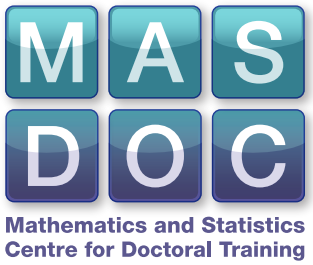


Consistency and intractable likelihood for jump diffusions and generalised coalescent processes

by

Jere Juhani Koskela

Thesis

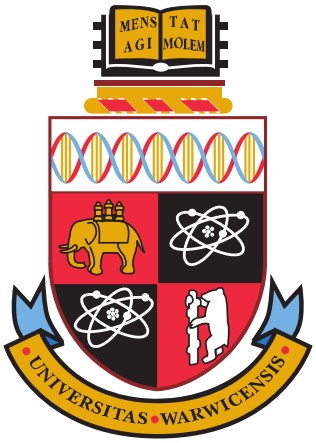
Submitted for the degree of

Doctor of Philosophy

Mathematics Institute

The University of Warwick

November 2016



Contents

List of Tables

List of Figures

Acknowledgments

I am deeply grateful to my supervisors, Dario Spanò and Paul Jenkins, for introducing me to mathematical population genetics, and for their support and guidance throughout my graduate studies. I would also like to thank my examiners, Matthias Birkner and Christian Robert, for significantly improving this thesis through their scrutiny and comments, as well as Julia Brettschneider for her help with the viva arrangements. In addition, thanks to everyone who has helped me develop my understanding the material contained herein; I'll mention Alison Etheridge, Ayalvadi Ganesh, Adam Griffin, Adam Johansen, Jerome Kelleher, Felipe Medina Aguayo, Murray Pollock, Yun Song, and Tim Sullivan, though the list is certainly incomplete. I'm also grateful to the support given to me by the University of Warwick, and the staff in its Mathematics Institute and Department of Statistics. Finally, my thanks to my family and friends for their invaluable support.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented (including data generated and data analysis) was carried out by the author.

Parts of this thesis have been published by the author:

- [Koskela et al., 2015a]
- [Koskela et al., 2015b]
- [Koskela et al., 2015c]
- [Koskela et al., 2016]

Abstract

This thesis has two related aims: establishing tractable conditions for posterior consistency of statistical inference from non-IID data with an intractable likelihood, and developing Monte Carlo methodology for conducting such inference. Two prominent classes of models, jump diffusions and generalised coalescent processes, are considered throughout. Both are motivated by population genetics applications.

Posterior consistency of nonparametric inference is established for joint inference of drift and compound Poisson jump components of unit volatility jump diffusions in arbitrary dimension under an identifiability assumption. This assumption is straightforward to verify in the diffusion case, but difficult to check in general for jump diffusions. A similar consistency result is established under somewhat weaker conditions for Λ-coalescent processes whenever time series data is available. I also show that Λ-coalescent inference cannot be consistent if observations are contemporaneous, in stark contrast to the more classical case of the Kingman coalescent.

I also introduce the notion of *reverse time* sequential Monte Carlo (SMC), which has previously been applied to Kingman and Λ-coalescents. Here, reverse time SMC is presented as a generic algorithm, and general conditions under which it is effective are developed. In brief, it is well suited to integration over paths which begin at a mode of the target distribution, and terminate in the tails. These innovations are used to design new SMC algorithms for generalised coalescent processes, as well as non-coalescent examples including evaluating a containment probability of the hyperbolic diffusion, an overflow probability in a queueing model and finding an initial infection in an epidemic network model.

Abbreviations

- ABC: approximate Bayesian computation
- ATM: asynchronous transfer mode
- CPU: central processing unit
- CSD: conditional sampling distribution
- DNA: deoxyribonucleic acid
- ESS: effective sample size
- GPU: graphical processing unit
- IID: independent and identically distributed
- LHS: left hand side
- MCMC: Markov chain Monte Carlo
- MRCA: most recent common ancestor
- PAC: product of approximate conditionals
- RHS: right hand side
- SDE: stochastic differential equation
- SIS: susceptible-infected-susceptible
- SMC: sequential Monte Carlo

Chapter 1

Introduction

Analysis of the likelihood function has been central to statistics for a century [Fisher, 1912, 1922] due to the fact that it (along with the prior in the Bayesian setting) encodes all of the signal contained in a data set about the data-generating mechanism. Hence analysis of the likelihood yields point estimators of parameter values, confidence or credible sets as well as estimators of any other quantity of interest, along with quantitative information about the accuracy of estimates, at least in principle.

Let $\{\mathbb{P}_\theta, \theta \in \Theta\}$ be a parametric family of statistical models, and $x_{1:n} = (x_1, \ldots, x_n)$ denote an observed data set. The likelihood function is the joint probability

$$L(\theta; x_{1:n}) = \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n),$$

which is not a tractable function in general. When the observations are independent, the likelihood decomposes into the substantially more tractable product form:

$$L(\theta; x_{1:n}) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i).$$

However, there are a myriad of statistical applications in which the independence assumption is either restrictive, or outright false. Asymptotically, the assumption of independence can be relaxed to the much more permissive regularity conditions of *local asymptotic normality* [Le Cam, 1953, 1956, 1960], under which correlated observations can be treated as arising from a joint Gaussian distribution. Thus it is only necessary to estimate the mean vector and covariance matrix under the Gaussian assumption.

However, local asymptotic normality still constrains the scope of possible models, and the associated Gaussian approximation is only valid asymptotically. If the necessary regularity conditions do not hold, or there is insufficient data to justify

an asymptotic analysis, there is no reason to expect the likelihood function to be tractable. Intractable likelihood functions also arise in e.g. statistical mechanics, inference from diffusions and missing data problems, all of which have a wide range of applications.

The tractability of the likelihood function is important for (at least) two reasons:

1. Maximising the likelihood function is a concrete way of obtaining *maximum likelihood estimators*, $\widehat{\theta}$, for parameters, θ.

2. Analysis of the likelihood function is central to proving desirable properties of these estimators, such as consistency, unbiasedness, efficiency etc.

In recent decades the desire to carry out these two procedures for increasingly complex models and data sets have motivated the development of statistical methods which circumvent the need for an exact likelihood function. An incomplete list of examples addressing point 1. includes the celebrated Metropolis-Hastings algorithm [Metropolis et al., 1953; Hastings, 1970], which only requires likelihood evaluations up to a normalising constant; the sequential Monte Carlo [Doucet and Johansen, 2011] and sequential Monte Carlo sampler [Del Moral et al., 2006] algorithms, which are well suited to missing data problems, rare event simulation and filtering; and exact simulation algorithms for inference from partially observed diffusions with intractable transition probabilities [Beskos et al., 2006, 2009].

A minimal requirement for good statistical inference is that the estimator $\widehat{\theta}$ is *consistent*, i.e. that $\widehat{\theta} \to \theta$ as $n \to \infty$, in some appropriate sense. Intuitively, the notion of consistency corresponds to it being possible to learn the truth from data. In the Bayesian setting consistency can be expressed as the posterior distribution, $\mathbb{P}(\theta|x_{1:n}) \propto Q(\theta)\mathbb{P}_\theta(x_{1:n})$, concentrating on a neighbourhood of the parameter which generates the data, where $Q(\theta)$ is the prior. Standard conditions to ensure posterior consistency are formulated in terms of Kullback-Leibler divergences and exponentially consistent hypothesis tests [Schwartz, 1965], which are difficult to verify when the likelihood is intractable. Moreover, many of the natural parameter sets of processes with intractable likelihood are infinite dimensional — consider for example function-valued coefficients of SDEs — and in this *nonparametric* setting posterior consistency is a very delicate property [Diaconis and Freedman, 1986].

The aim of this thesis is twofold:

1. to determine verifiable conditions under which posterior consistency holds for Bayesian nonparametric inference when the likelihood is intractable, and

2. to derive optimised, unbiased sequential Monte Carlo inference algorithms for intractable inference problems.

Both aims are motivated by inference problems in population genetics, where nonparametric inference arises naturally e.g. for the so-called Λ-coalescent family [Pitman, 1999; Sagitov, 1999] and where both Markov chain Monte Carlo [Kuhner et al., 1995; Wilson and Balding, 1998; Felsenstein et al., 1999; Drummond et al., 2002] and sequential Monte Carlo [Griffiths and Tavaré, 1994a,b,c; Stephens and Donnelly, 2000] have a well established role. Despite the motivating application, both results will be presented in some considerable generality: nonparametric consistency for discretely observed jump diffusions as well as Λ-coalescents, and the sequential Monte Carlo algorithms for generic, stopped Markov chains, of which coalescent models are an example. The derivation of the sequential Monte Carlo algorithms will yield very efficient but biased pseudo-likelihood algorithms as a byproduct, and these will also be investigated briefly. Sequential Monte Carlo will be the subject of Chapter 2, while Bayesian nonparametric consistency is developed in Chapter 3.

The key to the sequential Monte Carlo algorithms developed in this thesis will be the notion of time reversal: simulating trajectories of Markov chains in reverse time. This makes it easy to condition the trajectories to hit sets of small, or even zero probability, which makes the methods well suited for rare event simulation. Time reversal is at the core of sequential Monte Carlo inference in population genetics, and the idea of viewing the optimal sequential Monte Carlo algorithm as the time-reversal of the coalescent process has appeared before in [Birkner et al., 2011], but the results of this thesis make the connection transparent enough to be easily generalisable.

1.1 Coalescent processes

The evolution of biological populations is a complex process shaped by the interplay of genetic drift through random mating, mutation, recombination, natural selection and many other forces both external and internal to the population in question. This thesis will focus solely on genetic drift and recurrent mutation, which nevertheless necessitates more degrees of freedom than could feasibly be specified in any comprehensively realistic model of a population. The key to successful modelling in the face of such complexity is robustness: use of models which capture the essential features of genetic evolution regardless of the fine details of the process itself. In this context such robustness is achieved by coalescent processes, which are a core subject of the thesis.

Coalescent processes have been a central tool in population genetic modelling and inference ever since their introduction by Kingman [1982a,b]. Kingman's coalescent is a model of the ancestry of lineages sampled from an infinite, panmictic population undergoing random mating. Ancestral trees are generated by merging each pair of lineages into a common ancestor at rate 1, thus reducing the number of lineages by one. The process terminates once the most recent common ancestor (MRCA) of all sampled lineages is reached, so that a realisation of Kingman's coalescent is a random, binary tree.

Kingman's coalescent is the attractor of a broad class of individual-based, finite population models of evolution. This is the class is conveniently described in terms of Cannings models [Cannings, 1974, 1975]. Consider a stationary population of fixed size $n \in \mathbb{N}$, undergoing random mating in discrete time with non-overlapping generations. At time $t \in \mathbb{N}$ individual $i \in \{1, \dots, n\} =: [n]$ produces a random number $n_i(t)$ of offspring, so that the generation at time $t+1$ is given by the random vector $(n_1(t), \dots, n_n(t))$ with $\sum_{i=1}^n n_i(t) = n$. The population is stationary, and offspring numbers between different generations are assumed independent, so that the vectors $\{n_1(t), \dots, n_n(t)\}_{t\in\mathbb{N}}$ are IID. Further, assume that each vector is exchangeable, so that for any permutation $\sigma \in S_n$ of $[n]$ it holds that

$$(n_1(t), \dots, n_n(t)) \stackrel{d}{=} (n_{\sigma(1)}(t), \dots, n_{\sigma(n)}(t)),$$

where $\stackrel{d}{=}$ indicates equality in distribution.

Kingman's coalescent is obtained by defining the time scale

$$c_n := \frac{\mathbb{E}\left[n_1(1)(n_1(1)-1)\right]}{n-1}$$

and thus the rescaled process

$$(\tilde{n}_1(t), \dots, \tilde{n}_n(t)) := (n_1(\lfloor t/c_n \rfloor), \dots n_n(\lfloor t/c_n \rfloor)). \tag{1.1}$$

If $c_n \to 0$ and

$$\frac{\mathbb{E}\left[n_1(1)(n_1(1)-1)(n_1(1)-2)\right]}{c_n n^2} \to 0 \tag{1.2}$$

as $n \to \infty$, then the rescaled population model (1.1) lies in the domain of attraction of Kingman's coalescent [Möhle and Sagitov, 2001; Birkner and Blath, 2009]. Note that (1.2) enforces the binary nature of Kingman's coalescent trees by ensuring that the probability of three or more lineages merging in one generation vanishes in the limit.

The assumptions of discrete time and non-overlapping generations have been made for ease of exposition. Neither assumption is necessary for obtaining convergence, although the time scale c_n may have to be altered when they don't hold. The assumption of a fixed population size can also be relaxed. For a detailed exposition on Kingman's coalescent and coalescent theory, the interested reader is directed to [Wakeley, 2009] and references therein. In particular, the effect of changing population size, crossover recombination [Griffiths and Marjoram, 1997], natural selection [Krone and Neuhauser, 1997] and spatial structure [Herbots, 1997] on ancestries can all be incorporated into the coalescent framework.

The domain of attraction of Kingman's coalescent is determined by two crucial assumptions: exchangeability of the offspring vectors and the moment conditions $c_n \to 0$ and (1.2). The former has a biological interpretation as a neutral, homogeneous population with no natural selection or population structure, and the latter as small family sizes compared to the size of the whole population. I will focus on two relaxations of these conditions: allowing high fecundity events in which a small number of ancestors give rise to a significant fraction of the whole population in a small number of generations, and spatial structure across a continuous geography. The resulting families of coalescent models under study are, respectively, the Λ- and Ξ-coalescents for high fecundity events, and spatial Λ-coalescents for geographical structure. As with Kingman's coalescent, changing population size, recombination [Birkner et al., 2012; Etheridge and Véber, 2012] and selection [Etheridge et al., 2010, 2014] can be incorporated into both families of coalescents, and a spatially structured version of the Λ-coalescent has also been derived [Heuer and Sturm, 2013]. However, none of these extensions are within the scope of this thesis.

1.1.1 Λ- and Ξ-coalescents

The Λ-coalescents, introduced by Donnelly and Kurtz [1999], Pitman [1999] and Sagitov [1999], generalise Kingman's coalescent by permitting multiple lineages to merge in one event. Such multiple mergers correspond to high fecundity reproduction events, in which a single individual becomes ancestral to a significant fraction of the whole population in a single generation. The merger rate of any k out of n lineages is given by

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

for some finite measure Λ on $[0,1]$, which can be taken to be a probability measure without loss of generality. Λ-coalescents model infinite population ancestries from

Cannings-like models with $c_n \to 0$,

$$\frac{n}{c_n}\mathbb{P}(n_1(1) > nx) \to \int_x^1 r^{-2}\Lambda(dr) \tag{1.3}$$

for $0 < x < 1$, and

$$\frac{\mathbb{E}\left[n_1(1)(n_1(1)-1)n_2(1)(n_2(1)-1)\right]}{c_n n^2} \to 0 \tag{1.4}$$

as $n \to \infty$ [Möhle and Sagitov, 2001]. Note that while (1.3) permits large family sizes and hence mergers involving more than two lineages with positive probability, (1.4) ensures only one merger can take place at any given time. Simultaneous mergers are ruled out.

Popular choices of Λ include $\Lambda = \delta_0$, which corresponds to Kingman's coalescent, $\Lambda = \delta_1$ leading to star-shaped genealogies, $\Lambda = \frac{2}{2+\psi^2}\delta_0 + \frac{\psi^2}{2+\psi^2}\delta_\psi$ where $\psi \in (0,1]$ [Eldon and Wakeley, 2006], $\Lambda = \text{Beta}(2-\alpha, \alpha)$ where $\alpha \in (1,2)$ [Schweinsberg, 2003; Birkner and Blath, 2008; Birkner et al., 2011], and $\Lambda(dr) = c\delta_0(dr) + \frac{1-c}{2}rdr$ where $c \in [0,1]$ [Durrett and Schweinsberg, 2005]. Birkner and Blath [2009] provide a review of Λ-coalescents.

The Λ-coalescents allow multiple mergers, but only permit one merger at a time. They are generalised further by the Ξ-coalescents, which permit any number of simultaneous, multiple mergers. Ξ-coalescents were introduced by Schweinsberg [2000] and Möhle and Sagitov [2001], and can be expressed in terms of a finite measure Ξ on the infinite simplex

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

Again, Ξ can be taken to be a probability measure without loss of generality.

Let $\lambda_{n;k_1,\ldots,k_p;s}$ denote the rate of jumps involving $p \geq 1$ mergers with sizes $k_1, \ldots, k_p$, 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

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

Note that if Ξ assigns full mass to the set $\{\mathbf{r} \in \Delta : r_2 = r_3 = \ldots = 0\}$, the resulting process is a Λ-coalescent.

Ξ-coalescents correspond to Cannings-type models for which $\lim_{n\to\infty} c_n = 0$ and the limits

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[(n_1(1))_{k_1} \dots (n_p(1))_{k_p}\right]}{c_n n^{k_1+\dots+k_p-p}} \tag{1.5}$$

exist for any $p \in \mathbb{N}$ and $k_1, \dots, k_p \in \mathbb{N}$, where $(n)_k := n(n-1)\dots(n-k+1)$ is the falling factorial. Any combination of p simultaneous mergers involving, respectively, $k_1, k_2, \dots, k_p$ lineages is permitted with positive probability provided the corresponding limit (1.5) is positive. The case $\lim_{n\to\infty} c_n = c > 0$ results in discrete time versions of Ξ-coalescents [Möhle and Sagitov, 2001].

1.1.2 Spatial Λ-coalescents

This section presents the spatial Λ-coalescent, introduced by Etheridge [2008] and Barton et al. [2010a] as a generalisation of Kingman's coalescent for structured populations in a continuous geography. Previous generalisations typically incorporate spatial structure by modelling the geography as a graph with panmictic populations at the vertices and migration along edges [Wright, 1931; Kimura, 1953]. Natural population habitats are continuous, which makes an accurate subdivision difficult to specify. The choice of graph structure can also have an effect on inference. Another alternative is the Isolation by Distance model [Wright, 1943; Malécot, 1948], which suffers from the "Pain in the Torus" [Felsenstein, 1975] of either extinction or unstable population growth and clustering. The "Pain in the Torus" can be avoided by local population density regulation which stabilises the population, but typically renders models intractable.

The spatial Λ-coalescent circumvents both these difficulties by being defined on a continuous geography, and achieving local density regulation tractably by modelling reproduction via extinction-recolonisation events driven by a space-time Poisson process, which is independent of the state of the population. For concreteness I focus on a two-dimensional geography, which I take to be a torus of side length $L > 0$ denoted by $\mathbb{T} := \mathbb{T}(L)$. Let $z_{1:n} := (z_1, \dots, z_n) \in \mathbb{T}(L)^n$ be the locations of $n \in \mathbb{N}$ sampled lineages. For simplicity I assume all locations are distinct.

The dynamics of the spatial Λ-coalescent are driven by a Poisson process N on $\mathbb{R} \times \mathbb{T}$ with rate $dt \otimes dx$. The points $(t, x) \in N$ model extinction-recolonisation events, with the two co-ordinates specifying the time and place of the event respectively. Tracing backwards in time, at each point $(t, x) \in N$, all lineages within $B_r(x)$, a closed ball of radius $r > 0$ centred at x, flip a coin with success probability $u \in (0, 1]$. Successful lineages merge to a common ancestor with location sampled uniformly from $B_r(x)$, while lineages which fail their coin flip are unaffected by

the event. Once the whole sample has merged into a single lineage (the MRCA), the process terminates yielding a random tree with nodes labelled by geographical locations and edges denoting jumps in spatial locations and mergers.

This is the so called disc model of the spatial Λ-coalescent process defined by the replacement kernel $u\mathbb{1}_{B_r(x)}(y)$, but it is straightforward to construct variants by using e.g. Gaussian or heavy-tailed replacement around each event centre x. The Gaussian replacement model has been studied in [Barton et al., 2010b], and the heavy-tailed case in [Berestycki et al., 2013].

The spatial Λ-coalescent is also the attractor of high density limits of a broad class of individual-based models. Examples of such families are described in [Etheridge and Kurtz, 2014]. Barton et al. [2013b] provide a review of spatial Λ-coalescents and related processes.

1.1.3 Mutation

In addition to describing ancestral relationships between lineages, coalescent processes can be used to tractably sample genetic types from populations. In the notation of Paul and Song [2010], suppose that the genetic material of an organism of interest is formed of k linearly arranged loci, or a *haplotype*. Suppose the state of a locus $l \in [k]$ can be one of a finite collection of alleles $E_l = \{1, \ldots, |E_l|\}$, and mutates at rate θ_l with mutant alleles sampled from a stochastic matrix $M^{(l)}$. Let $\theta := \sum_{l\in[k]} \theta_l$ denote the total mutation rate, $\mathcal{H} := E_1 \times \ldots \times E_k$ denote the set of possible haplotypes, and let M denote the stochastic matrix on $\mathcal{H}$ formed as a mixture distribution from weights $\{\theta_l/\theta\}_{l\in[k]}$ and mixture components $\{M^{(l)}\}_{l\in[k]}$. Assume also that M has a unique stationary distribution m. For a haplotype $h \in \mathcal{H}$ let $h[l]$ denote the allele at locus l of haplotype h, and let $S_l^a(h)$ denote the haplotype obtained from h by substituting allele a at locus l.

When $\theta > 0$ and M is irreducible, all haplotypes will persist in the population and mutation is called recurrent. Given a realised coalescent tree, haplotypes from a stationary population can be generated by sampling a haplotype for the MRCA from m, and propagating it along the edges of the tree with mutations occurring at rate θ and mutant haplotypes sampled from M. Non-stationary samples can be obtained by changing the distribution of the MRCA haplotype, but I will focus on the stationary case in this thesis.

When incorporating mutation, the coalescent processes introduced above can be viewed as stochastic processes $\Pi := (\Pi_t)_{t\geq 0}$ taking values in $\mathcal{P}_n^{\mathcal{H}}$, the set of $\mathcal{H}$-labelled partitions of $[n]$. I will use Π to refer to a generic coalescent process, Π^Λ for Λ-coalescents, Π^Ξ for Ξ-coalescents and Π^{SL} for spatial Λ-coalescents. The

coalescent starts from the unlabelled, trivial partition $\psi_n := \{\{1\}, \{2\}, \ldots, \{n\}\}$, mergers of lineages correspond to merging the corresponding blocks in the partition, and haplotype labels can be propagated along the realised tree as described above once the MRCA is reached. Of course, in the spatial case it is also necessary to label the partitions with their spatial locations while the lineages are coalescing, and the resulting tree will be labelled both with haplotypes and locations.

As with the coalescent processes, I will denote the resulting laws on spaces of labelled trees by $\mathbf{P}_n^{\Lambda}(\cdot)$, $\mathbf{P}_n^{\Xi}(\cdot)$ and $\mathbf{P}_n^{\mathrm{SL}}(\cdot)$ for Λ-, Ξ- and spatial Λ-coalescents respectively. The corresponding expectations will be denoted by $E_n^{\Lambda}[\cdot]$, $E_n^{\Xi}[\cdot]$ and $E_{z_{1:n}}^{\mathrm{SL}}[\cdot]$, where in the spatial Λ-coalescent case the vector $z_{1:n}$ denotes initial sampling locations. The symbols $\mathbf{P}_n(\cdot)$ and $E_n[\cdot]$ will refer to generic coalescent processes.

The finite alleles model is arguably the most realistic model of mutation, as it mimics the structure of DNA sequences when $E_l = \{A, C, G, T\}$ for each $l \in [k]$. However, it can result in very computationally intensive simulations and inference when the number of loci is large. I will focus on recurrent finite alleles mutation in this thesis, but conclude the section by mentioning some popular alternatives.

The infinite alleles model depicts the allele at each locus as a point along the unit interval. Mutations occur at rate θ, and result in sampling a new allele uniformly, so that all information of the parental allele is lost. This model is coarse, but was often appropriate before modern DNA sequencing became widespread, when it was only possible to determine whether or not two DNA segments were identical.

The infinite sites model is a refinement of the infinite alleles model, in which haplotypes are depicted as a continuous line segment. Mutations occur at rate θ, and result in a mutant allele at a uniformly sampled location along the haplotype. Sampled haplotypes are identified relative to a reference, usually the ancestral haplotype, and their state can be specified by listing all the mutant locations at which the reference and sample differ. Note that no location can ever mutate more than once, which is problematic because real genetic data sets frequently contain loci with three or more observed alleles. However, the infinite sites assumption can be reasonable when the number of sampled loci is large, and the model is computationally more tractable than the finite alleles model. Note also that the infinite alleles model can be obtained from the infinite sites model by simply recording whether or not at least one mutation has taken place along a sampled haplotype.

Finally, the stepwise model depicts the allele at a locus as a repeat count, again measured relative to a reference so that negative counts are possible. As before, mutations occur at rate θ, and result in the repeat count being increased or decreased by a set or random amount. This model is natural for modelling biological

microsatellites, which consist of repeating a fixed pattern of DNA, e.g. AT, a variable number of times, so that the state of a microsatellite locus might be $(AT)^m$ for any $m \in \mathbb{Z}$, again relative to a reference number of repeats.

1.2 Jump diffusions and duality

A successful model in population genetics consists of a historical model of ancestry, as well as a corresponding model of population allele frequencies forwards in time. When endowed with a set of haplotypes $\mathcal{H}$ and inheritance of haplotypes from parents, the Cannings models and their generalisations introduced in Section 1.1 are examples of finite population models of allele frequencies. Infinite population limits are most naturally expressed in terms of measure-valued jump diffusions. In this section I will introduce the Λ-Fleming-Viot, Ξ-Fleming-Viot and spatial Λ-Fleming-Viot processes, corresponding in the obvious way to the coalescent processes outlined in Section 1.1. I will also make the correspondence precise via a duality relation which connects each measure-valued allele frequency process to its corresponding coalescent.

1.2.1 Λ- and Ξ-Fleming-Viot processes

Let $\Delta_{\mathcal{H}} := \{\mathbf{x} \in [0,1]^{|\mathcal{H}|} : \sum_{i=1}^{d} x_i = 1\}$ denote the $|\mathcal{H}|$-dimensional probability simplex. The Λ-Fleming-Viot process $\mathbf{X}^{\Lambda} := (\mathbf{X}_t^{\Lambda})_{t\geq 0}$ with mutation rates $\{\theta_l\}_{l\in[L]}$, mutation matrix M and Λ-measure $\Lambda \in \mathcal{M}_1([0,1])$ is a $\Delta_{\mathcal{H}}$-valued jump diffusion with generator

$$\begin{aligned}\mathcal{G}^{\Lambda} f(\mathbf{x}) =& \frac{\Lambda(\{0\})}{2} \sum_{i,j\in\mathcal{H}} x_i(\delta_{ij} - x_j) \frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) + \theta \sum_{i,j\in\mathcal{H}} x_j(M_{ji} - \delta_{ij}) \frac{\partial}{\partial x_i} f(\mathbf{x}) \\ &+ \sum_{i\in\mathcal{H}} x_i \int_{(0,1]} \left[f((1-r)\mathbf{x} + r\mathbf{e}_i) - f(\mathbf{x}) \right] r^{-2} \Lambda(dr) \end{aligned} \tag{1.6}$$

acting on functions $f \in C^2(\Delta_{\mathcal{H}})$, where $\mathbf{e}_i$ is the canonical unit vector with a 1 in the i^{th} place and zeros elsewhere. It was introduced by Bertoin and Le Gall [2003], and models the distribution of haplotypes $\mathcal{H}$ in a large population undergoing recurrent mutation and random mating with high fecundity reproduction events, in which a single individual becomes ancestral to a non-trivial fraction $r \in (0,1]$ of the whole population. The effect of high fecundity events is modelled by the jump term in (1.6), and gives rise to the multiple mergers in Λ-coalescent ancestries. Without this jump term, i.e. when $\Lambda = \delta_0$, the process $(\mathbf{X}_t^{\Lambda})_{t\geq 0}$ reduces to the classical Wright-Fisher

diffusion (see e.g. [Durrett, 2008], chapters 7 and 8) on $\Delta_{\mathcal{H}}$ with recurrent mutation.

The law of a Λ-Fleming-Viot process with initial condition $\mathbf{x} \in \Delta_{\mathcal{H}}$ will be denoted by $\mathbb{P}^{\Lambda}_{\mathbf{x}}(\cdot)$, and expectation with respect to this law by $\mathbb{E}^{\Lambda}_{\mathbf{x}}[\cdot]$. I will suppress dependence on initial conditions whenever the stationary process is meant. For bounded $f : \Delta_{\mathcal{H}} \mapsto \mathbb{R}$, let $P^{\Lambda}_t f(\mathbf{x}) := \mathbb{E}^{\Lambda}_{\mathbf{x}}[f(\mathbf{X}^{\Lambda}_t)]$ be the associated transition semigroup, $p^{\Lambda}_t(\mathbf{x}, \mathbf{y})d\mathbf{y} := \mathbb{P}^{\Lambda}_{\mathbf{x}}(\mathbf{X}^{\Lambda}_t \in d\mathbf{y})$ be the transition density and $\pi^{\Lambda}(\mathbf{x})d\mathbf{x}$ the corresponding unique stationary density on $\Delta_{\mathcal{H}}$. The transition semigroup is Feller for any $\Lambda \in \mathcal{M}_1([0,1])$ [Bertoin and Le Gall, 2003], and all densities are assumed to exist with respect to a common dominating measure $d\mathbf{x}$.

The following duality between Λ-coalescents and Λ-Fleming-Viot jump diffusions was established in Bertoin and Le Gall [2003]:

$$\mathbb{E}^{\Lambda}\left[\prod_{h\in\mathcal{H}} \mathbf{X}^{\Lambda}_t(h)^{n_h}\right] = E^{\Lambda}_n\left[\prod_{h\in\mathcal{H}} m(h)^{|\Pi^{\Lambda}_t(h)|}\right], \tag{1.7}$$

where n_h denotes the number of observed individuals of haplotype $h \in \mathcal{H}$ sampled IID from the random measure $\mathbf{X}^{\Lambda}_t$, and $|\Pi^{\Lambda}_t(h)|$ denotes the number of blocks in partition Π^{Λ}_t of haplotype $h \in \mathcal{H}$. In words, it states that the distribution of the allele frequencies generated by a Λ-coalescent started from ψ_n coincides with the distribution of a multinomial sample drawn from the corresponding stationary Λ-Fleming-Viot process.

Birkner et al. [2009] constructed the Ξ-Fleming-Viot process, and established the same duality between it and the Ξ-coalescent. The Ξ-Fleming-Viot process is a jump diffusion similar to the Λ-Fleming-Viot process, but with a wider class of possible jumps reflecting the more general mergers of the Ξ-coalescent. I denote the Ξ-Fleming-Viot process by $\mathbf{X}^{\Xi} := (\mathbf{X}^{\Xi}_t)_{t\geq 0}$.

With state space as above, the Ξ-Fleming-Viot process is the jump diffusion with generator

$$\mathcal{G}^{\Xi} f(\mathbf{x}) = \frac{\Xi(\{\mathbf{0}\})}{2} \sum_{i,j\in\mathcal{H}} x_i(\delta_{ij} - x_j)\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) + \theta \sum_{i,j\in\mathcal{H}} x_j(M_{ji} - \delta_{ij})\frac{\partial}{\partial x_i} f(\mathbf{x})$$
$$\int_{\Delta}\int_{\mathcal{H}^{\mathbb{N}}} \left[f\left(1 - \|\mathbf{r}\|_1\mathbf{x} + \sum_{i=1}^{\infty} r_i \mathbf{e}_{h'_i}\right) - f(\mathbf{x})\right] \left(\sum_{h\in\mathcal{H}} x_h \delta_h\right)^{\otimes\mathbb{N}} (d\mathbf{h}')\frac{\Xi(d\mathbf{r})}{\|\mathbf{r}\|_2^2},$$

where Ξ is a probability measure on the infinite simplex

$$\Delta := \left\{\mathbf{r} \in [0,1]^{\mathbb{N}} : \sum_{i=1}^{\infty} r_i = 1\right\},$$

and $\|\mathbf{r}\|_p$ denotes the p-norm of $\mathbf{r}$. For more details, including the specification of a suitable class of test functions, see Proposition 1.3 of [Birkner et al., 2009]. The law, expectation, transition semigroup, density and stationary distribution of the Ξ-Fleming-Viot process will be denoted by $\mathbb{P}^{\Xi}_{\mathbf{x}}(\cdot)$, $\mathbb{E}^{\Xi}_{\mathbf{x}}[\cdot]$, P^{Ξ}_t, $p^{\Xi}_t(\mathbf{x},\mathbf{y})d\mathbf{y}$ and $\pi^{\Xi}(\mathbf{x})d\mathbf{x}$, respectively.

1.2.2 Spatial Λ-Fleming-Viot processes

The spatial Λ-Fleming-Viot process is the analogue of the Λ-Fleming-Viot process, and describes the allele frequencies at each point in the continuous geography. It was introduced by Barton et al. [2010a], who also established a duality between the spatial Λ-Fleming-Viot process and the spatial Λ-coalescent.

Recall the Poisson process N from Section 1.1.2. The spatial Λ-Fleming-Viot process $\mathbf{X}^{\mathrm{SL}} := (\mathbf{X}^{\mathrm{SL}}_t(x,\cdot))_{x\in\mathbb{T},t\geq 0}$ specifies a probability measure on $\Delta_{\mathcal{H}}$ at each location $x \in \mathbb{T}$ and time $t \geq 0$. This probability measure describes the allele frequencies in the population at that location and time.

The dynamics of $\mathbf{X}^{\mathrm{SL}}$ are driven by N, in that at each point $(t,x) \in N$ a parental location z is sampled uniformly from $B_r(x)$, and a parental haplotype h from $\mathbf{X}^{\mathrm{SL}}(t,z,\cdot)$. The surface $\mathbf{X}^{\mathrm{SL}}$ then undergoes the update

$$\mathbf{X}^{\mathrm{SL}}_t(y,\cdot) = \begin{cases} (1-u)\mathbf{X}^{\mathrm{SL}}_{t-}(y,\cdot) + u\delta_h(\cdot) & \text{if } y \in B_r(x) \\ \mathbf{X}^{\mathrm{SL}}_{t-}(y,\cdot) & \text{otherwise.} \end{cases}$$

The spatial analogue of the duality relation (1.7) has precisely the same interpretation as before: the distribution of a sample drawn at locations $z_{1:n}$ from a stationary Λ-Fleming-Viot process coincides with that obtained by running a spatial Λ-coalescent from those locations until merging to the MRCA, sampling an ancestral type from m and propagating types along the coalescent tree with mutations at rate θ sampled from M.

The duality relation (1.7), and its generalisations will play a central role in designing sequential Monte Carlo inference algorithms for evaluating likelihoods of multinomial samples from the allele frequency processes. They will also prove convenient in proving the nonparametric consistency results of Section 3.3.1.

1.2.3 General jump diffusions

In Chapter 3 I will show that that the consistency results for Λ-coalescents generalise naturally to consistency results for more general jump diffusions. Hence, this

section introduces the general formulation of a time homogeneous jump diffusion on a domain $\Omega \subseteq \mathbb{R}^d$ in preparation for stating these results.

Jump diffusions are a broad wide class of stochastic processes encompassing systems undergoing deterministic mean-field dynamics, microscopic diffusion and macroscopic jumps. Jump diffusions are used as models across broad spectrum of applications, such as economics and finance [Merton, 1976; Aase and Guttorp, 1987; Bardhan and Chao, 1993; Chen and Filipović, 2005; Filipović et al., 2007], biology [Kallianpur, 1992; Kallianpur and Xiong, 1994; Bertoin and Le Gall, 2003; Birkner et al., 2009] and engineering [Au et al., 1982; Bodo et al., 1987]. They also contain many important families of stochastic processes as special cases, including diffusions and Lévy processes.

A general time-homogeneous, d-dimensional jump diffusion $\mathbf{X} := (\mathbf{X}_t)_{t\geq 0}$ is the solution of a stochastic differential equation of the form

$$\begin{aligned} d\mathbf{X}_t &= b(\mathbf{X}_t)dt + \sigma(\mathbf{X}_t)d\mathbf{W}_t + c(\mathbf{X}_{t-}, d\mathbf{Z}_t) \\ \mathbf{X}_0 &= \mathbf{x}_0 \end{aligned} \tag{1.8}$$

where $\sigma : \Omega \mapsto \mathbb{R}^{d\times d}$ is known as the diffusion coefficient, $b : \Omega \mapsto \mathbb{R}^d$ as the drift and $c : \Omega \times \mathbb{R}^d \mapsto \mathbb{R}^d_0$ as the jump coefficient. The process $\mathbf{W} := (\mathbf{W}_t)_{t\geq 0}$ is a d-dimensional Brownian motion and $\mathbf{Z} := (\mathbf{Z}_t)_{t\geq 0}$ is a pure jump Lévy process with jumps in $\mathbb{R}^d_0 := \mathbb{R}^d \setminus \{\mathbf{0}\}$. The process $\mathbf{Z}$ is taken to be independent of $\mathbf{W}$, and has Lévy measure $M(d\mathbf{z})$ satisfying

$$\int_{\mathbb{R}^d_0} (\|c(\mathbf{x}, \mathbf{z})\|_2^2 \wedge 1) M(d\mathbf{z}) < \infty$$

for any $\mathbf{x} \in \mathbb{R}^d$. Note that the space in which $\mathbf{Z}$ takes values can also be a more general Lusin space [El Karoui and Lepeltier, 1977], but this possibility is not considered in this thesis for ease of notation.

Under regularity conditions summarised below, jump diffusions are recurrent, ergodic Feller-Markov processes with transition densities $p_t(\mathbf{x}, \mathbf{y})d\mathbf{y}$ and a unique stationary density $\pi(\mathbf{x})d\mathbf{x}$ with respect to the d-dimensional Lebesgue measure. For diffusions such conditions are widely available in standard textbooks, but the same is not true of jump diffusions, even though sufficient conditions are known. Hence I conclude this section by formalising them into a proposition under the simplifying assumption $\sigma \equiv \mathrm{Id}$. This is a strong assumption whenever $d > 1$, though some models which fail to satisfy it outright can still be treated via the Lamperti transform [Aït-Sahalia, 2008]. Sufficient conditions for the Lamperti transform to

be well defined are non-singularity of σ and the following symmetry condition [Yu, 2007; Aït-Sahalia, 2008]:

$$\frac{\partial(\sigma^{-1})_{ij}(\mathbf{x})}{\partial x_k} = \frac{\partial(\sigma^{-1})_{ik}(\mathbf{x})}{\partial x_j} \text{ for all } i,j,k \in \{1,\dots,d\}.$$

Whenever well defined, this transformation maps a diffusion with general σ to one with unit diffusion but altered drift and jump size distribution. Results which can be deduced for the transformed diffusion must then be mapped by via an inverse Lamperti transform to the original problem.

Proposition 1. *Assume that $c(\cdot,0) \equiv 0$, and that there exist constants $C_1, C_2, C_3, C_4 > 0$ such that*

$$\|b(\mathbf{x}) - b(\mathbf{y})\|_2^2 + \int_{\mathbb{R}_0^d} \|c(\mathbf{x},\mathbf{z}) - c(\mathbf{y},\mathbf{z})\|_2^2 M(d\mathbf{z}) \leq C_1 \|\mathbf{x}-\mathbf{y}\|_2^2 \tag{1.9}$$

$$\|c(\mathbf{x},\mathbf{z}) - c(\mathbf{x},\boldsymbol{\xi})\|_2^2 \leq C_2 \|\mathbf{z}-\boldsymbol{\xi}\|_2^2 \tag{1.10}$$

For every $\mathbf{x} \in \Omega : \|\mathbf{x}\|_2 > C_3$ the following holds: $\mathbf{x} \cdot b(\mathbf{x}) \leq -C_4 \|\mathbf{x}\|_2^2$ (1.11)

$$\int_{\mathbb{R}_0^d : \|\mathbf{z}\|_2 > 1} \|\mathbf{z}\|_2^2 M(dz) < \infty. \tag{1.12}$$

Then (1.8) *has a unique, ergodic weak solution $\mathbf{X}$ with the Feller and Markov properties. Furthermore, $\mathbf{X}$ has a unique stationary density $\pi^{b,\nu}(\mathbf{x})d\mathbf{x}$ with a finite second moment, and the associated semigroup $P_t^{b,\nu}$ has transition densities $p_t^{b,\nu}(\mathbf{x},\mathbf{y})d\mathbf{y}$.*

Proof. Existence and uniqueness of $\mathbf{X}$ are obtained from (1.9), as well as the linear growth bounds implied by Lipschitz continuity, by Theorem 6.2.9 of [Applebaum, 2004]. Theorem 6.4.6 of [Applebaum, 2004] gives the Markov property under the same conditions. Finally, the corollary in Appendix 1 of [Kolokoltsov, 2004] yields the Feller property. In turn, the Feller property and the fact that $\log(1 + \|\boldsymbol{\xi}\|_2)^{-1}\|\boldsymbol{\xi}\|_2^2 \to \infty$ as $\|\boldsymbol{\xi}\|_2 \to \infty$ mean that the hypotheses of Theorem 1.1 of [Schilling and Wang, 2013] are fulfilled, so that $\mathbf{X}$ has bounded transition densities with respect to the Lebesgue measure.

Existence and uniqueness of $\pi^{b,\nu}$, as well as ergodicity of $\mathbf{X}$ will follow from Theorem 2.1 of [Masuda, 2007], the hypotheses of which will now be verified. Along with $c(\cdot,0) \equiv 0$, conditions (1.9) and (1.10) above imply Assumption 1 of [Masuda, 2007]. Now, for every $u \in (0,1)$ let

$$b^u(\mathbf{x}) := b(\mathbf{x}) - \int_{u < \|\mathbf{z}\|_1 \leq 1} c(\mathbf{x},\mathbf{z}) M(d\mathbf{z}).$$

Assumption 2(a)' of [Masuda, 2009] requires $\mathbf{X}$ to admit bounded transition densities, and the diffusion which solves

$$d\mathbf{X}_t^u = b^u(\mathbf{X}_t^u)dt + \sigma(\mathbf{X}_t^u)d\mathbf{W}_t$$

to be irreducible for each $u > 0$. Boundedness of the transition density of $\mathbf{X}$ was established above, and irreducibility of $\mathbf{X}^u$ holds because $\sigma \equiv 1$ by Theorem 2.3 of [Stramer and Tweedie, 1997].

Next, I will verify Assumptions 3 and 3* of [Masuda, 2007] by checking the conditions of Lemma 2.4' of [Masuda, 2009]. The diffusion coefficient is constant, and hence $o(\|x\|_2^{1-q/2})$ for any $q \in (0, 2)$. Condition (1.12) is the corresponding hypothesis of [Masuda, 2009], and both $\|\mathbf{x}\|_2^{q-2}\mathbf{x} \cdot b(\mathbf{x}) \to -\infty$ and $\|\mathbf{x}\|_2^{-2}\mathbf{x} \cdot b(\mathbf{x}) \leq -C_4$ follow from (1.11). Hence, Assumptions 3 and 3* of [Masuda, 2007] hold. This yields ergodicity (and mixing) by Theorem 2.1 of [Masuda, 2007], and second moments of the stationary distribution (and exponential mixing) by Theorem 2.2 of [Masuda, 2007].

It remains to show the invariant measure has a density. By combining Proposition 5.1.9 and Theorem 5.1.8 of [Fornaro, 2004] it can be seen that invariant measures of irreducible strong Feller processes are equivalent to the associated transition probabilities, which is sufficient in this case. Assumption 1 of [Masuda, 2007] and Assumption 2(a)' of [Masuda, 2009] imply irreducibility of $\mathbf{X}$ (c.f. Claim 1 on page 42 of [Masuda, 2007]). Condition (1.9) guarantees the strong Feller property by Theorem 2.3 of [Wang, 2010]. Hence the invariant measure has a density with respect to the transition densities, and thus also the Lebesgue measure. This concludes the proof. □

1.3 Sequential Monte Carlo

Sequential Monte Carlo (SMC) is a very general technique for sampling from a sequence of complicated distributions of increasing dimension and known pointwise up to a normalising constant; for an introduction see e.g. Doucet et al. [2001]; Liu [2001]; Doucet and Johansen [2011]. Briefly, a "cloud" of weighted particles is extended from one distribution to the next by a combination of sequential importance sampling and resampling. Each set of weighted particles then forms an empirical approximation of each subsequent distribution, provided adjacent distributions are sufficiently similar. This is typically the case when interest is in inference from a sequence of observations from e.g. a hidden Markov model, and thus SMC finds widespread use in this context, known as filtering (see e.g. Doucet et al. [2001];

Liu [2001]; Del Moral [2004]; Fearnhead [2008]; Doucet and Johansen [2011], and references therein).

