COMPUTATIONAL INFERENCE BEYOND KINGMAN'S COALESCENT

JERE KOSKELA, *

PAUL JENKINS ** *** AND

DARIO SPANÒ,** **** *University of Warwick*

Abstract

Full likelihood inference under Kingman's coalescent is a computationally challenging problem to which importance sampling (IS) and the product of approximate conditionals (PAC) methods have been applied successfully. Both methods can be expressed in terms of families of intractable conditional sampling distributions (CSDs), and rely on principled approximations for accurate inference. Recently, more general Λ- and Ξ-coalescents have been observed to provide better modelling fits to some genetic data sets. We derive families of approximate CSDs for finite sites Λ- and Ξ-coalescents, and use them to obtain 'approximately optimal' IS and PAC algorithms for Λ-coalescents, yielding substantial gains in efficiency over existing methods.

Keywords: Lambda-coalescent; xi-coalescent; product of approximate conditionals; importance sampling; conditional sampling distribution; population genetics

2010 Mathematics Subject Classification: Primary 60G09
Secondary 93E10; 92D25

1. Introduction

Importance sampling (IS) has a well-established role in population genetic inference as a means of approximating likelihoods. In this context the method was introduced by Griffiths and Tavaré, who derived a recursion for quantities of interest under Kingman's coalescent [25] and simulated a Markov chain to approximate its solution [17]–[20]. Their approach was identified as IS by Felsenstein *et al.* [14], which led Stephens and Donnelly to derive the optimal proposal distribution in terms of a family of conditional sampling distributions (CSDs) [41]. The CSDs are inaccessible in general, but the authors introduced an approximation which yielded dramatic improvements in efficiency and accuracy of IS algorithms. In addition, IS has been investigated and applied to genetic problems such as demographic and other parameter inferences in [10], [13], [15], [16], [21], [22] and [24].

Approximating the CSDs for various generalisations of Kingman's coalescent has received plenty of attention, both as a means of deriving approximations to the optimal importance sampling algorithm and due to the product of approximate conditionals (PAC) method introduced in [26]. De Iorio and Griffiths derived an approximation to finite alleles CSDs based on the Fleming–Viot generator [8], [9] and Paul and Song provided a genealogical interpretation

and included crossover recombination [31]. Further approximations based on hidden Markov models have been obtained in [32] and [40], and applied in [38].

Kingman's coalescent only permits binary mergers of ancestral lineages. The Λ-coalescents, introduced by Pitman [33] and Sagitov [34], generalise Kingman's coalescent by permitting multiple lineages to merge in one event. The rate of coalescence of any k out of n lineages is given by

$$\lambda_{n,k} := \int_0^1 r^k (1-r)^{n-k} \frac{1}{r^2} \Lambda(\mathrm{d}r)$$

for some finite measure Λ on [0, 1], which can be taken to be a probability measure without loss of generality. Popular choices of Λ include $\Lambda = \delta_0$, which corresponds to Kingman's coalescent, $\Lambda = \delta_1$ leading to star–shaped genealogies, $\Lambda = (2/(2+\psi^2))\delta_0 + (\psi^2/(2+\psi^2))\delta_\psi$, where $\psi \in (0, 1]$ [12] and $\Lambda = \text{Beta}(2 - \alpha, \alpha)$, where $\alpha \in (1, 2)$ [37]. See [3] for a review.

Investigations by Boom *et al.* [7], Árnason [1], Eldon and Wakeley [12], and Birkner and Blath [2] have concluded that Λ-coalescents can provide better descriptions of some populations than Kingman's coalescent, particularly among marine species. Thus, similar strategies of inference have been developed for them. An analogue of the Griffiths–Tavaré recursion (see (2.1) in Section 2) for Λ-coalescents was derived by Birkner and Blath in [2]. In a subsequent paper [5] the authors characterised the optimal IS proposal distribution in terms of a family of Green's functions related to the time-reversal of the Λ-coalescent, and used their representation to obtain an approximately optimal algorithm for the infinite sites model of mutation. Steinrücken *et al.* [39] contains a detailed discussion of inference under Beta-coalescents and their applicability to marine populations.

The Λ-coalescent family allows for multiple mergers, but only permits one merger at a time. They are generalised further by the Ξ-coalescents, which permit any number of simultaneous, multiple mergers. The Ξ-coalescents were introduced by Möhle and Sagitov [29] and Schweinsberg [36] and can be expressed in terms of a finite measure Ξ (again, a probability measure without loss of generality) on the infinite simplex

$$\Delta = \left\{ \boldsymbol{r} = (r_1, r_2, \dots) \in [0, 1]^{\mathbb{N}} : \sum_{i=1}^{\infty} r_i \leq 1 \right\}.$$

We denote by $\lambda_{n;k_1,\dots,k_p;s}$ the rate of jumps involving $p \geq 1$ mergers with sizes $k_1, \dots, k_p$, with $s = n - \sum_{i=1}^{p} k_i$ lineages not participating in any merger. The total number of lineages before the mergers is denoted by n. This rate is given as

$$\lambda_{n;k_1,\dots,k_p;s} = \int_{\Delta} \sum_{l=0}^{s} \binom{s}{l} \sum_{i_1 \in \mathbb{N}} \cdots \sum_{i_{p+l} \in \mathbb{N}} r_{i_1}^{k_1} \cdots r_{i_p}^{k_p} r_{i_{p+1}} \cdots r_{i_{p+l}} \frac{(1 - \sum_{i=1}^{\infty} r_i)^{s-l}}{\sum_{i=1}^{\infty} r_i^2} \Xi(\mathrm{d}\boldsymbol{r}).$$

The Ξ-coalescents have also been used to model genealogies of marine organisms [35] and populations undergoing mass extinctions [42], although the question of which measures Ξ are biologically relevant remains open. Note that if Ξ assigns full mass to the set $\{\boldsymbol{r} \in \Delta : r_2 = r_3 = \cdots = 0\}$ the resulting process is a Λ-coalescent.

In this paper we characterise the optimal IS proposal distribution for finite sites, finite alleles Λ- and Ξ-coalescents in terms of their respective families of CSDs, and derive principled approximations to the CSDs for both coalescent families. The rest of the paper is laid out as follows. In Section 2 we provide a heuristic description of Λ-coalescents and derive their optimal proposal distributions. In Section 3 we present principled derivations for approximate

Λ-coalescent CSDs, based on the finite alleles Λ-Fleming–Viot generator and on genealogical considerations. In Section 4 we present simulation studies using both IS and PAC algorithms on data sets simulated from a finite sites Λ-coalescent. In Section 5 we generalise the IS algorithm for the Ξ-coalescent family, and derive the analogous approximate CSDs. We conclude with a discussion in Section 6.

2. The Λ-coalescent and its optimal proposal distributions

In the notation of [31], let $L = \{1, \ldots, |L|\}$ be a set of loci, E_l be the finite set of alleles at locus $l \in L$, θ_l be the mutation rate at locus $l \in L$ and $P^{(l)}$ be a family of stochastic matrices giving transition probabilities of mutations at locus $l \in L$. Let $\theta := \sum_{l \in L} \theta_l$ be the total mutation rate, $\mathcal{H} := E_1 \times \cdots \times E_{|L|}$ be the set of haplotypes and

$$\Delta_{\mathcal{H}} = \left\{ \boldsymbol{x} = (x_1, \ldots, x_{|\mathcal{H}|}) \in [0, 1]^{|\mathcal{H}|} : \sum_{h \in \mathcal{H}} x_h = 1 \right\}$$

be the space of probability vectors of allele frequencies. Define a sample by $\boldsymbol{n} = (n_h)_{h \in \mathcal{H}} \in \mathbb{N}^{|\mathcal{H}|}$ and let $n := \sum_{h \in \mathcal{H}} n_h$. Let $\boldsymbol{e}_h$ be the canonical unit vector with a 1 in position h and 0 elsewhere. For $l \in L$ and $h \in \mathcal{H}$ let $h[l]$ define the allele at locus l of haplotype h. Finally, for $a \in E_l$ let $S_l^a(h)$ be the haplotype obtained from h by overwriting locus l by allele a.

The dynamics of a finite sample of individuals from the stationary Λ-Fleming–Viot process under the finite alleles model of mutation can be described as follows. For a rigorous account, see [11].

Consider a sample of n typed lineages and associate to each lineage a unique level from $\{1, \ldots, n\}$. Let Π be a Poisson process on $\mathbb{R}_+ \times [0, 1]$ with rate $\mathrm{d}t \otimes r^{-2}\Lambda(\mathrm{d}r)$. At each $(t, r) \in \Pi$ every lineage flips a coin with success probability r, and all successful lineages 'look down' and copy the type of the participating lineage with the lowest level. Independently, the type of each lineage mutates at locus l with rate θ_l and jumps drawn from $P^{(l)}$. This particle system embeds the Λ-coalescent, with coalescence events traced along the look-down-and-copy jumps.

