Conditional Sampling Distributions for Coalescent Models Incorporating Recombination

Joshua Paul

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2013-42
http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-42.html

May 1, 2013

Conditional Sampling Distributions for Coalescent Models Incorporating Recombination

by

Joshua Samuel Paul

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Computer Science

and the Designated Emphasis

in

Computational and Genomic Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Yun S. Song, Chair
Professor Lior Pachter
Professor Rasmus Nielsen

Fall 2012

Conditional Sampling Distributions for
Coalescent Models Incorporating Recombination

Joshua Samuel Paul

Abstract

Conditional Sampling Distributions for
Coalescent Models Incorporating Recombination

by

Joshua Samuel Paul

Doctor of Philosophy in Computer Science
and the Designated Emphasis in
Computational and Genomic Biology

University of California, Berkeley
Professor Yun S. Song, Chair

With the volume of available genomic data increasing at an exponential rate, we have unprecedented ability to address key questions in molecular evolution, historical demography, and epidemiology. Central to such investigations is population genetic inference, which seeks to quantify the genetic relationship of two or more individuals provided a stochastic model of evolution. A natural and widely-used model of evolution is Kingman's coalescent (Kingman, 1982a), which explicitly describes the genealogical relationship of the individuals, with various extensions to account for complex biological phenomena. Statistical inference under the coalescent, however, remains a challenging computational problem. Modern population genetic methods must therefore realize a balance between computational efficiency and fidelity to the underlying model. A promising class of such methods employ the conditional sampling distribution (CSD).

The CSD describes the probability of sampling an individual with a particular genomic sequence, provided that a collection of individuals from the population, and their corresponding sequences, has already been observed. Critically, the true CSD is generally inaccessible, and it is therefore necessary to use an approximate CSD in its place; such an approximate CSD is ideally both accurate and computationally efficient. In this thesis, we undertake a theoretical and algorithmic investigation of the CSD for coalescent models incorporating mutation, homologous (crossover) recombination, and population structure with migration.

Motivated by the work of De Iorio and Griffiths (2004a), we propose a general technique for algebraically deriving an approximate CSD directly from the underlying population genetic model. The resulting CSD admits an intuitive coalescent-like genealogical interpretation, explicitly describing the genealogical relationship of the conditionally sampled individual to the previously sampled individuals. We make use of the genealogical interpretation to introduce additional approximations, culminating in the sequentially Markov CSD (SMCSD), which models the conditional genealogical relationship site-by-site across the genomic sequence. Critically, the SMCSD can be cast as a hidden Markov model (HMM), for which efficient algorithms exist; by further specializing the general HMM methods to the SMCSD, we obtain optimized algorithms with substantial practical benefit. Finally, we empirically validate both the accuracy and computational efficiency of our proposed CSDs, and demonstrate their utility in several applied contexts.

For my parents, Lin and Dave

Contents

Acknowledgements

When I started at Berkeley in the Fall of 2007, I had little idea of what was in store for me. I knew that being a graduate student would engender a unique set of challenges, and I hoped to be successful, but I did not anticipate the multitude of emotions that would be involved – from the highest of highs to the lowest of lows and back again. Were it not for the support I received from those around me, I would not have made it through the first year at Berkeley.

First and foremost, I would like express my sincere gratitude to my advisor, Yun Song, who is truly a remarkable mentor and person. Yun introduced me to mathematical and population genetics, and provided me all of the raw material upon which my research was built; he also fostered a group dynamic that at once emphasized progress in several key areas and encouraged me to pursue my own academic interests (and have some good fun). Finally, Yun was an unwavering advocate: whenever I felt that my research had reached a dead end, Yun reminded me of all the progress I had made, and all of the exciting work that remained to be done.

I would also like to acknowledge the past and present Song group: Junming Yin, Wei-Chun Kao, Paul Jenkins, Ma'ayan Bresler, Andrew Chan, Anand Bhaskar, Matthias Steinrücken, Chris Hallsworth, Kelley Harris, Sara Sheehan, and Jack Kamm. Our conversations and interactions in group meeting, journal club, and the occasional Song group whiskey/white russian party were an invaluable part of my graduate school experience. I am particularly indebted to Matthias, Anand, and Jack, with whom I have closely collaborated and developed personal relationships – rare is the conversation that I don't learn something from these gentlemen.

I would like to thank the members of my qualifying and dissertation committee, Lior Pachter, Rasmus Nielsen, and Mike Jordan, for their insightful comments and advice on my research. The Center for Theoretical and Evolutionary Genomics, including the faculty, postdocs, and students, has been a boundless source of stimulating conversation and new ideas. Similarly, the Designated Emphasis in Computational and Genomic Biology has provided invaluable opportunities to interact and share research with my colleagues. It is truly an honor to have worked with such an extraordinary group of people during my time at Berkeley.

Finally, thank you to all of my friends and family, who provided me with constant support and the occasional reminder that there is more to life than research. My parents, Linda Mandelco and David Stahl, inspired me to pursue education and, along with the rest of my family, Krista, Jessica, Karl, and of course Dagny, encouraged me to see it through. My Berkeley roommates, Nathan, Paul, Greg, and Jelena, and long-time friends, Luke, Nick, Andy, Dave, and John have been a force for balance – we've had our share of good times over the past five years. Lastly, I am extraordinarily grateful for my fiancée, Tam Crane, who I met a week prior to starting at Berkeley, and who has been my anchor, my champion, and my closest friend ever since. Thank you, Tam, I could not have wished for anyone else with whom to share this journey.

Chapter 1

Introduction & Preliminaries

In the past decade, advances in technology have reduced the cost of genomic DNA sequencing by several orders of magnitude. As a direct result, the volume of available genomic data, both for humans and other organisms, is expanding exponentially. In principle, this influx of data provides a means of answering a great many questions: What is the demographic history of humankind, and did early humans interbreed with our Neanderthal forebears? What are the genomic abnormalities that contribute to a complex genetic disease, such as cancer? What are the roles of natural selection and other evolutionary forces, such as mutation and recombination, in shaping the genome?

A common thread running through these questions, and many more, is the requirement that many individuals belonging to a population, or species, be examined *jointly*. Such analyses are within the domain of population genetics, which is generally concerned with the genetic/genomic architecture of a population subject to a stochastic model of evolution. The model of evolution is typically assumed to be a Wright-Fisher diffusion, which naturally models the stochastic effects of genetic drift, and can also accommodate models of mutation, recombination, natural selection, and population demography. The Wright-Fisher diffusion is *prospective* in the sense that it describes the evolution of a population forwards in time; in many cases there also exists a dual model, the coalescent, which is *retrospective* in the sense that it describes the genealogical relationship for a sample of individuals within the same population backwards in time.

Both the Wright-Fisher diffusion and the coalescent have been used fruitfully in population genetics, both to understand the theoretical implications of various modes of evolution, and in the context of statistical inference to begin providing answers to data-driven questions, such as those introduced above. Despite being mathematical idealizations of natural evolution, statistical inference under these models remains a challenging computational problem. With the quantity of genomic data rapidly increasing, it is therefore critical to develop practicable statistical methods that realize a balance between computational efficiency and fidelity to the underlying model. A promising class of such methods employ the conditional sampling distribution (CSD).

The CSD describes the probability of sampling an individual with a particular genetic/genomic sequence, given that a collection of individuals from the population, and their corresponding sequences, has already been observed. Critically, the CSD is intuitively appealing and well-suited to approximation; statistical procedures requiring the joint analysis of many individuals can then be rephrased in terms of one or more CSDs, and approximations used thereafter. In this thesis, we undertake a theoretical and algorithmic investigation of the CSD for coalescent models incorporating recombination, with the objective of developing highly accurate approximations that remain

computationally practicable, even for genomic-scale data. The outcome of our research is a family of statistically well-motivated CSDs, and a corresponding efficient algorithmic framework. We also demonstrate the utility of our approximate CSDs in the context of several applications.

The structure of this thesis is as follows. In the remainder of this chapter, we provide an introduction to both the Wright-Fisher diffusion and the coalescent. These models are described in a general setting, including mutation, recombination, and population structure; notably, selective neutrality is assumed throughout, so that all individuals in the population have equal reproductive opportunity. Of particular importance is the probability of obtaining a sample, and we discuss two standard methods for exact computation of this quantity; the first derives directly from the Wright-Fisher diffusion, and the second from the genealogical interpretation provided by the coalescent. Finally, we formally introduce the CSD, and describe several commonly used approximations.

In Chapter 2, we develop the approximate CSD $\hat{\pi}_{\mathrm{PS}}$. Analogous to the sampling probability discussed in Chapter 1, $\hat{\pi}_{\mathrm{PS}}$ can be constructed either by an approximation to the Wright-Fisher diffusion, or from an intuitive genealogical process, the trunk-conditional coalescent. We investigate the resulting CSD in several limits and special cases, and provide evidence that it is a reasonable approximation. We also consider the recursive expression for the conditional sampling probability (CSP), and guided by the trunk-conditional coalescent, which describes the genealogical relationship of the conditionally sampled individual to the previously sampled individuals, propose additional approximations with desirable computational properties. These approximations culminate in the sequentially Markov CSD $\hat{\pi}_{\mathrm{SMC}}$, for which the sequence of site-by-site conditional genealogical relationships is assumed to be Markov. Finally, we relate the CSDs $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{SMC}}$ to previously-proposed CSDs, and conclude that $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{SMC}}$ more faithfully approximate the true CSD.

In Chapter 3, we more fully consider practical algorithms for computing the CSPs associated with $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{SMC}}$. We show that, for a single conditionally sampled individual, the computation associated with $\hat{\pi}_{\mathrm{PS}}$ is asymptotically super-exponential in the number of sites. Due to the Markov construction of $\hat{\pi}_{\mathrm{SMC}}$, the model can be cast as a hidden Markov model (HMM), and the associated computation is asymptotically linear in the number of sites, representing an impressive theoretical speedup. Making use of additional observations about the specific form of the HMM associated with $\hat{\pi}_{\mathrm{SMC}}$, we obtain an optimized algorithm that is, in practice, several orders of magnitude faster than the traditional dynamic programming algorithm used for HMM computation.

In Chapter 4, we empirically investigate the accuracy and computationally efficiency of our proposed CSDs. In concordance with our earlier theoretical conjecture, we find that our CSDs are generally more accurate than previously-proposed CSDs; importantly, the observed improvement in accuracy is amplified as the number of sites increases, an important consideration for application to genomic-scale data. Moreover, using our optimized algorithms for $\hat{\pi}_{\mathrm{SMC}}$, we find that the time required to evaluate the CSP is, for large genomic datasets, substantially less than for previously proposed CSDs. We also demonstrate the utility of our CSD in the context of two well-known applications, importance sampling and approximate likelihood inference, and describe and evaluate several extensions and algorithmic improvements in these settings. Additionally, we describe a novel application of our CSD for approximate inference of the genealogy relating several individuals at a particular site.

Finally, in Chapter 5, we discuss our results and propose several promising future research directions. We remark that although we do not explicitly answer any of the questions posed above, we believe that the theoretical methods and results presented herein have immediate application in these important research areas.

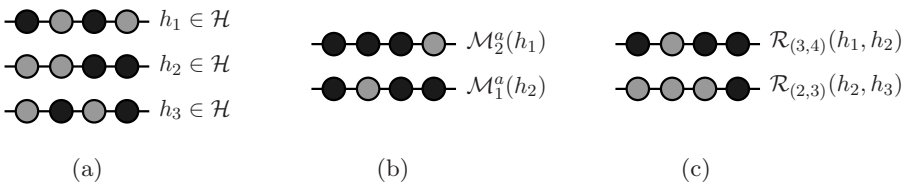


Figure 1.1. Illustration of fully-specified haplotypes, and the mutation and recombination operations. In this case, $L = \{1, 2, 3, 4\}$, $B = \{(1, 2), (2, 3), (3, 4)\}$, and $A_\ell = A = \{\text{light grey}, \text{dark grey}\}$ for each $\ell \in L$. (a) Three haplotypes $h_1, h_2, h_3 \in \mathcal{H}$. The loci of each haplotype are represented by filled circles, with the color representing the allelic type at that locus. (b) Example of two mutation operations, $\mathcal{M}_2^a(h_1), \mathcal{M}_1^a(h_2) \in \mathcal{H}$, where $a = \text{dark grey} \in A$. (c) Example of two recombination operations, $\mathcal{R}_{(3,4)}(h_1, h_2), \mathcal{R}_{(2,3)}(h_2, h_3) \in \mathcal{H}$.

1.1 Haplotypes and Sample Configurations

We begin by formalizing what is meant by the genomic/genetic "sequence" carried by an individual in a population. Without loss of generality, we consider a population of haploid individuals, so that there exists a single sequence, or haplotype, carried by each individual, and assume that this haplotype comprises a finite number of loci, and that there are a finite number of possible alleles at each locus. Each individual in the population thus carries a haplotype with the same structure, but with potentially different alleles. This model is often referred to as *finite-sites finite-alleles*. Denote the set of loci by $L = \{1, \ldots, k\}$, and the finite set of alleles available at locus $\ell \in L$ by A_ℓ. The space of haplotypes, denoted by $\mathcal{H}$, is then given by $\mathcal{H} = A_1 \times \cdots \times A_k$. Further, given a haplotype $h \in \mathcal{H}$, denote by $h[\ell] \in A_\ell$ the allele at locus $\ell \in L$, and by $h[\ell : \ell']$ the sub-haplotype for the range of loci $\ell \leq \ell'$. See Figure 1.1(a) for an example.

There are two key biological mechanisms by which the haplotypes carried by individuals within a population vary. The first, mutation, occurs when the descendant of an individual carries a haplotype with a different allele than the parental haplotype at some locus $\ell \in L$. The second, homologous recombination, occurs when the descendant of two individuals carries a haplotype that is a mosaic of the parental haplotypes. In principle, a recombination event can occur between any pair of adjacent loci; the set of recombination breakpoints is denoted by $B = \{(1, 2), \ldots, (k-1, k)\}$. Note that we only consider crossover recombination, in which a single breakpoint $b \in B$ is selected. These mechanisms are formalized by the following operators, illustrated in Figures 1.1(b) and 1.1(c).

Mutation: Given $h \in \mathcal{H}$, $\ell \in L$, and $a \in A_\ell$, define $\mathcal{M}_\ell^a(h) \in \mathcal{H}$ as the haplotype derived from h by substituting the allele at locus ℓ by a.

Recombination: Given $h, h' \in \mathcal{H}$ and $b = (\ell, \ell + 1) \in B$, define $\mathcal{R}_b(h, h') \in \mathcal{H}$ as the haplotype derived by concatenating $h[1, \ell]$ and $h'[\ell + 1, k]$.

We represent a configuration of fully-specified haplotypes by a vector $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$, where n_h is the number of haplotypes of type h in the sample. The total number of haplotypes is then denoted $n = |\mathbf{n}| = \sum_{h \in \mathcal{H}} n_h$. Finally, we denote by $\mathbf{e}_h$ the singleton configuration comprising a single haplotype of type h.

Partially-specified haplotypes It will frequently be necessary to employ haplotypes for which the alleles at one or more loci are *unspecified*. We denote an unspecified allele by $\bullet$, so that the space

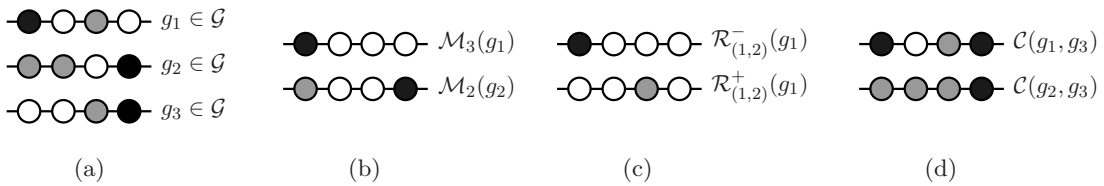


Figure 1.2. Illustration of partially-specified haplotypes, and the mutation, recombination, and coalescence operations, in the setting of Figure 1.1. (a) Three partially-specified haplotypes $g_1, g_2, g_3 \in \mathcal{G}$. Unspecified alleles are indicated by unfilled circles. (b) Example of two mutation operations, $\mathcal{M}_3(g_1), \mathcal{M}_2(g_2) \in \mathcal{G}$. (c) Example of two recombination operations, $\mathcal{R}^-_{(1,2)}(g_1), \mathcal{R}^+_{(1,2)}(g_1) \in \mathcal{G}$. (d) Example of two coalescence operations $\mathcal{C}(g_1, g_3), \mathcal{C}(g_2, g_3) \in \mathcal{G}$. Note that $g_1 \curlywedge g_3$ and $g_2 \curlywedge g_3$, and so the operations are well-defined.

of partially-specified k-locus haplotypes, denoted $\mathcal{G}$, is given by $\mathcal{G} = (A_1 \cup \{\bullet\}) \times \cdots \times (A_k \cup \{\bullet\}) \supset \mathcal{H}$. For $g \in \mathcal{G}$, we denote by $L(g) \subset L$ the subset of loci specified by g, and by $B(g)$ the set of breakpoints between the leftmost and the rightmost loci in $L(g)$. See Figure 1.2(a) for an example.

It is also necessary to revise the mutation and recombination operators for use with partially-specified haplotypes, and to introduce an operator for combining, or coalescing, two partially-specified haplotypes. Letting $g, g' \in \mathcal{G}$, we say that g and g' are compatible, and write $g \curlywedge g'$, if $g[\ell] = g'[\ell]$ for all $\ell \in L(g) \cap L(g')$.

Mutation: Given $g \in \mathcal{G}$, $\ell \in L(g)$, define $\mathcal{M}_\ell(h) \in \mathcal{G}$ as the haplotype derived from h by substituting an unspecified allele at locus ℓ.

Recombination: Given $g \in \mathcal{G}$ and $b = (\ell, \ell+1) \in B(g)$, define $\mathcal{R}^-_b(g) \in \mathcal{G}$ as the haplotype derived by concatenating the sub-haplotype $g[1, \ell]$ and $g_\bullet[\ell+1, k]$, where $g_\bullet \in \mathcal{G}$ has $g[\ell] = \bullet$ for all $\ell \in L$. Similarly, define $\mathcal{R}^+_b(g) \in \mathcal{G}$ as the haplotype derived by concatenating the sub-haplotype $g_\bullet[1, \ell]$ and $g[\ell+1, k]$

Coalescence: Given $g, g' \in \mathcal{G}$ with $g \curlywedge g'$, define $\mathcal{C}(g, g')$ as the haplotype derived by setting, for each $\ell \in L$

$$\mathcal{C}(g, g')[\ell] = \begin{cases} g[\ell] = g'[\ell], & \text{if } \ell \in L(g) \cap L(g'), \\ g[\ell], & \text{if } \ell \in L(g) \backslash L(g'), \\ g'[\ell], & \text{if } \ell \in L(g') \backslash L(g), \\ \bullet, & \text{if } \ell \notin L(g) \cup L(g'). \end{cases} \tag{1.1}$$

These modified operators are illustrated in Figures 1.2(b), 1.2(c), and 1.2(d). Analogous to a configuration of fully-specified haplotypes, we represent a configuration of partially-specified haplotypes by a vector $\mathbf{n} = (n_g)_{g \in \mathcal{G}}$, where n_g is the number of partially-specified haplotypes of type g in the sample.

The notation introduced in this section, though incomplete, forms a core that will be specialized or generalized to particular domains in the subsequent sections and chapters. We remark at the outset that the notation has been chosen to be as informative as possible without being overly cumbersome. As such, when no confusion arises, certain symbols will be re-used in different contexts. For example, we have used the symbol $\mathbf{n}$ to designate both fully- and partially-specified haplotype configurations, and will use the symbol again for configurations of haplotypes for which each haplotype resides in particular population subdivision or deme. For reference, a table of commonly used notation is provided in Appendix A.

1.2 Wright-Fisher Diffusion

The Wright-Fisher diffusion forms the basis of much of classical population genetics, and is most easily understood as mathematical idealization of the venerable discrete-time discrete-space Wright-Fisher process. The latter applies to a finite and constant-sized population of $2N$ haplotypes, corresponding to N diploid individuals, which is assumed to evolve in discrete, non-overlapping generations. We assume selective neutrality so that each haplotype is assumed to have equal reproductive opportunity. For the moment, we also assume that the population is not structured, and disregard mutation and recombination. Thus, each haplotype in a given generation is an identical copy of a single parental haplotype in the previous generation, and the parental haplotype is chosen uniformly at random. Iterating this procedure for each subsequent generation, the *count* of each haplotype in the population is modeled as a discrete-time Markov process. See Figure 1.3(a) for a realization of this process.

Though the discrete Wright-Fisher process is an intuitively appealing model of evolution, it is generally difficult to obtain associated theoretical results, particularly in the context of statistical inference. In the remainder of this section, we consider the limiting behavior of the Wright-Fisher model as $N \to \infty$. By also appropriately scaling time, we recover the Wright-Fisher diffusion, a continuous-time Markov process that models the *proportion* of each haplotype in the population, and is more amenable to mathematical analysis. We also add mutation, recombination, and population structure to the discrete Wright-Fisher process, and characterize the associated Wright-Fisher diffusions. Finally, in each case we derive a recursion for the sampling probability of a sample configuration directly from the Wright-Fisher diffusion.

1.2.1 Construction and sampling probabilities

Before proceeding, we briefly introduce diffusion processes and the associated mathematical techniques; for a more thorough introduction to diffusion processes, see Karlin and Taylor (1981). Let $\{\mathbf{X}(t)\}_{t\geq 0}$ be a continuous-time Markov process with continuous state space Δ. We say that $\{\mathbf{X}(t)\}_{t\geq 0}$ is a *diffusion process* if the sample paths are almost surely continuous. Hereafter, we consider diffusion processes that are time-homogeneous, so that the behavior of the process does not depend on the current time.

For ease of exposition, we consider the state space Δ associated with the Wright-Fisher diffusion. Recalling that the Wright-Fisher diffusion models the proportion x_h of each haplotype $h \in \mathcal{H}$ in the population, the state space is the $\mathcal{H}$-simplex

$$\Delta = \left\{ \mathbf{x} = (x_h)_{h\in\mathcal{H}} \;\middle|\; x_h \geq 0 \text{ for all } h \in \mathcal{H} \text{ and } \sum_{h\in\mathcal{H}} x_h = 1 \right\}. \tag{1.2}$$

Letting $f : \Delta \to \mathbb{R}$ be an arbitrary, bounded, twice-differentiable function with continuous second derivatives, we define the *generator* $\mathscr{L}$ of the diffusion process,

$$\mathscr{L}f(\mathbf{x}) = \lim_{t\to 0} \frac{1}{t} \cdot \mathbb{E}[f(\mathbf{X}(t)) - f(\mathbf{X}(0))|\mathbf{X}(0) = \mathbf{x}]. \tag{1.3}$$

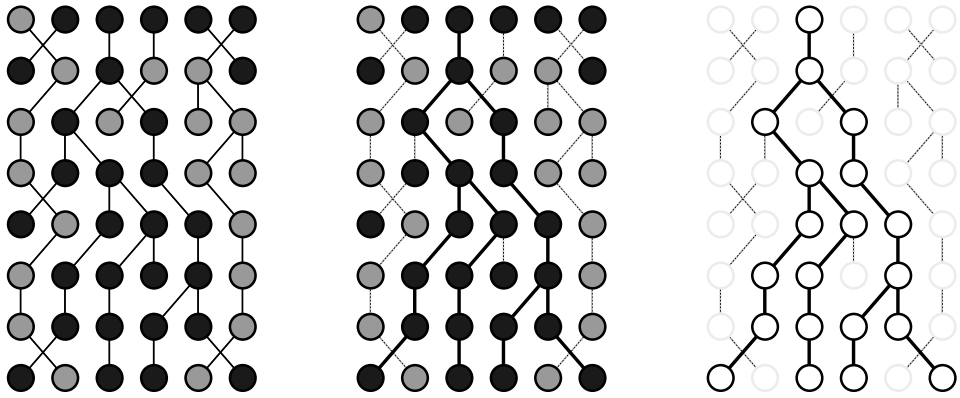


Figure 1.3. Illustration of the discrete Wright-Fisher process for a constant-sized population of 1-locus haplotypes, disregarding mutation and recombination. (a) Realization of the process for $2N = 6$ haplotypes over 8 generations. Each non-overlapping generation of haplotype is represented as a row, with the most ancient generation at the top. Each haplotype in a given generation is produced by choosing a parental haplotype uniformly at random from the haplotypes of the previous generation, and copying the type. The choice of parental haplotype is indicated by a line connecting each haplotype to its parent. (b) The genealogical relationship for a sample of 4 haplotypes in the final generation, produced by considering the ancestral haplotypes for each sample haplotype. When two or more haplotypes in a generation have a common parental haplotype, they are said to coalesce, and in this way, the genealogy forms a tree. (c) A genealogy for a sample can be produced directly for untyped haplotypes, which are represented by an unfilled circle. Starting with the most recent generation, each sample haplotype selects a parental haplotype uniformly at random. If two or more haplotypes coalesce, there are fewer ancestral haplotypes in the previous generation. This process is iterated until a single ancestral haplotype remains.

Observe that, using multi-dimensional Taylor expansion, the conditional expectation can be written

$$\begin{aligned}\mathbb{E}[f(\mathbf{X}(t)) - f(\mathbf{X}(0))|\mathbf{X}(0) = \mathbf{x}] \\ = \sum_{h\in\mathcal{H}} \mathbb{E}[X_h(t) - X_h(0)|\mathbf{X}(0) = \mathbf{x}]\frac{\partial}{\partial x_h} f(\mathbf{x}) \\ + \frac{1}{2}\cdot \sum_{h\in\mathcal{H}}\sum_{h'\in\mathcal{H}} \mathbb{E}[(X_h(t) - X_h(0))(X_{h'}(t) - X_{h'}(0))|\mathbf{X}(0) = \mathbf{x}]\frac{\partial^2}{\partial x_h \partial x_{h'}} f(\mathbf{x}) + o(t),\end{aligned} \tag{1.4}$$

where the $o(t)$ term is by the almost sure continuity of sample paths. We also define the time-homogenous *infinitesimal mean* and *infinitesimal covariance*,

$$\mu_h(\mathbf{x}) = \lim_{t\to 0}\frac{1}{t}\cdot \mathbb{E}[X_h(t) - X_h(0)|\mathbf{X}(0) = \mathbf{x}], \tag{1.5}$$

$$\sigma^2_{h,h'}(\mathbf{x}) = \lim_{t\to 0}\frac{1}{t}\cdot \mathbb{E}[(X_h(t) - X_h(0))(X_{h'}(t) - X_{h'}(0))|\mathbf{X}(0) = \mathbf{x}]. \tag{1.6}$$

The infinitesimal mean and covariance can be interpreted as the component-wise mean and covariance associated with the random variable $(\mathbf{X}(t) - \mathbf{X}(0))$ for small values of t, given that $\mathbf{X}(0) = \mathbf{x} \in \Delta$. Intuitively, these quantities describe the instantaneous stochastic evolution of the process. Making use of (1.4) along with definitions (1.5) and (1.6), the expression (1.3) for the generator may be written

$$\mathscr{L} f(\mathbf{x}) = \sum_{h\in\mathcal{H}} \mathcal{L}_h \frac{\partial}{\partial x_h} f(\mathbf{x}), \tag{1.7}$$

where

$$\mathcal{L}_h f(\mathbf{x}) = \mu_h(\mathbf{x}) f(\mathbf{x}) + \frac{1}{2}\cdot \sum_{h'\in\mathcal{H}} \sigma^2_{h,h'}(\mathbf{x})\frac{\partial}{\partial x_{h'}} f(\mathbf{x}). \tag{1.8}$$

The generator can thus be expressed in terms of the infinitesimal mean and covariance. Finally, if the diffusion admits a stationary distribution $\mathbf{X}$, then $\mathbb{E}[f(\mathbf{X}(t))|\mathbf{X}(0) = \mathbf{X}] = f(\mathbf{X})$, and therefore, making use of the definition (1.3) of the generator,

$$\mathbb{E}[\mathscr{L} f(\mathbf{X})] = \mathbb{E}\left[\sum_{h\in\mathcal{H}} \mathcal{L}_h \frac{\partial}{\partial x_h} f(\mathbf{X})\right] = 0. \tag{1.9}$$

This final result will form the basis for much of the remainder of this section.

Construction

Having introduced the relevant definitions and results for diffusion processes, we briefly describe the construction of the Wright-Fisher diffusion from the discrete Wright-Fisher process. Recall that the discrete Wright-Fisher process describes the evolution of a finite population of $2N$ haplotypes in non-overlapping generations. Denote the composition of the population after i generations by $\mathbf{Y}^{(N)}(i) = (Y_h^{(N)}(i))_{h\in\mathcal{H}}$, where $Y_h^{(N)}(i)$ is the random count of haplotypes with type $h \in \mathcal{H}$, so that $\sum_{h\in\mathcal{H}} Y_h^{(N)}(i) = 2N$. Because a given generation of haplotypes is constructed directly from previous generation, the discrete stochastic process $\{\mathbf{Y}^{(N)}(i)\}_{i\in\mathbb{N}}$ is Markov.

Next, define the continuous-time process $\{\mathbf{X}^{(N)}(t)\}_{t\geq 0}$ by scaling the discrete process,

$$\mathbf{X}^{(N)}(t) = \frac{\mathbf{Y}^{(N)}(\lfloor 2Nt \rfloor)}{2N}. \tag{1.10}$$

In particular, time is re-scaled in units of $2N$ generations, and $\mathbf{X}^{(N)}(t)$ is the vector of haplotype proportions after $\lfloor 2Nt \rfloor$ generations. The Markov property for the continuous-time process is inherited from the discrete process. Finally, we consider the limiting process as the population size $2N$ approaches infinity: it is possible to show that there exists diffusion process $\{\mathbf{X}(t)\}_{t\geq 0}$ with continuous state space Δ such that $\{\mathbf{X}^{(N)}(t)\}_{t\geq 0} \to \{\mathbf{X}(t)\}_{t\geq 0}$ in the limit $N \to \infty$. The process $\{\mathbf{X}(t)\}_{t\geq 0}$ is then the desired Wright-Fisher diffusion.

Observe that although this explanation provides intuition about the construction of the Wright-Fisher diffusion, it remains a substantial mathematical task to formally describe and prove the required convergence to a diffusion; see Donnelly (1986) for an excellent introduction. In general, the stochastic behavior of the resulting Wright-Fisher diffusion depends on the details of the evolution modeled by the discrete Wright-Fisher process. The infinitesimal mean and covariance can be obtained by considering the definitions with respect to the process $\{\mathbf{X}^{(N)}(t)\}_{t\geq 0}$ and taking the limit as $N \to \infty$. In the Sections 1.2.2 and 1.2.2, we provide concrete examples of the Wright-Fisher diffusion for specific evolutionary models.

Sampling distribution

Assuming the existence of well-defined Wright-Fisher diffusion, which models the time-evolution of haplotype proportions $\{\mathbf{X}(t)\}_{t\geq 0}$, we are then interested in the *sampling distribution* associated with the diffusion, and in particular the ordered sampling distribution $q(\cdot)$ assuming the diffusion has reached stationarity.

Let $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ be a sample configuration, and $\mathbf{x} = (x_h)_{h\in\mathcal{H}} \in \Delta$ be a haplo type proportion vector. The ordered sampling probability for $\mathbf{n}$ *conditioned* on haplotype proportions $\mathbf{x}$ is then given by the ordered multinomial probability

$$q(\mathbf{n}|\mathbf{x}) = \prod_{h\in\mathcal{H}} x_h^{n_h}. \tag{1.11}$$

Observe that, though $\mathbf{n}$ does not prescribe a particular ordering on haplotypes, the sequence of random haplotypes is exchangeable, and so the function $q(\mathbf{n}|\mathbf{x})$ is well-defined. The ordered sampling probability for $\mathbf{n}$ is then defined with respect to the stationary distribution of the Wright-Fisher diffusion, given by the random vector $\mathbf{X}$,

$$q(\mathbf{n}) = \mathbb{E}[q(\mathbf{n}|\mathbf{X})]. \tag{1.12}$$

Intuitively, $q(\mathbf{n})$ represents the probability of randomly sampling $|\mathbf{n}| = n$ haplotypes from the population drawn from the stationary distribution of the Wright-Fisher diffusion. In the general case, there is no known analytic form for $q(\mathbf{n})$. However, taking $f(\mathbf{x}) = q(\mathbf{n}|\mathbf{x})$ using the key identity (1.9), we obtain the expression

$$\mathbb{E}\left[\sum_{h\in\mathcal{H}} \mathcal{L}_h \frac{\partial}{\partial x_h} q(\mathbf{n}|\mathbf{X})\right] = \sum_{h\in\mathcal{H}} \mathbb{E}\left[\mathcal{L}_h \frac{\partial}{\partial x_h} q(\mathbf{n}|\mathbf{X})\right] = 0 \tag{1.13}$$

As will be demonstrated in the remainder of this section, in conjunction with the particulars of the Wright-Fisher model under consideration, specified by the infinitesimal mean and covariance, (1.5) and (1.6), the expression (1.13) gives rise to a *recursive* expression for the ordered sampling probability $q(\mathbf{n})$. It is similarly possible to obtain expressions for the unordered sampling probability, and these expressions will generally be related to the corresponding expressions for the ordered sampling probability by a combinatorial factor. For simplicity, we subsequently consider only the ordered sampling probability.

1.2.2 Multiple-locus, single-deme

We begin by considering a multiple-locus setting, including mutation and recombination (Ewens, 2004). Recall that in the discrete Wright-Fisher process, each haplotype in a given generation is constructed from the haplotypes of the previous generation. Incorporating recombination and mutation, construction of each haplotype occurs independently, in the following two steps,

1. With probability $(1-r)$, a haplotype selects a single parental haplotype from the previous generation. With probability r the haplotype selects two parental haplotypes, and is the product of crossover recombination; the recombination breakpoint $b \in B$ is selected with probability r_b, where $\sum_{b\in B} r_b = 1$.

2. Having selected one or both parental haplotypes, mutation at each locus $\ell \in L$ occurs with probability u_ℓ according the $(|A_\ell| \times |A_\ell|)$-dimensional matrix $\Phi^{(\ell)}$.

Following the procedure outlined in Section 1.2.1, it is possible to derive the associated Wright-Fisher diffusion by re-scaling time, and taking the limit as the population size $N \to \infty$. In order to obtain a non-degenerate diffusion, it is necessary to assume the mutation and recombination probabilities vary inversely with the population size $2N$, so that for all $\ell \in L$ and $b \in B$, $4Nu_\ell \to \theta_\ell$ and $4Nrr_b \to \rho_b$, where θ_ℓ is the *scaled mutation rate* and ρ_b is the *scaled recombination rate*. The Wright-Fisher diffusion then has infinitesimal mean and covariance,

$$\mu_h(\mathbf{x}) = \frac{1}{2}\Bigg\{ \sum_{\ell\in L} \theta_\ell \sum_{a\in A_\ell} x_{\mathcal{M}_\ell^a(h)} \big(\Phi^{(\ell)}_{a,h[\ell]} - \delta_{h,\mathcal{M}_\ell^a(h)}\big) + \sum_{b\in B} \rho_b \Bigg[\sum_{h'\in\mathcal{H}} x_{\mathcal{R}_b(h,h')} x_{\mathcal{R}_b(h',h)} - x_h \Bigg] \Bigg\} \tag{1.14}$$

$$\sigma^2_{h,h'}(\mathbf{x}) = x_h(\delta_{h,h'} - x_{h'}). \tag{1.15}$$

Having characterized the Wright-Fisher diffusion, we can use the technique described in Section 1.2.1 to obtain the following result,

Proposition 1.1. *Let $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ with $|\mathbf{n}| = n$. Then the ordered sampling probability $q(\mathbf{n})$ obtained using the diffusion generator technique described in Section 1.2.1 is given by the following recursion*

$$\begin{aligned} q(\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{h\in\mathcal{H}} n_h \Bigg\{ & (n_h - 1) q(\mathbf{n} - \mathbf{e}_h) \\ & + \sum_{\ell\in L} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} q(\mathbf{n} - \mathbf{e}_h + q(\mathcal{M}_\ell^a(h))) \\ & + \sum_{b\in B} \rho_b \sum_{h'\in\mathcal{H}} q(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h)}) \Bigg\}, \end{aligned} \tag{1.16}$$

where $\mathcal{N} = n\big(n - 1 + \sum_{\ell\in L}\theta_\ell + \sum_{b\in B}\rho_b\big)$.