Sequential importance sampling consists of sampling from a sequence of proposal distributions to build up a single, high-dimensional realisation. The proposals are typically not the conditional distributions of the model of interest, and so samples must be reweighted by the Radon-Nikodym derivative of the model and the proposal. Let $X_{1:n} := (X_1, \ldots, X_n)$ be a random vector with law P, and suppose it is of interest to evaluate an intractable functional

$$\begin{aligned}\mathbb{E}[f(X_{1:n})] &= \int f(x_{1:n}) P(dx_{1:n}) \\ &= \int \ldots \int f(x_{1:n}) \bigotimes_{i=1}^{n} P(dx_i | X_{1:i-1} = x_{1:i-1}), \end{aligned} \tag{1.13}$$

with the convention that $x_{1:0} = 0$. This expectation can be approximated by the sample mean of function evaluations of f on data $\{x_{1:n}^{(i)}\}_{i=1}^{k}$ with $x_{i:n}^{(j)} \overset{\text{IID}}{\sim} P$, but this approach can lead to very high variance if the dominant contributions to the integral are from regions which are unlikely under P. Variance can be reduced by introducing a proposal distribution Q with $P \ll Q$, and estimating (1.13) with

$$\widehat{I} := \frac{1}{k} \sum_{j=1}^{k} f(x_{1:n}^{(j)}) \bigotimes_{i=1}^{n} \frac{P(dx_i^{(j)} | X_{1:i-1}^{(j)} = x_{1:i-1}^{(j)})}{Q(dx_i^{(j)} | X_{1:i-1}^{(j)} = x_{1:i-1}^{(j)})} =: \frac{1}{k} \sum_{j=1}^{k} f(x_{1:n}^{(j)}) w_n(x_{1:n}^{(j)}),$$

with $x_{1:n}^{(j)} \overset{\text{IID}}{\sim} Q$.

Sequential Monte Carlo involves the combination of sequential importance sampling with a resampling step, where particles $\{x_{1:n}^{(j)}\}_{j=1}^{k}$ and weights $\{w(x_{1:n}^{(j)})\}_{j=1}^{k}$ are built up in parallel. The weighted collection can then be resampled at intermediary steps to discard particles with low weight and duplicate promising ones with high weight. Good choices of Q and resampling schedule can dramatically reduce the variance of estimators, and can be shown to be asymptotically efficient under mild conditions [Cérou et al., 2011]. On the other hand, poor choices of Q can yield estimators with higher variance than naïve Monte Carlo [Glasserman and Wang, 1997].

Good design of a resampling schedule is also crucial, as without resampling the variance of estimators typically increases exponentially in the number of sequential steps [Doucet et al., 2001; Liu, 2001]. In the simplest case particles are resampled multinomially between the i^{th} and $(i+1)^{\text{th}}$ sequential step with probabilities proportional to their respective importance weights $\{w(x_{1:i}^{(j)})\}_{j=1}^{k}$. Then all weights are

equalised, subject to preserving the total weights of the particle ensemble, and the resulting collection of particles is treated as the starting point of generating the $(i+1)^{\text{th}}$ sample. While intuitive, multinomial resampling is outperformed in terms of variance by other, more complicated resampling mechanisms [Douc et al., 2005]. It is also typically detrimental to resample at every stage of the algorithm. The most common heuristics for when resampling should be performed are based on the effective sample size (ESS) of the ensemble falling below a specified threshold [Kong et al., 1994].

1.3.1 SMC for coalescent processes

It is of great interest to estimate parameters describing the evolutionary history of a population from a sample of DNA sequences. This can be done by assuming an appropriate coalescent model, writing down a likelihood as a function of the parameters and evaluating the likelihood for various parameter values. Such an approach is common to both frequentist and Bayesian analyses, with the additional stage of prescribing a prior distribution on parameters in the Bayesian case.

Let $\mathbf{n} \in \mathbb{N}^{|\mathcal{H}|}$ denote an observed configuration of allele frequencies of size $n := \sum_{h\in\mathcal{H}} n_h$, and $\theta \in \Theta$ be a parameter value and a parameter space, respectively. Let $A \in \mathcal{A}$ denote, respectively, a realisation of the coalescent ancestry and the space of possible ancestries, i.e. the support of $\mathbf{P}_n$. For $A \in \mathcal{A}$ define $p_{\mathbf{n}}(A) = 1$ if the haplotypes at the leaves of A gives rise to allele frequencies $\mathbf{n}$, and $p_{\mathbf{n}}(A) = 0$ otherwise. The likelihood $L(\theta; \mathbf{n})$ can be decomposed by conditioning on the ancestry as

$$L(\theta; \mathbf{n}) = E_n\left[\mathbb{1}_{\mathbf{n}}(A)|\theta\right] = \int_{\mathcal{A}} p_{\mathbf{n}}(A)\mathbf{P}_n(A|\theta)dA \tag{1.14}$$

The likelihood function L will be endowed with a superscript Λ, Ξ or SL when a particular coalescent model is meant.

The space of coalescent trees is prohibitively large for all but trivially small data sets, so in practice the integral on the RHS of (1.14) is approximated by Monte Carlo, or replaced by a pseudo-likelihood from a simpler model. A naïve Monte Carlo approximation is obtained by simulating IID ancestries $A_1, \ldots, A_p$ from $\mathbf{P}_n$, and counting the fraction that yield alleles at the leaves of the tree compatible with $\mathbf{n}$. The problem with this approach is that the hitting probability of leaves compatible with $\mathbf{n}$ is vanishingly small for any realistic data sets, so that the number of samples required for stable estimates is prohibitively large, and SMC is a technique which has been successfully used to overcome this difficulty to some extent. Alternative approaches include Markov chain Monte Carlo (MCMC), approximate Bayesian

computation, and pseudo-likelihood methods, which will be outlined in the next section. Developing good approximations to (1.14) for large, whole-genome data sets under any but the simplest genetic models remains an open problem.

SMC 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 [1994a,b,c, 1999], who derived a recursion for quantities of interest under Kingman's coalescent [Kingman, 1982a,b] and simulated a Markov chain to approximate its solution. Their approach was identified as importance sampling by Felsenstein et al. [1999]. SMC has been investigated and applied to genetic problems such as demographic and other parameter inferences [Griffiths and Marjoram, 1996; Fearnhead and Donnelly, 2001; De Iorio et al., 2005; Griffiths et al., 2008; Gorur and Teh, 2008; Hobolth et al., 2008; Jenkins and Griffiths, 2011].

In brief, the idea of SMC for coalescent processes is to sample the ancestry A sequentially in reverse time, so that every sample will be compatible with the leaves and promising ancestries can be prioritised in favour of ones which are incompatible with the dynamics of the coalescent. Let $\{A_i\}_{i=0}^{K}$ denote the state of the ancestry i mutation or coalescence events into the past, so that A_0 denotes the coalescent leaves, A_K denotes the haplotype of the MRCA and the other A_i's are intermediate configurations along the tree. Let $\mathbb{Q}$ be a proposal distribution satisfying $\text{supp}(\mathbf{P}_n) \subseteq \text{supp}(\mathbb{Q})$. Then (1.14) can be decomposed into the sequential updates

$$L(\theta; \mathbf{n}) = \int_{\mathcal{A}} m(A_K) \prod_{i=0}^{K-1} \frac{\mathbf{P}_n(A_i|A_{i+1}, \theta)}{\mathbb{Q}(A_{i+1}|A_i, \theta)} \mathbb{Q}(dA_{i+1}|A_i, \theta) \tag{1.15}$$

Note that the indicator function $p_{\mathbf{n}}(A)$ is no longer needed as all trees will be compatible with the data by construction. Coalescent histories $\{\{A_i^{(j)}\}_{i=0}^{K_j}\}_{j=1}^{p}$ can now be sampled according to the proposal distribution $\mathbb{Q}$, and the likelihood can be approximated by the weighted average $\widehat{L}(\theta; \mathbf{n}) = p^{-1} \sum_{j=1}^{p} w_{K_j}$ where $w_K = \frac{\mathbf{P}_n(A|A_K, \theta)}{\mathbb{Q}(A|\theta)} m(A_K)$ is the importance weight associated with the j^{th} fully reconstructed coalescent tree.

Choice of proposal distribution $\mathbb{Q}$ is crucially important to the efficiency of the SMC algorithm. The optimal proposal (in terms of estimator variance) is the conditional distribution $\mathbf{P}_n(A|\mathbf{n})$, in which case all particle weights equal the true likelihood by Bayes' theorem:

$$\frac{\mathbf{P}_n(A|A_K, \theta) m(A_K)}{\mathbf{P}_n(A|\mathbf{n}, \theta)} = \frac{\mathbf{P}_n(A|\theta)\mathbf{P}_n(\mathbf{n}|\theta)}{\mathbf{P}_n(\mathbf{n}|A, \theta)\mathbf{P}_n(A|\theta)} = \mathbf{P}_n(\mathbf{n}|\theta),$$

since $\mathbf{P}_n(\mathbf{n}|A, \theta) = 1$ by construction. The resulting estimator is exact with proba-

bility one, and has zero variance.

Unfortunately, the conditional distribution is as intractable as the likelihood. The typical approach of approximating the optimal proposal distribution using large deviations [Sadowsky and Bucklew, 1990] is also rendered intractable by the high dimension of the space of coalescent trees. Progress can be made by designing proposal distributions which mimic the features of the optimal proposal distribution. This was done for Kingman's coalescent by Stephens and Donnelly [2000], who expressed the optimal sequential proposal distribution in terms of a family of conditional sampling distributions (CSDs). The CSDs are also intractable in general, but the authors introduced an approximation which yielded dramatic improvements in efficiency and accuracy in comparison to earlier SMC algorithms. However, even optimised SMC algorithms fail to scale to modern data sets, so that scalable but biased methods have also received much interest.

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 by Li and Stephens [2003]. The PAC algorithm is an example of the scalable but biased methods mentioned above. De Iorio and Griffiths [2004a,b] derived an approximation to finite alleles CSDs based on the Fleming-Viot generator while Paul and Song [2010] provided a genealogical interpretation and included crossover recombination. Further approximations based on hidden Markov models have been obtained by Paul et al. [2011] and Steinrücken et al. [2013b], and applied by Sheehan et al. [2013].

In the context of coalescent processes, resampling should take into account the weight of the particle *and* the progress it has made towards the MRCA. This can be achieved by introducing a sequence of intermediate sets; propagating all samples until they hit the next set; and performing resampling based on current weights once all particles have been stopped. This has been alternatively termed multilevel SMC or stopping-time resampling in Section 12.2 of [Del Moral, 2004] and in [Chen et al., 2005] respectively. It is natural to define the sets based on the number of coalescence and mutation events encountered by the partially reconstructed tree. This approach was investigated by Chen et al. [2005] and Jenkins [2012], and found to yield dramatic improvements to the accuracy of SMC algorithms.

1.3.2 Alternatives to SMC

In this section I will review three other computational approaches to population genetic inference: Markov chain Monte Carlo (MCMC), approximate Bayesian com-

putation (ABC) and product of approximate conditionals (PAC). I will present a brief simulation study of the PAC algorithm for Λ-coalescents in Section 2.4, for which this serves as an introduction, and the other two methods are included for completeness. For further details of MCMC and ABC for population genetics, the interested reader is directed to [Marjoram and Tavaré, 2006], and references therein.

MCMC involves constructing a Markov chain on $\mathcal{A}$ whose stationary distribution is the conditional law of the coalescent given the observed data, $\mathbf{P}_n(\cdot|\mathbf{n})$. The a simulated trajectory of this Markov chain can be used as an autocorrelated sample from $\mathbf{P}_n(\cdot|\mathbf{n})$. Provided the chain has been run sufficiently long, the collection of samples closely approximates a sample from the target posterior distribution.

A typical implementation of this idea is the Metropolis-Hastings algorithm [Metropolis et al., 1953; Hastings, 1970], which is based on an arbitrary transition kernel $q : \mathcal{A} \mapsto \mathcal{M}_1(\mathcal{A})$. If the current state of the chain is A, a step is generated by sampling a proposal $A' \sim q(A, \cdot)$, and the proposed step is accepted with probability

$$1 \wedge \frac{\mathbf{P}_n(A'|\mathbf{n})q(A', A)}{\mathbf{P}_n(A|\mathbf{n})q(A, A')}.$$

The ratio of conditional distributions can be evaluated by Bayes' theorem, even though the numerator and denominator individually are intractable. If the move is rejected then the chain remains at A. This algorithm can be extended to parameter inference by extending the state space of the chain to the product space $\mathcal{A} \times \Theta$, introducing transition kernels on the extended space, and marginalising the resulting sample over $\mathcal{A}$.

Kuhner et al. [1995] and Felsenstein et al. [1999] developed a Metropolis-Hastings algorithm for Kingman's coalescent in which moves are proposed by sampling a node uniformly at random in the coalescent tree, disconnecting it from it's parent and reattaching it after an exponentially distributed waiting time to a uniformly sampled parent edge alive at the time of the merger, or the MRCA if it survives past the current MRCA. Other proposal mechanisms have also been investigated for Kingman's coalescent, and it's various generalisations [Wilson and Balding, 1998; Nielsen, 2000; Wilson et al., 2003].

Like SMC, MCMC algorithms are unbiased but suffer the high dimensionality of the space of coalescent trees. Hence faster, heuristic methods have become prominent for modern genetic data sets. ABC and PAC algorithms are two examples of such heuristic methods.

The idea of ABC was introduced by Tavaré et al. [1997] and developed fully by Pritchard et al. [1999]. In the simplest setting it consists of proposing parameters

$\theta \in \Theta$ from a prior $Q(d\theta)$, and then data $\mathbf{n}|\theta \sim L(d\mathbf{n}|\theta)$ from the model. The proposal is accepted if $\mathbf{n} = \mathbf{n}^*$, where $\mathbf{n}^*$ denotes the observation. The distribution of the accepted proposals is the posterior $Q(d\theta|\mathbf{n})$.

However, the data is often extremely high-dimensional and hence the probability that $\mathbf{n} = \mathbf{n}^*$ is vanishingly small, or 0 when simulated data has a density. The innovation of the ABC algorithm is to introduce a summary statistic $S(\mathbf{n})$ and a tolerance $\varepsilon > 0$, and accept the proposal when $\|S(\mathbf{n}) - S(\mathbf{n}^*)\| < \varepsilon$ in some norm. The algorithm is exact if S is a sufficient statistic and $\varepsilon = 0$, but typically neither of these requirements is feasible and the resulting algorithm is approximate. The basic rejection ABC algorithm has been improved upon in numerous ways [Beaumont et al., 2002; Marjoram et al., 2003; Sisson et al., 2007; Beaumont et al., 2009]. The interested reader is directed to [Beaumont, 2010] for a review of ABC algorithms applied to evolutionary and ecological problems.

The PAC algorithm of Li and Stephens [2003] is based on a decomposition of the likelihood into a product of CSDs, and substituting a tractable, heuristic version of the CSDs in place of the intractable, true CSDs. Suppose that observed allele frequencies $\mathbf{n}$ are generated by haplotypes $h_{1:n} = (h_1, \ldots, h_n) \in \mathcal{H}^n$. Then the likelihood can be written as

$$L(\theta|\mathbf{n}) = \mathbf{P}_n(h_{1:n}|\theta) = \pi(h_n|h_{1:n-1}, \theta)\pi(h_{n-1}|h_{1:n-2}, \theta) \ldots \pi(h_2|h_1, \theta)\pi(h_1, \theta),$$

where $\pi(h|h_{1:n})$ denotes the distribution of the type of the $(n+1)^{\text{th}}$ leaf of a coalescent tree, given the types of the first n leaves.

Approximate CSDs are typically not exchangeable so that the value of the approximation depends on the choice of ordering of haplotypes. This difficulty could be overcome completely by averaging over all possible orderings, but the computational cost is too high for practical data sets. Instead, a common approach is to average over a small, random subset of permutations [Li and Stephens, 2003]. Coalescent processes and their dual diffusions have been instrumental in designing heuristic approximations to CSDs [Fearnhead and Donnelly, 2001; De Iorio and Griffiths, 2004a,b; Paul and Song, 2010; Paul et al., 2011; Sheehan et al., 2013; Steinrücken et al., 2013b], and the resulting PAC algorithms are widely used due to their scalability.

1.4 Bayesian nonparametric inference

As outlined above, Bayesian inference rests on the specification of a prior distribution Q on a set of models $\{\mathbb{P}_\theta : \theta \in \Theta\}$. This prior represents existing information or

beliefs about the parameter set Θ. The nonparametric development is to allow Θ to be infinite dimensional, yielding greater modelling flexibility at the cost of greater analytical and computational challenges. Typical nonparametric model sets include spaces of functions (e.g. $C_b^2(\Omega)$ for some $\Omega \subseteq \mathbb{R}^d$) or measures (e.g. $\mathcal{M}_1([0,1])$). I will abuse terminology and refer to elements of such spaces as "parameters", with the understanding that they may be infinite dimensional.

Given $n \in \mathbb{N}$ data points $x_{1:n}$, the central object of Bayesian inference is the posterior distribution, defined for sets $A \subset \Theta$ as

$$Q(A|x_{1:n}) := \frac{\int_A \mathbb{P}_\theta(x_{1:n})Q(d\theta)}{\int_\Theta \mathbb{P}_\theta(x_{1:n})Q(d\theta)}. \tag{1.16}$$

It is well known that in the Bayesian setting the posterior contains all the information about the parameter carried by the data. In the nonparametric setting neither existence nor uniqueness of the posterior is guaranteed. When a unique posterior does exist, it is given by (1.16). I will neglect issues of existence and uniqueness, and simply assume that a unique posterior exists.

The next consideration is computing the posterior. This can be done analytically for so called *conjugate* pairs of priors and likelihoods. Theses are pairs for which the prior and posterior both belong to the same family with altered parameters. Examples in the parametric case include beta priors and binomial likelihoods for estimating the binomial success probability, Gaussian priors and likelihoods for estimating the mean, and Gamma priors with Poisson likelihoods for estimating the rate. A nonparametric example is the Dirichlet process prior [Ferguson, 1973] with IID observations for estimating the unknown sampling distribution. Conjugate families impose very restrictive assumptions on the prior, and are often not available, most prominently whenever the likelihood is intractable. When conjugate priors are not available or appropriate, the posterior typically has to be approximated numerically, using Monte Carlo or other methods. This setting is the focus of this thesis.

In addition to reflecting prior information, the prior distribution Q can be seen as specifying a model for learning parameters from data. From this perspective, it makes sense to ask that $Q(\cdot|x_{1:n})$ concentrates on the "true", data generating parameter θ_0 as n increases, reflecting the potential to learn the true model from a sufficient amount of data. This property is known as *posterior consistency*, which can be stated more formally as

$$\lim_{n\to\infty} Q(\{\theta : \|\theta - \theta_0\| > \varepsilon\}|x_{1:n}) = 0$$

for any $\varepsilon > 0$ and some norm $\|\cdot\|$ on Θ, for some appropriate norm, topology and mode of convergence. In the nonparametric setting, posterior consistency is an intricate property which depends in subtle ways on Θ and Q. However, it is also regarded as a minimal requirement for well justified Bayesian inference [Diaconis and Freedman, 1986]. The typical properties required of the prior and the parameter space for posterior consistency to hold are a prior mass condition, i.e. Q must place sufficient mass in a neighbourhood of θ_0 (and, in particular, not exclude it), and regularity conditions to suitably limit the "size" of Θ.

Stronger, and more analytically demanding forms of posterior asymptotics consist of identification of contraction rates for consistent posteriors, and ultimately Bernstein-von Mises theorems:

$$\sup_{A \in \mathcal{B}(\Theta)} |Q(A|x_{1:n}) - \mu(\hat{\theta}, \Sigma)(A)| \to 0,$$

where μ is a Gaussian measure on Θ (see Section 6.3 of [Dashti and Stuart, 2016] for an overview of Gaussian measures on infinite dimensional spaces), $\hat{\theta}$ is an efficient estimator of the posterior mean and Σ is the posterior covariance. Contraction rates are typically established by constructing hypothesis tests with exponentially small error probability Schwartz [1965], for example in [Ghosal et al., 2000; Ghosal and van der Vaart, 2007; Gugushvili et al., 2015; Nickl and Söhl, 2015] in various nonparametric settings. The main drawback of this approach is the need to be able to construct exponentially consistent tests, which is rarely possible when the likelihood is intractable.

The Bernstein-von Mises theorem bridges the Bayesian and frequentist worlds by enabling the computation of asymptotic, frequentist estimators and confidence regions from the posterior. The earliest proof in a parametric setting was published by Doob [1949], and the modern form for parametric statistics was developed by Le Cam [1986]. Like contraction rates, nonparametric versions of Bernstein-von Mises theorems are an emerging and challenging area of research [Castillo and Nickl, 2013, 2014]. In this thesis I focus on establishing posterior consistency, and neglect more advanced notions of posterior contraction.

I will conclude this section by presenting an example nonparametric prior. Analytic formulae are rarely available in infinite dimensional spaces, and priors are typical specified by providing a sampling algorithm instead. Perhaps the most famous nonparametric prior is the Dirichlet process [Ferguson, 1973], the support of which consists of a.s. discrete probability measures on a general space Ω. Let ζ be a probability measure on Ω, and fix $\alpha > 0$. The following constructive definition

is due to Sethuraman [1994], and is called the *stick-breaking construction*:

- Sample $\{z_i\}_{i\in\mathbb{N}} \overset{\text{IID}}{\sim} \zeta$.
- Sample $\{\tilde{\beta}_i\}_{i\in\mathbb{N}} \overset{\text{IID}}{\sim} \text{Beta}(1,\alpha)$.
- For each $i \in \mathbb{N}$ set $\beta_i = \prod_{k=1}^{i-1}(1-\tilde{\beta}_k)\tilde{\beta}_i$ with the convention $\prod_{k=1}^{0} = 1$.
- A draw from the Dirichlet process is given by $\sum_{i=1}^{\infty} \beta_i \delta_{z_i}(\cdot)$.

The stick-breaking construction can be used to sample Dirichlet processes in practice using truncation with exponentially small error [Ishwaran and James, 2001], and an exact algorithm is available for sampling from measures generated by Dirichlet process priors [Papaspiliopoulos and Roberts, 2008]. A prior placing full mass on absolutely continuous densities can be obtained by using a Dirichlet process as a mixing measure for suitable kernels [Lo, 1984].

For a further overview of Bayesian nonparametric statistics, the interested reader is directed to [Hjort et al., 2010], and references therein.

Chapter 2

Sequential Monte Carlo in reverse time

2.1 Introduction

In this section I will specialise the SMC algorithm introduced in Section 1.3 for estimating functionals of Markov chains. Section 2.2 will then introduce the general reverse-time proposal distribution, of which the coalescent exposition in Section 1.3.1 turns out to be an example. The reverse-time SMC algorithm is most advantageous when the initial condition of the chain is typical and the terminal conditions lie in a set which is rare under the law of the chain. Hence reverse-time SMC can also be regarded as an example of rare event simulation. The interested reader is directed to e.g. Rubino and Tuffin [2009] and references therein for more details.

Consider the canonical Markov chain

$$\left(\Omega := \prod_{n=0}^{\infty} \Omega_n, \mathcal{F} := \bigotimes_{n=0}^{\infty} \mathcal{F}_n, \{X_n\}_{n=0}^{\infty}, \mathbb{P}_\mu\right) \tag{2.1}$$

where $\mathbb{P}_\mu$ is defined via its finite dimensional distributions as

$$\mathbb{P}_\mu(X_{0:n} \in dx_{0:n}) = \mu(dx_0) \prod_{i=0}^{n-1} P(x_i, dx_{i+1}). \tag{2.2}$$

Here $x_{0:n} := (x_0, \ldots, x_n)$, and P is a given transition kernel. I assume both P and μ can be evaluated pointwise, but not that (2.1) is stationary or even has a stationary distribution.

Two space-time sets are needed to specify the rare event problem. Let $I \subset$

$\mathbb{N} \times \Omega$ be an entrance set, and without loss of generality suppose $\mu(I) = 1$. Let $T \subset \mathbb{N} \times \Omega$ be the target set, which is assumed to have finite expected hitting time under the dynamics of the chain $\{n, X_n\}_{n\in\mathbb{N}}$. For a set $A \in \mathcal{F}$, let $\tau_A := \inf\{n \geq 0 : X_n \in A\}$ denote the hitting time of $\{X_n\}_{n=0}^{\infty}$ with initial distribution μ. The problem in question is estimating expectations of trajectories of the form

$$\mathbb{E}_\mu[f(\tau_T, X_{0:T})|\tau_T < \tau_I]$$

for integrable functions f. I emphasize that these trajectories are defined between the *last exit time* of I and the *hitting time* of T. In particular, re-entry into I before hitting T is not permitted. As an example, let T depend only on space, and consider the hitting probability (resp. density) of a point $x \in T$ whenever Ω is discrete (resp. continuous), before any other point of T. The corresponding functional is

$$\mathbb{E}_\mu[f(\tau_T, X_{0:T})|\tau_T < \tau_I] = \mathbb{E}_\mu[\mathbb{1}_{\{x\}}(X_{\tau_T})|\tau_T < \tau_I].$$

Similar rare events in the case of in the case of homogeneous Markov chains and recurrent initial sets were termed dynamic rare events by Johansen et al. [2006].

In the following Section I show how the reverse-time approach can be used to design proposal distributions based on the time-reversal of the process of interest. The distribution of the time-reversal can be expressed via the Green's function (c.f. Nagasawa's formula (2.4)). The Green's function is typically at least as difficult to compute as the quantity of interest, but progress can be made by introducing an ad-hoc approximation, and substituting it into Nagasawa's formula. The better the approximation, the more efficient the algorithm. Moreover, because the Green's functions appear in (2.4) only as ratios, conditioning arguments can often be used to cancel these ratios to a low-dimensional quantity that can be approximated directly (c.f. Proposition 2 in Section 2.2). This avoids the need to design high-dimensional proposal distributions even when the state space is itself high-dimensional. The method can be expected to be particularly efficient in contexts where

(i) the function f depends only on the terminal point in T, or a small length of trajectory preceding it,

(ii) the dimension and/or volume of the regions in T which contribute non-negligibly to $\mathbb{E}_\mu[f(\tau_T, f(X_{0,\tau_T})]$ are small, while I is high dimensional and/or has large volume,

(iii) the majority of the contribution to $\mathbb{E}_\mu[f(\tau_T, X_{0:\tau_T})]$ arises from a region of low conditional probability given that the chain has hit T, and

(iv) the process of interest is high-dimensional and transitions only alter a small number of components at a time.

Properties (i) and (ii) ensure that time-reversal is an effective strategy. In the extreme case where f is the indicator function of a singleton in T (corresponding to estimation of a conditional hitting density), I is a set which is hit by the reverse-time dynamics in finite time with probability 1, and T is a reverse-time entrance boundary, the optimal proposal distribution leading to zero variance estimators is the unconditional time-reversal. These conditions are very restrictive, but typically satisfied by coalescent processes. On the other hand, all of the examples in Section 2.6 violate at least one of them, which demonstrates that reverse-time proposals can still yield efficient algorithms under the milder conditions (i)-(iv).

Property (iii) is helpful in ensuring that T acts like a reverse-time entrance boundary with high probability, as proposal trajectories will naturally be pushed away from the rare hitting point of T and back towards I. Property (iv) means that it is only necessary to come up with a proposal distribution for the coordinates which differ between transitions, given the value of all other coordinates. This dimensionality reduction can greatly reduce the difficulty of designing proposal distributions in high dimension. Proposition 2 in Section 2.2 provides a precise formulation, and Sections 2.6.2 and 2.6.3 contain concrete examples.

The choice of resampling schedule also has a strong impact on the efficiency of the SMC algorithm. Few theoretical guidelines are available, though developments have been made in determining good schedules adaptively [Cérou et al., 2012; Jasra et al., 2014]. I will neglect the problem of optimising the resampling schedule, except to mention that implementing the method of Cérou et al. [2012] requires a reaction coordinate, i.e. a tractable function

$$g : \Omega \mapsto \mathbb{R} \tag{2.3}$$

which captures closeness of a particle to the target set, e.g. via mapping positions closer to the target to higher values in a monotonic way. Reaction coordinates can be difficult to derive in high dimension, and reverse time SMC does not require one to be implemented, but when one is available then the results of Cérou et al. [2012] apply. In contrast, the particle MCMC approach of Jasra et al. [2014] is directly applicable.

2.2 Time-reversal as a SMC proposal distribution

This section reviews some relevant facts about time-reversal of Markov chains and introduces the reverse-time SMC proposal distribution. Concrete examples are given in Sections 2.3, 2.5 and 2.6.

Define the time-reversal of (2.1) by extending the chain to the negative time-axis, and letting

$$\left(\prod_{n=0}^{-\infty}\Omega_n, \bigotimes_{n=0}^{-\infty}\mathcal{F}_n, \left\{\widetilde{X}_n\right\}_{n=0}^{-\infty}, \widetilde{\mathbb{P}}_\nu\right)$$

denote the reverse-time chain. Note that the initial time-indices are set to 0 by convention, and are not necessarily intended to coincide with the starting time of (2.1). The law $\widetilde{\mathbb{P}}_\nu$ is again defined via its finite dimensional distributions as

$$\widetilde{\mathbb{P}}_\nu(dx_{0:-n}) = \nu(dx_0)\prod_{i=0}^{-n+1}\widetilde{P}(x_i, dx_{i-1}).$$

Here, ν is the initial distribution of the reverse time chain, that is, the law of the random variable $X_{\tau_T}|\tau_T < \tau_I$. For simplicity, all the transition kernels above, and Green's functions below, are assumed absolutely continuous with respect to the same reference measure (e.g. Lebesgue or counting measure), and the same notation is used for both the kernels/Green's functions and their densities.

The reverse transition kernel is related to its forward counterparts via Nagasawa's formula (c.f. Chapter III.46 of [Rogers and Williams, 1994]):

$$\widetilde{P}(x_i, x_{i-1}) = \frac{G(\mu, x_{i-1})}{G(\mu, x_i)}P(x_{i-1}, x_i), \tag{2.4}$$

where for a measurable set A

$$G(\mu, A) := \mathbb{E}_\mu\left[\sum_{i=0}^{\tau_T}\mathbb{1}_A(X_i)\right] =: \int_I\int_A g(x,y)dy\mu(dx)$$

is the Green's function of (2.1), and $\mathbb{E}_\mu$ denotes expectation with respect to $\mathbb{P}_\mu$. When $A = \{z\}$ is a null set (with respect to the reference measure), $G(\mu, z)$ is defined as a density via the kernel $g(x, z)$:

$$G(\mu, z) := \int_I g(x,z)\mu(dx),$$

which is assumed to exist. Nagasawa's formula can be seen as a generalisation of the detailed balance condition to non-stationary chains; when $\mathbf{X}$ admits a unique

stationary distribution π, the stopping time τ_T is deterministic and the chain is at stationarity, then

$$\frac{G(\mu, y)}{G(\mu, x)} = \frac{\pi(y)}{\pi(x)},$$

and (2.4) is the detailed balance condition. Reverse time proposal distributions akin to (2.4) have been studied previously in [Birkner et al., 2011] for Λ-coalescents under infinite sites mutation.

The Green's functions in (2.4) cannot be computed in most cases of interest, but their qualitative behaviour can often be described. I assume that such a description is available, and that it is possible to write down a family of tractable functions with similar qualitative behaviour. These will be referred to as approximate Green's functions. It is not necessary for the match to be very precise, because importance weights will correct for the mismatch, though better approximations yield more efficient SMC algorithms.

The strategy for defining a reverse-time SMC proposal is as follows:

1. Design an approximate Green's function $\widehat{G}(\mu, x)$ to be substituted into (2.4) to yield an approximate reverse-time transition kernel $\widehat{P}$ and a proposal Markov chain

$$\left(\prod_{n=0}^{-\infty} \Omega_n, \prod_{n=0}^{-\infty} \mathcal{F}_n, \left\{ \widehat{X}_n \right\}_{n=0}^{-\infty}, \widehat{\mathbb{P}}_\nu \right) \tag{2.5}$$

   where $\widehat{\mathbb{P}}_\nu$ is defined from its finite dimensional distributions via $\widehat{P}$ as before.

2. If necessary, modify $\widehat{\mathbb{P}}_\nu$ locally to incorporate first hitting time constraints by preventing (2.5) from returning to T upon leaving it and from entering R at all.

3. If necessary, introduce further local modifications to the proposal distribution to ensure (2.5) hits I in finite time with $\widehat{\mathbb{P}}_\nu$-probability 1 so that the reverse-time chain terminates in finite time with certainty.

These steps can seem laborious because of their generality, but as will be seen in Sections 2.3, 2.5 and 2.6, they can be feasibly carried out in many cases of interest. Steps 2 and 3 could be incorporated automatically and more efficiently by considering the time-reversal of an appropriately h-transformed version of (2.1). However, the h-transform is typically intractable, whereas local modifications are widely implementable and still result in efficient algorithms when the dominant contribution to $\mathbb{E}_\mu[f(\tau_T, X_{0:\tau_T})]$ arises from a region of low $\mathbb{P}_\mu$-probability. This is because the ratio of Green's functions will drive sample paths away from T and towards I even without conditioning on no re-entry.

Proposition 2 presents a practical way of designing approximate ratios of Green's functions for a wide class of models. For notational simplicity I assume a countable state space, but the same argument holds for continuous state spaces provided the Green's densities $g(x, z)$ exist.

Proposition 2. *Consider a transition of the Markov chain* (2.1) *from* x_{n-1} *to* x_n*, and suppose the state space can be partitioned so that* $x_{n-1} = (z, y)$ *and* $x_n = (z, \bar{y})$*. Note that the decomposition may depend on the exact pair* (x_{n-1}, x_n)*, and need to coincide across pairs. Assume that the conditional sampling distribution (CSD)*

$$\pi(y|z) := \mathbb{P}_\mu(Y_n = y | Z_n = z)$$

is independent of $n \in \mathbb{N}$ *for* $\mathbb{P}_\mu$*-almost every* z*. Then the ratio of Green's functions in* (2.4) *cancels to the ratio of CSDs:*

$$\frac{G(\mu, (z, y))}{G(\mu, (z, \bar{y}))} = \frac{\pi(y|z)}{\pi(\bar{y}|z)}.$$

Proof. Let ∂ be a cemetery state, and define the process

$$X_t^\partial = \begin{cases} X_t & \text{if } t \leq \tau_T \\ \partial & \text{otherwise} \end{cases}.$$

Note that the laws of $\{X_n^\partial\}_{n=0}^{\tau_T}$ and $\{X_n\}_{n=0}^{\tau_T}$ coincide, and thus so do their Green's functions evaluated at states $(z, y) \in (T \cup \partial)^c$. Hence, for any such state, Fubini's theorem and conditioning on $Z_n = z$ yield

$$\begin{aligned} G(\mu, (z, y)) &= \mathbb{E}_\mu \left[\sum_{n=0}^{\infty} \mathbb{1}_{\{z,y\}}(Z_n^\partial, Y_n^\partial) \right] = \sum_{n=0}^{\infty} \mathbb{E}_\mu \left[\mathbb{1}_{\{z,y\}}(Z_n^\partial, Y_n^\partial) \right] \\ &= \sum_{n=0}^{\infty} \mathbb{E}_\mu \left[\mathbb{E}_\mu \left[\mathbb{1}_{\{y\}}(Y_n^\partial) | Z_n^\partial = z \right] \mathbb{1}_{\{z\}}(Z_i^\partial) \right], \end{aligned}$$

where $(Z_n^\partial, Y_n^\partial) := X_n^\partial$. Now $\pi(y|z) = \mathbb{E}_\mu \left[\mathbb{1}_{\{y\}}(Y_n^\partial) | Z_n^\partial = z \right]$ is independent of n by assumption. Thus

$$G(\mu, (z, y)) = \pi(y|z) \sum_{n=0}^{\infty} \mathbb{E}_\mu \left[\mathbb{1}_{\{z\}}(Z_n) \mathbb{1}_{n:\infty}(\tau_T) \right] = \pi(y|z) \sum_{n=0}^{\infty} \sum_{t=0}^{\infty} \mathbb{P}_\mu(Z_n = z, \tau_T = t).$$

Now note that the final double sum cancels whenever the Green's functions are evaluated as ratios, which completes the proof. □

Remark 1. The hypothesis of Proposition 2 is a weak stationarity condition, and is relatively mild. However, since ad hoc approximations to the true ratio of Green's functions is all that is required, it is possible to extend the scope of the reverse time framework by defining the proposal distribution based on a family of *approximate* CSDs $\widehat{\pi}(y|z)$ even when Proposition 2 fails. This is because the importance weights will still correct for any bias, resulting in a valid, unbiased algorithm. Because of their lower dimension, approximate CSDs can be much easier to design than either proposal kernels $\{Q(\cdot,\cdot)\}$ or approximate Greens functions $\{\widehat{G}(\mu,\cdot)\}$. Indeed, this dimensionality reduction in the design task is one of the main advantages of the reverse-time framework.

Choosing a family of approximate CSDs $\{\widehat{\pi}(\cdot|\cdot)\}$ and applying Proposition 2 to (2.4) yields proposal transition probabilities of the form

$$\widehat{P}(x,y) = \frac{\widehat{\pi}(y \setminus x | y \cap x)}{C(x)\widehat{\pi}(x \setminus y | x \cap y)} P(y,x),$$

where $C(x)$ is a normalising constant, $x \cap y$ is the vector of coordinates for which $x_i = y_i$, and $x \setminus y$ is the vector of coordinates of x for which $x_i \neq y_i$. The corresponding incremental importance weight at step n is

$$\frac{C(x)\widehat{\pi}(x \setminus y | x \cap y)}{\widehat{\pi}(y \setminus x | x \cap y)}.$$

Once a proposal chain has been constructed, functionals of interest can be unbiasedly estimated as

$$\begin{aligned}\mathbb{E}_\mu[f(X_{0:\tau_T})] &\approx \frac{1}{N}\sum_{j=1}^{N} f(x^{(j)}_{0:\tau_T^{(j)}}) \frac{d\mathbb{P}_\mu}{d\widehat{\mathbb{P}}_\nu}(x^{(j)}_{0:\tau_T^{(j)}}) \\ &= \frac{1}{N}\sum_{j=1}^{N} f(x_{0:\tau_T^{(j)}}) \frac{\mu(x_0^{(j)})}{\nu(x_{\tau_T})} \prod_{n=1}^{\tau_T} \frac{\widehat{\pi}(x_{n-1}^{(j)} \setminus x_n^{(j)} | x_n^{(j)} \cap x_{n-1}^{(j)})}{\widehat{\pi}(x_n^{(j)} \setminus x_{n-1}^{(j)} | x_n^{(j)} \cap x_{n-1}^{(j)})} C(x_n^{(j)}), \end{aligned} \tag{2.6}$$

with the convention $0/0 = 0$ for trajectories that are incompatible with the forward dynamics of $\mathbf{X}$, and where $\{x^{(j)}_{0:\tau_T^{(j)}}\}_{j=1}^N$ is a sample from the SMC algorithm that uses (2.5) as its proposal mechanism.

Remark 2. Approximating (2.6) can be a computationally daunting task if f takes non-negligible values across a set of trajectories with end points in a high-dimensional or large (in terms of Lebesgue-volume) subset of T. In such cases the reverse-time approach cannot always be expected to be competitive with forwards-

in-time algorithms, particularly if the initial set I is also small and hence difficult for the reverse-time chain to hit.

Remark 3. Normalising constants $C(x)$ in (2.6) would all be identically equal to one if an algorithm using the true CSD could be implemented. Thus, the realised values of these constants for a given approximate CSD could be used to design proposal distributions adaptively from trial runs, at least for discrete systems where the constants can easily be computed. However, developing this strategy is beyond the scope of this thesis.

I conclude this section with a formal specification of the reverse-time SMC algorithm (Algorithm 1).

2.3 SMC for Λ- and Ξ-coalescents

Investigations by Boom et al. [1994], Árnason [2004], Eldon and Wakeley [2006], and Birkner and Blath [2008] 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 to those outlined in Section 1.3.1 have been developed for them. An analogue of the Griffiths-Tavaré recursion for Λ-coalescents was derived by Birkner and Blath [2008]. In a subsequent paper Birkner et al. [2011] characterised the optimal SMC 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., 2013a] contains a detailed discussion of inference under Beta-coalescents and their applicability to marine populations.

Ξ-coalescents have also been used to model genealogies of marine organisms [Sargsyan and Wakeley, 2008] and populations undergoing mass extinctions [Taylor and Véber, 2009], although the question of which measures Ξ are biologically relevant remains open.

The coalescent inference problem can be cast into the rare event simulation framework by regarding the coalescent as a branching process starting from a single lineage, and branching outwards to grow a random labelled tree. The likelihood is then given as the probability that this labelled tree hits a configuration that is compatible with the observed haplotype frequencies before the tree grows larger than the observed sample. The dynamics of the branching tree processes can be obtained from lookdown constructions, which embed both the coalescent and the

Algorithm 1 Reverse-time multilevel SMC algorithm for $\mathbb{E}_\mu[f(\tau_T, X_{0:\tau_T})]$.

Require: Particle number N, approximate conditional distributions $\{\widehat{\pi}(\cdot|\cdot)\}$, stopping times $\{\{\tau_i^j\}_{i=n}^1\}_{j=1}^N$ such that $0 = \tau_T^j \leq \tau_n^j \leq \tau_{n-1}^j \leq \ldots \leq \tau_1^j = \tau_I^j$.

for j = 1 to N **do**
  Sample $X_0^j \sim \nu$. ▷ Initial particle locations.
  Set $w_j = 1/\nu(X_0^j)$. ▷ Initial importance weights.
  Set $k_j = 0$, $\bar{k}_j = 0$. ▷ Time indices.
  Set $A_j = j$. ▷ Ancestral indices are used for resampling.
end for
for i = n to 1 **do**
  if $\mathrm{ESS}(w_1, \ldots, w_N) < \frac{N}{2}$ **then** ▷ Resample if the ESS is too low.
    for j = 1 to N **do**
      Sample $A_j \sim \sum_{k=1}^N w_k \delta_k$. ▷ Sample ancestral particle $\propto$ weights.
      Set $\bar{k}_j = k_{A_j}$.
      Set $w_j = 1$. ▷ Equalise weights.
    end for
  end if
  for j = 1 to N **do**
    if $\bar{k}_j < \tau_i^j$ **then** ▷ First step if next level not yet hit.
      Set $k_j = \bar{k}_j + 1$.
      Sample $X_{k_j}^j \sim \dfrac{\widehat{\pi}\left(\cdot \backslash X_{\bar{k}_j}^{A_j} \middle| \cdot \cap X_{\bar{k}_j}^{A_j}\right)}{\widehat{\pi}\left(X_{\bar{k}_j}^{A_j} \backslash \cdot \middle| \cdot \cap X_{\bar{k}_j}^{A_j}\right)} \dfrac{P\left(\cdot, X_{\bar{k}_j}^{A_j}\right)}{C\left(X_{\bar{k}_j}^{A_j}\right)} \mathbb{1}_{T^c \cap R^c}(\cdot)$.
      Set $w_j = \dfrac{\widehat{\pi}\left(X_{\bar{k}_j}^{A_j} \backslash X_{k_j}^j \middle| X_{k_j}^j \cap X_{\bar{k}_j}^{A_j}\right) C\left(X_{\bar{k}_j}^{A_j}\right)}{\widehat{\pi}\left(X_{k_j}^j \backslash X_{\bar{k}_j}^{A_j} \middle| X_{k_j}^j \cap X_{\bar{k}_j}^{A_j}\right)}$.
      Set $A_j = j$.
    end if
    while $k_j < \tau_i^j$ **do** ▷ Propagate until the next level is hit.
      Set $k_j = k_j + 1$.
      Sample $X_{k_j}^j \sim \dfrac{\widehat{\pi}\left(\cdot \backslash X_{k_j-1}^j \middle| \cdot \cap X_{k_j-1}^j\right)}{\widehat{\pi}\left(X_{k_j-1}^j \backslash \cdot \middle| \cdot \cap X_{k_j-1}^j\right)} \dfrac{P\left(\cdot, X_{k_j-1}^j\right)}{C\left(X_{k_j-1}^j\right)} \mathbb{1}_{T^c \cap R^c}(\cdot)$.
      Set $w_j = w_j \dfrac{\widehat{\pi}\left(X_{k_j-1}^j \backslash X_{k_j}^j \middle| X_{k_j}^j \cap X_{k_j-1}^j\right) C\left(X_{k_j-1}^j\right)}{\widehat{\pi}\left(X_{k_j}^j \backslash X_{k_j}^j \middle| X_{k_j}^j \cap X_{k_j-1}^j\right)}$.
    end while
  end for
end for
for j = 1 to N **do**
  Set $w_j = w_j \mu\left(X_{k_j}^j\right)$. ▷ Account for entrance law μ.
end for
return $\sum_{j=1}^N \dfrac{w_j f(\tau_T^j, X_{0:\tau_T^j}^j)}{\sum_{j=1}^N w_j}$. ▷ Unbiased estimator of $\mathbb{P}_\mu(\tau_T < \tau_R)$.

dual Fleming-Viot process into the same countable particle system [Donnelly and Kurtz, 1996, 1999; Birkner et al., 2009].