Let $(H_i)_{i=0}^{-T}$ define the sequence of type configurations of the ancestral sample after i events, whether they be mutations or coalescences. H_0 contains the frequency counts of the observed sample, H_{-T} is the type of the most recent common ancestor (MRCA) and the other H_is correspond to intermediate states along the ancestral tree. Note that the likelihood $\mathbb{P}(H_0)$ can be written as

$$\mathbb{P}(H_0) = \sum_{\mathcal{A}} \mathbb{P}\{H_0 \mid A\}\mathbb{P}(A),$$

where $\mathcal{A}$ is the space of possible ancestries and $\mathbb{P}\{H_0 \mid A\} = 1$ if the leaves of A are compatible with H_0, and 0 otherwise. These ancestries can be decomposed into a sequence of updates as above to give

$$\mathbb{P}(H_0) = \sum_{H_0} \sum_{H_{-1}} \cdots \sum_{H_{-T}} \mathbb{P}\{H_0 \mid A\} \prod_{i=-1}^{-T} \mathbb{P}\{H_{i+1} \mid H_i\}\mathbb{P}(H_{-T}), \tag{2.1}$$

where $\mathbb{P}(H_{-T})$ is the invariant distribution of the mutation operator obtained by viewing mutation as a mixture Markov chain on $\mathcal{H}$ with weights $\{\theta_l/\theta\}_{l \in L}$ and mixture components

$\{P^{(l)}\}_{l \in L}$, and

$$\mathbb{P}\{H_{i+1} \mid H_i\}$$

$$= \begin{cases} \dfrac{\theta_l}{n\theta - q_{n_i n_i}}((n_i)_{S_l^a(h)} + 1 - \delta_{ah[l]})P_{ah[l]}^{(l)} & \text{if } H_{i+1} = H_i - e_{S_l^a(h)} + e_h, \\[2ex] \dbinom{n_i}{k} \dfrac{\lambda_{n_i,k}}{n\theta - q_{n_i n_i}} \dfrac{(n_i)_h - k + 1}{n_i - k + 1} & \text{if } H_{i+1} = H_i + (k-1)e_h, \end{cases}$$

where $-q_{nn} = \sum_{j=1}^{n-1} \binom{n}{n-j+1}\lambda_{n,n-j+1}$ is the total rate of coalescence of n untyped lineages, n_i is the number of lineages in H_i, and $(n_i)_h$ is the number of lineages of type h in H_i. See [2] for a detailed derivation.

As with Kingman's coalescent, (2.1) can be approximated by sampling N independent ancestors from the stationary distribution of the mutation mechanism, generating an ancestral tree A from each ancestor and counting the proportion of trees with leaves that are compatible with H_0. However, obtaining a nonzero estimator with reasonable probability requires a prohibitively large number of simulations as likelihoods of $\mathcal{O}(10^{-10})$, or much smaller still, are typical even among modest data sets. A better approach is to start with the data, propose mutations and coalescences backwards in time until the MRCA is reached, and, thus, ensure every simulated tree is compatible with the observed leaves. This yields

$$\mathbb{P}(H_0) = \sum_{H_{-T}} \cdots \sum_{H_{-1}} \prod_{i=-1}^{-T} \frac{\mathbb{P}\{H_{i+1} \mid H_i\}}{\mathbb{Q}(H_i \mid H_{i+1})} \mathbb{P}(H_{-T}) \mathbb{Q}(H_i \mid H_{i+1})$$

$$= \mathbb{E}\left\{ \prod_{i=-1}^{-T} \frac{\mathbb{P}\{H_{i+1} \mid H_i\}}{\mathbb{Q}(H_i \mid H_{i+1})} \mathbb{P}(H_{-T}) \right\}, \tag{2.2}$$

where $\mathbb{Q}(\cdot \mid H_{i+1})$ is an arbitrary proposal distribution satisfying mild support conditions, and the expectation is with respect to $\otimes_{i=-1}^{-T} \mathbb{Q}(H_i \mid H_{i+1})$. Equation (2.2) can be approximated by the IS estimator

$$\widehat{p}(H_0) = \frac{1}{N} \sum_{j=1}^{N} \prod_{i=-1}^{-T_j} \frac{\mathbb{P}\{H_{i+1}^{(j)} \mid H_i^{(j)}\}}{\mathbb{Q}(H_i^{(j)} \mid H_{i+1}^{(j)})} \mathbb{P}(H_{-T_j}^{(j)}),$$

where $\{\{H_i^{(j)}\}_{i=0}^{-T_j}\}_{j=1}^{N}$ is an independent and identically distributed (i.i.d.) sample of sequentially constructed coalescent trees from the distributions $\mathbb{Q}(H_i \mid H_{i+1})$.

The following theorem is a Λ-coalescent analogue of Theorem 1 of [41]. A similar result, using ratios of Green's functions instead of CSDs, is presented in Lemma 2.2 of [5].

Theorem 2.1. *Let $\pi(m \mid n)$ define the sampling distribution of the next m individuals, given the types of the first n from a population evolving according to the stationary Λ-Fleming–Viot process. Then the optimal proposal distributions $\mathbb{Q}^*$ are given by*

$$\mathbb{Q}^*(H_i \mid H_{i+1}) \propto \begin{cases} (n_{i+1})_h \theta_l \dfrac{\pi(e_{S_l^a(h)} \mid H_{i+1} - e_h)}{\pi(e_h \mid H_{i+1} - e_h)} P_{ah[l]}^{(l)} & \text{if } H_i = H_{i+1} - e_h + e_{S_l^a(h)}, \\[3ex] \dfrac{\binom{(n_{i+1})_h}{k}\lambda_{n_{i+1},k}}{\pi((k-1)e_h \mid H_{i+1} - (k-1)e_h)} & \text{if } H_i = H_{i+1} - (k-1)e_h, \end{cases}$$

where the first term ranges over all possible mutations for all haplotypes present in the sample, and the second over all present haplotypes and $k \in \{2, \ldots (n_{i+1})_h\}$.

Proof. The argument giving the mutation term is identical to that in Theorem 1 of [41] and is omitted.

For the coalescence term suppose the n lineages evolve according to the lookdown construction of [11], and define the types of the n particles at time t by $D_n(t) = (h_1, \ldots, h_n)$. Define Υ_k as the event that in the last δ units of time there was a merger involving lineages $n-k+1, n-k+2, \ldots, n$. To simplify the presentation let $h_{\{i:\,j\}} := (h_i, h_{i+1}, \ldots, h_{j-1}, h_j)$. Then

$$
\begin{aligned}
&\mathbb{P}\{\Upsilon_k \mid D_n(t) = (h_{\{1:\,n-k\}}, h, \ldots, h)\} \\
&= \sum_{g_{\{2:\,k\}} \in \mathcal{H}^{k-1}} \frac{\mathbb{P}\{\Upsilon_k \cap D_n(t-\delta) = (h_{\{1:\,n-k\}}, h, g_{\{2:\,k\}}) \cap D_n(t) = (h_{\{1:\,n-k\}}, h, \ldots, h)\}}{\mathbb{P}\{D_n(t) = (h_{\{1:\,n-k\}}, h, \ldots, h)\}} \\
&= \sum_{g_{\{2:\,k\}} \in \mathcal{H}^{k-1}} \frac{\mathbb{P}\{D_n(t-\delta) = (h_{\{1:\,n-k\}}, h, g_{\{2:\,k\}})\}\delta\lambda_{n,k}}{\mathbb{P}\{D_n(t) = (h_{\{1:\,n-k\}}, h, \ldots, h)\}} + o(\delta) \\
&= \frac{\delta\lambda_{n,k}}{\pi((k-1)e_h \mid D_n(t) - (k-1)e_h)} + o(\delta).
\end{aligned}
$$

By exchangeability every set of k lineages coalesces at this same rate, so the total rate is obtained by multiplying by $\binom{n_h}{k}$.

Remark 2.1. It is tempting to simplify the situation further by decomposing

$$
\pi((k-1)e_h \mid \boldsymbol{n} - (k-1)e_h) = \prod_{j=0}^{k-2} \pi(e_h \mid \boldsymbol{n} - (k-1+j)e_h)
$$

and, thus, requiring only univariate CSDs. In general a decomposition like this requires exchangeability, which the CSDs satisfy but typically approximations do not. However, in the Λ-coalescent setting the argument being decomposed will always consist of only one type of allele. Permuting lineages which feature only in the sample being conditioned on does not affect the outcome even for nonexchangeable families of distributions, so in this particular context univariate CSDs are sufficient. Note that this will not be the case for Ξ-coalescents, since simultaneous mergers of several types of lineages is permitted.

3. Approximating the Λ-coalescent CSDs

An approximation to the CSDs for Kingman's coalescent was derived in [8] by noting that the Fleming–Viot generator can be written component-wise as $\mathcal{L} = \sum_{h \in \mathcal{H}} \mathcal{L}_h$ and then assuming that there exists a probability measure, and an expectation $\widehat{\mathbb{E}}$ with respect to that measure, such that the standard stationarity condition $\mathbb{E}\{\mathcal{L}f(\boldsymbol{X})\} = 0$ holds component-wise,

$$
\widehat{\mathbb{E}}\{\mathcal{L}_h f(\boldsymbol{X})\} = 0 \text{ for every } h \in \mathcal{H} \text{ and } f \in C^2(\Delta_{\mathcal{H}}).
$$

Substituting the probability of an ordered sample $q(\boldsymbol{n} \mid \boldsymbol{x}) = \prod_{h \in \mathcal{H}} x_h^{n_h}$ yields a recursion whose solution is defined as the approximate CSD. The same argument can be applied to the Λ-Fleming–Viot process to define approximate CSDs for the Λ-coalescent.

