Introduction

The \( \Lambda \)-coalescent family ([Pe09], [Sag99]) models the ancestry of a sample of haplotypes from a large population with an infinite variance family size distribution. Recent studies have indicated that these coalescents sometimes provide better modeling fits to high-frequency populations, such as Atlantic cod or Pacific oysters, than the seminal Kingman’s coalescent ([BBB94], [Azu04], [BB06], [BB88], [BB811], [BB12]).

Coalescent likelihoods are highly intractable, but have a simple form given the ancestral tree. Thus a standard approach is to treat the ancestry as missing data. This requires integrating over all possible ancestral trees: also an intractable computation. However, the integral can be approximated via importance sampling (IS). Stephens and Donnelly identified the optimal importance sampling proposal distribution for Kingman’s coalescent ([SD00]), and provided a heuristic approximation to it to obtain an approximately optimal algorithm. This paper summarises the extension of their derivation to cover the whole \( \Lambda \)-coalescent family. Full details, and a further extensions to even more general \( \Lambda \)-coalescents can be found in ([KJS14]).

Example: the Beta\( (2-\alpha,\alpha) \)-coalescent

The Beta\( (2-\alpha,\alpha) \)-coalescent ([SD03]) can be obtained as the high-density limit of a finite population which evolves as follows. Suppose there are \( N \) individuals evolving in discrete generations, each with a haplotype drawn from a finite set \( S \) (e.g. \( M = \{T, C, A, G\} \) if we are modeling a single locus of DNA). Each individual produces a random number of potential offspring distributed according to a power law tail \( r^{-\alpha} \), \( \alpha \in (0,1) \). Offspring inherit the type of their parent.

The next generation is formed by sampling \( N \) of these offspring without replacement. Those offspring not sampled are assumed dead, so that the population is of constant size. Measuring time in units of \( N \) generations and letting \( N \rightarrow \infty \) yields a population whose type-frequency are described by the \( |H| \)-dimensional Beta\( (2-\alpha,\alpha) \)-Fleming-Viot jump-diffusion, the ancestors of samples are given by Beta\( (2-\alpha,\alpha) \)-coalescent trees.

General \( \Lambda \)-coalescents with recurrent mutation

More generally, the \( \Lambda \)-coalescent family is parameterised by probability measures on the unit interval, \( \alpha \in M_1([0,1]) \). This measure determines the coalescence algorithm. This posterior summarises the extension of their derivation to cover the whole \( \Lambda \)-coalescent family. Full details, and a further extensions to even more general \( \Lambda \)-coalescents can be found in ([KJS14]).

Ancestry as missing data

Let \( n=(n_1,\ldots,n_{|H|}) \) denote the observed allele frequencies of \( |H| \) lineages, and \( A \) denote the ancestral tree annotated with mutations. The likelihood can be written

\[
P(n) = \frac{\prod a_i^{|H|}}{\prod a_i} \mathcal{P}(n|A) \mathcal{D}(A|\Lambda),
\]

where \( \mathcal{P}(n|A) = 1 \) if the leaves of \( A \) are compatible with \( n \) and 0 otherwise. The number of transitions required to reach the MRCA of \( A \), the leaves of the tree, \( A_\Lambda \), is the root and the other \( A \)'s are intermediate states separated by mutations or coalescences. For example, if we have \( T = 5, n_1, n_2, n_3, n_4, n_5 \), then \( A \sim \{ P, R, A \} \). \( A \sim \{ R, B \} \) and \( A \sim \{ B, R \} \), where \( B, P, R \) stand for blue, green, purple and red.

The conditional distributions in (1) can be written in closed form, but the integral is intractable and naive Monte Carlo fails because leaves compatible with the data are exceedingly rare. Progress can be made by introducing an IS proposal distribution \( \mathcal{Q} \) starting from the data and proceeding backwards in time:

\[
P(n) = \int \prod_i \mathcal{Q}(n_i|A_i|\Lambda) \mathcal{D}(A_i|\Lambda) dA_i.
\]

The factor \( \mathcal{P}(n|A) \) is not needed because all trees will be compatible with the data by construction.

The optimal proposal distribution

It is a standard result in IS that the optimal proposal distribution is the conditional law of the process given the data. In this context that means the law of the \( \Lambda \)-coalescent conditioned on the observed leaves. This distribution can be written

\[
Q^*\mathcal{A}(n_i|\Lambda) = \frac{\int \mathcal{Q}(n_i|a_i|\Lambda) \mathcal{D}(a_i|\Lambda) da_i}{\int \mathcal{Q}(n_i|a_i|\Lambda) \mathcal{D}(a_i|\Lambda) da_i}
\]

for \( a_i \) the canonical unit vector in direction \( i \), the \( n_i \) refer to allele frequencies in \( A_i \) and \( \mathcal{Q} \) the conditional sampling distribution (CSD) of further haplotypes given observed configuration (A), (Theorem 1).

Approximate conditional sampling distributions

The true CSDs are as intractable as the likelihood, but they can be approximated in a principled way. These approximations can be substituted for \( \mathcal{Q} \) in (9), yielding an approximately optimal proposal distributions and hence an approximately optimal IS algorithm. The optimal algorithm produces zero variance estimators whose value is the exact likelihood, so algorithms that are close can be expected to be very efficient. Following the approach of [PS10], fix the observation configuration in time so that it incurs no mutations or coalescences. Let the branch describing the further allele undergo mutations as usual, and be absorbed into the observed configuration at a constant rate. Upon absorption it proceeds backwards in time:

\[
\phi(\mathcal{A}(n)|\Lambda) = \begin{cases} \phi_0(\mathcal{A}(n)|\Lambda) & \text{if } \mathcal{A}(n) = A \sim \{ k \} \text{ for } k \in \{2,\ldots,n\} \\ \phi_{n-k}(\mathcal{A}(n)|\Lambda) & \text{if } \mathcal{A}(n) = A \sim \{ k \} \text{ for } k \in \{1,\ldots,n\} \\ \mathcal{Q}(\mathcal{A}(n)|\Lambda) & \text{otherwise} \end{cases}
\]

where \( \lambda_{n-k} \) denotes the number of lineages in the conditioning configuration. The resulting approximate CSD can be thought of as the stationary distribution of a Markov chain on \( \mathcal{A}(n) \) with transition matrix

\[
\Phi = \begin{bmatrix}
\phi(1|\Lambda) & \ldots & \phi(n-1|\Lambda) & \phi(n|\Lambda)
\end{bmatrix}
\]

where \( N = |H| \times |H| \) matrix with each row equal to \( n \), \( n \), \( n \), \( n \) (KJS14), Proposition 1). The multi-haplotype CSDs required in \( \Phi \) can be obtained from the univariate ones by the standard product formula of conditional probabilities. The approximate CSDs are not exchangeable, but this does not cause an ambiguity because \( \mathcal{Q} \) only requires evaluating multi-\( \Lambda \)-coalescent configurations for all haplotypes are equal and permutations leave the expression unchanged.

Acknowledgements and References

Jere Koskela is a member of the Mathematics and Statistics Doctoral Training Centre, funded by EPSRC grant number EP/K023364/1.