Proof. By (1.8), and the infinitesimal mean and covariance given in (1.14) and (1.15),

$$\begin{aligned}\mathcal{L}_h\frac{\partial}{\partial x_h}f(\mathbf{x}) = \frac{1}{2}\Bigg\{ & x_h\sum_{h'\in\mathcal{H}}(\delta_{h,h'} - x_{h'})\frac{\partial}{\partial x_{h'}}\frac{\partial}{\partial x_h}f(\mathbf{x}) \\ & + \sum_{\ell\in L}\theta_\ell\sum_{a\in A_\ell} x_{\mathcal{M}^a_\ell(h)}\big(\Phi^{(\ell)}_{a,h[\ell]} - \delta_{h,\mathcal{M}^a_\ell(h)}\big)\frac{\partial}{\partial x_h}f(\mathbf{x}) \\ & + \sum_{b\in B}\rho_b\Bigg[\sum_{h'\in\mathcal{H}} x_{\mathcal{R}_b(h,h')}x_{\mathcal{R}_b(h',h)} - x_h\Bigg]\frac{\partial}{\partial x_h}f(\mathbf{x})\Bigg\},\end{aligned} \tag{1.17}$$

Setting $f(\mathbf{x}) = q(\mathbf{n}|\mathbf{x})$ in (1.17), and taking the expectation,

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}_h\frac{\partial}{\partial x_h}q(\mathbf{n}|\mathbf{X})\right] = n_h\cdot\frac{1}{2}\Bigg\{ & (n_h - 1)q(\mathbf{n} - \mathbf{e}_h) \\ & + \sum_{\ell\in L}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,h[\ell]}q(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}^a_\ell(h)}) \\ & + \sum_{b\in B}\rho_b\sum_{h'\in\mathcal{H}} q(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h)})\big) \\ & - \Bigg((n-1) + \sum_{\ell\in L}\theta_\ell + \sum_{b\in B}\rho_b\Bigg)q(\mathbf{n})\Bigg\}\end{aligned} \tag{1.18}$$

Summing (1.18) over haplotypes $h \in \mathcal{H}$, and making use of the key identity (1.13), the desired result (1.17) is obtained. □

In principle, repeated application of the recursion (1.16) yields a system of coupled linear equations, which can be be solved to obtain an explicit value for the ordered sampling probability $q(\mathbf{n})$. Observe, however, that the final term on the right hand side of (1.16), associated with recombination, is proportional to $q(\mathbf{n}')$, where $|\mathbf{n}'| = n + 1 > n = |\mathbf{n}|$. By induction, the resulting system of equations contains a variable for $q(\mathbf{n}')$ where $|\mathbf{n}'|$ is arbitrarily large. The system of equations is therefore infinite, and cannot be solved numerically.

Thus, although Proposition 1.1 is an important theoretical result, it does not enable explicit evaluation of $q(\mathbf{n})$. In order to obtain a recursion amenable to evaluation of $q(\mathbf{n})$, it is necessary to extend the analysis to partially-specified haplotypes. In particular, let $\mathbf{n} = (n_g)_{g\in\mathcal{G}}$ be a sample configuration of partially-specified haplotypes. Then conditional on $\mathbf{x} \in \Delta$, the ordered sampling probability is

$$q(\mathbf{n}|\mathbf{x}) = \prod_{g\in\mathcal{G}} y_g^{n_g}, \tag{1.19}$$

where $y_g = \sum_{h\in\mathcal{H}:h\curlywedge g} x_h$ is the total proportion of fully-specified haplotypes that subsume the partially-specified haplotype $g \in \mathcal{G}$. Then defining $q(\mathbf{n}) = \mathbb{E}[q(\mathbf{n}|\mathbf{X})]$ as before, it is possible to derive the following more general form of (1.16),

Proposition 1.2. *Let* $\mathbf{n} = (n_g)_{g\in\mathcal{G}}$ *with* $|\mathbf{n}| = n$. *Then the ordered sampling probability* $q(\mathbf{n})$ *obtained using the diffusion generator technique described in Section 1.2.1 is given by the following recursion*

$$\begin{aligned} q(\mathbf{n}) = \frac{1}{\mathcal{N}}\sum_{g\in\mathcal{G}} n_g \Bigg\{ & \sum_{g'\in\mathcal{G}:g'\curlywedge g} (n_{g'} - \delta_{g,g'}) q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{C}(g,g')}) \\ & + \sum_{\ell\in L(g)} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,g[\ell]} q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}^a_\ell(g)}) \\ & + \sum_{b\in B(g)} \rho_b q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{R}^-_b(g)} + \mathbf{e}_{\mathcal{R}^+_b(g))}) \Bigg\}, \end{aligned} \tag{1.20}$$

where $\mathcal{N} = \sum_{g\in\mathcal{G}} n_g\big(n - 1 + \sum_{\ell\in L(g)}\theta_\ell + \sum_{b\in B(g)}\rho_b\big)$.

Proof. Begin by observing the following identities, which are immediate from the product rule,

$$\frac{\partial}{\partial x_h} q(\mathbf{n}|\mathbf{x}) = \sum_{g\in\mathcal{G}:g\curlywedge h} n_g q(\mathbf{n} - \mathbf{e}_g|\mathbf{x}), \tag{1.21}$$

$$\frac{\partial^2}{\partial x_h \partial x_{h'}} q(\mathbf{n}|\mathbf{x}) = \sum_{g\in\mathcal{G}:g\curlywedge h} \sum_{g'\in\mathcal{G}:g'\curlywedge h'} n_g(n_{g'} - \delta_{g,g'}) q(\mathbf{n} - \mathbf{e}_g - \mathbf{e}_{g'}|\mathbf{x}). \tag{1.22}$$

Recalling the definition of $q(\cdot|\mathbf{x})$, it is also possible to obtain reduction identities, such as

$$\sum_{h\in\mathcal{H}:h\curlywedge g} q(\mathbf{n} + \mathbf{e}_h|\mathbf{x}) = q(\mathbf{n}|\mathbf{x}) \sum_{h\in\mathcal{H}:h\curlywedge g} x_h = q(\mathbf{n}|\mathbf{x}) q(\mathbf{e}_g|\mathbf{x}) = q(\mathbf{n} + \mathbf{e}_g|\mathbf{x}), \tag{1.23}$$

Making use of these identities, and setting $f(\mathbf{x}) = q(\mathbf{n}|\mathbf{x})$ in the diffusion generator (1.17),

$$\begin{aligned} \mathbb{E}\left[\mathcal{L}_h \frac{\partial}{\partial x_h} q(\mathbf{n}|\mathbf{X})\right] = \sum_{g\in\mathcal{G}:g\curlywedge h} n_g \cdot \frac{1}{2} \Bigg\{ & \sum_{g'\in\mathcal{G}:g'\curlywedge h} (g' - \delta_{g,g'}) q(\mathbf{n} - \mathbf{e}_g - \mathbf{e}_{g'} + \mathbf{e}_h) \\ & + \sum_{\ell\in L} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}^a_\ell(h)}) \\ & + \sum_{b\in B} \rho_b q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{R}^-_b(h)} + \mathbf{e}_{\mathcal{R}^+_b(h)}) \\ & - \left(n - 1 + \sum_{\ell\in L}\theta_\ell + \sum_{b\in B}\rho_b\right) q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_h) \Bigg\}. \end{aligned} \tag{1.24}$$

Summing (1.24) over haplotypes $h \in \mathcal{H}$, and making use of the key identity (1.13), the desired result (1.20) is obtained. □

As in (1.16), in computing $q(\mathbf{n})$ using (1.20), the final term on the right hand side is proportional to $q(\mathbf{n}')$, where $|\mathbf{n}'| = n + 1 > n = |\mathbf{n}|$. However, defining $L(\mathbf{n}) = \sum_{g\in\mathcal{G}} n_g \cdot |L(g)|$ to be the total number of *specified loci*, $L(\mathbf{n}') = L(\mathbf{n})$. Moreover, it can be be checked that each term on the right hand side proportional to $q(\mathbf{n}')$, for some $\mathbf{n}'$, has $L(\mathbf{n}') \le L(\mathbf{n})$. Thus, the system of equations contains only variables of the form $q(\mathbf{n}')$ for which $L(\mathbf{n}') \le L(\mathbf{n})$. As a result, repeated application of (1.20) yields a *finite* system of coupled linear equations, which can be numerically solved for the desired value $q(\mathbf{n})$.

Parent independent mutation

We shall also frequently be interested in *parent independent mutation* (PIM) models: when a mutation occurs, the mutant allele does not depend on the parental allele. Formally, a stochastic mutation matrix Φ exhibits PIM if there exists a vector $(\Phi_a)_{a\in A}$ with $\sum_{a\in A}\Phi_a = 1$, and $\Phi_{a',a} = \Phi_a$ for all $a' \in A$. Given a PIM model at locus $\ell \in L$, the term of the recursion (1.20) associated with mutation can be simplified,

$$\begin{aligned}\sum_{a\in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}^a_\ell(h)}) &= \Phi^{(\ell)}_{h[\ell]} \mathbb{E}\Big[\sum_{a\in A_\ell} q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}^a_\ell(h)}|\mathbf{X})\Big] \\ &= \Phi^{(\ell)}_{h[\ell]} \mathbb{E}\Big[q(\mathbf{n} - \mathbf{e}_a|\mathbf{X}) \sum_{a\in A_\ell} q(\mathbf{e}_{\mathcal{M}^a_\ell(h)}|\mathbf{X})\Big] \\ &= \Phi^{(\ell)}_{h[\ell]} q(\mathbf{n} - \mathbf{e}_a + \mathbf{e}_{\mathcal{M}_\ell(h)}),\end{aligned} \tag{1.25}$$

where the second and third equalities are by properties of the ordered multinomial distribution $q(\cdot|\mathbf{x})$ similar to (1.23). As a result, given a PIM model at every locus $\ell \in L$, identity (1.25) can be used to re-write (1.20) as follows,

$$\begin{aligned}q(\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{g\in\mathcal{G}} n_g \Big\{ &\sum_{g'\in\mathcal{G}: g' \curlywedge g} (n_{g'} - \delta_{g,g'}) q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{C}(g,g')}) \\ &+ \sum_{\ell\in L(g)} \theta_\ell \Phi^{(\ell)}_{h[\ell]} q(\mathbf{n} - \mathbf{e}_a + \mathbf{e}_{\mathcal{M}_\ell(h)}) \\ &+ \sum_{b\in B(g)} \rho_b q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{R}^-_b(g)} + \mathbf{e}_{\mathcal{R}^+_b(g))}) \Big\},\end{aligned} \tag{1.26}$$

where $\mathcal{N} = \sum_{g\in\mathcal{G}} n_g \big(n - 1 + \sum_{\ell\in L(g)} \theta_\ell + \sum_{b\in B(g)} \rho_b\big)$. Thus, assuming a PIM model at each locus confers both a mathematical and computational benefit. Importantly, any bi-allelic mutation model can be transformed into a PIM model. Consider an arbitrary model of mutation on the alleles A with $|A| = 2$, and specified by parameters θ and Φ,

$$\begin{aligned}\theta = \theta_0, \Phi = \begin{pmatrix} 1 - p_{12} & p_{12} \\ p_{21} & 1 - p_{21} \end{pmatrix} \\ \longrightarrow \theta_{\text{PIM}} = \theta_0(p_{12} + p_{21}), \Phi_{\text{PIM}} = \begin{pmatrix} \frac{p_{21}}{p_{12}+p_{21}} & \frac{p_{12}}{p_{12}+p_{21}} \\ \frac{p_{21}}{p_{12}+p_{21}} & \frac{p_{12}}{p_{12}+p_{21}} \end{pmatrix}.\end{aligned} \tag{1.27}$$

It can be verified that the resulting PIM model, specified by parameters θ_{PIM} and Φ_{PIM}, yields precisely the same recursive expression for $q(\mathbf{n})$.

Specialization to one-locus case

In the one-locus case, the space of haplotypes can be represented by the (finite) space of alleles $\mathcal{H} = A$, and each haplotype by a single allele $a \in A$. Moreover, recombination is not applicable, and the single scaled mutation rate is represented by θ. Given a one-locus configuration $\mathbf{n} = (n_a)_{a\in A}$, the recursion (1.16) for the ordered sampling probability $q(\mathbf{n})$ reduces to

$$q(\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{a\in A} n_a \Big\{ (n_a - 1) q(\mathbf{n} - \mathbf{e}_a) + \theta \sum_{a'\in A} \Phi_{a',a} q(\mathbf{n} - \mathbf{e}_a + \mathbf{e}_{a'}) \Big\}, \tag{1.28}$$

where $\mathcal{N} = n(n-1+\theta)$. Even for this relatively simple case, in order to explicitly evaluate $q(\mathbf{n})$, it remains necessary numerically solve a system of linear equations, generated by repeated application of (1.28). However, if we assume a PIM model, then the recursion (1.26) for $q(\mathbf{n})$ reduces to

$$q(\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{a \in A} n_a \Big\{ (n_a - 1 + \theta\Phi_a) q(\mathbf{n} - \mathbf{e}_a) \Big\}, \tag{1.29}$$

where $\mathcal{N} = n(n-1+\theta)$. Observe that each term on the right hand side of (1.29) proportional to $q(\mathbf{n}')$ has $|\mathbf{n}'| = n - 1 < n = |\mathbf{n}|$, where the inequality is strict. Consequently, there exists a partial order associated with the dependence of variables generated by repeated application of (1.29), and we refer to the recursion as *proper*. The quantity $q(\mathbf{n})$ can therefore be directly evaluated using dynamic programming or memoization, without the need to construct and numerically solve a coupled system of linear equations. Moreover, in this case, the recursion can be solved analytically, yielding the celebrated Wright Sampling Formula (Wright, 1949),

Proposition 1.3 (Wright Sampling Formula)**.** *Let $\mathbf{n} = (n_a)_{a \in A}$ be a one-locus configuration. Then the sampling probability $q(\mathbf{n})$ for a one-locus PIM model is given by*

$$q(\mathbf{n}) = \frac{1}{\theta_{(n)}} \prod_{a \in A} (\theta\Phi_a)_{(n_a)}, \tag{1.30}$$

where $x_{(i)} = (x)(x+1)(x+2)\cdots(x+i-1)$ denotes a rising factorial.

Proof. Substitute (1.30) into (1.29). □

The Wright Sampling Formula represents the only known closed-form formula for the sampling probability in the finite-locus finite-alleles setting. Recently, however, Bhaskar et al. (2012) have produced an asymptotic expansion for approximating the sampling probability for an irreducible model of mutation for four or fewer alleles.

Limiting distributions

Returning to the more general setting, we suppose that $\rho_b = \rho$ for all $b \in B$, and consider the limit $\rho \to \infty$. Intuitively, in the limit of infinite recombination, there should not exist correlation between the alleles at different loci. This is formalized in the following result,

Proposition 1.4. *Let $\mathbf{n} = (n_g)_{g \in \mathcal{G}}$ with $|\mathbf{n}| = n$, and suppose $\rho_b = \rho$ for all $b \in B$. In the limit $\rho \to \infty$, the ordered sampling probability $q(\mathbf{n})$ is given by*

$$q(\mathbf{n}) = \prod_{\ell \in L} q(\mathbf{n}[\ell]), \tag{1.31}$$

where $\mathbf{n}[\ell]$ is the one-locus configuration induced by $\mathbf{n}$ at locus $\ell \in L$, and $q(\mathbf{n}[\ell])$ is the one-locus ordered sampling probability given in (1.28).

Proof. We refer the reader to the proof of Proposition 2.6, which is entirely analogous. □

Thus, computing the sampling probability for a k-locus configuration can be efficiently performed by computing the product of the sampling probabilities for k one-locus configurations. Moreover, given a PIM model at each locus, the resulting one-locus sampling probabilities can be computed efficiently and exactly, yielding an exact result. Such asymptotic considerations have recently been extended (Jenkins and Song, 2009, 2010, 2012; Bhaskar and Song, 2012) to provide approximate expressions for the sampling probability for finite values of ρ.

1.2.3 Multiple-locus, multiple-deme

We now extend the analysis to the setting of a structured population including migration. We assume that there exists a finite set of demes $\mathcal{D}$, and that each haplotype resides in a particular deme. Recall that in the discrete Wright-Fisher process, each haplotype in a given generation is constructed from the haplotypes of the previous generation. This process, including mutation and recombination, can be extended to accommodate population structure and migration as follows. Denote the number of haplotypes in each deme $d \in \mathcal{D}$ by N_d, so that $2N = \sum_{d\in\mathcal{D}} N_d$. Then sampling a haplotype within deme $d \in \mathcal{D}$ proceeds by first selecting a parental deme $d' \in \mathcal{D}$ with probability $v_{dd'}$. Having selected a parental deme, the parental haplotype, or haplotypes in the case of recombination, are selected from the parental deme, and mutation occurs as described in Section 1.2.2.

As before, it is possible to derive the associated Wright-Fisher diffusion by re-scaling time, and taking the limit as population size $N \to \infty$. In order to obtain a non-degenerate diffusion, it is necessary to assume that the number of haplotypes N_d in each deme $d \in \mathcal{D}$ increases with N, so that $N_d/N \to \kappa_d$, the relative deme size, with $\sum_{d\in\mathcal{D}} \kappa_d = 1$. Similarly, it is necessary to assume that $v_{dd'}$ varies inversely with the population size for all $d' \neq d$ so that $4Nv_{dd'} \to \upsilon_{dd'}$, the *scaled migration rate*. Define the total migration rate associated with deme $d \in \mathcal{D}$ by $\upsilon_d = \sum_{d'\neq d} \upsilon_{dd'}$. The limiting Wright-Fisher diffusion has the expanded state space

$$\Delta = \left\{ \mathbf{x} = (x_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}} \;\middle|\; x_{d,h} \geq 0 \text{ for all } d \in \mathcal{D}, h \in \mathcal{H} \text{ and } \sum_{h\in\mathcal{H}} x_{d,h} = 1 \text{ for all } d \in \mathcal{D} \right\}, \quad (1.32)$$

where $x_{d,h}$ is the proportion of haplotype $h \in \mathcal{H}$ within deme $d \in \mathcal{D}$. As in Section 1.2.1, the diffusion generator can be written as a summation; for a bounded, twice-differentiable function with continuous second derivatives $f : \Delta \to \mathbb{R}$,

$$\mathscr{L}f(\mathbf{x}) = \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} \mathcal{L}_{d,h} \frac{\partial}{\partial x_{d,h}} f(\mathbf{x}), \quad (1.33)$$

where the generator component for $d \in \mathcal{D}$ and $h \in \mathcal{H}$ is given by

$$\mathcal{L}_{d,h}f(\mathbf{x}) = \mu_{d,h}(\mathbf{x})f(\mathbf{x}) + \frac{1}{2} \cdot \sum_{d'\in\mathcal{D}} \sum_{h'\in\mathcal{H}} \sigma^2_{(d,h),(d',h')}(\mathbf{x}) \frac{\partial}{\partial x_{d',h'}} f(\mathbf{x}), \quad (1.34)$$

and the associated infinitesimal mean and covariance are given by

$$\begin{aligned} \mu_{d,h}(\mathbf{x}) = \frac{1}{2} \Bigg\{ & \sum_{\ell\in L} \theta_\ell \sum_{a\in A_\ell} x_{d,\mathcal{M}^a_\ell(h)} \big(\Phi^{(\ell)}_{a,h[\ell]} - \delta_{h,\mathcal{M}^a_\ell(h)} \big) \\ & + \sum_{b\in B} \rho_b \Bigg[\sum_{h'\in\mathcal{H}} x_{d,\mathcal{R}_b(h,h')} x_{d,\mathcal{R}_b(h',h)} - x_{d,h} \Bigg] \\ & + \Bigg[\sum_{\substack{d'\in\mathcal{D} \\ d'\neq d}} \upsilon_{dd'} x_{d',h} - \upsilon_d x_{d,h} \Bigg] \Bigg\} \end{aligned} \quad (1.35)$$

$$\sigma^2_{(d,h),(d',h')}(\mathbf{x}) = x_{d,h}(\delta_{h,h'} - x_{d,h'})\kappa_d^{-1} \cdot \delta_{d,d'}. \quad (1.36)$$

In the extended setting of a structured population, a sample configuration is denoted by the vector $\mathbf{n} = (n_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$, where $n_{d,h}$ is the number of haplotypes of type h within deme d in the sample.

The sample configuration of haplotypes within deme $d \in \mathcal{D}$ is denoted by $\mathbf{n}_d$, and the number of haplotypes in the deme by $n_d = |\mathbf{n}_d|$. Finally, we use $\mathbf{e}_{d,h}$ to denote the singleton structured sample configuration comprising a single haplotype of type h in deme d.

Let $\mathbf{n} = (n_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$ be a structured sample configuration, and $\mathbf{x} = (x_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}} \in \Delta$ be a haplotype proportion vector. The ordered sampling probability for $\mathbf{n}$ *conditioned* on haplotype proportions $\mathbf{x}$ is then given by the ordered multinomial probability

$$q(\mathbf{n}|\mathbf{x}) = \prod_{d\in\mathcal{D}} \prod_{h\in\mathcal{H}} x_{d,h}^{n_{d,h}}. \tag{1.37}$$

Finally, taking $f(\mathbf{x}) = q(\mathbf{n}|\mathbf{x})$ and using the key identity (1.9) we obtain the expression

$$\mathbb{E}\Big[\sum_{d\in\mathcal{D}}\sum_{h\in\mathcal{H}} \mathcal{L}_{d,h}\frac{\partial}{\partial x_{d,h}} q(\mathbf{n}|\mathbf{X})\Big] = \sum_{d\in\mathcal{D}}\sum_{h\in\mathcal{H}} \mathbb{E}\Big[\mathcal{L}_{d,h}\frac{\partial}{\partial x_{d,h}} q(\mathbf{n}|\mathbf{X})\Big] = 0 \tag{1.38}$$

which is the population structure analogue of (1.13) described in Section 1.2.1, and yields

Proposition 1.5. *Let $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ be a structured sample configuration, with $|\mathbf{n}| = n$ and $|\mathbf{n}_d| = n_d$ for each $d \in \mathcal{D}$. Then the ordered sampling probability $q(\mathbf{n})$ obtained using the diffusion generator technique described in Section 1.2.1 is given by the following recursion*

$$\begin{aligned} q(\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{d\in\mathcal{D}}\sum_{h\in\mathcal{H}} n_{d,h} \Bigg\{ & (n_{d,h}-1)\kappa_d^{-1} q(\mathbf{n}-\mathbf{e}_{d,h}) \\ & + \sum_{\ell\in L}\theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} q(\mathbf{n}-\mathbf{e}_{d,h}+\mathbf{e}_{d,\mathcal{M}^a_\ell(h)}) \\ & + \sum_{b\in B}\rho_b \sum_{h'\in\mathcal{H}} q(\mathbf{n}-\mathbf{e}_{d,h}+\mathbf{e}_{d,\mathcal{R}_b(h,h')}+\mathbf{e}_{d,\mathcal{R}_b(h',h))}) \\ & + \sum_{\substack{d'\in\mathcal{D}\\ d'\neq d}} \upsilon_{dd'} q(\mathbf{n}-\mathbf{e}_{d,h}+\mathbf{e}_{d',h}) \Bigg\} \end{aligned} \tag{1.39}$$

where $\mathcal{N} = \sum_{d\in\mathcal{D}}\sum_{h\in\mathcal{H}} n_{d,h}\big((n_d-1)\kappa_d^{-1} + \sum_{\ell\in L}\theta_\ell + \sum_{b\in B}\rho_b + \upsilon_d\big)$.

Proof. Applying the generator component (1.34), with infinitesimal mean and covariance given by (1.35) and (1.36), to $f(\mathbf{x}) = q(\mathbf{n}|\mathbf{x})$, and taking the expectation,

$$\begin{aligned} \mathbb{E}\Big[\mathcal{L}_{d,h}\frac{\partial}{\partial x_{d,h}} q(\mathbf{n}|\mathbf{X})\Big] = n_{d,h}\cdot\frac{1}{2}\Bigg\{ & (n_{d,h}-1)\kappa_d^{-1} q(\mathbf{n}-\mathbf{e}_{d,h}) \\ & + \sum_{\ell\in L}\theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} q(\mathbf{n}-\mathbf{e}_{d,h}+\mathbf{e}_{d,\mathcal{M}^a_\ell(h)}) \\ & + \sum_{b\in B}\rho_b \sum_{h'\in\mathcal{H}} q(\mathbf{n}-\mathbf{e}_{d,h}+\mathbf{e}_{d,\mathcal{R}_b(h,h')}+\mathbf{e}_{d,\mathcal{R}_b(h',h))}) \\ & + \sum_{d'\neq d} \upsilon_{dd'} q(\mathbf{n}-\mathbf{e}_{d,h}+\mathbf{e}_{d',h}) \\ & - \Big((n_d-1)\kappa_d^{-1} + \sum_{\ell\in L}\theta_\ell + \sum_{b\in B}\rho_b + \sum_{d'\neq d}\upsilon_{dd'}\Big) q(\mathbf{n}) \Bigg\} \end{aligned} \tag{1.40}$$

Summing (1.40) over demes $d \in \mathcal{D}$ and haplotypes $h \in \mathcal{H}$, and making use of the key identity (1.38), the desired result (1.39) is obtained. □

Once again, though Proposition 1.5 is an important theoretical result, it does not enable explicit evaluation of $q(\mathbf{n})$ for a structured sample configuration $\mathbf{n}$. As in Section 1.2.2, it is necessary to extend the analysis to partially-specified haplotypes, which yields the following generalized recursion for a structured sample configuration on partially-specified haplotypes,

Proposition 1.6. *Let $\mathbf{n} = (n_{d,g})_{d\in\mathcal{D},g\in\mathcal{G}}$ be a structured sample configuration, with $|\mathbf{n}| = n$ and $|\mathbf{n}_d| = n_d$ for each $d \in \mathcal{D}$. Then the ordered sampling probability $q(\mathbf{n})$ obtained using the diffusion generator technique described in Section 1.2.1 is given by the following recursion*

$$\begin{aligned} q(\mathbf{c}) = \frac{1}{\mathcal{N}} \sum_{d\in\mathcal{D}} \sum_{g\in\mathcal{H}} n_{d,g} \Bigg\{ & \sum_{g'\in\mathcal{G}:g'\curlywedge g} (n_{d,g'} - \delta_{g,g'})\kappa_d^{-1} q(\mathbf{n} - \mathbf{e}_{d,g} + \mathbf{e}_{d,\mathcal{C}(g,g')}) \\ & + \sum_{\ell\in L(g)} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,g[\ell]} q(\mathbf{n} - \mathbf{e}_{d,g} + \mathbf{e}_{d,\mathcal{M}^a_\ell(g)}) \\ & + \sum_{b\in B(g)} \rho_b q(\mathbf{n} - \mathbf{e}_{d,g} + \mathbf{e}_{\mathcal{R}^-_b(g)} + \mathbf{e}_{\mathcal{R}^+_b(g))}) \\ & + \sum_{\substack{d'\in\mathcal{D}\\ d'\neq d}} \upsilon_{dd'} q(\mathbf{n} - \mathbf{e}_{d,g} + \mathbf{e}_{d',g}) \Bigg\}. \end{aligned} \tag{1.41}$$

where $\mathcal{N} = \sum_{d\in\mathcal{D}} \sum_{g\in\mathcal{G}} n_{d,h}\big((n_d - 1)\kappa_d^{-1} + \sum_{\ell\in L(g)} \theta_\ell + \sum_{b\in B(g)} \rho_b + \upsilon_d\big)$.

Proof. The proof is analogous to the proof of Proposition 1.2, with the necessary extension to a structured population provided in the proof of Proposition 1.5. □

It is reassuring that, for single deme $\mathcal{D} = \{1\}$ with $\kappa_1 = 1$, Propositions 1.5 and 1.6 are precisely equivalent to the analogous propositions 1.1 and 1.2, respectively, described in Section 1.2.2. Moreover, it is possible to extend the recursion (1.39) to a PIM model as described in Section 1.2.2. Similarly, assuming $\rho_b = \rho$ for all $b \in B$, the limit $b \to \infty$ produces the same decomposition into one-locus sampling probabilities described in Proposition 1.4.

1.3 The Coalescent

The Wright-Fisher diffusion, as a prospective model, is intuitively appealing as it models the evolution of a population forward in time. However, if we consider a finite sample of individuals from the present, there is no direct way to understand their relationship in such a population-centric context. For example, in order to directly sample of a collection of haplotypes from the Wright-Fisher diffusion, it is necessary to first explicitly simulate the population-wide diffusion proportions for a period of time sufficient to ensure stationarity, and then sample the desired haplotypes conditional on the proportions. The coalescent provides a complementary approach to the Wright-Fisher diffusion, in that it is retrospective, and operates directly on a finite sample; the outcome of the coalescent is a *genealogy* that explicitly relates the haplotypes of the sample.

As for the Wright-Fisher diffusion, the coalescent is most easily understood as a mathematical idealization of the discrete-time discrete-space Wright-Fisher process. Consider a realization of the

discrete Wright-Fisher process on a constant-size population of $2N$ one-locus haplotypes, disregarding mutation and recombination, as illustrated Figure 1.3(a). The genealogical relationship for a subset of haplotypes in the present generation can be extracted, as illustrated in Figure 1.3(b); when two or more haplotypes in a generation have a common parental haplotype in the previous generation, they are said to *coalesce*. Importantly, this genealogical structure can be produced more directly. Starting in the present generation and assuming that the haplotypes are *untyped*, meaning that the allelic type at each locus is not stated, the discrete Wright-Fisher process asserts that each haplotype selects a parental haplotype from the previous generation uniformly at random. If one or more haplotypes coalesce, there are fewer ancestral haplotypes in the previous generation. This process, illustrated in Figure 1.3(c), is iterated for the ancestral haplotypes in each generation until a single ancestral haplotype remains, the most recent common ancestor (MRCA) of the sample, yielding the desired genealogical structure.

Though the formulation of a coalescent process based directly on the discrete Wright-Fisher process is intuitively appealing, it is generally difficult to obtain associated theoretical results. In the remainder of this section, we consider the limiting behavior of the discrete coalescent process as $N \to \infty$. By also appropriately scaling time, we recover the coalescent process, a continuous-time Markov process that models the genealogical structure of a random sample of untyped haplotypes from the present, that is more amenable to mathematical analysis. Given the genealogical structure, it is straightforward to directly sample the type of the MRCA from the appropriate stationary distribution, and propagate this type forward in time, ultimately producing a typed sample and the associated genealogy. In this context, we consider the discrete Wright-Fisher process incorporating mutation, recombination, and population structure, and characterize the associated coalescent models. Finally, we describe a methodology for deriving recursive expressions for the sampling probability directly from the coalescent process, and use it to provide a genealogical interpretation for the sampling probabilities derived from the Wright-Fisher diffusion in the previous section.

1.3.1 Construction and sampling probabilities

In order to provide some intuition, we begin with a construction of Kingman's coalescent (Kingman, 1982a,b). Consider the discrete Wright-Fisher process for $2N$ haplotypes, disregarding mutation, recombination, and population structure, and the procedure described above for sampling a genealogy for n untyped haplotypes. In each generation, a number $i \leq n$ of the $2N$ haplotypes are ancestral to the haplotypes of the sample. Because each pair of haplotypes have a common parental haplotype in the previous generation with probability $1/2N$, the probability of p_{ij} of $j \leq i$ ancestral haplotypes in the previous generation is given by

$$p_{ij} = \begin{cases} 1 - \binom{i}{2}\frac{1}{2N} + o(N^{-1}), & \text{if } j = i, \\ \binom{i}{2}\frac{1}{2N} + o(N^{-1}), & \text{if } j = i-1, \\ o(N^{-1}), & \text{if } j < i-1. \end{cases} \tag{1.42}$$

The discrete process on the number of ancestral haplotypes is Markov, since the transition probability depends only on the current number ancestral haplotypes. Scaling time so that one unit of time is equivalent to $2N$ generations, precisely as was done in the construction of the Wright-Fisher diffusion, the waiting time $T_i^{(N)}$ while there are i individuals has distribution

$$\Pr(T_i^{(N)} \leq t) = 1 - p_{ii}^{\lfloor 2Nt \rfloor} = 1 - \left(1 - \binom{i}{2}\frac{1}{2N} + o(N^{-1})\right)^{\lfloor 2Nt \rfloor}. \tag{1.43}$$

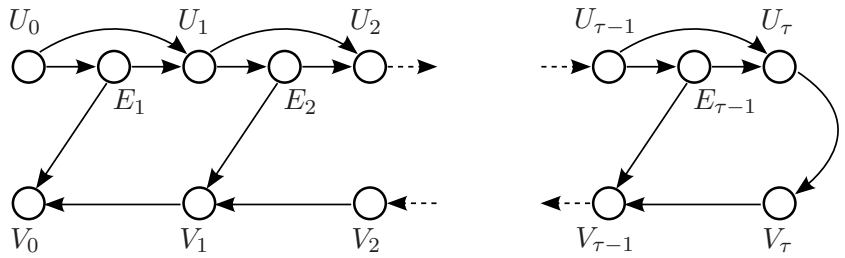


Figure 1.4. Graphical model representation of the generalized procedure for sampling a collection of haplotypes. The random variable U_i denotes the random untyped haplotype configuration following, backward in time, the i-th genealogical event, E_i; the random variable V_i denotes the corresponding typed haplotype configuration. The right-facing arrows correspond to the backward phase, in which the genealogical event $E_i = e$ is chosen conditional on $U_{i-1} = u$ from the distribution with density $p(\cdot|u)$, so that $U_{i+1} = e(u)$. The left-facing arrows correspond to the forward phase, in which the typed configuration V_i is sampled conditional on $V_{i+1} = v$ and $E_{i+1} = e$. Thus, beginning with an untyped configuration, $U_0 = \hat{n}$, this process ultimately yields the desired sample configuration V_0.

As $N \to \infty$, the waiting times converge in distribution $T_i^{(N)} \to T_i$ where $\Pr(T_i \leq t) = 1 - \exp\left(\binom{i}{2} t\right)$, so that T_i is distributed exponentially with parameter $\binom{i}{2}$. Moreover, when a transition occurs, the number of ancestral haplotypes almost surely decreases to $i - 1$. Thus, the number of haplotypes ancestral to the sample is a pure death process, backwards in time, where the death rates are given by $\binom{i}{2}$ for each $i = n, \ldots, 2$. Each transition in the pure death process corresponds to a coalescence event, wherein two ancestral haplotypes have a common ancestor. By symmetry, each pair of untyped haplotypes is equally likely to have coalesced.

The resulting process, here constructed from the discrete Wright-Fisher process by scaling time in units of $2N$ generations and taking the limit $N \to \infty$, is Kingman's coalescent. Much as for the discrete process, a realization of Kingman's coalescent can be succinctly represented as a bifurcating tree genealogy. The leaves of the tree correspond to the n untyped haplotypes for which the genealogy was constructed, each bifurcation corresponds to a coalescence of two untyped haplotypes, and the root of the tree corresponds to the untyped MRCA haplotype. Observe that the topology of the tree is entirely determined by the waiting times $\{T_i\}_{i=n,\ldots,2}$ and the pair of haplotypes chosen to coalesce at each transition.

Sampling for coalescent processes

We next consider a more general class of coalescent processes, which are able to accommodate genealogical events such as mutation, recombination, and migration, in addition to coalescence. Much as for Kingman's coalescent, a general coalescent process is naturally cast as continuous-time Markov process, starting with a collection of untyped haplotypes in the present, and proceeding backward in time, with each transition corresponding to a genealogical event; when a single untyped haplotype, the MRCA, remains, the process is terminated. A realization of the coalescent process is then a genealogy relating the haplotypes.

In order to formulate a probabilistic description of such processes, it is convenient to first introduce the concept of an untyped haplotype configuration. Recalling that an untyped haplotype has an unstated allelic type at each locus, denote by $\hat{n}$ an *ordered* collection of untyped haplotypes; equivalently, each of the haplotypes of $\hat{n}$ may be uniquely *labeled*. In the context of the coalescent process, it is also necessary to assume an ordering, or equivalently a labeling, for the haplotypes in a

typed configuration $\mathbf{n}$; a particular typed configuration therefore induces an untyped configuration. Because the haplotypes within the typed and untyped configurations are exchangeable, the sampling distributions we consider do not require an explicit representation of the haplotype labeling.

Provided a labeled untyped configuration of haplotypes $\hat{n}$, we denote a genealogical realization of the coalescent process by $\mathcal{A}_{\hat{n}}$. Importantly, haplotypes within the genealogy, including the MRCA, are also untyped, and so we refer to the genealogy itself as untyped. Provided a specific type for the MRCA, it is possible to stochastically propagate this type forward in time along the genealogy. For example, at a coalescence event, each of the descendant haplotypes is identical to the ancestral haplotype; other genealogical events, such as mutation, stochastically alter the descendant haplotype from the ancestral haplotype. Moreover, in the absence of natural selection, there is no correlation among the alleles of a single sampled haplotype, and so it straightforward to sample the specific type of the MRCA haplotype from the stationary distribution of the Wright-Fisher diffusion. In this way, it is possible to obtain the types of each haplotype in the genealogy, including the previously-untyped configuration of haplotypes $\hat{n}$. The result is a labeled typed configuration $\mathbf{n}$ associated with $\hat{n}$ and a corresponding typed genealogy $\mathcal{A}_{\mathbf{n}}$.

