [EG87].

(i) Run an n-Λ-coalescent. Obtain a *rooted* coalescent tree.

(ii) On this rooted tree with n leaves (*numbered* from 1 to n), place mutations along the branches at rate r (note that in the "Kingman world", this parameter is customarily called $\theta/2$).

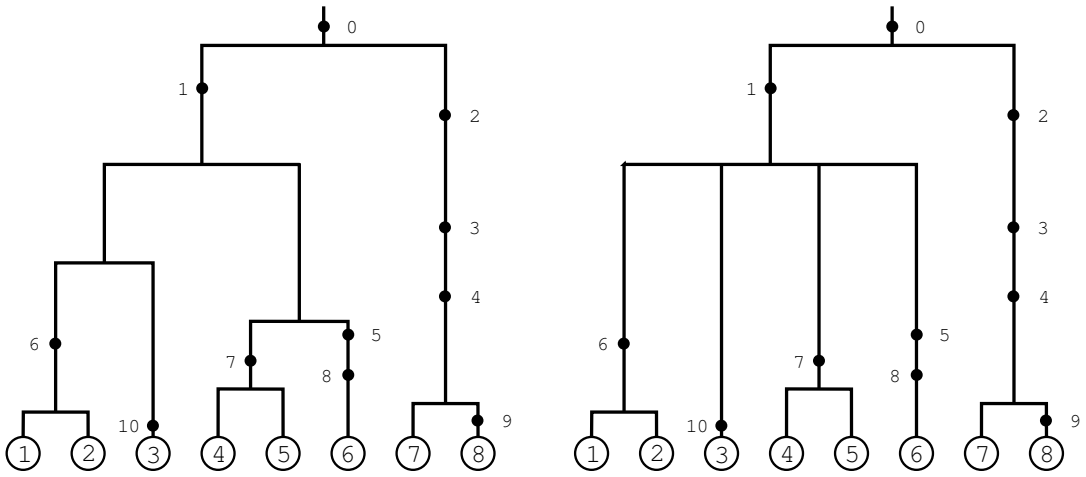
(iii) *Label* these mutations randomly: Given there are s mutations in total, attach randomly (i.e. according to the uniform distribution) the labels from $1, \ldots, s$ to these mutations.

(iv) Turn this coalescent tree with *labelled* mutations and *numbered* leaves into a "genetree" by breaking edges at mutations, resulting in vertices of degree 2, and then moving the branching points inwards until they reach the nearest mutation. Ignore the lengths of the edges.

(v) A *type* is the sequence of labels of mutations observed following the path backwards from a leaf to the root. *Enumerate* the *different* types randomly to obtain a set of sequences $\{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$, where $d \leq n$ is the number of different types.

In the following, we suppress the distinction between $\mathbf{t}$ and its equivalence class $[\mathbf{t}]$ under "$\sim$" defined in (2.2).

Figure 3.1: Two coalescent trees consistent with the example from Subsection 2.1.



(a) Binary coalescent tree (b) Coalescent tree with multiple collisions

(vi) Let $A_i \subset \{1, \dots, n\}$ be the random set of the numbers (being attached to leaves in Step 2) which have type $i \in \{1, \dots, d\}$. We obtain a random pair $(\mathbf{T}, \mathbf{A})$, where $\mathbf{A} = (A_1, \dots, A_d)$ is an *ordered* random partition.

(vii) Finally, let

$$p(\mathbf{t}, \mathbf{a}) := \mathbb{P}\{(\mathbf{T}, \mathbf{A}) = (\mathbf{t}, \mathbf{a})\}.$$

Note that, by the symmetry of the coalescent,

$$p\big(\mathbf{t}, (a_1, \dots, a_d)\big) = p\big(\mathbf{t}, (\pi(a_1), \dots, \pi(a_d))\big) \quad \text{for any permutation } \pi \in S_n. \tag{3.1}$$

We call such pairs $(\mathbf{t}, \mathbf{a})$ a *numbered sample configuration with ordered types*. Later, it will be useful to consider only the *frequencies* of the ordered types, i.e. define a map

$$\phi : (\mathbf{t}, \mathbf{a}) \mapsto (\mathbf{t}, \mathbf{n}),$$

where $\mathbf{n} = (n_1, \dots, n_d) := (\#a_1, \dots, \#a_d)$, i.e. $\sum_{i=1}^{d} n_i = n$. We denote its probability distribution by

$$\begin{aligned} p^0\big((\mathbf{t}, \mathbf{n})\big) &:= p\big(\phi^{-1}(\mathbf{t}, \mathbf{n})\big) \\ &= \frac{n!}{n_1! \cdots n_d!} p\big((\mathbf{t}, \mathbf{a})\big) \end{aligned} \tag{3.2}$$

for any $(\mathbf{t}, \mathbf{a}) \in \phi^{-1}(\mathbf{t}, \mathbf{n})$ by the observation in (3.1).

For notational simplicity, we introduce the following slightly ambiguous operations: By $\mathbf{a} - \mathbf{e}_i$, we mean a partition obtained from $\mathbf{a}$ by removing one element from the *set* a_i (with implicit adjustments so that the result is a partition of $\{1, \dots, n-1\}$). Note that we will not be concerned with the fact which element we actually remove, since, by (3.1), the type probability p will not depend on the actual choice. Similarly, by $\mathbf{a} - (k-1)\mathbf{e}_i$ we mean the partition obtained from $\mathbf{a}$ by removing $k-1$ elements from a_i (certainly, this only makes sense if $\#a_i \geq k$). Finally, $\mathbf{a} + \mathbf{e}_i$ will be the partition obtained from $\mathbf{a}$ by adding an arbitrary element of $\mathbb{N}$ to the set a_i that is not yet contained in any other set $a_l, l = 1 \dots d$.

4 Genealogical tree probabilities for Λ-coalescents in the infinite-sites model

In this section, we obtain recursions for the probability of given type configuration of a sample based on the probabilistic model discussed above. These recursions then lead to a Monte-Carlo method to (approximately) compute the probability of configurations under various Λ-coalescents.

We will distinguish two cases. In the first case, we will consider ordered labelled samples of type $(\mathbf{t}, \mathbf{a})$, which take the full information contained in the partition $\mathbf{a}$ into account. In the second case, we restrict to numbered ordered configurations of the type $(\mathbf{t}, \mathbf{n})$, which only count the multiplicities $\mathbf{n}$.

4.1 Ordered labelled samples

It is in principle possible to compute the exact probabilities of a given type configuration $(\mathbf{t}, \mathbf{a})$ via a recursive formula.

Theorem 4.1. *We have, for given* $(\mathbf{t}, \mathbf{a})$*,*

$$\begin{aligned} p(\mathbf{t},\mathbf{a}) = &\frac{1}{nr+\sum_{k=2}^{n}\binom{n}{k}\lambda_{n,k}} \sum_{i:\, n_i\geq 2} \sum_{k=2}^{n_i} \binom{n_i}{k} \lambda_{n,k}\, p(\mathbf{t}, \mathbf{a}-(k-1)\mathbf{e}_i) \\ &+ \frac{r}{nr+\sum_{k=2}^{n}\binom{n}{k}\lambda_{n,k}} \sum_{\substack{k:\, n_k=1,\, x_{k0}\ distinct \\ s(\mathbf{x}_k)\neq \mathbf{x}_j\ \forall j}} p(s_k(\mathbf{t}), \mathbf{a}) \\ &+ \frac{r}{nr+\sum_{k=2}^{n}\binom{n}{k}\lambda_{n,k}} \sum_{\substack{k:\, n_k=1, \\ x_{k0}\ distinct}} \sum_{j: s(\mathbf{x}_k)=\mathbf{x}_j} p(r_k(\mathbf{t}), r_k(\mathbf{a}+\mathbf{e}_j)), \end{aligned} \tag{4.1}$$

where $\mathbf{e}_j$ *denotes* j*-th unit vector,* $s_k(\mathbf{t})$ *deletes first coordinate of the* k*-th sequence in* t*,* $s(\mathbf{x}_k)$ *removes the first coordinate from the sequence* $\mathbf{x}_k$*,* $r_k(\mathbf{t})$ *removes* k*-th sequence from* $\mathbf{t}$*, and* x_{k0} *'distinct' means that* $x_{k0} \neq x_{ij}$ *for all* $(i,j) \neq (k,0)$*. The boundary condition for the root is* $p(\{0\},(1)) = 1$*.*

Proof. Similar to the Kingman-case by conditioning on the last event in the coalescent history, taking multiple mergers into account. The first term on the right-hand side corresponds to a (multiple) collision of lineages of the same type, hence requiring multiplicity at least two, the second term refers to the event that a mutation is removed from a type (necessarily a singleton), whose ancestral type is not visible in the sample at present. Finally, the third term corresponds to removing a mutation from a type whose ancestor is already present in the sample. □