Let $\{\widetilde{\Pi}_k\}_{k\in\mathbb{N}}$ denote the jump skeleton of the coalescent, and define $\tau_n := \inf_{k\geq 0}\{|\widetilde{\Pi}_k| \geq n\}$. From the hitting probability perspective, the coalescent inference problem is characterised by the sets

$$\begin{aligned} I &= \{\mathbf{e}_h : h \in \mathcal{H}\}, && \text{the MRCA, and} \\ T &= \{\mathbf{m} \in \mathbb{N}^{|\mathcal{H}|} : |\mathbf{m}| \geq n+1\}, && \text{exceeding } n \text{ leaves,} \end{aligned}$$

and the test function

$$\mathbb{E}_\mu[f(\tau_T, \tilde{\Pi}_{0:\tau_T})] = \mathbb{E}_\mu[\mathbb{1}_{\{\tilde{\Pi}_{\tau_{n+1}-1}=\mathbf{n}\}}],$$

i.e. the probability that the last configuration of leaves before growing beyond size n matches the observed data.

I will begin this section with the simpler Λ-coalescents, and cover SMC for Ξ-coalescents in Section 2.3.3 after a simulation study for Λ-coalescents in Section 2.3.2.

Recall the decomposition of the coalescent likelihood (1.15) and the notation introduced in Section 1.1.3, and note that for the Λ-coalescent the conditional transition probabilities can be written as

$$\mathbf{P}_n^\Lambda(A_{i+1}|A_i) = \begin{cases} \frac{\theta_l}{n\theta+q_{n_i n_i}}(n_i^{(S_l^a(h))} + 1 - \delta_{ah[l]})M_{ah[l]}^{(l)} & \text{if } A_{i+1} = A_i - \mathbf{e}_{S_l^a(h)} + \mathbf{e}_h \\ \binom{n_i}{k}\frac{\lambda_{n_i,k}}{n\theta+q_{n_i n_i}}\frac{n_i^{(h)}-k+1}{n_i-k+1} & \text{if } A_{i+1} = A_i + (k-1)\mathbf{e}_h \end{cases} \tag{2.7}$$

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 A_i and $n_i^{(h)}$ is the number of lineages of type h in A_i. See [Birkner and Blath, 2008] for a detailed derivation. Here, and in Theorem 1 below, the MRCA corresponds to time index 0, and higher indices denote the number of mutations or coalescence events in the tree, counting up from the MRCA towards the leaves.

The following theorem is a Λ-coalescent analogue of Theorem 1 of Stephens and Donnelly [2000]. Lemma 2.2 of [Birkner et al., 2011] presents a similar result using ratios of Greens functions instead of CSDs.

Theorem 1. *Let $\pi(\mathbf{m}|\mathbf{n})$ denote 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* $\widetilde{\mathbf{P}}_n$ *are given by*

$$\widetilde{\mathbf{P}}_n(A_{i-1}|A_i) \propto \begin{cases} n_{i-1}^{(h)}\theta_l \frac{\pi(\mathbf{e}_{S_l^a(h)}|A_i-\mathbf{e}_h)}{\pi(\mathbf{e}_h|A_i-\mathbf{e}_h)} M_{ah[l]}^{(l)} & \text{if } A_{i-1} = A_i - \mathbf{e}_h + \mathbf{e}_{S_l^a(h)} \\ \binom{n_{i-1}^{(h)}}{k} \frac{\lambda_{n_{i+1},k}}{\pi((k-1)\mathbf{e}_h|A_i-(k-1)\mathbf{e}_h)} & \text{if } A_{i-1} = A_i - (k-1)\mathbf{e}_h \end{cases}$$

where the first term ranges over all possible mutations for all haplotypes present in the sample, and the second over all present haplotypes and $k \in \{2, \dots n_i^{(h)}\}$.

Proof. The argument giving the mutation term is identical to that in Theorem 1 of Stephens and Donnelly [2000] and is omitted.

For the coalescence term suppose the n lineages evolve according to the lookdown construction of Donnelly and Kurtz [1999], and denote the types of the n particles at time t by $D_n(t) = (h_1, \dots, h_n)$. Define Υ_k as the event that in the last δ units of time there was a merger involving lineages $n-k+1, n-k+2, \dots, n$. To simplify the presentation let $h_{i:j} := (h_i, h_{i+1}, \dots, h_{j-1}, h_j)$. Then

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

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 4. It is tempting to simplify the situation further by decomposing

$$\pi((k-1)\mathbf{e}_h | \mathbf{n} - (k-1)\mathbf{e}_h) = \prod_{j=0}^{k-2} \pi(\mathbf{e}_h | \mathbf{n} - (k-1+j)\mathbf{e}_h)$$

and thus requiring only univariate CSDs. In general a decomposition like this requires exchangeability, which the CSDs satisfy but approximations typically do not. However, in the Λ-coalescent setting the argument being decomposed will always consist of only one type of haplotype. Permuting lineages which feature only in the sample being conditioned on does not affect the outcome even for non-exchangeable 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.

2.3.1 Approximate CDSs for Λ-coalescents

An approximation to the CSDs for Kingman's coalescent was derived in De Iorio and Griffiths [2004a] by noting that the Fleming-Viot generator can be written component-wise as $\mathcal{G}^{\delta_0} = \sum_{h\in\mathcal{H}} \mathcal{G}_h^{\delta_0}$, then assuming that there exists a probability measure and an expectation $\widehat{\mathbb{E}}^{\delta_0}$ with respect to that measure, such that the standard stationarity condition $\mathbb{E}^{\delta_0}[\mathcal{G}^{\delta_0} f(\mathbf{X})] = 0$ holds component-wise:

$$\widehat{\mathbb{E}}^{\delta_0}[\mathcal{G}_h^{\delta_0} f(\mathbf{X})] = 0 \text{ for every } h \in \mathcal{H} \text{ and } f \in C^2(\Delta_{\mathcal{H}}).$$

Substituting the probability of an ordered sample $q(\mathbf{n}|\mathbf{x}) = \prod_{h\in\mathcal{H}} x_h^{n_h}$ 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 2. *Let $\widehat{\pi}^{\Lambda}(\mathbf{m}|\mathbf{n})$ denote the approximate Λ-coalescent CSD as defined above. It solves the recursion*

$$\begin{aligned}
&m\left[\frac{\Lambda(\{0\})(n+m-1)}{2} + \theta + \frac{1}{n+m}\sum_{k=2}^{n+m}\binom{n+m}{k}\lambda_{n+m,k}\right]\widehat{\pi}^{\Lambda}(\mathbf{m}|\mathbf{n}) \\
&= \sum_{h\in\mathcal{H}} m_h\left[\frac{\Lambda(\{0\})(n_h+m_h-1)}{2}\widehat{\pi}^{\Lambda}(\mathbf{m}-\mathbf{e}_h|\mathbf{n})\right. \\
&\quad + \sum_{l\in[k]}\theta_l \sum_{a\in E_l} M^{(l)}_{ah[l]}\widehat{\pi}^{\Lambda}(\mathbf{m}-\mathbf{e}_h+\mathbf{e}_{S_l^a(h)}|\mathbf{n}) \\
&\quad + \frac{1}{n_h+m_h}\left\{\sum_{k=2}^{m_h+1}\binom{n_h+m_h}{k}\lambda_{n+m,k}\widehat{\pi}^{\Lambda}(\mathbf{m}-(k-1)\mathbf{e}_h|\mathbf{n})\right. \\
&\quad \left.\left. + \sum_{k=m_h+2}^{n_h+m_h}\binom{n_h+m_h}{k}\lambda_{n+m,k}\frac{\widehat{\pi}^{\Lambda}(\mathbf{m}-m_h\mathbf{e}_h|\mathbf{n}-(k-m_h-1)\mathbf{e}_h)}{\widehat{\pi}^{\Lambda}((k-m_h-1)\mathbf{e}_h|\mathbf{n}-(k-m_h-1)\mathbf{e}_h)}\right\}\right]. \quad (2.8)
\end{aligned}$$

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

$$\begin{aligned}\mathcal{G}^{\Lambda} f(\mathbf{x}) = &\sum_{h\in\mathcal{H}} \frac{\Lambda(\{0\})x_h}{2} \sum_{h'\in\mathcal{H}} (\delta_{hh'} - x_{h'}) \frac{\partial^2}{\partial x_h \partial x_{h'}} f(\mathbf{x}) \\ &+ \sum_{h\in\mathcal{H}} \sum_{l\in[k]} \theta_l \sum_{a\in E_l} x_{S_l^a(h)} \left(M^{(l)}_{ah[l]} - \delta_{ah[l]} \right) \frac{\partial}{\partial x_h} f(\mathbf{x}) \\ &+ \sum_{h\in\mathcal{H}} x_h \int_{(0,1]} \{ f((1-r)\mathbf{x} + r\mathbf{e}_h) - f(\mathbf{x}) \} \, r^{-2} \Lambda(dr) =: \sum_{h\in\mathcal{H}} \mathcal{G}_h^{\Lambda} f(\mathbf{x}).\end{aligned} \tag{2.9}$$

Substituting $q(\mathbf{n}|\mathbf{x})$ yields the following three terms on the RHS

$$\sum_{h\in\mathcal{H}} \frac{\Lambda(\{0\})n_h}{2} \left[(n_h - 1) q(\mathbf{n} - \mathbf{e}_h|\mathbf{x}) - \sum_{h'\in\mathcal{H}} (n_{h'} - \delta_{hh'}) q(\mathbf{n}|\mathbf{x}) \right] \tag{2.10}$$

$$+ \sum_{h\in\mathcal{H}} n_h \sum_{l\in[k]} \theta_l \left[\sum_{a\in E_l} M^{(l)}_{ah[l]} q(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{S_l^a(h)}|\mathbf{x}) - q(\mathbf{n}|\mathbf{x}) \right] \tag{2.11}$$

$$\begin{aligned}+ \int_{(0,1]} \Bigg\{ &\sum_{h\in\mathcal{H}} \sum_{p=0}^{n_h} \binom{n_h}{p} r^p (1-r)^{n-p} q(\mathbf{n} - (p-1)\mathbf{e}_h|\mathbf{x}) \\ &- \sum_{p=0}^{n} \binom{n}{p} r^p (1-r)^{n-p} q(\mathbf{n}|\mathbf{x}) \Bigg\} r^{-2} \Lambda(dr).\end{aligned}$$

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

$$\sum_{h\in\mathcal{H}} \left\{ \sum_{p=2}^{n_h} \binom{n_h}{p} \lambda_{n,p} q(\mathbf{n} - (p-1)\mathbf{e}_h|\mathbf{x}) - \frac{n_h}{n} \sum_{p=2}^{n} \binom{n}{p} \lambda_{n,p} q(\mathbf{n}|\mathbf{x}) \right\}. \tag{2.12}$$

Substituting (2.10), (2.11) and (2.12) into (2.9) and rearranging gives

$$\begin{aligned}&\sum_{h\in\mathcal{H}} n_h \left[\frac{\Lambda(\{0\})(n-1)}{2} + \theta + \frac{1}{n} \sum_{k=2}^{n} \binom{n}{k} \lambda_{n,k} \right] q(\mathbf{n}|\mathbf{x}) \\ &= \sum_{h\in\mathcal{H}} n_h \Bigg\{ \frac{\Lambda(\{0\})(n_h - 1)}{2} q(\mathbf{n} - \mathbf{e}_h|\mathbf{x}) + \sum_{l\in[k]} \theta_l \sum_{a\in E_l} M^{(l)}_{ah[l]} q(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{S_l^a(h)}|\mathbf{x}) \\ &\qquad + \frac{1}{n_h} \sum_{p=2}^{n_h} \binom{n_h}{p} \lambda_{n,p} q(\mathbf{n} - (p-1)\mathbf{e}_h|\mathbf{x}) \Bigg\}.\end{aligned} \tag{2.13}$$

The component-wise vanishing property implies

$$\widehat{\mathbb{E}}^{\Lambda}\left[\sum_{h\in\mathcal{H}} m_h \mathcal{G}_h^{\Lambda} q(\mathbf{n}|\mathbf{X}^{\Lambda})\right] = \sum_{h\in\mathcal{H}} m_h \widehat{\mathbb{E}}^{\Lambda}[\mathcal{G}_h^{\Lambda} q(\mathbf{n}|\mathbf{X}^{\Lambda})] = 0,$$

so that (2.13) becomes

$$\begin{aligned}
&m\left[\frac{\Lambda(\{0\})(n-1)}{2} + \theta + \frac{1}{n}\sum_{k=2}^{n}\binom{n}{k}\lambda_{n,k}\right]\widehat{\mathbb{E}}^{\Lambda}[q(\mathbf{n}|\mathbf{X}^{\Lambda})] \\
&\quad = \sum_{h\in\mathcal{H}} m_h \Bigg\{ \frac{\Lambda(\{0\})(n_h-1)}{2}\widehat{\mathbb{E}}^{\Lambda}[q(\mathbf{n}-\mathbf{e}_h|\mathbf{X}^{\Lambda})] \\
&\qquad + \sum_{l\in[k]}\theta_l \sum_{a\in E_l} M^{(l)}_{ah[l]}\widehat{\mathbb{E}}^{\Lambda}[q(\mathbf{n}-\mathbf{e}_h+\mathbf{e}_{S_l^a(h)}|\mathbf{X}^{\Lambda})] \\
&\qquad + \frac{1}{n_h}\sum_{p=2}^{n_h}\binom{n_h}{p}\lambda_{n,p}\widehat{\mathbb{E}}^{\Lambda}[q(\mathbf{n}-(p-1)\mathbf{e}_h|\mathbf{X}^{\Lambda})]\Bigg\}.
\end{aligned}$$

Note that $\pi^{\Lambda}(\mathbf{m}|\mathbf{n}) = \mathbb{E}^{\Lambda}[q(\mathbf{n}+\mathbf{m}|\mathbf{X}^{\Lambda})]/\mathbb{E}^{\Lambda}[q(\mathbf{n}|\mathbf{X}^{\Lambda})]$ so that substituting $\mathbf{n} \mapsto \mathbf{n}+\mathbf{m}$, assuming that $\mathbb{E}^{\Lambda} = \widehat{\mathbb{E}}^{\Lambda}$ and dividing by $\mathbb{E}^{\Lambda}[q(\mathbf{n}|\mathbf{X}^{\Lambda})]$ gives the desired recursion. □

Corollary 1. *The univariate approximate CSDs* $\hat{\pi}(\mathbf{e}_h|\mathbf{n})$ *satisfy*

$$\begin{aligned}
&\left[\frac{\Lambda(\{0\})n}{2} + \theta + \frac{1}{n+1}\sum_{k=2}^{n+1}\binom{n+1}{k}\lambda_{n+1,k}\right]\widehat{\pi}^{\Lambda}(\mathbf{e}_h|\mathbf{n}) \\
&= \frac{n_h}{2}\left(\Lambda(\{0\}) + \lambda_{n+1,2}\right) + \sum_{l\in[k]}\theta_l\sum_{a\in E_l} M^{(l)}_{ah[l]}\widehat{\pi}^{\Lambda}(\mathbf{e}_{S_l^a(h)}|\mathbf{n}) \\
&\quad + \frac{1}{n_h+1}\sum_{p=3}^{n_h+1}\binom{n_h+1}{p}\frac{\lambda_{n+1,p}}{\widehat{\pi}^{\Lambda}((p-2)\mathbf{e}_h|\mathbf{n}-(p-2)\mathbf{e}_h)}.
\end{aligned} \tag{2.14}$$

Proof. The result follows by substituting $\mathbf{m} = \mathbf{e}_h$ into (2.8). □

As per Remark 4 it is sufficient to work with the simpler recursion (2.14) as opposed to the full recursion (2.8). However, because of the denominator in the final term of (2.14), 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 1. Setting $\Lambda = \delta_0$ in (2.14) results in the approximate CSDs derived in Stephens and Donnelly [2000] 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. I denote this proposal distribution by $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{SD}}$.

Paul and Song [2010] introduced the *trunk ancestry*, which can be used to obtain an approximation which makes better use of the Λ-coalescent structure. I will briefly recall the definition of the trunk ancestry here before using it to define a second approximate CSD family.

Definition 2. The trunk ancestry $A^*(\mathbf{n})$ of a sample $\mathbf{n}$ is a deterministic, degenerate process started from $\mathbf{n}$ and evolving backwards in time but undergoing no dynamics.

In the trunk ancestry, the lineages that form $\mathbf{n}$ 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 RHS of (2.14), corresponding to pairwise mergers and mutations, can be interpreted genealogically as the rates with which the $(n+1)^{\text{th}}$ lineage mutates and is absorbed into $A^*(\mathbf{n})$ by a pairwise merger. The third term corresponds to a multiple merger between the $(n+1)^{\text{th}}$ lineage and two or more lineages in $\mathbf{n}$ of the same type. Because this last term involves coalescence between lineages in $\mathbf{n}$ it does not have an interpretation in terms of $A^*(\mathbf{n})$. However it can be forced into this framework by noting that the only relevant information is the time of absorption of the $(n+1)^{\text{th}}$ lineage and the type of the lineage(s) in $\mathbf{n}$ with which it merges. The following definition introduces dynamics for mutation and absorption into the trunk ancestry, which closely mimic the rates in (2.14). Hence, it can be expected to yield a good, tractable approximation to (2.14).

Definition 3. Let $\widehat{\pi}^{\Lambda,\mathrm{K}}(\mathbf{e}_h|\mathbf{n})$ be the distribution of the type of a lineage which, when traced backwards in time, encounters mutation events with rates θ_l according to the transition matrix $M^{(l)}$ at each locus $l \in L$ and is absorbed into $A^*(\mathbf{n})$ with rate

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

choosing its parent uniformly upon absorption. A parental type being thus acquired, the mutation events can be resolved forwards in time, yielding a random type at the leaf. The corresponding SMC proposal distribution is denoted by $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{K}}$.

Proposition 3. *$\widehat{\pi}^{\Lambda,K}(\mathbf{e}_h|\mathbf{n})$ solves the equations*

$$\left[\frac{\Lambda(\{0\})n}{2}+\theta+\frac{1}{n+1}\sum_{p=2}^{n+1}\binom{n+1}{p}\lambda_{n+1,p}\right]\widehat{\pi}^{\Lambda,K}(\mathbf{e}_h|\mathbf{n})=\frac{\Lambda(\{0\})n_h}{2}$$
$$+\sum_{l\in[k]}\theta_l\sum_{a\in E_l}M^{(l)}_{ah[l]}\widehat{\pi}^{\Lambda,K}(\mathbf{e}_{S^a_l(h)}|\mathbf{n})+\frac{n_h}{n(n+1)}\sum_{p=2}^{n+1}\binom{n+1}{p}\lambda_{n+1,p}$$

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

$$\frac{\theta M+\left[\Lambda(\{0\})/2+\frac{1}{n(n+1)}\sum_{p=2}^{n+1}\binom{n+1}{p}\lambda_{n+1,p}\right]N}{\theta+\frac{\Lambda(\{0\})n}{2}+\frac{1}{n+1}\sum_{p=2}^{n+1}\binom{n+1}{p}\lambda_{n+1,p}},\tag{2.15}$$

where N is the $|\mathcal{H}|\times|\mathcal{H}|$ matrix with each row equal to $(n_1,\ldots,n_{|\mathcal{H}|})$, and M is the $|\mathcal{H}|\times|\mathcal{H}|$ stochastic matrix corresponding to the mixture distribution of $\mathcal{H}$ with mixture weights θ_l/θ and mixture components $M^{(l)}$, suitably extended from $|E_l|\times|E_l|$ matrices to $|\mathcal{H}|\times|\mathcal{H}|$ matrices by adding zero entries as appropriate.

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

Note that $\widehat{\mathbf{P}}^{\Lambda,\mathrm{K}}_n$ has a very similar form to $\widehat{\mathbf{P}}^{\Lambda,\mathrm{SD}}_n$, and as a consequence of the linearity in N in (2.15) the efficient Gaussian quadrature approximation of Appendix A in [Stephens and Donnelly, 2000] can be applied to both with minor modifications for $\widehat{\mathbf{P}}^{\Lambda,\mathrm{K}}_n$.

2.3.2 Λ-coalescent simulation study

In this section I present an empirical comparison between the SMC algorithms defined by $\widehat{\mathbf{P}}^{\Lambda,\mathrm{K}}_n$ and $\widehat{\mathbf{P}}^{\Lambda,\mathrm{SD}}_n$, as well as the generalised Griffiths-Tavaré proposal distribution of Birkner and Blath [2008] which will be denoted by $\widehat{\mathbf{P}}^{\Lambda,\mathrm{GT}}_n$. Simulated samples have been generated using the efficient sampling algorithm provided in Section 1.4.4 of [Birkner and Blath, 2009]. Approximate CSDs have been evaluated using a Gauss quadrature of order four. See Appendix A of [Stephens and Donnelly, 2000] for details.

Simulated haplotypes consist of 15 loci with binary alleles denoted $\{0,1\}$ and mutation matrix $M^{(l)}=\begin{pmatrix}0&1\\1&0\end{pmatrix}$ at each locus. The coalescent is a Beta$(2-\alpha,\alpha)$-coalescent for $\alpha\in(1,2)$. All simulations have been run on a single core on a Toshiba

laptop, and make use of a stopping time resampling scheme 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.

The total mutation rate is $\theta = 0.1$ spread evenly among all 15 loci. The coalescent is The Beta$(0.5, 1.5)$-coalescent corresponding to $\alpha = 1.5$. The data consists of 100 sampled haplotypes, 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.

I will consider inferring both θ and α individually, assuming all other parameters are known and that $\theta_l = \theta/15$ for every $l \in [15]$. Eight independent simulations of 30 000 particles each were run on an evenly spaced grid of mutation rates spanning the interval $\theta \in [0.025, 0.2]$. 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 2.1.

The most striking observation is that both approximate CSD proposals yield algorithms which are two orders of magnitude faster than the Griffiths-Tavaré proposal algorithm. Moreover, it is clear that the α-surface obtained from $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{GT}}$ has not yet converged. The wide confidence envelope at the left hand edge and the lack of monotonicity at the right hand edge of the $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{GT}}$ θ-surface are indicative of poorer performance when inferring θ as well.

The runtimes of $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{SD}}$ and $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{K}}$ are very similar for both parameters, and all four surfaces from these proposals are good approximations of the truth. In the θ-case the accuracy of the two is very similar, but in the α-case $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{K}}$ 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. Figure 2.2 shows a joint likelihood heat map for the two parameters constructed from a grid of simulations of 30 000 particles from the $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{K}}$ 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.

The performance of the $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{SD}}$ proposal can be expected to deteriorate with more demanding data sets, and when the true model is very different from Kingman's coalescent. 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,

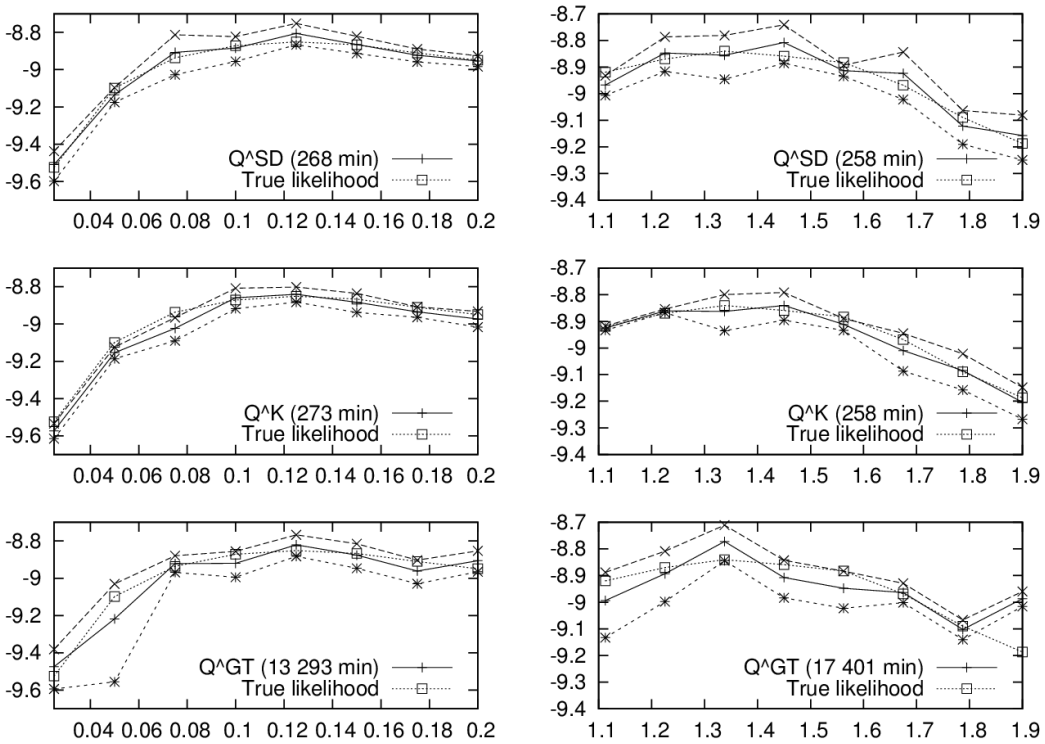


Figure 2.1: Simulated log-likelihood surfaces from 30 000 particles with ±2SE confidence envelopes based on assuming IID Gaussian weights. The left column is for θ and the right for α. The true surfaces are based on a 1 000 000 particle simulation using the $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{K}}$ proposal distribution.

each a single mutation removed from the main group. The sizes of these groups are 3, 1, 1, 1. The results are shown in Figure 2.3.

The algorithm using $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{K}}$ is noticeably faster when inferring θ, and slightly faster when inferring α. It also produces substantially more accurate estimates than $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{SD}}$ for small values of θ. 30 000 particle runs have not yielded an accurate estimate for large values of θ from either algorithm. The α-surface from $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{SD}}$ looks superficially better, but both surfaces are very similar and good matches to the true likelihood.

This deterioration of the performance of $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{SD}}$ 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 seems plausible that the high values of the

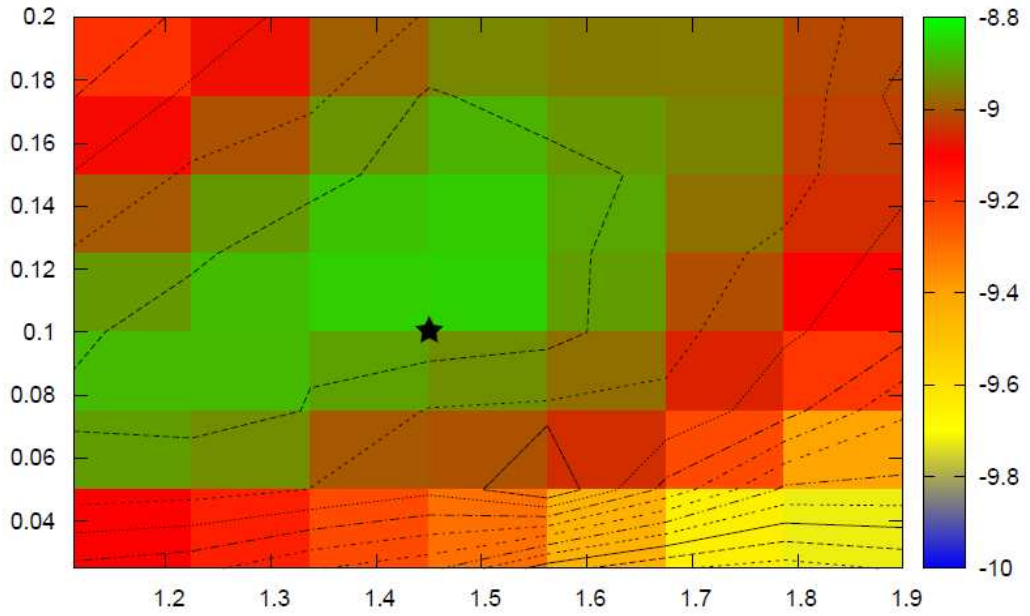


Figure 2.2: Simulated likelihood surface for the Beta$(2-\alpha, \alpha)$-coalescent family from 30 000 particles using $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{K}}$. Values of α run on the x-axis, and values of θ on the y-axis. The surface is interpolated from an 8×8 grid of independent simulations. The star denotes the MLE, which must lie on one of the grid points.

likelihood estimator near $\theta = 0.06$ using $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{SD}}$ coincide with the MLE for this data set under the assumption $\Lambda = \delta_0$, and is thus an artefact of the mismatch between the proposal distribution and the target likelihood. Hence $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{K}}$ is the recommended proposal distribution in practice.

The reported run times in Figures 2.1 and 2.3 suggest that the SMC algorithm using $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{K}}$ as its proposal distribution remains 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 [Griffiths and Tavaré, 1994c] or bridge sampling [Meng and Wong, 1996] are employed to reduce the number of independent simulations. There is also a strong dependence on model parameters: fast coalescence (or low α in this setting) corresponds to faster simulation runs, and both high mutation rate and large haplotype space will result in a slower algorithm.

2.3.3 Ξ-coalescents

The important tools in deriving the optimal proposal distributions $\widehat{\mathbf{P}}_n^{\Lambda,\mathrm{K}}$ and the approximate CSDs $\widehat{\pi}^{\Lambda,\mathrm{K}}(\cdot|\cdot)$ were, respectively, the lookdown construction Donnelly

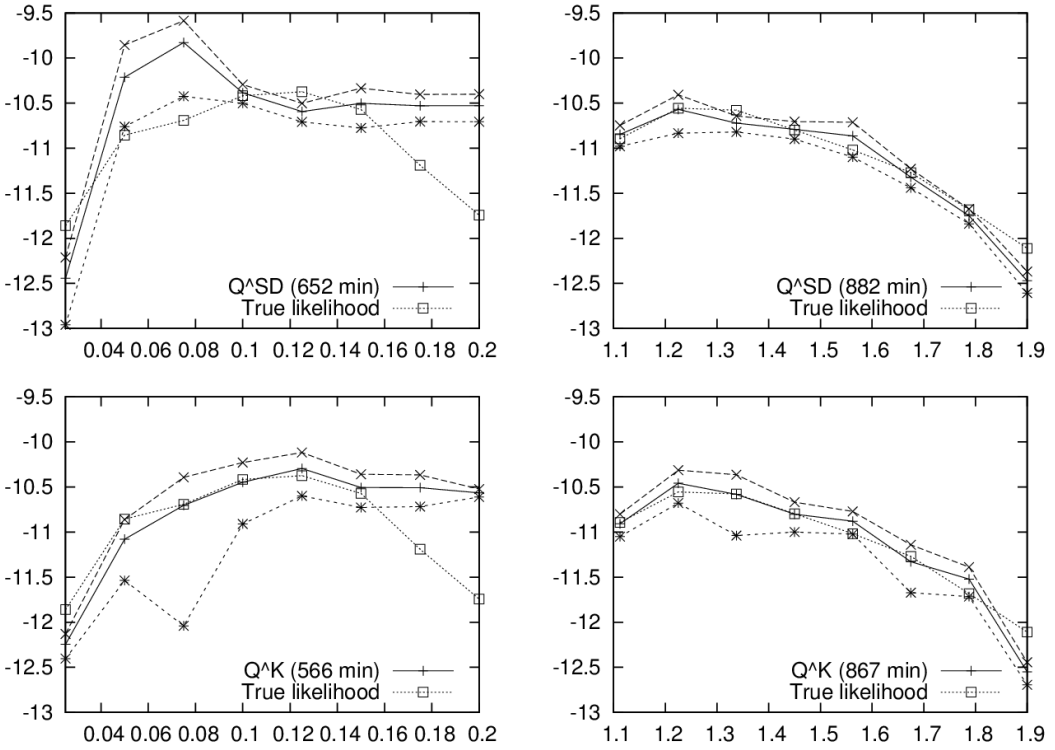


Figure 2.3: Simulated log-likelihood surfaces from 30 000 particles with ±2SE confidence envelopes assuming IID Gaussian weights. The left column is for θ and the right for α. The downward spike in the lower confidence boundary of the bottom left surface is an artefact caused by a negative value of the estimate, where the real part of the logarithm has been plotted. The values of the standard errors have no such spike.

and Kurtz [1996, 1999] and the trunk ancestry Paul and Song [2010]. Both of these are also available for the Ξ-coalescent, and in this section I make use of them to extend the SMC algorithm to this family of coalescent processes.

A lookdown construction for the Ξ-coalescent and the Ξ-Fleming-Viot process was derived by Birkner et al. [2009] and can be described as follows. For ease of notation I assume $\Xi(\{\mathbf{0}\}) = 0$. If Ξ does have an atom at zero, its treatment is identical to the Λ-case.

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

$$dt \otimes \|\mathbf{r}\|_2^{-2} \Xi(d\mathbf{r}) \otimes du^{\otimes \mathbb{N}}$$

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

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

At each $(t_j,(r_{j1},r_{j2},\dots),(u_{j1},u_{j2},\dots)) \in N^{\Xi}$ group the n particles such that all particles $d\in[n]$ with $g(\mathbf{r}_j,u_{jd})=k$ 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 similarly to the Λ-coalescent.

This lookdown construction will be instrumental in establishing the following recursion, which is a finite sites analogue of the sampling recursion derived by Möhle [2006] for the infinite alleles model. The coefficients of the recursion will form the forwards-in-time transition probabilities of a branching and mutating particle system, analogously to (2.7) in the Λ-coalescent case.

Theorem 3. *The likelihood of type frequencies $\mathbf{n}\in\mathbb{N}^{|\mathcal{H}|}$ sampled from the stationary Ξ-Fleming-Viot process solves*

$$\begin{aligned}\mathbf{P}_n^{\Xi}(\mathbf{n}) =& \frac{1}{g_n+n\theta}\Bigg\{ \sum_{h:n_h>0}\sum_{l\in[k]}\theta_l \sum_{a\in E_l}\left(n_{S_l^a(h)}+1-\delta_{ah[l]}\right) M^{(l)}_{ah[l]}\mathbf{P}_n^{\Xi}(\mathbf{n}-\mathbf{e}_h+\mathbf{e}_{S_l^a(h)}) \\ &+ \sum_{k_1=1}^{n_1}\cdots\sum_{k_{|\mathcal{H}|}=1}^{n_{|\mathcal{H}|}}\sum_{\pi^1\in P_{n_1}^{k_1}}\cdots\sum_{\pi^{|\mathcal{H}|}\in P_{n_{|\mathcal{H}|}}^{k_{|\mathcal{H}|}}} \mathbb{1}_{[n]}\left(\sum_{h\in\mathcal{H}}k_h\right)\binom{n}{|\pi_1^1|,|\pi_2^1|,\dots,|\pi_{|\mathcal{H}|}^{|\mathcal{H}|}|} \\ &\times \binom{|\vee_{h\in\mathcal{H}}\pi^h|}{|\pi^1|,\dots,|\pi^{|\mathcal{H}|}|}^{-1} \lambda_{n;K(\vee_{h\in\mathcal{H}}\pi^h);S(\vee_{h\in\mathcal{H}}\pi^h)}\mathbf{P}_k^{\Xi}(\mathbf{k})\Bigg\}, \end{aligned} \tag{2.16}$$

with the convention that $\sum_{k=1}^{0} f(k) = f(0)$ and with boundary condition $\mathbf{P}_1^{\Xi}(\mathbf{e}_h) = m(h)$, where m is the stationary distribution of the mutation process, which is assumed to exist and be unique. Here $P_{n_h}^{k_h}$ denotes the set of equivalence relations on $n_h\in\mathbb{N}$ elements with $k_h\leq n_h$ equivalence classes, $\pi^h=(\pi_1^h\dots\pi_{k_h}^h)$ 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. When $\vee_{h\in\mathcal{H}}\pi^h$ consists of only singletons, the whole corresponding summand is set to 0 by convention. 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

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

Proof. The proof is the same as in Section 1.4.1 of Birkner and Blath [2009], adapted here from the Λ-coalescent to the Ξ-coalescent. Let p denote 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

$$p(y_1,\ldots,y_n) = \frac{1}{g_n + n\theta} \Bigg\{ \sum_{i=1}^{n} \sum_{l \in [k]} \theta_l \sum_{a \in E_l} M^{(l)}_{ah[y_i]} p(y_1,\ldots,y_{i-1},S^a_l(y_i),y_{i+1},\ldots,y_n) + \sum_{\pi \in P(\mathbf{y})} \lambda_{n;K(\pi);S(\pi)} p(\gamma_\pi(y_1,\ldots,y_n)) \Bigg\} \tag{2.17}$$

where $P(\mathbf{y})$ is the set of equivalence relations describing permissible mergers for the sample $\mathbf{y} = (y_1,\ldots,y_n)$ (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 look-down-and-copy event denoted by the equivalence relation π takes place.

By exchangeability, only a vector of type frequencies $\mathbf{n} = (n_1,\ldots,n_{|\mathcal{H}|})$ is needed. For such a vector, define the canonical representative as

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

and the likelihood as

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

The following two identities yield the desired recursion when substituted into (2.17):

$$\begin{aligned}
n_h\binom{n}{n_1,\ldots,n_{|\mathcal{H}|}}&p(\kappa(\mathbf{n}-\mathbf{e}_h+\mathbf{e}_{S_l^a(h)}))\\
&=(n_{S_l^a(h)}+1-\delta_{ah[l]})p^0(\mathbf{n}-\mathbf{e}_h+\mathbf{e}_{S_l^a(h)}),\\
\binom{n}{n_1,\ldots,n_{|\mathcal{H}|}}&\prod_{h\in\mathcal{H}}\binom{n_h}{|\pi_1^h|,\ldots,|\pi_{k_h}^h|}p(\kappa(\mathbf{k}))\\
&=\binom{n}{|\pi_1^1|,|\pi_2^1|,\ldots,|\pi_{|\mathcal{H}|}^{|\mathcal{H}|}|}\binom{k}{k_1,\ldots,k_{|\mathcal{H}|}}^{-1}p^0(\mathbf{k}).
\end{aligned}$$

□

The coefficients of (2.16) are the Ξ-coalescent analogues of the forwards probabilities (2.7). The solution to (2.16) can be approximated by importance sampling as in the Λ-coalescent case, and the following theorem is a straightforward extension of Theorem 1.

Theorem 4. *The optimal proposal distributions for recursion* (2.16)*, denoted* $\widetilde{\mathbf{P}}_n^{\Xi}$*, are*

$$\widetilde{\mathbf{P}}_n^{\Xi}(A_{i-1}|A_i)\propto$$
$$\begin{cases}
n_h\theta_l\frac{\pi(\mathbf{e}_{S_l^a(h)}|A_i-\mathbf{e}_h)}{\pi(\mathbf{e}_h|A_i-\mathbf{e}_h)}M_{ah[l]}^{(l)} & \text{if } A_{i-1}=A_i-\mathbf{e}_h+\mathbf{e}_{S_l^a(h)}\\
\sum_{\pi^1\in P_{n_1}^{k_1}}\cdots\sum_{\pi^{|\mathcal{H}|}\in P_{n_{|\mathcal{H}|}}^{k_{|\mathcal{H}|}}}\prod_{h\in\mathcal{H}}\binom{n_h}{|\pi_1^h|,\ldots,|\pi_{k_h}^h|}\frac{\lambda_{n;K(\vee_{h\in\mathcal{H}}\pi^h);S(\vee_{h\in\mathcal{H}}\pi^h)}}{\pi(\mathbf{n}-\mathbf{k}|\mathbf{k})} & \\
\text{if } A_i=\mathbf{n} \text{ and } A_{i-1}=\mathbf{k} \text{ for } k_h\in[n_h] \text{ and } \sum_{h\in\mathcal{H}}k_h<n &
\end{cases}$$

where n *and* n_h *denote haplotype frequencies of* A_i*.*

Proof. The argument is identical to the proof of Theorem 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 4 are intractable, 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 I present a derivation from the trunk ancestry $A^*(\mathbf{n})$.

Definition 4. Let $\widehat{\pi}^{\Xi,K}(\mathbf{e}_h|\mathbf{n})$ be the CSD obtained by letting the $(n+1)^{\text{th}}$ lineage mutate with rates $\{\theta_l\}_{l\in L}$ with transition matrices $\{M^{(l)}\}_{l\in L}$, and be absorbed into $A^*(\mathbf{n})$ with rate

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

choosing its parent uniformly upon absorption. A parental type being thus acquired, the mutation events can be resolved forwards in time, yielding a random type at the leaf.

Proposition 4. *The approximate CSDs $\widehat{\pi}^{\Xi,K}(\mathbf{e}_h|\mathbf{n})$ solve the following recursion:*

$$\begin{aligned}&\left[\theta+\frac{1}{n+1}\sum_{k=1}^{n}\sum_{\pi\in P_{n+1}^k}\binom{n+1}{|\pi_1|,\ldots,|\pi_k|}\lambda_{n+1;K(\pi);S(\pi)}\right]\widehat{\pi}_{\Xi,K}(\mathbf{e}_h|\mathbf{n})\\ &=\frac{n_h}{n(n+1)}\sum_{p=1}^{n}\sum_{\pi\in P_{n+1}^p}\binom{n+1}{|\pi_1|,\ldots,|\pi_p|}\lambda_{n+1;K(\pi);S(\pi)}\\ &\quad+\sum_{l\in[k]}\theta_l\sum_{a\in E_l}P_{ah[l]}^{(l)}\widehat{\pi}^{\Xi,K}(\mathbf{e}_{S_l^a(h)}|\mathbf{n})\end{aligned}$$

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

$$\frac{\theta M+\left\{\frac{1}{n(n+1)}\sum_{p=1}^{n}\sum_{\pi\in P_{n+1}^p}\binom{n+1}{|\pi_1|,\ldots,|\pi_p|}\lambda_{n+1;K(\pi);S(\pi)}\right\}N}{\theta+\frac{1}{n+1}\sum_{p=1}^{n}\sum_{\pi\in P_{n+1}^p}\binom{n+1}{|\pi_1|,\ldots,|\pi_p|}\lambda_{n+1;K(\pi);S(\pi)}}.$$

where N and M are as in Proposition 3.

Proof. The proof is identical to Proposition 3 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 4 is no longer valid and multivariate approximate CSDs $\widehat{\pi}^{\Xi,K}(\mathbf{m}|\mathbf{n})$ must also be specified. This is most naturally done by averaging over all permutations of the lineages in $\mathbf{m}$, but this is computationally infeasible for all but very small samples $\mathbf{m}$. 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 [Möhle and Sagitov, 2003; Birkner et al., 2013] seems computationally feasible. In such a scenario care must be taken when using driving value or bridge sampling methods, because while a measure Ξ placing full mass on an n-dimensional simplex can be used to drive a simulation for any other measure Ξ placing full mass on a $k \leq n$ dimensional simplex, the converse will result in mutually singular target and proposal distributions.

2.4 An alternative to SMC: product of approximate conditionals

In this section I will introduce two PAC algorithms for Λ-coalescents making use of, respectively, $\widehat{\pi}^{\Lambda,\mathrm{K}}$ and a modification $\widehat{\pi}^{\Lambda,\mathrm{K2}}$, as well as investigate their efficiency and accuracy. The CSD $\widehat{\pi}^{\Lambda,\mathrm{K2}}$ is defined as the distribution of the haplotype of a lineage which encounters mutations with rates $\{\theta_l\}_{l\in[k]}$ as before, and is absorbed into $A^*(\mathbf{n})$ with rate

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

choosing its parent uniformly, and inheriting the parental haplotype. Mutations are then resolved forwards in time from the parental haplotype with transition matrices $\{M^{(l)}\}_{l\in[k]}$.