The coalescent processes we consider are time-homogeneous, so the behavior of the process does not depend on the current time. Consequently, embedded within the continuous-time Markov process is a discrete-time Markov process comprising the transitions within the continuous-time process, but not the waiting times. This suggests a methodology for sampling a typed haplotype configuration, which we present in some generality. Denote by U_i the random labeled untyped configuration following, backward in time, the i-th genealogical event E_i, and by V_i the corresponding typed haplotype configuration. Formally, the objective is to sample the typed configuration V_0 conditioned on the labeled untyped configuration $U_0 = \hat{n}$. The sampling procedure is naturally broken into two phases:

Backward phase: Conditioned on $U_i = u$, the distribution of possible genealogical events, backward in time, is specified by the time-homogenous coalescent process and has density denoted $p(\cdot|u)$ with support $\mathcal{E}(u)$. Moreover, for each genealogical event $e \in \mathcal{E}(u)$, then $E_i = e$ in conjunction with $U_{i-1} = u$ specifies a particular labeled untyped configuration, $U_i = e(u)$. For each i sequentially, starting with $i = 1$, suppose $U_{i-1} = u$ is known, and sample $E_i = e$ according to the density $p(\cdot|u)$, so that $U_i = e(u)$. This process is stopped when $U_i = u$ comprises a single haplotype, $|u| = 1$, and the stopping time τ is set to i.

Forward phase: As has been described, it is possible to sample a single haplotype from the stationary distribution, with density denoted $p(\cdot)$. Moreover, conditioned on $V_{i+1} = v$ and $E_i = e$, it is possible sample the typed configuration V_i from a distribution specified by the coalescent process, with density denoted $p(\cdot|v, e)$ and support $\mathcal{V}(v, e)$. Thus, sample the typed configuration V_τ according to the density $p(\cdot)$; for each $i = \tau - 1, \ldots, 0$, suppose that $V_{i+1} = v$ and $E_{i+1} = e$ are known, and sample V_i according to the density $p(\cdot|v, e)$. This ultimately yields the desired sample for V_0.

This generalized sampling procedure is depicted as a graphical model in Figure 1.4. Note that a genealogical event $e \in \mathcal{E}(u)$ operates on *labeled* haplotypes. Intuitively, a realization of the discrete-time Markov process is a typed genealogy $\mathcal{A}_{\mathbf{n}}$ with timing information removed. In the subsequent sections and chapters, we provide concrete examples of the densities associated with this procedure.

Finally, we remark that it is often convenient to interpret a coalescent process as a *genealogical process*. In this context, we envision a labeled *lineage* associated with each ancestral haplotype,

tracing a path backward in time to produce the genealogy $\mathcal{A}_{\hat{n}}$. Genealogical events, such as a coalescence event, then affect one or more lineages directly. For example, in the case of Kingman's coalescent, the genealogical process is succinctly described by stating that each pair of lineages coalesce with rate 1. Thus, while there are i remaining lineages in $\mathcal{A}_{\hat{n}}$, each associated with an ancestral haplotype, the total rate of coalescence is $\binom{i}{2}$, and the process is identical to that described above. For more complex coalescent processes, incorporating mutation, recombination, and migration, such a genealogical interpretation, though formally identical, provides a more concise and intuitive description of the process.

Sampling probabilities

Now let $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ be a sample configuration. As for the Wright-Fisher diffusion, we are interested in determining the ordered sampling probability $q(\mathbf{n})$ associated with the sampling procedure described above. Intuitively, this can be accomplished by integrating over all possible genealogies, as sampled by the above procedure, that are consistent with $\mathbf{n}$. Due to the Markov structure of the procedure, it is generally possible to factor the computation to obtain a recursion for the ordered sampling probability. This technique is generally referred to as the backward/forward procedure, which we here derive in some generality.

We assume that the haplotypes in $\mathbf{n}$ are ordered, or equivalently that each haplotype is uniquely labeled. The associated labeled untyped configuration is denoted by $\hat{n}$, and $q(\mathbf{n})$ is then the probability of $V_0 = \mathbf{n}$ conditioned on $U_0 = \hat{n}$. Partitioning with respect to the most recent genealogical event $E_1 = e \in \mathcal{E}(\hat{n})$,

$$q(\mathbf{n}) = \Pr(V_0 = \mathbf{n}|U_0 = \hat{n}) = \sum_{e\in\mathcal{E}(\hat{n})} \Pr(V_0 = \mathbf{n}|U_0 = \hat{n}, E_1 = e)p(e|\hat{n}) \tag{1.44}$$

Recall that $U_0 = \hat{n}$ and $E_1 = e$ uniquely determine the previous untyped configuration $U_1 = e(\hat{n})$. Thus, partitioning with respect to $V_1 = \mathbf{n}'$ such that $\mathbf{n} \in \mathcal{V}(\mathbf{n}', e)$,

$$\begin{aligned}\Pr(V_0 = \mathbf{n}|U_0 = \hat{n}, E_1 = e) &= \Pr(V_0 = \mathbf{n}|U_0 = \hat{n}, E_1 = e, U_1 = e(\hat{n})) \\ &= \sum_{\mathbf{n}':\mathbf{n}\in\mathcal{V}(\mathbf{n}',e)} p(\mathbf{n}|\mathbf{n}', e)\Pr(V_1 = \mathbf{n}'|U_1 = e(\hat{n}))\end{aligned} \tag{1.45}$$

where the final equality makes use of two conditional independence assertions. Finally, the untyped configuration associated with $\mathbf{n}'$ must be $e(\hat{n})$, and so by time homogeneity,

$$\Pr(V_1 = \mathbf{n}'|U_1 = e(\hat{n})) = \Pr(V_0 = \mathbf{n}'|U_0 = e(\hat{n})) = q(\mathbf{n}'). \tag{1.46}$$

Putting these results together, we obtain the desired recursive expression for $q(\mathbf{n})$

$$q(\mathbf{n}) = \sum_{e\in\mathcal{E}(\hat{n})} p(e|\hat{n}) \sum_{\mathbf{n}':\mathbf{n}\in\mathcal{V}(\mathbf{n}',e)} p(\mathbf{n}|\mathbf{n}', e)q(\mathbf{n}'). \tag{1.47}$$

Recall that we have constructed the coalescent as a limit of the discrete Wright-Fisher process, and that the same limit was used to construct the Wright-Fisher diffusion. We therefore expect to obtain an identical recursion for $q(\mathbf{n})$; in the subsequent sections, we show that this is the case.

1.3.2 Multiple-locus, single-deme

Recall from Section 1.2.2 that the discrete Wright-Fisher process can be generalized to haplotypes comprising multiple loci, and allowing for mutation to occur at locus $\ell \in L$ with probability u_ℓ, and recombination to occur at breakpoint $b \in B$ with probability $r \cdot r_b$. Viewing the process backward in time, two or more haplotypes in a given generation may have common parental haplotypes in the previous generation; such coalescence events decrease the number of haplotypes ancestral to a sample. In contrast, haplotypes formed by recombination have two parental haplotypes in the previous generation, and thereby increase the number of haplotypes ancestral to a sample. Mutation does not affect the number of haplotypes ancestral to a sample.

It is possible to directly obtain the genealogy for a labeled untyped configuration $\hat{n}$ by considering the discrete Wright-Fisher process backward in time, assuming a finite population of $2N$ haplotypes. By scaling time in units of $2N$ generations and considering the limit $N \to \infty$, a process similar to Kingman's coalescent is obtained, which incorporates both mutation and recombination (Hudson, 1983). As in the Wright-Fisher diffusion, it is necessary to assume that the mutation and recombination probabilities vary inversely with N, so that for all $\ell \in L$ and $b \in B$, $4Nu_\ell \to \theta_\ell$ and $4Nrr_b \to \rho_b$ as $N \to \infty$; θ_ℓ and ρ_b are the scaled mutation and recombination rates, respectively. The resulting stochastic process is the *coalescent with recombination*, and has the following genealogical interpretation,

Coalescence: Each pair of lineages coalesce with rate 1.

Mutation: Each lineage undergoes mutation at locus $\ell \in L$ with rate $\theta_\ell/2$ according to the stochastic matrix $\Phi^{(\ell)}$.

Recombination: Each lineage undergoes recombination at breakpoint $b \in B$ with rate $\rho_b/2$.

When a recombination event occurs, the number of lineages increases by 1. Thus, the number of ancestral lineages is a birth-death process; when there are i ancestral lineages the process has death rate $\binom{i}{2}$, corresponding to coalescence events, and birth rate $i \cdot \sum_{b \in B} \rho_b/2$, corresponding to recombination events. The process continues until a single ancestral lineage, the MRCA, remains. The resulting untyped genealogy $\mathcal{A}_{\hat{n}}$ is no longer a bifurcating tree, but rather a graph, known as the *ancestral recombination graph* (ARG). See Figure 1.5(a) for an illustration of an ARG.

Observe that, unlike recombination events, mutation events do not affect the underlying *topology* of the ARG $\mathcal{A}_{\hat{n}}$. It is therefore equivalent to sample an ARG using the following two step procedure: first, sample the ARG topology using the coalescence with recombination process without mutation events; second, realize the mutation events at each locus $\ell \in L$ as a Poisson process on the underlying topology with rate $\theta_\ell/2$. Importantly, given an ARG topology, it is straightforward to integrate over the possible realizations of the mutation process; the state space of ARGs can therefore be reduced in the statistical inference setting.

The effect of recombination within the ARG $\mathcal{A}_{\hat{n}}$ is to produce alternative genealogies for the loci to left and right of the recombination breakpoint $b \in B$. Consequently, for any locus $\ell \in L$, there is embedded within the ARG $\mathcal{A}_{\hat{n}}$ a marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ describing the genealogical relationship of the haplotypes at the single locus ℓ. The marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ can be recovered by tracing each lineage backward in time, starting from the present; when a recombination event is encountered, only the ancestral lineage associated with locus ℓ is retained. Thus, a marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ is once again a bifurcating tree, as illustrated in Figure 1.5(b).

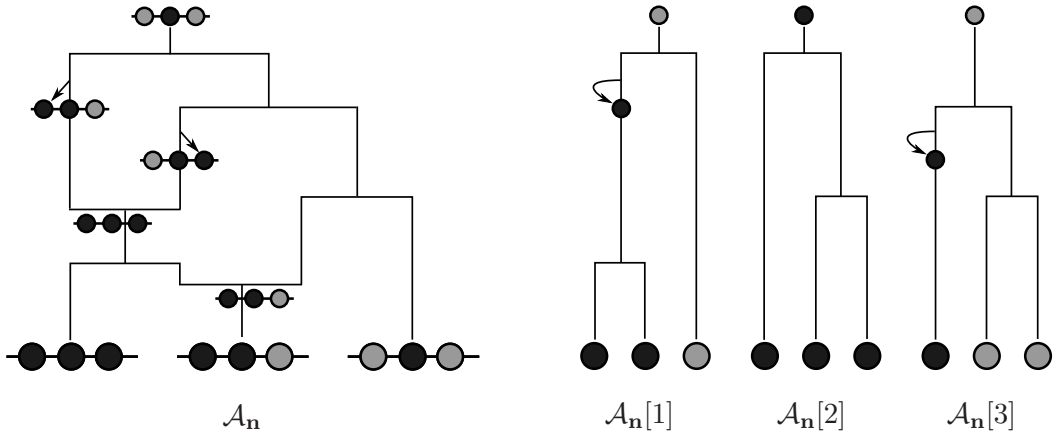


Figure 1.5. An illustration of an ARG genealogy and the associated marginal genealogies. (a) A typed ARG $\mathcal{A}_{\mathbf{n}}$ for a configuration $\mathbf{n}$ of 3-locus haplotypes, with $|\mathbf{n}| = 3$. Mutation events, along with the locus and resulting haplotype, are indicated by small arrows. Recombination events have occurred when a single descendant lineage has two ancestral lineages; the recombination breakpoint is indicated by the location of vertical segment relative to the resulting haplotype. It can be verified that the sample $\mathbf{n}$ is obtained by starting at the MRCA and tracing the type of each lineage forward in time. (b) The marginal genealogies $\mathcal{A}_{\mathbf{n}}[\ell]$ associated with each locus $\ell \in L$, obtained by considering only those lineages ancestral to the sample at locus ℓ. Each marginal genealogy is a bifurcating tree, and is correlated with other marginal genealogies by the coalescent with recombination process.

Given an untyped genealogy $\mathcal{A}_{\hat{n}}$, the type of the MRCA can be directly sampled and propagated forward in time, yielding a typed configuration $\mathbf{n}$ and the corresponding typed genealogy $\mathcal{A}_{\mathbf{n}}$. As described in Section 1.3.1, the time information within $\mathcal{A}_{\hat{n}}$ is not used to generate $\mathbf{n}$, and so it is only necessary to directly sample the genealogical events of $\mathcal{A}_{\hat{n}}$. Starting with an untyped configuration $\hat{n}$, the possible genealogical events $\mathcal{E}(\hat{n})$ include coalescence, mutation, and recombination. Let $e \in \mathcal{E}(\hat{n})$ be a genealogical event, and suppose $\mathbf{n}'$ is a typed configuration with associated untyped configuration $e(\hat{n})$,

Coalescence: Suppose $e \in \mathcal{E}(\hat{n})$ is a coalescence event. The untyped configuration $e(\hat{n})$ is derived from $\hat{n}$ by replacing the appropriate two labeled haplotypes with a single labeled haplotype, so that $|e(\hat{n})| = |\hat{n}| - 1$. Moreover $\mathcal{V}(\mathbf{n}', e)$ comprises a single typed configuration derived from $\mathbf{n}'$ by replacing the appropriate labeled haplotype $h \in \mathcal{H}$ with two identical labeled haplotypes,

$$\mathcal{V}(\mathbf{n}', e) = \{\mathbf{n}' - \mathbf{e}_h + \mathbf{e}_h + \mathbf{e}_h\} = \{\mathbf{n}' + \mathbf{e}_h\}. \tag{1.48}$$

Mutation: Suppose $e \in \mathcal{E}(\hat{n})$ is a mutation event at locus $\ell \in L$. The untyped configuration $e(\hat{n})$ is derived from $\hat{n}$ by replacing the appropriate labeled haplotype with a labeled haplotype, so that $|e(\hat{n})| = |\hat{n}|$. Moreover, $\mathcal{V}(\mathbf{n}', e)$ comprises a typed configuration for each allele $a \in A_\ell$, derived from $\mathbf{n}'$ by replacing the appropriate labeled haplotype $h \in \mathcal{H}$ with the labeled haplotype $\mathcal{M}_\ell^a(h)$,

$$\mathcal{V}(\mathbf{n}', e) = \{\mathbf{n}' - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}_\ell^a(h)} : a \in A_\ell\}, \tag{1.49}$$

and $p(\mathbf{n}' - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}_\ell^a(h)}|\mathbf{n}', e) = \Phi^{(\ell)}_{h[\ell],a}$.

Recombination: Suppose $e \in \mathcal{E}(\hat{n})$ is a recombination event at breakpoint $b \in B$. The untyped configuration $e(\hat{n})$ is derived from $\hat{n}$ by replacing the appropriate labeled haplotype with two labeled haplotypes, so that $|e(\hat{n})| = |\hat{n}| + 1$. Moreover $\mathcal{V}(\mathbf{n}', e)$ comprises a single typed configuration derived from $\mathbf{n}'$ by replacing the appropriate two labeled haplotypes $h, h' \in \mathcal{H}$ with the labeled haplotype $\mathcal{R}_b(h, h')$,

$$\mathcal{V}(\mathbf{n}', e) = \{\mathbf{n}' - \mathbf{e}_h - \mathbf{e}_{h'} + \mathbf{e}_{\mathcal{R}_b(h,h')}\}. \tag{1.50}$$

Finally, supposing that $|\hat{n}| = n$, the density $p(\cdot|\hat{n})$ is obtained considering the minimum of the exponential random variables associated with each event,

$$p(e|\hat{n}) = \begin{cases} 2/\mathcal{N}, & \text{for } e \text{ coalescence of two lineages,} \\ \theta_\ell/\mathcal{N}, & \text{for } e \text{ mutation of a lineage at locus } \ell \in L, \\ \rho_b/\mathcal{N}, & \text{for } e \text{ recombination of a lineage at breakpoint } b \in B, \end{cases} \tag{1.51}$$

where the normalizing constant $\mathcal{N} = n\big(n - 1 + \sum_{\ell \in L} \theta_\ell + \sum_{b \in B} \rho_b\big)$ is twice the total rate associated with all events. Using these densities, sampling a typed haplotype configuration proceeds by first sampling the events of an untyped genealogy $\mathcal{A}_{\hat{n}}$, sampling a type for the MRCA, and stochastically propagating this type down the genealogy.

Having characterized the sampling process associated with the coalescent with recombination, the technique described in Section 1.3.1 yields the following result,

Proposition 1.7. *Let $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$ with $|\mathbf{n}| = n$. Then the ordered sampling probability $q(\mathbf{n})$ obtained using the coalescent-based method in Section 1.3.1 is given by the following recursion*

$$\begin{aligned} q(\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{h \in \mathcal{H}} n_h \Bigg\{ & (n_h - 1) q(\mathbf{n} - \mathbf{e}_h) \\ & + \sum_{\ell \in L} \theta_\ell \sum_{a \in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} q(\mathbf{n} - \mathbf{e}_h + q\mathcal{M}_\ell^a(h)) \\ & + \sum_{b \in B} \rho_b \sum_{h' \in \mathcal{H}} q(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h))}) \Bigg\}, \end{aligned} \tag{1.52}$$

where $\mathcal{N} = n\big(n - 1 + \sum_{\ell \in L} \theta_\ell + \sum_{b \in B} \rho_b\big)$.

Proof. We use the technique described in Section 1.3.1. Define $\hat{n}$ to be the labeled untyped configuration associated with an arbitrary labeling of $\mathbf{n}$. Then we consider each event $e \in \mathcal{E}(\hat{n})$,

Coalescence: Suppose $e \in \mathcal{E}(\hat{n})$ is a coalescence event, specifying two labeled haplotypes $h, h' \in \mathcal{H}$ in $\mathbf{n}$. Since coalescence can only occur between identical haplotypes, $\{\mathbf{n}' : \mathbf{n} \in \mathcal{V}(\mathbf{n}', e)\} = \{\mathbf{n} - \mathbf{e}_h\}$ if $h = h'$ and is otherwise empty. As a result,

$$\Pr(V_0 = \mathbf{n}|U_0 = \hat{n}, E_1 = e) = \delta_{h,h'} \cdot q(\mathbf{n} - \mathbf{e}_h). \tag{1.53}$$

Mutation: Suppose $e \in \mathcal{E}(\hat{n})$ is a mutation event at locus $\ell \in L$, specifying the labeled haplotype $h \in \mathcal{H}$ in $\mathbf{n}$. Then $\{\mathbf{n}' : \mathbf{n} \in \mathcal{V}(\mathbf{n}', e)\} = \{\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}_\ell^a(h)} : a \in A_\ell\}$, and as a result,

$$\Pr(V_0 = \mathbf{n}|U_0 = \hat{n}, E_1 = e) = \sum_{a \in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} q(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}_\ell^a(h)}). \tag{1.54}$$

Recombination: Suppose $e \in \mathcal{E}(\hat{n})$ is a recombination event at locus $b \in L$, specifying the labeled haplotype $h \in \mathcal{H}$ in $\mathbf{n}$. Then $\{\mathbf{n}' : \mathbf{n} \in \mathcal{V}(\mathbf{n}', e)\} = \{\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h)} : h' \in \mathcal{H}\}$, and as result,

$$\Pr(V_0 = \mathbf{n} | U_0 = \hat{n}, E_1 = e) = \sum_{h' \in \mathcal{H}} q(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h)}). \tag{1.55}$$

The latter expression in each case is obtained by using (1.45) in conjunction with the known expressions for $p(\mathbf{n}|\mathbf{n}', e)$. Recall that each genealogical event $e \in \mathcal{E}(\hat{n})$ specifies haplotypes according to a *labeling*, and without regard to type. Thus, using the general recursion (1.47), via (1.44), in conjunction with the known density (1.51), the desired recursion (1.52) is obtained. □

Recall that we constructed coalescent with recombination as a limit of the discrete Wright-Fisher process, and that precisely the same limit was used to construct the Wright-Fisher diffusion; it is therefore reassuring that the recursion for the ordered sampling probability $q(\mathbf{n})$ obtained from the coalescent-based approach (1.52) is identical to that obtained from the diffusion-based approach (1.16). It is nonetheless remarkable that such different methodologies, reflecting entirely complementary interpretations, can be used to deduce the same result.

As in Section 1.2.2, explicit evaluation of $q(\mathbf{n})$ is not possible by repeated application of (1.52). We therefore consider a modification to the coalescent with recombination that directly produces a reduced recursion amenable to explicit evaluation. To this end, observe that, due to intervening recombination events, it is possible for a locus on a particular lineage within an untyped ARG $\mathcal{A}_{\hat{n}}$ to have no descendant loci in the untyped configuration $\hat{n}$; we describe such loci as *non-ancestral*. In sampling a typed haplotype configuration $\mathbf{n}$ associated with the untyped ARG $\mathcal{A}_{\hat{n}}$, non-ancestral loci can be left unspecified as, by definition, their type has no effect on $\mathbf{n}$. It is therefore unnecessary for the ARG to encode the genealogical history for such non-ancestral loci.

We modify the coalescent with recombination to explicitly incorporate the ancestral state of the loci on each lineage of the untyped ARG $\mathcal{A}_{\hat{n}}$. Beginning with the untyped configuration $\hat{n}$, every locus is ancestral by definition. Proceeding backward in time, the ancestral state of each lineage can be determined as follows,

- Given that a lineage undergoes recombination at breakpoint $b = (\ell, \ell + 1) \in B$, the set of ancestral loci for the two ancestral lineages is the intersection of set of ancestral loci for the descendant lineage with the sets $1 : \ell$ and $\ell + 1 : k$, respectively.
- Given a coalescence between two lineages, the set of ancestral loci of ancestral lineage is the union of the sets of ancestral loci of the two descendant lineages.

As stated above, it is unnecessary for the ARG to encode the genealogical history of non-ancestral loci, and we can therefore augment the ordinary coalescent model with the following controls,

- Mutation events at a non-ancestral locus of an untyped lineage are not allowed.
- Recombination events that produce an untyped lineage that is entirely non-ancestral are not allowed.

Using the modified coalescent with recombination, a *reduced* ARG is obtained. As before, it is then possible to sample a type for the MRCA haplotype and stochastically propagate it forward in

time. By construction, this reduced process yields the same distribution on sample configurations as the full process. The method of Section 1.3.1 applied to the modified coalescent then yields the following result,

Proposition 1.8. *Let* $\mathbf{n} = (n_g)_{g\in\mathcal{G}}$ *with* $|\mathbf{n}| = n$. *Then the ordered sampling probability* $q(\mathbf{n})$ *obtained using the coalescent-based method in Section 1.3.1 in conjunction with the reduced coalescent with recombination is given by the following recursion*

$$\begin{aligned} q(\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{g\in\mathcal{G}} n_g \Bigg\{ & \sum_{g'\in\mathcal{G}: g' \curlywedge g} (n_{g'} - \delta_{g,g'}) q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{C}(g,g')}) \\ & + \sum_{\ell\in L(g)} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,g[\ell]} q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}^a_\ell(g)}) \\ & + \sum_{b\in B(g)} \rho_b q(\mathbf{n} - \mathbf{e}_g + \mathbf{e}_{\mathcal{R}^-_b(g)} + \mathbf{e}_{\mathcal{R}^+_b(g))}) \Bigg\}, \end{aligned} \tag{1.56}$$

where $\mathcal{N} = \sum_{g\in\mathcal{G}} n_g \big(n - 1 + \sum_{\ell\in L(g)} \theta_\ell + \sum_{b\in B(g)} \rho_b\big)$.

Proof. As described above, each labeled untyped haplotype contains additional information about which loci are non-ancestral. Note that for a labeled typed configuration $\mathbf{n}$, those loci that are unspecified are considered non-ancestral in the corresponding untyped configuration as, for computing the sampling probability $q(\mathbf{n})$, their specific allelic value is irrelevant. Using this observation and the modified genealogical process described above, the proof of this proposition is analogous to the proof of Proposition 1.7. □

Once again, the expression (1.56) derived by the coalescent-based methodology is identical to the expression (1.20) derived from the diffusion-based methodology. Finally, recall the mathematical simplification (1.25) obtained when using a PIM model; given a mutation at locus $\ell \in L$, the specified allele at locus ℓ in the descendant haplotype is replaced with an unspecified allele in the ancestral haplotype. By analogy with the reduced ARG, we expect that locus ℓ is non-ancestral in the ancestral haplotype. Indeed, the type of the descendant allele does not depend on the ancestral allele, by definition of a PIM model, and so the ancestral haplotype is formally non-ancestral at locus ℓ. It is thus possible to refine the genealogical process for a PIM model, and so obtain the mathematical simplification (1.25) genealogically.

1.3.3 Multiple-locus, multiple-deme

Recall from Section 1.2.3 that the discrete Wright-Fisher process can be further generalized to a structured population with migration, for which there exist a finite set of $\mathcal{D}$, and the number of haplotypes in deme $d \in \mathcal{D}$ is given by N_d. To allow for migration between demes, in sampling a haplotype in deme $d \in \mathcal{D}$, the parental deme $d' \in \mathcal{D}$ is sampled with probability $v_{dd'}$.

It is possible to directly obtain the genealogy for a sample of *untyped* haplotypes by considering the discrete Wright-Fisher process backward in time, assuming a finite population of $2N$ haplotypes. By scaling time in units of $2N$ generations and considering the limit $N \to \infty$, a coalescent process is obtained, which incorporates mutation, recombination, and migration (Notohara, 1990). As in the Wright-Fisher diffusion, it is necessary to assume that $N_d/N \to \kappa_d$ and $4Nv_{dd'} \to \upsilon_{dd'}$ for all $d, d' \in \mathcal{D}$ with $d' \neq d$, where κ_d is the relative deme size and $\upsilon_{dd'}$ is the scaled migration rate.

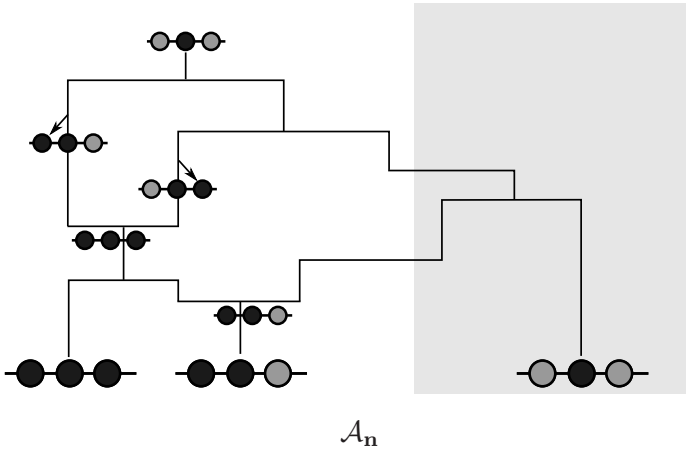


Figure 1.6. An illustration of population structured ARG $\mathcal{A}_{\mathbf{n}}$ for a configuration $\mathbf{n}$ of 3-locus haplotypes, with $|\mathbf{n}| = 3$, in two demes. The first deme, from which 2 haplotypes are sampled, is shown with a white background, and the second deme, from which 1 haplotype is sampled, with a light grey background. Mutation and recombination events are indicated as in Figure 1.5, and migration events are indicated by a horizontal transition of a lineage from one deme into another. It can be verified that the sample $\mathbf{n}$ is obtained by starting at the MRCA and tracing the type of each lineage forward in time.

The resulting stochastic process is the coalescent with recombination and migration, and has the following genealogical interpretation. Within each deme $d \in \mathcal{D}$,

Coalescence: Each pair of lineages coalesce with rate κ_d^{-1}.

Mutation: Each lineage undergoes mutation at locus $\ell \in L$ with rate $\theta_\ell/2$ according to the stochastic matrix $\Phi^{(\ell)}$.

Recombination: Each lineage undergoes recombination at breakpoint $b \in B$ with rate $\rho_b/2$.

Migration: Each lineage migrates to deme d' with rate $\upsilon_{dd'}/2$.

The outcome of this process is a generalized ARG, within which each lineage resides in a particular deme, as illustrated in Figure 1.6. Coalescence events can only occur between lineages in the same deme, and recombination events produces ancestral lineages in the same deme as the descendant lineage. Finally, migration events have the effect of moving a lineage, backward in time, from one deme into another.

The procedure for sampling described in Section 1.3.2 can be generalized to this setting by incorporating a genealogical event for migration. In addition, it is necessary to label haplotypes in both typed and untyped configurations by the deme in which they reside. Let $\hat{n}$ be such an untyped configuration, and $e \in \mathcal{E}(\hat{n})$ a genealogical event. Supposing that e is a coalescence, mutation, or recombination event, the description given in Section 1.3.2 suffices. Otherwise,

Migration: Suppose $e \in \mathcal{E}(\hat{n})$ is a migration event from $d \in \mathcal{D}$ to $d' \in \mathcal{D}$, backward in time. The untyped configuration $e(\hat{n})$ is derived from $\hat{n}$ by replacing the appropriate labeled untyped haplotype in deme d with a labeled untyped haplotype in deme d'. Given a typed configuration $\mathbf{n}'$ with associated untyped configuration $e(\hat{n})$, $\mathcal{V}(\mathbf{n}', e)$ comprises a single configuration

derived from $\mathbf{n}'$ by replacing the appropriate labeled haplotype $h \in \mathcal{H}$ in deme d' with an identical labeled haplotype in deme d,

$$\mathcal{V}(\mathbf{n}', e) = \{\mathbf{n}' - \mathbf{e}_{d',h} + \mathbf{e}_{d,h}\}. \tag{1.57}$$

Supposing that $|\hat{n}| = n$ and $|\hat{n}_d| = n_d$ for all $d \in \mathcal{D}$, the density $p(\cdot|\hat{n})$ is obtained considering the minimum of the exponential random variables associated with each event,

$$p(e|\hat{n}) = \begin{cases} 2\kappa_d^{-1}/\mathcal{N}, & \text{for } e \text{ coalescence of two lineages in deme } d \in \mathcal{D}, \\ \theta_\ell/\mathcal{N}, & \text{for } e \text{ mutation of a lineage at locus } \ell \in L, \\ \rho_b/\mathcal{N}, & \text{for } e \text{ recombination of a lineage at breakpoint } b \in B, \\ \upsilon_{dd'}/\mathcal{N}, & \text{for } e \text{ migration of a lineage from deme } d \text{ to deme } d', \end{cases} \tag{1.58}$$

where the normalizing constant $\mathcal{N} = \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} n_{d,h}\big((n_d - 1)\kappa_d^{-1} + \sum_{\ell\in L} \theta_\ell + \sum_{b\in B} \rho_b + \upsilon_d\big)$ is twice the total rate associated with all events. Having characterized the sampling process associated with the coalescent with recombination, the technique described in Section 1.3.1 yields the following result,

Proposition 1.9. *Let $\mathbf{n} = (n_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$ be a structured sample configuration, with $|\mathbf{n}| = n$ and $|\mathbf{n}_d| = n_d$ for each $d \in \mathcal{D}$. Then the ordered sampling probability $q(\mathbf{n})$ obtained using the coalescent-based method in Section 1.3.1 is given by the following recursion*

$$\begin{aligned} q(\mathbf{c}) = \frac{1}{\mathcal{N}} \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} n_{d,h} \Bigg\{ & (n_{d,h} - 1)\kappa_d^{-1} q(\mathbf{n} - \mathbf{e}_{d,h}) \\ & + \sum_{\ell\in L} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} q(\mathbf{n} - \mathbf{e}_{d,h} + \mathbf{e}_{d,\mathcal{M}^a_\ell(h)}) \\ & + \sum_{b\in B} \rho_b \sum_{h'\in\mathcal{H}} q(\mathbf{n} - \mathbf{e}_{d,h} + \mathbf{e}_{d,\mathcal{R}_b(h,h')} + \mathbf{e}_{d,\mathcal{R}_b(h',h)}) \\ & + \sum_{\substack{d'\in\mathcal{D} \\ d'\neq d}} \upsilon_{dd'} q(\mathbf{n} - \mathbf{e}_{d,h} + \mathbf{e}_{d',h}) \Bigg\} \end{aligned} \tag{1.59}$$

where $\mathcal{N} = \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} n_{d,h}\big((n_d - 1)\kappa_d^{-1} + \sum_{\ell\in L} \theta_\ell + \sum_{b\in B} \rho_b + \upsilon_d\big)$.

Proof. We use the technique described in Section 1.3.1 and exemplified in the proof of Proposition 1.7. Define $\hat{n}$ to be the labeled untyped configuration associated with an arbitrary labeling of $\mathbf{n}$, and let $e \in \mathcal{E}(\hat{n})$ be a genealogical event. If e is a coalescence, mutation, or recombination event, the description in the proof of Proposition 1.7 suffices; otherwise,

Migration: Suppose $e \in \mathcal{E}(\hat{n})$ is a migration event from deme $d \in \mathcal{D}$ to deme $d' \in \mathcal{D}$, backward in time, specifying the labeled haplotype $h \in \mathcal{H}$ in $\mathbf{n}$. Then $\{\mathbf{n}' : \mathbf{n} \in \mathcal{V}(\mathbf{n}', e)\} = \{\mathbf{n} - \mathbf{e}_{d,h} + \mathbf{e}_{d',h}\}$, and as result,

$$\Pr(V_0 = \mathbf{n} | U_0 = \hat{n}, E_1 = e) = q(\mathbf{n} - \mathbf{e}_{d,h} + \mathbf{e}_{d',h}). \tag{1.60}$$

Thus, using the general recursion (1.47), via (1.44), in conjunction with the known density (1.58), the desired recursion (1.59) is obtained. □

Once again, the recursion for the ordered sampling probability $q(\mathbf{n})$ obtained from the coalescent-based approach (1.59) is identical to that obtained from the diffusion-based approach (1.39). Moreover, though explicit evaluation of $q(\mathbf{n})$ is not possible by repeated application of (1.59), the reduced recursion (1.41) obtained using the diffusion generator technique can be obtained directly by considering a reduced coalescent with recombination and migration process, analogous to the one described in Section 1.3.2.

1.3.4 Sequentially Markov coalescent

Though an ARG $\mathcal{A}_{\hat{n}}$ is most naturally sampled starting in the present and proceeding backward in time, as described in Sections 1.3.2 and 1.3.3, Wiuf and Hein (1999) demonstrated that it is also possible to sample an ARG *sequentially*, beginning from the left-most locus and proceeding to the right. Though Wiuf and Hein describe this procedure for an infinite sites model, it is straightforward to translate the technique to the finite-sites, finite-alleles model of present interest. For simplicity, we consider the coalescent with recombination of Section 1.3.2, but note that the technique can be generalized to migration.

Recall that, embedded within an ARG $\mathcal{A}_{\hat{n}}$, there is a one-locus marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ describing the genealogical relationship of the configuration $\hat{n}$ at locus $\ell \in L$. We similarly define the embedded marginal ARG $\mathcal{A}_{\hat{n}}[1:\ell]$, which describes the genealogical relationship of the configuration $\hat{n}$ at the loci $\{1,\dots,\ell\}$. The marginal ARG $\mathcal{A}_{\hat{n}}[1:\ell]$ can be extracted from an ARG $\mathcal{A}_{\hat{n}}$ by preserving only those lineages that are ancestral to loci $\{1,\dots,\ell\}$. The key insight of Wiuf and Hein is that it is possible to sample the marginal ARGs directly, using a sequential process. Specifically, conditioned on the marginal ARG $\mathcal{A}_{\hat{n}}[1:\ell-1]$, the marginal ARG $\mathcal{A}_{\hat{n}}[1:\ell]$ can be sampled using the following process:

1. Recombination events, with breakpoint $b = (\ell-1,\ell) \in B$ are realized as a Poisson process with rate $\rho_b/2$ on the marginal genealogy $\mathcal{A}_{\hat{n}}[\ell-1]$ embedded within the marginal ARG for loci $\mathcal{A}_{\hat{n}}[1:\ell-1]$.

2. At each recombination event, a new lineage associated with locus ℓ is created. Proceeding backward in time, each of the new lineages associated with locus ℓ coalesce with the existing lineages in the marginal ARG $\mathcal{A}_{\hat{n}}[1:\ell-1]$, and with each other, at rate 1.

The resulting genealogy is a marginal ARG $\mathcal{A}_{\hat{n}}[1:\ell]$ consistent with $\mathcal{A}_{\hat{n}}[1:\ell-1]$. Observe that we have not incorporated the mutation process into the construction; as described in Section 1.3.2, mutation events at each locus can be incorporated subsequently. Thus, beginning with the marginal genealogy $\mathcal{A}_{\hat{n}}[1]$, sampled directly, according to Kingman's coalescent, it is possible to inductively sample the marginal ARG $\mathcal{A}_{\hat{n}}[1:\ell]$. Ultimately, this process yields the desired ARG $\mathcal{A}_{\hat{n}}$.