4.2 Numbered ordered samples

Recall from (3.2), using the notation of Theorem 4.1, that

$$p^0(\mathbf{t},\mathbf{n}) = \frac{n!}{n_1!\cdots n_d!} p(\mathbf{t},\mathbf{a}). \tag{4.2}$$

Thus, for the types and multiplicities $(\mathbf{t},\mathbf{n})$, we obtain

$$\begin{aligned}
p^0(\mathbf{t},\mathbf{n}) = \frac{1}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} & \sum_{i:\, n_i\geq 2}\sum_{k=2}^{n_i}\binom{n_i}{k} \\
& \times \lambda_{n,k}\frac{n!}{n_1!\cdots n_d!}\frac{n_1!\cdots(n_i-k+1)!\cdots n_d!}{(n-k+1)!}p^0(\mathbf{t},\mathbf{n}-(k-1)\mathbf{e}_i) \\
+ \frac{r}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} & \sum_{\substack{k:\, n_k=1,\ x_{k0}\ distinct \\ s(\mathbf{x}_k)\neq \mathbf{x}_j\ \forall j}} p^0(s_k(\mathbf{t}),\mathbf{n}) \\
+ \frac{r}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} & \sum_{\substack{k:\, n_k=1, \\ x_{k0}\ distinct}}\ \sum_{j:s(\mathbf{x}_k)=\mathbf{x}_j}\frac{n!}{n_1!\cdots n_d!} \\
& \times \frac{n_1!\cdots(n_j+1)!\cdots n_d!}{n!}p^0(r_k(\mathbf{t}),r_k(\mathbf{n}+\mathbf{e}_j)).
\end{aligned}$$

Since

$$\begin{aligned}
&\binom{n_i}{k}\frac{n!}{n_1!\cdots n_d!}\frac{n_1!\cdots(n_i-k+1)!\cdots n_d!}{(n-k+1)!} \\
&\qquad = \frac{n_i!}{k!(n_i-k)!}\frac{n!(n_i-k+1)!}{n_i!(n-k+1)!} = \binom{n}{k}\frac{n_i-k+1}{n-k+1},
\end{aligned}$$

rearrangement leads to

Corollary 4.2. *For given $(\mathbf{t},\mathbf{n})$, we have*

$$\begin{aligned}
p^0(\mathbf{t},\mathbf{n}) = \frac{1}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} & \sum_{i:\, n_i\geq 2}\sum_{k=2}^{n_i}\binom{n}{k}\lambda_{n,k}\frac{n_i-k+1}{n-k+1}p^0(\mathbf{t},\mathbf{n}-(k-1)\mathbf{e}_i) \\
+ \frac{r}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} & \sum_{\substack{k:\, n_k=1,\ x_{k0}\ distinct \\ s(\mathbf{x}_k)\neq \mathbf{x}_j\ \forall j}} p^0(s_k(\mathbf{t}),\mathbf{n}) \\
+ \frac{r}{nr+\sum_{k=2}^n \binom{n}{k}\lambda_{n,k}} & \sum_{\substack{k:\, n_k=1, \\ x_{k0}\ distinct}}\ \sum_{j:s(\mathbf{x}_k)=\mathbf{x}_j}(n_j+1)p^0(r_k(\mathbf{t}),r_k(\mathbf{n}+\mathbf{e}_j)),
\end{aligned} \tag{4.3}$$

with the usual boundary condition for the root, i.e. $p^0(\{0\},(1))=1$.

Remark. Regarding our second equivalence relation "$\approx$", defined in (2.3), the probability $p^*([\mathbf{t}]_\approx,\mathbf{n})$ of observing a particular unordered and unlabelled tree is related to $p^0(\mathbf{t},\mathbf{n})$ via a combinatorial factor

$$p^*([\mathbf{t}]_\approx,\mathbf{n}) = \frac{1}{a(\mathbf{t},\mathbf{n})}p^0(\mathbf{t},\mathbf{n}),$$

where, with $\mathbf{t}_\sigma := (\mathbf{x}_{\sigma(1)},\ldots,\mathbf{x}_{\sigma(d)})$, $\mathbf{n}_\sigma = (n_{\sigma(1)},\ldots,n_{\sigma(d)})$,

$$a(\mathbf{t},\mathbf{n}) = \#\big\{\sigma\in S_d : \mathbf{t}_\sigma\sim\mathbf{t}, \mathbf{n}=\mathbf{n}_\sigma\big\}$$

is the number of permutations of the types which leave the combinatorial structure unchanged (see [GT95]). □

Remark. In the case of Kingman's coalescent, we recover from (4.3) the following recursion, which is due to Ethier and Griffiths, see [EG87] and [G89] (and replace r by $\theta/2$):

$$\begin{aligned} p^0(\mathbf{t},\mathbf{n}) &= \frac{1}{nr+\binom{n}{2}} \sum_{k:\, n_k\geq 2} \binom{n}{2} \frac{n_k-1}{n-1} p^0(\mathbf{t},\mathbf{n}-\mathbf{e}_k) \\ &\quad + \frac{r}{nr+\binom{n}{2}} \sum_{\substack{k:\, n_k=1,\, x_{k0}\ distinct \\ s(\mathbf{x}_k)\neq \mathbf{x}_j\ \forall j}} p^0(s_k(\mathbf{t}),\mathbf{n}) \\ &\quad + \frac{r}{nr+\binom{n}{2}} \sum_{\substack{k:\, n_k=1, \\ x_{k0}\ distinct}} \sum_{j: s(\mathbf{x}_k)=\mathbf{x}_j} (n_j+1) p^0(r_k(\mathbf{t}), r_k(\mathbf{n}+\mathbf{e}_j)) \end{aligned} \tag{4.4}$$

wit the same boundary condition as above. □

Remark. For samples of size $n=2$, the recursion (4.3) can easily be solved explicitly (and of course independently of Λ, as long as $\Lambda([0,1])=1$): We have

$$p^0\big((0),(2)\big) = \frac{1}{1+2r} \quad \text{and} \tag{4.5}$$

$$p^0\big((\mathbf{x}_1,\mathbf{x}_2),(1,1)\big) = 2\binom{j(1)+j(2)}{j(1)} \Big(\frac{r}{1+2r}\Big)^{j(1)+j(2)} \frac{1}{1+2r} \tag{4.6}$$

for $\mathbf{x}_1=(x_{10},\dots,x_{1j(1)})$, $\mathbf{x}_2=(x_{20},\dots,x_{2j(2)})$ (and all entries distinct except $x_{1j(1)}=x_{2j(2)}=0$). □

4.3 Unrooted genealogical trees

If the ancestral types at segregating sites are not known, the data only determine an *unrooted* tree Q, as discussed in Subsection 2.2. The probability of an observation $(Q,\mathbf{a})$ is then given by

$$p(Q,\mathbf{a}) = \sum_{T: C(T)=C(T_0)} p(T,\mathbf{a}), \tag{4.7}$$

where T_0 is any rooted tree compatible with Q (and the sum has no. of segregating sites $+1$ summands), or with unlabelled samples

$$p^0(Q,\mathbf{n}) = \sum_{T: C(T)=C(T_0)} p^0(T,\mathbf{n}). \tag{4.8}$$

By combining (4.8) and (4.3) and re-arranging as in [GT95], Sect. 2.2, we obtain

$$\begin{aligned} p^0(Q,\mathbf{n}) &= \frac{1}{nr+\sum_{\ell=2}^n \binom{n}{\ell}\lambda_{n,\ell}} \sum_{i:\, n_i\geq 2} \sum_{k=2}^{n_i} \binom{n}{k} \lambda_{n,k} \frac{n_i-k+1}{n-k+1} p^0(Q,\mathbf{n}-(k-1)\mathbf{e}_i) \\ &\quad + \frac{r}{nr+\sum_{\ell=2}^n \binom{n}{\ell}\lambda_{n,\ell}} \sum_{\substack{k:\, n_k=1,\, |k|=1 \\ k\to j,\, m_{kj}>1}} p^0(Q-\mathbf{e}_{kj},\mathbf{n}) \\ &\quad + \frac{r}{nr+\sum_{\ell=2}^n \binom{n}{\ell}\lambda_{n,\ell}} \sum_{\substack{k:\, n_k=1,\, |k|=1 \\ k\to j,\, m_{kj}=1}} \sum_{j: s(\mathbf{x}_k)=\mathbf{x}_j} (n_j+1) p^0(Q-\mathbf{e}_{kj}, r_k(\mathbf{n}+\mathbf{e}_j)), \end{aligned} \tag{4.9}$$

where $|k| = 1$ means that the degree of vertex k is 1, $k \to j$ means that vertex k is joined to vertex j, and finally, in the last term on the right-hand side, vertex k is removed from Q. The boundary condition is $p^0(Q,(1)) = 1$ for the tree consisting of one vertex only.