Theorem 3.1. *Let* $\hat{\pi}(\boldsymbol{m} \mid \boldsymbol{n})$ *define the approximate* Λ-*coalescent CSD as defined above. It solves the following recursion:*

$$
m\left[\frac{\Lambda(\{0\})(n+m-1)}{2} + \theta + \frac{1}{n+m}\sum_{k=2}^{n+m}\binom{n+m}{k}\lambda_{n+m,k}\right]\hat{\pi}(\boldsymbol{m} \mid \boldsymbol{n})
$$

$$
= \sum_{h\in\mathcal{H}} m_h\left[\frac{\Lambda(\{0\})(n_h+m_h-1)}{2}\hat{\pi}(\boldsymbol{m}-\boldsymbol{e}_h \mid \boldsymbol{n})\right.
$$

$$
+ \sum_{l\in L}\theta_l\sum_{a\in E_l}P_{ah[l]}^{(l)}\hat{\pi}(\boldsymbol{m}-\boldsymbol{e}_h+\boldsymbol{e}_{S_l^a(h)} \mid \boldsymbol{n})
$$

$$
+ \frac{1}{n_h+m_h}\left\{\sum_{k=2}^{m_h+1}\binom{n_h+m_h}{k}\lambda_{n+m,k}\hat{\pi}(\boldsymbol{m}-(k-1)\boldsymbol{e}_h \mid \boldsymbol{n})\right.
$$

$$
+ \sum_{k=m_h+2}^{n_h+m_h}\binom{n_h+m_h}{k}\lambda_{n+m,k}
$$

$$
\left.\left.\times\frac{\hat{\pi}(\boldsymbol{m}-m_h\boldsymbol{e}_h \mid \boldsymbol{n}-(k-m_h-1)\boldsymbol{e}_h)}{\hat{\pi}((k-m_h-1)\boldsymbol{e}_h \mid \boldsymbol{n}-(k-m_h-1)\boldsymbol{e}_h)}\right\}\right].
\tag{3.1}
$$

Proof. The generator of the Λ-Fleming–Viot jump-diffusion can be written as

$$
\mathcal{L}f(\boldsymbol{x}) = \sum_{h\in\mathcal{H}}\frac{\Lambda(\{0\})x_h}{2}\sum_{h'\in\mathcal{H}}(\delta_{hh'}-x_{h'})\frac{\partial^2}{\partial x_h\partial x_{h'}}f(\boldsymbol{x})
$$

$$
+ \sum_{h\in\mathcal{H}}\sum_{l\in L}\theta_l\sum_{a\in E_l}x_{S_l^a(h)}(P_{ah[l]}^{(l)}-\delta_{ah[l]})\frac{\partial}{\partial x_h}f(\boldsymbol{x})
$$

$$
+ \sum_{h\in\mathcal{H}}x_h\int_{(0,1]}\{f((1-r)\boldsymbol{x}+r\boldsymbol{e}_h)-f(\boldsymbol{x})\}r^{-2}\Lambda(\mathrm{d}r)
$$

$$
=: \sum_{h\in\mathcal{H}}\mathcal{L}_hf(\boldsymbol{x}).
\tag{3.2}
$$

Substituting $q(\boldsymbol{n} \mid \boldsymbol{x})$ yields the following three terms on the right-hand side:

$$
\sum_{h\in\mathcal{H}}\frac{\Lambda(\{0\})n_h}{2}\left[(n_h-1)q(\boldsymbol{n}+\boldsymbol{e}_h \mid \boldsymbol{x})-\sum_{h'\in\mathcal{H}}(n_{h'}-\delta_{hh'})q(\boldsymbol{n} \mid \boldsymbol{x})\right]
\tag{3.3}
$$

$$
+ \sum_{h\in\mathcal{H}}n_h\sum_{l\in L}\theta_l\left[\sum_{a\in E_l}P_{ah[l]}^{(l)}q(\boldsymbol{n}-\boldsymbol{e}_h+\boldsymbol{e}_{S_l^a(h)} \mid \boldsymbol{x})-q(\boldsymbol{n} \mid \boldsymbol{x})\right]
\tag{3.4}
$$

$$
+ \int_{(0,1]}\left\{\sum_{h\in\mathcal{H}}\sum_{k=0}^{n_h}\binom{n_h}{k}r^k(1-r)^{n-k}q(\boldsymbol{n}-(k-1)\boldsymbol{e}_h \mid \boldsymbol{x})\right.
$$

$$
\left.- \sum_{k=0}^{n}\binom{n}{k}r^k(1-r)^{n-k}q(\boldsymbol{n} \mid \boldsymbol{x})\right\}r^{-2}\Lambda(\mathrm{d}r).
$$

The $k = 0$ terms inside the integral cancel because $\sum_{h \in \mathcal{H}} x_h = 1$ and the $k = 1$ terms cancel because $\sum_{h \in \mathcal{H}} n_h = n$, which means the third summand can be written as

$$\sum_{h \in \mathcal{H}} \left\{ \sum_{k=2}^{n_h} \binom{n_h}{k} \lambda_{n,k} q(\boldsymbol{n} - (k-1)\boldsymbol{e}_h \mid \boldsymbol{x}) - \frac{n_h}{n} \sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k} q(\boldsymbol{n} \mid \boldsymbol{x}) \right\}. \tag{3.5}$$

Substituting (3.3), (3.4), and (3.5) into (3.2) and rearranging, we obtain

$$\sum_{h \in \mathcal{H}} \left[\frac{\Lambda(\{0\})(n-1)}{2} + \theta + \frac{1}{n} \sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k} \right] q(\boldsymbol{n} \mid \boldsymbol{x})$$

$$= \sum_{h \in \mathcal{H}} \left\{ \frac{\Lambda(\{0\})(n_h - 1)}{2} q(\boldsymbol{n} - \boldsymbol{e}_h \mid \boldsymbol{x}) + \sum_{l \in L} \theta_l \sum_{a \in E_l} P_{ah[l]}^{(l)} q(\boldsymbol{n} - \boldsymbol{e}_h + \boldsymbol{e}_{S_l^a(h)} \mid \boldsymbol{x}) \right.$$

$$\left. + \frac{1}{n_h} \sum_{k=2}^{n_h} \binom{n_h}{k} \lambda_{n,k} q(\boldsymbol{n} - (k-1)\boldsymbol{e}_h \mid \boldsymbol{x}) \right\}. \tag{3.6}$$

The component-wise vanishing property implies

$$\widehat{\mathbb{E}} \left\{ \sum_{h \in \mathcal{H}} m_h \mathcal{L}_h q(\boldsymbol{n} \mid \boldsymbol{X}) \right\} = \sum_{h \in \mathcal{H}} m_h \widehat{\mathbb{E}} \{ \mathcal{L}_h q(\boldsymbol{n} \mid \boldsymbol{X}) \} = 0$$

so that (3.6) becomes

$$m \left[\frac{\Lambda(\{0\})(n-1)}{2} + \theta + \frac{1}{n} \sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k} \right] \widehat{\mathbb{E}} \{ q(\boldsymbol{n} \mid \boldsymbol{X}) \}$$

$$= \sum_{h \in \mathcal{H}} m_h \left\{ \frac{\Lambda(\{0\})(n_h - 1)}{2} \widehat{\mathbb{E}} \{ q(\boldsymbol{n} - \boldsymbol{e}_h \mid \boldsymbol{X}) \} \right.$$

$$+ \sum_{l \in L} \theta_l \sum_{a \in E_l} P_{ah[l]}^{(l)} \widehat{\mathbb{E}} \{ q(\boldsymbol{n} - \boldsymbol{e}_h + \boldsymbol{e}_{S_l^a(h)} \mid \boldsymbol{X}) \}$$

$$\left. + \frac{1}{n_h} \sum_{k=2}^{n_h} \binom{n_h}{k} \lambda_{n,k} \widehat{\mathbb{E}} \{ q(\boldsymbol{n} - (k-1)\boldsymbol{e}_h \mid \boldsymbol{X}) \} \right\}.$$

Note that $\pi(\boldsymbol{m} \mid \boldsymbol{n}) = \mathbb{E}\{q(\boldsymbol{n}+\boldsymbol{m} \mid \boldsymbol{X})\}/\mathbb{E}\{q(\boldsymbol{n} \mid \boldsymbol{X})\}$ so that substituting $\boldsymbol{n} \mapsto \boldsymbol{n}+\boldsymbol{m}$, assuming that $\mathbb{E} = \widehat{\mathbb{E}}$ and dividing by $\mathbb{E}\{q(\boldsymbol{n} \mid \boldsymbol{X})\}$ provides the desired recursion.

Corollary 3.1. *The univariate approximate CSDs $\hat{\pi}(\boldsymbol{e}_h \mid \boldsymbol{n})$ satisfy*

$$\left[\frac{\Lambda(\{0\})n}{2} + \theta + \frac{1}{n+1} \sum_{k=2}^{n+1} \binom{n+1}{k} \lambda_{n+1,k} \right] \hat{\pi}(\boldsymbol{e}_h \mid \boldsymbol{n})$$

$$= \frac{n_h}{2} (\Lambda(\{0\}) + \lambda_{n+1,2}) + \sum_{l \in L} \theta_l \sum_{a \in E_l} P_{ah[l]}^{(l)} \hat{\pi}(\boldsymbol{e}_{S_l^a(h)} \mid \boldsymbol{n})$$

$$+ \frac{1}{n_h+1} \sum_{k=3}^{n_h+1} \binom{n_h+1}{k} \frac{\lambda_{n+1,k}}{\hat{\pi}((k-2)\boldsymbol{e}_h \mid \boldsymbol{n} - (k-2)\boldsymbol{e}_h)}. \tag{3.7}$$

Proof. The result follows by substituting $\boldsymbol{m} = \boldsymbol{e}_h$ into (3.1). $\qquad \square$

As per Remark 2.1 it is sufficient to work with the simpler recursion (3.7) as opposed to the full recursion (3.1). However, because of the denominator in the final term of (3.7) the resulting system of equations still contains as many unknowns as the recursion for the full likelihood. Hence, further approximations are needed to obtain a family of proposal distributions which is feasible to evaluate and sample.