Note that because $\widehat{\pi}^{\Lambda,\mathrm{K2}}(\cdot|\mathbf{n})$ depends nonlinearly on the exact frequencies $\{n_h\}_{h\in\mathcal{H}}$, the precomputations which were possible for all other CSDs introduced thus far are not possible for it. See Proposition 1 of [Stephens and Donnelly, 2000] for details. For an SMC algorithm this loss of efficiency in evaluating the CSDs would be devastating, but PAC algorithms are fast enough to remain feasible. The increased speed of PAC algorithms has also enabled the order of the Gauss quadrature used to approximate the CSDs to be increased from four to ten for both families, resulting in more accurate approximations of PAC-estimators.

Neither approximate CSD family is exchangeable, so the estimates of the

likelihood depend on the order in which the count data **n** is conditioned upon. Fixing a representative ordering is inadequate, because it is well known that the ordering can substantially influence the PAC estimator [Li and Stephens, 2003]. This issue is partially addressed by averaging estimates across 1 000 uniformly sampled random permutations of the data, following the approach of Li and Stephens [2003] as well as subsequent works making use of the PAC method. The number of permutations is substantially larger than what has been used for PAC models based on Kingman's coalescent, and proved necessary in trial runs (results not shown), but comes at little additional cost. The results of applying these PAC algorithms to both simulated data sets from Section 2.3.2 are summarised in Figures 2.4 and 2.5.

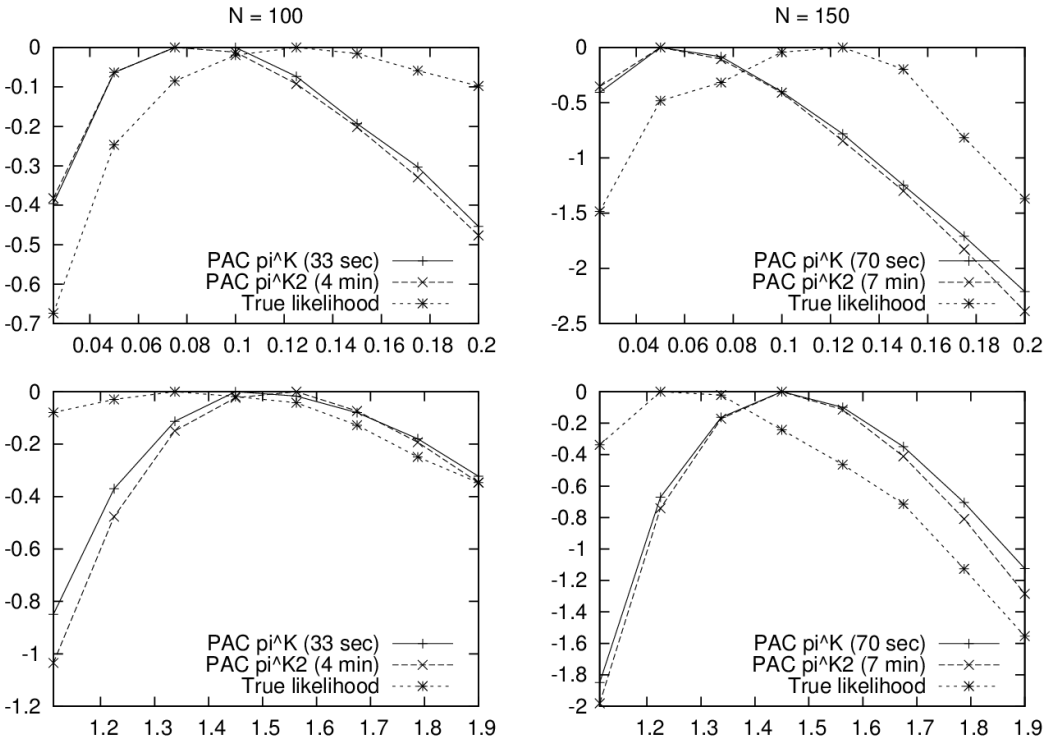


Figure 2.4: The PAC log-likelihood surfaces normalised to 0 following Li and Stephens [2003]. The true likelihood surfaces in the left column are those from Figure 2.1, and the true surfaces in the right column are those from Figure 2.3.

The results of the PAC simulations are mixed. Both PAC algorithms are extremely fast, and likely to remain feasible even for large data sets, but the PAC likelihood estimates are consistently too low by many orders of magnitude. Nevertheless, the PAC MLEs in Figure 2.4 are close to the true maximisers, particularly for the smaller data set in the left column. On the other hand, the joint PAC

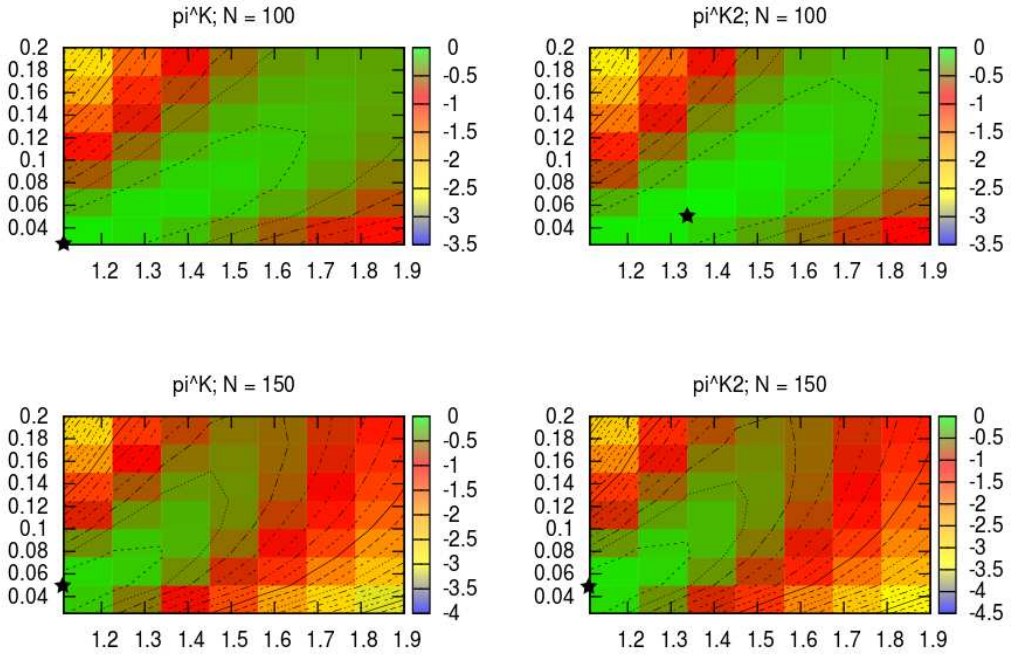


Figure 2.5: The PAC joint log-likelihood surfaces normalised to 0. Locations of MLEs are indicated by stars. Figure 2.2 provides a suitable SMC comparison to the top row.

likelihood surfaces in Figure 2.5 broadly capture the diagonal shape seen in Figure 2.2, but the location of the PAC MLEs is significantly offset. The two PAC methods perform very similarly in the one-dimensional problems in Figure 2.4, but the 2D surface obtained from $\widehat{\pi}^{\Lambda,\mathrm{K2}}$ is a better fit than that from $\widehat{\pi}^{\Lambda,\mathrm{K}}$ for the smaller sample. For the larger sample the two surfaces are nearly identical.

The run times in Figure 2.4 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 SMC, 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. These preliminary simulations moti-

vate the undertaking, and confirm that the PAC method is able to provide useful, principled and fast results for Λ-coalescents in some cases.

2.5 SMC for spatial Λ-coalescents

As with the panmictic coalescent processes in the previous section, the likelihood of an observed haplotype configuration $\mathbf{n}$ from the spatial Λ-coalescent is intractable. However, it too can be cast into the framework of stopped Markov processes by conditioning on N, the Poisson process driving the extinction-recolonisation events, and using the lookdown construction [Véber and Wakolbinger, 2015] to yield the following integral recursion, which is the analogue of the sampling recursion (2.16).

A standing assumption of this section will be that all sampling locations are distinct, which also means that any two lineages can be distinguished from one another. This is in stark contrast to the earlier panmictic coalescents, for which exchangeability guaranteed that the only relevant information was haplotype frequencies. To reflect this difference, I abuse notation and denote a sample $\mathbf{n} \in (\mathbb{T}, \mathcal{H})^n$ as an unordered set of pairs of locations and haplotypes. The operation $\mathbf{n} \oplus (z, h)$ denotes the sample $\mathbf{n}$ with a further lineage (z, h) added to it, while $n \ominus (z, h)$ denotes $\mathbf{n}$ with the entry (z, h) removed. The assumed distinctness of sampling locations ensures that this notation causes no ambiguity.

The coefficients of the following recursion once again form transition probabilities of a forwards-in-time branching and mutating particle system started at the MRCA, analogously to (2.7) in the Λ-coalescent case. Identifying these transition probabilities will enable the use of reverse time SMC to approximate hitting probabilities, or spatial Λ-coalescent likelihoods.

Proposition 5. *The likelihood of an observed configuration* $\mathbf{P}_n^{SL}(\mathbf{n})$ *solves*

$$\begin{aligned}\mathbf{P}_n^{SL}(\mathbf{n}) = &\frac{1}{L^2+\theta n}\sum_{i=1}^{n}\sum_{l\in[k]}\theta_l\sum_{a\in E_l}M^{(l)}_{ah_i[l]}\mathbf{P}_n^{SL}(\mathbf{n}\ominus(z_i,h_i)\oplus(z_i,S_l^a(h_i)))\\
&+\frac{1}{L^2+\theta n}\sum_{h\in\mathcal{H}}\int_{\mathbb{T}(L)}\int_{B_r(x)}\sum_{\substack{J\subseteq N^x_{\mathbf{n}}:\\ h'=h\ \forall(z,h')\in J}}\frac{u^{|J|}(1-u)^{|N^x_{\mathbf{n}}\ominus J|}}{(|J|\vee 1)\pi r^2}\\
&\times\mathbf{P}^{SL}_{n-|J|+\mathbb{1}_{\{1,2,\dots\}}(|J|)}(\mathbf{n}\ominus J\oplus(z,h)\mathbb{1}_{\emptyset^c}(J))dzdx, \qquad (2.18)\end{aligned}$$

with boundary condition $\mathbf{P}_1^{SL}((z,h)) = m(h)$ *for any* $z \in \mathbb{T}(L)$*, where*

$$N^x_{\mathbf{n}} := \{(z,h)\in\mathbf{n} : z\in B_r(x)\}$$

is the set of lineages within the disc of the extinction-recolonisation region $B_r(x)$.

Remark 5. The notation $\mathbf{P}^{\mathrm{SL}}_{n-|J|+\mathbb{1}_{\{1,2,\dots\}}(|J|)}(\mathbf{n} \ominus J \oplus (z,h)\mathbb{1}_{\emptyset^c}(J))$ denotes $\mathbf{P}^{\mathrm{SL}}_n(\mathbf{n})$ when J is empty, and $\mathbf{P}^{\mathrm{SL}}_{n-|J|+1}(\mathbf{n} \ominus J \oplus (z,h))$ when J is non-empty.

Proof. Consider a configuration $\mathbf{n}$ given by the n particles with the lowest levels in the lookdown construction, and trace their trajectories backwards in time until the first event is encountered. If the event in reverse time is a mutation from a type h to $S^a_l(h)$ then, forwards in time, the corresponding event happens at rate $\theta_l M^{(l)}_{ah[l]}$.

If the event is an extinction-recolonisation event at $x \in \mathbb{T}(L)$ in which $0 < m \leq n$ particles lie within $B_r(x)$ and $0 \leq k \leq m$ particles jump, then the corresponding jump forwards in time happens with probability $u^k(1-u)^{m-k}$. The event that the particles which jump have the locations required by $\mathbf{n}$ happens with density $(\pi r^2)^{-k}$, but interpreting the jumps as coalescences means that only the location of the particle with the lowest level plays a role. Hence the locations of $k-1$ particles can be integrated out, as they correspond to the same coalescence event, leaving a total density of $(\pi r^2)^{-1}$ provided that at least one particle jumped.

The ordering of levels is only relevant in that the particle with the lowest level must end up at the correct parental location. The levels of all other particles play no role in the coalescence event. Hence mixing uniformly over levels produces a factor of k^{-1} when at least one particle jumped. Finally, extinction-recolonisation events centred at $x \in \mathbb{T}(L)$ happen at rate 1 when mixed over realisations of N, which completes the proof. □

The coalescence rates in (2.18) are computationally intractable due to the $\int_{\mathbb{T}(L)} \cdots dx$-integral over possible extinction-recolonisation event centres. To make progress I extend the state space to include the sample configuration $\mathbf{n}$ and the almost surely finite, ordered vector of event centres $\mathbf{x} := \{x_k\}_{k=1}^K$ connecting the sample to the MRCA. The joint likelihood $(\mathbf{P}^{\mathrm{SL}}_n \otimes N)(\mathbf{n}, \mathbf{x})$ can then be used to recover the marginal likelihood of interest via

$$\mathbf{P}^{\mathrm{SL}}_n(\mathbf{n}) = \int (\mathbf{P}^{\mathrm{SL}}_n \otimes N)(\mathbf{n}, \mathbf{x}) d\mathbf{x} = \int \mathbf{P}^{\mathrm{SL}}_{n,\mathbf{x}}(\mathbf{n}) N(d\mathbf{x}), \tag{2.19}$$

where I have abused notation and let $N(d\mathbf{x})$ denote the marginal distribution of $\mathbf{x}$. Since the MRCA is reached using only finitely many events with probability 1 and $N(d\mathbf{x})$ is easy to sample, it is sufficient to obtain a sample $\{\mathbf{x}^i\}_{i=1}^p$ of vectors of

event centres as well as a conditional estimator $\widehat{\mathbf{P}}^{\mathrm{SL}}_{n,\mathbf{x}}(\mathbf{n})$, and estimate (2.19) via

$$\widehat{\mathbf{P}}^{\mathrm{SL}}_{n}(\mathbf{n}) = \frac{1}{p}\sum_{i=1}^{p} \widehat{\mathbf{P}}^{\mathrm{SL}}_{n,\mathbf{x}^{\mathbf{i}}}(\mathbf{n}).$$

Note that the number of required event centres is random, but new event centres can be sampled on-line as necessary due to the independence structure of the Poisson process which drives the events. It will not be necessary to store the locations of events once their effect on the genealogy has been resolved.

The analogue of (2.18) for the conditioned process given $\mathbf{x}$ is readily obtained from the lookdown construction by mixing only on the times of the events of N while retaining the conditioning upon locations of centres:

$$\begin{aligned}\mathbf{P}^{\mathrm{SL}}_{n,\mathbf{x}}((\mathbf{n},q)) = \frac{1}{L^2+\theta n}\sum_{i=1}^{n}\sum_{l\in[k]}\theta_l \sum_{a\in E_l} M^{(l)}_{ah_i[l]}\mathbf{P}^{\mathrm{SL}}_{n,\mathbf{x}}(\mathbf{n}\ominus(z_i,h_i)\oplus(z_i,S^a_l(h_i)),q)\\ + \frac{L^2}{L^2+\theta n}\sum_{h\in\mathcal{H}}\int_{B_r(x_k)}\sum_{\substack{J\subseteq N^{x_q}_{\mathbf{n}}:\\ h'=h\ \forall(z,h')\in J}}\frac{u^{|J|}(1-u)^{|N^{x_q}_{\mathbf{n}}\ominus J|}}{(|J|\vee 1)\pi r^2}\\ \times \mathbf{P}^{\ \mathrm{SL}}_{n-|J|+\mathbb{1}_{\{1,2,\ldots\}}(|J|),\mathbf{x}}((\mathbf{n}\ominus J\oplus(z,h)\mathbb{1}_{\emptyset^c}(J),q+1))dz, \qquad (2.20)\end{aligned}$$

where the index $q\in\mathbb{N}$ has been introduced to track the index of the next event in $\mathbf{x}$. The associated boundary condition is $\mathbf{P}^{\mathrm{SL}}_{n,\mathbf{x}}(((z,h),q)) = m(h)$ for any $\mathbf{x}$, $q\in\mathbb{N}$ and $z\in\mathbb{T}(L)$. The coefficient L^2 in front of the second term arises as the rate of arrival of the next event, given the location of its centre.

The coefficients in (2.20) are specified in closed form, and can be interpreted as the forwards transition probabilities

$$\begin{aligned}\mathbf{P}^{\mathrm{SL}}_{n,\mathbf{x}}((\mathbf{n},q)|(\mathbf{n}\ominus(z,h)\oplus(z,S^a_l(h)),q)) &= \frac{\theta_l}{L^2+\theta n}M^{(l)}_{ah[l]},\\ \mathbf{P}^{\mathrm{SL}}_{n,\mathbf{x}}((\mathbf{n},q+1)|(\mathbf{n}\ominus J\oplus(z,h),q)) &= \frac{L^2}{L^2+\theta n}\frac{u^{|J|}(1-u)^{|N^{x_q}_{\mathbf{n}}\ominus J|}}{|J|\pi r^2},\\ \mathbf{P}^{\mathrm{SL}}_{n,\mathbf{x}}((\mathbf{n},q+1)|(\mathbf{n},q)) &= \frac{L^2}{L^2+\theta n}(1-u)^{|N^{x_q}_{\mathbf{n}}|}.\end{aligned}$$

Denote the corresponding stochastic process by $\{(X_j,q_j)\}_{j\in\mathbb{N}}$ with X_0 being the MRCA and $q_0=K$, where K is unknown number of events of N needed to reach the MRCA. Let $\tau_1=0$ and $\{\tau_n\}_{n\geq 2}$ be a family of stopping times defined inductively via

$$\tau_n := \inf\{j\geq\tau_{n-1} : |X_j|\geq n\}.$$

Analogously to Section 2.3, the sets defining the trajectory of interest are

$$I = \{((z,h),K) : (z,h) \in \mathbb{T}(L) \times \mathcal{H}\}, \text{ the MRCA},$$
$$T = \{(X_{\tau_{n+1}}, q_{\tau_{n+1}})\}, \text{ any sample whose size exceeds } n,$$

and the quantity of interest is

$$\mathbb{E}_\mu[f(\tau_T, X_{0:\tau_T})] = \mathbb{E}_\mu[\mathbb{1}_{\{X_{\tau_{n+1}-1}=\mathbf{n}\}}],$$

i.e. the probability of the observed state $\mathbf{n}$ occurring immediately before the sample size exceeds n.

As in Section 2.3, the optimal (in terms of estimator variance) reverse-time proposal distribution can be expressed in terms of the forwards transition probabilities and a family of CSDs $\{\pi_z^{\mathbf{x},q}(\cdot|\mathbf{n})\}_{z\in\mathbb{T}(L)}$, which can be interpreted as the distribution of a genetic type sampled from a point $z \in \mathbb{T}(L)$ between the times of the $(q-1)^{\text{th}}$ and q^{th} events of a given a realisation $\mathbf{x}$ given an observed sample $\mathbf{n}$. I immediately introduce the first simplifying assumption and assume the only dependence on $\mathbf{x}$ and q is via x_q, the centre of the q^{th} event. The resulting approximate family is denoted by $\{\widehat{\pi}_z^{x_q}(\cdot|\mathbf{n})\}_{z\in\mathbb{T}(L)}$.

The corresponding approximation to the optimal proposal distribution can then be written as

$$\widehat{\mathbf{P}}^{\text{SL}}_{n,\mathbf{x}}((\mathbf{n}\ominus(z,h)\oplus(z,S_l^a(h)),k)|(\mathbf{n},k)) \propto \frac{\widehat{\pi}_z^{x_q}(S_l^a(h)|\mathbf{n}\ominus(z,h))}{\widehat{\pi}_z^{x_q}(h|\mathbf{n}\ominus(z,h))}\frac{\theta_l}{L^2+\theta n}M^{(l)}_{ah[l]},$$
$$\widehat{\mathbf{P}}^{\text{SL}}_{n,\mathbf{x}}((\mathbf{n},k+1)|(\mathbf{n},k)) \propto \frac{L^2}{L^2+\theta n}(1-u)^{|N_{\mathbf{n}}(x_k)|},$$
$$\widehat{\mathbf{P}}^{\text{SL}}_{n,\mathbf{x}}((\mathbf{n}\ominus J\oplus(z,h),k+1)|(\mathbf{n},k)) \propto \frac{\widehat{\pi}_z^{x_q}(h|\mathbf{n}\ominus J)}{\widehat{\pi}^{x_q}_{\mathbf{z}_J}(\underbrace{h,\ldots,h}_{|J|}|\mathbf{n}\ominus J)}\frac{L^2}{L^2+\theta n}\frac{u^{|J|}(1-u)^{|N_{\mathbf{n}}(x_q)\ominus J|}}{\pi r^2},$$

where $\mathbf{z}_J$ denotes the vector of locations of lineages in J, and $\widehat{\pi}^{x_q}_{\mathbf{z}_J}(h,\ldots,h|\mathbf{n})$ is the multivariate extension of $\widehat{\pi}_z^{x_q}(h|\mathbf{n})$. The proof of optimality of this proposal distribution (allowing for the simplifying assumptions on the CSD introduced above) is identical in form to Theorem 1 of Stephens and Donnelly [2000], or either Theorem 1 or 4 above, and is omitted. The construction of the approximate CSDs below will ensure that $\widehat{\pi}^{x_q}_{\mathbf{z}_J}(\cdot|\mathbf{n})$ is invariant under permutations of J whenever all lineages in J share a common type. These are the only kinds of mergers permitted by the spatial Λ-coalescent, so that the multivariate extension will be well-defined for all necessary

arguments.

I will construct the CSDs for coalescence and mutation separately, and focus first on the univariate $\widehat{\pi}_z^{x_q}(h|\mathbf{n})$ for the mutation term. In this case I ignore spatial structure and view a sample $\mathbf{n}$ as arising from a standard Λ-coalescent with non-standard coalescence rates given by

$$\lambda_{n,p} = u^p(1-u)^{n-p}. \tag{2.21}$$

A family $\widehat{\pi}^{\Lambda,\mathrm{K}}(\cdot|\mathbf{n})$ of approximate CSDs for Λ-coalescents was derived in Section 2.3.1, and can be used here with the modification (2.21) to yield a tractable family $\widehat{\pi}_z^{x_q}(\cdot|\mathbf{n}) = \widehat{\pi}^{\Lambda,\mathrm{K}}(\cdot|\mathbf{h}(\mathbf{n}))$, where $\mathbf{h}(\mathbf{n}) \in \mathbb{N}^d$ is the vector of haplotype frequencies in the sample $\mathbf{n}$.

For the coalescence terms I neglect all lineages outside the disc $B_r(x_q)$, and treat the lineages in $B_r(x_q)$ as a sample from a Λ-coalescent with coalescence rates given by (2.21). Thus the type and number of lineages to merge can be sampled from $\widehat{\pi}_z^{x_q}(\cdot|\mathbf{n}) = \widehat{\pi}^{\Lambda,\mathrm{K}}(\cdot|\mathbf{h}(N_{\mathbf{n}}^{x_q}))$, followed by selecting the precise set of lineages to merge uniformly at random among all those of the correct type. Finally, the parental location is sampled uniformly from $B_r(x_q)$. This construction is uniquely defined because only lineages inside the disc $B_r(x_q)$ are allowed to coalesce, so ignoring lineages outside of it causes no ambiguity.

Remark 6. Introducing approximations of the conditional distribution of parental locations, and of the random environment Π, can be expected to yield more efficient algorithms than that outlined above. I attempted several heuristic models based on mixtures of Gaussian distributions centred on the locations of remaining lineages, but obtained importance weights with very high (and seemingly infinite) variance (results not shown). Deriving practically useful approximations remains an important open problem in likelihood-based inference for the spatial Λ-coalescent.

The multivariate extensions $\widehat{\pi}_{\mathbf{z}_J}^{x_q}(h,\ldots,h|\mathbf{n}\ominus J)$ can be defined from the univariate distributions via

$$\widehat{\pi}_{\mathbf{z}_J}^{x_q}(h,\ldots,h|\mathbf{n}\ominus J) := \prod_{i=1}^{|J|} \widehat{\pi}_{(\mathbf{z}_J)_i}^{x_q}(h\big|\mathbf{n}\ominus J\oplus_{j=1}^{i-1}((\mathbf{z}_J)_j,h)). \tag{2.22}$$

Despite the fact that the approximate CSDs are not exchangeable, the product on the RHS of (2.22) is uniquely defined because it depends on the vector $\mathbf{z}_J$ only through the fact that all locations satisfy $(\mathbf{z}_J)_i \in B_r(x_q)$, and it is only evaluated for sets of lineages that are identical in type. Analogously to Remark 4, these two facts

ensure that $\widehat{\pi}^{x_q}_{\mathbf{z}_J}(h,\dots,h|\mathbf{n})$ is invariant under permutations of J. This would not be true if more than a single parent were permitted in each extinction-recolonisation event, because then simultaneous mergers of several types of lineages to different parents would be possible. Strategies for defining approximate CSDs for such spatial Ξ-coalescents might include fixing a canonical permutation to ensure (2.22) is well-defined, or averaging over a small number of random permutations as in [Li and Stephens, 2003].

In summary, the proposal distribution is defined as

$$\widehat{\mathbf{P}}^{\mathrm{SL}}_{n,\mathbf{x}}((\mathbf{n}\ominus(z,h)\oplus(z,S^a_l(h)),q)|(\mathbf{n},q)) \propto \frac{\widehat{\pi}^{\Lambda,\mathrm{K}}(S^a_l(h)|\mathbf{h}(\mathbf{n}\ominus(z,h)))}{\widehat{\pi}^{\Lambda,\mathrm{K}}(h|\mathbf{h}(\mathbf{n}\ominus(z,h)))}\frac{\theta_l}{L^2+\theta n}M^{(l)}_{ah[l]},$$

$$\widehat{\mathbf{P}}^{\mathrm{SL}}_{n,\mathbf{x}}((\mathbf{n},q+1)|(\mathbf{n},q)) \propto \frac{L^2}{L^2+\theta n}(1-u)^{|N_{\mathbf{n}}(x_q)|},$$

$$\widehat{\mathbf{P}}^{\mathrm{SL}}_{n,\mathbf{x}}((\mathbf{n}\ominus J\oplus(z,h),q+1)|(\mathbf{n},q)) \propto \frac{u^{|J|}(1-u)^{|N_{\mathbf{n}}(x_q)\ominus J|}}{\widehat{\pi}^{\mathrm{K}}(\underbrace{h,\dots,h}_{|J|-1}|\mathbf{h}(N^{x_k}_{\mathbf{n}}\ominus J\oplus(z,h)))}\frac{L^2}{L^2+\theta n}\frac{1}{\pi r^2}.$$

Efficient algorithms for simulating samples from the spatial Λ-coalescent have been developed by Kelleher et al. [2013, 2014] and were used to simulate a sample of 100 lineages uniformly distributed within a torus of side length $L = 10$. The type space consists of 10 binary loci, with mutations flipping a randomly chosen locus at rate $\theta = 10^{-3}$. The resulting sample is depicted in Figure 2.6.

Natural parameters of the model are the mutation rate θ, the radius r and the impact u. The latter two can be related to classical population genetics parameters: Wright's neighbourhood size is $\mathcal{N} = \frac{1}{u}$, and the variance per unit time of the spatial location of a lineage traced backwards in time is $\sigma^2 = \frac{\pi u r^4}{2}$ [Barton et al., 2013a]. I consider inferring all three parameters separately, assuming that the other two are known. Likelihood surfaces for all three parameters are shown in Figure 2.7. Stopping-time resampling has been used in all simulation runs, with trajectories being stopped whenever the sample size first hits or falls below $\{99, 98, \dots, 3, 2\}$ and resampling performed if the effective sample size is lower than one half of the number of particles. Note that because of multiple mergers it is possible for a trajectory to hit multiple stopping times at once, in which case it will remain stopped until all particles reach the same level provided it survives all intermediate resampling steps.

Figure 2.7 demonstrates that reverse time SMC can produce convergent estimators for very small likelihoods, even if the computational run times are daunting. The mutation rate θ can be identified to within an order of magnitude, and the

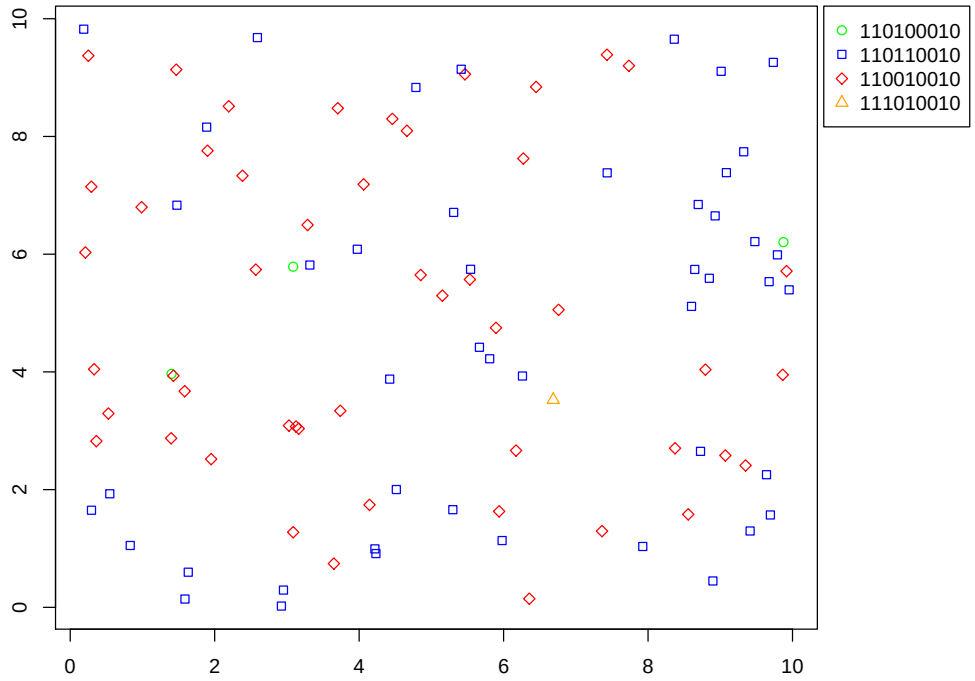


Figure 2.6: Sampling locations and observed types of the simulated observation from a spatial Λ-coalescent. The model parameters are $L = 10, \theta = 10^{-3}, r = 1.0$ and $u = 0.3$. The type space consists of binary vectors of length 10, with mutations flipping a randomly chosen element.

impact u to within a factor of 2. Radii which are too small can also be ruled out clearly, though large radii cannot be excluded similarly. This is most likely due to the relatively small ratio of the torus side length L to the event radius r, with larger tori resulting in greater radius identifiability. Simulations by Guindon et al. [2016] reached similar conclusions, with accurately estimated neighbourhood sizes $\mathcal{N} = 1/u$ and higher uncertainty estimates of dispersal rate $\sigma^2 = \pi u r^4/2$, at comparable computational cost.

As with SMC algorithms in general these can be greatly reduced by parallelising the algorithm, which is typically trivial, or by reducing the number of independent simulations through use of driving values [Griffiths and Tavaré, 1994c] or bridge sampling [Meng and Wong, 1996], at least in the case of a fixed radius r. Mismatches in radius cause mutual singularity of the proposal and target distribu-

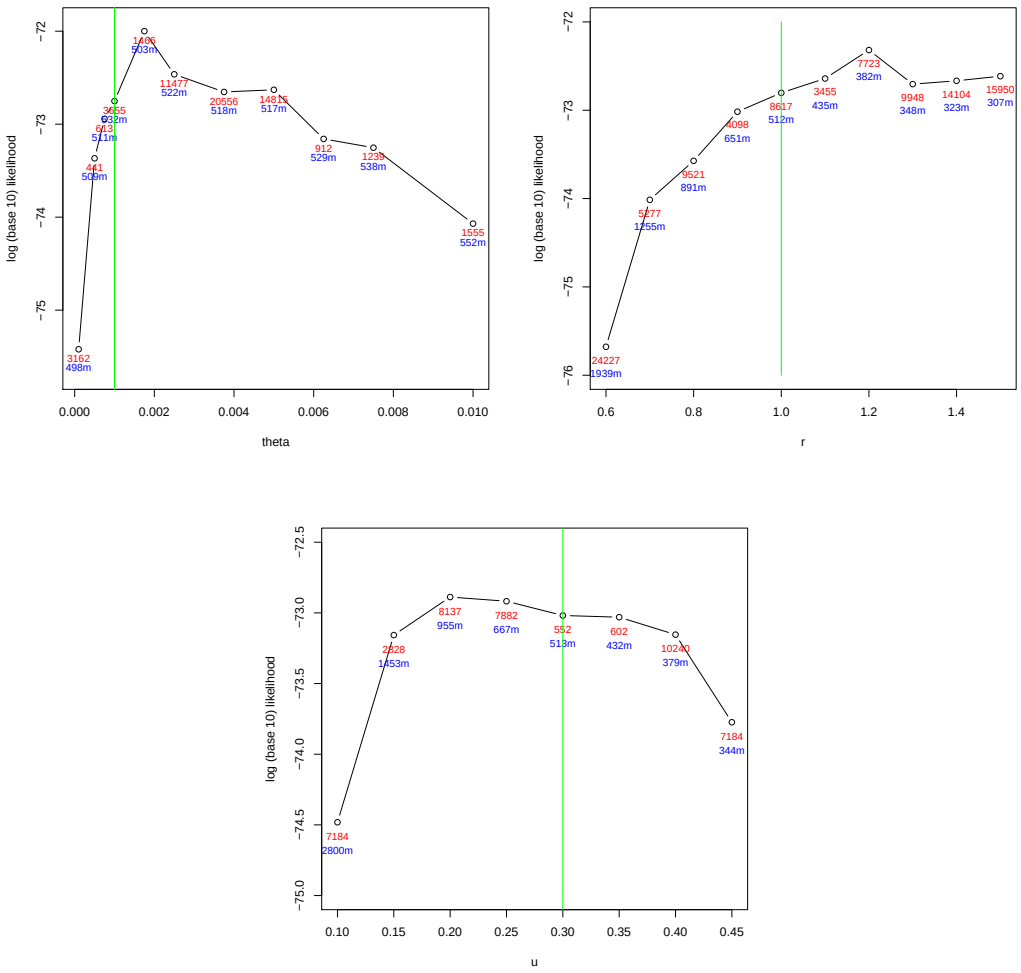


Figure 2.7: Simulated likelihood surfaces for each of the three parameters of the spatial Λ-coalescent, assuming the other two parameters are known and using 4 million particles. Points correspond to independent simulations, and are labelled with their effective sample sizes and run times using 12 cores on the MidPlus cluster Minerva. Some noise is still clearly present in the surfaces, but their general shapes are identifiable.

tions by allowing mergers in the model with larger radius which are not possible in the smaller radius. However, the mutation rate θ and impact u can be varied freely, provided both are strictly positive.

2.6 Other examples of reverse time SMC

As outlined in Section 2.2, time reversal is not a tool that is exclusive to coalescent processes, although it is most well studied in this context. Indeed, use of time reversal and approximate CSDs to write down proposal distributions is a completely general method. In this section I demonstrate this fact by deriving families of approximate CSDs for a diffusion process, a queueing model, and an epidemic model on a network, and present accompanying simulation studies for estimating intractable hitting probabilities.

2.6.1 Containment probabilities of a hyperbolic diffusion

The one-dimensional hyperbolic diffusion is the solution of the SDE

$$dX_t = \frac{-X_t}{\sqrt{1+X_t^2}}dt + dW_t, \tag{2.23}$$

where $(W_t)_{t\geq 0}$ is a Brownian motion. It was introduced by Barndorff-Nielsen [1978] in connection to hyperbolic distributions in geostatistical modelling [Barndorff-Nielsen, 1977], and its heavier-than-Gaussian tails have also made it a popular model in mathematical finance [Bibby and Sørensen, 2003].

The transition probabilities of the diffusion are intractable, but the stationary distribution is known to be the hyperbolic distribution

$$\pi(x) = \frac{1}{2K_1(1)}e^{-\sqrt{1+x^2}}, \tag{2.24}$$

where K_1 is the modified Bessel function of the second kind. I assume that the diffusion is started at stationarity, and focus on the probability that a trajectory lies in an interval (l_0, u_0) at time 0, and hits interval (l_t, u_t) at time $t \in \mathbb{N}$, without leaving the strip obtained by connecting l_0 to l_t and u_0 to u_t with straight lines at intermediate times. Similar containment probabilities have been studied e.g. in [Casella and Roberts, 2008] in the context of double barrier option pricing. The sets defining the event of interest are

$$\begin{aligned} I &= \{0\} \times (l_0, u_0), \\ T &= \left(\bigcup_{s\in(0,t)} \{s\} \times \left\{ \frac{l_t - l_0}{t}s + l_0, \frac{u_t - u_0}{t}s + u_0 \right\} \right) \cup (\{t\} \times (l_t, u_t)), \end{aligned}$$

the initial distribution is

$$\mu(\cdot) = \pi(\cdot | l_0 < \cdot < u_0),$$

and the quantity of interest is

$$\mathbb{E}_\mu[f(\tau_T, X_{0:\tau_T})] = \mathbb{E}_\mu[\mathbb{1}_{\{t\}}(\tau_T)],$$

i.e. the probability that the diffusion remains contained within the strip, and only hits the set T at the end at time t. I will consider a discretisation of (2.23), and use the Euler scheme with grid spacing $\delta > 0$ to define a family of approximate transition densities forwards in time:

$$P_\delta(x, y) = P_\delta((m, x), (n, y)) = \frac{\mathbb{1}_{\{m+\delta\}}(n)}{\sqrt{2\pi\delta}} \exp\left(-\frac{1}{2\delta}\left[y - x\left\{1 - \frac{\delta}{\sqrt{1+x^2}}\right\}\right]^2\right). \tag{2.25}$$

Note that in this case (2.25) is also the Milstein scheme because of the unit diffusion coefficient. The discretised transition density (2.25) and the unconditional stationary distribution (2.24) can be used to define a discretised reverse time proposal:

$$\widehat{P}_\delta(y, x) = \widehat{P}_\delta((n, y), (m, x)) \propto \frac{\pi(x)}{\pi(y)} P_\delta(x, y) \mathbb{1}_{\left\{\left(\frac{l_t - l_0}{t} m + l_0, \frac{u_t - u_0}{t} m + u_0\right)\right\}}(y).$$

I assume for simplicity that Δ divides t exactly, and consider the analogous discretisation of the target set T. I have also neglected the issue of bias due to unobserved boundary crossings between time discretisation points, though more sophisticated interpolation schemes [Gobet, 2000] could also be implemented.

This family of proposal distributions can be normalised numerically, and sampled by proposing $x\left(1 - \frac{\delta}{\sqrt{1+x^2}}\right)$ from a $\mathcal{N}(y, \delta)$ proposal distribution, solving for x and accepting the proposal with probability $e^{-\sqrt{1+x^2}}$. In the above definition, Step 2 of the strategy outlined in Section 2.2 has been implemented by automatically rejecting proposed values outside the permitted strip. Dynamic resampling, in which particles are resampled whenever their effective sample size falls below half the particle number, was also employed.

Figure 2.8 presents estimated probabilities of excursions containment as a function of the height of the end interval. The resulting effective sample size is not monotonically decreasing in the rarity of the terminal interval due to the fact that rarer intervals push the reverse time dynamics to the mode more rapidly. The increase in run time along the x-axis in Figure 2.8 is due to the exponential decay of the acceptance probability in the rejection sampler used to generate proposals. A

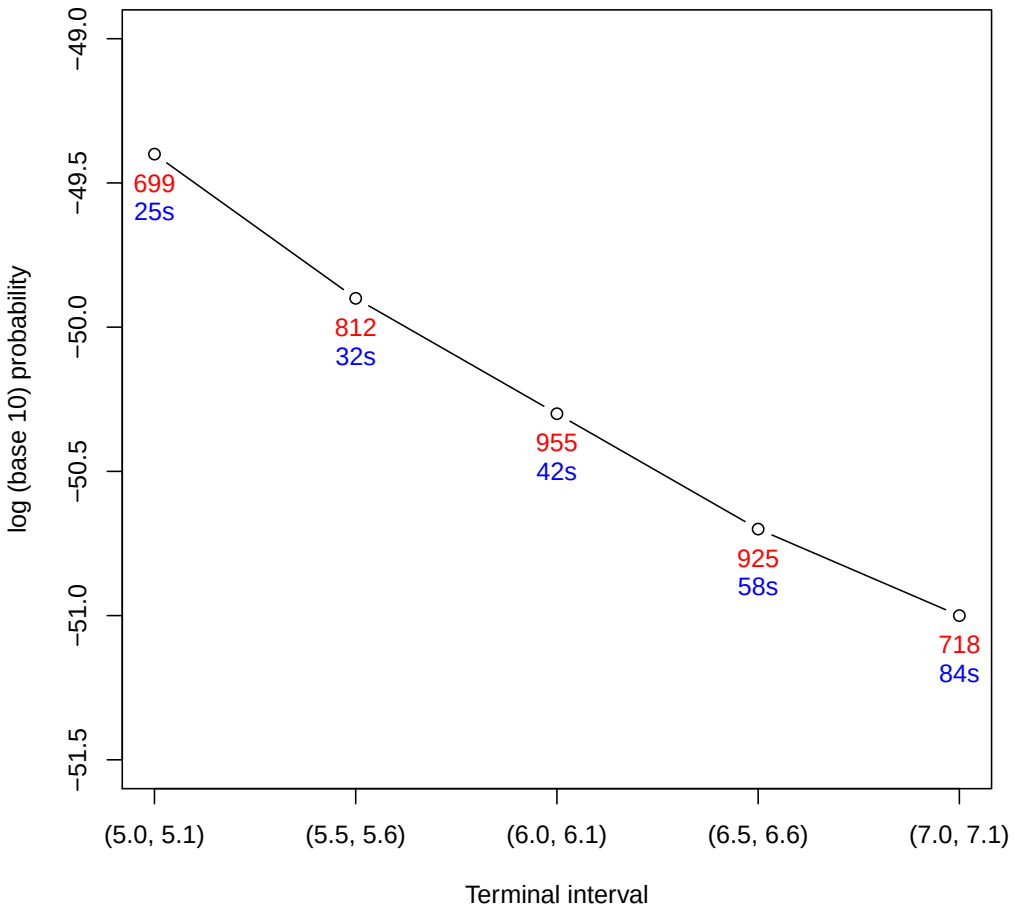


Figure 2.8: Simulated containment probabilities of the hyperbolic diffusion with initial interval $(l_0, u_0) = (-1, 1)$, trajectory length $t = 10$, time discretisation $\Delta = 0.01$, $N = 1000$ particles, and terminal window (l_t, u_t) given on the x-axis. Each experiment corresponds to an independent simulation, and is labelled with a run time on an Intel i5-2520M 2.5 GHz processor, and the effective sample size.

more uniformly efficient proposal sampler would result in run times which are more or less independent of the height, and thus the rarity, of the terminal condition as well.

2.6.2 Hitting probabilities of ATM queueing networks

The second example is the ATM (asynchronous transfer mode) network studied by Glasserman et al. [1999] in the context of rare events. The network consists of d sources, each of which is either on or off. Sources which are off do nothing, while sources which are on produce packets at rate λ. Packets are serviced by a common server with rate μ using the first-in-first-out policy. Off sources turn on at rate α_0 and on sources turn off at rate α_1. The state of the system is specified as $(i, j) \in \mathbb{N}_0 \times [d]$, where i denotes the number of packets in the queue and j the number of on sources.

Glasserman et al. [1999] estimated the probability of the queue length hitting a barrier $b \in \mathbb{N}$ before emptying, given an empty initial queue and $d\alpha_0/(\alpha_0 + \alpha_1)$ on sources. Reverse-time SMC could be used for this example by summing over all possible numbers of terminal open sources, but this results in a d-fold increase in computational burden and hence cannot be expected to be competitive with a forwards-in-time approach. I focus instead on the joint probability of an initially empty queue hitting a barrier b before emptying with exactly k sources open at the hitting time, and assume the initial number of open sources is $\mathrm{Bin}(d, \alpha_0/(\alpha_0 + \alpha_1))$-distributed. In this scenario a forwards-in-time algorithm would face the same difficulties as a reverse-time algorithm does in the scenario of Glasserman et al. [1999].