Remarks 1.) Note that it may be possible to draw inference about ancestral states at some or all segregating sites by comparing likelihoods for various positions of the root.

2.) As above, recursion (4.9) can be solved explicitly for samples of size $n = 2$. In fact, the only information about the two sequences in the infinitely-many-sites model is then captured by the number of segregating sites (i.e. the number of mutations), say, s. Hence, by a slight abuse of notation, we have

$$p^0((0),(2)) = \frac{1}{1+2r},$$

and

$$p^0((s)),(1,1)) = 2\left(\frac{2r}{1+2r}\right)^s \frac{1}{1+2r}, \quad s = 1,2,\ldots \tag{4.10}$$

in keeping with the idea that two samples are separated by a geometric number of mutations. □

5 Finite- and infinite alleles recursions

In this section, we provide similar recursions for the finite- and infinite alleles models of mathematical genetics. The *finitely-many-alleles* recursions can either be derived using Donnelly and Kurtz' [DK99] modified lookdown construction, assuming a given underlying generalised Λ-Fleming-Viot process, or via calculations based on the generator of the population model, as in described [DIG04a] for the Kingman-case. A detailed derivation of the recursions, using both approaches, can be found in [BB07].

Here, we consider type changes, or mutations, occurring at rate $r > 0$, and let $P = (P_{ij})$ denote a stochastic transition matrix on the corresponding finite type space E with $\#E =: d \geq 1$, and with equilibrium μ. This means that if a mutation occurs, type i mutates to type j with probability P_{ij}. Silent mutations are allowed (i.e. $P_{jj} \geq 0$). Due to exchangeability, we will merely be interested in the type frequency probability $p^0(\mathbf{n})$. So, the only relevant information is (of course) how many samples were of which type. For $\mathbf{n} = (n_1,\ldots,n_d) \in \mathbb{Z}_+^d$, denote $\#\mathbf{n} := n_1 + \cdots + n_d$. Let $p^0(\mathbf{n})$ be the probability that in a sample of size $\#\mathbf{n}$, there are exactly n_j of type j, $j = 1,\ldots,d$.

Theorem 5.1. *Abbreviate $n := \#\mathbf{n}$, and write $\mathbf{e}_k$ for the k-th canonical unit vector of $\mathbb{Z}^d$. Then, the recursion for p^0 is*

$$\begin{aligned} p^0(\mathbf{n}) &= \frac{r}{r_n}\sum_{j=1}^{d}\sum_{i=1}^{d}(n_i + 1 - \delta_{ij})P_{ij}p^0(\mathbf{n} - \mathbf{e}_j + \mathbf{e}_i) \\ &\quad + \frac{1}{r_n}\sum_{\substack{j=1 \\ n_j \geq 2}}^{d}\sum_{k=2}^{n_j}\binom{n}{k}\lambda_{n,k}\frac{n_j - k + 1}{n - k + 1}p^0(\mathbf{n} - (k-1)\mathbf{e}_j) \end{aligned} \tag{5.1}$$

with boundary conditions $p^0(\mathbf{e}_j) = \mu_j$. In the Kingman case, this agrees with (3) in [DIG04a].

In the *infinitely-many alleles case*, one assumes that every mutation, which occurs along the coalescent tree with rate $r > 0$, leads to an entirely new type, no other information is being retained. If we take a sample of $n \in \mathbb{N}$ genes, it is natural to ask for the probability $p^0(\mathbf{n})$ to sample a specific, non-ordered allele configuration $\mathbf{n} = (n_1,\ldots,n_\ell)$, where $\ell \leq n$ is the number of different types observed in the sample, and n_i, for $i \in \{1,\ldots,\ell\}$ is the number of times that type i is being

observed. Let $\tilde{\mathbf{n}}_j = (n_1, \ldots, n_{j-1}, n_{j+1}, \ldots, n_\ell)$. Using coalescent arguments, it is possible obtain the following recursion, see [M06], Theorem 3.1.

Theorem 5.2 (Möhle)**.** *The probability of a non-ordered allele configuration* $\mathbf{n} = (n_1, \ldots, n_\ell)$ *satisfies the recursion given by* $p^0(1) = 1$ *and*

$$p^0(\mathbf{n}) = \frac{nr}{\sum_{k=2}^n \binom{n}{k}\lambda_{n,k} + nr} \sum_{\substack{j=1 \\ n_j=1}}^{\ell} \frac{1}{\ell} p^0(\tilde{\mathbf{n}}_j)$$
$$+ \frac{1}{\sum_{k=2}^n \binom{n}{k}\lambda_{n,k} + nr} \sum_{k=2}^{n} \sum_{\substack{j=1 \\ n_j \geq k}}^{\ell} \binom{n}{k} \lambda_{n,k} \frac{n_j - k + 1}{n - k + 1} p^0(\mathbf{n} - (k-1)\mathbf{e}_j). \tag{5.2}$$

In the Kingman-case, this recursion can be solved explicitly and leads to an alternative formulation of the famous *Ewens sampling formula*, see [E79]. It seems that the only other case in which an explicit solution is known is the case $\Lambda = \delta_1$, in which the genealogy is star-shaped.

6 A Monte Carlo method for the computation of the likelihoods in the infinite-sites model

We first derive a simple Monte-Carlo approximation of the exact sampling probabilities in the infinite-sites model by simulating a Markov chain backwards along the sample paths of the coalescent (essentially based on [GT94b], see also [T01]). Note that this can be viewed as an integration over all paths of Algorithm 1 (see Section 7.2) which lead to the observed configuration – these correspond to "coalescent histories" as considered in [SD00], [DIG04a].

6.1 An unbiased estimator for $p^0(t, \mathbf{n})$

Given ordered types and frequencies $(\mathbf{t}, \mathbf{n})$, we define the *tree complexity* of $(\mathbf{t}, \mathbf{n})$ as

$$c[(\mathbf{t}, \mathbf{n})] = \sum_{i=1}^{d} n_i + \sum_{i=1}^{d} \#\mathbf{x}_i \in \mathbb{N}, \tag{6.1}$$

where, for $1 \leq i \leq d$, $\#\mathbf{x}_i$ denotes the length of the sequence $\mathbf{x}_i$ (exclusive of the root).

Note that the tree complexity is the sum of the sample size and the number of segregating sites. This definition transfers in the obvious way also to the pair of ordered types and partitions $(\mathbf{t}, \mathbf{a})$. It is clear that the tree complexity is independent of the choice of a representative $\mathbf{t}$ from the equivalence class $[\mathbf{t}]$ and hence well-defined. If $c[(\mathbf{t}, \mathbf{n})] = 1$, the minimum for a non-vanishing tree, then the tree consists only of its root with multiplicity one, i.e. $(\mathbf{t}, \mathbf{n}) = (\{0\}, (1)) =: \mathbf{t}_0$. We write

$$(\mathbf{t}', \mathbf{n}') \prec (\mathbf{t}, \mathbf{n})$$

if $(\mathbf{t}', \mathbf{n}')$ can be reached from $(\mathbf{t}, \mathbf{n})$ by either removing one mutation or a coalescence event, see below. In this case, $c[(\mathbf{t}', \mathbf{n}')] < c[(\mathbf{t}, \mathbf{n})]$. Hence observe that the recursions (4.4) and (4.3) are proper recursions in the sense that they strictly decrease the tree complexity in each step.

The following lemma is an appropriate version of the corresponding Lemma 6.1 in [T01].

