Approximately optimal importance sampling for Λ–coalescents

Jere Koskela, Paul Jenkins and Dario Spanò

Mathematics Institute and Department of Statistics, University of Warwick

Introduction

The Λ–coalescent family ([Pit99], [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–fecundity populations, such as Atlantic cod or Pacific oysters, than the seminal Kingman's coalescent ([BBB94], [Árn04], [EW06], [BB08], [BBS11], [SBB13]).

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 poster summarises the extension of their derivation to cover the whole Λ–coalescent family. Full details, and a further extensions to even more general Ξ–coalescents can be found in [KJS14].

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

The Beta$(2-\alpha,\alpha)$–coalescent [Sch03] 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 $\mathcal{H}$ (e.g. $\mathcal{H}=\{T,C,A,G\}$ 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 $r^{-\alpha}, \alpha\in[1,2)$. 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 $N\to\infty$ yields a population whose type–frequencies are described by the $|\mathcal{H}|$–dimensional Beta$(2-\alpha,\alpha)$–Fleming–Viot jump–diffusion, and the ancestries of samples are given by Beta$(2-\alpha,\alpha)$–coalescent trees.

General Λ–coalescents with recurrent mutation

More generally, the Λ–coalescent family is parametrised by probability measures on the unit interval: $\Lambda\in\mathcal{M}_1([0,1])$. This measure determines the coalescence rates. When there are $n\in\mathbb{N}$ lineages, any $k\in\{2,\dots,n\}$ of them will merge at rate

$$\lambda_{n,k}=\int_0^1 x^{k-2}(1-x)^{n-k}\Lambda(dx).$$

In addition to Beta measures, other famous examples are $\Lambda=\delta_0$, i.e. Kingman's coalescent, and $\Lambda=\frac{2}{2+\psi^2}\delta_0+\frac{\psi^2}{2+\psi^2}\delta_\psi$, where $\psi\in(0,1]$ [EW06].

Mutation can be incorporated into Λ–coalescents similarly to Kingman's coalescent. Conditional on an ancestral tree, mutations occur along branches as points of a Poisson process with rate $\theta>0$. In the finite alleles context mutation probabilities for each parental haplotype can be specified by a stochastic matrix P. Figure 1 depicts a realisation of a Λ–coalescent tree with four leaves.

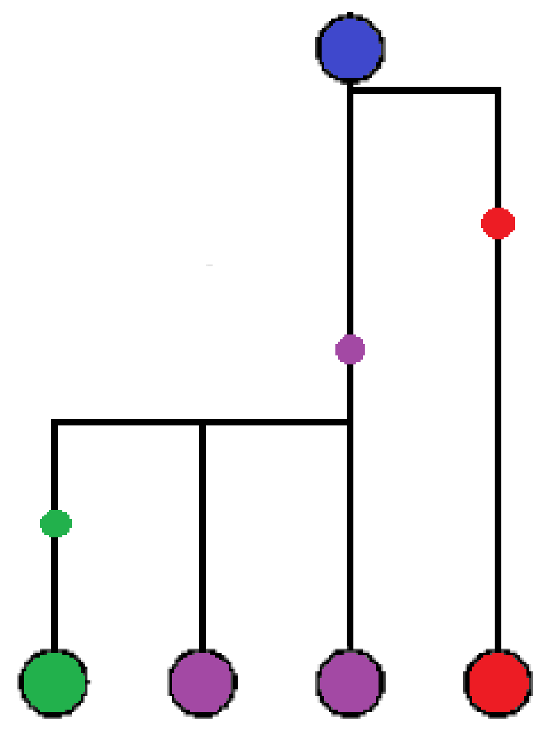

Figure 1 : A Λ–coalescent tree annotated with mutations. The most recent common ancestor (MRCA) is sampled from the stationary law of the mutation stochastic matrix P, and happens to be blue. Three mutations take place along the leaves of the tree, resulting in the haplotype configuration green, purple, purple, red at the leaves of the tree.

Ancestry as missing data

Let $\mathbf{n}=(n_1,n_2,\dots,n_{|\mathcal{H}|})$ denote the observed allele frequencies of n lineages, and A denote the ancestral tree annotated with mutations. The likelihood can be written

$$\mathbb{P}(\mathbf{n})=\int_{\mathcal{A}}\mathbb{P}(\mathbf{n}|A)\mathbb{P}(dA)=\int\dots\int\mathbb{P}(\mathbf{n}|A)\bigotimes_{i=0}^{-T+1}\mathbb{P}(dA_i|A_{i-1})\mathbb{P}(dA_{-T}),\qquad(1)$$

where $\mathbb{P}(\mathbf{n}|A)=1$ if the leaves of A are compatible with $\mathbf{n}$ and 0 otherwise, T is the number of transitions required to reach the MRCA, A_0 are the leaves of the tree, A_{-T} is the root and the other A_i's are intermediate states separated by mutations or coalescences. For example, in Figure 1 we have $T=5$, $A_0=(G,P,P,R)$, $A_{-1}=(P,P,P,R)$, $A_{-2}=(P,R)$, $A_{-3}=(B,R)$, $A_{-4}=(B,B)$ and $A_{-5}=(B)$, where B,G,P and 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 $\mathbb{Q}$ starting from the data and proceeding backwards in time:

$$\mathbb{P}(\mathbf{n})=\int\dots\int\prod_{i=0}^{-T+1}\frac{\mathbb{P}(A_i|A_{i-1})}{\mathbb{Q}(A_{i-1}|A_i)}\bigotimes_{i=0}^{-T+1}\mathbb{Q}(dA_{i-1}|A_i)\mathbb{P}(dA_{-T}).$$