We next consider the sequence of marginal genealogies $(\mathcal{A}_{\hat{n}}[\ell])_{\ell\in L}$ embedded with the ARG $\mathcal{A}_{\hat{n}}$. Though the procedure proposed by Wiuf and Hein (1999) produces these marginal genealogies *sequentially*, the procedure is explicitly non-Markov. In constructing the marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$, though the first step depends only on the marginal genealogy $\mathcal{A}_{\hat{n}}[\ell-1]$, the second step depends on the entire marginal ARG $\mathcal{A}_{\hat{n}}[1:\ell-1]$. Intuitively, this dependence corresponds to the potential for coalescence events that link marginal genealogies at non-adjacent loci. McVean and Cardin (2005) showed that the non-Markov process can be well-approximated by a Markov process on the marginal genealogies. As for the full sequential construction, McVean and Cardin describe this procedure for an infinite sites model, and we translate to a finite-sites model. Given the marginal genealogy $\mathcal{A}_{\hat{n}}[\ell-1]$, the marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ can be approximately sampled as follows:

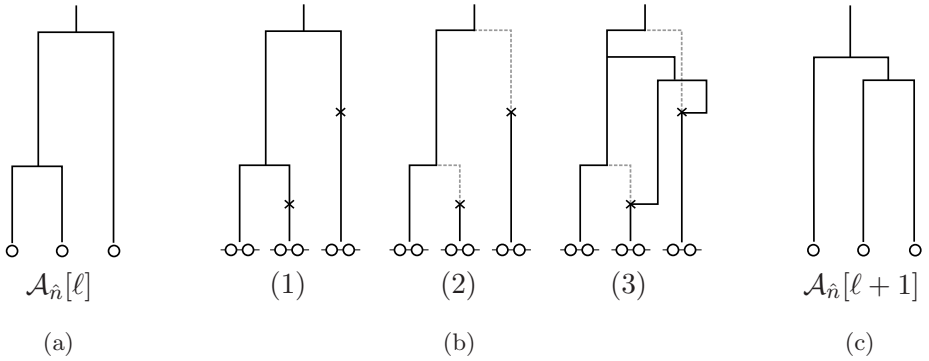


Figure 1.7. Illustration of Markov transition procedure for the SMC. (a) The untyped marginal genealogy $\mathcal{A}_{\hat{n}}[\ell-1]$ at locus $\ell-1 \in L$. (b) Conditional on the marginal genealogy $\mathcal{A}_{\hat{n}}[\ell-1]$, the marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ is sampled by (1) realizing recombination events, with breakpoint $b = (\ell-1, \ell) \in B$, as a Poisson process with rate $\rho_b/2$ on $\mathcal{A}_{\hat{n}}[\ell-1]$, (2) removing the lineages associated with locus $\ell-1$ ancestral to each recombination event, (3) creating a new lineage associated with locus ℓ at each breakpoint, and allowing each such lineage to coalesce with existing lineages in the marginal genealogy, and with each other, at rate 1. (c) The resulting untyped marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ at locus $\ell \in L$

1. Recombination events, with breakpoint $b = (\ell-1, \ell) \in B$ are realized as a Poisson process with rate $\rho_b/2$ on $\mathcal{A}_{\hat{n}}[\ell-1]$.

2. At each recombination event, the lineage associated with locus $\ell-1$ ancestral to the event is removed.

3. At each recombination event, a new lineage associated with locus ℓ is created. Proceeding backward in time, each of the new lineages associated with locus ℓ coalesce with the existing lineages in the marginal genealogy for locus $\ell+1$, and with each other, at rate 1.

See Figure 1.7 for an illustration. The sequence of marginal genealogies $(\mathcal{A}_{\hat{n}}[\ell])_{\ell \in L}$ is thus constructed directly, without requiring intermediate marginal ARGs for multiple loci. This process is called the *sequentially Markov coalescent* (SMC). Critically, though the resulting joint distribution on marginal genealogies is only approximate, due to the Markov assumption, the marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ at each locus $\ell \in L$ is correctly distributed as Kingman's coalescent. Moreover, it has been empirically demonstrated (McVean and Cardin, 2005; Marjoram and Wall, 2006) that the effect on the joint distribution of marginal genealogies using the SMC in place of the coalescent with recombination is minimal; McVean and Cardin (2005) conjecture, but do not formally prove, that SMC is equivalent to a modification to the coalescent with recombination in which coalescence is disallowed between lineages that do not contain *overlapping* ancestral loci.

Recall that mutation events can be realized on the marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ at locus $\ell \in L$ independently; and given the one-locus marginal genealogy, the allelic type for the MRCA can be sampled independently, and propagated forward in time, yielding a typed configuration at locus ℓ. Consequently, there is an evident procedure for sampling a typed configuration *sequentially*, starting from the left-most locus and proceeding to the right. While sampling directly from the coalescent with recombination requires explicit construction of the graph-like ARG, the procedure associated with the SMC is Markov on the tree-like marginal genealogies, and therefore confers considerable mathematical and computational simplicity. We note, however, that while it is straightforward to

sample the marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ conditioned on $\mathcal{A}_{\hat{n}}[\ell-1]$ using the procedure provided above, deriving an analytic form for the associated transition density remains a challenging open problem.

1.4 Conditional Sampling Distribution

Having described both the Wright-Fisher diffusion and the coalescent process in the previous sections, we now formally introduce the conditional sampling distribution (CSD). Conditioned on a sample configuration $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$, the CSD describes the probability distribution on one or more additionally sampled haplotypes. Intuitively, the configuration $\mathbf{n}$ is informative for the composition of the population, which is, in turn, informative for the additionally sampled haplotypes. We shall be interested in understanding this distribution, with the objective of deriving approximate distributions that facilitate computation.

Denoting a conditionally sampled configuration by $\mathbf{c} = (c_h)_{h\in\mathcal{H}}$, the *ordered* conditional sampling probability (CSP) is denoted $\pi(\mathbf{c}|\mathbf{n})$, and by the definition of conditional probability,

$$\pi(\mathbf{c}|\mathbf{n}) = \frac{q(\mathbf{c}+\mathbf{n})}{q(\mathbf{n})}. \tag{1.61}$$

Thus, it is possible to compute the CSP $\pi(\mathbf{c}|\mathbf{n})$ using the recursions provided in the previous sections for $q(\cdot)$. Making use of the exact analytic expression for $q(\mathbf{n})$ for a one-locus PIM model provided in Proposition 1.3,

Proposition 1.10 (Conditional Wright Sampling Formula)**.** *Let $\mathbf{c} = (c_a)_{a\in A}$ and $\mathbf{n} = (n_a)_{a\in A}$ be one-locus configurations with $|\mathbf{c}| = c$ and $|\mathbf{n}| = n$. For a PIM model, the CSP $\pi(\mathbf{c}|\mathbf{n})$ is given by*

$$\pi(\mathbf{c}|\mathbf{n}) = \frac{1}{(\theta+n)_{(c)}} \prod_{a\in A} (\theta\Phi_a + n_a)_{(c_a)}, \tag{1.62}$$

where $x_{(i)} = (x)(x+1)(x+2)\cdots(x+i-1)$ denotes a rising factorial.

Proof. Substitution of (1.30) into the CSP definition (1.61). □

Similarly, supposing $\rho_b = \rho$ for all $b \in B$, and considering the limit $\rho \to \infty$ described in Proposition 1.4,

Proposition 1.11. *Let $\mathbf{n} = (n_g)_{g\in\mathcal{G}}$ with $|\mathbf{c}| = c$ and $\mathbf{n} = (n_g)_{g\in\mathcal{G}}$ with $|\mathbf{n}| = n$, and suppose $\rho_b = \rho$ for all $b \in B$. Then in the limit that $\rho \to \infty$, the CSP $\pi(\mathbf{c}|\mathbf{n})$ can be decomposed as follows*

$$\pi(\mathbf{c}|\mathbf{n}) = \prod_{\ell\in L} \pi(\mathbf{c}[\ell]|\mathbf{n}[\ell]), \tag{1.63}$$

where $\mathbf{c}[\ell]$ and $\mathbf{n}[\ell]$ is the one-locus configuration induced by $\mathbf{c}$ and $\mathbf{n}$ at locus $\ell \in L$, and $\pi(\mathbf{c}[\ell]|\mathbf{n}[\ell])$ is the one-locus CSP.

Proof. Substitution of (1.31) into the CSP definition (1.61). □

Thus, computing the CSP for a k-locus configuration can be efficiently accomplished by computing the product of the CSPs for k one-locus configurations. Moreover, given a PIM model

at each locus, the resulting one-locus CSPs can be computed efficiently and exactly using (1.62), yielding an exact result.

Apart from these special cases, there are no known analytic formulas for computing the true CSP. By enumerating the finite set of configurations $\mathbf{c}$ and computing $\pi(\mathbf{c}|\mathbf{n})$ for each of them, it is also possible to sample from the true CSD. In contrast to the genealogical process described in Section 1.3 for the unconditional sampling distribution, however, there is not a known efficient procedure, genealogical or otherwise, for sampling from the true CSD.

As a result, exact computation for the true CSD using known methods is at least as challenging as the analogous computation for the unconditional sampling distribution. Nonetheless, we hope that by *approximating* the CSD, it is possible to obtain approximate, though computationally tractable, solutions for many population genetic problems of interest. As will be demonstrated in Chapter 4, in some cases it is even possible to correct these approximations using Monte Carlo techniques. Conditioned on the sample configuration $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ with $|\mathbf{n}| = n$, consider the following two extreme CSDs associated with a single haplotype,

Independence: The conditionally sampled haplotype is entirely independent of the previously sampled configuration $\mathbf{n}$. Letting $h \in \mathcal{H}$,

$$\hat{\pi}(\mathbf{e}_h|\mathbf{n}) = q(\mathbf{e}_h). \tag{1.64}$$

Complete Dependence: The conditionally sampled haplotype is chosen uniformly at random from the previously sampled configuration $\mathbf{n}$. Letting $h \in \mathcal{H}$,

$$\hat{\pi}(\mathbf{e}_h|\mathbf{n}) = \frac{n_h}{n}. \tag{1.65}$$

The first of these specifies that there is no dependence on the previously sampled configuration, which is trivially true when $n = 0$, and generally becomes a worse approximation with increasing n. On the other hand, the second specifies complete dependence on the previously sampled configuration, which is trivially true if $\mathbf{n}$ is precisely representative of the entire population, occurring in the limit $n \to \infty$; this approximation generally becomes worse with decreasing n, and ultimately is not defined for $n = 0$.

Hereafter, we consider approximate CSDs that, as the true CSD, are intermediate between these two extremes; the conditionally sampled haplotype should be similar to previously sampled haplotypes, with variation introduced by the processes of mutation and recombination. Intuitively, the recombination process breaks the conditionally sampled haplotype into several pieces, each similar to a single previously sampled haplotype, with additional variation introduced by the mutation process. The conditionally sampled haplotype is thus often referred to as an *imperfect mosaic* of the previously sampled haplotypes.

Several approximate CSDs following this general model have been proposed, three of which we introduce in some detail. It is important to note that these CSDs, though computationally appealing, have limited theoretical connection to the coalescent. In Chapter 2, we introduce a general methodology for constructing approximate CSDs directly from the Wright-Fisher diffusion, or alternatively from a genealogical process closely related to the coalescent with recombination.

1.4.1 Stephens and Donnelly

Stephens and Donnelly (2000) proposed the following CSD, which we denote by $\hat{\pi}_{\text{SD}}$, applicable

in the absence of recombination so that $\rho_b = 0$ for all $b \in B$. Let $\mathbf{n} = (\mathbf{n}_h)_{h\in\mathcal{H}}$ be a sample configuration with $|\mathbf{n}| = n$. Conditional on $\mathbf{n}$, a haplotype is sampled using the following procedure,

1. Choose a haplotype h from $\mathbf{n}$ uniformly at random.
2. Letting $\Theta = \sum_{\ell\in L}\theta_\ell$, mutate the haplotype a geometric number of times, with parameter $n/(n+\Theta)$; a mutation occurs at locus $\ell \in L$ with probability θ_ℓ/Θ, and according to stochastic mutation matrix $\Phi^{(\ell)}$.

Thus, as n increases, the number of mutations decreases, concordant with our earlier intuition. Letting $\eta \in \mathcal{H}$, it is possible to compute the CSP,

$$\hat{\pi}_{\text{SD}}(\mathbf{e}_\eta|\mathbf{n}) = \sum_{h\in\mathcal{H}} \frac{n_h}{n} \sum_{\mathbf{m}\in\mathbb{N}^k} F^{(n)}(h,\eta,\mathbf{m}), \tag{1.66}$$

where the vector $\mathbf{m} = (m_\ell)_{\ell\in L}$ indicates the number of mutations at each locus, and $F(h,\eta,\mathbf{m})$ is the probability of h mutating to η with $\mathbf{m}$ mutations,

$$F^{(n)}(h,\eta,\mathbf{m}) = \binom{m}{\mathbf{m}}\left[\prod_{\ell\in L}\left(\frac{\theta_\ell}{n+\Theta}\right)^{m_\ell}\left[\left(\Phi^{(\ell)}\right)^{m_\ell}\right]_{h[\ell],\eta[\ell]}\right]\frac{n}{n+\Theta}, \tag{1.67}$$

where $m = \sum_{\ell\in L} m_\ell$ and $\binom{m}{\mathbf{m}}$ is the multinomial coefficient. Though this form is mathematically elegant, it is challenging to compute numerically. Stephens and Donnelly observe that, by elementary properties of Poisson processes, the mutational procedure is equivalent to drawing a time $t \in \mathbb{R}_{\geq 0}$ from an exponential distribution with rate parameter n, and applying m_ℓ mutations at each locus $\ell \in L$, where the values of m_ℓ are independent and Poisson distributed with mean $\theta_\ell t$. Thus, the CSD $\hat{\pi}_{\text{SD}}(\mathbf{e}_\eta|\mathbf{n})$ can alternatively be expressed

$$\hat{\pi}_{\text{SD}}(\mathbf{e}_\eta|\mathbf{n}) = \sum_{h\in\mathcal{H}} \frac{n_h}{n}\int_{\mathbb{R}_{\geq 0}} ne^{-nt}\prod_{\ell\in L} G_\ell(h[\ell],\eta[\ell],t)dt, \tag{1.68}$$

where $G_\ell(a,a',t)$ is the probability of mutation from allele a to a' at locus ℓ,

$$G_\ell^{(n)}(a,a',t) = e^{-\theta_\ell t}\sum_{m=0}^{\infty}\frac{(\theta_\ell t)^m}{m!}\cdot\left[\left(\Phi^{(\ell)}\right)^m\right]_{a,a'}. \tag{1.69}$$

By using Gaussian quadrature, it is possible to approximate the integral in (1.68) as a summation over a finite number of values of t, and the value $G_\ell(a,a',t)$ can be numerically approximated for each such value of t. This provides a computationally tractable method for obtaining a highly accurate approximation to the CSP $\hat{\pi}_{\text{SD}}(\mathbf{e}_\eta|\mathbf{n})$.

Specialization to one-locus case

In the one-locus case, the space of haplotypes can be represented by the (finite) space of alleles $\mathcal{H} = A$, and each haplotype by a single allele $a \in A$. The single scaled mutation rate is represented by θ so that $\Theta = \theta$. Letting $\mathbf{n} = (n_a)_{a\in A}$ be a one-locus configuration, and $\alpha \in A$, the general solution (1.66) reduces to

$$\hat{\pi}_{\text{SD}}(\mathbf{e}_\alpha|\mathbf{n}) = \sum_{a\in A}\frac{n_a}{n}\sum_{m=0}^{\infty}\left(\frac{\theta}{n+\theta}\right)^m\frac{n}{n+\theta}\left[\Phi^m\right]_{a,\alpha}. \tag{1.70}$$

Moreover, for a PIM model, we have that $\Phi^m = \Phi$ for $m \geq 1$, and therefore,

$$\hat{\pi}_{\text{SD}}(\mathbf{e}_\alpha|\mathbf{n}) = \sum_{a \in A} \frac{n_a}{n} \frac{n}{n+\theta} \left(\delta_{\alpha,a} + \frac{\theta \Phi_\alpha}{n} \right) = \frac{n_\alpha + \theta \Phi_\alpha}{n+\theta} \tag{1.71}$$

Substituting $\mathbf{c} = \mathbf{e}_\alpha$ into (1.62), it can be verified that for the one-locus PIM model, $\hat{\pi}_{\text{SD}} = \pi$. Though this result is promising, it can be empirically demonstrated that $\hat{\pi}_{\text{SD}}$ is not generally exact, even in the one-locus case for a general model of mutation.

1.4.2 Fearnhead and Donnelly

Fearnhead and Donnelly (2001) proposed a generalization of the method of Stephens and Donnelly (2000) incorporating recombination, which we denote by $\hat{\pi}_{\text{FD}}$. Let $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$ be a sample configuration with $|\mathbf{n}| = n$. Conditional on $\mathbf{n}$, a haplotype is sampled using the following procedure,

1. Recombination occurs at each breakpoint $b \in B$ independently with probability $\rho_b/(n+\rho_b)$. The recombination process splits the haplotype into one or more intervals.

2. Each haplotype interval is sampled independently according to the procedure proposed by Stephens and Donnelly, and detailed above.

3. The sampled haplotype intervals are joined to produce a sampled haplotype.

As above, as n increases, the number of recombinations decreases, concordant with our earlier intuition. In order to compute the CSP, it is necessary to integrate over the possible realizations of recombination events, taking the product over the probabilities of each induced haplotype interval.

Considering a particular set of recombination events, and recalling the alternative interpretation (1.68) of $\hat{\pi}_{\text{SD}}$, each of the induced haplotype intervals is independently characterized by a haplotype chosen uniformly at random from $\mathbf{n}$, and a time chosen according to an exponential distribution with rate n. Because the recombination events are independent, the sequence of haplotype and time pairs associated with each locus is Markov. Making use of this observation, Fearnhead and Donnelly (2001) provide an efficient dynamic programming algorithm for computing the CSP. As in the corresponding CSP computation for $\hat{\pi}_{\text{SD}}$, this algorithm relies on Gaussian quadrature.

Finally, observe that when $\rho_b = 0$ for all $b \in B$, the recombination process does not split the sampled haplotype, and so $\hat{\pi}_{\text{FD}} = \hat{\pi}_{\text{SD}}$. Alternatively, suppose that $\rho_b = \rho$ for all $b \in B$; in the limit that $\rho \to \infty$, recombination occurs at each breakpoint almost surely, and therefore each locus $\ell \in L$ is independently sampled according to the one-locus CSD $\hat{\pi}_{\text{SD}}$. Additionally assuming a PIM model, recall that the one-locus CSP $\hat{\pi}_{\text{SD}} = \pi$, so that $\hat{\pi}_{\text{FD}} = \pi$.

1.4.3 Li and Stephens

Li and Stephens (2003) propose a straightforward modification to $\hat{\pi}_{\text{FD}}$, and we denote the resulting CSD by $\hat{\pi}_{\text{LS}}$. Let $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$ be a sample configuration with $|\mathbf{n}| = n$. Conditional on $\mathbf{n}$, a haplotype is sampled using the following procedure,

1. Recombination occurs at each breakpoint $b \in B$ independently with probability $1 - \exp(-\rho_b/n)$. The recombination process splits the haplotype into one or more intervals.

2. Each such haplotype interval is sampled independently by choosing a haplotype h from $\mathbf{n}$ uniformly at random, and mutating each locus $\ell \in L$ within the haplotype interval independently with probability $\theta_\ell/(\theta_\ell + n)$.

3. The sampled haplotype intervals are joined to produce a sampled haplotype.

As for $\hat{\pi}_{\text{FD}}$, the CSP associated with $\hat{\pi}_{\text{LS}}$ can be efficiently computed using a dynamic programming algorithm. Because Gaussian quadrature is not required, computation of the CSP associated with $\hat{\pi}_{\text{LS}}$ is a (small) constant factor faster than computation of the CSP associated with $\hat{\pi}_{\text{FD}}$. However, unlike $\hat{\pi}_{\text{FD}}$, $\hat{\pi}_{\text{LS}}$ is not identical to $\hat{\pi}_{\text{SD}}$ in the absence of recombination, and for a one-locus PIM model, $\hat{\pi}_{\text{LS}} \neq \hat{\pi}_{\text{SD}} = \pi$. We thus anticipate that $\hat{\pi}_{\text{LS}}$ is less accurate than $\hat{\pi}_{\text{FD}}$ in order to provide the aforementioned computational benefit. This claim is empirically investigated in Chapter 4.

Chapter 2

Theory

In this chapter, we describe two related techniques for obtaining an approximate conditional sampling distribution (CSD) in a principled way. The development of these techniques parallels the development of the sampling probabilities in Chapter 1. We first consider an approximation to the diffusion generator technique described in Section 1.2, and use it to derive an approximate CSD, $\hat{\pi}_{\text{PS}}$, for the coalescent with recombination, both with and without population structure and migration. We then consider a genealogical process for conditional sampling, closely related to the coalescent process described in Section 1.3, and show that the associated distribution is once again the CSD $\hat{\pi}_{\text{PS}}$. The genealogical process is of particular importance as it provides an intuitive generative process for $\hat{\pi}_{\text{PS}}$ in much the same way the coalescent serves as a generative process for the sampling distribution.

We derive recursive expressions for the conditional sampling probability (CSP) associated with $\hat{\pi}_{\text{PS}}$, for models of evolution incorporating mutation, recombination, and population structure. As for the sampling probabilities discussed in Chapter 1, explicit evaluation of the CSP by repeated application of the recursive expressions is computationally intractable for all but very small datasets. Guided by the genealogical process for $\hat{\pi}_{\text{PS}}$, we propose several genealogical approximations in order to improve the computational complexity of CSP evaluation. These approximations culminate with the sequentially Markov CSD $\hat{\pi}_{\text{SMC}}$, for which the sequence of *marginal conditional genealogies* is assumed to be Markov, analogous to the sequentially Markov coalescent described in Section 1.3.4. Finally, we relate the CSDs $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{SMC}}$ to previously-proposed CSDs, and conclude that $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{SMC}}$ more precisely model the true CSD.

2.1 Diffusion-Generator Approximation

The diffusion-generator approximation was introduced by De Iorio and Griffiths (2004a), where it was used to algebraically derive, directly from the diffusion, the one-locus CSD $\hat{\pi}_{\text{SD}}$, proposed by Stephens and Donnelly (2000); the same approximation has also been used (De Iorio and Griffiths, 2004b) to derive a one-locus CSD in the setting of structured populations. Griffiths et al. (2008) extended the diffusion-generator approximation to derive a two-locus CSD, including recombination. Their technique relies on an *ad hoc* symmetry argument, however, and cannot be generalized to more than two loci; moreover, their technique is limited to parent independent mutation (PIM) models.

More recently, Paul and Song (2010) generalized the diffusion-generator approximation to an

arbitrary number of loci and arbitrary finite-alleles mutation model. In this section, we describe the generalized diffusion-generator technique, and apply it to the general finite-locus finite-alleles settings, both with and without population structure. We show that, in the one-locus case, the resulting CSDs are the same as those derived by De Iorio and Griffiths (2004a,b).

2.1.1 Mathematical technique

Recall from Section 1.2 that the Wright-Fisher diffusion for a finite-locus finite-alleles model, for which the space of haplotypes is denoted $\mathcal{H}$, has state space given by the standard $\mathcal{H}$-simplex

$$\Delta = \left\{ \mathbf{x} = (x_h)_{h\in\mathcal{H}} \;\middle|\; x_h \geq 0 \text{ for all } h \in \mathcal{H} \text{ and } \sum_{h\in\mathcal{H}} x_h = 1 \right\}, \tag{2.1}$$

where x_h is the proportion of haplotype $h \in \mathcal{H}$. Letting $f : \Delta \to \mathbb{R}$ be an arbitrary, bounded, twice-differentiable function with continuous second derivatives, the diffusion generator can be decomposed into a summation

$$\mathscr{L} f(\mathbf{x}) = \sum_{h\in\mathcal{H}} \mathcal{L}_h \frac{\partial}{\partial x_h} f(\mathbf{x}), \tag{2.2}$$

where the form (1.8) of $\mathcal{L}_h$ depends on the infinitesimal mean (1.5) and covariance (1.6) associated with the Wright-Fisher diffusion. Let $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ be a haplotype configuration, and recall that the ordered sampling probability $q(\mathbf{n})$ can be expressed $q(\mathbf{n}) = \mathbb{E}[q(\mathbf{n}|\mathbf{X})]$ where the expectation is with respect to the stationary distribution of the Wright-Fisher diffusion, and $q(\mathbf{n}|\mathbf{x})$ is the ordered multinomial probability (1.12) of sampling $\mathbf{n}$ conditioned on haplotype proportions $\mathbf{x} \in \Delta$. Finally, applying a general result (1.9) for $f(\mathbf{x}) = q(\mathbf{n}|\mathbf{x})$,

$$\sum_{h\in\mathcal{H}} \mathbb{E}\left[\mathcal{L}_h \frac{\partial}{\partial x_h} q(\mathbf{n}|\mathbf{X}) \right] = 0, \tag{2.3}$$

and this result can be used to derive a recursive expression for $q(\mathbf{n})$. We now assume the existence of distribution and associated expectation operator $\hat{\mathbb{E}}$ such that (2.3) holds *component-wise*; that is, for an arbitrary $h \in \mathcal{H}$,

$$\hat{\mathbb{E}}\left[\mathcal{L}_h \frac{\partial}{\partial x_h} q(\mathbf{n}|\mathbf{X}) \right] = 0. \tag{2.4}$$

Observe that this is a stronger assertion than (2.3), and need not generally be true. We refer to this assumption as the *diffusion-generator approximation*; critically, this is precisely the assumption used by De Iorio and Griffiths (2004a,b), and is the only approximation required for the development of our approximate CSD. Let $\mathbf{c} = (c_h)_{h\in\mathcal{H}}$, and analogous to the definition (1.61) of the CSD π, define the approximate CSD $\hat{\pi}_{\text{PS}}$

$$\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n}) = \frac{\hat{q}(\mathbf{c}+\mathbf{n})}{\hat{q}(\mathbf{n})}, \tag{2.5}$$

where $\hat{q}(\mathbf{n}) = \hat{\mathbb{E}}[q(\mathbf{n}|\mathbf{X})]$ is an approximate sampling probability. Using the diffusion generator approximation (2.4), we propose the following *re-weighted* version of (2.3),

$$\begin{aligned} &\hat{\mathbb{E}}\left[\sum_{h\in\mathcal{H}} \frac{c_h}{c_h+n_h} \mathcal{L}_h \frac{\partial}{\partial x_h} q(\mathbf{c}+\mathbf{n}|\mathbf{X}) \right] \\ &\qquad = \sum_{h\in\mathcal{H}} \frac{c_h}{c_h+n_h} \hat{\mathbb{E}}\left[\mathcal{L}_h \frac{\partial}{\partial x_h} q(\mathbf{c}+\mathbf{n}|\mathbf{X}) \right] = 0, \end{aligned} \tag{2.6}$$

with the final equality by (2.4). Analogous to the way (2.3) produces a recursive equation for the sampling probability $q(\mathbf{n})$, the latter equation (2.6) produces a recursive equation for the approximate sampling probability $\hat{q}(\mathbf{c}+\mathbf{n})$. By construction, the resulting equation is recursive only on haplotypes within configuration $\mathbf{c}$; thus, dividing by $\hat{q}(\mathbf{n})$ and making use of definition (2.5) yields a recursive expression for the CSP $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$. As we shall see, the fact that the CSP $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$ is recursive only on the conditional sample $\mathbf{c}$ confers a critical computational benefit.

General mathematical results

Because the proposed CSP $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$ is approximate, it is reasonable to question whether it satisfies several important properties of the distribution π. For example, we can show that the approximate CSPs are properly normalized by considering an arbitrary $\mathbf{n}$ and $c > 0$. Summing over all ordered configurations of $\mathbf{c}$ with $|\mathbf{c}| = c$,

$$\begin{aligned}\sum_{\mathbf{c}:|\mathbf{c}|=c} \hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n}) &= \frac{1}{\hat{q}(\mathbf{n})} \sum_{\mathbf{c}:|\mathbf{c}|=c} \hat{q}(\mathbf{c}+\mathbf{n}) \\ &= \frac{1}{\hat{q}(\mathbf{n})} \hat{\mathbb{E}}\left[q(\mathbf{n}|\mathbf{X}) \sum_{\mathbf{c}:|\mathbf{c}|=c} q(\mathbf{c}|\mathbf{X}) \right] = \frac{1}{\hat{q}(\mathbf{n})} \hat{\mathbb{E}}\left[q(\mathbf{n}|\mathbf{X}) \right] = 1,\end{aligned} \tag{2.7}$$

where the penultimate equality is by the fact that $q(\cdot|\mathbf{x})$ is the properly normalized ordered multinomial distribution. Thus, $\hat{\pi}_{\text{PS}}(\cdot|\mathbf{n})$ is a probability distribution, and we can henceforth refer to $\hat{\pi}_{\text{PS}}$ as a CSD. Moreover, if $n = 0$, then the key generating equation (2.6) reduces to (1.9), and so the resulting CSD $\hat{\pi}_{\text{PS}}(\cdot|\mathbf{n})$ is actually *exact*; in this case, $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n}) = \pi(\mathbf{c}|\mathbf{n}) = q(\mathbf{c})$.

Letting $\mathbf{c}$ and $\mathbf{n}$ be arbitrary configurations, (2.6) does not depend on an ordering within the configuration $\mathbf{c}$. The derived ordered CSP $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$ is therefore *exchangeable* with respect to the conditionally sampled configuration $\mathbf{c}$, and so our convention of representing the ordered configuration $\mathbf{c}$ as an unordered vector is well-defined. Finally, we consider the exchangeability property for $\hat{q}$. Given configurations $\mathbf{c}$ and $\mathbf{n}$, and a haplotype h with $c_h > 0$, exchangeability would dictate that

$$\hat{q}(\mathbf{c}+\mathbf{n}) \stackrel{?}{=} \hat{q}((\mathbf{c}-\mathbf{e}_h)+(\mathbf{n}+\mathbf{e}_h)) = \hat{q}(\mathbf{c}'+\mathbf{n}'), \tag{2.8}$$

where $\mathbf{c}' = \mathbf{c} - \mathbf{e}_h$ and $\mathbf{n}' = \mathbf{n} + \mathbf{e}_h$. By looking at the form of (2.6), the key approximation generating the recursion for $\hat{q}$, the necessary exchangeability between $\mathbf{c}$ and $\mathbf{n}$ is not evident. In fact, it is simple to empirically demonstrate that in the general case $\hat{q}(\mathbf{c}+\mathbf{n}) \neq \hat{q}(\mathbf{c}'+\mathbf{n}')$. Even though our construction of the well-formed and exchangeable conditional distribution $\hat{\pi}_{\text{PS}}(\cdot|\mathbf{n})$ uses the distribution $\hat{q}$, the distribution $\hat{q}$ is not itself exchangeable, and therefore not well-defined for our convention of representing the ordered configuration as a vector. Let $\mathbf{n} = \mathbf{e}_{h_1} + \cdots + \mathbf{e}_{h_n}$, and σ be a permutation of $\{1, \ldots, n\}$; in general, we would like to write

$$\begin{aligned}\hat{q}(\mathbf{e}_{h_1} + \cdots + \mathbf{e}_{h_n}) &= \hat{q}(\mathbf{e}_{h_1}) \hat{\pi}_{\text{PS}}(\mathbf{e}_{h_2}|\mathbf{e}_{h_1}) \cdots \hat{\pi}_{\text{PS}}(\mathbf{e}_{h_n}|\mathbf{e}_{h_1} + \cdots + \mathbf{e}_{h_{n-1}}), \text{ and} \\ \hat{q}(\mathbf{e}_{h_{\sigma(1)}} + \cdots + \mathbf{e}_{h_{\sigma(n)}}) &= \hat{q}(\mathbf{e}_{h_{\sigma(1)}}) \hat{\pi}_{\text{PS}}(\mathbf{e}_{h_{\sigma(2)}}|\mathbf{e}_{h_{\sigma(1)}}) \cdots \hat{\pi}_{\text{PS}}(\mathbf{e}_{h_{\sigma(n)}}|\mathbf{e}_{h_{\sigma(1)}} + \cdots + \mathbf{e}_{h_{\sigma(n-1)}}),\end{aligned}$$

but as a consequence of this shortcoming, $\hat{q}(\mathbf{e}_{h_1} + \cdots + \mathbf{e}_{h_n}) \neq \hat{q}(\mathbf{e}_{h_{\sigma(1)}} + \cdots + \mathbf{e}_{h_{\sigma(n)}})$. Therefore, it is non-trivial to approximate the sampling probability $q(\mathbf{n}) \approx \hat{q}(\mathbf{n})$ using a decomposition into approximate CSDs, as the result will generally depend on the *ordering* of the sample and therefore on the ordering of the decomposition.

2.1.2 Multiple-locus, single-deme

Given the general form of diffusion-generator technique described, we derive the following result for multiple loci with recombination.

Theorem 2.1. *Let* $\mathbf{c} = (c_h)_{h\in\mathcal{H}}$ *with* $|\mathbf{c}| = c$, *and* $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ *with* $|\mathbf{n}| = n$. *Then the CSP* $\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n})$ *obtained using the approximate diffusion-generator technique described in Section 2.1.1 is given by the following recursive expression,*

$$\begin{aligned}\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}}\sum_{h\in\mathcal{H}} c_h\Bigg\{&(c_h+n_h-1)\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_h|\mathbf{n})\\ &+\sum_{\ell\in L}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,h[\ell]}\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_h+\mathbf{e}_{\mathcal{M}^a_\ell(h)}|\mathbf{n})\\ &+\sum_{b\in B}\rho_b\sum_{h'\in\mathcal{H}}\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_h+\mathbf{e}_{\mathcal{R}_b(h,h')}+\mathbf{e}_{\mathcal{R}_b(h',h))}|\mathbf{n})\Bigg\},\end{aligned} \tag{2.9}$$

where $\mathcal{N} = c\big(c+n-1+\sum_{\ell\in L}\theta_\ell+\sum_{b\in B}\rho_b\big)$.

Proof. Recalling the specifics of the diffusion generator (1.17), apply the key equation (2.6). In conjunction with the component-wise expectation (1.18), this yields

$$\begin{aligned}0 = \sum_{h\in\mathcal{H}} c_h\frac{1}{2}\Bigg\{&(c_h+n_h-1)\hat{q}(\mathbf{c}+\mathbf{n}-\mathbf{e}_h)\\ &+\sum_{\ell\in L}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,h[\ell]}\hat{q}(\mathbf{c}+\mathbf{n}-\mathbf{e}_h+\mathbf{e}_{\mathcal{M}^a_\ell(h)})\\ &+\sum_{b\in B}\rho_b\sum_{h'\in\mathcal{H}}\hat{q}(\mathbf{c}+\mathbf{n}-\mathbf{e}_h+\mathbf{e}_{\mathcal{R}_b(h,h')}+\mathbf{e}_{\mathcal{R}_b(h',h))})\\ &-\Bigg((c+n-1)+\sum_{\ell\in L}\theta_\ell+\sum_{b\in B}\rho_b\Bigg)\hat{q}(\mathbf{c}+\mathbf{n})\Bigg\}.\end{aligned} \tag{2.10}$$

Dividing by $\hat{q}(\mathbf{n})$ and using the definition (2.5) of $\hat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n})$, the desired result (2.9) is obtained. □

Observe that, as in Section 1.2.2, the system of linear equations resulting from repeated application of the recursion (2.9) is of infinite size, and therefore cannot be numerically solved. Therefore, though we consider Theorem 2.1 to be a primary result, it does not enable explicit evaluation of the CSP $\hat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n})$.

In order to establish a practicable formulation, it is necessary to extend this result to partially-specified haplotypes. Let $\mathbf{n} = (n_g)_{g\in\mathcal{G}}$ be a sample configuration of partially-specified haplotypes. Then conditional on $\mathbf{x}\in\Delta$, the ordered sampling probability is

$$q(\mathbf{n}|\mathbf{x}) = \prod_{g\in\mathcal{G}} y_g^{n_g}, \tag{2.11}$$

where $y_g = \sum_{h\in\mathcal{H}:h\curlywedge g} x_h$ is the total proportion of fully-specified haplotypes that subsume the partially-specified haplotype $g\in\mathcal{G}$. Defining the ordered sampling probability $\hat{q}(\cdot)$ and the CSP $\hat{\pi}_{\mathrm{PS}}(\cdot|\cdot)$ as in Section 2.1.1, it is possible to derive the following more general form of (2.9),

Theorem 2.2. *Let* $\mathbf{c} = (c_g)_{g\in\mathcal{G}}$ *with* $|\mathbf{c}| = c$, *and* $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ *with* $|\mathbf{n}| = n$. *Then the CSP* $\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n})$ *obtained using the approximate diffusion-generator technique described in Section 2.1.1 is given by the following recursive expression,*

$$\begin{aligned}\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}}\sum_{g\in\mathcal{G}} c_g\Bigg\{&\Bigg(\sum_{h\in\mathcal{H}:h\lambda g} n_h\Bigg)\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_g|\mathbf{n})\\ &+\sum_{g'\in\mathcal{G}:g'\lambda g}(c_{g'}-\delta_{g,g'})\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_g+\mathbf{e}_{\mathcal{C}(g,g')}|\mathbf{n})\\ &+\sum_{\ell\in L(g)}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,g[\ell]}\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_g+\mathbf{e}_{\mathcal{M}^a_\ell(g)}|\mathbf{n})\\ &+\sum_{b\in B(g)}\rho_b\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_h+\mathbf{e}_{\mathcal{R}^-_b(h)}+\mathbf{e}_{\mathcal{R}^+_b(h))}|\mathbf{n})\Bigg\},\end{aligned} \tag{2.12}$$

where $\mathcal{N} = \sum_{g\in\mathcal{G}} c_g\big(c+n-1+\sum_{\ell\in L(g)}\theta_\ell+\sum_{b\in B(g)}\rho_b\big)$.