Lemma 6.1. *Let* $\{X_k, k \geq 0\}$ *be a Markov chain on the space of ordered types with corresponding frequencies, denoted by* $(\mathcal{T}, \mathcal{N})$*, and with transitions* $\mathbb{Q} = (q_{(\mathbf{t},\mathbf{n}),(\mathbf{t}',\mathbf{n}')})$ *such that the hitting time*

$$\tau = \inf\left\{k \geq 0 \,:\, X_k = (\{0\}, (1))\right\}$$

for any given initial state $(\mathbf{t},\mathbf{n})$ *in* $(\mathcal{T},\mathcal{N})$ *is bounded by some constant* $0 \le K_1(\mathbf{t},\mathbf{n}) < \infty$. *Let* $f : (\mathcal{T},\mathcal{N}) \to [0,\infty)$ *be a measurable function and define*

$$u_{(\mathbf{t},\mathbf{n})}(f) = \mathbb{E}_{(\mathbf{t},\mathbf{n})} \prod_{k=0}^{\tau} f(X_k) \tag{6.2}$$

for all $X_0 = (\mathbf{t},\mathbf{n}) \in (\mathcal{T},\mathcal{N})$, *so that*

$$u_{(\{0\},(1))}(f) = f\big(\{0\},(1)\big).$$

Then

$$u_{(\mathbf{t},\mathbf{n})}(f) = f\big((\mathbf{t},\mathbf{n})\big) \sum_{\substack{(\mathbf{t},\mathbf{n})\in(\mathcal{T},\mathcal{N}) \\ (\mathbf{t}',\mathbf{n}')\prec(\mathbf{t},\mathbf{n})}} q_{(\mathbf{t},\mathbf{n}),(\mathbf{t}',\mathbf{n}')} u_{(\mathbf{t}',\mathbf{n}')}(f) \tag{6.3}$$

for all $(\mathbf{t},\mathbf{n}) \in (\mathcal{T},\mathcal{N})\backslash(\{0\},(1))$. *Conversely, the unique solution of (6.3) is given by (6.2).*

Remark. If the transitions $\mathbb{Q} = (q_{(\mathbf{t}',\mathbf{n}'),(\mathbf{t},\mathbf{n})})$ are only positive if $c[(\mathbf{t}',\mathbf{n}')] < c[(\mathbf{t},\mathbf{n})]$, then

$$\tau = \inf\big\{k \ge 0 \,:\, X_k = (\{0\},(1))\big\}$$

is always bounded from above by the tree complexity of the initial state. □

Proof. Since $\tau \le c[(\mathbf{t},\mathbf{n})]$, the expected value remains finite for each initial condition. Now, compute

$$\begin{aligned}
u_{(\mathbf{t},\mathbf{n})}(f) &= \mathbb{E}_{(\mathbf{t},\mathbf{n})} \prod_{k=0}^{\tau} f(X_k) \\
&= f(\mathbf{t},\mathbf{n})\, \mathbb{E}_{(\mathbf{t},\mathbf{n})} \left[\mathbb{E}_{(\mathbf{t},\mathbf{n})} \left[\prod_{k=1}^{\tau} f(X_k) \,\middle|\, X_1 \right] \right] \\
&= f(\mathbf{t},\mathbf{n})\, \mathbb{E}\big[u_{X_1}(f)\big] \\
&= f(\mathbf{t},\mathbf{n}) \sum_{\substack{(\mathbf{t}',\mathbf{n}')\in(\mathcal{T},\mathcal{N}) \\ (\mathbf{t}',\mathbf{n}')\prec(\mathbf{t},\mathbf{n})}} q_{(\mathbf{t},\mathbf{n}),(\mathbf{t}',\mathbf{n}')} u_{(\mathbf{t}',\mathbf{n}')}(f),
\end{aligned}$$

as required. □

The result provides a simulation method for solving recursions of type (6.3): simulate a trajectory of the chain X starting at $(\mathbf{t},\mathbf{n})$ until it hits the root $(\{0\},(1))$ at time τ, compute the value of the product $\prod_{k=0}^{\tau} f(X_k)$ and repeat this many times. Averaging these values provides an unbiased and consistent estimate of $u_{(\mathbf{t},\mathbf{n})}(f)$ in terms of an approximation of the expected value $\mathbb{E}_{(\mathbf{t},\mathbf{n})} \prod_{k=0}^{\tau} f(X_k)$ by the strong law of large numbers. Lemma 6.1 states that this expectation is a solution to the recursion in question.

Corollary 6.2. *For ordered types and frequencies* $(\mathbf{t},\mathbf{n})$, *define*

$$u_{(\mathbf{t},\mathbf{n})}(f) = p^0(\mathbf{t},\mathbf{n})$$

and for $c[(\mathbf{t},\mathbf{n})] > 1$, *put*

$$\begin{aligned}
f(\mathbf{t},\mathbf{n}) = \frac{1}{r_n} \Bigg(& \sum_{\substack{k:\, n_k=1,\, x_{k0}\ distinct \\ s(\mathbf{x}_k)\neq\mathbf{x}_j\ \forall j}} r \qquad + \sum_{\substack{k:\, n_k=1, \\ x_{k0}\ distinct}} \sum_{j:s_k(\mathbf{x}_k)=\mathbf{x}_j} r(n_j+1) \\
& + \sum_{1\le i\le d:\, n_i\ge 2} \sum_{k=2}^{n_i} \binom{n}{k} \lambda_{n,k} \frac{n_i-k+1}{n-k+1} \Bigg),
\end{aligned} \tag{6.4}$$

where

$$r_n = nr + \sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k}. \tag{6.5}$$

Furthermore, let

$$u_{(\{0\},(1))}(f) = f\big(\{0\},(1)\big) = 1. \tag{6.6}$$

Consider a Markov-Chain $\{X_l = (\mathbf{t}(l), \mathbf{n}(l))\}$ *on* $(\mathcal{T},\mathcal{N})$ *with transitions*

$$(\mathbf{t},\mathbf{n}) \to \begin{cases} (s_k(\mathbf{t}),\mathbf{n}) & w.\ p. \quad \frac{r}{r_n f(\mathbf{t},\mathbf{n})} \quad \text{if} \quad n_k = 1,\ x_{k0}\ \text{dist.},\ s(\mathbf{x}_k) \neq \mathbf{x}_j\ \forall j, \\ (r_k(\mathbf{t}), r_k(\mathbf{n}+\mathbf{e}_j)) & w.\ p. \quad \frac{r(n_j+1)}{r_n f(\mathbf{t},\mathbf{n})} \quad \text{if} \quad n_k = 1,\ x_{k0}\ \text{dist.},\ s(\mathbf{x}_k) = \mathbf{x}_j, \\ (\mathbf{t}, \mathbf{n}-(k-1)\mathbf{e}_i) & w.\ p. \quad \frac{1}{r_n f(\mathbf{t},\mathbf{n})}\binom{n}{k}\lambda_{n,k}\frac{n_i-k+1}{n-k+1} \quad \text{if} \quad 2 \le k \le n_i. \end{cases}$$

Then,

$$p^0(\mathbf{t},\mathbf{n}) = \mathbb{E}_{(\mathbf{t},\mathbf{n})} \prod_{l=0}^{\tau} f(\mathbf{t}(l),\mathbf{n}(l)). \tag{6.7}$$

Proof. This is the immediate application of Lemma 6.1, noting that, as in the last remark, starting from $(\mathbf{t},\mathbf{n})$, the stopping time τ is bounded by $c[(\mathbf{t},\mathbf{n})] < \infty$. Note that one might prefer to stop at $n = 2$ in view of (4.5 – 4.6). □

Simulating independent copies and taking the average now yields an unbiased estimator of $p^0(\mathbf{t},\mathbf{n})$. Note that a similar result holds for the recursion w.r.t. $(\mathbf{t},\mathbf{a})$.

To compute $p^0(Q,\mathbf{n})$ in the unrooted case, one can either estimate each term in (4.8) using the method above, or implement an analogous Monte-Carlo scheme based on (4.9) and a Markov-Chain $\{Y(l), l = 0,1,2,\dots\}$ on the space $(\mathcal{Q},\mathcal{N})$ of unrooted trees with node multiplicities as below. Note that the complexity of a tree as defined in (6.1) does not depend on the position of the root, and is thus well-defined for unrooted trees.

Corollary 6.3. *With the notation of Subsection 4.3, put* $f((0),(1)) = 1$*, and for* $(Q,\mathbf{n}) \in (\mathcal{Q},\mathcal{N})$ *with* $c[(Q,\mathbf{n})] > 1$*, set*

$$f(Q,\mathbf{n}) = \frac{1}{r_n}\Bigg(\sum_{i:\, n_i \ge 2} \sum_{k=2}^{n_i} \binom{n}{k} \lambda_{n,k} \frac{n_i - k + 1}{n - k + 1} + \sum_{\substack{k:\, n_k = 1,\, |k| = 1 \\ k \to j,\, m_{kj} > 1}} r \quad + \sum_{\substack{k:\, n_k = 1,\, |k| = 1 \\ k \to j,\, m_{kj} = 1}} \sum_{j: s(\mathbf{x}_k) = \mathbf{x}_j} r\,(n_j + 1) \Bigg)$$

where r_n *is defined in (6.5). Consider a Markov-Chain* $\{Y_l = (Q(l),\mathbf{n}(l))\}$ *on* $(\mathcal{Q},\mathcal{N})$ *with transitions*

$$(Q,\mathbf{n}) \to \begin{cases} (Q - \mathbf{e}_{kj}, \mathbf{n}) & w.\ p. \quad \frac{r}{r_n f(\mathbf{t},\mathbf{n})} \quad \text{if} \quad n_k = 1,\ |k| = 1,\ k \to j,\ m_{kj} > 1 \\ (Q - \mathbf{e}_{kj}, r_k(\mathbf{n}+\mathbf{e}_j)) & w.\ p. \quad \frac{r(n_j+1)}{r_n f(\mathbf{t},\mathbf{n})} \quad \text{if} \quad n_k = 1,\ |k| = 1,\ k \to j,\ m_{kj} = 1 \\ (Q, \mathbf{n}-(k-1)\mathbf{e}_i) & w.\ p. \quad \frac{1}{r_n f(\mathbf{t},\mathbf{n})}\binom{n}{k}\lambda_{n,k}\frac{n_i-k+1}{n-k+1} \quad \text{if} \quad 2 \le k \le n_i. \end{cases}$$