The sets which define the event of interest are

$$I = \bigcup_{j=0}^{d} \{(0, j)\},$$
$$T = \bigcup_{j=0}^{d} \{(0, j)\} \cup \{(b, j)\},$$

the quantity of interest is the hitting probability

$$\mathbb{E}_\mu[f(\tau_T, X_{0:\tau_T})] = \mathbb{E}_\mu[\mathbb{1}_{\{(b,k)\}}(X_{\tau_T})],$$

and the initial law is

$$\mu(\{(0, j)\}) = \binom{d}{j} \left(\frac{\alpha_0}{\alpha_0 + \alpha_1}\right)^j \left(\frac{\alpha_1}{\alpha_0 + \alpha_1}\right)^{d-j}.$$

To define the proposal distribution it is only necessary to specify approximate conditional distributions of i given j and j given i. These are denoted by $\widehat{\pi}_i(i|j)$ and

$\widehat{\pi}_j(j|i)$ respectively, and chosen to be

$$\widehat{\pi}_i(i|j) \propto \left(\frac{\lambda j}{\mu}\right)^i \qquad \text{for } i \in [b] \text{ and } j \in [d],$$

$$\widehat{\pi}_j(j|i) \propto \widehat{\pi}_i(i|j)\binom{d}{j}\left(\frac{\alpha_0}{\alpha_0+\alpha_1}\right)^j\left(\frac{\alpha_1}{\alpha_0+\alpha_1}\right)^{d-j} \qquad \text{for } i \in [b] \text{ and } j \in [d].$$

The former is the true distribution of a queue with arrival rate λj and service rate μ whenever $\mu > \lambda j$, and well-defined otherwise as well because the range of possible values of i is finite. The latter is obtained from the former via Bayes' rule. These probabilities also implicitly define the rule for choosing which coordinate to update: both $\widehat{\pi}_i(i|j)$ and $\widehat{\pi}_j(j|i)$ are evaluated for all moves allowed by the current state of the system, and a move is sampled proportional to the resulting probabilities.

I also employ stopping time resampling again, with simulations stopped every time a new minimum queue length is reached in reverse time. Once all simulations have been stopped, resampling takes place if the effective sample size is below half of the number of particles.

Figure 2.9 presents simulated hitting probabilities of a queue length of 30 across all possible fixed numbers of terminal on sources. Despite some residual noise the shape and magnitude of the surface can be distinguished clearly, and the effective sample size shows at most weak decay as the estimated probability decreases. This is because increased problem difficulty (as measured by the rarity of the event of interest) is compensated for automatically by stronger drift towards the mode by the reverse-time dynamics.

2.6.3 Initial infection in a susceptible-infected-susceptible network

Consider a finite network with vertices V and undirected edges E, and with vertices labelled as either susceptible (S) or infected (J). For a vertex $v \in V$, let $l(v) \in \{J, S\}$ denote its label, $N_v := \{v' \in V : (v, v') \in E\}$ denote its neighbourhood and, for $a \in \{J, S\}$, let

$$N_v^a := \{v' \in V : (v, v') \in E \text{ and } l(v') = a\}$$

denote the sub-neighbourhood with label a. Then the susceptible-infected-susceptible (SIS) epidemic evolves as follows.

Every infected node is cured with rate $\beta > 0$, at which point it immediately becomes susceptible again. A susceptible node is infected by each infected neighbour at rate $\alpha > 0$, so that a vertex $v \in V$ becomes infected at total rate $\alpha|N_v^J|$. These types of dynamics on networks are popular models e.g. for the spread of biological

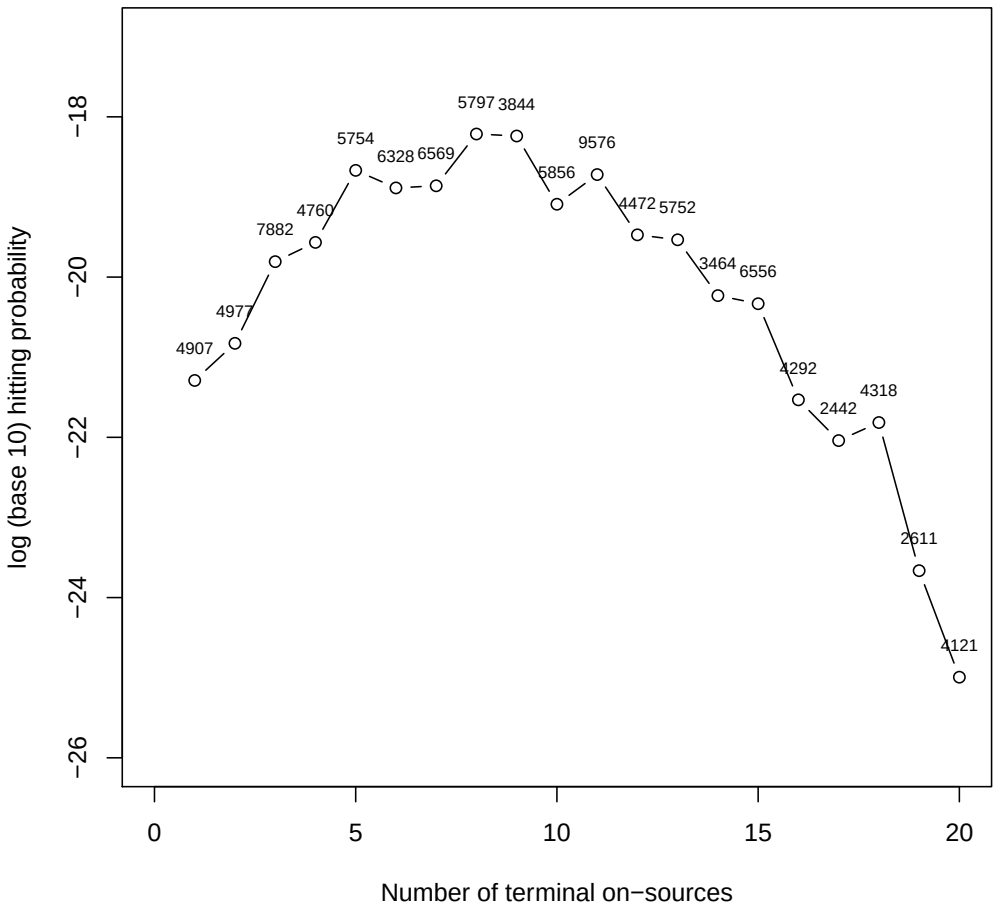


Figure 2.9: Simulated hitting probabilities of an ATM network with parameters $d = 20$, $b = 30$, $\lambda = 0.5$, $\mu = 10.0$, $\alpha_0 = 1.0$, $\alpha_1 = 3.0$. An independent simulation of 500 000 particles was run for each value of k for a runtime of around 20 minutes for each k on an Intel i5-2520M 2.5 GHz processor. Estimates are labelled with their corresponding effective sample sizes.

epidemics in structured populations [Moore and Newman, 2000; Pastor-Satorras and Vespignani, 2001; Ganesh et al., 2015], malware in computer networks [Shah and Zaman, 2010], and rumours in social networks [Fuchs and Yu, 2015; Shah and Zaman, 2016], and are also sometimes referred to as contact processes. In addition, let $\gamma > 0$ be the rate at which a new infection enters the network, infecting one uniformly sampled node. Such new infections are assumed to only enter when all

vertices are susceptible, i.e. only one infection can exist in the population at one time.

Suppose that there is no infection in the initial population, and that small infections go undetected. An infection is defined as large once it infects at least $\lfloor |V|/10 \rfloor$ nodes. Assume that the labels of all nodes are immediately observed as soon as a large infection arises. Infection times are not observed, nor is any information about the history of the infection, such as whether a vertex that is now susceptible was previously infected. The object of interest is inferring the initial location of the observed large infection, which may no longer be infected itself. Point estimators for similar inference problems have been studied in [Shah and Zaman, 2010; Fuchs and Yu, 2015; Shah and Zaman, 2016]. Suppose that $\beta \gg \alpha$, so that the epidemic is subcritical and large infections are rare. Corresponding inference for supercritical infections falls outside the scope of reverse-time SMC as outlined in this thesis, because the initial state of no infection is rare while the target state of a large infection is typical.

More formally, consider the jump skeleton of the above continuous time Markov process, let $l_t(v)$ denote the label of vertex $v \in V$ at time $t \in \mathbb{N}_0$, and let the Markov chain $\{X_t\}_{t=0}^{\tau_T}$ be given as

$$X_t = \{v \in V : l_t(v) = J\},$$

i.e. the set of infected vertices at time t. Then the initial condition I is the empty set, the initial distribution is $\mu(\emptyset) = 1$, the target set $T := \{X : |X| = \lfloor |V|/10 \rfloor\}$ is the set of observed epidemics that are sufficiently large to be detected, and the quantity of interest is the likelihood of the location of the initial infection given an observed infection $X_{\tau_T} = \mathbf{v}^*$:

$$\begin{aligned}\mathbb{E}_\mu[f(\tau_T, X_{0:\tau_T})] = |V|\mathbb{E}_\mu[\mathbb{1}_{\{\mathbf{v}^*\}}(X_{\tau_T})\mathbb{1}_{\{v\}}(X_1)] &= \frac{\mathbb{P}_\mu(X_1 = v, X_{\tau_T} = \mathbf{v}^*)}{\mathbb{P}_\mu(X_1 = v)} \\ &= \mathbb{P}_\mu(X_{\tau_T} = \mathbf{v}^* | X_1 = v).\end{aligned}$$

Note that approximating $\mathbb{E}_\mu[f(X_{0:\tau_T})]$ using a forwards-in-time algorithm is challenging because it can be difficult to know a priori which nodes are likely to be the one initially infected, and hence the algorithm may spend much effort sampling trajectories of low probability. The problem also lacks a natural reaction coordinate (2.3) because nodes can be uninfected and reinfected, which makes driving samples towards the observed configuration difficult. Neither of these problems causes any difficulty in reverse time: a reaction coordinate is not needed and sampled trajecto-

ries drift towards initial locations of high probability automatically.

It remains to specify the proposal distribution, which is done by specifying the CSD of the label of one vertex given the labels of all the others. Conditioned on the labels of all other vertices, vertex $v \in V$ becomes infected at fixed rate N_v^J and susceptible with rate β, so a natural choice of approximate CSD is

$$\hat{\pi}(l(v)|\{l(v')\}_{v' \neq v}) = \begin{cases} \frac{\alpha|N_v^J|+\varepsilon}{\alpha|N_v^J|+\beta+\varepsilon} & \text{if } l(v) = J \\ \frac{\beta}{\alpha|N_v^J|+\beta+\varepsilon} & \text{if } l(v) = S \end{cases},$$

where $\varepsilon > 0$ is a regularisation term correcting for the fact that isolated individuals can become infected in reverse time, corresponding to an infection spreading outwards and all connecting individuals becoming uninfected before the final, leaf one. An approximate CSD based on a larger neighbourhood size would yield a more accurate approximation at greater computational cost. Note that this formula also captures the final transition from a single infected to a fully susceptible graph, with probability proportional to

$$\mathbb{P}_\mu(X_1 = \{v\}) \frac{\hat{\pi}(S|\{S, \ldots, S\})}{\hat{\pi}(J|\{S, \ldots, S\})} = \frac{\beta}{|V|\varepsilon},$$

because the probability of a fully susceptible graph acquiring an infection at site $v \in V$ in the next time step is $|V|^{-1}$.

Figure 2.10 shows a estimated likelihood surface produced from SMC output for the initial infected location, along with a plot of the observed infection. The surface shows a high degree of monotonicity, and concentrates around the observed epidemic as expected.

2.7 Discussion

In this chapter I have presented a general framework for designing SMC proposal distributions which proceed backwards in time. Time-reversal makes it straightforward to ensure realisations of the process hit desired regions of the state space, essentially irrespective of the probability assigned to them by the law of the process of interest. Even the extreme case of conditioning paths on a terminal point of probability 0 can be dealt with easily. This makes time-reversal a natural and efficient choice when the end point of a path is known with high accuracy, but its initial distribution is diffuse. As most existing rare event and path simulation algorithms make the opposite assumptions about initial and terminal conditions, time-reversal can

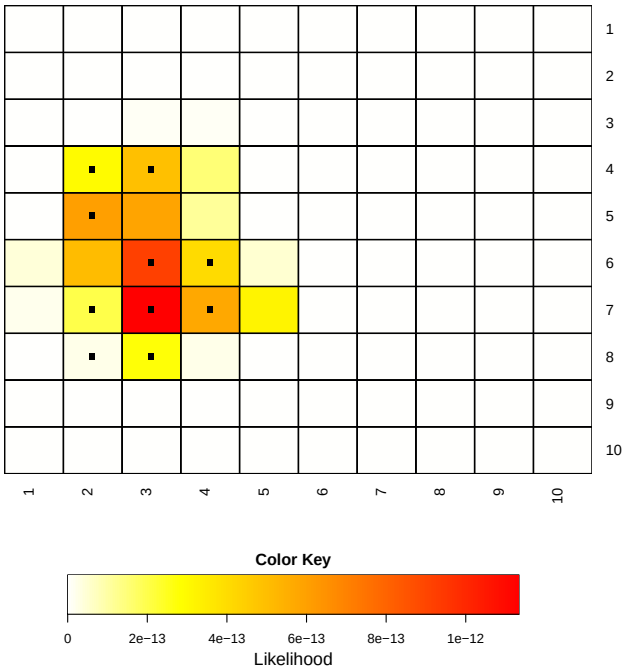


Figure 2.10: Simulated likelihood surface for the location of the initial infected on a 10 x 10 nearest neighbour network with $\alpha = 1$, $\beta = 12$, $\gamma = 1$, $\varepsilon = 10^{-8}$ and using 10 000 particles for a run time of 40 minutes on an Intel i5-2520M 2.5 GHz processor. The black dots denote the observed infection, and the true initial location is row 7, column 2.

be expected to be a useful tool in extending the scope of simulation-based inference and computation.

Expressing the law of the reverse-time process via Nagasawa's formula (2.4) often leads to a substantial reduction in the dimensionality of the design task of defining a proposal distribution via Proposition 2. The difficulty of designing efficient proposal distributions in high dimension is a central barrier to practical SMC, so this cancellation of dimensions is an important advantage. Furthermore, I want to emphasize that it is not inherently linked to time-reversal: re-weighting jump probabilities by an approximate stationary distribution and cancelling out common co-ordinates would lead to a forwards-in-time proposal defined by low dimensional approximate CSDs. For rare terminal conditions a reverse-time approach is easier because the conditional and unconditioned stationary distributions share the qualitative behaviour of rapidly leaving the rare state for a stationary mode. A forwards-in-time algorithm would have to use CSDs approximating the behaviour of

an appropriate Doob's h-transform in order to drive the process away from modes and into the rare state. Nevertheless, analogues of Proposition 2 can be a useful design tool beyond the scope of this thesis.

All example simulations considered in this chapter have had the property that the proposal distribution could be normalised numerically, so that proposals could be sampled via standard methods. This property is computationally convenient, but is often not necessary since importance weights typically only need to be evaluated up to a normalising constant. In such cases samples from unnormalised proposals can be generated via Metropolis-Hastings and the only modification to Algorithm 1 is a step to self-normalise weights after line 28.

Section 2.6 presents examples, each of which would be inefficient to solve using forwards in time methods. For the diffusion containment problem in Section 2.6.1, a forwards in time proposal distribution would have to mimic an intractable h-transform. For the ATM queue in Section 2.6.2 a forwards in time method would have to explicitly average over initial conditions. Finally, the SIS network model in Section 2.6.3 lacks a natural reaction coordinate (2.3), so that it is difficult to design proposal distributions which drive an empty network towards the observed configuration efficiently. All three problems are tackled easily by the reverse time approach.

In Section 2.3 I presented reverse-time SMC 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 De Iorio and Griffiths [2004a], De Iorio and Griffiths [2004b], Paul and Song [2010]. Furthermore, they outperform existing algorithms for Λ-coalescent inference. 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, and through paralellisation, which can be done very effectively as is typical for SMC algorithms. In particular, use of GPUs for parallel Monte Carlo simulations has been found to speed up computations by up to 500 fold in comparison to serial simulation [Lee et al., 2010].

The limits on coalescent data sets that can be feasibly analysed using SMC 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 simulations in Section 2.4 suggest it can also a viable approach for Λ-coalescents. Much work has been done on sophisticated approximations to CSDs for Kingman's coalescent with recombination and other features, and results in Section 2.4 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 coalescents with generators that differ from those studied in this thesis only by additive terms, so the machinery used here can be applied more generally with little added difficulty.

Section 2.5 demonstrates that reverse-time SMC can handle very high dimensional missing data and conditioning on events of extremely small probability, albeit at high computational cost. The algorithm introduced in Section 2.5 is also the first published inference algorithm for the spatial Λ-coalescent, which solves a number of long-standing problems in spatially structured population genetics. It was made possible by a recursion of the form (2.20) in combination with the reverse-time framework, and the cancellation of dimensions outlined in Proposition 2. As such, it provides a concrete example of a model of practical interest which is rendered computationally tractable by this method.

Chapter 3

Bayesian nonparametric inference

3.1 Introduction

Bayesian nonparametric statistics differs from classical, parametric inference in only one way: parametric inference is concerned with learning a finite number of parameters from data, while nonparametric statistics allows the parameter space to be infinite dimensional. This relaxation enables learning function- or measure-valued quantities of interest from observations, without the restrictive assumption of a parametric family of candidates. However, this flexibility often comes at the cost of considerable analytic and computational complexity.

Recall that the Bayesian inference procedure is to specify a prior Q for a quantity of interest b, and make inferences from the posterior given observed data $x_{1:n}$ as

$$Q(db|x_{1:n}) = \frac{\mathbb{P}(x_{1:n}|b)Q(db)}{\int \mathbb{P}(x_{1:n}|b)Q(db)}.$$

To fix intuition, consider for example the problem of inferring the drift coefficient of the scalar SDE

$$dX_t = b(X_t)dt + dW_t$$

from discretely sampled observations from trajectories at stationarity. The drift function b must satisfy regularity conditions for the model to be well defined and possess a unique stationary distribution to sample, but such drifts are still function-valued parameters which cannot be captured in full generality by finitely many parameters. The first technical difficulty brought about by the infinite dimensional setting is that existence and uniqueness of the posterior are not guaranteed. More-

over, the highly desirable property of posterior consistency, $Q(U_{b_0}^c|x_{1:n}) \to 0$ as $n \to \infty$, where U_{b_0} is an open neighbourhood of the data generating parameter b_0, is non-trivial and depends in subtle ways on details of the topology, mode of convergence and the prior [Diaconis and Freedman, 1986].

This section investigates Bayesian posterior consistency for the more general, d-dimensional jump diffusion models

$$d\mathbf{X}_t = b(\mathbf{X}_t)dt + d\mathbf{W}_t + c(\mathbf{X}_{t-}, d\mathbf{Z}_t),$$

introduced in Section 1.2.3 (see (1.8)), as well as the genetically motivated Λ-Fleming-Viot processes which are specified via their generators,

$$\begin{aligned}\mathcal{G}^\Lambda f(\mathbf{x}) = \theta \sum_{i,j=1}^{d} x_j(M_{ji} - \delta_{ij}) f_i(\mathbf{x}) + \frac{\Lambda(\{\mathbf{0}\})}{2} \sum_{i,j=1}^{d} x_i(\delta_{ij} - x_j) f_{ij}(\mathbf{x}) \\ + \sum_{i=1}^{d} \int_{(0,1]} x_i\{f((1-r)\mathbf{x} + r\mathbf{e}_i) - f(\mathbf{x})\} r^{-2} \Lambda(dr),\end{aligned}$$

as in Section 1.2.1 (see (1.6)). I will give criteria for joint inference of the drift and jump components in the jump diffusion case, and for inferring Λ assuming θ and M are known in the Λ-Fleming-Viot case. With the exception of an identifiability condition, which seems difficult to check in general, all consistency criteria are tractable and can be verified. I will also provide a parametric algorithm for practical nonparametric inference for Λ-Fleming-Viot processes. This parametric approach is similar to the method of *likelihood-informed subspaces* for accelerating MCMC inference in the case of Gaussian measures [Cui et al., 2014], with the further advantage that the parametric method presented here captures the likelihood fully resulting in no truncation or approximation error.

3.2 Jump diffusions

As outlined above, Bayesian nonparametric consistency is highly sensitive to details of the prior, topology and mode of convergence. Hence it is important to specify these details before moving on to discuss consistency.

For Borel sets $A \in \mathcal{B}(\mathbb{R}_0^d)$, let $c^*(\mathbf{x}, \cdot)$ denote the pull-back of the jump coefficient $c(\mathbf{x}, \cdot)$:

$$c^*(\mathbf{x}, A) := \{\mathbf{z} \in \mathbb{R}_0^d : c(\mathbf{x}, \mathbf{z}) \in A\}.$$

Definition 5. Let $\Theta = \{(b, \nu)|b : \Omega \mapsto \mathbb{R}^d, \nu : \Omega \times \mathbb{R}_0^d \mapsto \mathbb{R}_+\}$ denote a set of

pairs of drift functions $b(\mathbf{x})$ and Lévy measures $\nu(\mathbf{x}, d\mathbf{z}) = M(c^*(\mathbf{x}, d\mathbf{z}))$ with each pair satisfying the hypotheses of Proposition 1 with uniformly bounded Lipschitz constant C_1 in (1.9). Furthermore, suppose that for each $\mathbf{x} \in \Omega$ and any pair of Lévy measures $(\cdot, \nu), (\cdot, \nu') \in \Theta$, the measures $\nu(\mathbf{x}, \cdot) \sim \nu'(\mathbf{x}, \cdot)$ are equivalent with Radon-Nikodym density satisfying

$$0 < \inf_{\mathbf{x} \in \Omega, \mathbf{z} \in \mathbb{R}_0^d} \left\{ \frac{d\nu'}{d\nu} \right\} \leq \sup_{\mathbf{x} \in \Omega, \mathbf{z} \in \mathbb{R}_0^d} \left\{ \frac{d\nu'}{d\nu} \right\} < \infty,$$

and that either

1. all Lévy measures supported by Θ are finite, or
2. there exists an open set A containing the origin such that $\nu(\mathbf{x}, \cdot)|_A = \nu'(\mathbf{x}, \cdot)|_A$ uniformly in $(\cdot, \nu) \neq (\cdot, \nu') \in \Theta$.

Remark 7. The conditions of Definition 5 mean that the unit diffusion coefficient and the infinite intensity component of the Lévy measure can be thought of as known confounders of the joint inference problem for the drift function and the finite intensity compound Poisson component of the Lévy measure.

The following assumption ensures that the drift function and Lévy measure can be uniquely identified from discrete data.

Assumption 1. For any pair $(b, \nu) \neq (b', \nu') \in \Theta$ and any $\delta > 0$ there exists an $\mathbf{x} \in \Omega$ and a bounded, measurable function $f : \Omega \mapsto \mathbb{R}$ such that $P_\delta^{b,\nu} f(\mathbf{x}) \neq P_\delta^{b',\nu'} f(\mathbf{x})$. Note that both sides of the inequality are real numbers for a fixed $\mathbf{x}$. In particular, identifying $P_\delta^{b,\nu}$ is equivalent to identifying (b, ν).

Assumption 1 states that the mapping $(b, \nu) \mapsto P_\delta^{b,\nu}$ is injective, or that the same semigroup cannot arise from two different generators. If this were to happen, then it would be impossible to tell the corresponding drift and jump coefficients apart from discrete data, so that consistency necessarily fails. Transition densities of jump diffusions are typically intractable, which makes verifying Assumption 1 challenging in general. One approach which can sometimes be fruitful is to verifying that the mapping $(b, \nu) \mapsto \pi^{b,\nu}$ is injective, because

$$\lim_{k \mapsto \infty} (\underbrace{P_\delta^{b,\nu} \circ \ldots \circ P_\delta^{b,\nu}}_{k\text{-fold}}) f(\mathbf{x}) = \lim_{k \mapsto \infty} P_{k\delta}^{b,\nu} f(\mathbf{x}) = \int_\Omega f(\mathbf{y}) \pi^{b,\nu}(\mathbf{y}) d\mathbf{y}$$

by the semigroup property and ergodicity. Of course, the stationary densities of jump diffusions are also intractable in general. It seems likely that in a case where

identifiability fails but all other consistency criteria hold, the posterior will converge to be supported on the set of all pairs (b, ν) that give rise to the semigroup generating the data, with weights proportional to the prior densities of the pairs, at least subject to the set of these pairs being sufficiently regular. However, this conjecture has not been verified rigorously.

The topology under consideration is defined as in [van der Meulen and van Zanten, 2013; Gugushvili and Spreij, 2014] by specifying a subbase determined by the semigroups $P_t^{b,\nu}$. For details about the notion of a subbase, and other topological concepts, see e.g. [Dudley, 2002].

Definition 6. Fix a sampling interval $\delta > 0$ and a finite measure $\rho \in \mathcal{M}_f(\Omega)$ with positive mass in all non-empty, open sets. For any $(b, \nu) \in \Theta$, $\varepsilon > 0$ and $f \in C_b(\Omega)$ define the set

$$U_{f,\varepsilon}^{b,\nu} := \{(b', \nu') \in \Theta : \|P_\delta^{(b',\nu')} f - P_\delta^{b,\nu} f\|_{1,\rho} < \varepsilon\},$$

where $\|\cdot\|_{1,\rho}$ is the $L^1(\rho)$-norm. A weak topology on Θ is generated by requiring that the family $\{U_{f,\varepsilon}^{b,\nu} : f \in C_b(\Omega), \varepsilon > 0, (b, \nu) \in \Theta\}$ is a subbase of the topology.

The following lemma is a direct analogue of Lemma 3.2 of [van der Meulen and van Zanten, 2013]:

Lemma 1. *The topology generated by a subbase of sets of the form $U_{f,\varepsilon}^{b,\nu}$ is Hausdorff.*

Proof. Consider $(b, \nu) \neq (b', \nu') \in \Theta$. By Assumption 1 there exists $f \in C(\Omega)$ and $\mathbf{x} \in \Omega$ such that $P_\delta^{b,\nu} f(\mathbf{x}) \neq P_\delta^{b',\nu'} f(\mathbf{x})$, and hence by continuity a non-empty, open set $J \subset \Omega$ where $P_\delta^{b,\nu} f$ and $P_\delta^{b',\nu'} f$ differ. Hence $\|P_\delta^{b,\nu} f - P_\delta^{b',\nu'} f\|_{1,\rho} > \varepsilon$ for some $\varepsilon > 0$ so that the neighbourhoods $U_{f,\varepsilon/2}^{b,\nu}$ and $U_{f,\varepsilon/2}^{b',\nu'}$ are disjoint. □

I am now in a position to formally define posterior consistency.

Definition 7. Let $\mathbf{x}_{0:n} := (\mathbf{x}_0, \ldots, \mathbf{x}_n)$ denote $n + 1$ samples observed at times $0, \delta, \ldots, \delta n$ from $\mathbf{X}$ at stationarity, i.e. with initial distribution $\mathbf{X}_0 \sim \pi^{b,\nu}$. Weak posterior consistency holds if $Q(U_{b_0,\nu_0}^c | \mathbf{x}_{0:n}) \to 0$ with $\mathbb{P}^{b_0,\nu_0}$-probability 1 as $n \to \infty$, where U_{b_0,ν_0} is any open neighbourhood of $(b_0, \nu_0) \in \Theta$.

3.2.1 Posterior consistency

This section contains the statement and proof of the posterior consistency result for jump diffusions.

Theorem 5. *Let $\mathbf{x}_{0:n}$ be as in Definition 7, and suppose that the prior Q is supported on a set Θ which satisfies the conditions of Definition 5, as well as Assumption 1. If*

$$Q\Bigg((b,\nu)\in\Theta:\\ \frac{1}{2}\left(\|b_0-b\|_{2,\pi^{b_0,\nu_0}}+\left\|\int_{\mathbb{R}_0^d}\left[\frac{d\nu_0}{d\nu}(\cdot,\mathbf{z})-1\right]\mathbb{1}_{(0,1]}(\|\mathbf{z}\|_2)\mathbf{z}\nu(\cdot,d\mathbf{z})\right\|_{2,\pi^{b_0,\nu_0}}\right)^2\\ +\left\|\int_{\mathbb{R}_0^d}\left[\log\left(\frac{d\nu_0}{d\nu}(\cdot,\mathbf{z})\right)-\frac{d\nu_0}{d\nu}(\cdot,\mathbf{z})+1\right]\nu_0(\cdot,d\mathbf{z})\right\|_{1,\pi^{b_0,\nu_0}}<\varepsilon\Bigg)>0 \qquad (3.1)$$

for any $\varepsilon>0$ and any $(b_0,\nu_0)\in\Theta$ (which are thought of as the true parameters generating the data), then weak posterior consistency holds for Q on Θ.

Proof. The proof of Theorem 1 is a generalisation of the proof of Theorem 3.5 of van der Meulen and van Zanten [2013] and Theorem 1 of Gugushvili and Spreij [2014]. For $(b,\nu)\in\Theta$ let $\mathrm{KL}(b_0,\nu_0;b,\nu)$ denote the Kullback-Leibler divergence between $p_\delta^{b_0,\nu_0}$ and $p_\delta^{b,\nu}$:

$$\mathrm{KL}(b_0,\nu_0;b,\nu):=\int_\Omega\int_\Omega\log\left(\frac{p_\delta^{b_0,\nu_0}(\mathbf{x},\mathbf{y})}{p_\delta^{b,\nu}(\mathbf{x},\mathbf{y})}\right)p_\delta^{b_0,\nu_0}(\mathbf{x},\mathbf{y})\pi^{b_0,\nu_0}(\mathbf{x})d\mathbf{y}d\mathbf{x},$$

and for two probability measures P and P' on the same σ-field let $K(P,P'):=\mathbb{E}_P\left[\log\left(\frac{dP}{dP'}\right)\right]$. The law of a random object Z under a probability measure P is denoted by $\mathcal{L}(Z|P)$.

The following two properties are required:

1. A prior mass condition at the "truth": $Q((b,\nu)\in\Theta:\mathrm{KL}(b_0,\nu_0;b,\nu)<\varepsilon)>0$ for any $\varepsilon>0$.

2. Uniform equicontinuity of the semigroups $\{P_\delta^{b,\nu}f:(b,\nu)\in\Theta\}$ for $f\in\mathrm{Lip}(\Omega)$, the set of Lipschitz functions on Ω. The test functions employed in [van der Meulen and van Zanten, 2013; Gugushvili and Spreij, 2014] were $f\in C_b(\Omega)$, but by the Portemanteau theorem these families both determine weak convergence so there is no discrepancy.

These two properties will be established in Lemmas 2 and 3 below, which are the necessary generalisations of Lemmas 5.1 and A.1 of [van der Meulen and van Zanten, 2013], respectively.

Lemma 2. *Condition* (3.1) *implies that* $Q((b,\nu)\in\Theta : \mathrm{KL}(b_0,\nu_0;b,\nu)<\varepsilon)>0$ *for any* $\varepsilon>0$.

Proof. As in Lemma 5.1 of [van der Meulen and van Zanten, 2013] it will be sufficient to bound $\mathrm{KL}(b_0,\nu_0;b,\nu)$ from above by a constant multiple of

$$\frac{1}{2}\left(\|b_0-b\|_{2,\pi^{b_0,\nu_0}}+\left\|\int_{\mathbb{R}_0^d}\left[\frac{d\nu_0}{d\nu}(\cdot,\mathbf{z})-1\right]\mathbb{1}_{(0,1]}(\|\mathbf{z}\|_2)\mathbf{z}\nu(\cdot,d\mathbf{z})\right\|_{2,\pi^{b_0,\nu_0}}\right)^2$$
$$+\left\|\int_{\mathbb{R}_0^d}\left[\log\left(\frac{d\nu_0}{d\nu}(\cdot,\mathbf{z})\right)-\frac{d\nu_0}{d\nu}(\cdot,\mathbf{z})+1\right]\nu_0(\cdot,d\mathbf{z})\right\|_{1,\pi^{b_0,\nu_0}}.$$

A formal calculation yields

$$\begin{aligned}
&\int_\Omega\int_\Omega\log\left(\frac{\pi^{b_0,\nu_0}(\mathbf{x})p_\delta^{b_0,\nu_0}(\mathbf{x},\mathbf{y})}{\pi^{b,\nu}(\mathbf{x})p_\delta^{b,\nu}(\mathbf{x},\mathbf{y})}\right)p_\delta^{b_0,\nu_0}(\mathbf{x},\mathbf{y})\pi^{b_0,\nu_0}(\mathbf{x})d\mathbf{y}d\mathbf{x}\\
&\quad=K(\pi^{b_0,\nu_0},\pi^{b,\nu})+\mathrm{KL}(b_0,\nu_0;b,\nu)=K(\mathcal{L}(\mathbf{X}_0,\mathbf{X}_\delta|\mathbb{P}^{b_0,\nu_0}),\mathcal{L}(\mathbf{X}_0,\mathbf{X}_\delta|\mathbb{P}^{b,\nu}))\\
&\quad\le K(\mathcal{L}((\mathbf{X}_t)_{t\in[0,\delta]}|\mathbb{P}^{b_0,\nu_0}),\mathcal{L}((\mathbf{X}_t)_{t\in[0,\delta]}|\mathbb{P}^{b,\nu}))\\
&\quad=K(\pi^{b_0,\nu_0},\pi^{b,\nu})+\mathbb{E}^{b_0,\nu_0}\left[\log\left(\frac{d\mathbb{P}^{b_0,\nu_0}_{\mathbf{X}_0}}{d\mathbb{P}^{b,\nu}_{\mathbf{X}_0}}((\mathbf{X}_t)_{t\in[0,\delta]})\right)\right] \qquad (3.2)
\end{aligned}$$

by the conditional version of Jensen's inequality.

The aim is to identify the Radon-Nikodym derivative using Theorem 2.4 of [Cheridito et al., 2005], the hypotheses of which will now be verified. The local boundedness assumptions of [Cheridito et al., 2005] follow from Lipschitz continuity (1.9). Moreover, let $\{\Omega_n\}_{n=1}^\infty$ denote a sequence of bounded, open subsets of Ω such that $\Omega_1\subset\Omega_2\subset\ldots$ and $\cup_{n\ge1}\Omega_n=\Omega$. Then Lipschitz continuity, and the assumed finiteness of the Radon-Nikodym derivatives in Definition 5 ensure that there exists a sequence of finite constants $\{K_N\}_{n=1}^\infty$ such that

$$\sup_{\mathbf{x}\in\Omega_n}\left\{\left\|b_0(\mathbf{x})-b(\mathbf{x})-\int_{\mathbb{R}_0^d}\left[\frac{d\nu_0}{d\nu}(\mathbf{x},\mathbf{z})-1\right]\mathbb{1}_{(0,1]}(\|\mathbf{z}\|_2)\mathbf{z}\nu(\mathbf{x},d\mathbf{z})\right\|_2\right\}<K_n \qquad (3.3)$$

$$\sup_{\mathbf{x}\in\Omega_n}\left\{\int_{\mathbb{R}_0^d}\left[\frac{d\nu_0}{d\nu}(\mathbf{x},\mathbf{z})\log\left(\frac{d\nu_0}{d\nu}(\mathbf{x},\mathbf{z})\right)-\frac{d\nu_0}{d\nu}(\mathbf{x},\mathbf{z})+1\right]\nu(\mathbf{x},d\mathbf{z})\right\}<K_n \qquad (3.4)$$

for each $(b,\nu)\in\Theta$ and each $n\in\mathbb{N}$. In particular, the conditions in Remark 2.5 of [Cheridito et al., 2005] are satisfied. Hence Theorem 2.4 of [Cheridito et al., 2005] holds, and the Radon-Nikodym derivative on the RHS of (3.2) can be expressed as $\mathbb{E}^{b_0,\nu_0}[\log(\mathcal{E}(L_\delta))]$, where $\mathcal{E}$ is the Doléans-Dade stochastic exponential and the

process $L := (L_t)_{t\in[0,\delta]}$ is given as

$$L_t = \int_0^t \int_{\mathbb{R}_0^d} \left[\frac{d\nu_0}{d\nu}(\mathbf{X}_{s-}, \mathbf{z}) - 1\right] (\mathbf{Z}^\nu(\mathbf{X}_{s-}, d\mathbf{z}, ds) - \nu(\mathbf{X}_{s-}, d\mathbf{z})ds)$$
$$+ \int_0^t b_0(\mathbf{X}_s) - b(\mathbf{X}_s) - \int_{\mathbb{R}_0^d} \left(\frac{d\nu_0}{d\nu}(\mathbf{X}_{s-}, \mathbf{z}) - 1\right) \mathbb{1}_{(0,1]}(\|\mathbf{z}\|_2)\mathbf{z}\nu(\mathbf{X}_{s-}, d\mathbf{z})d\mathbf{X}_s^c,$$

where $(\mathbf{X}_s^c)_{s\geq 0}$ is the continuous martingale part of $\mathbf{X}$, i.e. a Brownian motion in this setting, and $\mathbf{Z}^\nu(\mathbf{x}, \cdot, \cdot)$ is a Poisson random measure with intensity $\nu(\mathbf{x}, d\mathbf{z}) \otimes ds$. Note that under $\mathbb{P}^{b_0,\nu_0}$ the process L is a local martingale, L^c is a continuous local martingale with quadratic variation

$$\langle L^c \rangle_t = \int_0^t \left\| b_0(\mathbf{X}_s) - b(\mathbf{X}_s) - \int_{\mathbb{R}_0^d} \left(\frac{d\nu_0}{d\nu}(\mathbf{X}_{s-}, \mathbf{z}) - 1\right) \mathbb{1}_{(0,1]}(\|\mathbf{z}\|_2)\mathbf{z}\nu(\mathbf{X}_{s-}, d\mathbf{z}) \right\|_2^2 ds$$

and jump discontinuities of L can be written as

$$\Delta L_t = \left[\frac{d\nu_0}{d\nu}(\mathbf{X}_{t-}, \Delta\mathbf{X}_t) - 1\right] \mathbb{1}_{(0,\infty)}(\|\Delta\mathbf{X}_t\|_2),$$

where $\Delta\mathbf{X}_t$ denotes a jump discontinuity of $\mathbf{X}$ at time t. Now, the expected quadratic variation of $\langle L^c \rangle_t$ can be bounded by

$$\mathbb{E}^{b_0,\nu_0}[\langle L^c \rangle_t] \leq \int_0^t \mathbb{E}^{b_0,\nu_0}[\|b_0(\mathbf{0}) + b(\mathbf{0}) + 2C_1\mathbf{X}_s + K\|_2^2]ds$$

for some constant $K > 0$, using (1.9), the uniform upper and lower bounds on $\frac{d\nu_0}{d\nu}$, and the fact that either ν and ν_0 are equivalent and either both finite, or $\frac{d\nu_0}{d\nu} \equiv 1$ on a neighbourhood of 0 and ν is finite on any open set not containing the origin. The stationary density has a first moment by Proposition 1, so that $\mathbb{E}^{b_0,\nu_0}[\langle L^c \rangle_t] \leq K't$ for some other constant $K' > 0$. Likewise,

$$\mathbb{E}^{b_0,\nu_0}\left[\sum_{t:\|\Delta\mathbf{X}_t\|_2\neq 0} \Delta L_t^2\right] = \int_0^t \mathbb{E}^{b_0,\nu_0}\left[\int_{\mathbb{R}_0^d} \left(\frac{d\nu_0}{d\nu}(\mathbf{X}_{s-}, \mathbf{z}) - 1\right)^2 \nu(\mathbf{X}_s, d\mathbf{z})\right] ds$$

is finite due to the aforementioned conditions on ν_0 and ν. Thus L has expected quadratic variation

$$\mathbb{E}^{b_0,\nu_0}[\langle L \rangle_t] = \mathbb{E}^{b_0,\nu_0}\left[\sum_{t:\|\Delta\mathbf{X}_t\|_2\neq 0} \Delta L_t^2 \nu(\mathbf{X}_s, d\mathbf{z})ds + \langle L^c \rangle_t\right] < \infty$$

for any $t > 0$, and is a true $\mathbb{P}^{b_0,\nu_0}$-martingale by Corollary 3 on page 73 of [Protter, 2005]. Then, the Radon-Nikodym term in (3.2) can be written as

$$\begin{aligned}
&\mathbb{E}^{b_0,\nu_0}\left[\log\left(\frac{d\mathbb{P}_{\mathbf{x}}^{b_0,\nu_0}}{d\mathbb{P}_{\mathbf{x}}^{b,\nu}}((\mathbf{X}_t)_{t\in[0,\delta]})\right)\right] = \mathbb{E}^{b_0,\nu_0}[\log(\mathcal{E}(L_t))] \\
&= \mathbb{E}^{b_0,\nu_0}\left[L_\delta - L_0 - \frac{1}{2}\langle L^c\rangle_\delta + \sum_{t:\Delta\mathbf{X}_t\neq 0}\{\log(1+\Delta L_t) - \Delta L_t\}\right] \\
&= \mathbb{E}^{b_0,\nu_0}\left[\frac{-1}{2}\int_0^\delta \Big\| b_0(\mathbf{X}_t) - b(\mathbf{X}_t) \right. \\
&\qquad - \int_{\mathbb{R}_0^d}\left(\frac{d\nu_0}{d\nu}(\mathbf{X}_{t-},\mathbf{z}) - 1\right)\mathbb{1}_{(0,1]}(\|\mathbf{z}\|_2)\mathbf{z}\nu(\mathbf{X}_{t-},d\mathbf{z})\Big\|_2^2 dt \\
&\qquad \left. + \sum_{0\le t\le\delta:\Delta\mathbf{X}_t\neq 0}\left\{\log\left(\frac{d\nu_0}{d\nu}(\mathbf{X}_{t-},\Delta\mathbf{X}_t)\right) - \left(\frac{d\nu_0}{d\nu}(\mathbf{X}_{t-},\Delta\mathbf{X}_t) - 1\right)\right\}\right] \\
&\le \delta\left[\frac{1}{2}\left(\|b_0 - b\|_{2,\pi^{b_0,\nu_0}} + \left\|\int_{\mathbb{R}_0^d}\left(\frac{d\nu_0}{d\nu}(\cdot,\mathbf{z}) - 1\right)\mathbb{1}_{(0,1]}(\|\mathbf{z}\|_2)\mathbf{z}\nu(\cdot,d\mathbf{z})\right\|_{2,\pi^{b_0,\nu_0}}\right)^2\right. \\
&\qquad \left. + \left\|\int_{\mathbb{R}_0^d}\left\{\log\left(\frac{d\nu_0}{d\nu}(\cdot,\mathbf{z})\right) - \frac{d\nu_0}{d\nu}(\cdot,\mathbf{z}) + 1\right\}\nu_0(\cdot,d\mathbf{z})\right\|_{1,\pi^{b_0,\nu_0}}\right], \qquad (3.5)
\end{aligned}$$

where the first equality follows from Theorem 2.4 of [Cheridito et al., 2005], the second by definition of $\mathcal{E}$ for jump diffusion processes, and the remainder of the calculation by stationarity and because ν_0 is the compensator of the Poisson random measure driving the jumps of $\mathbf{X}$ under $\mathbb{P}^{b_0,\nu_0}$. The result now follows from (3.2) and (3.5). □

Lemma 3. *For each $\delta > 0$ and $f \in \mathrm{Lip}(\Omega)$, the collection $\{P_\delta^{b,\nu} f : (b,\nu) \in \Theta\}$ is locally uniformly equicontinuous: for any compact $K \in \Omega$ and $\varepsilon > 0$ there exists $\gamma := \gamma(\varepsilon, f, \delta) > 0$ such that*

$$\sup_{(b,\nu)\in\Theta}\sup_{\substack{\mathbf{x},\mathbf{y}\in K:\\ \|\mathbf{x}-\mathbf{y}\|_2<\gamma}} |P_\delta^{b,\nu} f(\mathbf{x}) - P_\delta^{b,\nu} f(\mathbf{y})| < \varepsilon.$$

Proof. Theorem 2.2 of [Wang, 2010] uses a coupling argument to establish global equicontinuity for jump diffusions satisfying (1.9). A sufficient condition is that for

some constant $\beta \in (0,1)$ there exists a constant $C_\beta > 0$ such that

$$\|b(\mathbf{x}) - b(\mathbf{y})\|_2(1 + \|\mathbf{x}-\mathbf{y}\|_2) + C_\beta\|\mathbf{x}-\mathbf{y}\|_2(1+\|\mathbf{x}-\mathbf{y}\|_2)^2 \\ + \frac{(1+\|\mathbf{x}-\mathbf{y}\|_2)(\frac{1}{2} + \|\mathbf{x}-\mathbf{y}\|_2(1+\|\mathbf{x}-\mathbf{y}\|_2))\int_{\mathbb{R}_0^d}\|c(\mathbf{x},\mathbf{z}) - c(\mathbf{y},\mathbf{z})\|_2^2 M(d\mathbf{z})}{\|\mathbf{x}-\mathbf{y}\|_2} \leq 1$$

whenever $\|\mathbf{x}-\mathbf{y}\|_2 < \beta$. By (1.9), the LHS can be bounded above by

$$\sqrt{C_1}\beta(1+\beta) + \left(\frac{1}{2} + \beta(1+\beta)\right)\beta(1+\beta)C_1 + C_\beta\beta(1+\beta)^2,$$

which can clearly be made arbitrarily small by choosing a sufficiently small β. Note that a uniform bound on the Lipschitz constant C_1 is required because, to leading order, $\beta \sim C_1^{-1}$. A uniform bound on C_1 implies uniform equicontinuity, since then both β and C_β can be chosen uniformly. □

The remainder of the proof follows as in [van der Meulen and van Zanten, 2013]. It suffices to show that $Q(B|\mathbf{x}_{0:n}) \to 0$ with $\mathbb{P}^{b_0,\nu_0}$-probability 1 for $f \in \mathrm{Lip}(\Omega)$ and $B := \{(b,\nu) \in \Theta : \|P_\delta^{b,\nu}f - P_\delta^{b_0,\nu_0}f\|_{1,\rho} > \varepsilon\}$. To that end, fix $f \in \mathrm{Lip}(\Omega)$ and $\varepsilon > 0$ and thus the set B. Lemma 2 implies that Lemma 5.2 of [van der Meulen and van Zanten, 2013] holds, so that if, for measurable subsets $C_n \subset \Theta$, there exists $c > 0$ such that

$$e^{nc}\int_{C_n}\frac{\pi^{b,\nu}(\mathbf{x}_0)}{\pi^{b_0,\nu_0}(\mathbf{x}_0)}\prod_{i=1}^n\frac{p_\delta^{b,\nu}(\mathbf{x}_{i-1},\mathbf{x}_i)}{p_\delta^{b_0,\nu_0}(\mathbf{x}_{i-1},\mathbf{x}_i)}Q(db,d\nu) \to 0$$

$\mathbb{P}^{b_0,\nu_0}$-a.s. then $Q(C_n|\mathbf{x}_{0:n}) \to 0$ $\mathbb{P}^{b_0,\nu_0}$-a.s. as well. Likewise, Lemma 3 implies Lemma 5.3 of [van der Meulen and van Zanten, 2013]: there exists a compact subset $K \subset \Omega$, $N \in \mathbb{N}$ and compact, connected sets $I_1, \ldots, I_N$ that cover K such that

$$B \subset \bigcup_{j=1}^N B_j^+ \cup \bigcup_{j=1}^N B_j^-,$$

where

$$B_j^+ := \left\{(b,\nu) \in \Theta : P_\delta^{b,\nu}f(\mathbf{x}) - P_\delta^{b_0,\nu_0}f(\mathbf{x}) > \frac{\varepsilon}{4\rho(K)} \text{ for every } I_j\right\},$$
$$B_j^- := \left\{(b,\nu) \in \Theta : P_\delta^{b,\nu}f(\mathbf{x}) - P_\delta^{b_0,\nu_0}f(\mathbf{x}) < \frac{-\varepsilon}{4\rho(K)} \text{ for every } I_j\right\}.$$

Thus it is only necessary to show $Q(B_j^\pm|\mathbf{x}_{0:n}) \to 0$ $\mathbb{P}^{b_0,\nu_0}$-almost surely. Define the

stochastic process

$$D_n := \left(\int_{B_j^+} \frac{\pi^{b,\nu}(\mathbf{x}_0)}{\pi^{b_0,\nu_0}(\mathbf{x}_0)} \prod_{i=1}^{n} \frac{p_\delta^{b,\nu}(\mathbf{x}_{i-1}, \mathbf{x}_i)}{p_\delta^{b_0,\nu_0}(\mathbf{x}_{i-1}, \mathbf{x}_i)} Q(db, d\nu) \right)^{1/2}.$$

Now $D_n \to 0$ exponentially fast as $n \to \infty$ by an argument identical to that used to prove Theorem 3.5 of [van der Meulen and van Zanten, 2013]. The same is also true of the analogous stochastic process defined by integrating over B_j^-, which completes the proof. □

3.2.2 An example prior

The conditions of Theorem 5 are verifiable in the sense that they do not depend on intractable quantities (with the exception of Assumption 1), but it is not immediately clear whether a prior Q satisfying its assumptions exists, in particular in the infinite dimensional setting. The following example demonstrates that there is at least one family of priors which satisfies these assumptions: independent discrete net priors of Ghosal et al. [1997] for $b(\cdot)$ and $c(\cdot,\cdot)$, and a further, independent Dirichlet process mixture model prior [Lo, 1984] for $M(\cdot)$. Discrete net priors were also used in both van der Meulen and van Zanten [2013] and Gugushvili and Spreij [2014] to demonstrate the existence of priors for nonparametric inference of drifts for diffusions.