Definition 3.1. Setting $\Lambda = \delta_0$ in (3.7) results in the approximate CSDs derived in [41] for Kingman's coalescent. This approximation ignores the dynamics of the Λ-coalescent but results in a valid IS proposal distribution that still simulates Λ-coalescent trees. The corresponding IS proposal distribution is denoted by $\mathbb{Q}^{\mathrm{SD}}$.

Paul and Song [31] introduced the *trunk ancestry*, which can be used to obtain an approximation which makes better use of the Λ-coalescent structure. We briefly recall the definition of the trunk ancestry here before using it to define a second approximate CSD family.

Definition 3.2. The trunk ancestry $\mathcal{A}^*(\boldsymbol{n})$ of a sample $\boldsymbol{n}$ is a deterministic, degenerate process started from $\boldsymbol{n}$ and evolving backwards in time but undergoing no dynamics.

In the trunk ancestry, the lineages that form $\boldsymbol{n}$ do not mutate or coalesce, and, hence, do not reach a MRCA. Instead they form an ancestral forest or 'trunk' that extends infinitely into the past.

The first two terms on the right-hand side of (3.7), corresponding to pairwise mergers and mutations, can be interpreted genealogically as the rates with which the $(n + 1)$th lineage mutates and is absorbed into $\mathcal{A}^*(\boldsymbol{n})$ by a pairwise merger. The third term corresponds to a multiple merger between the $(n + 1)$th lineage and two or more lineages in $\boldsymbol{n}$ of the same type. Because this last term involves coalescence between lineages in $\boldsymbol{n}$ it does not have an interpretation in terms of $\mathcal{A}^*(\boldsymbol{n})$. However, it can be forced into this framework by noting that the only relevant information is the time of absorption of the $(n + 1)$th lineage and the type of the lineage(s) in $\boldsymbol{n}$ with which it merges. Motivated by the trunk ancestral interpretation, we expect the following recursion to be a good, tractable approximation to (3.7).

Definition 3.3. Let $\hat{\pi}^{\mathrm{K}}(\boldsymbol{e}_h \mid \boldsymbol{n})$ be the distribution of the type of a lineage which, when traced backwards in time, mutates with rates θ_l according to the transition matrix $P^{(l)}$ at each locus $l \in L$ and is absorbed into $\mathcal{A}^*(\boldsymbol{n})$ with rate

$$\frac{\Lambda(\{0\})n}{2} + \frac{1}{n+1} \sum_{k=2}^{n+1} \binom{n+1}{k} \lambda_{n+1,k},$$

choosing its parent uniformly upon absorption. The corresponding IS proposal distribution is denoted by $\mathbb{Q}^{\mathrm{K}}$.

Proposition 3.1. *It holds that $\hat{\pi}^{\mathrm{K}}(\boldsymbol{e}_h \mid \boldsymbol{n})$ satisfies the equations*

$$\left[\frac{\Lambda(\{0\})n}{2} + \theta + \frac{1}{n+1} \sum_{k=2}^{n+1} \binom{n+1}{k} \lambda_{n+1,k} \right] \hat{\pi}^{\mathrm{K}}(\boldsymbol{e}_h \mid \boldsymbol{n})$$

$$= \frac{\Lambda(\{0\})n_h}{2} + \sum_{l \in L} \theta_l \sum_{a \in E_l} P^{(l)}_{ah[l]} \hat{\pi}^{\mathrm{K}}(\boldsymbol{e}_{S^a_l(h)} \mid \boldsymbol{n}) + \frac{n_h}{n(n+1)} \sum_{k=2}^{n+1} \binom{n+1}{k} \lambda_{n+1,k}$$

and is the stationary distribution of the Markov chain on $\mathcal{H}$ with transition matrix

$$\frac{P + [\Lambda(\{0\})/2 + (1/n(n+1)) \sum_{k=2}^{n+1} \binom{n+1}{k} \lambda_{n+1,k}]N}{\sum_{l \in L} \theta_l + (\Lambda(\{0\})n/2) + (1/(n+1)) \sum_{k=2}^{n+1} \binom{n+1}{k} \lambda_{n+1,k}}, \tag{3.8}$$

where N is the $|\mathcal{H}| \times |\mathcal{H}|$ matrix with each row equal to $(n_1, \ldots, n_{|\mathcal{H}|})$ and P is the transition probability matrix on $\mathcal{H}$ formed as a mixture of the matrices $\{P^{(l)}\}_{l \in L}$ with weights $\{\theta_l\}_{l \in L}$.

Proof. The simultaneous equations follow by tracing the $(n + 1)$th lineage backwards in time and decomposing based on the first event, and the transition matrix follows immediately from the simultaneous equations.

Note that $\mathbb{Q}^K$ has a very similar form to $\mathbb{Q}^{SD}$, and as a consequence of the linearity in N in (3.8) the efficient Gaussian quadrature approximation of Appendix A of [41] can be applied to both with minor modifications for $\mathbb{Q}^K$.

4. Simulation study

In this section we present an empirical comparison between the IS algorithms defined by $\mathbb{Q}^{SD}$ and $\mathbb{Q}^K$, and the generalised Griffiths–Tavaré proposal distribution from [2] which will be denoted by $\mathbb{Q}^{GT}$. We will also introduce two PAC algorithms making use of, respectively, $\hat{\pi}^K$ and a refinement to be specified below, and investigate their accuracy. Simulated samples have been generated using the efficient sampling algorithm provided in Section 1.4.4 of [3]. Approximate CSDs have been evaluated using a Gauss quadrature of order four (see Appendix A of [41] for details).

Simulated chromosomes consist of 15 loci with two possible alleles denoted $\{0, 1\}$ and mutation matrix $P^{(l)} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ at each locus. The coalescent is a Beta$(2 - \alpha, \alpha)$-coalescent. All simulations have been run on a single core on a Toshiba^{TM} laptop, and make use of a stopping time resampling scheme [23] with resampling checks made at hitting times of all sample sizes reaching $B = \{n - 5, n - 10, \ldots, 5\}$. This generic resampling regime has been chosen for simplicity and without regard for any particular proposal distribution.

4.1. Experiment 1

The total mutation rate is $\theta = 0.1$ spread evenly among all 15 loci. The coalescent is specified as $\alpha = 1.5$. The data consists of 100 sampled chromosomes, 95 of which share a single type, four lineages, a second type one mutation away from the main block, and a single lineage is of a third type one different mutation removed from the main block.

We consider inferring both θ and α individually, assuming that all other parameters are known and that $\theta_l = \theta/L$ for every $l \in L$. Eight independent simulations of 30 000 particles each were run on an evenly spaced grid of mutation rates spanning the interval $[0.025, 0.2]$. The same simulations were then repeated on an evenly spaced grid spanning $\alpha \in [1.1125, 1.9]$. The resulting likelihood surfaces are shown in Figure 1.

The most striking observation is that both approximate CSDs yield algorithms which are two orders of magnitude faster than the Griffiths–Tavaré scheme. Moreover, it is clear that the α-surface obtained from $\mathbb{Q}^{GT}$ has not fully converged. The wide confidence envelope at the left-hand edge and the lack of monotonicity at the right-hand edge of the $\mathbb{Q}^{GT}$ θ-surface are indicative of poorer performance when inferring θ as well.

The runtimes of $\mathbb{Q}^{SD}$ and $\mathbb{Q}^K$ are very similar in both cases, and all four surfaces from these proposals are good approximations of the truth. In the θ-case the accuracy of the two is
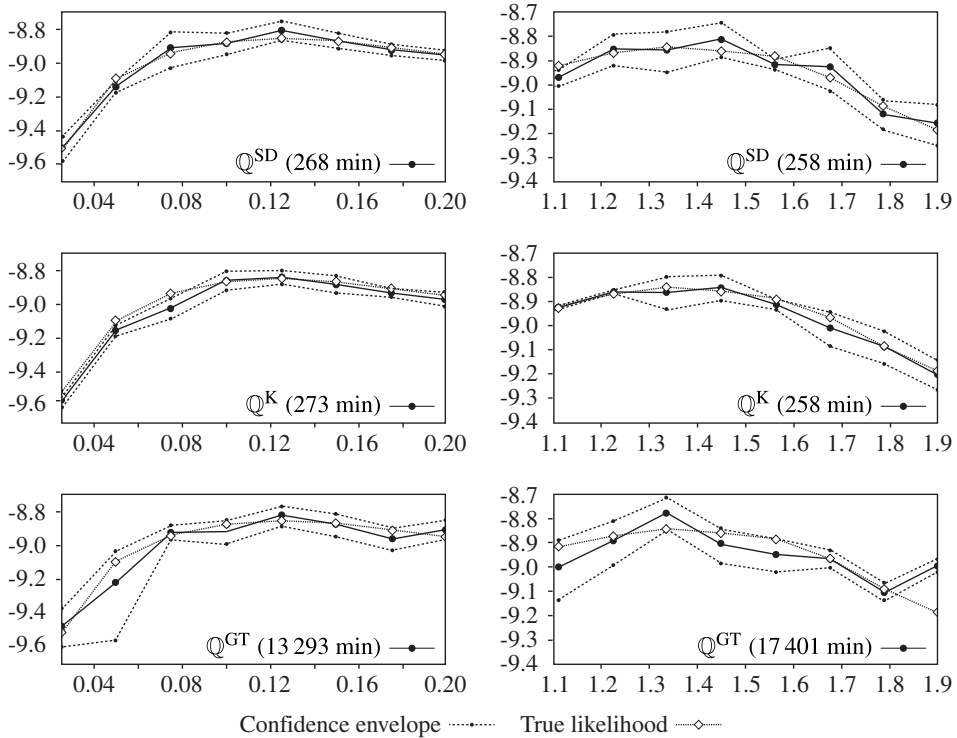
FIGURE 1: Simulated log-likelihood surfaces from 30 000 particles with ±2SE confidence envelopes. The left column is for θ and the right for α. The true surfaces are based on a 1 000 000 particle simulation using the $\mathbb{Q}^K$ proposal distribution.
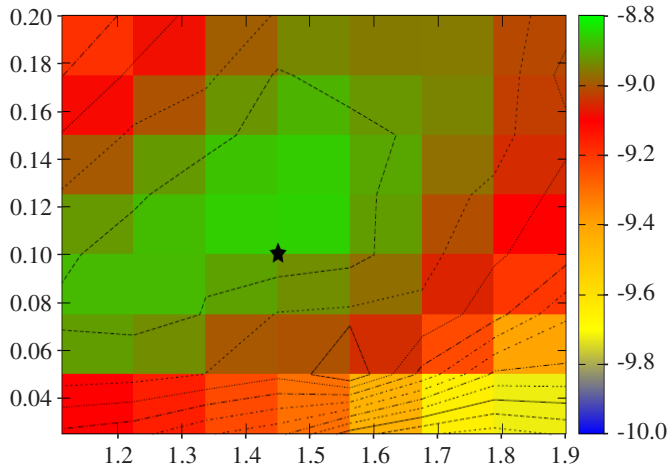