Then, with $\tau := \min\{l : (Q(l),\mathbf{n}(l)) = ((0),(1))\}$*,*

$$p^0(Q,\mathbf{n}) = \mathbb{E}_{(Q,\mathbf{n})} \prod_{l=0}^{\tau} f(Q(l),\mathbf{n}(l)).$$

6.2 Simulation of likelihood surfaces with pre-specified driving values.

It is possible to obtain simultaneous likelihoods for a variety of values for (r, Λ) using a single realization of the Markov-chain X only.

Lemma 6.4. *Let $\{X_k, k \geq 0\}$ be a Markov chain with state space $(\mathcal{T}, \mathcal{N})$ and with transitions $\mathbb{Q} = (q_{(\mathbf{t},\mathbf{n}),(\mathbf{t}',\mathbf{n}')})$ such that the hitting time*

$$\tau = \inf\big\{k \geq 0 \,:\, X_k = (\{0\}, (1))\big\}$$

for any given initial state $(\mathbf{t}, \mathbf{n})$ in $(\mathcal{T}, \mathcal{N})$ is bounded by some constant $0 \leq K_2(\mathbf{t}, \mathbf{n}) < \infty$. Let $g : (\mathcal{T}, \mathcal{N}) \times (\mathcal{T}, \mathcal{N}) \to [0, \infty)$ be a measurable function and define

$$u_{(\mathbf{t},\mathbf{n})}(g) = \mathbb{E}_{(\mathbf{t},\mathbf{n})} \prod_{k=0}^{\tau-1} g(X_k, X_{k+1}) \tag{6.8}$$

for all $X_0 = (\mathbf{t}, \mathbf{n}) \in (\mathcal{T}, \mathcal{N}))$, with $u_{(\{0\},(1))}(g) = 1$. Then, for all $(\mathbf{t}, \mathbf{n}) \in (\mathcal{T}, \mathcal{N})\backslash(\{0\}, (1))$,

$$u_{(\mathbf{t},\mathbf{n})}(g) = \sum_{\substack{(\mathbf{t},\mathbf{n})\in(\mathcal{T},\mathcal{N}) \\ (\mathbf{t}',\mathbf{n}')\prec(\mathbf{t},\mathbf{n})}} g\big((\mathbf{t}, \mathbf{n}), (\mathbf{t}', \mathbf{n}')\big) q\big((\mathbf{t}, \mathbf{n}), (\mathbf{t}', \mathbf{n}')\big) u_{(\mathbf{t}',\mathbf{n}')}(g) \tag{6.9}$$

and this set of equations has the unique solution (6.8).

Proof. Similar to the proof of Lemma 6.1. □

We follow the spirit of Corollary 6.2 and suitably rewrite (4.3). To this end, define $p^0_{(r,\Lambda)}(\mathbf{t}, \mathbf{n})$ to be the probability of observing the unordered, labelled tree $(\mathbf{t}, \mathbf{n})$ if the underlying mutation rate is r and the genealogy is governed by a Λ-coalescent.

Corollary 6.5. *Let (r, Λ) and $(r^*, \Lambda^*) \in \mathbb{R}_+ \times \mathcal{M}([0, 1])$ be given. For ordered types and frequencies $(\mathbf{t}, \mathbf{n})$, define $f_{(r,\Lambda)}(\mathbf{t}, \mathbf{n})$ through (6.4) – (6.6) and similarly $f_{(r^*,\Lambda^*)}(\mathbf{t}, \mathbf{n})$. Consider a Markov-Chain $\{X_l = (\mathbf{t}(l), \mathbf{n}(l))\}$ on $(\mathcal{T}, \mathcal{N})$ with transitions $q_{(r^*,\Lambda^*)}$ given by*

$$(\mathbf{t}, \mathbf{n}) \to \begin{cases} (s_k(\mathbf{t}), \mathbf{n}) & w.\ p. \quad \frac{r^*}{r^*_n f_{(r*,\Lambda*)}(\mathbf{t},\mathbf{n})} \quad \text{if} \quad n_k = 1,\ x_{k0}\ \text{dist.},\ s(\mathbf{x}_k) \neq \mathbf{x}_j\ \forall j, \\ (r_k(\mathbf{t}), r_k(\mathbf{n} + \mathbf{e}_j)) & w.\ p. \quad \frac{r^*(n_j+1)}{r^*_n f_{(r*,\Lambda*)}(\mathbf{t},\mathbf{n})} \quad \text{if} \quad n_k = 1,\ x_{k0}\ \text{dist.},\ s(\mathbf{x}_k) = \mathbf{x}_j, \\ (\mathbf{t}, \mathbf{n} - (k-1)\mathbf{e}_i) & w.\ p. \quad \frac{1}{r^*_n f_{(r*,\Lambda*)}(\mathbf{t},\mathbf{n})} \binom{n}{k} \lambda^*_{n,k} \frac{n_i-k+1}{n-k+1} \quad \text{if} \quad 2 \leq k \leq n_i. \end{cases}$$

Then, defining

$$g_{(r,\Lambda),(r^*,\Lambda^*)}((\mathbf{t}, \mathbf{n}), (\mathbf{t}', \mathbf{n}')) = f_{(r,\Lambda)}(\mathbf{t}, \mathbf{n}) \frac{q_{(r,\Lambda)}\Big((\mathbf{t}, \mathbf{n}), (\mathbf{t}', \mathbf{n}')\Big)}{q_{(r^*,\Lambda^*)}\Big((\mathbf{t}, \mathbf{n}), (\mathbf{t}', \mathbf{n}')\Big)},$$

one has

$$p^0_{(r,\Lambda)}(\mathbf{t}, \mathbf{n}) = \mathbb{E}^{(r^*,\Lambda^*)}_{(\mathbf{t},\mathbf{n})} \prod_{k=0}^{\tau-1} g_{(r,\Lambda),(r^*,\Lambda^*)}(X_k, X_{k+1}), \tag{6.10}$$

provided that the parameters (r, Λ), (r^*, Λ^*) fulfil the condition

$$f_{(r,\Lambda)}(\mathbf{t}, \mathbf{n}) q_{(r,\Lambda)}\Big((\mathbf{t}, \mathbf{n}), (\mathbf{t}', \mathbf{n}')\Big) > 0 \ \Rightarrow\ q_{(r^*,\Lambda^*)}\Big((\mathbf{t}, \mathbf{n}), (\mathbf{t}', \mathbf{n}')\Big) > 0. \tag{6.11}$$

Again, this gives rise to a simulation algorithm, this time based on (r^*, Λ^*) rather than the "target" (r, Λ).

Proof. We may rewrite (4.3) as

$$p^0_{(r,\Lambda)}(\mathbf{t},\mathbf{n}) = \sum_{\substack{(\mathbf{t}',\mathbf{n}'):\\ (\mathbf{t}',\mathbf{n}')\prec(\mathbf{t},\mathbf{n})}} f_{(r,\Lambda)}(\mathbf{t},\mathbf{n}) q_{(r,\Lambda)}\Big((\mathbf{t},\mathbf{n}),(\mathbf{t}',\mathbf{n}')\Big) p^0_{(r,\Lambda)}(\mathbf{t}',\mathbf{n}') \tag{6.12}$$

for the obvious choice for $q_{(r,\Lambda)}$. Furthermore, using (6.11), (6.12) may be recast as

$$p^0_{(r,\Lambda)}(\mathbf{t},\mathbf{n}) = \sum_{\substack{(\mathbf{t}',\mathbf{n}'):\\ (\mathbf{t}',\mathbf{n}')\prec(\mathbf{t},\mathbf{n})}} f_{(r,\Lambda)}(\mathbf{t},\mathbf{n}) \frac{q_{(r,\Lambda)}\Big((\mathbf{t},\mathbf{n}),(\mathbf{t}',\mathbf{n}')\Big)}{q_{(r^*,\Lambda^*)}\Big((\mathbf{t},\mathbf{n}),(\mathbf{t}',\mathbf{n}')\Big)} q_{(r^*,\Lambda^*)}\Big((\mathbf{t},\mathbf{n}),(\mathbf{t}',\mathbf{n}')\Big) p^0_{(r,\Lambda)}(\mathbf{t}',\mathbf{n}'), \tag{6.13}$$