Firstly, let Θ_b be a collection of uniformly Lipschitz functions from Ω to $\mathbb{R}^d$, each satisfying (1.11) for some (not necessarily uniform) constants C_3 and C_4. Let $\Theta_b^{(m)} := \{b|_{\overline{B_0(m)}} : b \in \Theta_b\}$ be the set of restriction in Θ_b to the closed ball of radius m centred at the origin. By uniform equicontinuity and the Arzelà-Ascoli theorem, $\Theta_b^{(m)}$ is totally bounded in the uniform norm. Hence, for every n, it is possible to construct a finite ε_n-net $\Theta_b^{(m,n)}$ over $\Theta_b^{(m)}$, where $\{\varepsilon_n\}_{n\in\mathbb{N}}$ is a sequence of strictly positive numbers tending to 0. In other words, $\Theta_b^{(m,n)}$ is a finite set with the property that every element of $\Theta_b^{(m)}$ is within distance ε_n of some element of $\Theta_b^{(m,n)}$ in the supremum norm. Finally, every $b \in \Theta_b^{(m,n)}$ is extended to Ω by setting $b(\mathbf{x}) = b(P_{\overline{B_0(m)}}\mathbf{x}) - \mathbf{x} + P_{\overline{B_0(m)}}\mathbf{x}$ outside $\overline{B_0(m)}$, where $P_{\overline{B_0(m)}}$ is the orthogonal projection onto $\overline{B_0(m)}$. Now, a discrete net prior is constructed by fixing two probability mass functions on $\mathbb{N}$, $\{p_m\}_{m\in\mathbb{N}}$ and $\{q_n\}_{n\in\mathbb{N}}$, both of which assign positive mass to every positive integer. Then, a draw from the prior is generated by sampling $m \sim p_m$ and $n \sim q_n$, followed by $b|m,n \sim U(\Theta_b^{(m,n)})$. Samples from this prior are bounded, uniformly Lipschitz continuous, and satisfy (1.11) by construction.

Now let $J \subset \mathbb{R}^d$ be a fixed, compact domain including the origin, and let Θ_c

be a set of uniformly Lipschitz continuous functions $c : \Omega \times J \mapsto J$ which satisfy the following:

1. $c(\cdot, 0) \equiv 0$,
2. $c(\mathbf{x}, \cdot) : J \mapsto J$ is a surjection for each $\mathbf{x} \in \Omega$,
3. for any $c \in \Theta_c$, any $\mathbf{x} \in \Omega$ and $\mathbf{z} \in \mathbb{R}_0^d$, there exists an open ball of strictly positive radius centred at $\mathbf{z}$, $B_{\mathbf{z}}(\varepsilon)$, such that $c(\mathbf{x}, \mathbf{z}) \neq c(\mathbf{x}, \boldsymbol{\xi})$ for any $\mathbf{z} \neq \boldsymbol{\xi} \in B_{\mathbf{z}}(\varepsilon)$,
4. for each $c \in \Theta_c$ there exists $K_c > 0$ such that

$$\sup_{\mathbf{x}\in\Omega} \left\{ \sup_{\mathbf{z}\in J} \{c(\mathbf{x}, \mathbf{z})\} \right\} = \sup_{\mathbf{x}:\|\mathbf{x}\|_2 \leq K_c} \left\{ \sup_{\mathbf{z}\in J} \{c(\mathbf{x}, \mathbf{z})\} \right\}, \tag{3.6}$$

   and likewise for infima.

Condition 2. guarantees that $\nu(\mathbf{x}, d\mathbf{z})$ has a positive density everywhere whenever $M(d\mathbf{z})$ does, while condition 3. rules out atoms in $\nu(\mathbf{x}, d\mathbf{z})$. Let $\Theta_c^{(m)} := \{c|_{\overline{B_0(m)}} : c \in \Theta_c\}$ be the set of restrictions of the first coordinate to the ball $\overline{B_0(m)} \subset \Omega$, and let $\Theta_c^{(m,n)}$ be a $\tilde{\varepsilon}_n$-net over $\Theta_c^{(m)}$ for a strictly positive sequence $\tilde{\varepsilon}_n \searrow 0$. Each element of $\Theta_c^{(m,n)}$ can again be extended to a function on the whole $\Omega \times J$ by setting $c(\mathbf{x}, \mathbf{z}) = c(P_{\overline{B_0(m)}}\mathbf{x}, \mathbf{z})$ outside $\overline{B_0(m)}$, where $P_{\overline{B_0(m)}}$ denotes the orthogonal projection to $\overline{B_0(m)}$ as before. An independent discrete net can be used to define a prior for $c(\cdot, \cdot)$, by specifying two probability mass functions $\{\tilde{p}_m\}_{m\in\mathbb{N}}$ and $\{\tilde{q}_n\}_{n\in\mathbb{N}}$, both assigning positive mass to all positive integers, and sampling draws analogously to the discrete net prior on Θ_b.

Finally take the prior for the intensity measure $M(\cdot)$ to be a Dirichlet process mixture model [Lo, 1984]. Let $\phi_\tau(\mathbf{z})$ denote the d-dimensional centred Gaussian density with covariance matrix $\tau^{-1}\mathbb{I}_{d\times d}$ restricted to J, and renormalised to be a probability density. Let F be a probability measure on $(0, \infty)$ assigning positive mass to all non-empty open sets, and let $\mathrm{DP}(\zeta)$ denote the law of a Dirichlet process (c.f. Section 1.4) with the mean measure $\zeta \in \mathcal{M}_f(J)$, which is taken to be a probability measure with a finite first moment, independent of F. Let $\mathcal{D}_\Upsilon(J)$ denote the space of continuous, positive densities on J with total mass at most $\Upsilon > 0$. The Dirichlet process mixture model on $\mathcal{D}_\Upsilon(J)$ with truncated Gaussian mixture kernel ϕ_τ and mixing distribution $U(0, \Upsilon) \otimes F \otimes \mathrm{DP}(\zeta)$ is specified via the following sampling procedure:

1. Sample $P \sim \mathrm{DP}(\zeta)$. Then P is a discrete probability measure on $\mathbb{R}^d$ with

countably many atoms with DP(ζ)-probability 1 [Ferguson, 1973]. Let $\mathbf{z}_1, \mathbf{z}_2, \ldots$ denote these atoms in some fixed ordering.

2. Sample IID copies $\tau_1, \tau_2, \ldots \sim F$.

3. Sample $\alpha \sim U(0, \Upsilon)$.

4. Set $M(d\mathbf{z}) = \alpha \sum_{j=1}^{\infty} P(\mathbf{z}_j)\phi_{\tau_j}(\mathbf{z} - \mathbf{z}_j)d\mathbf{z}$.

Note that samples are finite measures with strictly positive densities on J, which also means they have second moments because J is compact.

Sampling all three components, $b(\cdot)$, $c(\cdot,\cdot)$ and $M(\cdot)$ independently from the priors specified above yields draws which almost surely satisfy (1.9) by uniform Lipschitz continuity of b and c, as well as a uniform bound on the total mass of M:

$$\begin{aligned}\|b(\mathbf{x}) - b(\mathbf{y})\|_2^2 + \int_J \|c(\mathbf{x},\mathbf{z}) - c(\mathbf{y},\mathbf{z})\|_2^2 M(d\mathbf{z}) \\ \leq C_b\|\mathbf{x}-\mathbf{y}\|_2^2 + \int_J C_c\|\mathbf{x}-\mathbf{y}\|_2^2 M(d\mathbf{z}) \leq (C_b + \Upsilon C_c)\|\mathbf{x}-\mathbf{y}\|_2^2.\end{aligned}$$

Condition (1.10) is immediate from the uniform Lipschitz continuity of c, and (1.11) holds by construction of the prior for b. The requirement that $c(\cdot, 0) \equiv 0$ holds by constructions. Finally, $\nu(\mathbf{x}, d\mathbf{z}) = M(c^*(\mathbf{x}, d\mathbf{z}))$ is a finite measure for each $\mathbf{x}$ because M is finite and $c^*(\mathbf{x},\mathbf{z})$ is a finite union of points by non-constancy, Lipschitz continuity and compactness of J. The Radon-Nikodym derivative $\frac{d\nu}{d\nu_0}$ exists for the same reason, and is bounded both from above and away from 0 by compactness of J and (3.6). Thus, the conditions of Definition 5 are fulfilled.

It remains to verify that (3.1) holds for this product prior. This will be achieved by controlling the three π^{b_0,ν_0}-norms separately, and showing that samples which result in all three taking arbitrarily small values are drawn with positive probability.

First, fix $b_0 \in \Theta_b$, $c_0 \in \Theta_c$ and $M_0 \in \mathcal{D}_\Upsilon(J)$, as well as $\varepsilon > 0$, and define

$$\|b\|_{m,\infty} := \sup_{\|\mathbf{x}\|_2 \leq m} \|b(\mathbf{x})\|_\infty. \tag{3.7}$$

Note that $\|\cdot\|_{m,\infty}$ is well defined for Lipschitz functions because they are locally bounded. Then

$$\begin{aligned}\|b_0 - b\|_{2,\pi^{b_0,\nu_0}}^2 &\leq \|b_0 - b\|_{m,\infty}^2 + \int_{\|\mathbf{x}\|_2 > m} \|b_0(\mathbf{x}) - b(\mathbf{x})\|_2^2 \pi^{b_0,\nu_0}(\mathbf{x})d\mathbf{x} \\ &\leq \|b_0 - b\|_{m,\infty}^2 + \int_{\|\mathbf{x}\|_2 > m} \|b_0(\mathbf{0}) + b(\mathbf{0}) + 2C_1\mathbf{x}\|_2^2 \pi^{b_0,\nu_0}(\mathbf{x})d\mathbf{x},\end{aligned}$$

by Lipschitz continuity. Now, choose m to be large enough that the second term on the RHS is bounded above by $\varepsilon/8$, which can be done because π^{b_0,ν_0} has second moments by Proposition 1. Likewise, the first term can be bounded by $\varepsilon/8$ by choosing n large enough that $\varepsilon_n \leq \varepsilon/8$. Note that by construction, the probability of sampling a corresponding function b from the prior is at least $p_m q_n > 0$, and that for such a b we have

$$\|b_0 - b\|_{2,\pi^{b_0,\nu_0}} \leq \sqrt{2\varepsilon/8} = \sqrt{\varepsilon}/2.$$

For the second norm, an elementary calculation using Jensen's inequality and the fact that $\|\mathbf{z}\|_2 \leq 1$ yields that

$$\begin{aligned}
&\Big\| \int_J \left[\frac{d\nu_0}{d\nu}(\cdot, \mathbf{z}) - 1 \right] \mathbb{1}_{(0,1]}(\|\mathbf{z}\|_2) \mathbf{z} \nu(\cdot, d\mathbf{z}) \Big\|_{2,\pi^{b_0,\nu_0}} \\
&\leq \int_\Omega \nu(\mathbf{x}, B_0(1)) \int_J \left(\frac{\nu_0(\mathbf{x},\mathbf{z})^2 - 2\nu_0(\mathbf{x},\mathbf{z})\nu(\mathbf{x},\mathbf{z}) + \nu(\mathbf{x},\mathbf{z})^2}{\nu(\mathbf{x},\mathbf{z})^2} \right) \nu(\mathbf{x}, d\mathbf{z}) \pi^{b_0,\nu_0}(\mathbf{x}) d\mathbf{x} \\
&\leq \int_\Omega \pi^{b_0,\nu_0}(\mathbf{x}) \nu(\mathbf{x}, B_0(1)) \Bigg(\Big| \int_{\mathbf{z}:\|\mathbf{z}\|_2 \leq 1} \frac{d\nu_0}{d\nu}(\mathbf{x},\mathbf{z}) \nu_0(\mathbf{x}, d\mathbf{z}) - \nu_0(\mathbf{x}, B_0(1)) \Big| \\
&\qquad\qquad + |\nu(\mathbf{x}, B_0(1)) - \nu_0(\mathbf{x}, B_0(1))| \Bigg) d\mathbf{x} \\
&= \int_{\mathbf{x}:\|\mathbf{x}\|_2 \leq m} \pi^{b_0,\nu_0}(\mathbf{x}) \nu(\mathbf{x}, B_0(1)) \Bigg(\Big| \int_{\mathbf{z}:\|\mathbf{z}\|_2 \leq 1} \frac{d\nu_0}{d\nu}(\mathbf{x},\mathbf{z}) \nu_0(\mathbf{x}, d\mathbf{z}) - \nu_0(\mathbf{x}, B_0(1)) \Big| \\
&\qquad\qquad + |\nu(\mathbf{x}, B_0(1)) - \nu_0(\mathbf{x}, B_0(1))| \Bigg) d\mathbf{x} \\
&+ \int_{\mathbf{x}:\|\mathbf{x}\|_2 > m} \pi^{b_0,\nu_0}(\mathbf{x}) \nu(\mathbf{x}, B_0(1)) \Bigg(\Big| \int_{\mathbf{z}:\|\mathbf{z}\|_2 \leq 1} \frac{d\nu_0}{d\nu}(\mathbf{x},\mathbf{z}) \nu_0(\mathbf{x}, d\mathbf{z}) - \nu_0(\mathbf{x}, B_0(1)) \Big| \\
&\qquad\qquad + |\nu(\mathbf{x}, B_0(1)) - \nu_0(\mathbf{x}, B_0(1))| \Bigg) d\mathbf{x}. \qquad (3.8)
\end{aligned}$$

Both $\nu(\mathbf{x}, \cdot)$ and $\nu_0(\mathbf{x}, \cdot)$ are finite measures with Radon-Nikodym derivative bounded from above and away from 0. Thus, finiteness of π^{b_0,ν_0} ensures that the second integral on the RHS can be made arbitrarily small by choosing large enough m. Now consider

$$\begin{aligned}
&|\nu(\mathbf{x}, B_0(1)) - \nu_0(\mathbf{x}, B_0(1))| = |M(c^*(\mathbf{x}, B_0(1))) - M_0(c_0^*(\mathbf{x}, B_0(1)))| \\
&\leq |M(c^*(\mathbf{x}, B_0(1))) - M(c_0^*(\mathbf{x}, B_0(1)))| + |M(c_0^*(\mathbf{x}, B_0(1))) - M_0(c_0^*(\mathbf{x}, B_0(1)))|,
\end{aligned}$$

and note that if $\|c - c_0\|_{m,\infty} \leq \gamma_1$ then

$$c^*(\mathbf{x}, B_0(1)) \subseteq c_0^*\left(\mathbf{x}, \left\{\boldsymbol{\xi} \in J : \inf_{\mathbf{y} \in B_0(1)} \{\|\boldsymbol{\xi} - \mathbf{y}\|_\infty\} \leq \gamma_1\right\}\right) \tag{3.9}$$

for any $\mathbf{x} : \|\mathbf{x}\|_2 \leq m$. The sets on the RHS are decreasing with decreasing $\gamma_1 > 0$ and of finite M-mass, so that continuity of measure gives $|M(c^*(\mathbf{x}, B_0(1))) - M(c_0^*(\mathbf{x}, B_0(1)))| \leq \gamma_2$ for some γ_2, which decreases to 0 as $\gamma_1 \searrow 0$. Likewise,

$$|M(c_0^*(\mathbf{x}, B_0(1))) - M_0(c_0^*(\mathbf{x}, B_0(1)))| \leq \|M - M_0\|_\infty.$$

Hence

$$|\nu(\mathbf{x}, B_0(1)) - \nu_0(\mathbf{x}, B_0(1))| \leq \gamma_2 + \|M - M_0\|_\infty,$$

which can be made arbitrarily small by first choosing a sufficiently large m, then a sufficiently small γ_2 as well as $c : \|c - c_0\|_{m,\infty} < \gamma_2$, and finally an $M : \|M - M_0\|_\infty < \gamma_3$ for sufficiently small γ_3.

Similarly, (3.9) gives that

$$\frac{d\nu_0}{d\nu}(\mathbf{x}, \mathbf{z}) = \frac{M_0(c_0^*(\mathbf{x}, \mathbf{z}))}{M(c^*(\mathbf{x}, \mathbf{z}))} \leq \frac{M_0(c^*(\mathbf{x}, \{\boldsymbol{\xi} \in J : \|\boldsymbol{\xi} - \mathbf{z}\|_\infty \leq \gamma_1\}))}{M(c^*(\mathbf{x}, \mathbf{z}))}$$

for $\mathbf{x} : \|\mathbf{x}\|_2 \leq m$ whenever $\|c - c_0\|_{m,\infty} < \gamma_1$. Hence taking such a c, as well as $M : \|M - M_0\|_\infty < \gamma_3$, and using continuity of measure yields the estimate

$$\frac{d\nu_0}{d\nu}(\mathbf{x}, \mathbf{z}) \leq \frac{M(c^*(\mathbf{x}, \mathbf{z})) + \gamma_3 + \gamma_4}{M(c^*(\mathbf{x}, \mathbf{z}))}$$

for some $\gamma_4 \searrow 0$ as $\gamma_1 \searrow 0$. The denominator is bounded from below, so that

$$\frac{d\nu_0}{d\nu}(\mathbf{x}, \mathbf{z}) \leq 1 + \frac{\gamma_3 + \gamma_4}{\inf_{\|\mathbf{x}\|_2 \leq m, \|\mathbf{z}\|_2 \leq 1}\{\nu(\mathbf{x}, \mathbf{z})\}},$$

which can be made arbitrarily close to 1 by choosing small enough γ_3 and γ_4. An analogous lower bound follows by reversing the roles of ν and ν_0, and lower bounding the Radon-Nikodym derivative instead. Thus

$$(1 - \gamma_3 - \gamma_4)\nu_0(\mathbf{x}, B_0(1)) \leq \int_{\mathbf{z}:\|\mathbf{z}\|_2 \leq 1} \frac{d\nu_0}{d\nu}(\mathbf{x}, \mathbf{z})\nu_0(\mathbf{x}, d\mathbf{z}) \leq \nu_0(\mathbf{x}, B_0(1))(1 + \gamma_3 + \gamma_4),$$

so that

$$\Big| \int_{\mathbf{z}:\|\mathbf{z}\|_2 \leq 1} \frac{d\nu_0}{d\nu}(\mathbf{x}, \mathbf{z})\nu_0(\mathbf{x}, d\mathbf{z}) - \nu_0(\mathbf{x}, B_0(1)) \Big| \leq \gamma_3 + \gamma_4.$$

Taken together, the above bounds imply that (3.8) can be bounded by $\sqrt{\varepsilon}/2$ by first choosing a large enough m, and then c and M such that $\|c - c_0\|_{m,\infty} < \gamma$ and $\|M - M_0\|_{m,\infty} < \gamma$ for sufficiently small $\gamma > 0$. Fix n such that $\tilde{\varepsilon}_n \leq \gamma$. Then a suitable c is sampled from the prior with probability at least $\tilde{p}_m \tilde{q}_n > 0$. The probability of sampling a suitable M is also positive, because by Theorem 1 of Bhattacharya and Dunson [2012], the support of the Dirichlet process mixture model is dense in $\mathcal{D}_\Upsilon(J)$.

The third norm in (3.1) can be treated identically to the second, because $x \mapsto \log(x)$ is continuous and J is compact. Hence, its value is also bounded by $\varepsilon/2$ with strictly positive Q-probability. Thus, with positive Q-probability

$$\begin{aligned}
&\frac{1}{2}\left(\|b_0 - b\|_{2,\pi^{b_0,\nu_0}} + \left\| \int_{\mathbb{R}_0^d} \left[\frac{d\nu_0}{d\nu}(\cdot, \mathbf{z}) - 1\right] \mathbb{1}_{(0,1]}(\|\mathbf{z}\|_2)\mathbf{z}\nu(\cdot, d\mathbf{z})\right\|_{2,\pi^{b_0,\nu_0}}\right)^2 \\
&+ \left\| \int_{\mathbb{R}_0^d} \left[\log\left(\frac{d\nu_0}{d\nu}(\cdot, \mathbf{z})\right) - \frac{d\nu_0}{d\nu}(\cdot, \mathbf{z}) + 1\right] \nu_0(\cdot, d\mathbf{z})\right\|_{1,\pi^{b_0,\nu_0}} \\
&< \frac{1}{2}\left(\frac{\sqrt{\varepsilon}}{2} + \frac{\sqrt{\varepsilon}}{2}\right)^2 + \frac{\varepsilon}{2} = \varepsilon,
\end{aligned}$$

and hence (3.1) holds.

3.2.3 Discussion

In this section I have shown that posterior consistency for joint, nonparametric Bayesian inference of drift and jump coefficients of jump diffusion SDEs from discrete data holds under criteria which can be readily checked in practice, subject to an identifiability assumption which is difficult to verify in general. This generalises previous work by [van der Meulen and van Zanten, 2013; Gugushvili and Spreij, 2014], in which similar results were proven for diffusions without jumps, in which setting identifiability can also be verified. Products of discrete net priors and Dirichlet process mixture models were shown to satisfy the conditions for consistency, provided that identifiability holds.

These results share the limitation of [van der Meulen and van Zanten, 2013; Gugushvili and Spreij, 2014] of being established for a weak topology, for which the martingale approach of [Walker, 2004; Lijoi et al., 2004] is well suited. An approach based on constructing hypothesis tests for the true coefficients (b_0, ν_0) with exponentially small error probability in n, such as that of [Ghosal and van der Vaart, 2007], would yield convergence in a stronger topology as well as rates of convergence, but it is not clear how to adapt their results to the diffusion or jump

diffusion settings. Currently, results in this direction are only available for scalar diffusions [van der Meulen et al., 2006; Panzar and van Zanten, 2009; Pokern et al., 2013; Nickl and Söhl, 2015].

Practical implementation of inference algorithms is beyond the scope of this thesis, but note that algorithms based on exact simulation for jump diffusions are available, at least in the scalar case [Casella and Roberts, 2011; Gonçalves, 2011; Pollock et al., 2015b]. Exact simulation of jump diffusions is an active area of research [Gonçalves and Roberts, 2013; Pollock et al., 2015a; Pollock, 2015] and well suited for applications in Monte Carlo inference algorithms, with preliminary results in the continuous diffusion setting indicating that nonparametric algorithms can be feasibly implemented [Papaspiliopoulos et al., 2012; van Zanten, 2013; van der Meulen et al., 2014].

3.3 Λ-coalescents

Let $Q \in \mathcal{M}_1(\mathcal{M}_1([0,1]))$ be a prior distribution for Λ, and $\mathbf{n} \in \mathbb{N}^{|\mathcal{H}|}$ denote observed haplotype frequencies of $n \in \mathbb{N}$ $\mathcal{H}$-labelled lineages generated by the Λ-coalescent. For Borel sets $B \in \mathcal{B}(\mathcal{M}_1([0,1]))$, define the posterior as

$$Q(B|\mathbf{n}) = \frac{\int_B \mathbf{P}_n^\Lambda(\mathbf{n}) Q(d\Lambda)}{\int_{\mathcal{M}_1([0,1])} \mathbf{P}_n^\Lambda(\mathbf{n}) Q(d\Lambda)}.$$

I will begin this section by showing that the posterior given contemporaneously observed haplotypes is necessarily inconsistent for any non-trivial prior. Since contemporaneous samples are typical in population genetics, this result is of great applied importance whenever Λ cannot be assumed known with certainty. Unlike the consistency result of Theorem 7, the inconsistency result of Theorem 6 is very universal and does not depend on mode of convergence or topology.

Theorem 6. *Let $\mathbf{n} \in \mathbb{N}^{|\mathcal{H}|}$ denote the observed haplotype frequencies in a sample of size $n \in \mathbb{N}$ generated by a Λ-coalescent started from n leaves at a fixed time, and let $\mathbf{x} := \lim_{n\to\infty} \frac{\mathbf{n}}{n} \in \mathcal{M}_1(\mathcal{H})$ denote the limiting observed relative haplotype frequencies. Then the limiting posterior is given by*

$$\lim_{n\to\infty} Q(B|\mathbf{n}) = \frac{\int_B \pi^\Lambda(\mathbf{x}) Q(d\Lambda)}{\int_{\mathcal{M}_1([0,1])} \pi^\Lambda(\mathbf{x}) Q(d\Lambda)}.$$

In particular, the RHS is positive for any $B \in \mathcal{B}(\mathcal{M}_1([0,1]))$ which has a non-null intersection with the support of Q, regardless of the $\Lambda \in \mathcal{M}_1([0,1])$ generating the

data.

Proof. Conditioning on the ancestral tree of the observed sample give the following representation for the posterior:

$$Q(A|\mathbf{n}) = \frac{\int_A \mathbf{P}_n^\Lambda(\mathbf{n})Q(d\Lambda)}{\int_{\mathcal{M}_1([0,1])} \mathbf{P}_n^\Lambda(\mathbf{n})Q(d\Lambda)} = \frac{\int_A E_n^\Lambda\left[\mathbb{1}_{\{\mathbf{n}\}}(\Pi_0)\right]Q(d\Lambda)}{\int_{\mathcal{M}_1([0,1])} E_n^\Lambda\left[\mathbb{1}_{\{\mathbf{n}\}}(\Pi_0)\right]Q(d\Lambda)}.$$

Using (1.7) the above can be written as

$$Q(A|\mathbf{n}) = \frac{\int_A \mathbb{E}^\Lambda\left[q(\mathbf{n}|\mathbf{X})\right]Q(d\Lambda)}{\int_{\mathcal{M}_1([0,1])} \mathbb{E}^\Lambda\left[q(\mathbf{n}|\mathbf{X})\right]Q(d\Lambda)},$$

where $q(\mathbf{n}|\mathbf{X}) := \binom{n}{n_1,\ldots,n_{|\mathcal{H}|}} \prod_{h\in\mathcal{H}} X_h^{n_h}$ is the multinomial sampling probability. I will show the requisite convergence of the numerator and denominator separately, and the result will follow by the algebra of limits.

Consider first the numerator. By Fubini's theorem

$$\int_A \mathbb{E}^\Lambda\left[q(\mathbf{n}|\mathbf{X})\right]Q(d\Lambda) = \int_{\Delta_\mathcal{H}} q(\mathbf{n}|\mathbf{y}) \int_A \pi^\Lambda(\mathbf{y})Q(d\Lambda)d\mathbf{y} =: \int_{\Delta_\mathcal{H}} q(\mathbf{n}|\mathbf{y})F_{Q;A}(\mathbf{y})d\mathbf{y},$$

where $F_{Q;A}(\mathbf{y})$ is a sub-probability density on $\Delta_\mathcal{H}$ since it is a mixture of probability densities. Hence $F_{Q;A}(\mathbf{y})d\mathbf{y}$ defines a finite measure on $\Delta_\mathcal{H}$, and $q(\mathbf{n}|\mathbf{y}) \leq 1$ so that by the Dominated Convergence theorem

$$\lim_{n\to\infty} \int_{\Delta_\mathcal{H}} q(\mathbf{n}|\mathbf{y})F_{Q;A}(\mathbf{y})d\mathbf{y} = \int_{\Delta_\mathcal{H}} \lim_{n\to\infty} q(\mathbf{n}|\mathbf{y})F_{Q;A}(\mathbf{y})d\mathbf{y}.$$

By the Law of Large Numbers $\mathbf{n} \sim \lfloor n\mathbf{x} \rfloor$ so that $q(\mathbf{n}|\mathbf{y}) \to q(\lfloor n\mathbf{x} \rfloor|\mathbf{y})$, and by Stirling's formula

$$q(\lfloor n\mathbf{x} \rfloor|\mathbf{y}) \sim \prod_{h\in\mathcal{H}} \left(\frac{y_h}{x_h}\right)^{nx_h},$$

or

$$\log(q(\lfloor n\mathbf{x} \rfloor|\mathbf{y})) \sim n\sum_{h\in\mathcal{H}} x_h \log\left(\frac{y_h}{x_h}\right) = -n\sum_{h\in\mathcal{H}} x_i \log\left(\frac{x_i}{y_i}\right) = -n\,\mathrm{KL}(\mathbf{x},\mathbf{y}),$$

where $\mathrm{KL}(\mathbf{x},\mathbf{y})$ denotes the Kullback-Leibler divergence between the probability mass functions $\mathbf{x}$ and $\mathbf{y}$. By Gibbs' inequality $\mathrm{KL}(\mathbf{x},\mathbf{y}) \geq 0$ and $\mathrm{KL}(\mathbf{x},\mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$, so that $q(\mathbf{n}|\cdot) \to \delta_\mathbf{x}(\cdot)$ almost surely. Hence

$$\int_{\Delta_\mathcal{H}} \lim_{n\to\infty} q(\mathbf{n}|\mathbf{y})F_{Q;A}(\mathbf{y})d\mathbf{y} = F_{Q;A}(\mathbf{x}) = \int_A \pi^\Lambda(\mathbf{x})Q(d\Lambda),$$

as required. The argument for the denominator is identical after substituting $\mathcal{M}_1([0,1])$ for the domain of integration A. □

Remark 8. There is an apparent contradiction between the negative conclusion of Theorem 6 and recent positive results [Spence et al., 2016, Theorems 2, 3, 4 and 5] showing that Λ-measures can often be identified from their site frequency spectra. The contradiction is resolved by noting that Spence et al. [2016] work directly with the expected site frequency spectrum, thereby sidestepping both the randomness of the ancestral tree and the randomness of the mutation process given the tree. Numerical investigations by Spence et al. [2016] show that their method is unreliable unless a modest number (10-100) of independent realisations of ancestral trees is available. Independent trees cannot be sampled from populations whose ancestry is described by any non-Kingman Λ-coalescent, even in the idealised scenario of an infinitely long genome in the presence of recombination. However, as noted by [Spence et al., 2016], in some cases the decay of correlations with increasing genome length is determined by the prelimiting model of evolution, and not necessarily the limiting Λ-coalescent. For example, the selective sweep model of Durrett and Schweinsberg [2005] can allow for asymptotically independent trees across a genome in the presence of multiple mergers for some combinations of parameters, in which case the identifiability results of Spence et al. [2016] hold.

The following example is an extension of a result by Der and Plotkin [2014], and demonstrates that the lack of consistency can have dramatic consequences for statistical identifiability even in the case of very simple priors.

Example 1. Consider $\mathcal{H} = \{0,1\}$, $M_{hh'} = 1/2$ for $h, h' \in \mathcal{H}$, and set $Q(d\Lambda) = \frac{1}{2}\delta_{\delta_0}(d\Lambda) + \frac{1}{2}\delta_{\delta_1}(d\Lambda)$. The stationary law $\pi^{\Lambda}(x)$ is known in the parent-independent, two-allele case for both of these atoms [Der and Plotkin, 2014]:

$$\pi^{\delta_0}(x) = \frac{\Gamma(2\theta)}{[\Gamma(\theta)]^2} x^{\theta-1}(1-x)^{\theta-1}$$
$$\pi^{\delta_1}(x) = \frac{1}{\theta}|1-2x|^{\frac{1-\theta}{\theta}},$$

so the expected limiting posterior probabilities can be computed assuming either data-generating measure. These are listed in Table 3.1 for some candidate values of θ, while Figure 3.1 depicts limiting posterior probabilities as functions of the observed allele frequencies. The extreme sensitivity of the posterior probabilities in Figure 3.1 is akin to the "Bayesian brittleness" investigated by Owhadi et al. [2015], resulting in inferences which are not robust to small changes in the observed allele

frequencies, prior probabilities or latent parameters.

θ	$\mathbb{E}^{\delta_0}[Q(\delta_0\|X)]$	$\mathbb{E}^{\delta_1}[Q(\delta_0\|X)]$
0.04	0.84	0.16
0.1	0.73	0.27
0.5	0.54	0.46
1	0.5	0.5
5	0.65	0.35
10	0.75	0.25
17	0.82	0.18

Table 3.1: Expected posterior probabilities given an infinite number of simultaneous observations in the parent-independent, two-allele model.

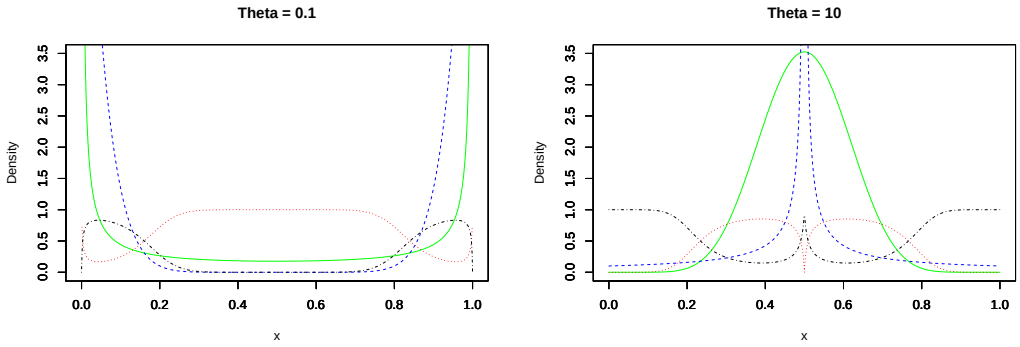


Figure 3.1: Limiting posterior probabilities as functions of the observed allele frequency x for $\theta = 0.1$ and 10 in the parent-independent, two-allele model. $Q(\delta_0|x)$ is dotted, $Q(\delta_1|x)$ is dash-dotted, $\pi^{\delta_0}(x)$ is solid and $\pi^{\delta_1}(x)$ is dashed. Note the extreme sensitivity of the posterior to the observed allele frequencies near $x = 0.5$ when $\theta = 10$ and near $x = 0$ or 1 when $\theta = 0.1$.

The fact that $\pi^{\delta_0}(x) = \pi^{\delta_1}(x)$ when $\theta = 1$ in Example 1 was pointed out by Der and Plotkin [2014] as proof of the fact that Λ-measures cannot in general be uniquely identified from independent draws from $\pi^{\Lambda}(x)$. These calculations illustrate that inference suffers from low power even when $\theta \neq 1$ if all observations are contemporaneous.

The inconsistency result of Theorem 6 holds for essentially arbitrary priors. The next aim is to show that the posterior can be consistent when the data set is a time series of increasing length. This does not contradict the non-identifiability claim of Der and Plotkin [2014], because they only consider independent draws from π^{Λ}. In contrast, in this setting it is information about transition probabilities $p^{\Lambda}_{\Delta}(\mathbf{x}, d\mathbf{y})$ which facilitates posterior consistency.

3.3.1 Posterior consistency

The topology and definition of weak posterior consistency are still as in Section 3.2, but I will restate them here for continuity of presentation.

Definition 8. Fix $\eta > 0$ and let $\mathcal{D}_\eta$ be a collection of Lebesgue densities on $[\eta, 1]$ satisfying $\inf_{r\in[\eta,1]} \phi(r) > 0$ and $\sup_{r\in[\eta,1]} \phi(r) < \infty$ for each $\phi \in \mathcal{D}_\eta$. Note that neither bound need be uniform in $\mathcal{D}(\eta)$. Assume that $\Lambda(dr) = \phi(r)dr$ for $\phi \in \mathcal{D}_\eta$, and denote the data generating density by ϕ_0.

Restricting the support of ϕ to $[\eta, 1]$ ensures that the Λ-coalescent can have no Kingman component, and that the corresponding Λ-Fleming-Viot process is a compound Poisson process with drift. Furthermore, most previously studied parametric families of Λ-measures, such as those mentioned in Section 1.1.1, are ruled out. However, Section 3.3.3 will show that the prior can be chosen to satisfy the conditions of Definition 8 and place mass arbitrarily close to any desired Λ-measure, or family of Λ-measures, in a way which will be made precise in Example 2.

Assumption 1 continues to be a standing assumption in this section, for the same reasons as in the previous section. It is restated below for continuity of presentation.

Assumption 2. For any pair $\phi \neq \phi' \in \mathcal{D}_\eta$ and any $\delta > 0$ there exists $\mathbf{x} \in \Delta_\mathcal{H}$ and $f : \Delta_\mathcal{H} \mapsto \mathbb{R}$ such that $P_\delta^{\phi'} f(\mathbf{x}) \neq P_\delta^{\phi} f(\mathbf{x})$. In particular, identifying P_δ^ϕ is equivalent to identifying ϕ.

Definition 9. Fix a sampling interval $\delta > 0$ and a finite Borel measure ρ on $\Delta_\mathcal{H}$ placing positive mass in all non-empty open sets. A weak topology on $\mathcal{D}_\eta$ is generated by requiring that open sets of the form

$$U_{f,\varepsilon}^\phi := \{\phi' : \|P_\delta^{\phi'} f(\mathbf{x}) - P_\delta^\phi f(\mathbf{x})\|_{1,\rho} < \varepsilon\},$$

for any $\phi \in \mathcal{D}_\eta$, $\varepsilon > 0$ and $f \in C(\Delta_\mathcal{H})$ form a subbase of the topology [Dudley, 2002].

Lemma 4. *The topology generated by a subbase of sets of the form $U_{f,\varepsilon}^\phi$ is Hausdorff.*

Proof. This proof is identical to that of Lemma 1 in Section 3.2. □

Definition 10. Let $\mathbf{n}_0, \ldots, \mathbf{n}_p$ denote $p + 1$ samples observed at times $0, \delta, \ldots, \delta m$ generated by a stationary Λ-coalescent, with each sample being of size $n \in \mathbb{N}$. See e.g. [Beaumont, 2003] for details of how temporally structured samples can be generated.

Theorem 7. *Let $\mathbf{n}_0, \ldots, \mathbf{n}_p$ be as in Definition 10 and $\mathbf{x}_0, \ldots, \mathbf{x}_p$ denote the observed limiting type frequencies as $n \to \infty$, i.e. $\mathbf{x}_i = \lim_{n\to\infty} \mathbf{n}_i/n$. Suppose that the prior Q is supported on a set $\mathcal{D}_\eta$ which satisfies the conditions in Definition 8, that Assumption 2 holds, and that*

$$Q\left(\phi \in \mathcal{D}_\eta : \int_\eta^1 \left\{ \left| \log\left(\frac{\phi_0(r)}{\phi(r)}\right) \right| + \left| \frac{\phi_0(r)}{\phi(r)} - 1 \right| \right\} r^{-2}\phi_0(r)dr < \varepsilon \right) > 0 \qquad (3.10)$$

for any $\varepsilon > 0$ and any $\phi_0 \in \mathcal{D}_\eta$. Then weak posterior consistency holds for Q on $\mathcal{D}_\eta$.

Remark 9. This result is similar to Theorem 5, and the proof will follow a similar structure. Both proofs of consistency require verification of a Kullback-Leibler condition for the prior, and uniform equicontinuity of the family of semigroups corresponding to densities supported by the prior. The former result is immediate by the same argument used to prove Lemma 2 in Section 3.2.1. The latter, Lemma 6 below, is different to its counterpart, Lemma 3 in Section 3.2.1. In particular, uniform Lipschitz continuity of the functions specifying jump sizes turns out not to be necessary here.

Proof. For fixed $p \in \mathbb{N}$, the same argument used to prove Theorem 6 yields that the following convergence holds $\mathbb{P}^{\phi_0}$-a.s. as $n \to \infty$:

$$\lim_{n\to\infty} Q(d\phi|\mathbf{n}_0, \ldots, \mathbf{n}_p) \propto \pi^\phi(\mathbf{x}_0) \prod_{i=1}^{p} p_\delta^\phi(\mathbf{x}_{i-1}, \mathbf{x}_i) Q(d\phi).$$

Hence it is sufficient to establish posterior consistency for discrete observations from a stationary Λ-Fleming-Viot process as $p \to \infty$. This is achieved by adapting the proof of Theorem 5. For $\phi \in \mathcal{D}_\eta$ let $\mathrm{KL}(\phi_0, \phi)$ denote the Kullback-Leibler divergence between $p_\delta^{\phi_0}$ and p_δ^ϕ:

$$\mathrm{KL}(\phi_0, \phi) := \int_{\Delta_\mathcal{H}} \int_{\Delta_\mathcal{H}} \log\left(\frac{p_\delta^{\phi_0}(\mathbf{x}, \mathbf{y})}{p_\delta^\phi(\mathbf{x}, \mathbf{y})}\right) p_\delta^{\phi_0}(\mathbf{x}, \mathbf{y}) \pi^{\phi_0}(\mathbf{x}) d\mathbf{y} d\mathbf{x},$$

and recall that for two probability measures P, P' on the same σ-field

$$K(P, P') := \mathbb{E}_P\left[\log\left(\frac{dP}{dP'}\right)\right].$$

The law of a random object Z under a probability measure P is denoted by $\mathcal{L}(Z|P)$.