Proof. Without loss of generality, write $\mathbf{c} = \mathbf{e}_{g_1}+\cdots+\mathbf{e}_{g_c}$ for $g_1,\ldots,g_c\in\mathcal{G}$. Let $f(\cdot)$ be an arbitrary real-valued function on fully-specified haplotype configurations, and define the linear map $\mathscr{S}_{\mathbf{c}}$,

$$\mathscr{S}_{\mathbf{c}}f = \sum_{\substack{h_1\in\mathcal{H}\\ h_1\lambda g_1}}\cdots\sum_{\substack{h_m\in\mathcal{H}\\ h_m\lambda g_m}} f(\mathbf{e}_{h_1}+\cdots+\mathbf{e}_{h_c}). \tag{2.13}$$

Then setting $f(\mathbf{c}') = \hat{\pi}_{\mathrm{PS}}(\mathbf{c}'|\mathbf{n})$

$$\begin{aligned}\mathscr{S}_{\mathbf{c}}f &= \sum_{\substack{h_1\in\mathcal{H}\\ h_1\lambda g_1}}\cdots\sum_{\substack{h_m\in\mathcal{H}\\ h_m\lambda g_m}}\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{h_1}+\cdots+\mathbf{e}_{h_c}|\mathbf{n})\\ &= \frac{1}{\hat{q}(\mathbf{n})}\hat{\mathbb{E}}\Bigg[\Bigg(\prod_{h\in\mathcal{H}}X_h^{n_h}\Bigg)\cdot\sum_{\substack{h_1\in\mathcal{H}\\ h_1\lambda g_1}}X_{h_1}\cdots\sum_{\substack{h_c\in\mathcal{H}\\ h_c\lambda g_c}}X_{h_c}\Bigg]\\ &= \frac{1}{\hat{q}(\mathbf{n})}\hat{\mathbb{E}}\Bigg[\Bigg(\prod_{h\in\mathcal{H}}X_h^{n_h}\Bigg)\cdot Y_{g_1}\cdots Y_{g_c}\Bigg] = \hat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n})\end{aligned} \tag{2.14}$$

Setting $f(\mathbf{c}') = \sum_{h\in\mathcal{H}} c'_h n_h\hat{\pi}_{\mathrm{PS}}(\mathbf{c}'-\mathbf{e}_h|\mathbf{n})$ and using a similar technique yields

$$\mathscr{S}_{\mathbf{c}}f = \sum_{g\in\mathcal{G}} c_g\sum_{\substack{h\in\mathcal{H}\\ h\lambda g}} n_h\cdot\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_g|\mathbf{n}). \tag{2.15}$$

And in the same way, setting $f(\mathbf{c}') = \sum_{h\in\mathcal{H}} c'_h(c_h-1)\hat{\pi}_{\mathrm{PS}}(\mathbf{c}'-\mathbf{e}_h|\mathbf{n})$ yields

$$\mathscr{S}_{\mathbf{c}}f = \sum_{g\in\mathcal{G}} c_g\sum_{\substack{g'\in\mathcal{G}\\ g\lambda g}}(c_{g'}-1)\cdot\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_g-\mathbf{e}_{g'}+\mathbf{e}_{\mathcal{C}(g,g')}|\mathbf{n}), \tag{2.16}$$

setting $f(\mathbf{c}') = \sum_{h\in\mathcal{H}} c'_h\sum_{\ell\in L}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,h[\ell]}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}'-\mathbf{e}_h+\mathbf{e}_{\mathcal{M}^a_\ell(h)}|\mathbf{n})$ yields

$$\mathscr{S}_{\mathbf{c}}f = \sum_{g\in\mathcal{G}} c_g\Bigg(\sum_{\ell\in L(g)}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,g[\ell]}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_g+\mathbf{e}_{\mathcal{M}^a_\ell(g)}|\mathbf{n})+\sum_{\ell\notin L(g)}\theta_\ell\hat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n})\Bigg), \tag{2.17}$$

setting $f(\mathbf{c}') = \sum_{b\in B}\rho_b \sum_{h'\in\mathcal{H}} \hat{\pi}_{\mathrm{PS}}(\mathbf{c}' - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h))}|\mathbf{n})$ yields

$$\mathscr{S}_{\mathbf{c}} f = \sum_{g\in\mathcal{G}} c_g \Bigg(\sum_{b\in B(g)} \rho_b \hat{\pi}_{\mathrm{PS}}(\mathbf{c} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b^-(h)} + \mathbf{e}_{\mathcal{R}_b^+(h))}|\mathbf{n}) + \sum_{b\notin B(g)} \rho_b \hat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n}) \Bigg). \tag{2.18}$$

Thus, regarding both the left and right hand sides of (2.9) as real-valued functions on full haplotype configuration and applying the linear map $\mathscr{S}_{\mathbf{c}}$ yields, in conjunction with the results just presented, the desired result (2.12). Observe that this proof explicitly depends on the result (2.9) of Theorem 2.1 □

Let $\mathbf{c} = (c_g)_{g\in\mathcal{G}}$ and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$, and denote the total number of specified loci in $\mathbf{c}'$ by $L(\mathbf{c}')$. Applying the recursion (2.12) to $\mathbf{c}$ and $\mathbf{n}$, each term on the right hand side is proportional to $\hat{\pi}_{\mathrm{PS}}(\mathbf{c}'|\mathbf{n})$ for some partially-specified configuration $\mathbf{c}'$, and $L(\mathbf{c}') \leq L(\mathbf{c})$. Consequently, repeated application of (2.12) yields a system of equations containing variables of the form $\hat{\pi}_{\mathrm{PS}}(\mathbf{c}'|\mathbf{n})$ for which $L(\mathbf{c}') \leq L(\mathbf{c})$. The resulting system is therefore finite, and can be numerically or algebraically solved for the desired value $\hat{\pi}(\mathbf{c}|\mathbf{n})$. The size of this linear system will be discussed in Section 3.1.

Finally, observe that Theorem 2.2 is applicable only when the configuration $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ is fully-specified. Obtaining a more general form of Theorem 2.2 for a partially-specified configuration $\mathbf{n}$ remains an important open problem.

Parent independent mutation

We shall also frequently be interested in *parent independent mutation* (PIM) models. Recall that a stochastic mutation matrix Φ exhibits PIM if there exists a vector $(\Phi_a)_{a\in A}$ with $\sum_{a\in A}\Phi_a = 1$, and $\Phi_{a',a} = \Phi_a$ for all $a' \in A$. Given a PIM model at locus $\ell \in L$, the term of the recursion (2.12) associated with mutation can be simplified,

$$\begin{aligned}
\sum_{a\in A_\ell} \Phi^{(\ell)}_{a,g[\ell]} \hat{\pi}_{\mathrm{PS}}(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}_\ell^a(g)}|\mathbf{n}) &= \Phi^{(\ell)}_{g[\ell]} \frac{1}{\hat{q}(\mathbf{n})} \hat{\mathbb{E}}\Bigg[\sum_{a\in A_\ell} q(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}_\ell^a(g)} + \mathbf{n}|\mathbf{X}) \Bigg] \\
&= \Phi^{(\ell)}_{g[\ell]} \frac{1}{\hat{q}(\mathbf{n})} \hat{\mathbb{E}}\Bigg[q(\mathbf{c} - \mathbf{e}_g + \mathbf{n}|\mathbf{X}) \sum_{a\in A_\ell} q(\mathbf{e}_{\mathcal{M}_\ell^a(g)}|\mathbf{X}) \Bigg] \\
&= \Phi^{(\ell)}_{g[\ell]} \hat{\pi}_{\mathrm{PS}}(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}_\ell(g)}|\mathbf{n}),
\end{aligned} \tag{2.19}$$

where the second and third equalities are by properties of the ordered multinomial distribution $q(\cdot|\mathbf{x})$ similar to (1.23). As a result, given a PIM model at every locus $\ell \in L$, identity (2.19) can be used to re-write (2.12) as follows,

$$\begin{aligned}
\hat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{g\in\mathcal{G}} c_g \Bigg\{ &\Bigg(\sum_{h\in\mathcal{H}: h \curlywedge g} n_h \Bigg) \hat{\pi}_{\mathrm{PS}}(\mathbf{c} - \mathbf{e}_g|\mathbf{n}) \\
&+ \sum_{g'\in\mathcal{G}: g' \curlywedge g} (c_{g'} - \delta_{g,g'}) \hat{\pi}_{\mathrm{PS}}(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{C}(g,g')}|\mathbf{n}) \\
&+ \sum_{\ell\in L(g)} \theta_\ell \Phi^{(\ell)}_{g[\ell]} \hat{\pi}_{\mathrm{PS}}(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}_\ell(g)}|\mathbf{n}) \\
&+ \sum_{b\in B(g)} \rho_b \hat{\pi}_{\mathrm{PS}}(\mathbf{c} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b^-(h)} + \mathbf{e}_{\mathcal{R}_b^+(h))}|\mathbf{n}) \Bigg\},
\end{aligned} \tag{2.20}$$

where $\mathcal{N} = \sum_{g\in\mathcal{G}} c_g\big(c + n - 1 + \sum_{\ell\in L(g)} \theta_\ell + \sum_{b\in B(g)} \rho_b\big)$. Thus, as for the sampling probability computations discussed in Section 1.2.1, assuming a PIM model at each locus confers both a mathematical and computational benefit. Nevertheless, it remains necessary to construct and numerically or algebraically solve a system of linear equations in order to evaluate the CSP $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$. In Section 2.2.2, we describe an additional approximation that obviates the need for solving a system. Finally, recall from Section 1.2.2 that any bi-allelic mutation model can be transformed into a PIM model, making (2.20) broadly applicable.

Specialization to one-locus case

In the one-locus case, the space of haplotypes can be represented by the (finite) space of alleles $\mathcal{H} = A$, and each haplotype by a single allele $a \in A$. Moreover, recombination is not applicable, and the single scaled mutation rate is represented by θ. Given a one-locus configurations $\mathbf{c} = (c_a)_{a\in A}$ and $\mathbf{n} = (n_a)_{a\in A}$, the recursion (2.9) for the CSP $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$ reduces to

$$\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{a\in A} c_a \Bigg\{ (c_a + n_a - 1)\hat{\pi}_{\text{PS}}(\mathbf{c} - \mathbf{e}_a|\mathbf{n}) + \theta \sum_{a'\in A} \Phi_{a',a}\hat{\pi}_{\text{PS}}(\mathbf{c} - \mathbf{e}_a + \mathbf{e}_{a'}|\mathbf{n}) \Bigg\} \tag{2.21}$$

where $\mathcal{N} = c(c + n - 1 + \theta)$. It is reassuring that given a haplotype $\alpha \in A$ and setting $\mathbf{c} = \mathbf{e}_\alpha$, we obtain the result obtained by De Iorio and Griffiths (2004a) using the same diffusion generator approximation. Further assuming a PIM model, the recursion (2.20) for $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$ reduces to

$$\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{a\in A} c_a \Bigg\{ (c_a + n_a - 1 + \theta\Phi_a)\hat{\pi}_{\text{PS}}(\mathbf{c} - \mathbf{e}_a|\mathbf{n}) \Bigg\} \tag{2.22}$$

where $\mathcal{N} = c(c + n - 1 + \theta)$. Observe that each term on the right hand side of (1.29) proportional to $\hat{\pi}_{\text{PS}}(\mathbf{c}'|\mathbf{n})$ has $|\mathbf{c}'| = c - 1 < c = |\mathbf{c}|$, where the inequality is strict. As in Section 1.2.2, the recursion is therefore proper, and the quantity $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$ can be directly evaluated using dynamic programming or memoization, without the need to construct and solve a coupled system of linear equations. Moreover, as for the one-locus PIM sampling probability, this recursion can be solved analytically,

Proposition 2.3. *Let $\mathbf{c} = (c_a)_{a\in A}$ and $\mathbf{n} = (n_a)_{a\in A}$ be one-locus configurations. Then the CSP $\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n})$ for a one-locus PIM model is given by*

$$\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n}) = \frac{1}{(\theta + n)_{(c)}} \prod_{a\in A} (\theta\Phi_a + n_a)_{(c_a)}, \tag{2.23}$$

where $x_{(i)} = (x)(x + 1)(x + 2)\cdots(x + i - 1)$ denotes a rising factorial.

Proof. Substitute (2.23) into (2.22). □

The analytic solution (2.23) is precisely the Conditional Wright Sampling Formula (1.62), and so for the one-locus PIM model, $\hat{\pi}_{\text{PS}} = \pi$. As we shall see, the correctness of the diffusion-generator technique is atypical. Nonetheless, this result provides some reassurance that our methodology, the diffusion generator approximation, is reasonable.

Specialization to two-locus case

For two loci, $L = \{1, 2\}$ and $B = \{(1, 2)\}$. Further assuming a PIM model, it is possible to derive the following closed-form solution of (2.20) when conditionally sampling a single haplotype,

Proposition 2.4. *Let $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ be fully-specified two-locus configuration, and $(a_1, a_2) \in \mathcal{H}$ a two-locus haplotype. Then the CSP $\hat{\pi}_{PS}(\mathbf{e}_{(a_1,a_2)}|\mathbf{n})$ for a two-locus PIM model is given by*

$$\begin{aligned}\hat{\pi}_{PS}(\mathbf{e}_{(a_1,a_2)}|\mathbf{n}) = \frac{1}{\mathcal{N}}\Bigg\{ & n_{(a_1,a_2)} + \theta_1\Phi^{(1)}_{a_1}\hat{\pi}_{PS}(\mathbf{e}_{a_1}|\mathbf{n}[1]) + \theta_2\Phi^{(2)}_{a_2}\hat{\pi}_{PS}(\mathbf{e}_{a_2}|\mathbf{n}[2]) \\ & + \rho_{(1,2)}\frac{2n+\theta_1+\theta_2}{2(n+1)+\theta_1+\theta_2}\hat{\pi}_{PS}(\mathbf{e}_{a_1}|\mathbf{n}[1])\hat{\pi}_{PS}(\mathbf{e}_{a_2}|\mathbf{n}[2])\Bigg\},\end{aligned} \tag{2.24}$$

where $\mathbf{n}[\ell]$ is the one-locus configuration induced by $\mathbf{n}$ at locus $\ell \in L$, and $\hat{\pi}_{PS}(\mathbf{e}_a|\mathbf{n}[\ell])$ is the one-locus CSP given in (2.22), and

$$\mathcal{N} = n + \theta_1 + \theta_2 + \rho_{(1,2)}\left(\frac{2n+\theta_1+\theta_2}{2(n+1)+\theta_1+\theta_2}\right). \tag{2.25}$$

Proof. Substitute (2.24) into (2.20). □

Though the one-locus CSPs comprising (2.24) are known to be exact, it is not the case that the CSP given by (2.24) is exact. It is interesting that, despite also using the diffusion generator approximation, Griffiths et al. (2008) obtain a distinct result, denoted $\hat{\pi}_{\text{GJS}}$

$$\begin{aligned}\hat{\pi}_{\text{GJS}}(\mathbf{e}_{(a_1,a_2)}|\mathbf{n}) = \frac{1}{\mathcal{N}'}\Bigg\{ & n_{(a_1,a_2)} + \theta_1\Phi^{(1)}_{a_1}\hat{\pi}(\mathbf{e}_{a_1}|\mathbf{n}[1]) + \theta_2\Phi^{(2)}_{a_2}\hat{\pi}(\mathbf{e}_{a_2}|\mathbf{n}[2]) \\ & + \frac{1}{2}\rho_{(1,2)}\left(\frac{n+\theta_1}{n+1+\theta_1} + \frac{n+\theta_2}{n+1+\theta_2}\right)\hat{\pi}(\mathbf{e}_{a_1}|\mathbf{n}[1])\hat{\pi}(\mathbf{e}_{a_2}|\mathbf{n}[2])\Bigg\},\end{aligned} \tag{2.26}$$

where $\mathcal{N}' = n + \theta_1 + \theta_2 + \frac{1}{2}\rho_{(1,2)}\left(\frac{n+\theta_1}{n+1+\theta_1} + \frac{n+\theta_2}{n+1+\theta_2}\right)$. To understand this disparity, observe that directly substituting $\mathbf{c} = \mathbf{e}_{(a_1,a_2)}$ into (2.20) immediately yields the term $\hat{\pi}(\mathbf{e}_{(a_1,\bullet)} + \mathbf{e}_{(\bullet,a_2)}|\mathbf{n})$, the probability of conditionally sampling *two* haplotypes. The generalized recursion (2.20) is directly applicable for $\mathbf{c} = \mathbf{e}_{(a_1,\bullet)} + \mathbf{e}_{(\bullet,a_2)}$. In contrast, Griffiths et al. (2008) derive and use a form of the recursion limited to conditionally sampling a single haplotype, and therefore must approximate this term using the symmetrized form:

$$\begin{aligned}\hat{\pi}(\mathbf{e}_{(a_1,\bullet)} + \mathbf{e}_{(\bullet,a_2)}|\mathbf{n}) &= \sum_{a_1'\in A_1}\sum_{a_2'\in A_2}\hat{\pi}(\mathbf{e}_{(a_1,a_2')} + \mathbf{e}_{(a_1',a_2)}|\mathbf{n}) \\ &\approx \sum_{a_1'\in A_1}\sum_{a_2'\in A_2}\frac{1}{2}\Big(\hat{\pi}(\mathbf{e}_{(a_1,a_2')}|\mathbf{n}+\mathbf{e}_{(a_1',a_2)})\hat{\pi}(\mathbf{e}_{(a_1',a_2)}|\mathbf{n}) \\ &\qquad + \hat{\pi}(\mathbf{e}_{(a_1',a_2)}|\mathbf{n}+\mathbf{e}_{(a_1,a_2')})\hat{\pi}(\mathbf{e}_{(a_1,a_2')}|\mathbf{n})\Big).\end{aligned} \tag{2.27}$$

Using this expression in place of the recursion (2.20) to evaluate $\hat{\pi}(\mathbf{e}_{(a_1,\bullet)} + \mathbf{e}_{(\bullet,a_2)}|\mathbf{n})$ yields the cited result (2.26). Note that the method employed by Griffiths et al. (2008) does not have an evident generalization to more than two loci, and requires the additional approximation (2.27).

Finally, we remark that it is possible, in principal, to obtain closed-form solutions for (2.12) for a non-PIM finite-alleles model, and even for more than two loci. There does not appear to be very much algebraic simplification possible in these cases, however, and the resulting solutions are tantamount to symbolically solving the system of equations generated using the recursion (2.12).

Limiting distributions

Returning to the more general setting, suppose that $\rho_b = \rho$, for all $b \in B$. We begin by investigating the CSD $\hat{\pi}_{\text{PS}}$ when $\rho = 0$. Setting $\mathbf{c} = \mathbf{e}_\eta$ for $\eta \in \mathcal{H}$, the recursion (2.12) yields the following simplified recursion for the single-haplotype CSP $\hat{\pi}_{\text{PS}}(\mathbf{e}_\eta|\mathbf{n})$,

$$\hat{\pi}_{\text{PS}}(\mathbf{e}_\eta|\mathbf{n}) = \frac{1}{n + \sum_{\ell \in L} \theta_\ell} \left(n_\eta + \sum_{\ell \in L} \theta_\ell \sum_{a \in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} \hat{\pi}_{\text{PS}}(\mathbf{e}_{\mathcal{M}^\eta_\ell(a)}|\mathbf{n}) \right). \tag{2.28}$$

Recall from Section 1.4.1 that Stephens and Donnelly's CSD $\hat{\pi}_{\text{SD}}$ is applicable in the absence of recombination (i.e. when $\rho = 0$). Despite the dissimilarity of Stephens and Donnelly's formulation (1.66) and the above recursion (2.28), the following proposition demonstrates that they are mathematically identical,

Proposition 2.5. *Let $\eta \in \mathcal{H}$ and $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$. Assuming $\rho_b = 0$ for all $b \in B$,*

$$\hat{\pi}_{PS}(\mathbf{e}_\eta|\mathbf{n}) = \hat{\pi}_{SD}(\mathbf{e}_\eta|\mathbf{n}). \tag{2.29}$$

Proof. We show that the expression (1.66) for $\hat{\pi}_{\text{SD}}(\mathbf{e}_\eta|\mathbf{n})$ solves the same recursion (2.28) as $\hat{\pi}_{\text{PS}}(\mathbf{e}_\eta|\mathbf{n})$. Removing the summand with $\mathbf{m} = \mathbf{0} \in \mathbb{N}^k$ in equation (1.66) yields:

$$\hat{\pi}_{\text{SD}}(\mathbf{e}_\eta|\mathbf{n}) = \sum_{h \in \mathcal{H}} \frac{n_h}{n} \left[F^{(n)}(h, \eta, \mathbf{0}) + \sum_{\mathbf{m} \in \mathbb{N}^k} \sum_{\ell \in L} \frac{m_\ell + 1}{m + 1} F^{(n)}(h, \eta, \mathbf{m} + \mathbf{e}_\ell) \right]. \tag{2.30}$$

Additionally, we have that $F^{(n)}(h, \eta, \mathbf{0}) = \delta_{h,\eta} \cdot n/(n + \Theta)$, and

$$F^{(n)}(h, \eta, \mathbf{m} + \mathbf{e}_\ell) = \frac{m + 1}{m_\ell + 1} \frac{\theta_\ell}{n + \Theta} \sum_{a \in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} \cdot F(h, \mathcal{M}^\eta_\ell(a), \mathbf{m}).$$

Substituting these identities into (2.30) yields the recursion

$$\hat{\pi}_{\text{SD}}(\mathbf{e}_\eta|\mathbf{n}) = \frac{1}{n + \Theta} \left(n_\eta + \sum_{\ell \in L} \theta_\ell \sum_{a \in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} \hat{\pi}_{\text{SD}}(\mathbf{e}_{\mathcal{M}^\eta_\ell(a)}|\mathbf{n}) \right),$$

which is identical to the recursion (2.28), proving the proposition. □

This result generalizes a similar result (De Iorio and Griffiths, 2004a) demonstrating the equivalence of $\hat{\pi}_{\text{SD}}$ to the diffusion-generator method in the one-locus case. Moreover, the equivalence provides a method for exact computation of the CSP $\hat{\pi}_{\text{SD}}(\mathbf{e}_\eta|\mathbf{n})$. Conversely, using the Gaussian quadrature method proposed by Stephens and Donnelly and described in Section 1.4.1, it provides a fast method for approximating $\hat{\pi}_{\text{PS}}$ in the absence of recombination. As will be demonstrated in Section 2.3.2, this is special case of a more general class of approximations related to the sequentially Markov coalescent. Finally, recall from Section 1.4.2 that in the absence of recombination, Fearnhead and Donnellys CSD $\hat{\pi}_{\text{FD}}$ coincides with $\hat{\pi}_{\text{SD}}$ by construction, and so $\hat{\pi}_{\text{PS}} = \hat{\pi}_{\text{FD}} = \hat{\pi}_{\text{SD}}$.

We next consider the limit $\rho \to \infty$. In this setting, we derive a result analogous to Proposition 1.4, showing that the CSP for $\hat{\pi}_{\text{PS}}$ can be decomposed into a product of one-locus CSPs.

Proposition 2.6. *Let* $\mathbf{c} = (c_h)_{h\in\mathcal{H}}$ *and* $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$*, and suppose* $\rho_b = \rho$ *for all* $b \in B$*. In the limit* $\rho \to \infty$*, the CSP* $\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n})$ *is given by*

$$\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n}) = \prod_{\ell\in L} \hat{\pi}_{PS}(\mathbf{c}[\ell]|\mathbf{n}[\ell]), \tag{2.31}$$

where $\mathbf{c}[\ell]$ *and* $\mathbf{n}[\ell]$ *are the one-locus configurations induced by* $\mathbf{c}$ *and* $\mathbf{n}$ *at locus* $\ell \in L$*, and* $\hat{\pi}_{PS}(\mathbf{c}[\ell]|\mathbf{n}[\ell])$ *is the one-locus CSP given in (2.21).*

Proof. Let $\mathbf{c}' = (c'_g)_{g\in\mathcal{G}}$, and define $B(\mathbf{c}') = \sum_{g\in\mathcal{G}} c'_g|B(g)|$ to be the total number of *valid* breakpoints in $\mathbf{c}'$. For $B(\mathbf{c}') > 0$, and in the limit that $\rho \to \infty$, the key recursion (2.12) produces

$$\hat{\pi}_{\mathrm{PS}}(\mathbf{c}'|\mathbf{n}) = \frac{1}{B(\mathbf{c}')} \sum_{g\in\mathcal{G}} n_g \sum_{b\in B(g)} \hat{\pi}_{\mathrm{PS}}(\mathbf{c}' - \mathbf{e}_g + \mathbf{e}_{\mathcal{R}_b^-(g)} + \mathbf{e}_{\mathcal{R}_b^+(g)}|\mathbf{n}),$$

and repeated application of this equation yields the identity

$$\hat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n}) = \hat{\pi}_{\mathrm{PS}}(\mathbf{c}^*|\mathbf{n}), \tag{2.32}$$

where $\mathbf{c}^*$ is derived from $\mathbf{c}$ by recombination at every possible breakpoint. More precisely, for $\ell \in L$ and $a \in A_\ell$, define $c_{\ell,a}$ to be the number of haplotypes in $\mathbf{c}$ with allele a at locus ℓ, and $u_\ell(a) \in \mathcal{G}$ to be the haplotype with allele a at locus ℓ and unspecified elsewhere. Then

$$\mathbf{c}^* = \sum_{\ell\in L} \mathbf{c}^*_\ell, \text{ where } \mathbf{c}^*_\ell = \sum_{a\in A_\ell} c_{\ell,a} \cdot \mathbf{e}_{u_\ell(a)}. \tag{2.33}$$

Since $B(\mathbf{c}^*) = 0$, (2.12) in conjunction with (2.32) yields

$$\begin{aligned}\sum_{\ell\in L} \big(c(n+\theta_\ell)\big)\hat{\pi}_{\mathrm{PS}}(\mathbf{c}^*|\mathbf{n}) = \sum_{\ell\in L}\sum_{a\in A_\ell} c_{\ell,a}\Big[\big(n_{\ell,a} + (c_{\ell,a}-1)\big)\hat{\pi}_{\mathrm{PS}}\Big(\big(\mathbf{c}^*_\ell - \mathbf{e}_{u_\ell(a)}\big) + \textstyle\sum_{\ell'\neq\ell}\mathbf{c}^*_{\ell'}|\mathbf{n}\Big) \\ + \theta_\ell \sum_{a'\in A_\ell} \Phi^{(\ell)}_{a',a}\hat{\pi}_{\mathrm{PS}}\Big(\big(\mathbf{c}^*_\ell - \mathbf{e}_{u_\ell(a)} + \mathbf{e}_{u_\ell(a')}\big) + \textstyle\sum_{\ell'\neq\ell}\mathbf{c}^*_{\ell'}|\mathbf{n}\Big)\Big].\end{aligned} \tag{2.34}$$

Observe that (2.34) is a sum of independent recursions, each for a particular locus $\ell \in L$. Consequently, it can be verified that the solution for the recursion is the product of solutions for each recursion summand,

$$\hat{\pi}_{\mathrm{PS}}(\mathbf{c}^*|\mathbf{n}) = \prod_{\ell\in L} \hat{\pi}_{\mathrm{PS}}(\mathbf{c}^*_\ell|\mathbf{n}) = \prod_{\ell\in L} \hat{\pi}_{\mathrm{PS}}(\mathbf{c}[\ell]|\mathbf{n}[\ell]).$$

In conjunction with (2.32), this produces the desired result □

Recall from Section 1.4.2 that Fearnhead and Donnelly's CSD $\hat{\pi}_{\mathrm{FD}}$ exhibits the same limiting decomposition, and the one-locus CSD $\hat{\pi}_{\mathrm{FD}}$ coincides with the one-locus CSDs $\hat{\pi}_{\mathrm{SD}}$ and $\hat{\pi}_{\mathrm{PS}}$. These facts imply that $\hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{FD}}$ in the limit $\rho \to \infty$. Moreover, by Proposition 1.11, the true CSD π can be identically decomposed; coupled with the fact that the one-locus CSDs $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{FD}}$ are exact for PIM models, we may conclude that for PIM models in the limit $\rho \to \infty$, $\hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{FD}} = \pi$.

2.1.3 Multiple-locus, multiple-deme

We now extend diffusion generator approximation to the setting of a structured population including migration. Recall from Section 1.2.3 that the Wright-Fisher diffusion for a finite-locus finite-alleles model including population structure over a finite set of demes, denoted $\mathcal{D}$, has state space

$$\Delta = \left\{ \mathbf{x} = (x_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}} \;\middle|\; x_{d,h} \geq 0 \text{ for all } d \in \mathcal{D}, h \in \mathcal{H} \text{ and } \sum_{h\in\mathcal{H}} x_{d,h} = 1 \text{ for all } d \in \mathcal{D} \right\}, \quad (2.35)$$

where $x_{d,h}$ is the proportion of haplotype $h \in \mathcal{H}$ within deme $d \in \mathcal{D}$. As before, the diffusion generator can be decomposed into a summation

$$\mathscr{L}f(\mathbf{x}) = \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} \mathcal{L}_{d,h} \frac{\partial}{\partial x_{d,h}} f(\mathbf{x}), \quad (2.36)$$

where the form (1.34) of $\mathcal{L}_{d,h}$ includes the infinitesimal mean (1.35) and covariance (1.36) associated with the Wright-Fisher diffusion. Let $\mathbf{n} = (n_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$ be a structured haplotype configuration, and recall that $q(\mathbf{n}|\mathbf{x})$ is the ordered multinomial probability (1.37) of sampling $\mathbf{n}$ conditioned on haplotype proportions $\mathbf{x} \in \Delta$. Analogous to the technique described in Section 2.1.1, we assume the existence of distribution and associated expectation $\hat{\mathbb{E}}$ such that (1.38) hold *component-wise*,

$$\hat{\mathbb{E}}\left[\mathcal{L}_{d,h} \frac{\partial}{\partial x_{d,h}} q(\mathbf{n}|\mathbf{X}) \right] = 0. \quad (2.37)$$

Note that this is a generalization of the diffusion-generator approximation (2.4) to a structured population, and that in the case that $|\mathcal{D}| = 1$, reduces to (2.4). Using the generalized diffusion generator approximation (2.37), we propose the following *re-weighted* version of (1.38),

$$\begin{aligned} &\hat{\mathbb{E}}\left[\sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} \frac{c_{d,h}}{c_{d,h} + n_{d,h}} \mathcal{L}_{d,h} \frac{\partial}{\partial x_{d,h}} q(\mathbf{c} + \mathbf{n}|\mathbf{X}) \right] \\ &\quad = \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} \frac{c_{d,h}}{c_{d,h} + n_{d,h}} \hat{\mathbb{E}}\left[\mathcal{L}_{d,h} \frac{\partial}{\partial x_{d,h}} q(\mathbf{c} + \mathbf{n}|\mathbf{X}) \right] = 0, \end{aligned} \quad (2.38)$$

As before, (2.38) produces a recursive equation for the sampling probability $\hat{q}(\mathbf{c}+\mathbf{n})$, and dividing by $\hat{q}(\mathbf{n})$ yields a recursive equation for the CSP $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$. We note that all of the mathematical results described in Section 2.1.1 for the diffusion generator technique continue to hold in this more general setting.

Given the generalized form of diffusion-generator technique described, we derive the following result for multiple loci with recombination and migration.

Theorem 2.7. *Let $\mathbf{c} = (c_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$ with $|\mathbf{c}| = c$, and $\mathbf{n} = (n_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$ with $|\mathbf{n}| = n$. Then the CSP $\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n})$ obtained using the approximate diffusion-generator technique is given by the following*

recursive expression,

$$\begin{aligned}\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} c_{d,h} \Bigg\{ & (c_{d,h} + n_{d,h} - 1)\kappa_d^{-1} \hat{\pi}_{PS}(\mathbf{c} - \mathbf{e}_{d,h}|\mathbf{n}) \\ & + \sum_{\ell\in L} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} \hat{\pi}_{PS}(\mathbf{c} - \mathbf{e}_{d,h} + \mathbf{e}_{d,\mathcal{M}^a_\ell(h)}|\mathbf{n}) \\ & + \sum_{b\in B} \rho_b \sum_{h'\in\mathcal{H}} \hat{\pi}_{PS}(\mathbf{c} - \mathbf{e}_{d,h} + \mathbf{e}_{d,\mathcal{R}_b(h,h')} + \mathbf{e}_{d,\mathcal{R}_b(h',h))}|\mathbf{n}) \\ & + \sum_{\substack{d'\in\mathcal{D} \\ d'\neq d}} \upsilon_{dd'} \hat{\pi}_{PS}(\mathbf{c} - \mathbf{e}_{d,h} + \mathbf{e}_{d',h}|\mathbf{n}) \Bigg\}\end{aligned} \tag{2.39}$$

where $\mathcal{N} = \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} c_{d,h}\big((c_d + n_d - 1)\kappa_d^{-1} + \sum_{\ell\in L} \theta_\ell + \sum_{b\in B} \rho_b + \upsilon_d\big)$.

Proof. Recalling the specifics of the diffusion generator (1.34), with infinitesimal mean and covariance given by (1.35) and (1.36), apply the key equation (2.38). In conjunction with the componentwise expectation (1.40), this yields

$$\begin{aligned}0 = \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} c_{d,h} \cdot \frac{1}{2} \Bigg\{ & (c_{d,h} + n_{d,h} - 1)\kappa_d^{-1} \hat{q}(\mathbf{c} + \mathbf{n} - \mathbf{e}_{d,h}) \\ & + \sum_{\ell\in L} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} \hat{q}(\mathbf{c} + \mathbf{n} - \mathbf{e}_{d,h} + \mathbf{e}_{d,\mathcal{M}^a_\ell(h)}) \\ & + \sum_{b\in B} \rho_b \sum_{h'\in\mathcal{H}} \hat{q}(\mathbf{c} + \mathbf{n} - \mathbf{e}_{d,h} + \mathbf{e}_{d,\mathcal{R}_b(h,h')} + \mathbf{e}_{d,\mathcal{R}_b(h',h))}) \\ & + \sum_{\substack{d'\in\mathcal{D} \\ d'\neq d}} \upsilon_{dd'} \hat{q}(\mathbf{c} + \mathbf{n} - \mathbf{e}_{d,h} + \mathbf{e}_{d',h}) \\ & - \Bigg((c_d + n_d - 1)\kappa_d^{-1} + \sum_{\ell\in L} \theta_\ell + \sum_{b\in B} \rho_b + \upsilon_d \Bigg) \hat{q}(\mathbf{c} + \mathbf{n}) \Bigg\}\end{aligned} \tag{2.40}$$

Dividing by $\hat{q}(\mathbf{n})$, and using the definition (2.5) of $\hat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n})$, the desired result (2.39) is obtained. □

Once again, though Theorem 2.7 is an important theoretical result, it does not enable explicit evaluation of $\hat{\pi}(\mathbf{c}|\mathbf{n})$ for a structured sample configurations $\mathbf{c}$ and $\mathbf{n}$. As in Section 2.1.2, it is necessary to extend the analysis to partially-specified haplotypes, which yields the following generalized recursion for a structured sample configuration on partially-specified haplotypes,

Theorem 2.8. *Let* $\mathbf{c} = (c_{d,g})_{d\in\mathcal{D},g\in\mathcal{G}}$ *with* $|\mathbf{c}| = c$, *and* $\mathbf{n} = (n_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$ *with* $|\mathbf{n}| = n$. *Then the CSP* $\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n})$ *obtained using the approximate diffusion-generator technique is given by the following*

recursive expression,

$$\begin{aligned}\hat{\pi}_{PS}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}}\sum_{d\in\mathcal{D}}\sum_{g\in\mathcal{G}} c_{d,g}\Bigg\{&\Bigg(\sum_{h\in\mathcal{H}:h\curlywedge g} n_{d,h}\Bigg)\kappa_d^{-1}\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_{d,g}|\mathbf{n})\\ &+\sum_{g'\in\mathcal{G}:g'\curlywedge g}(c_{d,g'}-\delta_{g,g'})\kappa_d^{-1}\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_{d,g}+\mathbf{e}_{d,\mathcal{C}(g,g')}|\mathbf{n})\\ &+\sum_{\ell\in L(g)}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,g[\ell]}\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_{d,g}+\mathbf{e}_{d,\mathcal{M}^a_\ell(g)}|\mathbf{n})\\ &+\sum_{b\in B(g)}\rho_b\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_{d,g}+\mathbf{e}_{d,\mathcal{R}_b^-(g)}+\mathbf{e}_{d,\mathcal{R}_b^+(g))}|\mathbf{n})\\ &+\sum_{\substack{d'\in\mathcal{D}\\ d'\neq d}}\upsilon_{dd'}\hat{\pi}_{PS}(\mathbf{c}-\mathbf{e}_{d,g}+\mathbf{e}_{d',g}|\mathbf{n})\Bigg\},\end{aligned} \tag{2.41}$$

where $\mathcal{N} = \sum_{d\in\mathcal{D}}\sum_{g\in\mathcal{H}} c_{d,g}\big((c_d+n_d-1)\kappa_d^{-1} + \sum_{\ell\in L(g)}\theta_\ell + \sum_{b\in B(g)}\rho_b + \upsilon_d\big)$.