hence

$$p^0_{(r,\Lambda)}(\mathbf{t},\mathbf{n}) = \sum_{\substack{(\mathbf{t}',\mathbf{n}'):\\ (\mathbf{t}',\mathbf{n}')\prec(\mathbf{t},\mathbf{n})}} g_{(r,\Lambda),(r^*,\Lambda^*)}((\mathbf{t},\mathbf{n}),(\mathbf{t}',\mathbf{n}')) q_{(r^*,\Lambda^*)}\Big((\mathbf{t},\mathbf{n}),(\mathbf{t}',\mathbf{n}')\Big) p^0_{(r,\Lambda)}(\mathbf{t}',\mathbf{n}'), \tag{6.14}$$

so that Lemma 6.4 may directly be applied to equation (6.14) and the Markov chain $X_l = (\mathbf{t}(l), \mathbf{n}(l))$ with driving values r^* and $(\lambda^*_{n,k})_{2\le k\le n}$ (coming from Λ^*) and transitions as above. Thus we arrive at the representation

$$p^0_{(r,\Lambda)}(\mathbf{t},\mathbf{n}) = \mathbb{E}^{(r^*,\Lambda^*)}_{(\mathbf{t},\mathbf{n})} \prod_{k=0}^{\tau-1} g_{(r,\Lambda),(r^*,\Lambda^*)}(X_k, X_{k+1}),$$

as required. □

With this result, many estimators for $p^0_{(r,\Lambda)}(\mathbf{t},\mathbf{n})$ for various values of (r,Λ), respecting the absolute continuity condition (6.11), can be obtained by simulating just one realization of the Markov chain with driving values (r^*, Λ^*). This seems computationally much more efficient than using different driving values. However, one should be aware that one obtains correlated estimates and that the variance of the estimator for $p^0_{(r,\Lambda)}(\mathbf{t},\mathbf{n})$ depends on (r^*, Λ^*).

Remarks. 1.) The same approach can be used to extend Corollary 6.3.

2.) There are obvious improvements of this method. Combining likelihoods in approximately optimal linear combinations of the (r_i, Λ_i) leads to a further reduction in variance (see [T01] for details). More advanced techniques such as a sophisticated *importance sampling* in the spirit of [SD00] or *bridge sampling* are currently under investigation by the authors and part of an ongoing research project.

7 An urn-like algorithm for generating samples

7.1 Reversing the block-counting process

In this section, we show how the so-called *block counting process*, which keeps track of the number of blocks of a coalescent-process, can be used to derive the site frequency spectrum for an n-sample in the infinite-sites model. The time-reversal of this process will later be useful in order to obtain urn-like algorithms to produce samples under the finite- and infinite-alleles model.

Let $\{\Pi_t\}_{t\geq 0}$ be a Λ-coalescent. We denote by $\{Y_t\}_{t\geq 0}$ the corresponding *block counting process*, i.e. $Y_t = \#$ of blocks of Π_t is a continuous-time Markov chain on $\mathbb{N}$ with jump rates

$$q_{ij} = \binom{i}{i-j+1}\lambda_{i,i-j+1}, \quad i > j \geq 1.$$

The total jump rate while in i is of course $-q_{ii} = \sum_{j=1}^{i-1} q_{ij}$. We write

$$p_{ij} := \frac{q_{ij}}{-q_{ii}} \tag{7.1}$$

for the jump probabilities of the *skeleton chain*, noting that (p_{ij}) is a stochastic matrix. Note that in order to reduce i classes to j classes, an $i-j+1$-merger has to occur. Let

$$g(n,m) := \mathbb{E}_n\left[\int_0^\infty \mathbf{1}_{\{Y_s=m\}}\, ds\right] \quad \text{for } n \geq m \geq 2 \tag{7.2}$$

be the expected amount of time that Y, starting from n, spends in m. Decomposing according to the first jump of Y, we find the following set of equations for $g(n,m)$:

$$g(n,m) = \sum_{k=m}^{n-1} p_{nk} g(k,m), \quad n > m \geq 2, \tag{7.3}$$

$$g(m,m) = \frac{1}{-q_{mm}}, \quad m \geq 2. \tag{7.4}$$

Let us write $Y^{(n)}$ for the process starting from $Y_0^{(n)} = n$. Let $\tau := \inf\{t : Y_t^{(n)} = 1\}$ be the time required to come down to only one class, and let

$$\tilde{Y}_t^{(n)} := Y_{(\tau-t)-}^{(n)}, \quad 0 \leq t < \tau$$

be the time-reversed path, where we define $\tilde{Y}_t^{(n)} = \partial$, some cemetery state, when $t \geq \tau$.

Proposition 7.1 (Time-reversal)**.** *With the above definitions, $\tilde{Y}^{(n)}$ is a continuous-time Markov chain on $\{2,\ldots,n\}\cup\{\partial\}$ with jump rates*

$$\tilde{q}_{ji}^{(n)} = \frac{g(n,i)}{g(n,j)} q_{ij}, \quad j < i \leq n,$$

and $\tilde{q}_{n\partial}^{(n)} = -q_{nn}$, where $g(n,m)$ is as in (7.2). The starting distribution *of $\tilde{Y}^{(n)}$ is given by*

$$\mathbb{P}\{\tilde{Y}_0^{(n)} = k\} = g(n,k) q_{k1},$$

for each k.

Proof. The result follows from Nagasawa's Formula, see e.g. [RW87], and the observation

$$\begin{aligned}\mathbb{P}\{\tilde{Y}_0^{(n)} = k\} &= \mathbb{P}_n\big\{\tilde{Y}^{(n)} \text{ hits } k, \text{ jumps to 1 from there}\big\} \\ &= \mathbb{P}_n\big\{\tilde{Y}^{(n)} \text{ hits } k\big\}\frac{q_{k1}}{-q_{kk}} \\ &= g(n,k) q_{k1}.\end{aligned}$$

Note that unless Λ is concentrated on $\{0\}$ (Kingman-case), the dynamics of $\tilde{Y}^{(n)}$ does depend on n. □

7.2 Generating samples

The stochastic mechanism described in Section 3 allows in principle to generate random samples in a two-step procedure by first simulating a Λ-coalescent tree with real branch lengths, and then superimposing mutations along the branches at rate r. However, from a computational point of view, it is more efficient to generate the genealogy "in one pass" from the root forwards to the leaves of the coalescent tree with the help of the reversed block counting process. This is achieved by the following algorithm. We write $\#\mathbf{n} := \sum_{i=1}^{d} n_i$, and denote $\tilde{q}_k^{(n)} := -\tilde{q}_{kk}^{(n)}$.

Algorithm 1.

1) Draw K according to the law of $\tilde{Y}_0^{(n)}$, i.e. $\mathbb{P}\{K = k\} = g(n,k)q_{k1}$. Begin with the a single "ancestral type" with multiplicity K, i.e. $\mathbf{t} = (\mathbf{x}_1), \mathbf{x}_1 = 0, \mathbf{n} = ((K))$, and so $d = 1$. Set $s := 1$.

2) Given $(\mathbf{t}, \mathbf{n})$, let $k := \#\mathbf{n}$, and draw a uniform random variable U on $[0,1]$.

   - If $U \leq \frac{kr}{kr+\tilde{q}_k^{(n)}}$, then draw one type, say I, according to the present frequencies.
     - If $n_I = 1$, replace $\mathbf{x}_I$ by $(s, x_{I0}, \ldots, x_{Ij(I)})$, increase s by 1.
     - If $n_I > 1$, create new type $\mathbf{x}_{d+1} = (s, x_{I0}, \ldots, x_{Ij(I)})$, set $n_{d+1} := 1$, increase s and d each by one, decrease n_I by one.
   - If $U > \frac{kr}{kr+\tilde{q}_k^{(n)}}$, then:
     - If $\#\mathbf{n} = n$, go to 4).
     - Otherwise, pick $J \in \{k+1, \ldots, n\}$ with $\mathbb{P}\{J = j\} = \tilde{q}_{\#\mathbf{n},j}^{(n)} / \tilde{q}_{\#\mathbf{n}}^{(n)}$. Choose one of the present types i (according to their present frequency), and add $J - \#\mathbf{n}$ copies of this type, i.e. replace $n_i := n_i + J - \#\mathbf{n}$.

3) Repeat 2).