FIGURE 2: Simulated likelihood surface from 30 000 particles using $\mathbb{Q}^K$. The surface is interpolated from an 8×8 grid of independent simulations. The star denotes the maximum–likelihood estimation (MLE), which must lie on one of the grid points.

very similar, but in the α-case $\mathbb{Q}^K$ yields noticeably tighter confidence bounds and a smoother surface. This is particularly true of low values of α, which correspond to Beta-coalescents that are very different from $\Lambda = \delta_0$.

Joint inference of α and θ is also of interest. In Figure 2 we show a joint likelihood heat map for the two parameters constructed from a grid of simulations of 30 000 particles from the $\mathbb{Q}^K$ proposal. The surface is flat due to the limited amount of information in 100 samples, but the maximum likelihood estimator is close to the true $(1.5, 0.1)$ and the surface shows a high degree of monotonicity.

4.2. Experiment 2

We expect the performance of the $\mathbb{Q}^{SD}$ proposal to deteriorate the further the true model is from Kingman's coalescent, and the more demanding the data set. To that end the one-dimensional inference problems for θ and α were repeated for a sample of 150 lineages with true parameters $\theta = 0.15$ and $\alpha = 1.2$. The data set consists of 144 lineages of a given type with four other types present, each a single mutation removed from the main group. The sizes of these groups are 3, 1, 1, and 1. The results are shown in Figure 3.

We see that $\mathbb{Q}^K$ is noticeably faster when inferring θ, and slightly faster when inferring α. It also produces substantially more accurate estimates than $\mathbb{Q}^{SD}$ for small values of θ. The 30 000 particle runs have not yielded an accurate estimate for large values of θ from either algorithm.
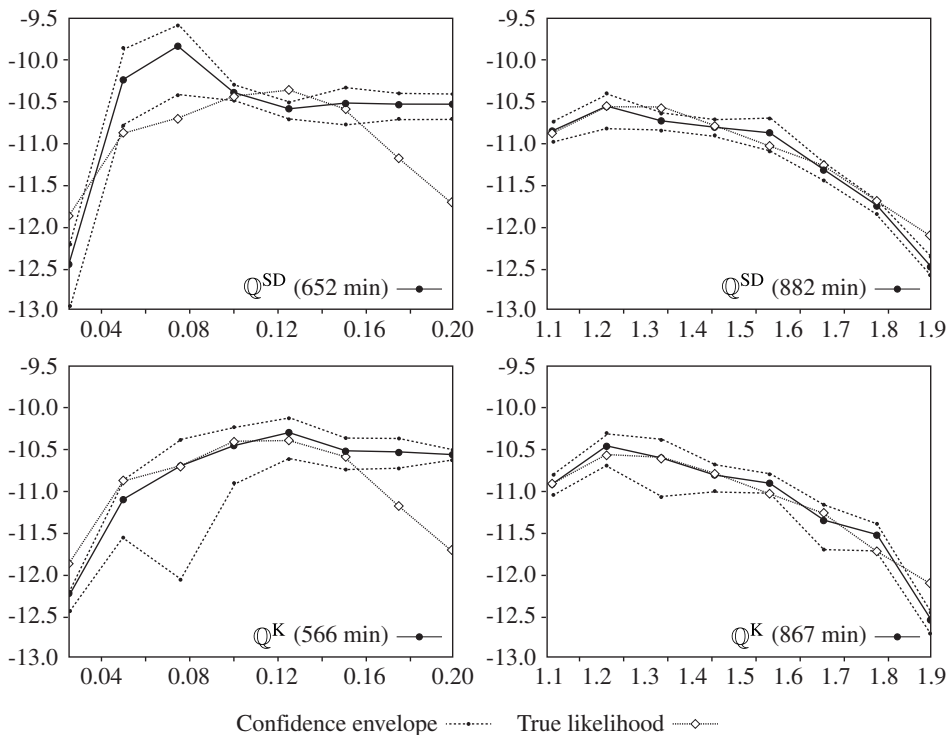

FIGURE 3: Simulated log-likelihood surfaces from 30 000 particles with $\pm2\text{SE}$ confidence envelopes. The left column is for θ and the right for α. The downward spike in the lower confidence boundary of the bottom left surface is an artifact caused by a negative value of the estimate, and the real part of the logarithm has been plotted. The values of the standard errors have no such spike.

The α-surface from $\mathbb{Q}^{\text{SD}}$ looks superficially better, but both surfaces are very similar and good matches to the true likelihood.

The deterioration of the performance of $\mathbb{Q}^{\text{SD}}$ is to be expected because the true Beta(0.8, 1.2)-coalescent is a significant departure from the $\Lambda = \delta_0$ assumption used to derive the corresponding approximate CSDs. Such coalescents are of particular interest because significantly more efficient implementations exist for Kingman's coalescent, and these should be preferred whenever the Kingman hypothesis of $\Lambda = \delta_0$ cannot be rejected. It is likely that the overestimated likelihood near $\theta = 0.06$ coincides with the MLE for this data set, had it been generated by Kingman's coalescent. Hence, $\mathbb{Q}^{\text{K}}$ is the recommended proposal distribution in practice.

Based upon the reported run times in Figures 1 and 3, we expect our IS algorithm to be feasible for samples containing hundreds of lineages formed of tens of loci, or an order of magnitude more if methods such as a driving value [19] or bridge sampling [27] are employed to reduce the number of independent simulations. There is also a strong dependence on model parameters: fast coalescence (or low α in our setting) corresponds to faster simulation runs, and both high mutation rate and large haplotype space will result in a slower algorithm.

4.3. Experiment 3: product of approximate conditionals

The IS algorithms used in the previous numerical experiments provide accurate results with reasonable computational cost, but the inference problems and data sets are of toy size and complexity. It is clear that these algorithms will be too slow for many problems of interest, such as genome-wide data or large sample sizes. The PAC method is a principled way of overcoming this restriction at the cost of asymptotic correctness, but with very significant improvements in speed. It is based on decomposing the likelihood of observed alleles $h_1, \ldots, h_n$ into a product of CSDs:

$$\mathbb{P}\{h_1, h_2, \ldots, h_n\} = \pi(h_n \mid h_1, \ldots, h_{n-1})\pi(h_{n-1} \mid h_1, \ldots, h_{n-2}) \times \cdots \times \pi(h_2 \mid h_1)\pi(h_1)$$

and then substituting in a tractable, approximate CSD to obtain computable estimates. We consider two different classes of CSDs: $\hat{\pi}^{\text{K}}(\cdot \mid \boldsymbol{n})$, and a modified version in which a lineage is absorbed into $\mathcal{A}^*(\boldsymbol{n})$ with rate

$$\sum_{h \in \mathcal{H}} \left\{ \frac{\Lambda(\{0\})n_h}{2} + \frac{1}{n_h + 1} \sum_{k=2}^{n_h + 1} \binom{n_h + 1}{k} \lambda_{n+1,k} \right\}$$

and inherits the type of the cluster n_h into which it is absorbed. This approximate CSD will be denoted by $\hat{\pi}^{\text{K2}}(\cdot \mid \boldsymbol{n})$. Note that because $\hat{\pi}^{\text{K2}}(\cdot \mid \boldsymbol{n})$ depends nonlinearly on the exact frequencies $(n_h)_{h \in \mathcal{H}}$, the precomputations which are possible for all other CSDs considered in this paper are not possible (see Proposition 1 of [41] for details). For an IS algorithm this loss of efficiency in evaluating the CSDs would be devastating, but PAC algorithms are fast enough to remain feasible. The order of the Gauss quadrature used to approximate the CSDs has also been increased to 10 for both families.