Proof. The proof is entirely analogous to the proof of Theorem 2.2, and so we do not reproduce it here. □

As in Section 2.1.2, it is possible to show the reduced recursion (2.41) yields a finite set of coupled linear equations, which can be numerically solved for the CSP. It is reassuring that, for single deme $\mathcal{D} = \{1\}$ with $\kappa_1 = 1$, Theorems 2.7 and 2.8 reduce to the analogous Theorems 2.1 and 2.2, respectively, described in Section 2.1.2.

Product migration rates

Given a model of migration on a set $\mathcal{D}$ of demes, the migration rate model is said to be a *product migration rate* (PMR) model if there exist vectors $(\upsilon_d^{(\mathrm{s})})_{d\in\mathcal{D}}$ and $(\upsilon_d^{(\mathrm{d})})_{d\in\mathcal{D}}$, with $\sum_d \upsilon_d^{(\mathrm{d})} = 1$, such that $\upsilon_{dd'} = \upsilon_d^{(\mathrm{s})}\upsilon_{d'}^{(\mathrm{d})}$ for all $d, d' \in \mathcal{D}$ with $d \neq d'$. Note that any migration model on $|\mathcal{D}| = 2$ demes is a PMR model. Given a PMR model, the term of the recursion associated with migration can be re-factored,

$$\begin{aligned}\sum_{\substack{d'\in\mathcal{D}\\ d'\neq d}}\upsilon_{dd'}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_{d,h}+\mathbf{e}_{d',h}|\mathbf{n}) &= \upsilon_d^{(\mathrm{s})}\sum_{\substack{d'\in\mathcal{D}\\ d'\neq d}}\upsilon_{d'}^{(\mathrm{d})}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_{d,h}+\mathbf{e}_{d',h}|\mathbf{n})\\ &= \upsilon_d^{(\mathrm{s})}\sum_{d'\in\mathcal{D}}\upsilon_{d'}^{(\mathrm{d})}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_{d,h}+\mathbf{e}_{d',h}|\mathbf{n}) - \upsilon_d^{(\mathrm{s})}\upsilon_d^{(\mathrm{d})}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n}).\end{aligned} \tag{2.42}$$

As a result, given a PMR model, identity (2.42) can be used to re-write (2.41) as follows,

$$\begin{aligned}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}}\sum_{d\in\mathcal{D}}\sum_{g\in\mathcal{G}} c_{d,g}\Bigg\{&\Bigg(\sum_{h\in\mathcal{H}:h\curlywedge g} n_{d,h}\Bigg)\kappa_d^{-1}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_{d,g}|\mathbf{n})\\ &+\sum_{g'\in\mathcal{G}:g'\curlywedge g}(c_{d,g'}-\delta_{g,g'})\kappa_d^{-1}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_{d,g}+\mathbf{e}_{d,\mathcal{C}(g,g')}|\mathbf{n})\\ &+\sum_{\ell\in L(g)}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,g[\ell]}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_{d,g}+\mathbf{e}_{d,\mathcal{M}^a_\ell(g)}|\mathbf{n})\\ &+\sum_{b\in B(g)}\rho_b\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_{d,g}+\mathbf{e}_{d,\mathcal{R}^-_b(g)}+\mathbf{e}_{d,\mathcal{R}^+_b(g))}|\mathbf{n})\\ &+\upsilon^{(\mathrm{s})}_d\sum_{d'\in\mathcal{D}}\upsilon^{(\mathrm{d})}_{d'}\hat{\pi}_{\mathrm{PS}}(\mathbf{c}-\mathbf{e}_{d,g}+\mathbf{e}_{d',g}|\mathbf{n})\Bigg\},\end{aligned} \tag{2.43}$$

where $\mathcal{N} = \sum_{d\in\mathcal{D}}\sum_{g\in\mathcal{H}} c_{d,g}\big((c_d+n_d-1)\kappa_d^{-1} + \sum_{\ell\in L(g)}\theta_\ell + \sum_{b\in B(g)}\rho_b + \upsilon^{(\mathrm{s})}_d\big)$. In this general case, a PMR model does provide some regularity to the recursive expression, but unlike PIM, does not appear to confer an advantage in evaluating the recursion. In the one-locus case, however, we shall see that a PMR model does allow an analytic solution.

Specialization to one-locus case

In the one-locus case, the space of haplotypes can be represented by the (finite) space of alleles $\mathcal{H} = A$, and each haplotype by a single allele $a \in A$. Moreover, recombination is not applicable, and the single scaled mutation rate is represented by θ. Let $\alpha \in A$ and $d \in \mathcal{D}$; given the one-locus configurations $\mathbf{e}_{d,\alpha}$ and $\mathbf{n} = (n_{d,a})_{d\in\mathcal{D},a\in A}$, the recursion (2.39) for the CSP $\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{d,\alpha}|\mathbf{n})$ reduces to

$$\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{d,\alpha}|\mathbf{n}) = \frac{1}{\mathcal{N}}\Bigg\{n_{d,\alpha}\kappa_d^{-1} + \theta\sum_{a'\in A_\ell}\Phi_{a',\alpha}\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{d,a'}|\mathbf{n}) + \sum_{\substack{d'\in\mathcal{D}\\ d'\neq d}}\upsilon_{dd'}\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{d',\alpha}|\mathbf{n})\Bigg\} \tag{2.44}$$

where $\mathcal{N} = n_d\kappa_d^{-1} + \theta + \upsilon_d$. This is precisely the result derived by De Iorio and Griffiths (2004b) under the same diffusion-generator approximation. Further assuming a PMR and PIM model,

$$\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{d,\alpha}|\mathbf{n}) = \frac{1}{\mathcal{N}}\Bigg\{n_{d,\alpha}\kappa_d^{-1} + \theta\Phi_\alpha + \upsilon^{(\mathrm{s})}_d\sum_{d'\in\mathcal{D}}\upsilon^{(\mathrm{d})}_{d'}\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{d',\alpha}|\mathbf{n})\Bigg\} \tag{2.45}$$

where $\mathcal{N} = n_d\kappa_d^{-1} + \theta + \upsilon^{(\mathrm{s})}_d$. In contrast to the single-deme case, this recursion is not proper; explicit evaluation by repeated application of (2.45) still requires solving a system. However, De Iorio and Griffiths (2004b) showed that there does exist an analytic solution,

Proposition 2.9. *Let $\alpha \in A$ and $d \in \mathcal{D}$, and let $\mathbf{n} = (n_a)_{a\in A}$ be a one-locus configuration. Then the CSP $\hat{\pi}_{PS}(\mathbf{e}_{d,\alpha}|\mathbf{n})$ for a one-locus PMR and PIM model is given by*

$$\hat{\pi}_{PS}(\mathbf{e}_{d,\alpha}|\mathbf{n}) = \frac{1}{\mathcal{N}_d}\Bigg\{n_{d,\alpha}\kappa_d^{-1} + \theta\Phi_\alpha + \upsilon^{(s)}_d\Bigg(\frac{\sum_{d'\in\mathcal{D}}(n_{d',\alpha}\kappa_{d'}^{-1} + \theta\Phi_\alpha)\upsilon^{(d)}_{d'}\mathcal{N}_{d'}^{-1}}{1-\sum_{d'\in\mathcal{D}}\upsilon^{(s)}_{d'}\upsilon^{(d)}_{d'}\mathcal{N}_{d'}^{-1}}\Bigg)\Bigg\}, \tag{2.46}$$

where $\mathcal{N}_d = n_d\kappa_d^{-1} + \theta + \upsilon^{(s)}_d$.

Proof. Substitute (2.46) into (2.45). □

We note that, in contrast to the single-deme case, for which we were able to obtain the conditional Wright Sampling Formula, the result of Proposition 2.9 is not the true CSP. Moreover, it remains an open problem to determine a general analytic solution analogous to (2.46) for multiple conditionally sampled haplotypes.

2.2 A Genealogical Interpretation

In this section, we describe a coalescent-like genealogical process (Paul and Song, 2010) for conditional sampling. As we demonstrate, the genealogical process induces the same CSD $\hat{\pi}_{\text{PS}}$ as the diffusion-generator approximation employed in the previous section. The genealogical process thus furnishes an intuitive generative process for the CSD $\hat{\pi}_{\text{PS}}$, analogous to the way that the coalescent serves as a generative process for the sampling distribution of Chapter 1.

Perhaps more importantly, the genealogical process suggests several *genealogical* approximations that might be made to improve the efficiency of computing the CSP associated with $\hat{\pi}_{\text{PS}}$; these approximations culminate in the sequentially Markov CSD, to be discussed in the following section. We first describe the genealogical process, and then demonstrate how it can be applied to the finite-locus finite-alleles setting, both with and without population structure.

2.2.1 The trunk-conditional coalescent

Recall from Section 1.3 that a realization of the coalescent process is a genealogy $\mathcal{A}_{\hat{n}}$ comprising a series of genealogical events (e.g. coalescence, mutation, and recombination) relating an untyped collection of haplotypes $\hat{n}$. The procedure for sampling $\mathcal{A}_{\hat{n}}$ is naturally described by continuous-time Markov process starting in the present and continuing backward in time. The state of the process is a collection of labeled untyped haplotypes, or lineages, ancestral to the haplotypes of $\hat{n}$; genealogical events then correspond to transitions in the process, and the state is modified according to the event. When a single lineage, corresponding to the most recent common ancestor (MRCA) of $\hat{n}$ remains, the process terminates. Given an untyped genealogy $\mathcal{A}_{\hat{n}}$, a type for the MRCA can be sampled from the stationary distribution of the Wright-Fisher diffusion, and propagated forward in time on the genealogy $\mathcal{A}_{\hat{n}}$. This yields a typed configuration $\mathbf{n}$ and the corresponding typed genealogy $\mathcal{A}_{\mathbf{n}}$. The embedded discrete-time process, comprising the genealogical events and corresponding typed and untyped configuration, is depicted as a graphical model in Figure 1.4.

Suppose that we wish to sample a collection of additional haplotypes *conditional* on having already observed the configuration $\mathbf{n}$. For the moment, assume that the typed genealogy $\mathcal{A}_{\mathbf{n}}$ associated with configuration $\mathbf{n}$ is known. The coalescent process can be extended to sample a *conditional genealogy* $\mathcal{C}_{\hat{c}}$ relating the conditionally sampled haplotypes of the untyped configuration $\hat{c}$ to each other and to the haplotypes of the observed configuration $\mathbf{n}$. Specifically, the continuous-time Markov process for sampling $\mathcal{C}_{\hat{c}}$ comprises the same genealogical events, within $\mathcal{C}_{\hat{c}}$, as the unconditional process, and also coalescence events involving a lineage in $\mathcal{C}_{\hat{c}}$ and a lineage in $\mathcal{A}_{\mathbf{n}}$. We refer to these latter coalescence events as *absorption* events, since the lineage of $\mathcal{C}_{\hat{c}}$ has been absorbed into the known genealogy $\mathcal{A}_{\mathbf{n}}$. When all of the lineages of $\mathcal{C}_{\hat{c}}$ have been absorbed into the genealogy $\mathcal{A}_{\mathbf{n}}$, the process terminates. Because $\mathcal{A}_{\mathbf{n}}$ is a typed genealogy, the type of each absorbed lineage in $\mathcal{C}_{\hat{c}}$ is known, and can then be propagated forward in time, yielding a typed configuration $\mathbf{c}$ and the corresponding typed conditional genealogy $\mathcal{C}_{\mathbf{c}}$. See Figure 2.1(a) for an illustration.

There are several complications with this approach. Foremost is that the genealogy $\mathcal{A}_{\mathbf{n}}$ associated with a sample $\mathbf{n}$ is typically unknown, and the posterior distribution for $\mathcal{A}_{\mathbf{n}}$ is generally inaccessible. Moreover, in order to sample the typed conditional genealogy $\mathcal{C}_{\mathbf{c}}$, the types of each of the lineages within $\mathcal{A}_{\mathbf{n}}$ must be fully-specified, and the genealogy must therefore be *unreduced*. Similarly, because the conditional genealogy may extend beyond the MRCA of the genealogy $\mathcal{A}_{\mathbf{n}}$, the genealogy $\mathcal{A}_{\mathbf{n}}$ must extend beyond the MRCA, and infinitely into the past. Finally, unlike the genealogical processes described in Chapter 1, the Markov process for generating $\mathcal{C}_{\mathbf{c}}$ depends on $\mathcal{A}_{\mathbf{n}}$, and is therefore time-inhomogeneous; as a result, the general methods developed in Section 1.3.1 for producing a recursive expression for the CSP are not applicable.

To address these complications, we approximate the unknown genealogy by $\mathcal{A}_{\mathbf{n}} = \mathcal{A}_0(\mathbf{n})$, where $\mathcal{A}_0(\mathbf{n})$ is the non-random trunk genealogy, within which lineages do not mutate, recombine, migrate, or coalesce with one another, and instead form a trunk extending infinitely into the past. Note that although $\mathcal{A}_0(\mathbf{n})$ is an improper genealogy, as there is no MRCA, the process for sampling $\mathcal{C}_{\hat{c}}$ remains well-defined. See Figure 2.1(b) for an illustration of the approximate conditional sampling process. In conjunction, the conditional process is modified so that the rate of each non-absorption event within $\mathcal{C}_{\hat{c}}$ is doubled. This modification may be interpreted as mitigating the effect of the assumption $\mathcal{A}_{\mathbf{n}} = \mathcal{A}_0(\mathbf{n})$; for example, mutations do not occur in $\mathcal{A}_0(\mathbf{n})$, but occur at double the rate within $\mathcal{C}_{\hat{c}}$. We refer to this genealogical process as the *trunk-conditional coalescent*.

Because the trunk genealogy $\mathcal{A}_0(\mathbf{n})$ is time-homogeneous, and extends infinitely into the past, the trunk-conditional coalescent is also time-homogeneous. Moreover, every lineage of $\mathcal{A}_0(\mathbf{n})$ is fully-specified; the type of each absorbed lineage is therefore known, and can be propagated forward in time, yielding a sample $\mathbf{c}$ and the corresponding typed conditional genealogy $\mathcal{C}_{\mathbf{c}}$. The trunk-conditional coalescent thus induces a CSD, which we denote by $\hat{\pi}_{\text{GEN}}$.

2.2.2 Multiple-locus, single-deme

Let $\hat{c}$ be an untyped haplotype configuration and $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$ a typed haplotype configuration with associated trunk genealogy $\mathcal{A}_0(\mathbf{n})$. The trunk-conditional coalescent with recombination then has the following genealogical interpretation. For lineages within the conditional genealogy $\mathcal{C}_{\hat{c}}$,

Mutation: Each lineage undergoes mutation at locus $\ell \in L$ with rate θ_ℓ according to the mutation transition matrix $\Phi^{(\ell)}$.

Recombination: Each lineage undergoes recombination at breakpoint $b \in B$ with rate ρ_b.

Coalescence: Each pair of lineages coalesce with rate 2.

Absorption: Each lineage is absorbed into a lineage of $\mathcal{A}_0(\mathbf{n})$ with rate 1.

This process continues until all lineages of $\mathcal{C}_{\hat{c}}$ have been absorbed into the trunk $\mathcal{A}_0(\mathbf{n})$. A conditional genealogy realized by this process is illustrated in Figure 2.1(b). Because mutation events do not affect the topology of the conditional genealogy $\mathcal{C}_{\hat{c}}$, it is equivalent to sample a conditional genealogy using a two step procedure: first, sample the conditional genealogy topology using the procedure above without mutation events; second, realize the mutation events at each locus $\ell \in L$ as a Poisson process on the underlying topology with rate θ_ℓ.

Given an untyped conditional genealogy $\mathcal{C}_{\hat{c}}$ and the corresponding trunk genealogy $\mathcal{A}_0(\mathbf{n})$, the type of each absorbed lineage is known, and can be propagated forward in time, yielding a typed conditional configuration $\mathbf{c}$. Because the conditional sampling process is time-homogenous, the

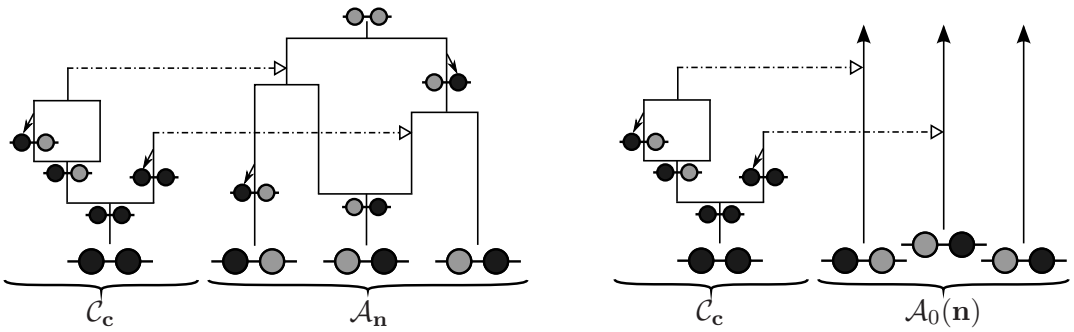


Figure 2.1. An illustration of the genealogical process for sampling a single haplotype conditional on configuration $\mathbf{n}$. (a) Idealized conditional sampling, for which the typed genealogy $\mathcal{A}_{\mathbf{n}}$ is known. An untyped conditional genealogy $\mathcal{C}_{\hat{c}}$ is sampled for untyped sample $\hat{c}$ using the unconditional genealogical procedure, and including absorption events, wherein an untyped lineage of $\mathcal{C}_{\hat{c}}$ is absorbed into a typed lineage of $\mathcal{A}_{\mathbf{n}}$, at rate 1. Absorption events are indicated by dot-dash arrows into $\mathcal{A}_{\mathbf{n}}$. It can be verified that the configuration $\mathbf{c}$ is obtained by tracing the type of each absorbed lineage forward in time. (b) Setting $\mathcal{A}_{\mathbf{n}} = \mathcal{A}_0(\mathbf{n})$, where $\mathcal{A}_0(\mathbf{n})$ is the improper trunk genealogy, within which lineages to not coalesce, mutate, or recombine. A similar procedure can be used to sample the conditional genealogy $\mathcal{C}_{\hat{c}}$, and to account for the absence of events within $\mathcal{A}_0(\mathbf{n})$, the rate of coalescence, mutation, and recombination within $\mathcal{C}_{\hat{c}}$ is doubled. It can be verified that the configuration $\mathbf{c}$ is obtained by tracing the type of each absorbed lineage forward in time.

time information within $\mathcal{C}_{\hat{c}}$ is not used to generate $\mathbf{c}$, and so it is only necessary to directly sample the genealogical events of $\mathcal{C}_{\hat{c}}$. Recalling the general construction of Section 1.3.1, starting with an untyped configuration $\hat{c}$, the possible genealogical events $\mathcal{E}(\hat{c})$ include coalescence, mutation, and recombination, and absorption. Let $e \in \mathcal{E}(\hat{c})$ be a genealogical event, and suppose $\mathbf{c}'$ is a typed configuration with associated untyped configuration $e(\hat{c})$,

Coalescence: Suppose $e \in \mathcal{E}(\hat{c})$ is a coalescence event. The untyped configuration $e(\hat{c})$ is derived from $\hat{c}$ by replacing the appropriate two labeled haplotypes with a single labeled haplotype, so that $|e(\hat{c})| = |\hat{c}| - 1$. Moreover $\mathcal{V}(\mathbf{c}', e)$ comprises a single typed configuration derived from $\mathbf{c}'$ by replacing the appropriate labeled haplotype $h \in \mathcal{H}$ with two identical labeled haplotypes,

$$\mathcal{V}(\mathbf{c}', e) = \{\mathbf{c}' - \mathbf{e}_h + \mathbf{e}_h + \mathbf{e}_h\} = \{\mathbf{c}' + \mathbf{e}_h\}. \tag{2.47}$$

Mutation: Suppose $e \in \mathcal{E}(\hat{c})$ is a mutation event at locus $\ell \in L$. The untyped configuration $e(\hat{c})$ is derived from $\hat{c}$ by replacing the appropriate labeled haplotype with a labeled haplotype, so that $|e(\hat{c})| = |\hat{c}|$. Moreover, $\mathcal{V}(\mathbf{c}', e)$ comprises a typed configuration for each allele $a \in A_\ell$, derived from $\mathbf{c}'$ by replacing the appropriate labeled haplotype $h \in \mathcal{H}$ with the labeled haplotype $\mathcal{M}_\ell^a(h)$,

$$\mathcal{V}(\mathbf{c}', e) = \{\mathbf{c}' - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}_\ell^a(h)} : a \in A_\ell\}, \tag{2.48}$$

and $p(\mathbf{c}' - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}_\ell^a(h)} | \mathbf{c}', e) = \Phi^{(\ell)}_{h[\ell],a}$.

Recombination: Suppose $e \in \mathcal{E}(\hat{c})$ is a recombination event at breakpoint $b \in B$. The untyped configuration $e(\hat{c})$ is derived from $\hat{c}$ by replacing the appropriate labeled haplotype with

two labeled haplotypes, so that $|e(\hat{c})| = |\hat{c}| + 1$. Moreover $\mathcal{V}(\mathbf{c}', e)$ comprises a single typed configuration derived from $\mathbf{c}'$ by replacing the appropriate two labeled haplotypes $h, h' \in \mathcal{H}$ with the labeled haplotype $\mathcal{R}_b(h, h')$,

$$\mathcal{V}(\mathbf{c}', e) = \{\mathbf{c}' - \mathbf{e}_h - \mathbf{e}_{h'} + \mathbf{e}_{\mathcal{R}_b(h,h')}\}. \tag{2.49}$$

Absorption: Suppose $e \in \mathcal{E}(\hat{c})$ is an absorption event. The untyped configuration $e(\hat{c})$ is derived from $\hat{c}$ by removing the appropriate labeled haplotype, so that $|e(\hat{c})| = |\hat{c}| - 1$. Moreover $\mathcal{V}(\mathbf{c}', e)$ comprises a single typed configuration derived from $\mathbf{c}'$ by adding the labeled haplotype $h \in \mathcal{H}$, where h is the type of the trunk lineage specified by the event.

$$\mathcal{V}(\mathbf{c}', e) = \{\mathbf{c}' + \mathbf{e}_h\}. \tag{2.50}$$

Finally, supposing that $|\hat{c}| = c$ and $|\mathbf{n}| = n$, the density $p(\cdot|\hat{c})$ is obtained considering the minimum of the exponential random variables associated with each event,

$$p(e|\hat{c}) = \begin{cases} 2/\mathcal{N}, & \text{for } e \text{ coalescence of two lineages,} \\ \theta_\ell/\mathcal{N}, & \text{for } e \text{ mutation of a lineage at locus } \ell \in L, \\ \rho_b/\mathcal{N}, & \text{for } e \text{ recombination of a lineage at breakpoint } b \in B, \\ 1/\mathcal{N}, & \text{for } e \text{ absorption of a lineage,} \end{cases} \tag{2.51}$$

where the normalizing constant $\mathcal{N} = c\big(c - 1 + n + \sum_{\ell \in L} \theta_\ell + \sum_{b \in B} \rho_b\big)$ is the total rate associated with all events. Having characterized the conditional sampling process associated with the trunk-conditional coalescent with recombination, the technique described in Section 1.3.1 yields the following result,

Theorem 2.10. *Let $\mathbf{c} = (c_h)_{h \in \mathcal{H}}$ with $|\mathbf{c}| = c$, and $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$ with $|\mathbf{n}| = n$. Then the CSP $\hat{\pi}_{GEN}(\mathbf{c}|\mathbf{n})$ obtained using the technique described in Section 1.3.1 in conjunction with the trunk-conditional coalescent with recombination is given by the following recursion*

$$\begin{aligned} \hat{\pi}_{GEN}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{h \in \mathcal{H}} c_h \Bigg\{ & (c_h + n_h - 1)\hat{\pi}_{GEN}(\mathbf{c} - \mathbf{e}_h|\mathbf{n}) \\ & + \sum_{\ell \in L} \theta_\ell \sum_{a \in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} \hat{\pi}_{GEN}(\mathbf{c} - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}^a_\ell(h)}|\mathbf{n}) \\ & + \sum_{b \in B} \rho_b \sum_{h' \in \mathcal{H}} \hat{\pi}_{GEN}(\mathbf{c} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h))}|\mathbf{n}) \Bigg\}, \end{aligned} \tag{2.52}$$

where $\mathcal{N} = c\big(c + n - 1 + \sum_{\ell \in L} \theta_\ell + \sum_{b \in B} \rho_b\big)$.

Proof. We use the technique described in Section 1.3.1. Define $\hat{c}$ to be the labeled untyped configuration associated with an arbitrary labeling of $\mathbf{c}$. Then we consider each event $e \in \mathcal{E}(\hat{n})$,

Coalescence: Suppose $e \in \mathcal{E}(\hat{c})$ is a coalescence event, specifying two labeled haplotypes $h, h' \in \mathcal{H}$ in $\mathbf{n}$. Since coalescence can only occur between identical haplotypes, $\{\mathbf{c}' : \mathbf{c} \in \mathcal{V}(\mathbf{c}', e)\} = \{\mathbf{c} - \mathbf{e}_h\}$ if $h = h'$ and is otherwise empty. As a result,

$$\Pr(V_0 = \mathbf{c}|U_0 = \hat{c}, E_1 = e) = \delta_{h,h'} \cdot q(\mathbf{c} - \mathbf{e}_h). \tag{2.53}$$

Mutation: Suppose $e \in \mathcal{E}(\hat{c})$ is a mutation event at locus $\ell \in L$, specifying the labeled haplotype $h \in \mathcal{H}$ in $\mathbf{c}$. Then $\{\mathbf{c}' : \mathbf{c} \in \mathcal{V}(\mathbf{c}', e)\} = \{\mathbf{c} - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}^a_\ell(h)} : a \in A_\ell\}$, and as a result,

$$\Pr(V_0 = \mathbf{c} | U_0 = \hat{c}, E_1 = e) = \sum_{a \in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} q(\mathbf{c} - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}^a_\ell(h)}). \tag{2.54}$$

Recombination: Suppose $e \in \mathcal{E}(\hat{c})$ is a recombination event at locus $b \in L$, specifying the labeled haplotype $h \in \mathcal{H}$ in $\mathbf{c}$. Then $\{\mathbf{c}' : \mathbf{c} \in \mathcal{V}(\mathbf{c}', e)\} = \{\mathbf{c} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h)} : h' \in \mathcal{H}\}$, and as result,

$$\Pr(V_0 = \mathbf{c} | U_0 = \hat{c}, E_1 = e) = \sum_{h' \in \mathcal{H}} q(\mathbf{c} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h)}). \tag{2.55}$$

Absorption: Suppose $e \in \mathcal{E}(\hat{c})$ is an absorption event at locus $b \in L$, specifying the labeled haplotype $h \in \mathcal{H}$ in $\mathbf{c}$ and $h' \in \mathcal{H}$ in $\mathbf{n}$. Since absorption can only occur between identical haplotypes $\{\mathbf{c}' : \mathbf{c} \in \mathcal{V}(\mathbf{c}', e)\} = \{\mathbf{c} - \mathbf{e}_h\}$ if $h = h'$ and is otherwise empty. As a result,

$$\Pr(V_0 = \mathbf{c} | U_0 = \hat{c}, E_1 = e) = \delta_{h,h'} \cdot q(\mathbf{c} - \mathbf{e}_h). \tag{2.56}$$

The latter expression in each case is obtained by using (1.45) in conjunction with the known expressions for $p(\mathbf{n}|\mathbf{n}', e)$. Recall that each genealogical event $e \in \mathcal{E}(\hat{n})$ specifies haplotypes according to a *labeling*, and without regard to type. Thus, using the general recursion (1.47), via (1.44), in conjunction with the known density (2.51), the desired recursion (2.52) is obtained. □

Observe that the recursive expression (2.52) for computing the CSP $\hat{\pi}_{\text{GEN}}(\mathbf{c}|\mathbf{n})$ is identical to the recursive expression (2.9) for computing $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$, and therefore $\hat{\pi}_{\text{GEN}} = \hat{\pi}_{\text{PS}}$. The trunk-conditional coalescent thus furnishes a genealogical interpretation for $\hat{\pi}_{\text{PS}}$. Moreover, recall that in the absence of recombination Proposition 2.5 states that $\hat{\pi}_{\text{PS}} = \hat{\pi}_{\text{SD}}$, and consequently the trunk-conditional coalescent also serves as an explicit genealogical interpretation for $\hat{\pi}_{\text{SD}}$. We note that it is remarkable that such different methodologies, reflecting distinct approximations to entirely complementary interpretations of the Wright-Fisher diffusion, can be used to deduce the same result. In Section 2.2.4, we investigate the relationship between these two approximations.

As in Section 1.3.2, we define a lineage within the conditional genealogy $\mathcal{C}_{\hat{c}}$ to be *non-ancestral* at locus $\ell \in L$ if, due to intervening recombination events in $\mathcal{C}_{\hat{c}}$, the locus has no descendant loci within the untyped configuration $\hat{c}$. Thus, in conditionally sampling a typed haplotype configuration, non-ancestral loci can be left unspecified, and it is unnecessary for $\mathcal{C}_{\hat{c}}$ to encode their genealogical history. By incorporating information about the non-ancestral loci into each lineage of the untyped conditional genealogy $\mathcal{C}_{\hat{c}}$, it is possible to specify a *reduced* trunk-conditional coalescent process. Applying the technique described in Section 1.3.1 to the reduced conditional genealogy then directly yields the more general form (2.12) of the recursion for the CSP $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$.

Limiting coalescence

In the one-locus setting discussed in Section 2.1.2, assuming a PIM model has the effect of making the CSP recursion proper, as there exists a partial order associated with the dependence of variables generated by repeated application of the recursive expression (2.22). As a result, the CSP can be evaluated using dynamic programming or memoization rather than numerically or algebraically

solving a system of equations. Moreover, we demonstrated in Proposition 2.3 that it is possible to obtain a closed-form solution to the recursion.

In contrast, assuming a PIM model in the more general multiple-locus case yields (2.20). Though we showed in Section 2.1.2 that repeated application of this recursion yields a set of coupled linear equations, the recursion is not proper. Examination of (2.20) in the context of the trunk-conditional coalescent reveals that it is the terms associated with coalescence events that make the recursion improper. For example, in a conditional genealogy it is possible for a lineage to undergo recombination, and for the two resulting lineages to then coalesce, thereby generating an identical configuration and precluding the existence of a partial order for the dependence of haplotype configurations on one another.

In order to prohibit this behavior, it is necessary to modify the trunk-conditional coalescent to disallow a certain class of coalescence events. We say that two untyped lineages are *overlap-coalesceable* if the sets of ancestral loci have a non-empty intersection; we then modify the trunk-conditional coalescent process so that coalescence events within $\mathcal{C}_{\hat{c}}$ are only allowed between pairs of lineages that are overlap-coalesceable. This modification alters the induced CSD, which we denote by $\hat{\pi}_{\text{LC}}$, where "LC" is an abbreviation for "limited coalescence".

Formally, given partially-specified haplotypes $g_1, g_2 \in \mathcal{G}$, analogous to the case for untyped haplotypes, we say that g_1 and g_2 are *overlap-coalesceable*, and write $g_1 \mathbin{\overline{\wedge}} g_2$ if $L(g_1) \cap L(g_2) \neq \emptyset$. Similarly, we say that g_1 and g_2 *overlap-compatible*, and write $g_1 \mathbin{\bar{\curlywedge}} g_2$, if $g_1 \curlywedge g_2$ and $g_1 \mathbin{\overline{\wedge}} g_2$. Let $\mathbf{c} = (c_g)_{g\in\mathcal{G}}$ with $|\mathbf{c}| = c$, and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ with $|\mathbf{n}| = n$; using the technique described in Section 1.3.1 and assuming a PIM model yields the following recursion

$$\begin{aligned}\hat{\pi}_{\text{LC}}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{g\in\mathcal{G}} c_g \Bigg\{ \Bigg(\sum_{h\in\mathcal{H}:h\curlywedge g} n_h \Bigg) \hat{\pi}_{\text{LC}}(\mathbf{c} - \mathbf{e}_g|\mathbf{n}) \\ + \sum_{g'\in\mathcal{G}:g' \mathbin{\bar{\curlywedge}} g} (c_{g'} - \delta_{g,g'}) \hat{\pi}_{\text{LC}}(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{C}(g,g')}|\mathbf{n}) \\ + \sum_{\ell\in L(g)} \theta_\ell \Phi^{(\ell)}_{g[\ell]} \hat{\pi}_{\text{LC}}(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}_\ell(g)}|\mathbf{n}) \\ + \sum_{b\in B(g)} \rho_b \hat{\pi}_{\text{LC}}(\mathbf{c} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b^-(g)} + \mathbf{e}_{\mathcal{R}_b^+(g))}|\mathbf{n}) \Bigg\},\end{aligned} \tag{2.57}$$

where $\mathcal{N} = \sum_{g\in\mathcal{G}} c_g \big(\sum_{g'\in\mathcal{G}:g \mathbin{\overline{\wedge}} g'} (c_{g'} - \delta_{g,g'}) + n + \sum_{\ell\in L(g)} \theta_\ell + \sum_{b\in B(g)} \rho_b \big)$. Note that the normalization constant requires lineages to be overlap-coalesceable, while the body of recursion requires lineages to be overlap-compatible.

The resulting recursion (2.57) for $\hat{\pi}_{\text{LC}}$ is proper. To see this, define $R(\mathbf{c}) = L(\mathbf{c}) + B(\mathbf{c})$, where $L(\mathbf{c}) = \sum_{g\in\mathcal{G}} c_g |L(g)|$ is the total number of specified loci and $B(\mathbf{c}) = \sum_{g\in\mathcal{G}} c_g |B(g)|$ is the total number of valid recombination breakpoints. Applying the recursion, each term on the right hand side is a scalar multiple of $\hat{\pi}(\mathbf{c}'|\mathbf{n})$ for some partially-specified configuration $\mathbf{c}'$. For the first term, representing an absorption event, $L(\mathbf{c}') < L(\mathbf{c})$ and $B(\mathbf{c}') \leq B(\mathbf{c})$. For the second term, representing a overlap-compatible coalescence, $L(\mathbf{c}') < L(\mathbf{c})$ and $B(\mathbf{c}') \leq B(\mathbf{c})$. For the third term, representing mutation $L(\mathbf{c}') < L(\mathbf{c})$ and $B(\mathbf{c}') \leq B(\mathbf{c})$, and for the fourth term, representing recombination, $L(\mathbf{c}') = L(\mathbf{c})$ and $B(\mathbf{c}') < B(\mathbf{c})$.

Therefore, in each case $R(\mathbf{c}') < R(\mathbf{c})$. As a result, there exists a partial-ordering on the dependence of the variables, and so the recursion is proper. The CSP $\hat{\pi}_{\text{LC}}(\mathbf{c}|\mathbf{n})$ can thus be computed using dynamic programming or memoization, and does not rely upon numerically or algebraically

solving of a system of coupled linear equations. Unlike the recursion (2.22) associated with the one-locus PIM model, no analytic solution for (2.57) is known.