As in the case of Theorem 5, the following two properties are required:

1. $Q(\phi \in \mathcal{D}_\eta : \mathrm{KL}(\phi_0, \phi) < \varepsilon) > 0$ for any $\varepsilon > 0$.

2. Uniform equicontinuity of the semigroups $\{P_\delta^\phi f : \phi \in \mathcal{D}_\eta\}$ for $f \in \mathrm{Lip}(\Delta_\mathcal{H})$, the set of Lipschitz functions on $\Delta_\mathcal{H}$.

The first will be established in Lemma 5 below. It is a straightforward adaptation of Lemma 2 in Section 3.2.1, and hence the proof is omitted.

Lemma 5. *Condition* (3.10) *implies that* $Q(\phi \in \mathcal{D}_\eta : \mathrm{KL}(\phi_0, \phi) < \varepsilon) > 0$ *for any* $\varepsilon > 0$.

The second condition can be established by verifying the hypotheses of Lemma 3 continue to hold in this setting without a positive definite diffusion coefficient.

Lemma 6. *For each* $\delta > 0$ *and* $f \in \mathrm{Lip}(\Delta_\mathcal{H})$, *the collection* $\{P_\delta^\phi f : \phi \in \mathcal{D}_\eta\}$ *is uniformly equicontinuous: for any* $\varepsilon > 0$ *there exists* $\gamma := \gamma(\varepsilon, f, \delta) > 0$ *such that*

$$\sup_{\phi \in \mathcal{D}_\eta} \sup_{\|\mathbf{x}-\mathbf{y}\|_2 < \gamma} |P_\delta^\phi f(\mathbf{x}) - P_\delta^\phi f(\mathbf{y})| < \varepsilon.$$

Proof. It is again sufficient to verify the hypotheses of Proposition 2.2 of Wang [2010], which establishes uniform equicontinuity for general Lévy-driven SDEs. It will be convenient to write the generator (1.6) in the following form:

$$\begin{aligned}\mathcal{G}^\phi f(\mathbf{x}) = \theta \sum_{h,h' \in \mathcal{H}} x_j (M_{h'h} - \delta_{hh'}) \frac{\partial}{\partial x_h} f(\mathbf{x}) \\ + \int_\eta^1 \int_0^1 \{f(\mathbf{x} + r\Psi(u,\mathbf{x})) - f(\mathbf{x}) - r\Psi(u,\mathbf{x})\nabla f(\mathbf{x})\} r^{-2} \phi(r) du dr,\end{aligned}$$

where $\Psi(u,\mathbf{x}) := \sum_{h \in \mathcal{H}} (\mathbb{1}_{(S_{h-1}^\mathbf{x}, S_h^\mathbf{x}]}(u) - x_h)\mathbf{e}_h$, $S_h^\mathbf{x} := \sum_{j=1}^h x_h$ and $S_0^\mathbf{x} = 0$, and where $\mathcal{H}$ has been endowed with a fixed but arbitrary ordering and associated with a set $\{1, \ldots, d\}$ for some $d \in \mathbb{N}$. As pointed out by Bertoin and Le Gall [2005], the two generators are equal because $\int_0^1 \Psi(u,\mathbf{x}) du = 0$ for any $\mathbf{x} \in \Delta_\mathcal{H}$.

For functions $f \in C^2(\Delta_\mathcal{H} \times \Delta_\mathcal{H})$ let $\mathcal{L}^\phi$ be the coupling generator

$$\begin{aligned}\mathcal{L}^\phi f(\mathbf{x},\mathbf{y}) := \theta \sum_{h,h' \in \mathcal{H}} (M_{h'h} - \delta_{hh'}) \left(x_{h'} \frac{\partial}{\partial x_h} f(\mathbf{x},\mathbf{y}) + y_{h'} \frac{\partial}{\partial y_h} f(\mathbf{x},\mathbf{y}) \right) \\ + \int_\eta^1 \int_0^1 \{f(\mathbf{x} + r\Psi(u,\mathbf{x}), \mathbf{y} + r\Psi(u,\mathbf{y})) - f(\mathbf{x},\mathbf{y}) \\ - r\Psi(u,\mathbf{x})\nabla_\mathbf{x} f(\mathbf{x},\mathbf{y}) - r\Psi(u,\mathbf{y})\nabla_\mathbf{y} f(\mathbf{x},\mathbf{y})\} r^{-2} \phi(r) du dr,\end{aligned}$$

where $\nabla_\mathbf{x}$ and $\nabla_\mathbf{y}$ are the gradient operators with respect to the $\mathbf{x}$ and $\mathbf{y}$ variables, respectively.

Fix the test function $g(z) := z(1+z)^{-1}$. It is straightforward to check that for this function

$$\begin{aligned}\mathcal{L}^{\phi} g(\|\mathbf{x}-\mathbf{y}\|_2) &\leq \frac{2B(\mathbf{x},\mathbf{y}) + \int_0^1 \|\Psi(u,\mathbf{x}) - \Psi(u,\mathbf{y})\|_2^2 du}{2\|\mathbf{x}-\mathbf{y}\|^2(1+\|\mathbf{x}-\mathbf{y}\|_2)} g(\|\mathbf{x}-\mathbf{y}\|_2) \\ &=: K(\mathbf{x},\mathbf{y}) g(\|\mathbf{x}-\mathbf{y}\|_2)\end{aligned}$$

uniformly in $\phi \in \mathcal{D}_\eta$, where $B(\mathbf{x},\mathbf{y}) := \theta(\mathbf{x}-\mathbf{y})^T(M - \mathbb{I}_{\mathcal{H}})(\mathbf{x}-\mathbf{y})$ and $\mathbb{I}_{\mathcal{H}}$ is the $|\mathcal{H}| \times |\mathcal{H}|$ identity matrix.

Proposition 2.2 of Wang [2010] requires that $K(\mathbf{x},\mathbf{y})\|\mathbf{x}-\mathbf{y}\|_2 \leq C_\gamma \|\mathbf{x}-\mathbf{y}\|_2$ for $\|\mathbf{x}-\mathbf{y}\|_2 \leq \gamma$, and some constant C_γ depending only on γ, θ and M. Note that $B(\mathbf{x},\mathbf{y}) \leq \theta \|M - \mathbb{I}_{\mathcal{H}}\|_2 \|\mathbf{x}-\mathbf{y}\|_2^2$, so that the result will follow if $\int_0^1 \|\Psi(u,\mathbf{x}) - \Psi(u,\mathbf{y})\|_2^2 du \leq C\|\mathbf{x}-\mathbf{y}\|_2$ for some constant C. Now

$$\begin{aligned}&\int_0^1 \|\Psi(u,\mathbf{x}) - \Psi(u,\mathbf{y})\|_2^2 du = 2 - \|\mathbf{x}-\mathbf{y}\|_2^2 - 2\sum_{h\in\mathcal{H}} \int_0^1 \mathbb{1}_{(S^{\mathbf{x}}_{h-1}, S^{\mathbf{x}}_h]}(u) \mathbb{1}_{(S^{\mathbf{y}}_{h-1}, S^{\mathbf{y}}_h]}(u) du \\ &= 2\sum_{h\in\mathcal{H}} \left(S^{\mathbf{x}}_h - S^{\mathbf{x}}_{h-1} - \int_0^1 \mathbb{1}_{(S^{\mathbf{x}}_{h-1}, S^{\mathbf{x}}_h]}(u) \mathbb{1}_{(S^{\mathbf{y}}_{h-1}, S^{\mathbf{y}}_h]}(u) du \right) - \|\mathbf{x}-\mathbf{y}\|_2^2,\end{aligned}$$

and hence it suffices to bound

$$\sum_{h\in\mathcal{H}} S^{\mathbf{x}}_h - S^{\mathbf{x}}_{h-1} - \int_0^1 \mathbb{1}_{(S^{\mathbf{x}}_{h-1}, S^{\mathbf{x}}_h]}(u) \mathbb{1}_{(S^{\mathbf{y}}_{h-1}, S^{\mathbf{y}}_h]}(u) du \leq C\|\mathbf{x}-\mathbf{y}\|_2.$$

Consider a single index $h \in \mathcal{H}$ and suppose without loss of generality that $S^{\mathbf{y}}_h \leq S^{\mathbf{x}}_h$ and if $S^{\mathbf{y}}_h = S^{\mathbf{x}}_h$ then $S^{\mathbf{x}}_{h-1} < S^{\mathbf{y}}_{h-1}$. There are two cases: either $(S^{\mathbf{y}}_{h-1}, S^{\mathbf{y}}_h] \subseteq (S^{\mathbf{x}}_{h-1}, S^{\mathbf{x}}_h]$ or $S^{\mathbf{y}}_{h-1} < S^{\mathbf{x}}_{h-1}$ and $S^{\mathbf{y}}_h < S^{\mathbf{x}}_h$. In the former case

$$S^{\mathbf{x}}_h - S^{\mathbf{x}}_{h-1} - \int_0^1 \mathbb{1}_{(S^{\mathbf{x}}_{h-1}, S^{\mathbf{x}}_h]}(u) \mathbb{1}_{(S^{\mathbf{y}}_{h-1}, S^{\mathbf{y}}_h]}(u) du = x_h - y_h \leq \|\mathbf{x}-\mathbf{y}\|_1,$$

while in the latter

$$\begin{aligned}&S^{\mathbf{x}}_h - S^{\mathbf{x}}_{h-1} - \int_0^1 \mathbb{1}_{(S^{\mathbf{x}}_{h-1}, S^{\mathbf{x}}_h]}(u) \mathbb{1}_{(S^{\mathbf{y}}_{h-1}, S^{\mathbf{y}}_h]}(u) du = S^{\mathbf{x}}_h - S^{\mathbf{x}}_{h-1} - (S^{\mathbf{y}}_h - S^{\mathbf{x}}_{h-1})^+ \\ &\leq S^{\mathbf{x}}_h - S^{\mathbf{y}}_h + S^{\mathbf{x}}_{h-1} - S^{\mathbf{y}}_{h-1} \leq \sum_{h'=h+1}^{|\mathcal{H}|} |x_{h'} - y_{h'}| + \sum_{h'=1}^{h-1} |x_{h'} - y_{h'}| \leq \|\mathbf{x}-\mathbf{y}\|_1\end{aligned}$$

because the L^1-distance between the sub-probability measures $(x_1, \ldots, x_{h-1})$ and $(y_1, \ldots, y_{h-1})$ on $[h-1]$ is at least as large as the difference in their total masses,

and likewise for the corresponding pair on $\{h+1,\dots,|\mathcal{H}|\}$. Summing over indices yields

$$\int_0^1 \|\Psi(u,\mathbf{x})-\Psi(u,\mathbf{y})\|_2^2 du \leq 2|\mathcal{H}|^{3/2}\|\mathbf{x}-\mathbf{y}\|_2$$

since $\|\mathbf{z}\|_1 \leq \sqrt{|\mathcal{H}|}\|\mathbf{z}\|_2$ by the Cauchy-Schwarz inequality. Hence by Proposition 2.2 of Wang [2010]

$$\sup_{\|\mathbf{x}-\mathbf{y}\|_2<\gamma} |P_\delta^\phi f(\mathbf{x}) - P_\delta^\phi f(\mathbf{y})| < \varepsilon$$

uniformly in $\phi \in \mathcal{D}_\eta$ as required. □

The remainder of the proof of Theorem 7 is identical to that of Theorem 5, and is omitted. □

Note that bounding the support of Λ away from 0 is not necessary: the proof could be adapted to equivalent collections of measures supported on $(0,1]$ for which (3.10) holds, e.g. by requiring $\phi(r) \sim r^2$ as $r \searrow 0$ for Q-a.e. $\phi \in \mathcal{D}_\eta$. However, any such condition is bound to be mathematically restrictive. In the next section I will show that given a data set of size n, it is natural to parametrise inference with the first $n-2$ moments of Λ because they fully capture the signal in the data set. Example 2 in Section 3.3.3 provides a family of priors which satisfy the hypotheses of Theorem 7, and whose support can be chosen to contain arbitrarily close approximations to finite moment sequences of any $\Lambda \in \mathcal{M}_1([0,1])$.

Remark 10. The hypotheses of Theorem 7 are strong, and thus it would be desirable to obtain a posterior contraction rate in addition to just consistency. In fact, methods akin to that employed in the proof have been extended to provide rates for compound Poisson processes [Gugushvili et al., 2015] and scalar diffusions on compact intervals [Nickl and Söhl, 2015]. However, extending either approach to this setting would require bounds of the form

$$\|\pi^\phi - \pi^{\phi_0}\|_2 \leq Cn^{-\beta}, \qquad \|\pi^\phi/\pi^{\phi_0} - 1\|_2 \leq \tilde{C}n^{-\tilde{\beta}}$$

for some constants $\beta, \tilde{\beta}, C, \tilde{C} > 0$. Since the Λ-Fleming-Viot stationary density is intractable in nearly all cases, it does not seem possible to extend this approach to obtain rates of posterior consistency.

3.3.2 A parametric approach to nonparametric inference

Consider a set of haplotype frequencies $\mathbf{n} \in \mathbb{N}^{|\mathcal{H}|}$ of size $n := \sum_{h\in\mathcal{H}} n_h$ generated by a Λ-coalescent with finite alleles mutation started from ψ_n.

Lemma 7. *The likelihood satisfies* $\mathbf{P}_n^{\Lambda}(\mathbf{n}) = \mathbf{P}_n^{\lambda_{3,3},\lambda_{4,4},\ldots,\lambda_{n,n}}(\mathbf{n})$. *That is,* Λ *is conditionally independent of* $\mathbf{n}$ *given* $\{\lambda_{p,p}\}_{p=3}^n$.

Proof. Let $q_{nn} = \sum_{p=1}^{n-1} \binom{n}{n-p+1}\lambda_{n,n-p+1}$ be the total merger rate of the Λ-coalescent with n blocks. It is well known that the Λ-coalescent likelihood is the unique solution to the recursion Möhle [2006]; Birkner and Blath [2008, 2009]:

$$\begin{aligned}\mathbf{P}_n^{\Lambda}(\mathbf{n}) =& \frac{\theta}{n\theta + q_{nn}} \sum_{h,h' \in \mathcal{H}} (n_{h'} - 1 + \delta_{hh'}) M_{h'h} \mathbf{P}_n^{\Lambda}(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{h'}) \\ &+ \frac{1}{n\theta + q_{nn}} \sum_{h:n_h \geq 2} \sum_{p=2}^{n_h} \binom{n}{p} \lambda_{n,p} \frac{n_h - p + 1}{n - p + 1} \mathbf{P}_{n-p+1}^{\Lambda}(\mathbf{n} - (p-1)\mathbf{e}_h), \quad (3.11)\end{aligned}$$

with boundary condition $\mathbf{P}_1^{\Lambda}(\mathbf{e}_h) = m(h)$. This recursion and its analogues were used in Chapter 2 to derive forwards transition probabilities for reverse time SMC. In this chapter the relevant observation is that repeated application of the recursion yields a closed system of linear equations for the likelihood. This is because all sample sizes on the RHS are equal to or smaller than the one on the LHS, and finiteness of $\mathcal{H}$ guarantees that only finitely many configurations of a given size are possible. This system is far too large to solve for all but very small sample sizes, but it is clear that the solution can depend on Λ only through the polynomial moments $\{\lambda_{q,p}\}_{p \leq q=2}^n$. Polynomial moments can be written as a linear combination of monomial moments:

$$\lambda_{q,p} = \sum_{j=0}^{q-p} \binom{q-p}{j} (-1)^j \lambda_{p+j,p+j}, \quad (3.12)$$

which means that only the monomial moments $\{\lambda_{p,p}\}_{p=2}^n$ are required. Since $\lambda_{2,2} = \int_0^1 \Lambda(dx) = 1$, the moments $\{\lambda_{k,k}\}_{k=3}^n$ are sufficient. □

Lemma 7 motivates the following definition:

Definition 11. Let $\sim_n$ be the equivalence relation on $\mathcal{M}_1([0,1])$ defined via

$$\Lambda_1 \sim_n \Lambda_2 \text{ if } \lambda_{p,p}^{(1)} = \lambda_{p,p}^{(2)} \text{ for } p \in \{3, \ldots, n\}$$

where $\lambda_{p,p}^{(i)} = \int_0^1 r^{p-2} \Lambda_i(dr)$. Let the equivalence classes of $\sim_n$ be called moment classes of order n.

In view of Lemma 7 it is natural to consider the problem of inferring Λ from $\mathbf{n}$ in the quotient space $\mathcal{M}_1([0,1])/\sim_n$, not in $\mathcal{M}_1([0,1])$. Moreover, requiring

all linear combinations of the form (3.12) to be non-negative guarantees a unique solution to the Hausdorff moment problem, so that each moment sequence bounded by 1 corresponds to some $\Lambda \in \mathcal{M}_1([0,1])$. Hence, the space $\mathcal{M}_1([0,1])/\sim_n$ can be parametrised by completely monotonic sequences of length $n-2$ with leading term $\lambda_{3,3} \leq 1$. This approach yields a compact, finite-dimensional parameter space which nevertheless captures all the signal in the data. Table 3.2 lists some moment sequences corresponding to popular families of Λ-measures.

Λ	δ_0	δ_1	$\text{Beta}(2-\alpha,\alpha)$	$U(0,1)$	$\frac{2}{2+\psi^2}\delta_0 + \frac{\psi^2}{2+\psi^2}\delta_\psi$	$c\delta_0 + \frac{(1-c)}{2} r dr$
$\lambda_{k,k}$	0	1	$\frac{(2-\alpha)_{(k-2)}}{(2)_{(k-2)}}$	$\frac{1}{k-1}$	$\frac{\psi^k}{2+\psi^2}$	$\frac{1-c}{2k}$

Table 3.2: Moment sequences of particular Λ-coalescents. Here $(a)_{(k)} := a(a+1)\dots(a+k-1)$ denotes the rising factorial.

Naturally, the prior Q ought to be chosen to yield tractable a tractable push-forward prior on finite moment sequences. These push-forward priors inherit posterior consistency whenever Q satisfies the conditions of Theorem 7 because finite moment sequences can be written as bounded functionals of Λ.

3.3.3 An example prior

This section provides an example family of priors which satisfy the consistency criteria of Theorem 5, and have tractable push-forward distributions on finite moment sequences.

Definition 12. Let $Q \in \mathcal{M}_1(\mathcal{M}_1([0,1]))$ be a prior distribution for Λ. Then the moments $\{\lambda_{p,p}\}_{p=3}^n$ have joint prior Q_n on the space of completely monotonic sequences of length $n-2$ given by

$$Q_n(\lambda_{3,3} \in dy_3, \dots, \lambda_{n,n} \in dy_n) := \int_{\mathcal{M}_1([0,1])} \prod_{p=3}^{n} \mathbb{1}_{\{dy_p\}} \left(\int_{(0,1]} r^{p-2} \Lambda(dr) \right) Q(d\Lambda). \quad (3.13)$$

The prior Q should to be chosen such that the RHS of (3.13) is tractable, and the following examples illustrate that such a choice is possible. Note also that this mapping of priors is not invertible: starting with any prior Q, computing Q_n and lifting it back to a distribution $\mathcal{M}_1([0,1])$ will yield a prior that is constant on equivalence classes of $\sim_n$.

Example 2. Fix $\eta > 0$ and $\alpha \in \mathcal{M}([\eta, 1])$ with finite mass and a strictly positive Lebesque density $\alpha(r)$. Suppose $\mathcal{D}_\eta$ satisfies the conditions of Definition 8, and in addition that every $\phi \in \mathcal{D}_\eta$ is continuous. Let $R(d\tau)$ be a probability measure on $(0, \infty)$ placing positive mass in all non-empty open sets. For $x \in [\eta, 1]$ and $\tau > 0$ let

$$q_{x,\tau}(r) := \frac{\mathbb{1}_{[\eta,1]}(r - x)h_\tau(r - x)}{h_\tau([\eta, 1])},$$

where h_τ is the Gaussian density on $\mathbb{R}$ with mean 0 and variance τ^{-1}.

Let $DP(\alpha)$ be the law of a Dirichlet process centred on α [Ferguson, 1973] and let Q be given by the Dirichlet process mixture distribution [Lo, 1984] with mixing distribution $DP(\alpha) \otimes R$ and mixture components $q_{x,\tau}$. See Section 3.2.2 for a specification of this prior.

The prior Q places full mass on equivalent densities bounded from above and away from 0 by construction, and satisfies (3.10) by the argument obtained by taking $b = b_0$ and $c = c_0$ in Section 3.2.2.

Next, the machinery of Regazzini et al. [2002] is used to give an explicit system of equations for the distribution function of Q_n under this choice of Q. Define the family of functions

$$g_p(x) := \int_\eta^1 r^p q_{x,\tau}(r)dr$$

for $p \in \mathbb{N}$ and $x \in [\eta, 1]$, as well as the vectors $\mathbf{g}_n(x) := (g_1(x), g_2(x), \ldots, g_n(x))$ and $\mathbf{s}_n := (s_1, \ldots, s_n) \in \mathbb{R}^n$. For brevity, for a measure ν and a function f let $\nu(f) := \int f d\nu$ whenever the integral exists.

Let γ_α be a Gamma random measure with parameter α, that is, a random finite measure on $[\eta, 1]$ such that for any measurable partition $\{A_1, \ldots, A_n\}$ the random variables $(\gamma_\alpha(A_1), \ldots, \gamma_\alpha(A_n))$ are independent and gamma distributed with common scale parameter 1 and respective shape parameters $\alpha(A_k)$. Let

$$h_n(\mathbf{s}_n; \mathbf{g}_n; \alpha) := \mathbb{E}\left[\exp\left(i\mathbf{s}_n \cdot \gamma_\alpha(\mathbf{g}_n)\right)\right]$$

be the characteristic function of $\gamma_\alpha(\mathbf{g}_n) := (\gamma_\alpha(g_1), \ldots, \gamma_\alpha(g_n))$. Note that $h_n(\mathbf{s}_n; \mathbf{g}_n; \alpha) = h_n(\mathbf{1}; \mathbf{s}_n \cdot \mathbf{g}_n; \alpha)$ and by [Regazzini et al., 2002, Proposition 10]

$$h_n(\mathbf{s}_n; \mathbf{g}_n; \alpha) = \exp\left(-\int_0^1 \log(1 - i\mathbf{s}_n \cdot \mathbf{g}_n)d\alpha\right). \tag{3.14}$$

Now let $F_n(\boldsymbol{\sigma}, \mathbf{g}_n, \alpha)$ be the joint distribution function of $F(\mathbf{g}_n) := (F(g_1), \ldots, F(g_n))$

under $DP(\alpha)$. The following trick was introduced in [Hannum et al., 1981, equation (2.9)]:

$$F_n(\boldsymbol{\sigma}, \mathbf{g}_n, \alpha) = F_n(\mathbf{0}, \gamma_\alpha(\mathbf{g}_n - \boldsymbol{\sigma}), \alpha)$$

for any $\boldsymbol{\sigma} \in \mathbb{R}^n$, so that it is sufficient to invert h_n at the origin to obtain F_n. This can be done using the multidimensional version of the Gurland inversion formula [Gurland, 1948, Theorem 3]:

Let $C_0, \ldots, C_n \in \mathbb{R}^{n+1}$ solve

$$\begin{aligned} C_n &= -1 \\ \sum_{k=0}^{n-r-1} \binom{n-r}{k} C_{r+k} &= 1 \text{ for } r \in \{0, \ldots, n-1\}. \end{aligned} \tag{3.15}$$

Then

$$\begin{aligned} (-1)^{n+1} 2^n F_n(\boldsymbol{\sigma}, \mathbf{g}_n, \alpha) &= C_0 \\ &+ \sum_{k=1}^{n} \frac{C_k}{(\pi i)^k} \sum_{1 \le j_1 < \ldots < j_k \le n} \int_0^\infty \cdots \int_0^\infty \frac{h_k(\mathbf{s}_k; g_{j_1} - \sigma_{j_1}, \ldots, g_{j_k} - \sigma_{j_k}; \alpha)}{s_1 \times \cdots \times s_k} d\mathbf{s}_k. \end{aligned}$$

The characteristic functions h_k and the constants C_k can be computed from (3.14) and (3.15) respectively, so that the RHS can be evaluated numerically for practical applications. Numerical methods are discussed in [Regazzini et al., 2002, Section 6].

Finally, I will demonstrate that the restrictive assumptions of Theorem 5 still allow inference for broad classes of moment sequences with arbitrarily small approximation errors. Let $\beta(r)$ be any non-negative probability density on $[0, 1]$, and define the truncation $\bar{\beta}(r) := \kappa(\eta, c, K)^{-1}(\beta(r) \vee c \wedge K)$, where κ is the normalising constant

$$\kappa(\eta, c, K) := \int_\eta^1 \beta(r) - (\beta(r) - K)^+ + (c - \beta(r))^+ dr.$$

Note that $Q(\phi \in \mathcal{D}_\eta : \|\phi - \bar{\beta}\|_\infty < \delta) > 0$ for any $\delta > 0$, and fix such a ϕ. Now consider the error on the k^{th} moment:

$$\begin{aligned} &\Big| \int_0^1 r^k \beta(r) dr - \int_\eta^1 r^k \phi(r) dr \Big| \le \\ &\beta((0, \eta)) + (1 - \eta)\delta + \frac{(\kappa(\eta, c, K) - 1)\beta([\eta, 1]) + c(1 - \eta)}{\kappa(\eta, c, K)} + \int_\eta^1 \frac{(\beta(r) - K)^+}{\kappa(\eta, c, K)} dr. \end{aligned}$$

Each term on the RHS can be made small by choosing η, c and δ sufficiently small,

and K sufficiently large because $\kappa \to 1$ as $\eta \to 0$, $c \to 0$ and $K \to \infty$, and

$$\int_\eta^1 (\beta(r) - K)^+ dr = 1 - \int_\eta^1 \beta(r) \wedge K dr \to \beta((0, \eta))$$

as $K \to \infty$ by the Monotone Convergence Theorem. A further approximation step also enables consideration of atoms by choosing β which places all of its mass in neighbourhoods of the desired locations for atoms. Hence it is possible to ensure the support of Q extends arbitrarily close to any desired moment sequences despite the restrictive assumptions on $\mathcal{D}_\eta$ in Theorem 7.

3.3.4 Robust bounds on functionals of Λ

Having established consistency criteria for the posterior and a finite parametrisation via $n-2$ leading moments, I now turn to what can be said about Λ based on inferring the parameters. It would be ideal if the diameter of moment classes shrunk with increasing n, as then it would be possible to fix a representative $\Lambda \in \mathcal{M}_1([0,1])$ with specified $n-2$ leading moments and control the remaining within-moment-class error. In Theorem 8 I show that such shrinking does not happen, and devote the remainder of the section to presenting quantities which can be controlled based on $n-2$ moments alone. I begin by recalling some standard results from the theory of orthogonal polynomials.

Definition 13. Suppose n is odd. Let $u := \frac{n-3}{2}$ and $\{\phi_k\}_{k=0}^u$ be the first $u+1$ Λ-orthogonal polynomials. Let $\{\xi_p\}_{p=1}^u$ be the zeros of ϕ_u.

Remark 11. It is a standard result that $\{\phi_p\}_{p=0}^{u-1}$ and $\{\xi_p\}_{p=1}^u$ are constant within moment classes of order at least n.

The following bounds on Λ in terms of its leading $n-2$ moments are classical:

Lemma 8 (Chebyshev-Markov-Stieltjes (CMS) inequalities)**.** *Define*

$$\rho_{u-1}(z) := \left(\sum_{p=0}^{u-1} |\phi_p(z)|^2 \right)^{-1}.$$

Then the following inequalities are sharp:

$$\Lambda([0, \xi_j]) \leq \sum_{p=1}^{j} \rho_{u-1}(\xi_p) \leq \Lambda([0, \xi_{j+1})) \text{ for } j \in [u],$$

where $\xi_{u+1} := 1$.

Theorem 8. *For any $n \in \mathbb{N}$ and any completely monotonic sequence of moments $\{\lambda_{p,p}\}_{p=3}^{n}$ with $\lambda_{3,3} \leq 1$ there exist uncountably many measures $\Lambda \in \mathcal{M}_1([0,1])$, all with leading moments $\{\lambda_{p,p}\}_{p=3}^{n}$ and all satisfying the CMS inequalities, such that any pair, Λ_x and Λ_y, satisfy $d_{TV}(\Lambda_x, \Lambda_y) = 2$, where d_{TV} is the total variation norm.*

Proof. It will be convenient to write the CMS inequalities in the following, equivalent form:

$$\begin{aligned}
&0 \leq \Lambda([0, \xi_1)) \leq \rho_{u-1}(\xi_1) \\
&0 \leq \Lambda([\xi_j, \xi_{j+1})) \leq \rho_{u-1}(\xi_j) + \rho_{u-1}(\xi_{j+1}) \text{ for } j \in [u-1] \\
&0 \leq \Lambda([\xi_u, 1]) \leq \rho_{u-1}(\xi_u)
\end{aligned}$$

where the last inequality follows from the fact that $\sum_{p=1}^{u} \rho_{u-1}(\xi_p) = 1$. This equality holds because $\sum_{p=1}^{u} \rho_{u-1}(\xi_p)$ is the sum of all order u Gauss quadrature weights, or equivalently the quadrature applied to the constant function 1, which is a polynomial of degree 0. The equality follows by recalling that Gauss quadrature is exact for polynomials of order up to $2u-1$.

Now let the measures Λ_x and Λ_y be described by sequences of $(u+1)$ weights $(x_0, x_1, \ldots, x_u)$ and $(y_0, y_1, \ldots, y_u)$, with the j^{th} weight denoting the mass that the corresponding measure places in the interval $[\xi_j, \xi_{j+1})$, with obvious adjustments for the rightmost boundary terms.

For brevity let $\zeta_j := \rho_{u-1}(\xi_j)$. Suppose first that u is odd, and let the vectors of weights be given as

$$\begin{aligned}
(x_0, x_1, x_2, x_3, x_4, \ldots, x_{u-1}, x_u) &= (\zeta_1, 0, \zeta_2 + \zeta_3, 0, \zeta_4 + \zeta_5, \ldots, 0, \zeta_u) \\
(y_0, y_1, y_2, y_3, y_4, \ldots, y_{u-1}, y_u) &= (0, \zeta_1 + \zeta_2, 0, \zeta_3 + \zeta_4, 0, \ldots, \zeta_{u-1} + \zeta_u, 0)
\end{aligned}$$

Both measures have total mass $\sum_{j=1}^{u} \zeta_j = 1$, and the interlacing masses have no overlap so $d_{TV}(\Lambda_x, \Lambda_y) = 2$. The case where m is even is similar. □

Remark 12. The same result holds in Kullback-Leibler divergence due to Pinsker's inequality:

$$d_{TV}(P, Q) \leq \sqrt{\frac{1}{2} K(P, Q)},$$

for probability measures P and Q, so that $K(\Lambda_x, \Lambda_y) \geq 8$ for Λ_x and Λ_y as in Theorem 8.

Despite this seemingly disappointing result, it is possible to make some con-

clusions about Λ based on $n-2$ moments. For example, the Kingman hypothesis can be tested in a robust way by checking whether the vector $(0,0,\ldots,0)$ lies in a desired credibility region of the posterior $\mathbf{P}_n^{\Lambda}(\cdot|\mathbf{n})$, and the plausibility of any other Λ of interest can be assessed similarly. More generally, it is possible to extremise a certain class of functionals subject to moment constraints obtained from a credibility region to obtain robust bounds for quantities of interest. I begin by recalling some relevant definitions.

Definition 14. For $p, n \in \mathbb{N}$, $\mathbb{R}_+$-valued constants $\{c_q\}_{q=1}^p$, a sequence $\{i_q\}_{q=1}^p$ of $\{3,\ldots,n\}$-valued indices and a binary sequence $\{j_q\}_{q=1}^p$ of zeros and ones, let

$$\mathcal{C}_p := \left\{\Lambda \in \mathcal{M}_1([0,1]) : (-1)^{j_q}\lambda_{i_q,i_q} \leq c_q \text{ for } q \in [p]\right\} \tag{3.16}$$

be a subset of $\mathcal{M}_1([0,1])$ with leading $n-2$ moments in a desired region specified by p linear inequalities. Let $\text{ext}(\mathcal{C}_p)$ be the extremal points of $\mathcal{C}_p$, i.e. those which cannot be written as non-trivial convex combinations of elements in $\mathcal{C}_p$, and

$$\mathcal{C}_D := \left\{\nu \in \mathcal{C}_p : \nu = \sum_{r=1}^{q} w_r \delta_{x_r} \text{ where } 1 \leq q \leq p+1, w_r \geq 0, x_r \in [0,1] \text{ and } \sum_{r=1}^{q} w_r = 1\right\}$$

be the set of discrete probability measures on $[0,1]$ with at most $p+1$ atoms. Here $\mathcal{C}_p$ should be thought of as a convex envelope expressed using finitely many linear constraints and containing a desired credibility region of finite, completely monotonic moment sequences. I postpone discussion of how an approximate credibility region can be obtained to the next section, and simply assume one is available.

Example 3. The extremal points of $\mathcal{M}_1([0,1])$ are the Dirac measures:

$$\text{ext}\left(\mathcal{M}_1([0,1])\right) = \{\delta_x : x \in [0,1]\}.$$

Definition 15. The functional $F : \mathcal{C} \mapsto \overline{\mathbb{R}}$ is measure-affine if, for every $\nu \in \mathcal{C}$ and $p \in \mathcal{M}_1(\text{ext}(\mathcal{C}))$ such that $\nu(E) = \int_{\text{ext}(\mathcal{C})} \gamma(E) p(d\gamma)$ for every $E \in \mathcal{B}([0,1])$, F is p-integrable and

$$F(\nu) = \int_{\text{ext}(\mathcal{C})} F(\gamma) p(d\gamma).$$

Intuitively, ν is a barycentre of $\mathcal{C}$ with weights on extremal points given by p, and F is measure-affine if it commutes with the operation of expressing ν as the weighted sum of extremal points. If $\mathcal{C}$ consis of finitely many points, this definition

coincides with the usual definition of affine functions.

The following two results are due to Winkler [1988]:

Lemma 9. *If $q : [0,1] \mapsto \overline{\mathbb{R}}$ is bounded on at least one side then $F : \nu \mapsto \mathbb{E}_\nu[q]$ is measure-affine.*

Lemma 10. *Let $\mathcal{C}_p$ be as in Definition 14 and $F : \mathcal{C}_p \mapsto \overline{\mathbb{R}}$ be measure-affine. Then*

$$\inf_{\nu \in \mathcal{C}_p} F(\nu) = \inf_{\nu \in \mathcal{C}_D} F(\nu) \tag{3.17}$$

$$\sup_{\nu \in \mathcal{C}_p} F(\nu) = \sup_{\nu \in \mathcal{C}_D} F(\nu). \tag{3.18}$$

Remark 13. The importance of Lemma 10 is that the optimisation problems on the RHS of (3.17) and (3.18) are finite-dimensional and can be solved numerically. Hence tight bounds for measure-affine functionals $F(\Lambda)$ over credibility regions can be computed in an assumption-free manner.

In order to specify $\mathcal{C}_p$ it remains to be able to approximate the posterior, which will be achieved via MCMC. This will be detailed in the next section. Before that, I conclude this section with a simple example computation.

Example 4. Suppose a posterior credible region is specified via two linear constraints as

$$\mathcal{C}_2 = \{\Lambda \in \mathcal{M}_1([0,1]) : \lambda_3 \leq 0.5 \text{ and } 0.3 \leq \lambda_4\},$$

and that the measure-affine functional of interest is the exponential:

$$F(\Lambda) := \int_0^1 e^{-r} \Lambda(dr).$$

Then the finite dimensional subspace $\mathcal{C}_D \subset \mathcal{C}_2$ consists of discrete probability measures on $[0,1]$ with at most three atoms:

$$\mathcal{C}_D = \left\{ \Lambda \in \mathcal{C}_2 : \Lambda = \sum_{k=1}^{p} w_k \delta_{r_k} \text{ where } 1 \leq p \leq 3, w_k \geq 0, r_k \in [0,1] \text{ and } \sum_{k=1}^{p} w_k = 1 \right\}.$$

This yields three maximisation/minimisation problems, one corresponding to each number of atoms, though in practice only the largest needs to be solved since the two others can be recovered as special cases. In this case, the constrained optimisation

problem is

$$\begin{aligned}
\text{Maximise/Minimise: } & ae^{-x} + be^{-y} + (1-a-b)e^{-z} \\
\text{Subject to: } & ax + by + (1-a-b)z \leq 0.5 \\
& -ax^2 - by^2 - (1-a-b)z^2 \leq -0.3 \\
& a, b \leq 1 \\
& -a, -b \leq 0 \\
& a + b \leq 1 \\
& x, y, z \leq 1 \\
& -x, -y, -z \leq 0.
\end{aligned}$$

Numerical evaluation in Mathematica yields the bounds $F(\Lambda) \in (0.620, 0.810)$.

3.3.5 A simulation study

Efficient methods for approximating the Λ-coalescent likelihood pointwise were presented in Section 2.3 and can be readily adapted to the form developed by Beaumont [2003] for time series data. These likelihood estimators can then be used in the pseudo-marginal Metropolis-Hastings algorithm [Beaumont, 2003; Andrieu and Roberts, 2009], in which the likelihood evaluations required in a standard Metropolis-Hastings algorithm are replaced with unbiased estimators. The resulting algorithm still targets the correct posterior and inherits the efficient exploration of parameter space of MCMC methods. Thus it is well-suited to high-dimensional situations with intractable likelihood.

Let $S(n)$ denote the space of completely monotonic sequences of length $n-2$, and for $\boldsymbol{\lambda} \in S(n)$ let $L(\boldsymbol{\lambda}; \mathbf{n})$ be the likelihood function, and $\widehat{L}(\boldsymbol{\lambda}; \mathbf{n})$ be an unbiased estimator, e.g. one obtained from SMC, as is often the case in practice. The pseudo-marginal Metropolis-Hastings algorithm is presented in Algorithm 2 below.

Algorithm 2 returns a sample of moment sequences S, whose limiting distribution is the posterior. A credible region C can be approximated from MCMC output, and used to form $\mathcal{C}_p$ as per (3.16). Measure-affine quantities of interest can then be maximised or minimised using finite computation by making use of Lemma 10.

By way of demonstration I focus on assessing the Kingman hypothesis, $\Lambda = \delta_0$, which can be robustly evaluated based upon whether or not $\lambda_{3,3} = 0$. The type space consists of 10 binary loci, or 2^{10} types, with mutations flipping a uniformly chosen locus. The total mutation rate is $\theta = 0.1$. Samples of 20 lineages

Algorithm 2 Pseudo-marginal Metropolis-Hastings for finite moment sequences

Require: Prior P_n, observation $\mathbf{n}$, transition kernel $K : S(n) \times S(n) \mapsto \mathbb{R}_+$ and $N \in \mathbb{N}$.

Initialise sample $S \leftarrow \emptyset$ and moment sequence $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}_0$.
Compute likelihood estimator $\widehat{L}(\boldsymbol{\lambda}; \mathbf{n})$.
for j = 1, ..., N **do**
    Sample $\boldsymbol{\lambda}' \sim K(\boldsymbol{\lambda}, \cdot)$.
    Compute likelihood estimator $\widehat{L}(\boldsymbol{\lambda}'; \mathbf{n})$.
    Set $a \leftarrow 1 \wedge \frac{K(\boldsymbol{\lambda}', \boldsymbol{\lambda})\widehat{L}(\boldsymbol{\lambda}'; \mathbf{n})P_n(\boldsymbol{\lambda}')}{K(\boldsymbol{\lambda}, \boldsymbol{\lambda}')\widehat{L}(\boldsymbol{\lambda}; \mathbf{n})P_n(\boldsymbol{\lambda})}$.
    Sample $U \sim U(0, 1)$.
    if $U < a$ **then**
        Set $S \leftarrow S \cup \{\boldsymbol{\lambda}'\}$, $\widehat{L}(\boldsymbol{\lambda}) \leftarrow \widehat{L}(\boldsymbol{\lambda}')$ and $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda}'$.
    else
        Set $S \leftarrow S \cup \{\boldsymbol{\lambda}\}$.
    end if
end for

return S

were generated at each of five time points from both the Kingman ($\lambda_{3,3} = 0$) and Bolthausen-Sznitman ($\Lambda = U(0, 1)$, $\lambda_{3,3} = 0.5$) coalescents. These are summarised in Table 3.3. Both data sets come from independent simulations, and are sampled from a population at stationarity. The Kingman coalescent is a classical model of genetic ancestry, while the Bolthausen-Sznitman coalescent has recently been suggested as a ancestral model for influenza and HIV [Neher and Hallatschek, 2013].

Time	Bolthausen-Sznitman	Kingman
0.0	20 x 1001001111	20 x 0010000000
0.5	19 x 1001001111	15 x 0010000000
	1 x 1101001111	5 x 0000000000
1.0	20 x 1001001111	8 x 0010000000
		6 x 0000000000
		6 x 0010001000
1.5	19 x 1001001111	10 x 0010000000
	1 x 1001101111	6 x 0000000000
		4 x 0010001000
2.0	19 x 1001001111	16 x 0010000000
	1 x 1001001110	4 x 0010001000

Table 3.3: Observed sequences sampled from the two models.

Let $\eta = 10^{-6}$ and let the prior for the density of Λ on $[\eta, 1]$ be a Dirichlet process mixture model of truncated Gaussian kernels [Ferguson, 1973; Lo, 1984]. The base measure is the uniform measure on $[\eta, 1]$, with total mass scaled to 0.1. Finally,

the prior for $\tau^{-1/2}$, the standard deviations of the truncated Gaussian kernels was chosen to be the Beta$(1.0, 3.0)$ distribution on $[0, 1]$. Truncating the maximal standard deviation at 1 excludes some very flat densities from the support of the prior, but the standard normal density is already very flat across $[\eta, 1]$ and the truncation was found to yield substantial gains in speed of convergence of algorithms. Note that neither data generating model lies in the support of this prior, but both can be well approximated by members of the support. The choice of hyperparameters was made because it yields a relatively flat marginal prior for $\lambda_{3,3}$, the quantity of interest (c.f. Figure 3.3), and the prior satisfies the requirements of the consistency result in Theorem 7.

I make use of the Sethumaran stick-breaking construction of the Dirichlet process [Sethuraman, 1994] and truncate after the first four atoms. See Section 1.4 for a brief specification of the stick-breaking construction. For the chosen of base measure and concentration parameter this results in a total variation truncation error of order $400e^{-30} \approx 3.7 \times 10^{-11}$ [Ishwaran and James, 2001]. Any truncation error could be avoided by pushing forward the prior directly onto the space of moment sequences as illustrated in Section 3.3.3. The cost is a more computationally expensive prior to sample and evaluate, as well as a higher dimensional parameter space consisting of 98 moments for these data sets. This strategy is not investigated further in this thesis.

The four atom truncation results in 11 parameters: four locations and standard deviations of truncated Gaussian kernels, and three stick break points. The fourth break point is set to fulfil the constraint of the weights summing to 1. Updates to these four parameters are proposed using a truncated Gaussian random walk on $[\eta, 1]^4 \times [0, 1]^4 \times [0, 1]^3$ with covariance matrix $0.0025\,\mathrm{Id}$. This scaling was found to result in a reasonable balance of acceptance probability and jump size for the first moment $\lambda_{3,3}$.

The likelihoods required for computing the acceptance probability a are approximated using a straightforward adaptation of the optimised importance sampling method of Section 2.3 to the time series setting of [Beaumont, 2003], but it is necessary to specify the number of particles in the approximation. More particles will result in more accurate approximations, but at greater computational cost. In [Doucet et al., 2015] the authors show that tuning the variance of the log likelihood estimator to 1.44 results in efficient algorithms under a wide range of assumptions. Preliminary simulations showed this was achieved in this setting by choosing 75 particles for Bolthausen-Sznitman data, and 180 particles for Kingman data.