The factor $\mathbb{P}(\mathbf{n}|A)$ 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 Λ–coalescent conditioned on the observed leaves. This distribution can be written

$$\mathbb{Q}^*(A_{i-1}|A_i)\propto\begin{cases}n_h\frac{\pi(\mathbf{e}_{h'}|A_i-\mathbf{e}_h)}{\pi(\mathbf{e}_h|A_i-\mathbf{e}_h)}\theta P_{h'h} & \text{if } A_{i-1}=A_i-\mathbf{e}_h+\mathbf{e}_{h'} \text{ for } h,h'\in\mathcal{H}\\ \binom{n_h}{k}\frac{\lambda_{n,k}}{\pi((k-1)\mathbf{e}_h|A_i-(k-1)\mathbf{e}_h)} & \text{if } A_{i-1}=A_i-(k-1)\mathbf{e_h} \text{ for } k\in\{2,\dots,n_h\} \text{ and } h\in\mathcal{H}\end{cases},$$

where $\mathbf{e}_h$ is the canonical unit vector in direction h, the n_h's refer to allele frequencies in A_i and $\pi(\cdot|A_i)$ is the conditional sampling distribution (CSD) of further haplotypes given observed configuration A_i ([KJS14], 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 π in $\mathbb{Q}^*$, 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 observed configuration in time so that it undergoes 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 chooses its parent uniformly at random, and inherits the type of the parent. Matching with the coalescence rate of the usual Λ–coalescent suggests an absorption rate of

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

where n now denotes the number of lineages in the conditioning configuration. The resulting approximate CSD $\hat{\pi}$ can be thought of as the stationary distribution of a Markov chain on $\mathcal{H}$ with transition matrix

$$\frac{\theta P+\left[\frac{\Lambda(\{0\})}{2}+\frac{1}{n(n+1)}\sum_{k=2}^{n+1}\binom{n+1}{k}\lambda_{n+1,k}\right]N}{\theta+\frac{\Lambda(\{0\})n}{2}+\frac{1}{n+1}\sum_{k=2}^{n+1}\binom{n+1}{k}\lambda_{n+1,k}},$$

where N is the $|\mathcal{H}|\times|\mathcal{H}|$ matrix with each row equal to $(n_1,n_2,\dots,n_{|\mathcal{H}|})$ ([KJS14], Proposition 1). The multi–haplotype CSDs required in $\mathbb{Q}^*$ 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 $\mathbb{Q}^*$ only requires evaluating multi–haplotype CSDs for configurations where all haplotypes are equal and permutations leave the expression unchanged.

Simulation study

100 lineages were simulated from the Beta$(0.5,1.5)$–coalescent with type space consisting of 15 binary loci: $\mathcal{H}=\{0,1\}^{15}$. The total mutation rate is 0.1, and at a mutation a locus chosen uniformly at random flips. The data contains three haplotypes with frequencies 95, 4 and 1, respectively. Both small blocks differ from the main one at only one locus.