Neither approximate CSD family is exchangeable, so the estimates of the likelihood depend on the order in which the count data $\boldsymbol{n}$ is conditioned upon. Following the approach of Li and Stephens [26], and subsequent works making use of the PAC method, we partially address this issue by averaging our estimates across 1000 uniformly random permutations of the data. The number of permutations is substantially greater than what has been used for PAC models based on Kingman's coalescent, but has proven necessary in trial runs (results not shown) and comes at little additional cost. The results of applying these PAC algorithms to both the individual and joint inference questions posed in experiments 1 and 2 are summarised in Figures 4 and 5.
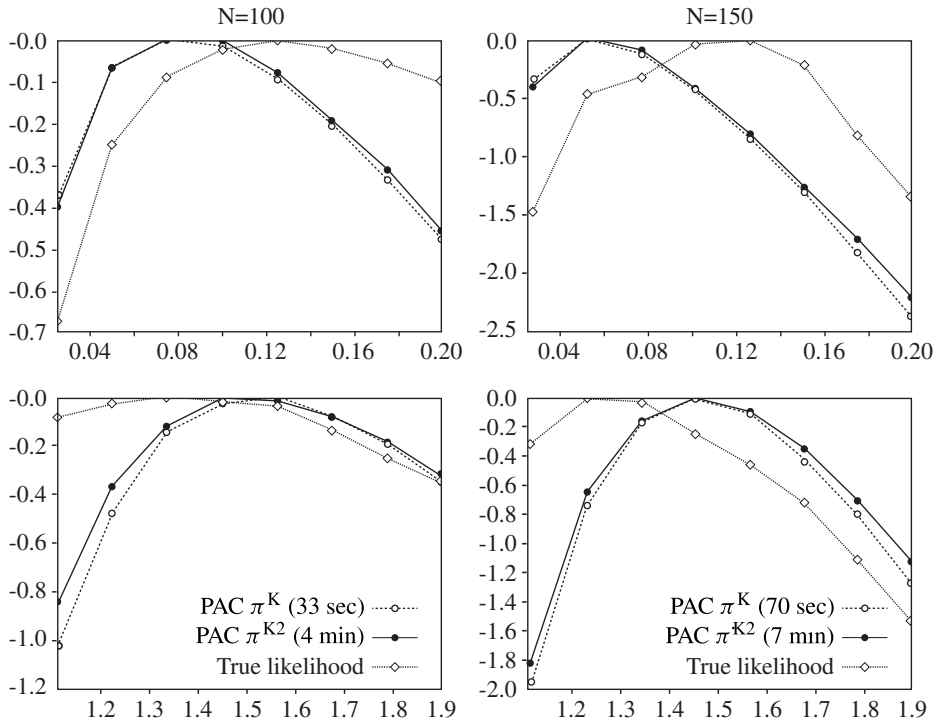
FIGURE 4: The PAC log-likelihood surfaces normalized to 0 following Li and Stephens [26]. The true likelihood surfaces in the left column are from experiment 1, and the true surfaces in the right column are from experiment 2.

The results of the PAC calculations seem mixed. Both PAC algorithms are extremely fast, and would likely remain feasible even for reasonably large data sets. The PAC likelihood estimates are consistently too low by many orders of magnitude. Nevertheless, the PAC MLEs in Figure 4 are strikingly close to the true maximisers, particularly for the smaller data set in the left column. On the other hand, the joint MLEs in Figure 5 are much further from the truth, although the surfaces still broadly capture the diagonal shape seen in Figure 2. It is also interesting to note that the two PAC methods perform very similarly in the one-dimensional problems in Figure 4, but the two-dimensional surface obtained from π^{K2} is a better fit than that from π^{K} for the smaller sample. For the larger sample the surfaces are nearly identical.

The run times in Figure 4 indicate that the PAC method will remain computationally feasible for substantially larger data sets than the IS algorithm, at least up to tens of thousands of lineages and/or thousands of loci. Of course, the accuracy of the PAC method to such data sets cannot be concluded from the trials presented here, and careful verification will be necessary on a case-by-case basis. In further contrast to IS, the runtime of the PAC algorithm is independent of the model parameters, and influenced only weakly by the size of the space of haplotypes.

A substantial amount of work will be required to develop a thorough understanding of the accuracy and pitfalls of these PAC algorithms, and whether or not the more advanced PAC algorithms developed for Kingman's coalescent can be adapted to the Λ-coalescent setting as well. Our preliminary simulations motivate this undertaking, and confirm that the PAC method can provide useful, principled and fast results for Λ-coalescents.
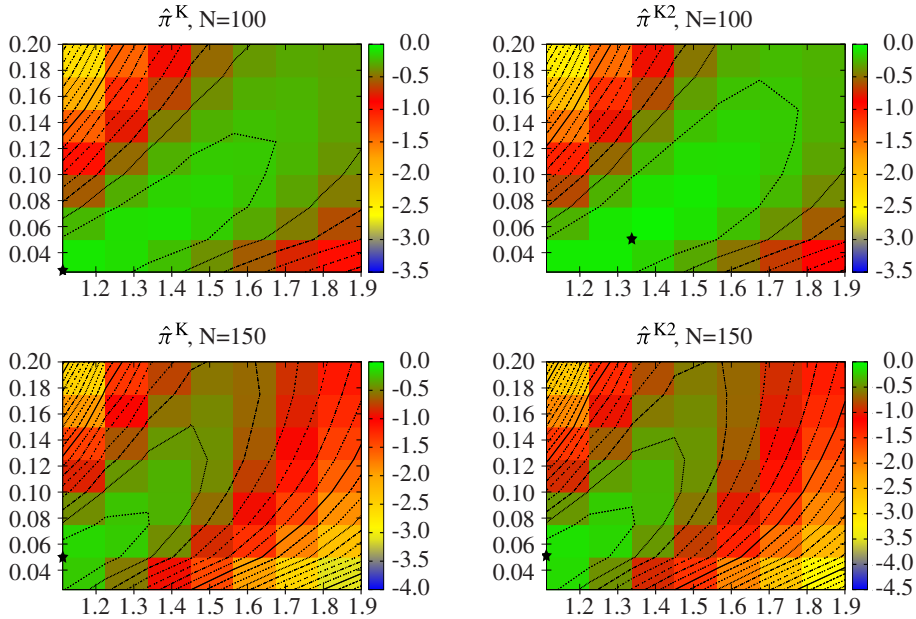
FIGURE 5: The PAC joint log-likelihood surfaces normalized to 0. Locations of MLEs are indicated by stars. Figure 2 provides a suitable IS comparison to the top row.

5. Importance sampling for Ξ-coalescents

The important tools in deriving the optimal proposal distributions $\mathbb{Q}^*(\cdot \mid \cdot)$ and the approximate CSDs $\hat{\pi}^K(\cdot \mid \cdot)$ were, respectively, the lookdown construction of [11] and the trunk ancestry of [31]. Both of these are also available for the Ξ-coalescent, and in this section we make use of them to extend the IS algorithm to this family of models.

A lookdown construction for the Ξ-coalescent and the Ξ-Fleming–Viot process was derived by Birkner *et al.* [6] and can be described as follows. For ease of notation, we assume $\Xi(\{\mathbf{0}\}) = 0$. If Ξ does have an atom at zero, its treatment is identical to the Λ-case.

Let Π_Ξ be a Poisson point process on $\mathbb{R}_+ \times \Delta \times [0, 1]^{\mathbb{N}}$ with rate

$$\mathrm{d}t \otimes \left(\sum_{i=1}^{\infty} r_i^2 \right)^{-1} \Xi(\mathrm{d}\boldsymbol{r}) \otimes \mathrm{d}r^{\otimes \mathbb{N}}$$

and associate to each lineage a level $\{1, \ldots, n\}$. Define the function

$$g(\boldsymbol{r}, u) := \begin{cases} \min\{j \in \mathbb{N} \colon \sum_{i=1}^{j} r_i \geq u\} & \text{if } u \leq \sum_{i=1}^{\infty} r_i, \\ \infty & \text{otherwise.} \end{cases}$$

At each $(t_j, (r_{j1}, r_{j2}, \ldots), (u_{j1}, u_{j2}, \ldots)) \in \Pi_\Xi$ group the n particles such that all particles $l \in \{1, \ldots, n\}$ with $g(\boldsymbol{r}_j, u_{jl}) = k$ form a family for each $k \in \mathbb{N}$. Among each family every particle copies the type of the particle with the lowest level. In addition each particle follows an independent mutation process as for the Λ-coalescent.

This lookdown construction will be instrumental in establishing the following recursion, which is an explicit version of (2.1) for Ξ-coalescents and a finite sites analogue of the sampling recursion presented in [28] for the infinite alleles model.