Remark 14. In the context of real data, when the true data generating parameters

are not known, optimising the number of particles using trial runs may require an infeasible amount of computation. In practice, adaptive algorithms, which optimise parameters online, can be used to circumvent this problem.

It is well known that the standard, exact pseudo-marginal algorithm suffers from "sticking" behaviour, where an unusually high likelihood estimator prevents the algorithm from moving for a macroscopic number of steps [Andrieu and Roberts, 2009]. The usual solution is to use a noisy version of the algorithm, in which the likelihood estimator is recomputed at each stage. This doubles the number of required likelihood evaluations and biases the algorithm into an incorrect stationary distribution, but can greatly reduce the variance of estimates. A comparison of independent runs of both the exact and noisy versions of the pseudo-marginal algorithm is presented in Figure 3.2. In addition, I also investigated the effect of delayed acceptance acceleration [Christen and Fox, 2005], in which proposed moves are first subjected to an accept-reject decision based on an approximate likelihood function that is cheap to compute. Only samples which are accepted at this first stage are subjected to an accept-reject decision based on the full likelihood estimates, or more specifically a slight modification to ensure that the delayed acceptance mechanism does not affect the stationary distribution of the algorithm. In the Λ-coalescent setting approximate likelihoods are readily available in the form of Product of Approximate Conditionals (or PAC) methods (c.f. Section 2.4), which were used to implement delayed acceptance chains.

Figure 3.2 shows trace plots of 20 000 steps from the four algorithms introduced above. The exact pseudo-marginal algorithm exhibits sticking behaviour as might be expected, but it is surprising to see that the noisy algorithm does not completely eliminate it. I conjecture that the remaining stickiness in the noisy trace plot is due to multiple, narrow modes in the 11 dimensional posterior. It is also clear that the bias in the noisy algorithm is confounding the signal in the data, as the traces are much more intermixed than those of the exact algorithm.

Both the noisy and exact pseudo-marginal algorithm are very computationally expensive to run, particularly for the Kingman data set due to the larger number of particles used to estimate likelihoods. Delayed acceptance acceleration reduces these run times as expected, particularly for the Kingman case. Both delayed acceptance algorithms also suffer from sticking, and show less clear separation of the traces than the exact algorithm. They also look very similar to each other.

Since it appears to be difficult to eliminate sticking behaviour in this case, leveraged the speed up obtained by making use of delayed acceptance and ran a further exact, delayed acceptance pseudo-marginal algorithm for 200 000 steps. A

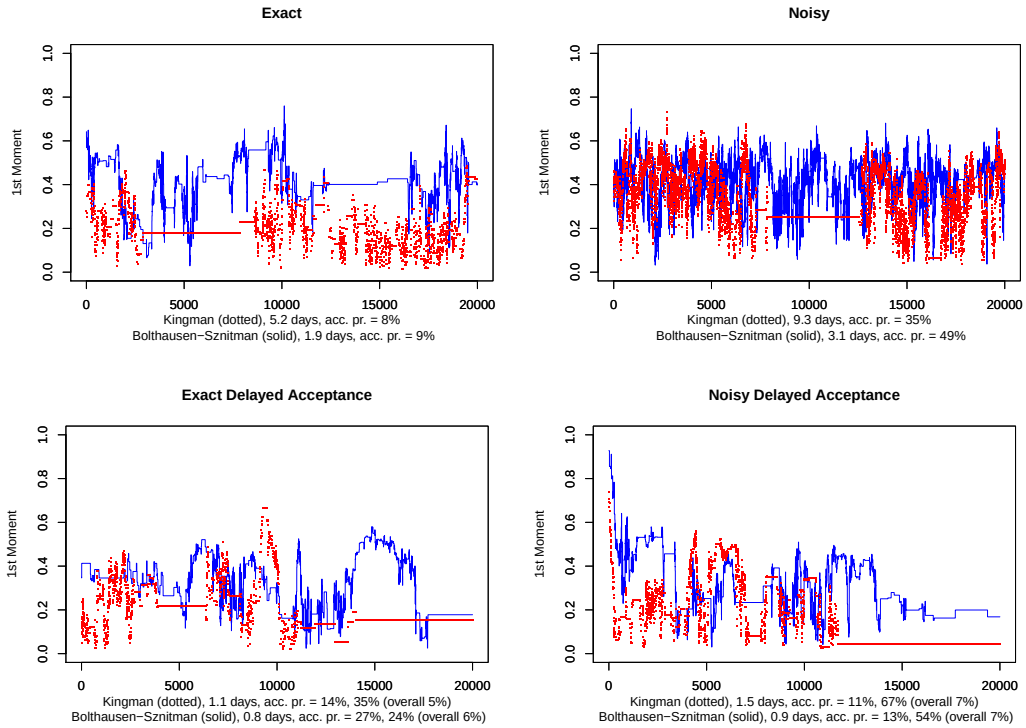


Figure 3.2: Trace plot of the pseudo-marginal algorithm (top left), the noisy algorithm (top right) and corresponding delayed acceptance algorithms (bottom row). Also shown are computation times (on a mid-range Toshiba laptop with an Intel i5 processor) and acceptance probabilities. Delayed acceptance runs show acceptance probabilities for both stages, as well as an overall probability. All runs are independent and initialised from the prior.

trace plot is shown in Figure 3.3. Sticking behaviour is still present, but on a much shorter scale relative to the run length. Run times are comparable to the noisy algorithm without delayed acceptance, and the Bolthausen-Sznitman trace is again clearly centred at a higher level than the Kingman trace. The output of this long run was thinned by a factor of 4 000 to reduce the effect of sticking to obtain 50 samples of first moments, which were used to plot the histograms in Figure 3.3.

It is clear from both the trace plots and histograms in Figure 3.3 that the run length is still not sufficient for fully converged estimates. However, both plots already show a clear shift of posterior modes toward the values generating the data. The red histogram is consistent with the Kingman coalescent, while the blue one is consistent with the Bolthausen-Sznitman coalescent. Moreover, approximate 95% credible intervals are $\lambda_{3,3} \in [0.1, 0.6]$ for the Bolthausen-Sznitman posterior, and $\lambda_{3,3} \in [\eta, 0.5]$ for the Kingman posterior. This suggests the relatively short time

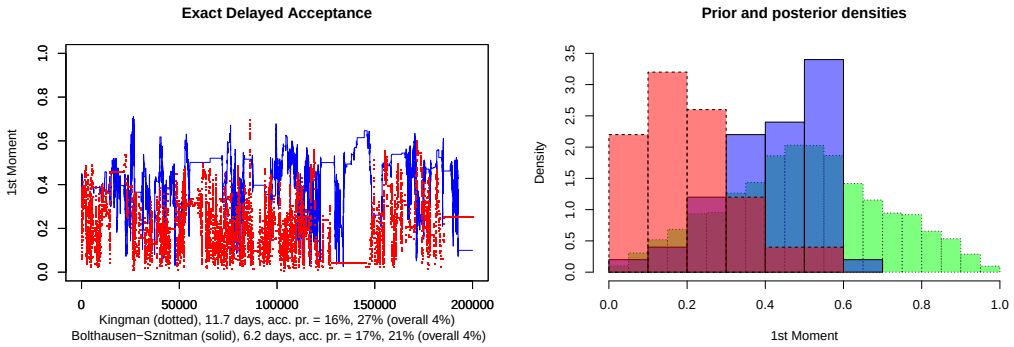


Figure 3.3: (Left) Trace plot of the delayed acceptance exact pseudo-marginal algorithm, along with computation times (on a mid-range Toshiba laptop with an Intel i5 processor) and acceptance probabilities for both stages as well as overall. Both runs are independent and initialised from the prior. (Right) Histograms of both Kingman (dashed) and Bolthausen-Sznitman (solid) posteriors estimated from 50 MCMC samples obtained by thinning the runs shown on the left. The estimated prior density is shown (dotted), based on 10 000 independent samples from the prior.

series is nevertheless sufficiently informative to reject the incorrect model in both cases.

3.3.6 Discussion

This section has presented a robust framework for Bayesian non-parametric inference for Λ-coalescent processes with time series data, and studied the feasibility of implementable families of algorithms for practical inference. Posterior consistency for time series data was obtained under verifiable conditions on the prior, provided that the Λ-measure is identifiable in the sense that the mapping $\Lambda \mapsto P_\delta^\Lambda$ is injective. Identifiability seems difficult to verify in practice, but the results from a numerical simulation using time series data were promising even without an identifiability proof. In contrast, as seen in Example 1, lack of consistency can lead to very low statistical power and high sensitivity of inference both to confounding parameters, such as mutation rate, and the observed allele frequencies. A theoretical guarantee of consistency is crucial as expressions for statistical power rely on intractable stationary distributions and transition densities of Λ-Fleming-Viot jump-diffusions, making the reliability of experiments without time series data very difficult to evaluate.

Efficient methods for importance sampling Λ-coalescent trees are available as outlined in Section 2.3, and references therein, and these can be used to generalise

the pseudo-marginal MCMC algorithms of [Beaumont, 2003] for temporally spaced data. The consistency conditions of Theorem 7 on the prior are sufficiently mild to permit the use of Dirichlet process mixture model priors, which can be readily truncated for implementable algorithms. Alternatively, parametrising the inference problem via truncated moment sequences leads to implementable algorithms with no discretisation or truncation error. This work provides a strong indication that time series data, and accompanying inference methods such as the one outlined above, should be adopted as standard whenever the coalescent generating the data cannot be assumed to be known.

Generalising of consistency result within the Λ-Fleming-Viot process class to include unknown drift, which can be used to model e.g. mutation, recombination and selection, as well as more general Λ-measures is of great interest. However, it is difficult for a number of reasons. Firstly, relaxing conditions on Λ near 0 while ensuring the integral in (3.1) remains finite is challenging. Likewise, it is well known that equivalent changes of measure for Lévy processes necessitate equivalent Lévy measures (see e.g. [Sato, 1999], Theorem 33.1), and this is also the condition needed for the jump-diffusions considered in [Cheridito et al., 2005]. The way in which the drift can be transformed while maintaining absolute continuity in [Cheridito et al., 2005] is also restrictive, and depends on the diffusion coefficient and Lévy compensator. Finally, any difference in diffusion coefficients will obviously destroy absolute continuity outright, so if there were an atom $\Lambda(\{0\}) > 0$, its size would have to be known with certainty.

It would also be of great interest to obtain contraction rates of the posterior under verifiable conditions. Obtaining rates is a challenging problem in non-IID Bayesian non-parametric inference, and existing results by Gugushvili et al. [2015] for compound Poisson processes and Nickl and Söhl [2015] for scalar diffusions do not seem generalisable. A different approach by Nguyen [2013] for mixing measures of infinite mixture models could present a promising directions of future work by viewing the Λ-coalescent tree as a mixture of merger events, but adaptation into the present setting is a formidable task and is beyond the scope of this paper.

The method of parametrising the unknown Λ-measure by its first $n-2$ moments when the data set is of size $n \in \mathbb{N}$ reflects the limited amount of signal in finite data. More precisely, the likelihood given a sample of size $n \in \mathbb{N}$ is constant within moment classes of order n, so that any variation in the posterior within these moment classes is due solely to the prior. Hence this parametrisation can be seen as regularising an under-determined inference problem in an infinite dimensional space by identifying an appropriate, data-driven, finite dimensional quotient space

in which to conduct inference. I believe this approach to have more broad applicability in non-parametric statistics as well as an alternative to direct regularisation by a prior in the infinite dimensional space, or to approximate projections onto finite dimensional subspaces [Cui et al., 2014].

The algorithms used to approximate the posterior and maximise/minimise quantities of interest given the posterior are highly computationally intensive, and this approach cannot be expected to be competitive with well-chosen parametric families when the number of observed lineages or loci is large. However, the simulations in Section 3.3.5 demonstrate that the assumption-free framework can be used to empirically evaluate the modelling fit of parametric families given moderately sized pilot data, for instance by ensuring that the family contains a candidate Λ which matches the MAP estimators of some small number of moments. Such parametric families can then be confidently used to process larger data sets. The pseudo-marginal method can also be adapted to incorporate unknown mutation parameters, recombination and other forces not considered in this paper, albeit at the cost of greater computational cost and lower parameter identifiability. This cost can be alleviated to a large extent by modern GPU and cluster computing approaches, because the importance sampling algorithm used to estimate likelihoods is readily parallelisable. For example, up to 500 fold speed up was reported by Lee et al. [2010] when computations were parallelised on GPUs instead of being run in serial on CPUs. Such gains in computation speed would make the algorithms employed in Section 3.3.5 practical for many realistic genetic data sets.

Chapter 4

Discussion

In this thesis I have extended Bayesian consistency results from unit diffusions with unknown drift [van der Meulen and van Zanten, 2013; Gugushvili and Spreij, 2014] to more general unit jump diffusions with unknown drift and Lévy measure (Section 3.2), as well as to Λ-coalescent processes (Section 3.3). Both extensions were subject to an identifiability assumption, which is easy to verify in the diffusion case but intractable in the jump diffusion and Λ-coalescent cases, due to the intractability of transition and stationary densities. I have also introduced the reverse-time sequential Monte Carlo framework, building on the work of Stephens and Donnelly [2000], as a method for sampling certain classes of complex distributions and rare events in an asymptotically exact way (Chapter 2). Both results were applied to complex coalescent models of population genetics, resulting in computationally feasible, theoretically sound inference algorithms.

A fundamental assumption underlying this work is that consistency, asymptotic exactness of algorithms and exact, interpretable models are of value. This assumption has yet to be examined, and the answer is not obvious. In fact, many modern algorithms are moving away from such restrictive criteria. Approximate Bayesian computation (ABC) [Beaumont, 2010] replaces likelihood evaluations with a noisy, binary approximation based on simulated summary statistics from the model. The noisy pseudo-marginal method introduced in Section 3.3.5 samples a biased target to improve mixing, in what can be seen as a bias-variance trade-off. Composite likelihoods [Varin et al., 2011] correspond to inference based on a misspecified model. A comprehensive list of methods would be too long to list here, but the common theme is abandoning consistency, unbiasedness or some notion of exactness or error control for a gain in computational efficiency. Likewise, the necessity of Bayesian posterior consistency is not always clear: it may be perfectly adequate

in practice for the posterior to concentrate on a sufficiently narrow neighbourhood close to the truth, rather than the truth exactly. For example, the two-parameter Poisson-Dirichlet prior is known to give rise to an inconsistent posterior for certain combinations of parameters [James, 2008], but is still widely used, for instance in Bayesian statistics, population genetics and physics [Pitman, 2006].

In population genetics specifically, the intractability of likelihood functions arising from coalescent models has motivated many approximations. Indeed, both ABC [Tavaré et al., 1997; Beaumont et al., 2002] and the (noisy) pseudo-marginal method [Beaumont, 2003] were originally motivated by problems in population genetic inference. In addition, the product of approximate conditionals (PAC) method (Section 2.4) is used very widely in genetics due to its scalability, despite the fact that there is no known error bound between PAC and coalescent likelihoods. A further concern is that the copying models typically used in PAC inference do not give rise to exchangeable data. Despite these limitations, PAC-based inference has been very successful in practice [Li and Stephens, 2003; Crawford et al., 2004; Stephens and Scheet, 2005; Li and Abecasis, 2006; Scheet and Stephens, 2006; Gay et al., 2007; Marchini et al., 2007; Hellenthal et al., 2008, 2009; Howie et al., 2009; Yin et al., 2009].

In most cases a model is at best a cartoon of the phenomenon under study, so it is perhaps not surprising that misspecified models or inexact methods can lead to useful inferences. However, careful analysis of models and methods is still needed to avoid pitfalls. For example, it is known that carelessly adopting the noisy version of a pseudo-marginal algorithm to speed up mixing of MCMC can turn an ergodic chain into a transient one [Medina-Aguayo et al., 2015]. This behaviour might be very computationally expensive to detect reliably without reliable, exact methodology for benchmarking, and an understanding of when bad behaviour should be expected can only been reached through careful analysis of precisely specified models.

Likewise, the inconsistency results and example calculations in Section 3.3 demonstrate a pitfall in naively sampling contemporaneous DNA when the reproduction dynamics of the population are uncertain. For decades population genetics has been focused on the domain of attraction of Kingman's coalescent, where these dynamics are assumed known, and hence contemporaneous samples presented no danger. In the Λ-coalescent world this received wisdom is false, and time series data is necessary. Still more work is needed in determining sufficient numbers of time points and samples, for instance by identifying concentration rates of the posterior, but it is clear that these questions cannot be addressed by checks based on simulation, and a failure to address them can lead to imprecise inference and compromised

decision making. For a model used in fields such as conservation ecology, immunology, cancer research and public health, the consequences can be dramatic. The PAC simulations in Section 2.4 further demonstrate that heuristic inference without a theoretically sound benchmark can lead to biased results and false confidence.

Ultimately, it seems unlikely that the challenges of inference from complex models and challenging data could be tackled without a grounding in statistical theory and algorithms with theoretical guarantees, even if the theorems and algorithms cannot be applied to the most challenging applications directly. They can certainly motivate and evaluate heuristics, and thus lead to improved, more reliable conclusions in applications.

Bibliography

K. K. Aase and P. Guttorp. Estimation in models for security prices. *Scand. Actuarial J.*, pages 211–224, 1987.

Y. Aït-Sahalia. Closed-form likelihood expansions for multivariate diffusions. *Ann. Stat.*, 36(2):906–937, 2008.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Stat.*, 37(2):697–725, 2009.

D. Applebaum. *Lévy processes and stochastic calculus*. Cambridge studies in advanced mathematics. Cambridge University Press, 2004.

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

S. P. Au, A. H. Haddad, and V. H. Poor. A state estimation algorithm for linear systems driven simultaneously by Weiner and Poisson processes. *IEEE Trans. Aut. Control*, Ac-27(3):617–626, 1982.

I. Bardhan and X. Chao. Pricing options on securities with discontinuous returns. *Stoch. Proc. Appl.*, 48(1):123–137, 1993.

O. Barndorff-Nielsen. Exponentially decreasing distributions for the logarithm of particle size. *Proc. R. Soc. London A*, 353:401–419, 1977.

O. Barndorff-Nielsen. Hyperbolic distributions and distributions on hyperbolae. *Scand. J. Statist.*, 5:151–157, 1978.

N. H. Barton, A. M. Etheridge, and A. Véber. A new model for evolution in a spatial continuum. *Electron. J. Probab.*, 15(7):162–216, 2010a.

N. H. Barton, J. Kelleher, and A. M. Etheridge. A new model for extinction and recolonization in two dimensions: quantifying phylogeography. *Evolution*, 64(9): 2701–2715, 2010b.

N. H. Barton, A. M. Etheridge, J. Kelleher, and A. Véber. Inference in two dimensions: allele frequencies versus lengths of shared sequence blocks. *Theor. Popln Biol.*, 87:105–119, 2013a.

N. H. Barton, A. M. Etheridge, and A. Véber. Modelling evolution in a spatial continuum. *J. Stat. Mech.*, P01002, 2013b.

M. A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160, 2003.

M. A. Beaumont. Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.*, 41:379–406, 2010.

M. A. Beaumont, W. Zhang, and D.J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035, 2002.

M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

N. Berestycki, A. M. Etheridge, and A. Véber. Large scale behaviour of the spatial Λ-Fleming-Viot process. *Ann. Inst. H. Poincaré Probab. Statist.*, 49(2):374–401, 2013.

J. Bertoin and J.-F. Le Gall. Stochastic flows associated to coalescent processes. *Probab. Theory Related Fields*, 126:261–288, 2003.

J. Bertoin and J.-F. Le Gall. Stochastic flows associated to coalescent processes II: stochastic differential equations. *Ann. Inst. Henri Poincare Probab. Stat.*, 41(3): 307–333, 2005.

A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. R. Statist. Soc. B*, 68(3):333–382, 2006.

A. Beskos, O. Papaspiliopoulos, and G. O. Roberts. Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. *Ann. Statist.*, 37(1): 223–245, 2009.

A. Bhattacharya and D. B. Dunson. Strong consistency of nonparametric Bayes density estimation on compact metrix spaces with applications to specific manifolds. *Ann. Inst. Stat. Math.*, 64:687–714, 2012.

B. M. Bibby and M. Sørensen. Hyperbolic processes in finance. In S. Rachev, editor, *Handbook of heavy tailed distributions in finance*, pages 211–248. Elsevier, Amsterdam, 2003.

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.

M. Birkner and J. Blath. Measure-valued diffusions, general coalescents and population genetic inference. In J. Blath, P. Mörters, and M. Scheutzow, editors, *Trends in Stochastic Analysis*, London Mathematical Society Lecture Note Series, pages 329–363. Cambridge University Press, 2009.

M. Birkner, J. Blath, M. Möhle, M. Steinrücken, and J. Tams. A modified lookdown construction for the Ξ-Fleming-Viot process with mutation and populations with recurrent bottlenecks. *Alea*, 6:25–61, 2009.

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.

M. Birkner, J. Blath, and B. Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, Early online November 12, 2012.

M. Birkner, J. Blath, and B. Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193(1):255–290, 2013.

B. A. Bodo, M. E. Thompson, and T. E. Unny. A review of stochastic differential equations for applications in hydrology. *Stoch. Hydrol. Hydraul.*, 2:81–100, 1987.

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.

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

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

B. Casella and G. O. Roberts. Exact Monte Carlo simulation of killed diffusions. *Adv. Appl. Probab.*, 40:273–291, 2008.

B. Casella and G. O. Roberts. Exact simulation of jump-diffusion processes with Monte Carlo applications. *Methodol. Comput. Appl. Probab.*, 13:449–473, 2011.

I. Castillo and R. Nickl. Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *Ann. Stat.*, 41(4):1999–2028, 2013.

I. Castillo and R. Nickl. On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *Ann. Stat.*, 42(5):1941–1969, 2014.

F. Cérou, P. Del Moral, and A. Guyader. A non–asymptotic variance theorem for unnormalized Feynman-Kac particle models. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47:629–649, 2011.

F. Cérou, P. Del Moral, T. Furon, and A. Guyader. Sequential Monte Carlo for rare event estimation. *Stat. Comput.*, 22:795–808, 2012.

L. Chen and D. Filipović. A simple model for credit migration and spread curves. *Finance Stochast.*, 9:211–231, 2005.

Y. Chen, J. Xie, and J. S. Liu. Stopping-time resampling for sequential Monte Carlo methods. *J. R. Stat. Soc. B*, 67:199–217, 2005.

P. Cheridito, D. Filipović, and M. Yor. Equivalent and absolutely continuous measure changes for jump-diffusion processes. *Ann. Appl. Probab.*, 15(3):1713–1732, 2005.

J. A. Christen and C. Fox. Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Stat.*, 14(4):795–810, 2005.

D. C. Crawford, T. Bhangale, N. Li, G. Hellenthal, M. J. Reider, D. A. Nickerson, and M. Stephens. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.*, 36:700–706, 2004.

T. Cui, J. Martin, Y. M. Marzouk, A. Solonen, and A. Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11): 114015, 2014.

M. Dashti and A. M. Stuart. The Bayesian approach to inverse problems. In R. Ghanem, D. Higdon, and H. Owhadi, editors, *Handbook of Uncertainty Quantification*. Springer, 2016.

M. De Iorio and R. C. Griffiths. Importance sampling on coalescent histories I. *Adv. in Appl. Probab.*, 36(2):417–433, 2004a.

M. De Iorio and R. C. Griffiths. Importance sampling on coalescent histories II: Subdivided population models. *Adv. in Appl. Probab.*, 36(2):434–454, 2004b.

M. De Iorio, R. C. Griffiths, L. Leblois, and F. Rousset. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor. Popln Biol.*, 68:41–53, 2005.

P. Del Moral. *Feynman-Kac formulae: genealogical and interacting particle systems with applications*. Springer, New York, 2004.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, 68:411–436, 2006.

R. Der and J. B. Plotkin. The equilibrium allele frequency distribution for a population with reproductive skew. *Genetics*, 196(4):1199–1216, 2014.

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *Ann. Statist.*, 14(1):1–26, 1986.

P. Donnelly and T. Kurtz. A countable representation of the Fleming-Viot measure-valued diffusion. *Ann. Probab.*, 24(2):698–742, 1996.

P. Donnelly and T. Kurtz. Particle representations for measure-valued population models. *Ann. Probab.*, 27(1):166–205, 1999.

J. Doob. Application of the theory of martingales. *Colloq. Intern. du C.N.R.S. (Paris)*, 13:23–27, 1949.

R. Douc, O. Cappé, and E. Moulines. Comparison of resampling schemes for particle filtering. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, pages 64–69, 2005.

A. Doucet and A. M. Johansen. A tutorial on particle filtering and smoothing: fifteen years later. In D. Crisan and B. Rozovsky, editors, *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press, 2011.

A. Doucet, J. F. G. De Freitas, and N. J. Gordon. *Sequential Monte Carlo methods in practice*. Springer, New York, 2001.

A. Doucet, M. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2):295–313, 2015.

A. J. Drummond, G. K. Nicholls, A. G. Rodrigo, and W. Solomon. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161:1307–1320, 2002.

R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge studies in advanced mathematics*. Cambridge University Press, revised reprint of the 1989 original edition, 2002.

R. Durrett. *Probability models for DNA sequence evolution*. Springer, 2008.

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

N. El Karoui and J. P. Lepeltier. Representation des processus ponctuels multivariés à l'aide dun processus d'e Poisson. *Z. Warsch. Verw. Gebiete*, 39:111–133, 1977.

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

A. M. Etheridge. Drift, draft and structure: Some mathematical models of evolution. *Banach Center Publ.*, 80:121–144, 2008.

A. M. Etheridge and T. G. Kurtz. Genealogical construction of population models. *Preprint*, arXiv:1402.6724 [math.PR], 2014.

A. M. Etheridge and A. Véber. The spatial Λ-Fleming-Viot process on a large torus: Genealogies in the presence of recombination. *Ann. Appl. Probab.*, 22(6): 2165–2209, 2012.

A. M. Etheridge, R. C. Griffiths, and J. E. Taylor. A coalescent dual process in a moran model with genic selection, and the Lambda coalescent limit. *Theor. Popln Biol.*, 78(2):77–92, 2010.

A. M. Etheridge, A. Véber, and F. Yu. Rescaling limits of the spatial Lambda-Fleming-Viot process with selection. *Preprint*, arXiv:1406.5884 [math.PR], 2014.

P. Fearnhead. Computational methods for complex stochastic systems: a review of some alternatives to MCMC. *Stat. Comput.*, 18:151–171, 2008.

P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159:1299–1318, 2001.

J. Felsenstein. A pain in the torus: some difficulties with models of isolation by distance. *The American Naturalist*, 109(967):359 – 368, 1975.

J. Felsenstein, M. K. Kuhner, J. Yamamoto, and P. Beerli. Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. *IMS Lect. Notes Monogr. Ser.*, 33:163–185, 1999.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *Ann. Stat.*, 1(2):209–230, 1973.

D. Filipović, P. Cheridito, and R. L. Kimmel. Market price of risk specifications for affine models: theory and evidence. *J. Financ. Econ.*, 83(1):123–170, 2007.

R. A. Fisher. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–160, 1912.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A*, 222:309–368, 1922.

S. Fornaro. *Regularity properties for second order partial differential operators with unbounded coefficients*. PhD thesis, Università del Salento, 2004.

M. Fuchs and P.-D. Yu. Rumor source detection for rumor spreading on random increasing trees. *Electron. Commun. Probab.*, 20(Article 2):1–12, 2015.

A. Ganesh, L. Massoulie, and D. Towsley. The effect of network topology on the spread of epidemics. In *Proc. 24th Annual Joint Conference of the IEEE Computer and Communication Societies (INFOCOM)*, volume 2, pages 1455–1466, 2015.

J. C. Gay, S. Myers, and G. McVean. Estimating meiotic gene conversion rates from population genetic data. *Genetics*, 177:881–894, 2007.

S. Ghosal and A. W. van der Vaart. Convergence rates of posterior distributions for non-i.i.d observations. *Ann. Statist.*, 35:192–223, 2007.

S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Non-informative priors via sieves and packing numbers. In S. Panchapakesan and N. Balakrishnan, editors, *Advances in statistical decision theory and applications*. Birkhäuser, 1997.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.

P. Glasserman and Y. Wang. Counterexamples in importance sampling for large deviations probabilities. *Ann. Appl. Probab.*, 7(3):731–746, 1997.

P. Glasserman, P. Heidelberg, P. Shahabuddin, and T. Zajic. Multi-level splitting for estimating rare event probabilities. *Oper. Res.*, 47:585–600, 1999.

E. Gobet. Weak approximation of killed diffusions using Euler schemes. *Stoch. Proc. Appl.*, 87(2):167–197, 2000.

F. B. Gonçalves. *Exact simulation and Monte Carlo inference for jump-diffusion processes*. PhD thesis, University of Warwick, 2011.

F. B. Gonçalves and G. O. Roberts. Exact simulation problems for jump-diffusions. *Methodol. Comput. Appl. Probab.*, 16(4):907–930, 2013.

D. Gorur and Y.W. Teh. An efficient sequential Monte Carlo algorithm for coalescent clustering. *NIPS*, 2008.

R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.*, 3:479–502, 1996.

R. C. Griffiths and P. Marjoram. An ancestral recombination graph. In P. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Human Evolution*, pages 257–270. Springer-Verlang, 1997.

R. C. Griffiths and S. Tavaré. Ancestral inference in population genetics. *Statist. Sci.*, 9:307–319, 1994a.

R. C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond. B*, 344:403–410, 1994b.

R. C. Griffiths and S. Tavaré. Simulating probability distributions in the coalescent. *Theor. Popln Biol.*, 46:131–159, 1994c.

R. C. Griffiths and S. Tavaré. The ages of mutations in gene trees. *Ann. Appl. Probab.*, 9:567–590, 1999.

R. C. Griffiths, P. A. Jenkins, and Y. S. Song. Importance sampling and the two-locus model with subdivided population structure. *Adv. in Appl. Probab.*, 40: 473–500, 2008.

S. Gugushvili and P. Spreij. Non-parametric Bayesian drift estimation for stochastic differential equations. *Lith. Math. J.*, 54(2):127–141, 2014.

S. Gugushvili, F. van der Meulen, and P. Spreij. Nonparametric Bayesian inference for multidimensional compound Poisson processes. *Mod. Stoch. Theory Appl.*, 2 (1):1–15, 2015.

S. Guindon, H. Guo, and D. Welch. Demographic inference under the coalescent in a spatial continuum. *Theor. Popul. Biol.*, 111:43–50, 2016.

J. Gurland. Inversion formulae for the distribution of ratios. *Ann. Math. Statist.*, 19:228–237, 1948.

R. C. Hannum, M. Hollander, and N. A. Langberg. Distributional results for random functionals of a dirichlet process. *Ann. Probab.*, 9:665–670, 1981.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

G. Hellenthal, A. Auton, and D. Falush. Inferring human colonization history using a copying model. *PLoS Genet.*, 4:e1000078, 2008.

G. Hellenthal, A. Auton, and D. Falush. An approximate likelihood for genetic data under a model with recombination and population splitting. *Theor. Popul. Biol.*, 75(4):331–345, 2009.

H. M. Herbots. The structured coalescent. In P. Donnelly and S. Tavaré, editors, *Progress in Population Genetics and Human Evolution*, pages 231–255. Springer-Verlang, 1997.

B. Heuer and A. Sturm. On spatial coalescents with multiple mergers in two dimensions. *Theor. Popln Biol.*, 87:90–104, 2013.

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors. *Bayesian nonparametrics*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, 2010.

A. Hobolth, M. Uyenoyama, and C. Wiuf. Importance sampling for the infinite sites model. *Stat. Appl. Genet. Mol.*, 7:Article 32, 2008.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, 5(6):e1000529, 2009.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.*, 96(453):161–173, 2001.

L. F. James. Large sample asymptotics for the two-parameter Poisson-Dirichlet process. In *Institute of Mathematical Statistics Collections*, volume 3, pages 187–199. 2008.

A. Jasra, N. Kantas, and A. Persing. Bayesian parameter inference for partially observed stopped processes. *Stat Comput*, 24:1–20, 2014.

P. A. Jenkins. Stopping-time resampling and population genetic inference under coalescent models. *Stat. Appl. Genet. Mol. Biol.*, 11(1):Article 9, 2012.

P. A. Jenkins and R. C. Griffiths. Inference from samples of DNA sequences using a two-locus model. *J. Comput. Biol.*, 18:109–127, 2011.

A. M. Johansen, P. Del Moral, and A. Doucet. Sequential Monte Carlo samplers for rare events. *Proceedings of the 6th international workshop on rare event simulation*, pages 256–267, 2006.

G. Kallianpur. Differential-equations models for spatially distributed neurons and propagation of chaos for interacting systems. *Math. Biosc.*, 112:207–224, 1992.

G. Kallianpur and J. Xiong. Asymptotic behaviour of a system of interacting nuclear-space-valued stochastic differential equations driven by Poisson random measures. *Appl. Math. Opt.*, 30:175–201, 1994.

J. Kelleher, N. H. Barton, and A. M. Etheridge. Coalescent simulation in continuous space. *Bioinformatics*, 29(7):955–956, 2013.

J. Kelleher, A. M. Etheridge, and N. H. Barton. Coalescent simulation in continuous space: algorithms for large neighbourhood size. *Theor. Popln Biol.*, 95:13–23, 2014.

M. Kimura. "Stepping stone" model of population. *Ann. Rep. Nat. Inst. Genetics Japan*, 3:62–63, 1953.

J. F. C. Kingman. The coalescent. *Stochast. Process. Applic.*, 13(3):235–248, 1982a.

J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Prob.*, 19A: 27–43, 1982b.

V. N. Kolokoltsov. On Markov processes with decomposable pseudo-differential generators. *Stoch. Stoch. Rep.*, 76(1):1–44, 2004.

A. Kong, J. S. Liu, and W. H. Wong. Sequential imputations and Bayesian missing data problems. *J. Am. Statist. Assoc.*, 89(425):278–288, 1994.

J. Koskela, P. A. Jenkins, and D. Spanò. Computational inference beyond Kingman's coalescent. *J. Appl. Probab.*, 52(2):519–537, 2015a.

J. Koskela, P. A. Jenkins, and D. Spanò. Bayesian non-parametric inference for Λ-coalescents: consistency and a parametric method. *Preprint*, arXiv:1512.00982, 2015b.

J. Koskela, D. Spanò, and P. A. Jenkins. Consistency of Bayesian nonparametric inference for discretely observed jump diffusions. *Preprint*, arXiv:1506.04709, 2015c.

J. Koskela, D. Spanò, and P. A. Jenkins. Inference and rare event simulation for stopped Markov processes via reverse time sequential Monte Carlo. *Preprint*, arXiv:1603.02834, 2016.

S. M. Krone and C. Neuhauser. Ancestral processes with selection. *Theor. Popln Biol.*, 51(3):210–237, 1997.

M. K. Kuhner, J. Yamamoto, and J. Felsenstein. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, 140:1421–1430, 1995.

L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics*, 1: 277–330, 1953.

L. Le Cam. On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symp. Math. Statist. Probab.*, 1:129–156, 1956.

L. Le Cam. Locally asymptotically normal families of distributions. *University of California Publications in Statistics*, 3:37–98, 1960.

L. Le Cam. *Asymptotic methods in statistical decision theory.* Springer, 1986.

A. Lee, C. Yau, M. B. Giles, A. Doucet, and C. C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced Monte Carlo methods. *J. Comp. Graph. Stat.*, 19(4):769–789, 2010.

N. Li and G. R. Abecasis. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.*, S79:2290, 2006.

N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165:2213–2233, 2003.

A. Lijoi, I. Prünster, and S. G. Walker. Extending Doob's consistency theorem to nonparametric densities. *Bernoulli*, 10(4):651–663, 2004.

J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer, New York, 2001.

A. Y. Lo. On a class of Bayesian nonparametric estimates. 1. density estimates. *Ann. Statist.*, 12:351–357, 1984.

G. Malécot. Les mathématiques de l'hérédité. *Masson et Cie*, 1948.

J. Marchini, B. Howie, S. R. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, 39(7):906–913, 2007.

P. Marjoram and S. Tavaré. Modern computational appraoches for analysing molecular genetic variation data. *Nat. Rev. Genet.*, 7:759–770, 2006.

P. Marjoram, J. Molitor, V. Plaganol, and S. Tavaré. Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 100:15324–15328, 2003.

H. Masuda. Ergodicity and exponential β-mixing bounds for multidimensional diffusions with jumps. *Stoch. Proc. Appl.*, 117:35–56, 2007.

H. Masuda. Erratum to "Ergodicity and exponential β-mixing bounds for multidimensional diffusions with jumps". *Stoch. Proc. Appl.*, 119:676–678, 2009.

F. Medina-Aguayo, A. Lee, and G. O. Roberts. Stability of noisy Metropolis-Hastings. *Preprint*, arxiv:1503.07066, 2015.

X. L. Meng and W. H. Wong. Simulating rations of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica*, 6:831–860, 1996.

R. C. Merton. Option pricing when underlying stock returns are discontinuous. *J. Financ. Econ.*, 3:125–144, 1976.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21: 1087–1092, 1953.

M. Möhle. On sampling distributions for coalescent processes with simultaneous multiple collisions. *Bernoulli*, 12(1):35–53, 2006.

M. Möhle and S. Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Ann. Probab.*, 29(4):1547–1562, 2001.

M. Möhle and S. Sagitov. Coalescent patterns in exchangeable diploid population models. *J. Math. Biol.*, 47:337–352, 2003.

C. Moore and M. E. J. Newman. Epidemics and percolation in small-world networks. *Phys. Rev. E*, 61:5678–5682, 2000.

R. A. Neher and O. Hallatschek. Genealogies of rapidly adapting populations. *Proc. Natl Acad. Sci.*, 110(2):437–442, 2013.

X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Stat.*, 41(1):370–400, 2013.

R. Nickl and J. Söhl. Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Preprint*, arXiv:1510.05526, 2015.

R. Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154:931–942, 2000.

H. Owhadi, C. Scovel, and T. Sullivan. Brittleness of Bayesian inference under finite information in a continuous world. *Electron. J. Statist.*, 9(1):1–79, 2015.

L. Panzar and H. van Zanten. Nonparametric Bayesian inference for ergodic diffusions. *J. Statist. Plann. Inference*, 139:4193–4199, 2009.

O. Papaspiliopoulos and G. O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

O. Papaspiliopoulos, Y. Pokern, G. O. Roberts, and A. M. Stuart. Nonparametric estimation of diffusions: a differential equations approach. *Biometrika*, 99:511–531, 2012.

R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, 2001.

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.

J. S. Paul, M. Steinrücken, and Y. S. Song. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics*, 187:1115–1128, 2011.

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

J. Pitman. *Combinatorial stochastic processes. Summer school on probability theory held in Saint-Flour, July 7–24, 2002. Lecture Notes in Math.*, volume 1875. Springer, Berlin, 2006.

Y. Pokern, A. M. Stuart, and H. van Zanten. Posterior consistency via precision operators for nonparametric drift estimation in SDEs. *Stochastic Process. Appl.*, 123:603–628, 2013.

M. Pollock. On the exact simulation of (jump) diffusion bridges. *Submitted*, arXiv:1505.03030, 2015.

M. Pollock, A. M. Johansen, and G. O. Roberts. On the exact and ε-strong simulation of (jump) diffusions. *Bernoulli*, To appear, 2015a.

M. Pollock, A. M. Johansen, and G. O. Roberts. Particle filtering for partially observed jump diffusions. *In preparation*, 2015b.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molec. Biol. Evol.*, 16:1791–1798, 1999.

P. Protter. *Stochastic integration and differential equations*. Stochastic modelling and applied probability. Springer-Verlag, second edition, 2005.

E. Regazzini, A. Guglielmi, and G. Di Nunno. Theory and numerical analysis for exact distributions of functionals of a Dirichlet process. *Ann. Stat.*, 30(5):1376–1411, 2002.

L. C. G Rogers and D. Williams. *Diffusions, Markov processes and martingales*, volume 1. Wiley, 2 edition, 1994.

G. Rubino and B. Tuffin. *Rare event simulation using Monte Carlo methods*. Wiley, 2009.

J. S. Sadowsky and J. A. Bucklew. On large deviation theory and asymptotically efficient Monte Carlo estimation. *IEEE Trans. Inf. Theory*, 36:579–588, 1990.

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

O. Sargsyan and J. Wakeley. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theor. Popln Biol.*, 7:104–114, 2008.

K.-I. Sato. *Lévy processes and infinitely divisible distributions*. Cambridge University Press, 1999.

P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotype phase. *Am. J. Hum. Genet.*, 78:629–644, 2006.

R. L. Schilling and J. Wang. Some theorems on Feller processes: transience, local times and ultracontractivity. *T. Am. Math. Soc.*, 365(6):3255–3286, 2013.

L. Schwartz. On Bayes procedures. *Z. Warsch. Verw. Gebiete*, 4:10–26, 1965.

J. Schweinsberg. Coalescents with simultaneous multiple collisions. *Electron. J. Probab.*, 5:1–50, 2000.

J. Schweinsberg. Coalescent processes obtained from super-critical Galton-Watson processes. *Stoch. Proc. Appl.*, 106:107–139, 2003.

J. Sethuraman. A constructive definition of Dirichlet priors. *Stat. Sinica*, 4:639–650, 1994.

D. Shah and T. Zaman. Detecting sources of computer viruses in networks: theory and experiment. In *Proc. ACM Sigmetrics*, volume 15, pages 5249–5262, 2010.

D. Shah and T. Zaman. Finding rumor sources on random trees. *arXiv preprint*, 1110.6230, 2016.

S. Sheehan, K. Harris, and Y. S. Song. Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics*, 194:647–662, 2013.

S. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA*, 104:1760–1765, 2007.

J. P. Spence, J. A. Kamm, and Y. S. Song. The site frequency spectrum for general coalescents. *Genetics*, 202:1549–1561, 2016.

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

M. Steinrücken, J. S. Paul, and Y. S. Song. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. *Theor. Popln Biol.*, 87:51–61, 2013b.

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

M. Stephens and P. Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.*, 76(3):449–462, 2005.

O. Stramer and R. L. Tweedie. Existence and stability of weak solutions to stochastic differential equations with non-smooth coefficients. *Stat. Sinica*, 7:577–593, 1997.

S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518, 1997.

J. E. Taylor and A. Véber. Coalescent processes in subdivided populations subject to recurrent mass extinctions. *Electron. J. Probab.*, 14:242–288, 2009.

F. van der Meulen and H. van Zanten. Consistent nonparametric Bayesian inference for discretely observed scalar diffusions. *Bernoulli*, 19(1):44–63, 2013.

F. van der Meulen, A. W. van der Vaart, and H. van Zanten. Convergence rates of posterior distributions for Brownian semimartingale models. *Bernoulli*, 12(5): 863–888, 2006.

F. van der Meulen, M. Schauer, and H. van Zanten. Reversible jump MCMC for non-parametric drift estimation for diffusion processes. *Comput. Stat. Data An.*, 71:615–632, 2014.

H. van Zanten. Nonparametric Bayesian methods for one-dimensional diffusion models. *Math. Biosci.*, 243(2):215–222, 2013.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Stat. Sinica*, 21:5–42, 2011.

A. Véber and A. Wakolbinger. The spatial Lambda-Fleming-Viot process: an event-based construction and a lookdown representation. *Ann. Inst. H. Poincar Probab. Statist.*, 51(2):570–598, 2015.

J. Wakeley. *Coalescent theory: an introduction.* Roberts & Co, 2009.

S. Walker. New approaches to Bayesian consistency. *Ann. Statist.*, 32:2028–2043, 2004.

J. Wang. Regularity of semigroups generated by Lévy type operators via coupling. *Stoch. Proc. Appl.*, 120(9):1680–1700, 2010.

I. J. Wilson and D. J. Balding. Genealogical inference from microsatellite data. *Genetics*, 150:499–510, 1998.

I. J. Wilson, M. E. Weale, and D. J. Balding. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society: Series A*, 166(2):155–201, 2003.

G. Winkler. Extreme points of moment sets. *Math. Oper. Res.*, 30(4):581–587, 1988.

S. Wright. Evolution in Mendelian populations. *Genetics*, 16(2):97–159, 1931.

S. Wright. Isolation by distance. *Genetics*, 28(2):114–138, 1943.

J. Yin, M. I. Jordan, and Y. S. Song. Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. *Bioinformatics*, 25(12): i231–i239, 2009.

J. Yu. Closed-form likelihood approximation and estimation of jump-diffusions with an application to the realignment risk of the Chinese Yuan. *J. Econometrics*, 141: 1245–1280, 2007.