Disallowing coalescence

As an extension to limiting coalescence to those lineages which are overlap-coalesceable in the conditional genealogical process, we next consider disallowing coalescence entirely, and denote the corresponding CSD by $\hat{\pi}_{\mathrm{NC}}$. Recall that in the more general case, a conditional genealogy $\mathcal{C}_{\hat{c}}$ comprises mutation, recombination, coalescence, and absorption events. Among these events, only coalescence has the effect of *coupling* two lineages backward in time; mutation, recombination, and absorption events have the non-coupling effect of modifying, splitting, and removing lineages, respectively. Intuitively then, in a genealogical process disallowing coalescence, separate lineages should behave independently. We formalize this intuition in the following proposition,

Proposition 2.11. *Let* $\mathbf{c} = \mathbf{e}_{g_1} + \cdots + \mathbf{e}_{g_c}$*, where* $g_1, \ldots, g_c \in \mathcal{G}$*, and* $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ *where* $|\mathbf{n}| = n$*. The CSP* $\hat{\pi}_{NC}(\mathbf{c}|\mathbf{n})$ *can be decomposed as follows,*

$$\hat{\pi}_{NC}(\mathbf{c}|\mathbf{n}) = \hat{\pi}_{NC}(\mathbf{e}_{g_1} + \cdots + \mathbf{e}_{g_c}|\mathbf{n}) = \prod_{i=1}^{c} \hat{\pi}_{NC}(\mathbf{e}_{g_i}|\mathbf{n}), \tag{2.58}$$

and for $\eta \in \mathcal{G}$*,*

$$\begin{aligned}\hat{\pi}_{NC}(\mathbf{e}_\eta|\mathbf{n}) = \frac{1}{\mathcal{N}}\Bigg\{ &\sum_{h\in\mathcal{H}:h\not\perp\eta} n_h + \sum_{\ell\in L(\eta)} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} \hat{\pi}_{NC}(\mathbf{e}_{\mathcal{M}^\eta_\ell(a)}|\mathbf{n}) \\ &+ \sum_{b\in B(\eta)} \rho_b \hat{\pi}_{NC}(\mathbf{e}_{\mathcal{R}^-_b(\eta)}) \hat{\pi}_{NC}(\mathbf{e}_{\mathcal{R}^+_b(\eta)}|\mathbf{n})\Bigg\},\end{aligned} \tag{2.59}$$

where $\mathcal{N} = n + \sum_{\ell\in L(\eta)} \theta_\ell + \sum_{b\in B(\eta)} \rho_b$*.*

Proof. Applying the technique described in Section 1.3.1 to the conditional genealogical process for which coalescence has been disallowed yields the following recursion for the CSP $\hat{\pi}_{\mathrm{NC}}(\mathbf{c}|\mathbf{n})$,

$$\begin{aligned}\hat{\pi}_{\mathrm{NC}}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{g\in\mathcal{G}} c_g \Bigg\{ &\Bigg(\sum_{h\in\mathcal{H}:h\not\perp g} n_h \Bigg) \hat{\pi}_{\mathrm{NC}}(\mathbf{c} - \mathbf{e}_g|\mathbf{n}) \\ &+ \sum_{\ell\in L(g)} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,g[\ell]} \hat{\pi}_{\mathrm{NC}}(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}^a_\ell(g)}|\mathbf{n}) \\ &+ \sum_{b\in B(g)} \rho_b \hat{\pi}_{\mathrm{NC}}(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{R}^-_b(g)} + \mathbf{e}_{\mathcal{R}^+_b(g))}|\mathbf{n})\Bigg\},\end{aligned} \tag{2.60}$$

where $\mathcal{N} = \sum_{g\in\mathcal{G}} c_g\big(n + \sum_{\ell\in L(g_i)}\theta_\ell + \sum_{b\in B(g_i)}\rho_b\big)$. And making use of the stated definition of $\mathbf{c}$,

$$\sum_{i=1}^{c}\Big(n + \sum_{\ell\in L(g_i)}\theta_\ell + \sum_{b\in B(g_i)}\rho_b\Big)\hat{\pi}_{\mathrm{NC}}(\mathbf{c}|\mathbf{n}) = \sum_{i=1}^{c}\Bigg\{\Bigg(\sum_{h\in\mathcal{H}:h\curlywedge g_i} n_h\Bigg)\hat{\pi}_{\mathrm{NC}}(\mathbf{c}-\mathbf{e}_{g_i}|\mathbf{n}) + \sum_{\ell\in L(g_i)}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,g_i[\ell]}\hat{\pi}_{\mathrm{NC}}(\mathbf{c}-\mathbf{e}_{g_i}+\mathbf{e}_{\mathcal{M}^a_\ell(g_i)}|\mathbf{n}) + \sum_{b\in B(g_i)}\rho_b\hat{\pi}_{\mathrm{NC}}(\mathbf{c}-\mathbf{e}_{g_i}+\mathbf{e}_{\mathcal{R}^-_b(g_i)}+\mathbf{e}_{\mathcal{R}^+_b(g_i))}|\mathbf{n})\Bigg\} \tag{2.61}$$

Observe that the latter expression is a sum of independent recursions, each for a particular haplotype $g_i \in \mathcal{G}$, and therefore has the solution given by (2.58). And by setting $\mathbf{c} = \mathbf{e}_\eta$ in (2.60), and applying (2.58) to the final term associated with recombination, (2.59) is obtained. □

Thus, disallowing coalescence confers a substantial computational simplification, as the state space of the recursion can be restricted to single-haplotype configurations. As in the case of limiting coalescence events to overlap-coalesceable lineages, further assuming a PIM model makes the recursion (2.59) proper, so that it can be computed using dynamic programming or memoization. The computational complexity of these methods will be discussed in Chapter 3. Observe that for a single conditionally sampled haplotype $\mathbf{c} = \mathbf{e}_\eta$ for $\eta \in \mathcal{H}$, events within the conditional genealogy cannot produce lineages that are overlap-coalesceable. As a result, $\hat{\pi}_{\mathrm{LC}} = \hat{\pi}_{\mathrm{NC}}$ for a single conditionally sampled haplotype.

Finally, we remark that disallowing and limiting coalescence is not as unreasonable as it first may seem; unlike the coalescent process, the conditional genealogical process does not rely on coalescence events to terminate (absorption events play the analogous role). Intuitively, the importance of modeling coalescence events within the conditional genealogy decreases with the ratio of the size of the conditional sample to the size of the observed sample; this is because absorption events become relatively more common than coalescence events as this ratio decreases. In many applications of the CSD, the size of conditional sample is indeed small, and the observed sample large.

2.2.3 Multiple-locus, multiple-deme

The approximate conditional sampling process can be further extended to population structure, including migration. Letting $\hat{c}$ be a structured untyped configuration and $\mathbf{n} = (n_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$ a structured typed configuration, lineages within the conditional genealogy $\mathcal{C}_{\hat{c}}$ exist in a particular deme $d \in \mathcal{D}$, and can migrate from deme to deme within $\mathcal{C}_{\hat{c}}$ prior to absorption. In contrast, lineages within the trunk genealogy do not migrate, and so the trunk genealogy $\mathcal{A}_0(\mathbf{n})$ can be decomposed into sub-trunk genealogies $\mathcal{A}_0(\mathbf{n}_d)$ for each deme $d \in \mathcal{D}$. Coalescence between lineages in $\mathcal{C}_{\hat{c}}$ can only occur if the lineages are in the same deme, and a lineage in deme d of $\mathcal{C}_{\hat{c}}$ can only be absorbed into a the sub-trunk genealogy $\mathcal{A}_0(\mathbf{n}_d)$. The trunk-conditional coalescent with recombination and migration then has the following genealogical interpretation. For lineages in deme $d \in \mathcal{D}$ of the conditional genealogy $\mathcal{C}_{\hat{c}}$,

Coalescence: Each pair of lineages coalesce with rate $2\kappa_d^{-1}$.

Mutation: Each lineage undergoes mutation at locus $\ell \in L$ with rate θ_ℓ according to the mutation transition matrix $\Phi^{(\ell)}$.

Recombination: Each lineage undergoes recombination at breakpoint $b \in B$ with rate ρ_b.

Migration: Each lineage migrates to deme d' with rate $\upsilon_{dd'}$

Absorption: Each lineage is absorbed into a lineage of $\mathcal{A}_0(\mathbf{n}_d)$ with rate κ_d^{-1}.

This genealogical process continues until all lineages of $\mathcal{C}_{\hat{c}}$ have been absorbed into the trunk $\mathcal{A}_0(\mathbf{n})$. A conditional genealogy realized by this process is illustrated in Figure 2.4(a).

The procedure for conditional sampling described in Section 2.2.2 can be generalized to this setting by incorporating a genealogical event for migration. Note that it is necessary to label haplotypes in both typed and untyped configurations by the deme in which they reside. Let $\hat{c}$ be such a structured untyped configuration, and $e \in \mathcal{E}(\hat{c})$ a genealogical event. Supposing that e is a coalescence, mutation, recombination, or absorption event, the description given in Section 2.2.2 suffices. Otherwise,

Migration: Suppose $e \in \mathcal{E}(\hat{c})$ is a migration event from $d \in \mathcal{D}$ to $d' \in \mathcal{D}$, backward in time. The untyped configuration $e(\hat{c})$ is derived from $\hat{c}$ by replacing the appropriate labeled untyped haplotype in deme d with a labeled untyped haplotype in deme d'. Given a typed configuration $\mathbf{c}'$ with associated untyped configuration $e(\hat{c})$, $\mathcal{V}(\mathbf{c}', e)$ comprises a single configuration derived from $\mathbf{c}'$ by replacing the appropriate labeled haplotype $h \in \mathcal{H}$ in deme d' with an identical labeled haplotype in deme d,

$$\mathcal{V}(\mathbf{n}', e) = \{\mathbf{c}' - \mathbf{e}_{d',h} + \mathbf{e}_{d,h}\}. \tag{2.62}$$

Finally, supposing that $|\hat{c}| = c$ and $|\hat{c}_d| = c_d$ and $|\hat{n}_d| = n_d$ for all $d \in \mathcal{D}$, the density $p(\cdot|\hat{c})$ is obtained considering the minimum of the exponential random variables associated with each event,

$$p(e|\hat{n}) = \begin{cases} 2\kappa_d^{-1}/\mathcal{N}, & \text{for } e \text{ coalescence of two lineages in deme } d \in \mathcal{D}, \\ \theta_\ell/\mathcal{N}, & \text{for } e \text{ mutation of a lineage at locus } \ell \in L, \\ \rho_b/\mathcal{N}, & \text{for } e \text{ recombination of a lineage at breakpoint } b \in B, \\ \upsilon_{dd'}/\mathcal{N}, & \text{for } e \text{ migration of a lineage from deme } d \text{ to deme } d', \\ \kappa_d^{-1}/\mathcal{N}, & \text{for } e \text{ absorption of a lineage,} \end{cases} \tag{2.63}$$

where the normalizing constant $\mathcal{N} = \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} c_{d,h}\big((c_d - 1 + n_d)\kappa_d^{-1} + \sum_{\ell\in L} \theta_\ell + \sum_{b\in B} \rho_b + \upsilon_d\big)$ is the total rate associated with all events. Having characterized the conditional sampling process associated with the trunk-conditional coalescent with recombination and migration, the technique described in Section 1.3.1 yields the following result,

Theorem 2.12. *Let* $\mathbf{c} = (c_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$ *with* $|\mathbf{n}| = n$, *and* $\mathbf{n} = (n_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$ *with* $|\mathbf{c}| = c$. *Then the CSP* $\hat{\pi}_{GEN}(\mathbf{c}|\mathbf{n})$ *obtained using the technique described in Section 1.3.1 in conjunction with the*

trunk-conditional coalescent with recombination and migration is given by the following recursion

$$\begin{aligned}\hat{\pi}_{GEN}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} c_{d,h} \Bigg\{ & (c_{d,h} + n_{d,h} - 1)\kappa_d^{-1} \hat{\pi}_{GEN}(\mathbf{c} - \mathbf{e}_{d,h}|\mathbf{n}) \\ & + \sum_{\ell\in L} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} \hat{\pi}_{GEN}(\mathbf{c} - \mathbf{e}_{d,h} + \mathbf{e}_{d,\mathcal{M}^a_\ell(h)}|\mathbf{n}) \\ & + \sum_{b\in B} \rho_b \sum_{h'\in\mathcal{H}} \hat{\pi}_{GEN}(\mathbf{c} - \mathbf{e}_{d,h} + \mathbf{e}_{d,\mathcal{R}_b(h,h')} + \mathbf{e}_{d,\mathcal{R}_b(h',h))}|\mathbf{n}) \\ & + \sum_{\substack{d'\in\mathcal{D} \\ d'\neq d}} \upsilon_{dd'} \hat{\pi}_{GEN}(\mathbf{c} - \mathbf{e}_{d,h} + \mathbf{e}_{d',h}|\mathbf{n}) \Bigg\}\end{aligned} \tag{2.64}$$

where $\mathcal{N} = \sum_{d\in\mathcal{D}} \sum_{h\in\mathcal{H}} c_{d,h}\big((c_d + n_d - 1)\kappa_d^{-1} + \sum_{\ell\in L} \theta_\ell + \sum_{b\in B} \rho_b + \upsilon_d\big)$*. This is identical to the recursion (2.39) obtained using the diffusion-generator approximation in Section 2.1.3.*

Proof. We use the technique described in Section 1.3.1 and exemplified in the proof of Theorem 2.10. Define $\hat{c}$ to be the labeled untyped configuration associated with an arbitrary labeling of $\mathbf{c}$, and let $e \in \mathcal{E}(\hat{c})$ be a genealogical event. If e is a coalescence, mutation, recombination, or absorption event, the description in the proof of Theorem 2.10 suffices; otherwise,

Migration: Suppose $e \in \mathcal{E}(\hat{c})$ is a migration event from deme $d \in \mathcal{D}$ to deme $d' \in \mathcal{D}$, backward in time, specifying the labeled haplotype $h \in \mathcal{H}$ in $\mathbf{n}$. Then $\{\mathbf{c}' : \mathbf{c} \in \mathcal{V}(\mathbf{c}', e)\} = \{\mathbf{c} - \mathbf{e}_{d,h} + \mathbf{e}_{d',h}\}$, and as result,

$$\Pr(V_0 = \mathbf{c}|U_0 = \hat{c}, E_1 = e) = q(\mathbf{c} - \mathbf{e}_{d,h} + \mathbf{e}_{d',h}). \tag{2.65}$$

Thus, using the general recursion (1.47), via (1.44), in conjunction with the known density (2.63), the desired recursion (2.64) is obtained. □

As in Section 2.2.2, the recursive expression (2.64) for computing the CSP $\hat{\pi}_{\text{GEN}}(\mathbf{c}|\mathbf{n})$ is identical to the recursive expression (2.39) for computing $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$, and therefore $\hat{\pi}_{\text{GEN}} = \hat{\pi}_{\text{PS}}$. The trunk-conditional coalescent thus furnishes a genealogical interpretation for $\hat{\pi}_{\text{PS}}$ in this more general setting of a structured population with migration. Moreover, by considering a reduced conditional genealogical process that accounts for non-ancestral loci, it is possible to directly obtain the more general form (2.41) of the recursion for the CSP $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$.

Finally, we remark that the conditional genealogical process can be modified, as in Section 2.2.2, so that coalescence events are limited or entirely disallowed. Though several key properties hold, including the haplotype decomposition (2.58) associated with $\hat{\pi}_{\text{NC}}$, the notable exception is that the CSP recursion associated with both $\hat{\pi}_{\text{LC}}$ and $\hat{\pi}_{\text{NC}}$, for a PIM model, is no longer proper; this is due to cycles in the dependence structure introduced by migration events. As a result, the CSPs associated with $\hat{\pi}_{\text{LC}}$ and $\hat{\pi}_{\text{NC}}$, including migration, must be evaluated by constructing and numerically or algebraically solving a system of coupled linear equations.

2.2.4 Interpretation

We have proposed a genealogical process, related to the coalescent with recombination, for conditional sampling. Importantly, the CSDs associated with the genealogical process are identical to

the CSDs derived, in Section 2.1, from the diffusion-generator approximation. We here investigate this connection, and provide some intuition for why the particular mathematical assumption used in the diffusion-generator approximation (2.4) is related to a genealogical process. We state the key result as a proposition, first suggested by Griffiths et al. (2008),

Proposition 2.13. *Let* $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ *be a sample configuration with associated untyped configuration* $\hat{n}$*. In the context of the coalescent process described in Section 1.3.1, denote by* Λ_h *the probabilistic event that the first genealogical event* E_1 *includes one of the* n_h *labeled haplotypes of type* h *in* $\mathbf{n}$*. Fixing* $h \in \mathcal{H}$*, the diffusion-generator approximation (2.4) applied to* h *is equivalent to assuming that the events* Λ_h *and* $V_0 = \mathbf{n}$ *are conditionally independent given the event* $U_0 = \hat{n}$*,*

$$\hat{\mathbb{E}}\left[\mathcal{L}_h \frac{\partial}{\partial x_h} q(\mathbf{n}|\mathbf{X})\right] = 0 \quad \Leftrightarrow \quad \widehat{\Pr}(\Lambda_h|V_0 = \mathbf{n}) = \Pr(\Lambda_h|U_0 = \hat{n}). \tag{2.66}$$

Proof. The following recursive expression is immediate from the technique of Section 1.3.1,

$$\begin{aligned}\Pr(V_0 = \mathbf{n}|U_0 = \hat{n}, \Lambda_h) = \frac{1}{\mathcal{N}}\Bigg\{ & (n_h - 1) q(\mathbf{n} - \mathbf{e}_h) \\ & + \sum_{\ell\in L} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,h[\ell]} q(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}^a_\ell(h)}) \\ & + \sum_{b\in B} \rho_b \sum_{h'\in\mathcal{H}} q(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h))}) \Bigg\},\end{aligned} \tag{2.67}$$

where $\mathcal{N} = n_h\big(n - 1 + \sum_{\ell\in L}\theta_\ell + \sum_{b\in B}\rho_b\big)$. Beginning with the diffusion-generator approximation (2.4), in conjunction with (1.18), we obtain

$$\hat{q}(\mathbf{n}) = \widehat{\Pr}(V_0 = \mathbf{n}|U_0 = \hat{n}) = \Pr(V_0 = \mathbf{n}|U_0 = \hat{n}, \Lambda_h), \tag{2.68}$$

where the first equality is by definition, and the second by mutual equality to (2.67). Applying Bayes Law, in conjunction with (2.68), then yields

$$\widehat{\Pr}(\Lambda_h|V_0 = \mathbf{n}) = \frac{\Pr(V_0 = \mathbf{n}|U_0 = \hat{n}, \Lambda_h)}{\widehat{\Pr}(V_0 = \mathbf{n}|U_0 = \hat{n})} \cdot \Pr(\Lambda_h|U_0 = \hat{n}) = \Pr(\Lambda_h|U_0 = \hat{n}). \tag{2.69}$$

Because each step can be reversed, the desired equivalence is established. □

This proposition furnishes a link between the diffusion-generator approximation and the genealogical interpretation, providing an intuitive statement about the distribution of genealogical event E_1, conditioned on the sample configuration $V_0 = \mathbf{n}$. The equivalent intermediate result (2.68) is also valuable, showing that the sampling probability $q(\mathbf{n})$ can be evaluated by choosing an arbitrary $h \in \mathcal{H}$, and conditioning on the genealogical event E_1 including one of the n_h haplotypes.

Now consider a haplotype configuration $\mathbf{c} + \mathbf{n}$. Applying the above logic, we may condition on the genealogical event E_1 including at least one haplotype within $\mathbf{c}$ to obtain a recursion for the sampling probability $q(\mathbf{c} + \mathbf{n})$ that does not include haplotypes in $\mathbf{n}$. This is the operation that is formalized by the approximate diffusion-generator technique, and in particular the weighted average provided in (2.6). In the genealogical context, this is precisely what the conditional sampling process accomplishes by only allowing events that include at least one haplotype within $\mathbf{c}$, and therefore do not disrupt the lineages associated with $\mathbf{n}$, thereby giving rise to the trunk genealogy .

2.3 Sequentially Markov CSD

Though we have not yet thoroughly discussed computation in the context of the approximate CSD $\hat{\pi}$, it should be intuitively clear that constructing and solving the system of linear equations associated with the recursion for $\hat{\pi}_{\text{PS}}$ is computationally challenging. We show in the next chapter that the computational complexity of these solutions is exponential in both the number of loci and the number of conditionally sampled individuals. Much as for the ordinary coalescent, the genealogical interpretation identified in the previous section suggests a key approximation, related to the sequentially Markov coalescent (SMC) introduced in Section 1.3.4. In this section, we describe the approximation, which yields the sequentially Markov CSD $\hat{\pi}_{\text{SMC}}$, then demonstrate how it can be applied to general finite-locus finite-alleles settings, both with (Steinrücken et al., 2012) and without population structure (Paul et al., 2011).

2.3.1 Marginal conditional genealogies

Recall from Section 1.3.4 that embedded within an ARG $\mathcal{A}_{\hat{n}}$, there is a sequence $(\mathcal{A}_{\hat{n}}[\ell])_{\ell \in L}$ of marginal genealogies, where each one-locus marginal genealogy $\mathcal{A}_{\hat{n}}[\ell]$ describes the genealogical relationship of the configuration $\hat{n}$ at locus $\ell \in L$. Wiuf and Hein (1999) demonstrated that it is possible to sample the marginal genealogies *sequentially*, starting from the left-most locus and proceeding to the right, in such a way that the joint distribution is identical to that obtained from the underlying coalescent model. Critically, the sequence of marginal genealogies produced by the method of Wiuf and Hein is not Markov. Intuitively, the non-Markov dependence corresponds to the potential for coalescence events that link marginal genealogies at non-adjacent loci.

McVean and Cardin (2005) showed that the non-Markov process of Wiuf and Hein can be well-approximated by a Markov process on the marginal genealogies, the SMC. The transition distribution for the approximate Markov process, as described in Section 1.3.4, is related to the two-locus distribution induced by the coalescent with recombination. Because each of the marginal genealogies in the Markov sequence is tree-like, the SMC confers substantial mathematical and computational simplicity relative to the coalescent with recombination, for which the entire graph-like ARG must be constructed. Moreover, it has been empirically demonstrated that the effect of this approximation is minimal (McVean and Cardin, 2005; Marjoram and Wall, 2006).

In much the same way, embedded within a conditional genealogy, there is a sequence $(\mathcal{C}_{\hat{c}}[\ell])_{\ell \in L}$ of *marginal conditional genealogies* (MCGs), where each one-locus MCG describes the genealogy, culminating with one or more absorptions into the trunk genealogy $\mathcal{A}_0(\mathbf{n})$, of the configuration $\hat{c}$ at locus $\ell \in L$. Though the sequence of MCGs is not Markov, we follow McVean and Cardin (2005) in constructing a Markov approximation, with transition distribution related to the two-locus transition distribution induced by the trunk-conditional coalescent with recombination. Using the transition distribution, the sequence of MCGs can be sampled directly. Recall from Section 2.2.2 that the mutation process does not affect the topology of the conditional genealogy; the sequence of MCGs $(\mathcal{C}_{\hat{c}}[\ell])_{\ell \in L}$ can therefore be produced *without* mutation events, which can be subsequently sampled at each locus independently.

We denote by $\hat{\pi}_{\text{SMC}}$ the CSD resulting from the sequentially Markov process. Critically, the conditional sampling process associated with $\hat{\pi}_{\text{SMC}}$ can be cast as a hidden Markov model (HMM). Suppose we wish to sample a typed configuration associated with $\hat{c}$, conditional on the observed configuration $\mathbf{n}$. At locus $\ell \in L$, the hidden state is the MCG at locus ℓ, without mutation events, which we denote by $s_\ell \in \mathcal{S}$, where $\mathcal{S}$ is the space of such MCGs. The corresponding observed state

is the one-locus typed configuration associated with $\hat{c}[\ell]$. It is necessary to specify the initial and transition distributions for the hidden state, and the emission distribution for the observed state,

Initial Distribution The random MCG at the first locus S_1 is drawn from the initial distribution, with density denoted $\zeta^{(\mathbf{n})}(\cdot)$, and is taken to be the one-locus marginal distribution on MCGs induced by the trunk-conditional coalescent.

Transition Distribution Given the MCG $S_{\ell-1} = s_{\ell-1}$, the random MCG S_ℓ is drawn from the transition distribution, with density denoted $\phi^{(\mathbf{n})}_{(\ell-1,\ell)}(\cdot|s_{\ell-1})$, and is taken to be the two-locus transition distribution induced by the trunk-conditional coalescent.

Emission Distribution Given the MCG $S_\ell = s_\ell$, the alleles at the ℓ-th locus of the conditionally sampled configuration $\hat{c}[\ell]$ are drawn from the emission distribution, with density denoted $\xi^{(\mathbf{n})}_\ell(\cdot|s_\ell)$, and is taken to be the distribution induced by the mutation process.

Now let $\mathbf{c}$ be a configuration, and consider computing the CSP $\hat{\pi}_{\text{SMC}}(\mathbf{c}|\mathbf{n})$. Recalling that there are k loci, the forward recursion (Cappé et al., 2005) for HMMs immediately yields

$$\hat{\pi}_{\text{SMC}}(\mathbf{c}|\mathbf{n}) = \int_{\mathcal{S}} f_k^{(\mathbf{c},\mathbf{n})}(s_k)\, ds_k, \tag{2.70}$$

where $f_\ell^{(\mathbf{c},\mathbf{n})}(\cdot)$ is defined (for $1 < \ell \leq k$) by

$$f_\ell^{(\mathbf{c},\mathbf{n})}(s_\ell) = \xi^{(\mathbf{n})}_\ell(\mathbf{c}[\ell]|s_\ell) \cdot \int_{\mathcal{S}} \phi^{(\mathbf{n})}_{(\ell-1,\ell)}(s_\ell|s_{\ell-1}) \cdot f_{\ell-1}^{(\mathbf{c},\mathbf{n})}(s_{\ell-1})\, ds_{\ell-1}, \tag{2.71}$$

with base case

$$f_1^{(\mathbf{c},\mathbf{n})}(s_1) = \xi^{(\mathbf{n})}_\ell(\mathbf{c}[1]|s_1) \cdot \zeta^{(\mathbf{n})}(s_1). \tag{2.72}$$

The MCG state space is continuous, however, and we generally cannot explicitly evaluate these integrals. In Chapter 3, we consider discretizing the state space, allowing $\hat{\pi}_{\text{SMC}}(\mathbf{c}|\mathbf{n})$ to be approximated efficiently and with high precision; the resulting CSP can be evaluated with computational complexity linear in the number of loci. In the remainder of this section, we apply the sequentially Markov approximation of $\hat{\pi}_{\text{SMC}}$, obtaining explicit characterizations of the hidden state space and expressions for the initial, transition, and emission densities.

Before proceeding, we remark that, for ease of notation, we generally suppress the dependence on $\mathbf{n}$ and $\mathbf{c}$ whenever possible. Thus, we typically write ζ, ϕ_b, and ξ_ℓ for the initial, transition, and emission densities, respectively. Similarly, for the forward density we write f_ℓ.

2.3.2 Single-deme, one-haplotype

Let $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ be a haplotype configuration, and consider sampling a single haplotype conditioned on $\mathbf{n}$ according to the trunk-conditional coalescent of Section 2.2.2. As discussed above, embedded within the conditional genealogy $\mathcal{C}_{\hat{c}}$ at locus $\ell \in L$ is an MCG $s_\ell \in \mathcal{S}$; disregarding mutation events, s_ℓ is entirely specified by two variables:

1. The *absorption time*, denoted $t_\ell \in \mathbb{R}_{\geq 0}$ (with $t_\ell = 0$ representing the present), at which the lineage associated with locus ℓ was absorbed into the trunk.

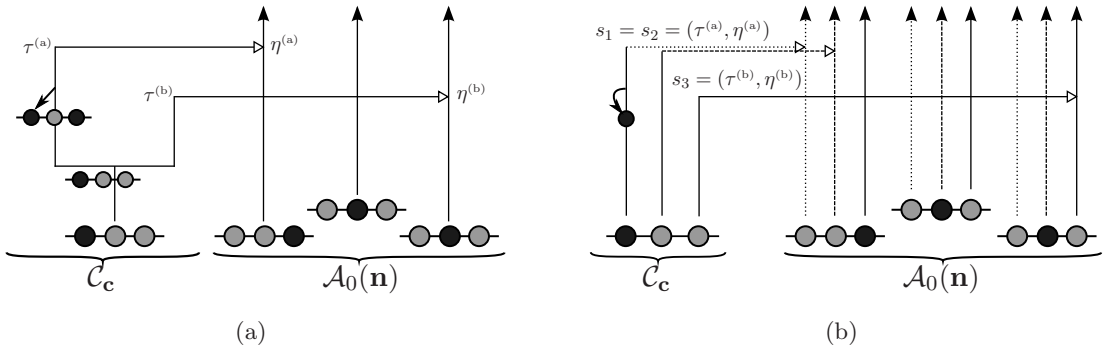


Figure 2.2. Illustration of the corresponding genealogical and sequential interpretations of a conditional genealogy $\mathcal{C}_{\mathbf{c}}$ with respect to the trunk genealogy $\mathcal{A}_0(\mathbf{n})$. (a) The genealogical interpretation. Absorption events, and the corresponding absorption time ($\tau^{(a)}$ and $\tau^{(b)}$) and haplotype ($\eta^{(a)}$ and $\eta^{(b)}$, respectively), are indicated by dot-dashed horizontal lines. (b) The corresponding sequential interpretation. The marginal conditional genealogies at the first, second, and third locus (s_1, s_2, and s_3) are indicated by dotted, dashed, and solid lines, respectively.

2. The *absorption haplotype*, denoted $h_\ell \in \mathcal{H}$, corresponding to the lineage in the trunk into which the lineage associated with locus ℓ was absorbed.

As a result the state space for the MCG can be represented $\mathcal{S} = \mathbb{R}_{\geq 0} \times \mathcal{H}$. We also write $S_\ell = (T_\ell, H_\ell)$ for the random MCG, and $s_\ell = (t_\ell, h_\ell) \in \mathcal{S}$ for the realized MCG at locus $\ell \in L$. See Figure 2.2 for an illustration.

We begin by considering the distribution of S_ℓ induced by the conditional genealogical process. Because the absorption process is Markov, T_ℓ and H_ℓ are independent, with T_ℓ distributed exponentially with parameter $n = |\mathbf{n}|$, and H_ℓ distributed uniformly over the n haplotypes of $\mathbf{n}$. Thus, the marginal density $\zeta(\cdot)$ is given by,

$$\zeta(s_\ell) = n_{h_\ell} e^{-nt_\ell}. \tag{2.73}$$

Conditioning on $S_{\ell-1} = s_{\ell-1} = (t_{\ell-1}, h_{\ell-1})$, the marginal conditional genealogy S_ℓ, for $\ell \geq 2$, is distributed according to a process analogous to that described in Section 1.3.4 for the SMC. Letting $b = (\ell - 1, \ell) \in B$,

1. Recombination breakpoints are realized as a Poisson process with rate ρ_b on the marginal conditional genealogy $s_{\ell-1}$.

2. Going backward in time, the lineage associated with locus $\ell - 1$ branching from each recombination breakpoint is removed, so that only the lineage more recent than the first (i.e. the most recent) breakpoint remains.

3. The lineage associated with locus ℓ branching from the first recombination breakpoint is subject to absorption into each lineage of $\mathcal{A}_0(\mathbf{n})$ at rate 1.

See Figure 2.3 for an illustration of this process. From this description, we deduce that there is no recombination between loci $\ell - 1$ and ℓ with probability $\exp(-\rho_b t_{\ell-1})$, and in this case the marginal conditional genealogy is unchanged, that is $S_\ell = s_{\ell-1}$. Otherwise, the time T_{r} of the first recombination breakpoint is distributed exponentially with parameter ρ_b, truncated at time $t_{\ell-1}$,

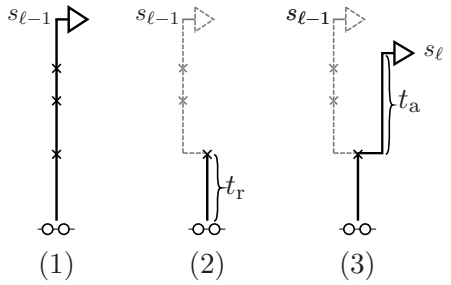


Figure 2.3. Illustration of the process for sampling the MCG S_ℓ conditioned on $S_{\ell-1} = s_{\ell-1}$. The MCG S_ℓ is sampled by (1) realizing recombination events, with breakpoint $b = (\ell-1, \ell) \in B$, as a Poisson process with rate ρ_b on the MCG $s_{\ell-1}$, (2) removing the lineage associated with locus $\ell - 1$ branching from each breakpoint, so that only the lineage more recent than the first breakpoint, at time $T_\mathrm{r} = t_\mathrm{r}$, remains, (3) creating a new lineage associated with locus ℓ at the first breakpoint, which is absorbed into a haplotype of $\mathbf{n}$ chosen uniformly at random, after time $T_\mathrm{a} = t_\mathrm{a}$ distributed exponentially with rate n. This produces the MCG $S_\ell = s_\ell$, with $t_\ell = t_\mathrm{r} + t_\mathrm{a}$.

and the additional time T_a until absorption is distributed exponentially with parameter n. Thus we have $S_\ell = (T_\mathrm{r} + T_\mathrm{a}, H_\ell)$, where H_ℓ is chosen uniformly at random from the sample $\mathbf{n}$. Taking a convolution of T_r and T_a, the transition density $\phi_b(\cdot|s_{\ell-1})$ is given by

$$\phi_b(s_\ell|s_{\ell-1}) = e^{-\rho_b t_{\ell-1}} \cdot \delta_{s_{\ell-1}, s_\ell} + \frac{n_{h_\ell}}{n} \int_0^{t_{\ell-1} \wedge t_\ell} \rho_b e^{-\rho_b t_\mathrm{r}} n e^{-n(t_\ell - t_\mathrm{r})} dt_\mathrm{r}, \tag{2.74}$$

where $t_{\ell-1} \wedge t_\ell$ denotes the minimum of $t_{\ell-1}$ and t_ℓ.

Finally, conditioning on $S_\ell = s_\ell$, recall that mutations are realized as a Poisson process (c.f. Stephens and Donnelly (2000)) with rate θ_ℓ. Thus, the number of mutations is Poisson-distributed, with mean $\theta_\ell t_\ell$, and each mutation proceeds according to $\Phi^{(\ell)}$. The emission density on alleles $\xi_\ell(\cdot|s_\ell)$ is therefore given by

$$\xi_\ell(a|s_\ell) = e^{-\theta_\ell t_\ell} \sum_{m=0}^{\infty} \frac{(\theta_\ell t_\ell)^m}{m!} \left[\left(\Phi^{(\ell)} \right)^m \right]_{h_\ell[\ell], a}. \tag{2.75}$$

Using these densities within the forward recursion given above provides, in principle, a method for computing $\hat{\pi}_{\mathrm{SMC}}$. In practice, there is no known analytic solution for the integrals, and so it is necessary to numerically approximate them. This technique is discussed in detail in Chapter 3. We next document several important properties satisfied by the densities and by the CSD $\hat{\pi}_{\mathrm{SMC}}$.

Equivalence to $\hat{\pi}_{\mathbf{LC}}$

Recall from Section 2.3.1 that the sequentially Markov assumption is violated by coalescence events, which introduce non-Markov dependence between the marginal genealogies at non-adjacent loci. With this as intuition, it is reasonable to conjecture (McVean and Cardin, 2005) that a genealogical process *disallowing* a certain class of coalescence events may be equivalent to the sequentially Markov coalescent. The disallowed coalescences are those between two lineages that do not share ancestral loci; formally, these are precisely the coalescence events between lineages that are not overlap-coalesceable, as described in Section 2.2.2. To the author's knowledge, no proof of this conjecture exists.

In the conditional sampling setting, the same intuition makes it reasonable to conjecture that $\hat{\pi}_{\text{SMC}}$ is equivalent to $\hat{\pi}_{\text{LC}}$ (see Section 2.2.2), for which the same class of coalescence events are disallowed within the conditional genealogy. In this most general case, the conjecture is again unproved. Recall that when conditionally sampling a single haplotype, $\hat{\pi}_{\text{LC}}$ is identical $\hat{\pi}_{\text{NC}}$, for which coalescence entirely disallowed within the conditional genealogy; in this special case, we can algebraically demonstrate that the conjecture is true,

Theorem 2.14. *Let $\eta \in \mathcal{H}$ and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$. Then the CSD $\hat{\pi}_{SMC}$ is equivalent to the CSD induced by the trunk-conditional coalescent with coalescence events disallowed,*

$$\hat{\pi}_{SMC}(\mathbf{e}_\eta|\mathbf{n}) = \hat{\pi}_{LC}(\mathbf{e}_h|\mathbf{n}) = \hat{\pi}_{NC}(\mathbf{e}_\eta|\mathbf{n}). \tag{2.76}$$

Sketch of Proof. The key idea of the proof is to introduce a *genealogical* recursion for the joint density function $g_\ell^{(\eta,\mathbf{n})}(s_\ell)$ associated with sampling the first ℓ loci of haplotype η (under $\hat{\pi}_{\text{NC}}$) and the marginal genealogy s_ℓ at the final locus. This recursion can be constructed following the lines of Griffiths and Tavaré (1994) to explicitly incorporate coalescent time.