Theorem 5.1. *The likelihood of type frequencies* $\boldsymbol{n} \in \mathbb{N}^{|\mathcal{H}|}$ *sampled from the stationary* Ξ-*Fleming–Viot process solves*

$$
\mathbb{P}(\boldsymbol{n}) = \frac{1}{g_n + n\theta} \Bigg\{ \sum_{h:\, n_h > 0} \sum_{l \in L} \theta_l \sum_{a \in E_l} (n_{S_l^a(h)} + 1 - \delta_{ah[l]}) \mathbb{P}\{\boldsymbol{n} - \boldsymbol{e}_h + \boldsymbol{e}_{S_l^a(h)}\}
$$

$$
+ \sum_{k_1 = 1}^{n_1} \cdots \sum_{k_{|\mathcal{H}|} = 1}^{n_{|\mathcal{H}|}} \sum_{\pi^1 \in P_{n_1}^{k_1}} \cdots \sum_{\pi^{|\mathcal{H}|} \in P_{n_{|\mathcal{H}|}}^{k_{|\mathcal{H}|}}} \mathbf{1}\left\{\sum_{h \in \mathcal{H}} k_h < n\right\} \binom{n}{|\pi_1^1|, |\pi_2^1|, \ldots, |\pi_{|\mathcal{H}|}^{|\mathcal{H}|}|}
$$

$$
\times \binom{|\vee_{h \in \mathcal{H}}\, \pi^h|}{|\pi^1|, \ldots, |\pi^{|\mathcal{H}|}|}^{-1}
$$

$$
\times \lambda_{n; K(\vee_{h \in \mathcal{H}}\pi^h); S(\vee_{h \in \mathcal{H}}\pi^h)} \mathbb{P}(\boldsymbol{k}) \Bigg\}, \tag{5.1}
$$

with the convention that $\sum_{k=1}^{0} f(k) = f(0)$ *and where* $\mathbf{1}$ *is the indicator function. Here* $P_{n_h}^{k_h}$ *denotes the set of equivalence relations on* $n_h \in \mathbb{N}$ *elements with* $k_h \le n_h$ *equivalence classes,* $\pi^h = (\pi_1^h \ldots \pi_{k_h}^h)$ *denotes such an equivalence relation so that* $\sum_{i=1}^{k_h} |\pi_i^h| = n_h$ *and* $\vee_{h \in \mathcal{H}}\pi^h$ *is the equivalence relation on* n *elements obtained from applying each* π^h *to the corresponding* n_h *elements. The vector* $K(\pi)$ *lists the sizes of all equivalence classes with more than one member,* $S(\pi)$ *is the number of classes with exactly one member, and* g_n *is the total coalescence rate of* n *untyped lineages given by*

$$
g_n = \sum_{a=1}^{n-1} \frac{n!}{a!} \sum_{b_1, \ldots, b_a \in \mathbb{N}, b_1 + \cdots + b_a = n} \frac{\lambda_{n; K(\boldsymbol{b}); S(\boldsymbol{b})}}{b_1! \times \cdots \times b_a!}.
$$

Proof. The proof is the same as the proof in Section 1.4.1 of [3], adapted here from the Λ-coalescent to the Ξ-coalescent. Let p define the distribution of the types of the first n levels of the stationary lookdown construction. Decomposing according to which event (whether mutation or a merger) occurred first when tracing backwards in time yields

$$
p(y_1, \ldots, y_n) = \frac{1}{g_n + n\theta} \Bigg\{ \sum_{i=1}^{n} \sum_{l \in L} \theta_l \sum_{a \in E_l} P_{ah[y_i]}^{(l)} p(y_1, \ldots, y_{i-1}, S_l^a(y_i), y_{i+1}, \ldots, y_n)
$$

$$
+ \sum_{\pi \in P(\boldsymbol{y})} \lambda_{n; K(\pi); S(\pi)} p(\gamma_\pi(y_1, \ldots, y_n)) \Bigg\}, \tag{5.2}
$$

where $P(\boldsymbol{y})$ is the set of equivalence relations describing permissible mergers for the sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ (that is, mergers where no equivalence class contains lineages of more than one type) and $\gamma_\pi(y_1, \ldots, y_n)$ is the vector of types which results in $(y_1, \ldots, y_n)$ if the lookdown-and-copy event denoted by the equivalence relation π takes place.

By exchangeability, we are only interested in the vector of $\boldsymbol{n} = (n_1, \ldots, n_{|\mathcal{H}|})$ type frequencies. For such a vector define the canonical representative as

$$
\kappa(\boldsymbol{n}) := (\underbrace{1, \ldots, 1}_{n_1}, \underbrace{2, \ldots, 2}_{n_2}, \ldots, \underbrace{|\mathcal{H}|, \ldots, |\mathcal{H}|}_{n_{|\mathcal{H}|}})
$$

and the likelihood as

$$
p^0(\boldsymbol{n}) := \binom{n}{n_1, \ldots, n_{|\mathcal{H}|}} p(\kappa(\boldsymbol{n})).
$$

Now we have the following identities:

$$n_h \binom{n}{n_1, \ldots, n_{|\mathcal{H}|}} p(\kappa(\boldsymbol{n} - \boldsymbol{e}_h + \boldsymbol{e}_{S_l^a(h)})) = (n_{S_l^a(h)} + 1 - \delta_{ah[l]}) p^0(\boldsymbol{n} - \boldsymbol{e}_h + \boldsymbol{e}_{S_l^a(h)}),$$

$$\binom{n}{n_1, \ldots, n_{|\mathcal{H}|}} \prod_{h \in \mathcal{H}} \binom{n_h}{|\pi_1^h|, \ldots, |\pi_{k_h}^h|} p(\kappa(\boldsymbol{k}))$$

$$= \binom{n}{|\pi_1^1|, |\pi_2^1|, \ldots, |\pi_{|\mathcal{H}|}^{|\mathcal{H}|}|} \binom{k}{k_1, \ldots, k_{|\mathcal{H}|}}^{-1} p^0(\boldsymbol{k}),$$

which, when substituted into (5.2), yield the desired recursion.

As in Section 2 we can consider approximating the solution to (5.1) by importance sampling, and the following theorem is a straightforward extension of Theorem 2.1.

Theorem 5.2. *The optimal proposal distributions for recursion (5.1), denoted $\mathbb{Q}_\Xi^*$, are*

$$\mathbb{Q}_\Xi^*(H_i \mid H_{i+1})$$

$$\propto \begin{cases} n_h \theta_l \dfrac{\pi(\boldsymbol{e}_{S_l^a(h)} \mid H_{i+1} - \boldsymbol{e}_h)}{\pi(\boldsymbol{e}_h \mid H_{i+1} - \boldsymbol{e}_h)} P_{ah[l]}^{(l)} & \text{if } H_i = H_{i+1} - \boldsymbol{e}_h + \boldsymbol{e}_{S_l^a(h)}, \\[2ex] \displaystyle\sum_{\pi^1 \in P_{n_1}^{k_1}} \cdots \sum_{\pi^{|\mathcal{H}|} \in P_{n_{|\mathcal{H}|}}^{k_{|\mathcal{H}|}}} \prod_{h \in \mathcal{H}} \binom{n_h}{|\pi_1^h|, \ldots, |\pi_{k_h}^h|} \\[1ex] \qquad\qquad \times \dfrac{\lambda_{n; K(\vee_{h \in \mathcal{H}} \pi^h); S(\vee_{h \in \mathcal{H}} \pi^h)}}{\pi(\boldsymbol{n} - \boldsymbol{k} \mid \boldsymbol{k})} & \text{if } H_{i+1} = \boldsymbol{n} \text{ and } H_i = \boldsymbol{k}, \end{cases}$$

where mergers of the second type range over $k_h = \{1, \ldots, n_h\}$ and $\sum_{h \in \mathcal{H}} k_h < n$.

Proof. The argument is identical to the proof of Theorem 2.1 taking into account the larger class of permitted simultaneous multiple mergers and, hence, different combinatorial coefficients.

As before, the CSDs used in the statement of Theorem 5.2 are not available, but any approximation to them will yield an unbiased algorithm and better approximations can be expected to correspond to more efficient algorithms. The generator of the Ξ-Fleming–Viot process is not as immediately tractable as its Fleming–Viot and Λ-Fleming–Viot counterparts so we abandon the generator-based approach of De Iorio and Griffiths (see [8] and [9]) and derive approximate CSDs from the trunk ancestry $\mathcal{A}^*(\boldsymbol{n})$.

Definition 5.1. Let $\hat{\pi}_\Xi^K(\boldsymbol{e}_h \mid \boldsymbol{n})$ be the CSD obtained by letting the $(n + 1)$th lineage mutate with rates $\{\theta_l\}_{l \in L}$ via transition matrices $\{P^{(l)}\}_{l \in L}$, be absorbed into $\mathcal{A}^*(\boldsymbol{n})$ with rate

$$\frac{1}{n+1} \sum_{k=1}^{n} \sum_{\pi \in P_{n+1}^k} \binom{n+1}{|\pi_1|, \ldots, |\pi_k|} \lambda_{n+1; K(\pi); S(\pi)},$$

and choose its parent uniformly upon absorption.

Proposition 5.1. *The approximate CSD $\hat{\pi}_{\Xi}^{K}(e_h \mid n)$ solves the following recursion:*

$$
\left[\theta + \frac{1}{n+1}\sum_{k=1}^{n}\sum_{\pi \in P_{n+1}^{k}}\binom{n+1}{|\pi_1|, \ldots, |\pi_k|}\lambda_{n+1;K(\pi);S(\pi)}\right]\hat{\pi}_{\Xi}^{K}(e_h \mid n)
$$

$$
= \frac{n_h}{n(n+1)}\sum_{k=1}^{n}\sum_{\pi \in P_{n+1}^{k}}\binom{n+1}{|\pi_1|, \ldots, |\pi_k|}\lambda_{n+1;K(\pi);S(\pi)} + \sum_{l \in L}\theta_l\sum_{a \in E_l}P_{ah[l]}^{(l)}\hat{\pi}_{\Xi}^{K}(e_{S_l^a(h)} \mid n)
$$

and is the stationary distribution of the Markov Chain on $\mathcal{H}$ with transition probability matrix

$$
\frac{P + \{1/n(n+1)\sum_{k=1}^{n}\sum_{\pi \in P_{n+1}^{k}}\binom{n+1}{|\pi_1|, \ldots, |\pi_k|}\lambda_{n+1;K(\pi);S(\pi)}\}N}{\theta + 1/(n+1)\sum_{k=1}^{n}\sum_{\pi \in P_{n+1}^{k}}\binom{n+1}{|\pi_1|, \ldots, |\pi_k|}\lambda_{n+1;K(\pi);S(\pi)}},
$$

where P and N are as in Proposition 3.1.