4) Finally, in order to create a numbered sample configuration with ordered types $(\mathbf{t}, \mathbf{a})$ from $(\mathbf{t}, \mathbf{n})$, pick uniformly an ordered partition $\mathbf{a}$ with $\#a_i = n_i$, $i = 1, \ldots, d$.

Remark. It is easy to adapt this algorithm to work in the finite- and infinitely-many alleles cases. In the case of parent-independent mutation, one can also use an algorithm which runs "backwards in time". Indeed, in order to simulate such a sample one follows lineages backwards. "Active" ancestral lineages are lost either by (possibly multiple) coalescence or when hitting their "defining" mutation. Details can be found in [BB07]. □

8 Illustration and discussion

8.1 Beta-coalescents

Recall that the genealogy of a sample from a large but finite population model of size N in the domain of attraction of the classical Fleming-Viot process is asymptotically described by Kingman's coalescent, if time is measured (backwards) in units of N/σ^2 generations, where σ^2 is the variance of the number of offspring per individual. Thus, if the variability of individual offspring numbers is very high, this limit may be inappropriate, and a multiple merger coalescent could be a more reasonable model for the genealogy of the sample under consideration.

The one-parameter family of multiple-merger coalescents described by (1.4) with $\alpha \in (1,2)$ can be used to describe the genealogy of a sample at a neutral locus in a scenario with (asymptotically) infinite variance of offspring distributions, and the parameter α describes the algebraic decay of the tail of the individual offspring distribution. This can be justified either by considering the time-changed genealogy of a (continuous-state) branching process of index α as in [BBC05], or more directly by a sequence of Cannings-type models as in [S03]: In each generation, let individuals generate potential offspring as in a supercritical Galton-Watson process, where the tail of the offspring distribution varies regularly with index α, i.e. the probability to have more than k children decays like Const.$\times k^{-\alpha}$. Among these, N are sampled without replacement to survive and form the next generation. Then ([S03], Thm. 4 and Lemma 13), if time is measured in units of $C_\alpha \times N^{\alpha-1}$ generations, the genealogy of a random sample is described by a Beta$(2-\alpha,\alpha)$-coalescent in the limit $N \to \infty$. In both approaches, the situation with finite variance of individual offspring numbers can be included as the boundary case $\alpha = 2$, which corresponds to Kingman's coalescent. Intuitively, smaller α corresponds to more extreme variability among offspring numbers.

Note that implicitly, the choice of α fixes the scaling of the individual mutation probability μ per generation: in a population of size N, this translates to a rate

$$r = C_\alpha N^{\alpha-1}\mu$$

with which mutations appear on the (limiting) Beta$(2-\alpha,\alpha)$-coalescent. In the case $\alpha = 2$, this is the familiar formula $r\ (= \theta/2) = 2N\mu$.

8.2 Likelihood surfaces

The Monte Carlo algorithm described by Corollaries 6.2 and 6.5 is implemented in `beta genetree`, which is, together with a technical report documenting the program, available from [B07]. By repeated calls of the program, it can be used to (approximately) compute likelihood surfaces for parametric families of coalescents. Here, we illustrate this by an application to four artificial (infinitely many sites) datasets, each of size $n = 100$, generated randomly using the algorithm described in Subsection 7.2 for a Beta$(2-\alpha,\alpha)$-coalescent, with $\alpha = 1.25, 1.5, 1.75, 2.0$ and mutation parameter $r = 2.0$. The rooted genetrees corresponding to the four datasets, drawn with the program `treepic` from Bob Griffith's `genetree` software suite, can be found in the appendix.

Figure 8.2 shows (approximately) the logarithm of the probability of observing each of the four datasets under a Beta$(2-\alpha,\alpha)$-coalescent, on which mutations appear with rate r, as a function of $(\alpha, r) \in (1,2] \times (0,4]$. Computation was based on a grid of 25×25 points in the α-r-plane, the value at each point is calculated by replacing the expected value on the right-hand side of (6.7) by an empirical average using 10^7 independent runs of the Markov chain.

Such likelihood surfaces can be used to find maximum-likelihood estimators for the parameters (α, r). Positions of maxima are given in the table below.

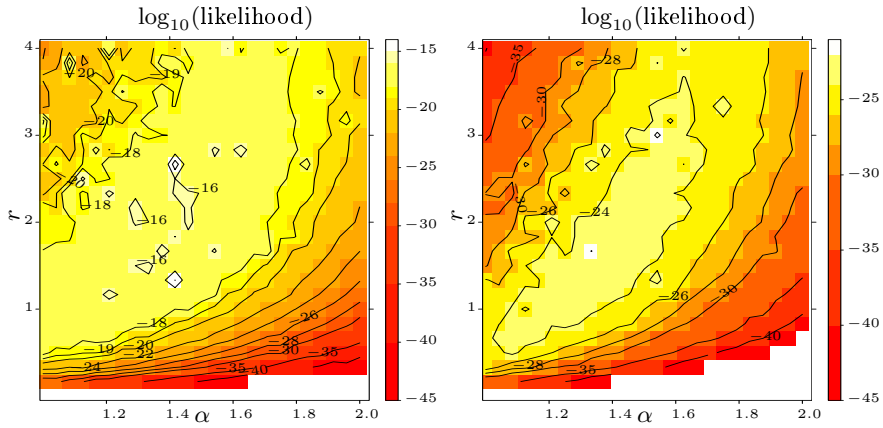
Dataset	(a)	(b)	(c)	(d)
true value of (α, r)	$(1.25, 2.0)$	$(1.5, 2.0)$	$(1.75, 2.0)$	$(2.0, 2.0)$
ML estimator $(\hat{\alpha}, \hat{r})$	$(1.4, 2.67)$	$(1.54, 3.0)^*$	$(1.63, 1.67)$	$(2.0, 2.17)$

*there is a comparable value at $(1.3, 1.67)$.

These results appear promising in that it seems possible to (at least on the principal level) recover α and r from a sample, and in particular to distinguish between Kingman- and multiple merger coalescents. Note that in the cases (a) — (c), corresponding to multiple merger behaviour, the maximal likelihood is assumed well away from the "Kingman axis" $\alpha = 2$, and the maximal value is at least two orders of magnitude larger than the highest value on the Kingman axis.

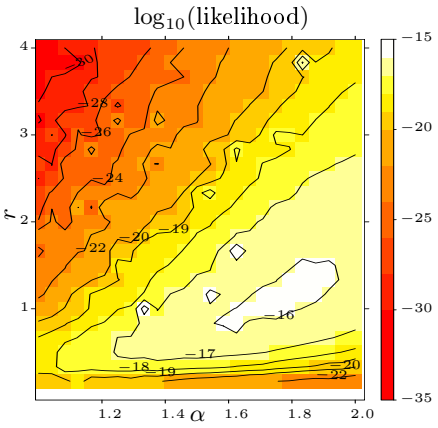
Remarks 1.) Investigation of statistical properties of these ML-estimators and comparison with estimators based on likelihoods of summary statistics (as in [EW06]) and on moment-estimators

Figure 8.1: Likelihood-surfaces for $\alpha = 1.25, 1.5, 1.75$ and 2 (Kingman case).

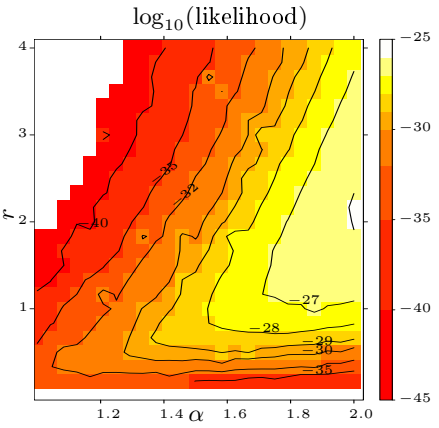


(a) 1.25

(b) 1.5



(c) 1.75



(d) 2

based on the frequency spectrum (e.g. a Watterson-like estimator of r for given Λ) will be treated in future work.

2.) The same method can obviously be applied to other families of Λ-coalescents, e.g. those described by (1.3) or (1.5). However, the choice of a class of coalescents for a given dataset should be based on biological modelling considerations. Further discussion and an application to "real" datasets will be subject of future work.

Acknowledgements We wish to thank Bob Griffiths, Jay Taylor (both Oxford) and Matthias Steinrücken (Berlin) for many most valuable discussions and in particular Matthias Steinrücken for his help with the simulations.

References

[A04] Árnason, E.: Mitochondrial Cytochrome b DNA Variation in the High-Fecundity Atlantic Cod: Trans-Atlantic Clines and Shallow Gene Genealogy, *Genetics* **166**, 1871–1885 (2004).