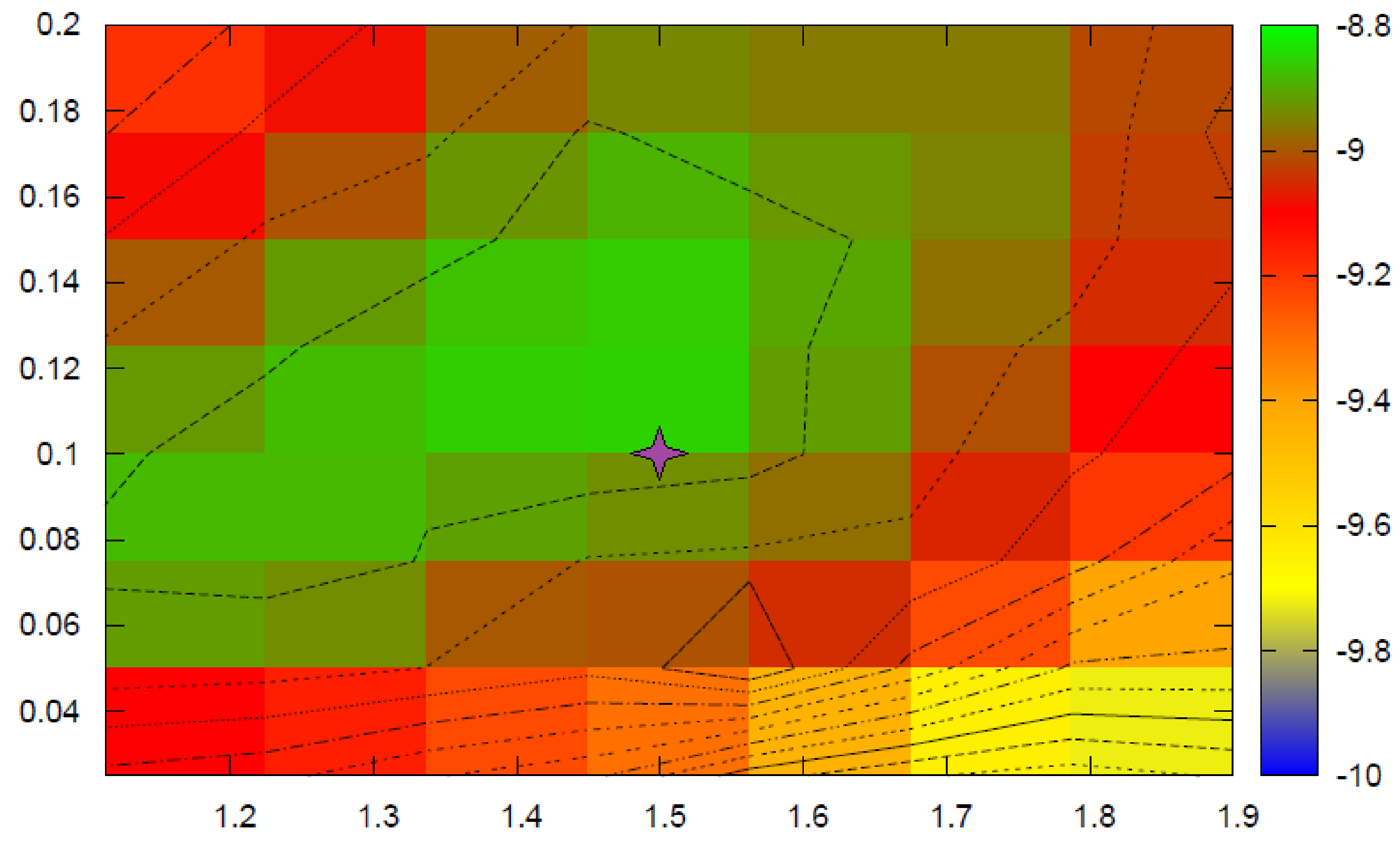

Figure 2 : A likelihood surface interpolated from an 8×8 grid of independent simulations of 30 000 particles each. The mutation rate is on the y–axis and α specifying the Beta$(2-\alpha,\alpha)$–measure is on the x–axis. The purple star denotes the true values. The surface is unimodal around the true value apart from a small, second mode due to noise in the estimators.

Acknowledgements and References

Jere Koskela is a member of the Mathematics and Statistics Doctoral Training Centre, funded by EPSRC grant number EP/HO23364/1.

[Árn04] E. Árnason. Mitochondrial cytochrome b DNA variation in high–fecundity Atlantic cod: trans–Atlantic clines and shallow gene genealogy. *Genetics*, 166:1871–1885, 2004

[BB08] M. Birkner and J. Blath. Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. *J. Math. Biol.*, 57(3):435–463, 2008

[BBS11] M. Birkner, J. Blath and M Steinrücken. Importance sampling for Lambda–coalescents in the infinitely many sites model. *Theor. Popln Biol.*, 79(4):155–173, 2011

[BBB94] J.D.G. Boom, E.G. Boulding and A.T Beckenback. Mitochondrial DNA variation in introduced populations of Pacific oyster, Crassostrea gigas, in British Columbia. *Can. J. Fish. Aquat. Sci.*, 51:1608–1614, 1994

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

[KJS14] J. Koskela, P.A. Jenkins and D. Spanò. Computational inference beyond Kingman's coalescent. *Preprint*, arXiv:1311, 2014

[PS10] J.S. Paul and Y.S. Song. A principled approach to deriving approximate conditional sampling distributions in population genetic models with recombination. *Genetics*, 186:321–338, 2010

[Pit99] J. Pitman. Coalescents with multiple collisions. *Ann. Probab.*, 27(4):1870–1902, 1999

[Sag99] S. Sagitov. The general coalescent with asynchronous mergers of ancestral lineages. *J. Appl. Probab.*, 36(4):1116–1125, 1999

[Sch03] J. Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stoch. Proc. Appl.*, 106:107–139, 2003

[SBB13] M. Steinrücken, M. Birkner and J. Blath. Analysis of DNA sequence variation within marine species using Beta–coalescents. *Theor. Popln Biol.*, 87:15–24, 2013

[SD00] M. Stephens and P. Donnelly. Inference in molecular population genetics. *J. R. Statist. Soc. B*, 62(4):605–655, 2000

MASDOC CDT, University of Warwick Mail: masdoc.info@warwick.ac.uk WWW: http://www2.warwick.ac.uk/fac/sci/masdoc/