Proof. The proof is identical to Proposition 3.1 and follows by considering the first event backwards in time encountered by the lineage.

Note that because simultaneous multiple mergers can take place, the decomposition in Remark 2.1 is no longer valid and multivariate approximate CSDs $\hat{\pi}_{\Xi}^{K}(m \mid n)$ must also be specified. This is most naturally done by averaging over all permutations of the lineages in m, but this is computationally infeasible for all but very small samples m. The PAC approach of averaging over a relatively small number of random permutations can be used to yield a more practical family, although algorithms will still be limited by the fact that evaluating the CSDs requires computing all equivalence classes on n elements. This burden can be alleviated considerably by assuming that the measure Ξ places full mass on a finite dimensional simplex, which amounts to restricting the number of permitted simultaneous mergers to the same, finite number. If this number is small compared to the size of the data set, far fewer terms will need to be computed at each stage of the algorithm but the model still allows for more general ancestral trees than any Λ-coalescent. In particular, the case of up to four simultaneous mergers arising in coalescent models of diploid populations (see [4], [30]) seems computationally feasible.

6. Discussion

In this paper we have developed novel IS algorithms for inference under the Λ- and Ξ-coalescent models, which retain the rigorous motivations of proposals that have been designed for Kingman's coalescent [8], [9], [31]. Furthermore, they outperform existing algorithms for Λ-coalescent inference, and, like all IS methods, are unbiased. It should be noted however that the greater modelling flexibility provided by Λ- and Ξ-coalescents comes with additional computational cost in comparison to the more restrictive Kingman's coalescent. The inference problems considered in this paper have consisted of small samples of chromosomes comprised of a small number of loci, each with a simple mutation model. While some cost is certainly unavoidable, these computations can be sped up considerably by reducing the number of independent simulations through making use of driving values [19] or bridge sampling [27]. It is also noteworthy that, as with IS algorithms in general, all of the algorithms used here can be parallelised very effectively.

The limits on data sets which can be feasibly analysed using IS are restrictive even under Kingman's coalescent, so alternate methods have been developed to tackle broader classes

of problems. The PAC method is a prime example, and our simulations suggest that it is also a viable approach for Λ-coalescents. Much work has been carried out on sophisticated approximations to CSDs for Kingman's coalescent with recombination and other features, and our results in Section 4.3 indicate that investigating similar approaches under Λ- and Ξ-coalescents is a fruitful direction for future research. Many of the generalisations of interest result in processes with generators that vary from those studied in this paper only by additive terms so we expect that the machinery used here can be applied more generally with little added difficulty.

Acknowledgement

References

[1] ÁRNASON, E. (2004). Mitochondrial cytochrome *b* DNA variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. *Genetics* **166,** 1871–1885.

[2] BIRKNER, M. AND BLATH, J. (2008). Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *J. Math. Biol.* **57,** 435–465.

[3] BIRKNER, M. AND BLATH, J. (2009). Measure-valued diffusions, general coalescents and population genetic inference. In *Trends in Stochastic Analysis* (London Math. Soc. Lecture Notes Ser. **353**), Cambridge University Press, pp. 329–363.

[4] BIRKNER, M., BLATH, J. AND ELDON, B. (2013). An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics* **193,** 255–290.

[5] BIRKNER, M., BLATH, J. AND STEINRÜCKEN, M. (2011). Importance sampling for Lambda-coalescents in the infinitely many sites model. *Theoret. Pop. Biol.* **79,** 155–173.

[6] BIRKNER, M. *et al.* (2009). A modified lookdown construction for the Xi–Fleming–Viot process with mutation and populations with recurrent bottlenecks. *ALEA Lat. Amer. J. Prob. Math. Statist.* **6,** 25–61.

[7] BOOM, J. D. G., BOULDING, E. G. AND BECKENBACK, A. T. (1994). Mitochondrial DNA variation in introduced populations of Pacific oyster, *Crassostrea gigas*, in British Columbia. *Canad. J. Fish. Aquat. Sci.* **51,** 1608–1614.

[8] DE IORIO, M. AND GRIFFITHS, R. C. (2004). Importance sampling on coalescent histories. I. *Adv. Appl. Prob.* **36,** 417–433.

[9] DE IORIO, M. AND GRIFFITHS, R. C. (2004). Importance sampling on coalescent histories. II. Subdivided population models. *Adv. Appl. Prob.* **36,** 434–454.

[10] DE IORIO, M., GRIFFITHS, R. C., LEBLOIS, R. AND ROUSSET, F. (2005). Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theoret. Pop. Biol.* **68,** 41–53.

[11] DONNELLY, P. AND KURTZ, T. G. (1999). Particle representations for measure-valued population models. *Ann. Prob.* **27,** 166–205.

[12] ELDON, B. AND WAKELEY, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics* **172,** 2621–2633.

[13] FEARNHEAD, P. AND DONNELLY, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159,** 1299–1318.

[14] FELSENSTEIN, J., KUHNER, M. K., YAMATO, J. AND BEERLI, P. (1999). *Likelihoods on Coalescents: A Monte Carlo Sampling Approach to Inferring Parameters from Population Samples of Molecular Data* (IMS Lect. Notes Monogr. Ser. **33**), Institute of Mathematical Statistics, Hayward, CA, pp. 163–185.

[15] GÖRÜR, D. AND TEH, Y. W. (2008). An efficient sequential Monte Carlo algorithm for coalescent clustering. In *Advances in Neural Information Processing Systems 21* (NIPS 2008), 8pp.

[16] GRIFFITHS, R. C. AND MARJORAM, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3,** 479–502.

[17] GRIFFITHS, R. C. AND TAVARÉ, S. (1994). Ancestral inference in population genetics. *Statist. Sci.* **9,** 307–319.

[18] GRIFFITHS, R. C. AND TAVARÉ, S. (1994). Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. London B* **344,** 403–410.

[19] GRIFFITHS, R. C. AND TAVARÉ, S. (1994). Simulating probability distributions in the coalescent. *Theoret. Pop. Biol.* **46,** 131–159.

[20] GRIFFITHS, R. C. AND TAVARÉ, S. (1999). The ages of mutations in gene trees. *Ann. Appl. Prob.* **9,** 567–590.

[21] GRIFFITHS, R. C., JENKINS, P. A. AND SONG, Y. S. (2008). Importance sampling and the two-locus model with subdivided population structure. *Adv. Appl. Prob.* **40,** 473–500.

[22] HOBOLTH, A., UYENOYAMA, M. K. AND WIUF, C. (2008). Importance sampling for the infinite sites model. *Statist. Appl. Genet. Mol. Biol.* **7,** Article 32.

[23] JENKINS, P. A. (2012). Stopping-time resampling and population genetic inference under coalescent models. *Statist. Appl. Genet. Mol. Biol.* **11,** Article 9.

[24] JENKINS, P. A. AND GRIFFITHS, R. C. (2011). Inference from samples of DNA sequences using a two-locus model. *J. Comput. Biol.* **18,** 109–127.

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

[26] LI, N. AND STEPHENS, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165,** 2213–2233.

[27] MENG, X.-L. AND WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6,** 831–860.

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

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

[30] MÖHLE, M. AND SAGITOV, S. (2003). Coalescent patterns in diploid exchangeable population models. *J. Math. Biol.* **47,** 337–352.

[31] PAUL, J. S. AND SONG, Y. S. (2010). A principled approach to deriving approximate conditional sampling distributions in population genetic models with recombination. *Genetics* **186,** 321–338.

[32] PAUL, J. S., STEINRÜCKEN, M. AND SONG, Y. S. (2011). An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics* **187,** 1115–1128.

[33] PITMAN, J. (1999). Coalescents with multiple collisions. *Ann. Prob.* **27,** 1870–1902.

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

[35] SARGSYAN, O. AND WAKELEY, J. (2008). A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theoret. Pop. Biol.* **74,** 104–114.

[36] SCHWEINSBERG, J. (2000). Coalescents with simultaneous multiple collisions. *Electron. J. Prob.* **5,** 50pp.

[37] SCHWEINSBERG, J. (2003). Coalescent processes obtained from supercritical Galton–Watson processes. *Stoch. Process. Appl.* **106,** 107–139.

[38] SHEEHAN, S., HARRIS, K. AND SONG, Y. S. (2013). Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* **194,** 647–662.

[39] STEINRÜCKEN, M., BIRKNER, M. AND BLATH, J. (2013). Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theoret. Pop. Biol.* **87,** 15–24.

[40] STEINRÜCKEN, M., PAUL, J. S. AND SONG, Y. S. (2013). A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theoret. Pop. Biol.* **87,** 51–61.

[41] STEPHENS, M. AND DONNELLY, P. (2000). Inference in molecular population genetics. *J. R. Statist. Soc. B* **62,** 605–655.

[42] TAYLOR, J. E. AND VÉBER, A. (2009). Coalescent processes in subdivided populations subject to recurrent mass extinctions. *Electron. J. Prob.* **14,** 242–288.