[BBS06] Berestycki, N.; Berestycki, J.; Schweinsberg, J.: Beta-coalescents and continuous stable random trees. *to appear in: Ann. Probab.*, (2006).

[BLG03] Bertoin, J.; Le Gall, J.-F.: Stochastic flows associated to coalescent processes. Probab. Theory Related Fields 126 (2003), no. 2, 261–288.

[BBC05] Birkner, M.; Blath, J.; Capaldo, M.; Etheridge, A.; Möhle, M.; Schweinsberg, J.; Wakolbinger, A.: Alpha-stable branching and Beta-coalescents. *Electron. J. Probab.* **10**, 303–325, (2005).

[B07] `http://www.wias-berlin.de/people/birkner/bgt`

[BB07] Birkner, M.; Blath, J.: Measure-valued diffusions, general coalescents and population genetic inference. To appear in: Trends in Stochastic Analysis – a Festschrift for Heinrich von Weizsäcker, (2007).

[BBB94] Boom, J. D. G.; Boulding, E. G.; Beckenbach, A. T.: Mitochondrial DNA variation in introduced populations of Pacific oyster, Crassostrea gigas, in British Columbia. *Can. J. Fish. Aquat. Sci.* 51:1608Ũ1614, (1994).

[Bo06] Bovier, A.: *Statistical mechanics of disordered systems*, Cambridge University Press, (2006).

[C74] Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. *Adv. Appl. Prob.* **6**, 260–290, (1974).

[C75] Cannings, C.: The latent roots of certain Markov chains arising in genetics: a new approach, II. Further haploid models. *Adv. Appl. Prob.* **7**, 264–282, (1975).

[D93] Dawson, D.: Lecture Notes, Ecole d'Eté de Probabilités de Saint-Flour XXI, Berlin, Springer, (1993).

[DIG04a] De Iorio, M. and Griffiths, R. C.: Importance sampling on coalescent histories I. *Adv. Appl. Prob.* **36**, 417-433, (2004).

[DK99] Donnelly, P.; Kurtz, T.: Particle representations for measure-valued population models. Ann. Probab. 27, no. 1, 166–205, (1999)

[DS05] Durrett, R.; Schweinsberg, J.: A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stoch. Proc. Appl.* **115**, 1628–1657 (2005)

[EW06] Eldon, B.; Wakeley, J.: Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* **172**, 2621–2633 (2006).

[E79] Ewens, W.J.: Mathematical population genetics. Berlin: Springer-Verlag, (1979).

[EG87] Ethier, S.; Griffiths, R.C.: The infinitely-many-sites model as a measure-valued diffusion. *Ann. Probab.* **15**, no. 2, 515–545, (1987).

[EK86] Ethier, S.; Kurtz, T.: Markov Processes: Characterization and Convergence. Wiley, (1986).

[EK93] Ethier, S.; Kurtz, T.: Fleming-Viot processes in population genetics, SIAM J. Control Optim. 31, no. 2, 345–386, (1993).

[FKY99] Felsenstein, J., Kuhner, M. K., Yamato, J. and Beerli, P.: Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. IMS Lecture Notes - Monograph Series, **33**, 163–185, (1999).

[G89] Griffiths, R. C.: Genealogical-tree probabilities in the infinitely-many-site model. *J. Math. Biol.* **27**, no. 6, 667–680, (1989).

[GT94a] Griffiths, R.C. and Tavaré, S.: Simulating probability distributions in the coalescent. *Theor. Pop. Biol.* **46**, 131–159, (1994).

[GT94b] Griffiths, R.C. and Tavaré, S.: Ancestral Inference in population genetics. *Statistical Science* **9**, 307–319, (1994).

[GT94c] Griffiths, R.C. and Tavaré, S.: Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society London*, Series B, **344**, 403–410, (1994).

[GT95] Griffiths, R.C. and Tavaré, S.: Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Math. Biosciences* **127**, 77–98 (1995).

[GT96a] Griffiths, R.C. and Tavaré, S.: Monte Carlo inference methods in population genetics. Monte Carlo and quasi-Monte Carlo methods. *Math. Comput. Modeling* 23 (1996), no. 8-9, 141–158.

[GT96b] Griffiths, R.C. and Tavaré, S.: Markov chain inference methods in population genetics. *Math. Comput. Modeling* **23**, 8/9, 141-158, (1996).

[GT97] Griffiths, R.C. and Tavaré, S.: Computational Methods for the coalescent. Progress in Population Genetics and Human Evolution, 165–182, Springer, (1997).

[G91] Gusfield, D.: Efficient algorithms for inferring evolutionary trees. *Networks* **21**, no. 1, 19–28 (1991).

[H90] Hudson, R.R.: Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1–44, (1990).

[HSW05] Hein, J.; Schierup, M. H.; Wiuf, C.: *Gene Genealogies, Variation and Evolution – A Primer in Coalescent Theory.* Oxford University Press, 2005.

[K69] KIMURA, M.: The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. *Genetics* **61**, 893–903, (1969).

[K82] KINGMAN, J. F. C.: The coalescent. *Stoch. Proc. Appl.* **13**, 235–248, (1982).

[M06] MÖHLE, M.: On sampling distributions for coalescent processes with simultaneous multiple collisions. *Bernoulli* **12**, 35–53, (2006).

[MS01] MÖHLE, M; SAGITOV, S.: A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.* **29**, 1547–1562, (2001).

[N01] NORDBORG, M.: Coalescent Theory. In Balding, Bishop, and Cannings, eds. *Handbook of Statistical genetics,* 179–208. Wiley, (2001).

[P99] PITMAN, J.: Coalescents with multiple collisions. *Ann. Probab.* **27** (4), 1870-1902, (1999).

[RW87] ROGERS, L.C.G.; WILLIAMS, D.: Diffusions, Markov Processes and Martingales. Vol. 1, 2nd ed., Wiley, (1994)

[S99] SAGITOV, S.: The general coalescent with asynchronous mergers of ancestral lines. *J. Appl. Probab.* **36** (4) 1116–1125, (1999).

[S00] SCHWEINSBERG, J.: A necessary and sufficient condition for the Λ-coalescent to come down from infinity. *Electron. Comm. Probab.* **5**, 1–11, (2000).

[S03] SCHWEINSBERG, J.: Coalescent processes obtained from supercritical Galton-Watson processes. *Stoch. Proc. Appl.* **106**, 107–139, (2003).

[SD00] STEPHENS, M; DONNELLY, P.: Inference in molecular population genetics. *J. Roy. Stat. Soc.* B. **62**, 605–655 (2000).

[SK88] STUDIER, J.; KEPPLER, K.: A note on the neighbor-joining algorithm of Saitou and Nei, *Mol. Biol. Evol.* **5**, 729–731 (1988).

[T01] TAVARÉ, S.: Ancestral Inference in Population Genetics. Springer Lecture Notes **1837**, 2001.

[W06] WAKELEY, J.: *Coalescent Theory.* To appear (2007).

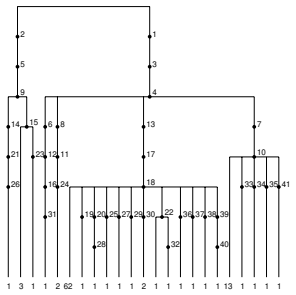
[WSS77] WATERMAN, M.S.; SMITH, T.F.; SINGH, M.; BEYER, W.A.: Additive Evolutionary Trees. *J. Theor. Bio.* **64**, 199-213 (1977).

[W75] WATTERSON, G. A.: On the number of segregating sites in genetical models without recombination. *Theoret. Popul. Biol.* **10**, 256–276 (1975).

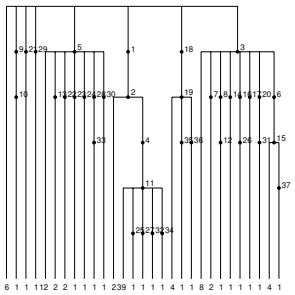
9 Appendix

9.1 Underlying genetrees

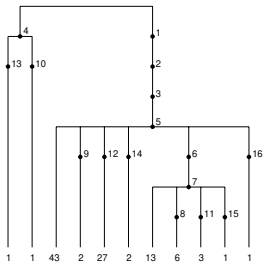
Figure 9.1: Genetrees corresponding to the four datasets analysed in Subsection 8.2 ($\alpha = 1.25, 1.5, 1.75$ and 2 (Kingman case)).



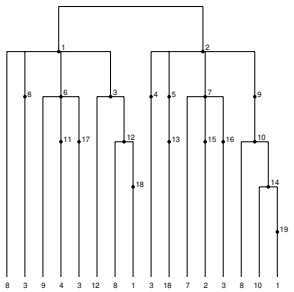
(a) 1.25



(b) 1.5



(c) 1.75



(d) 2