By partitioning with respect to the most recent event occurring at the last locus k, it is possible to inductively show that $f_\ell^{(\eta,\mathbf{n})}(s_\ell) = g_\ell^{(\eta,\mathbf{n})}(s_\ell)$. Moreover, the identity $\int g_k^{(\eta,\mathbf{n})}(s_\ell)ds_\ell = \hat{\pi}_{\text{NC}}(\mathbf{e}_\eta|\mathbf{n})$ can be verified, and thus we conclude that

$$\hat{\pi}_{\text{LC}}(\mathbf{e}_\eta|\mathbf{n}) = \hat{\pi}_{\text{NC}}(\mathbf{e}_\eta|\mathbf{n}) = \int g_k^{(\eta,\mathbf{n})}(s_k)ds_k = \int f_k^{(\eta,\mathbf{n})}(s_k)ds_k = \hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n}). \qquad \square$$

A full version of this proof is presented in Appendix B.1. We believe that this method of proof could, in principle, be extended to the more general case of conditionally sampling two or more haplotypes. Without further abstraction, however, it seems the requisite algebra would be overwhelming. We thus leave proof of this general conjecture as an open problem. Finally, we note that the demonstrated equivalence provides a method for *exact* computation of $\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n})$, providing a useful baseline to compare numerical approximations to.

Mathematical properties

We now demonstrate several other intuitively appealing properties of $\hat{\pi}_{\text{SMC}}$ for a single conditionally sampled haplotype. For example, the marginal and transition distributions described above satisfy the *detailed-balance condition*. Letting $b = (\ell-1, \ell) \in B$, and $s_{\ell-1}, s_\ell \in \mathcal{S}$ be arbitrary MCGs,

$$\begin{aligned}
&\phi_b(s_\ell|s_{\ell-1})\zeta(s_{\ell-1}) \\
&\quad= \left(e^{-\rho_b t_{\ell-1}} \cdot \delta_{s_{\ell-1},s_\ell} + \frac{n_{h_\ell}}{n}\int_0^{t_{\ell-1}\wedge t_\ell} \rho_b e^{-\rho_b t} n e^{-n(t_\ell - t)} dt\right)\left(n_{h_{\ell-1}} e^{-nt_{\ell-1}}\right) \\
&\quad= \left(e^{-\rho_b t_\ell} \cdot \delta_{s_\ell,s_{\ell-1}} + \frac{n_{h_{\ell-1}}}{n}\int_0^{t_\ell\wedge t_{\ell-1}} \rho_b e^{-\rho_b t} n e^{-n(t_{\ell-1} - t)} dt\right)\left(n_{h_\ell} e^{-nt_\ell}\right) \\
&\quad= \phi_b(s_{\ell-1}|s_\ell)\zeta(s_\ell)
\end{aligned} \tag{2.77}$$

The detailed-balance condition shows that Markov process is reversible, and that the distribution ζ is stationary under the given transition dynamics; that is, the invariance condition,

$$\int_{\mathcal{S}} \phi_b(s_\ell|s_{\ell-1})\zeta(s_{\ell-1})\, ds_{\ell-1} = \zeta(s_\ell) \cdot \int_{\mathcal{S}} \phi_b(s_{\ell-1}|s_\ell)\, ds_{\ell-1} = \zeta(s_\ell) \tag{2.78}$$

is satisfied. Thus, for $\hat{\pi}_{\text{SMC}}$, the random MCG S_ℓ is marginally distributed according to ζ for all loci $\ell \in L$, and in particular the marginal distribution of the absorption time T_ℓ is exponential with rate n. This parallels the fact that the marginal genealogies under the SMC (and the coalescent with recombination) are distributed according to Kingman's coalescent. Moreover, this property ensures that the CSP computation will yield the same result regardless of whether we proceed from left to right, as in (2.70), or from right to left.

Similarly, the transition density exhibits a consistency property, which we refer to as the *locus-skipping property*. Intuitively, this property states that transitioning directly from locus $\ell - 1$ to $\ell + 1$ can be accomplished by using the transition density parametrized with the sum of the recombination rates. Formally, letting $s_{\ell-1}$ and $s_{\ell+1}$ be arbitrary MCGs, it can be verified that

$$\int_{\mathcal{S}} \phi_{(\ell-1,\ell)}(s_\ell|s_{\ell-1})\phi_{(\ell,\ell+1)}(s_{\ell+1}|s_\ell)ds_\ell = \phi_{(\ell-1,\ell+1)}(s_{\ell+1}|s_{\ell-1}), \tag{2.79}$$

where $\phi_{(\ell-1,\ell+1)}$ is the transition density parameterized by $\rho_{(\ell-1,\ell)} + \rho_{(\ell,\ell+1)}$. As will be more thoroughly described in Chapter 3, this property is computationally useful, as it enables loci $\ell \in L$ for which $\eta[\ell]$ is unobserved to be skipped in computing the CSP $\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n})$.

Finally, it can be verified that the expectation of T_ℓ conditioned on $T_{\ell-1} = t_{\ell-1}$ is

$$\begin{aligned}\mathbb{E}[T_\ell|T_{\ell-1} = t_{\ell-1}] &= \int_0^\infty t_\ell \left(e^{-\rho_b t_{\ell-1}} \cdot \delta_{t_{\ell-1},t_\ell} + \int_0^{t_{\ell-1}\wedge t_\ell} \rho_b e^{-\rho_b t} n e^{-n(t_\ell - t)} dt \right) dt_\ell \\ &= \left(\frac{1}{\rho_b} + \frac{1}{n}\right)\left(1 - e^{-\rho_b t_{\ell-1}}\right),\end{aligned} \tag{2.80}$$

where $b = (\ell - 1, \ell) \in B$. Asymptotically, this expression provides several intuitive results.

- As $\rho_b \to \infty$, $\mathbb{E}[T_\ell|T_{\ell-1} = t_{\ell-1}] \to 1/n$. In this limit, recombination occurs immediately, and so $1/n$ is the expectation of the additional absorption time T_{a}.
- As $\rho_b \to 0$, $\mathbb{E}[T_\ell|T_{\ell-1} = t_{\ell-1}] \to t_{\ell-1}$. In this limit there is no recombination, and the absorption time does not change.
- As $t_{\ell-1} \to \infty$, $\mathbb{E}[T_\ell|T_{\ell-1} = t_{\ell-1}] \to 1/\rho_b + 1/n$. In this limit, recombination must occur, and the exponentially distributed time is not truncated, so the expectation is the sum of the expectations of two exponentials.
- As $t_{\ell-1} \to 0$, $\mathbb{E}[T_\ell|T_{\ell-1} = t_{\ell-1}] \to 0$. In this limit, no recombination can occur, and so the absorption time is unchanged.

Limiting distributions

We next set $\rho_b = \rho$, for all $b \in B$, and explore the properties of $\hat{\pi}_{\text{SMC}}$ when $\rho = 0$ and in the limit $\rho \to \infty$. Setting $\rho = 0$, the transition distribution reduces to $\phi_b(s_\ell|s_{\ell-1}) = \delta_{s_{\ell-1},s_\ell}$ for all $b = (\ell - 1, \ell) \in B$, and therefore $f_\ell(s_\ell) = \xi_\ell(\mathbf{c}[\ell]|s_\ell) f_{\ell-1}(s_\ell)$ and

$$\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n}) = \int_{\mathcal{S}} \zeta(s) \prod_{\ell \in L} \xi_\ell(\eta[\ell]|s) ds. \tag{2.81}$$

From a genealogical perspective, when $\rho = 0$, the only possible events are absorption and mutation; equivalently, it is possible to initially disregard mutation, and conditioned on the time of the

absorption event, sample mutation events (independently) for each locus. Thus, in the limit that $\rho = 0$, $\hat{\pi}_{\mathrm{SMC}}$ is equivalent to $\hat{\pi}_{\mathrm{PS}}$ and, by extension, $\hat{\pi}_{\mathrm{SD}}$ and $\hat{\pi}_{\mathrm{FD}}$. Note that the form (2.81) for $\hat{\pi}_{\mathrm{SMC}}$ when $\rho = 0$ is equivalent to the alternative form (1.68) for $\hat{\pi}_{\mathrm{SD}}$.

Similarly, in the limit $\rho \to \infty$, the transition distribution reduces to $\phi_b(s_\ell|s_{\ell-1}) = \zeta(s_\ell)$ for all $b = (\ell - 1, \ell) \in B$, and therefore $f_\ell(s_\ell) = \xi_\ell(\mathbf{c}[\ell]|s_\ell)\zeta(s_\ell) \int_{\mathcal{S}} f_{\ell-1}(s_{\ell-1})ds_{\ell-1}$ and

$$\hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta|\mathbf{n}) = \prod_{\ell \in L} \left[\int_{\mathcal{S}} \zeta(s_\ell)\xi_\ell(\eta[\ell]|s_\ell)ds_\ell \right] = \prod_{\ell \in L} \hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta[\ell]|\mathbf{n}[\ell]), \tag{2.82}$$

where $\hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta[\ell]|\mathbf{n}[\ell])$ is the one-locus CSP. Recalling Proposition 2.6, $\hat{\pi}_{\mathrm{PS}}$ enjoys the same limiting decomposition, and because $\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{PS}}$ in the one-locus case, we have that in the limit $\rho \to \infty$, $\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{FD}}$. Moreover, for a PIM model of mutation the CSDs are correct in this limit.

2.3.3 Multiple-deme, one-haplotype

We now demonstrate how the CSD $\hat{\pi}_{\mathrm{SMC}}$ described above can be extended to a structured population model including migration. Let $\mathbf{n} = (n_{d,h})_{d \in \mathcal{D}, h \in \mathcal{H}}$ be a structured sample, and consider sampling a single haplotype in deme $\alpha \in \mathcal{D}$ conditioned on $\mathbf{n}$, according to the trunk-conditional coalescent of Section 2.2.3. Embedded within the conditional genealogy at locus $\ell \in L$ is an MCG s_ℓ, and disregarding mutation events, s_ℓ is specified by the absorption time $t_\ell \in \mathbb{R}_{\geq 0}$ and haplotype $h_\ell \in \mathcal{H}$, as before, and also the migrational history Q_ℓ, which is represented by the sequence

$$Q_\ell = \big((t_0^{\mathrm{m}}, d_0^{\mathrm{m}}), (t_1^{\mathrm{m}}, d_1^{\mathrm{m}}), \ldots, (t_p^{\mathrm{m}}, d_p^{\mathrm{m}})\big), \tag{2.83}$$

where t_i^{m} and d_i^{m} are the time and destination deme of the i-th migration event (for ease of notation, the dependence on ℓ is not indicated), and $t_0^{\mathrm{m}} = 0$ and $d_0^{\mathrm{m}} = \alpha$, the deme from which the haplotype is sampled. It is possible that $p = 0$, corresponding to the case that the ancestral lineage associated with locus ℓ did not migrate prior to absorption. Thus, denoting the space of migrational histories by $\mathcal{Q}$, the state space for the MCG can be represented $\mathcal{S} = \mathbb{R}_{\geq 0} \times \mathcal{H} \times \mathcal{Q}$, and the MCG at locus $\ell \in L$ by $s_\ell = (t_\ell, h_\ell, Q_\ell) \in \mathcal{S}$.

As before, we begin by considering the distribution of S_ℓ induced by the conditional genealogical process. The migration and absorption dynamics at a single locus can be described by a continuous-time Markov process with a finite state space. The states can be divided into two groups: for each of the $d \in \mathcal{D}$, the state r_d corresponds to *residence* within deme d, and the state a_d corresponds to *absorption* into some haplotype within deme d. Letting $\mathcal{D} = \{1, 2, \ldots, q\}$, and ordering the states by $(r_1, \ldots, r_q, a_1, \ldots, a_q)$, the Markov process is specified by the following rate matrix,

$$Z = \begin{pmatrix} \Upsilon - A & A \\ 0 & 0 \end{pmatrix}, \tag{2.84}$$

where $\Upsilon = (\upsilon_{dd'}/2)_{d,d' \in \mathcal{D}}$ and $\upsilon_{dd} = \upsilon_d$, is the matrix of migration rates which govern the transitions between the first group of states (the residence states), and A is the diagonal matrix

$$A = \begin{pmatrix} \kappa_1^{-1} n_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \kappa_q^{-1} n_q \end{pmatrix} \tag{2.85}$$

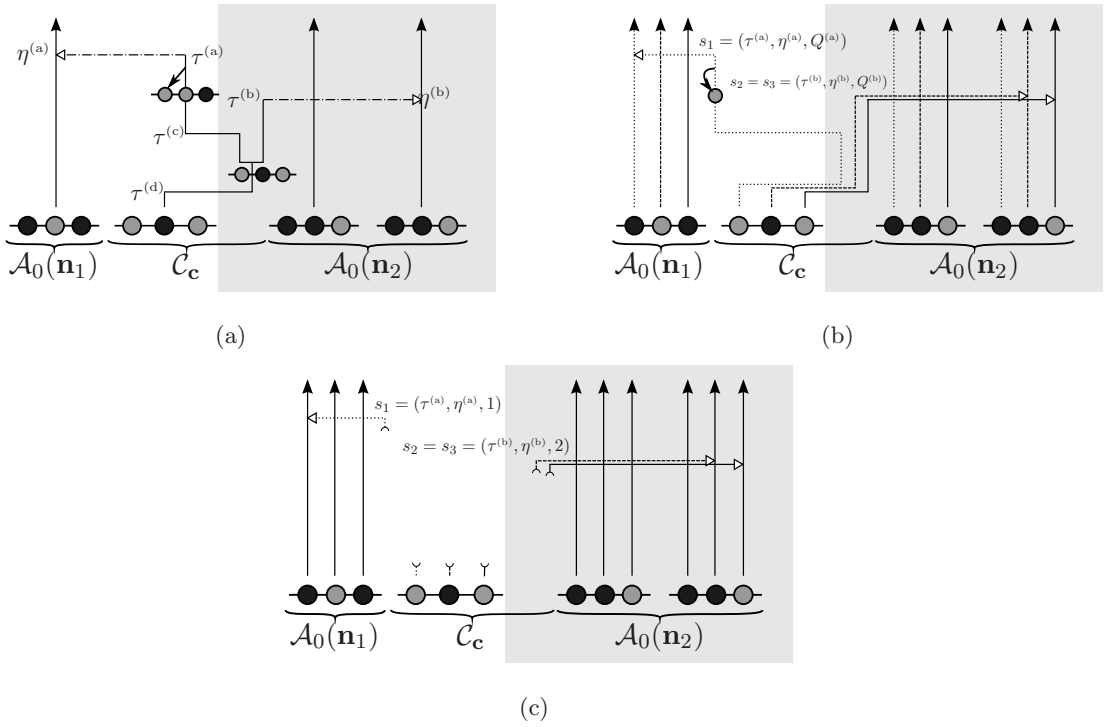


Figure 2.4. Illustration of the approximations to the conditional coalescent with recombination and migration, assuming two demes $\mathcal{D} = \{1, 2\}$, where deme $1 \in \mathcal{D}$ is shown in white and deme $2 \in \mathcal{D}$ is shown in light grey. The trunk genealogy $\mathcal{A}_0(\mathbf{n}_d)$ for each of the two demes $d \in \mathcal{D}$ is indicated, along with the conditional genealogy $\mathcal{C}_\mathbf{c}$. (a) The genealogical interpretation. Absorption events, and the corresponding absorption time ($\tau^{(a)}$ and $\tau^{(b)}$) and haplotype ($\eta^{(a)}$ and $\eta^{(b)}$), are indicated by dot-dashed horizontal lines. The times of the migration events ($\tau^{(c)}$ and $\tau^{(d)}$) are also indicated. (b) The corresponding sequential interpretation. The marginal genealogies (s_1, s_2, and s_3) at the first, second, and third locus are shown as dotted, dashed, and solid lines, respectively. We denote the two distinct migrational histories by $Q^{(a)} = \big((0,1),(\tau^{(d)},2),(\tau^{(c)},1)\big)$ and $Q^{(b)} = \big((0,1),(\tau^{(d)},2)\big)$. (c) The corresponding sequential interpretation where just the absorption time, deme, and haplotype are recorded. The gap in each MCG indicates that the specific migrational history is not preserved.

which governs the transition into the second group (the absorption states). The diagonal form of A ensures that the absorbed state a_d can be reached only if the ancestral lineage currently resides in deme d. The absorption rate within deme d is inversely proportional to the relative size of the deme, κ_d^{-1}, and proportional to the number of trunk-lineages n_d in deme d, as in the genealogical description in Section 2.2.3. Finally, because the absorption states are also absorbing in the context of the Markov chain, the rows of Z corresponding to these states are set to zero.

Using this process and the theory of continuous-time Markov processes, the marginal density $\zeta(\cdot)$ of the MCG s_ℓ is given by

$$\zeta(s_\ell) = \left(\prod_{i=1}^{p} Z(r_{d_{i-1}^{\mathrm{m}}}, r_{d_{i-1}^{\mathrm{m}}}, t_i^{\mathrm{m}} - t_{i-1}^{\mathrm{m}}) \right) \left(\frac{n_{d_p^{\mathrm{m}}, h_\ell}}{n_{d_p^{\mathrm{m}}}} \cdot Z(r_{d_p^{\mathrm{m}}}, a_{d_p^{\mathrm{m}}}, t_\ell - t_p^{\mathrm{m}}) \right), \tag{2.86}$$

where $Z(\alpha, \beta, t) = -\exp(t \cdot Z_{\alpha,\alpha}) \cdot Z_{\alpha,\beta} / Z_{\alpha,\alpha}$ is the probability of transitioning from state α to state β in time t for the process specified by Z. The first factor corresponds to each of the p migration events in Q_ℓ, and the second factor to the absorption event. Because the rates of absorption into each of the lineages within the absorption deme are identical, the absorption lineage is chosen uniformly at random within the absorption deme.

Conditioning on $S_{\ell-1} = s_{\ell-1} = (Q_{\ell-1}, t_{\ell-1}, h_{\ell-1})$, the MCG S_ℓ, for $\ell \geq 2$, is distributed according to a process similar to that described in Section 2.3.2. As before, there is no recombination between loci $\ell - 1$ and ℓ with probability $\exp(-\rho_b t_{\ell-1})$, and in this case $S_\ell = s_{\ell-1}$. Otherwise, the time T_{r} of the first recombination breakpoint is distributed exponentially with parameter ρ_b, truncated at time $t_{\ell-1}$. The lineage associated with locus ℓ is then subject to the marginal migration and absorption process, starting in the resident deme of the MCG $s_{\ell-1}$ at time T_{r}. Letting $b = (\ell - 1, \ell) \in B$, the transition density $\phi_b(\cdot | s_{\ell-1})$ is given by

$$\phi_b(s_\ell | s_{\ell-1}) = e^{-\rho_b t_{\ell-1}} \cdot \delta_{s_{\ell-1}, s_\ell} + \int_0^{t_{\ell-1} \wedge t_\ell} \delta_{Q_{\ell-1}[\downarrow t_{\mathrm{r}}], Q_\ell[\downarrow t_{\mathrm{r}}]} \cdot \rho_b e^{-\rho_b t_{\mathrm{r}}} \cdot \zeta(t_\ell - t_{\mathrm{r}}, h_\ell, Q_\ell[\uparrow t_{\mathrm{r}}]) dt_{\mathrm{r}}, \tag{2.87}$$

where we have denoted by $Q_\ell[\downarrow t]$ the sequence of migration events Q_ℓ truncated at time t, and by $Q_\ell[\uparrow t]$ the sequence of migration events induced by Q_ℓ starting at time t. Thus, in the second term, the δ factor ensures that, prior to the recombination event the sequence of migration events in $s_{\ell-1}$ and s_ℓ are identical.

Finally, because the mutation process does not depend on the deme in which a lineage resides, the emission density on alleles $\xi_\ell(\cdot | s_\ell)$ is identical to (2.75). In principle, $\hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_{d,h} | \mathbf{n})$ can thus be computed using the forward recursion detailed in Section 2.3.1. However, in practice, much of the mathematical and computational simplicity is lost due to the MCG state space $\mathcal{S}$, which has infinite dimension due to the presence of the migrational history. We next consider an additional approximation that enables practicable computation.

Absorption Deme Only

In order to reduce the MCG state space $\mathcal{S}$, we restrict the migrational history to the deme in which absorption occurred. As a result, the sequence of MCGs is no longer Markov, even under the sequentially Markov assumption. For example, suppose the absorption deme at locus ℓ is d; then knowledge that the absorption deme at locus $\ell - 1$ is $d' \neq d$ increases the probability that the absorption deme at locus $\ell + 1$ is d', introducing a non-Markov dependence. Nonetheless, it

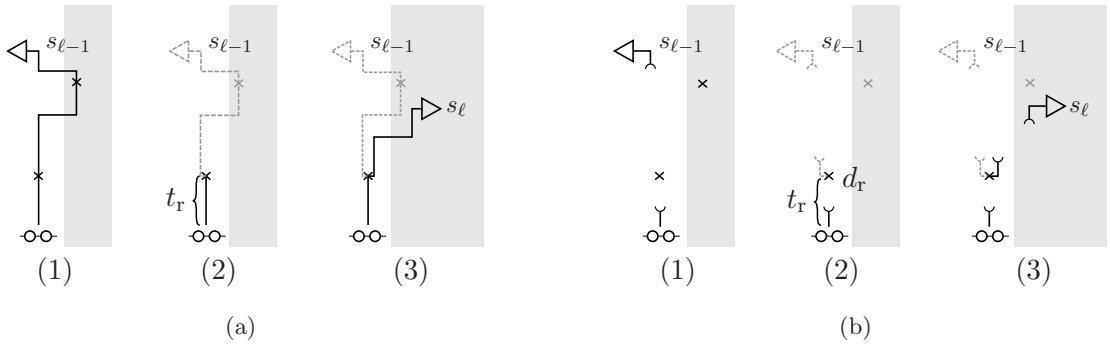


Figure 2.5. Illustration of the process for sampling the MCG S_ℓ conditioned on $S_{\ell-1} = s_{\ell-1}$, with population structure and migration. (a) Given the full migrational history $Q_{\ell-1}$, the MCG S_ℓ is sampled by (1) realizing recombination events, with breakpoint $b = (\ell-1, \ell) \in B$, as a Poisson process with rate ρ_b on the MCG $s_{\ell-1}$, (2) removing the lineage associated with locus $\ell-1$ branching from each breakpoint, so that only the lineage more recent than the first breakpoint, at time $T_\mathrm{r} = t_\mathrm{r}$, remains, (3) creating a new lineage associated with locus ℓ at the first breakpoint, and in the deme in which the recombination event occurred, and subjecting this lineage to migration and absorption events, producing the MCG $S_\ell = s_\ell$. (b) Given only the deme in which absorption occurred $D_{\ell-1} = d_{\ell-1}$, the process is similar to that above; in step (2) the deme D_r in which recombination occurred is not known, and so is sampled conditional on the absorption deme $D_{\ell-1} = d_{\ell-1}$ and recombination time $T_\mathrm{r} = t_\mathrm{r}$. The remainder of the process occurs as before, producing the MCG $S_\ell = s_\ell$.

is possible to further approximate this non-Markov process by a Markov process by integrating over the possible migrational histories consistent with the given absorption deme. We denote the resulting approximation to $\hat{\pi}_{\mathrm{SMC}}$ by $\hat{\pi}_{\text{SMC-ADO}}$, where "ADO" is an abbreviation for "absorption deme only".

Denote the absorption deme at locus ℓ by $d_\ell \in \mathcal{D}$, so that the reduced MCG state space is given by $\hat{\mathcal{S}} = \mathbb{R}_{\geq 0} \times \mathcal{H} \times \mathcal{D}$, and the MCG at locus ℓ is given by the triple $\hat{s}_\ell = (t_\ell, h_\ell, d_\ell) \in \hat{\mathcal{S}}$. As before, the migration and absorption distribution at a single locus are specified by the rate matrix Z. Because $\hat{s}_\ell$ only specifies the absorption deme, the reduced marginal density $\zeta(\cdot)$ is given by

$$\zeta(\hat{s}_\ell) = \frac{n_{d_\ell, h_\ell}}{n_{d_\ell}} \cdot \left[Z e^{Z t_\ell} \right]_{r_\alpha, a_{d_\ell}}. \tag{2.88}$$

Because the rates of absorption into each lineage of the absorption deme are identical, the absorption lineage is chosen uniformly at random within the absorption deme. By virtue of not incorporating information about the entire migration history, (2.88) is considerably simpler than (2.86).

Conditioning on $\hat{S}_{\ell-1} = \hat{s}_{\ell-1} = (t_{\ell-1}, h_{\ell-1}, d_{\ell-1})$, the MCG $\hat{S}_\ell$, for $\ell \geq 2$, is distributed according to a process similar to that described above. As before, there is no recombination at $b = (\ell-1, \ell)$ with probability $\exp(-\rho_b t_{\ell-1})$, and otherwise the time T_r of the first recombination breakpoint is distributed exponentially with parameter ρ_b, truncated at time $t_{\ell-1}$. The difference in this case is that the deme D_r in which recombination occurs is not known. Conditioned on $\hat{S}_{\ell-1} = \hat{s}_{\ell-1}$ and the time of recombination $T_\mathrm{r} = t_\mathrm{r}$, the density $f(\cdot | \hat{s}_{\ell-1}, t_\mathrm{r})$ of the deme in which recombination occurs D_r is given by,

$$f(d | \hat{s}_{\ell-1}, t_\mathrm{r}) = \frac{\left[e^{Z t_\mathrm{r}} \right]_{r_\alpha, r_d} \left[Z e^{Z(t_{\ell-1} - t_\mathrm{r})} \right]_{r_d, a_{d_{\ell-1}}}}{\left[Z e^{Z t_{\ell-1}} \right]_{r_\alpha, a_{d_{\ell-1}}}}. \tag{2.89}$$

Conditioned on the time $T_{\mathrm{r}} = t_{\mathrm{r}}$ and deme $D_{\mathrm{r}} = d_{\mathrm{r}}$ at which recombination occurs, the lineage associated with locus ℓ is then subject to the marginal migration and absorption process, starting in deme D_{r} at time T_{r}. This process yields the transition distribution $\phi_b(\cdot|\hat{s}_{\ell-1})$ given by

$$\begin{aligned}\phi_b(\hat{s}_\ell|\hat{s}_{\ell-1}) &= e^{-\rho_b t_{\ell-1}} \cdot \delta_{\hat{s}_{\ell-1},\hat{s}_\ell} \\ &\quad + \int_0^{t_{\ell-1}\wedge t_\ell} \rho_b e^{-\rho_b t_{\mathrm{r}}} \sum_{d_{\mathrm{r}}\in\mathcal{D}} f(d_{\mathrm{r}}|\hat{s}_{\ell-1}, t_{\mathrm{r}}) \left(\frac{n_{d_\ell,h_\ell}}{n_{d_\ell}} \cdot \left[Z e^{Z(t_\ell - t_{\mathrm{r}})} \right]_{r_{d_{\mathrm{r}}}, a_{d_\ell}} \right) dt_{\mathrm{r}}.\end{aligned} \tag{2.90}$$

Once again, the mutation process does not depend on the deme in which a lineage resides, and so the emission distribution on alleles $\xi_\ell(\cdot|\hat{s}_\ell)$ is again identical to (2.75). Thus, in principle, $\hat{\pi}_{\text{SMC-ADO}}(\mathbf{e}_{d,h}|\mathbf{n})$ can be approximated using the forward recursion detailed in Section 2.3.1, substituting in the reduced initial and transition distributions. We thoroughly describe a practical implementation for the recursion in Chapter 3, and consider the accuracy of the approximation to $\hat{\pi}_{\text{SMC-ADO}}$ in light of empirical results in Chapter 4.

2.3.4 Single-deme, two-haplotype

Finally, we demonstrate how the CSD $\hat{\pi}_{\text{SMC}}$ can be extended to conditionally sampling more than one haplotype. As before, let $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ and consider sampling two haplotypes conditioned on $\mathbf{n}$ according to the trunk-conditional coalescent of Section 2.2.2. Embedded within the conditional genealogy at locus $\ell \in L$ is an MCG s_ℓ, and disregarding mutation events, s_ℓ is specified by,

1. The MCG for the first haplotype, denoted by $m_\ell^{(1)} = (t_\ell^{(1)}, h_\ell^{(1)})$, comprising the absorption time $t_\ell^{(1)}$ and haplotype $h_\ell^{(1)}$.
2. The MCG for the second haplotype, denoted by $m_\ell^{(2)} = (t_\ell^{(2)}, h_\ell^{(2)})$, comprising the absorption time $t_\ell^{(2)}$ and haplotype $h_\ell^{(2)}$.
3. The coalescence time, denoted by $t_\ell^{(\mathrm{c})}$, within s_ℓ of the first and second conditionally sampled haplotype. We set $t_\ell^{(\mathrm{c})} = \emptyset$ to denote that there is no coalescence event within s_ℓ at locus ℓ.

See Figure 2.6 for an illustration. Observe that if the two haplotypes coalesce prior to absorption, the MCGs for each haplotype must be identical; formally $t^{(\mathrm{c})} \neq \emptyset$ implies that $m^{(1)} = m^{(2)} = m$, and moreover that $t^{(\mathrm{c})} < t^{(1)}, t^{(2)}$. By the contrapositive, $m^{(1)} \neq m^{(2)}$ implies that $t^{(\mathrm{c})} = \emptyset$. Thus, letting $\mathcal{M} = \mathbb{R}_{\geq 0} \times \mathcal{H}$, the MCG state space $\mathcal{S}$ is given by

$$\mathcal{S} = \left\{ (m^{(1)}, m^{(2)}, t^{(\mathrm{c})}) \in \mathcal{M} \times \mathcal{M} \times (\mathbb{R}_{\geq 0} \cup \emptyset) : t^{(\mathrm{c})} \neq \emptyset \Rightarrow m^{(1)} = m^{(2)} > t^{(\mathrm{c})} \right\}. \tag{2.91}$$

For ease of notation, we shall also frequently write, for $t \in \mathbb{R}_{\geq 0}$ and $s = (m^{(1)}, m^{(2)}, t^{(\mathrm{c})}) \in \mathcal{S}$, that $t < s$ to indicate that either $t^{(\mathrm{c})} = \emptyset$ and $t < t^{(1)}, t^{(2)}$ or $t^{(\mathrm{c})} \neq \emptyset$ and $t < t^{(\mathrm{c})} < t^{(1)}, t^{(2)}$.

In unconditionally sampling the MCG S_ℓ, the lineages associated with each of the two haplotypes are *free* in the sense that they are subject to the coalescence and absorption events specified by the genealogical process. In contrast, in sampling the MCG S_ℓ conditional upon $S_{\ell-1} = s_{\ell-1}$, the lineages associated with each of two haplotypes are initially *anchored* to the lineages of $s_{\ell-1}$. However, when a recombination event occurs on the shared lineage, the lineage associated with locus ℓ is no longer anchored, and becomes free. Though we did not require the terminology, we made use of this logic in Sections 2.3.2 and 2.3.3 in order to write down the marginal and transition distributions. Thus, letting $b = (\ell - 1, \ell) \in B$, we define the following densities,

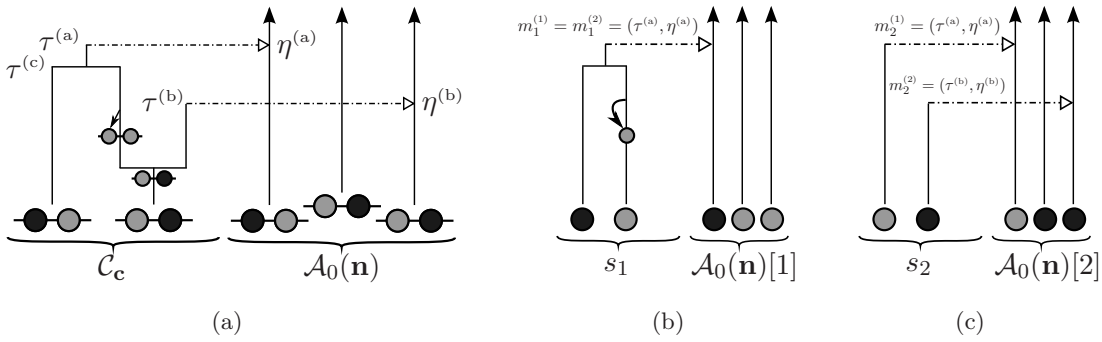


Figure 2.6. Illustration of the corresponding genealogical and sequential interpretations of a conditional genealogy $\mathcal{C}_{\mathbf{c}}$ with respect to the trunk genealogy $\mathcal{A}_0(\mathbf{n})$ for two conditionally sampled haplotypes. (a) The genealogical interpretation. Absorption events, and the corresponding absorption time ($\tau^{(a)}$ and $\tau^{(b)}$) and haplotype ($\eta^{(a)}$ and $\eta^{(b)}$, respectively), are indicated by dot-dashed horizontal lines. (b) The corresponding sequential interpretation. The marginal genealogies at the first and second locus (s_1 and s_2) are provided. Note that $t_1^{(c)} = \tau^{(c)}$ and $t_2^{(c)} = \emptyset$.

$f_t^{(f)}(m_\ell)$: The density associated with sampling the one-haplotype MCG M_ℓ conditioned on the lineage being free at time t.

$f_{b,t}^{(a)}(m_\ell|m_{\ell-1})$: The density associated with sampling the one-haplotype MCG M_ℓ conditioned on the lineage being anchored to $M_{\ell-1} = m_{\ell-1} \in \mathcal{M}$ at time t.

$f_t^{(f,f)}(s_\ell)$: The density associated with sampling the two-haplotype MCG S_ℓ conditioned on both of the lineages being free at time t.

$f_{b,t}^{(f,a)}(s_\ell|m_{\ell-1})$ $[f_{b,t}^{(a,f)}(s_\ell|m_{\ell-1})]$: The density associated with sampling the two-haplotype MCG S_ℓ conditioned on the lineage associated with haplotype 1 [respectively, haplotype 2] being free, and the lineage associated with haplotype 2 [respectively, haplotype 1] being anchored to the one-haplotype MCG $M_{\ell-1} = m_{\ell-1} \in \mathcal{M}$ at t.

$f_{b,t}^{(a,a)}(s_\ell|s_{\ell-1})$: The density associated with sampling the two-haplotype MCG S_ℓ conditioned on the lineages associated with both haplotypes being anchored to two-haplotype MCG $S_{\ell-1} = s_{\ell-1} \in \mathcal{S}$ at t.

Observe that $f_0^{(f)}(\cdot)$ and $f_{b,0}^{(a)}(\cdot|m_{\ell+1})$ are precisely the one-haplotype marginal and transition distributions discussed in Section 2.3.2. In precisely the same way, $f_0^{(f,f)}(s_\ell)$ and $f_{b,0}^{(a,a)}(s_{\ell+1}|s_\ell)$ are the two-haplotype marginal and transition distributions. We now demonstrate a technique for deriving expressions for these densities in a systematic way. The technique is a generalization of the basic reasoning used in the previous sections. Critically, it is possible for anchored lineages to become free via recombination, but free lineages cannot become anchored without reducing the total number of lineages; thus, it is possible to write densities involving more (anchored) lineages in terms of densities involving fewer (anchored) lineages.

We begin by re-deriving the one-haplotype densities in this more general setting. Considering first the density $f_t^{(f)}(\cdot)$, the single free lineage is absorbed into each lineage of the trunk genealogy $\mathcal{A}_0(\mathbf{n})$ at rate 1, so that the total rate is $|\mathbf{n}| = n$. Integrating over the time of the absorption event,

$$f_t^{(f)}(m_\ell) = \int_t^\infty \delta_{t_{\mathrm{a}},t_\ell} \frac{n_{h_\ell}}{n} n e^{-n(t_{\mathrm{a}}-t)} dt_{\mathrm{a}} = n_{h_\ell} e^{-n(t_\ell - t)} \tag{2.92}$$

for $t_\ell > t$. For the density $f^{(a)}_{b,t}(\cdot|m_\ell)$, the anchored lineage is subject to recombination at rate ρ_b; if recombination occurs, the lineage becomes free at the time of recombination. Integrating over these possibilities and the time of the recombination event,

$$\begin{aligned} f^{(a)}_{b,t}(m_\ell|m_{\ell-1}) &= e^{-\rho_b(t_{\ell-1}-t)}\delta_{m_\ell,m_{\ell-1}} + \int_t^{t_{\ell-1}} \rho_b e^{-\rho_b(t_\mathrm{r}-t)} f^{(f)}_{\ell,t_\mathrm{r}}(m_\ell)dt_\mathrm{r} \\ &= e^{-\rho_b(t_{\ell-1}-t)}\delta_{m_\ell,m_{\ell-1}} + n_{\ell-1}\int_t^{t_{\ell-1}\wedge t_\ell} \rho_b e^{-\rho_b(t_\mathrm{r}-t)}e^{-n(t_\ell-t_\mathrm{r})}dt_\mathrm{r}, \end{aligned} \tag{2.93}$$

for $t_{\ell-1}, t_\ell > t$, where the second equality is by direct substitution, taking into account the time boundary for $f^{(f)}_t(\cdot)$. As anticipated, these expressions are identical to those derived in Section 2.3.2 when setting $t = 0$.

Continuing with the two-haplotype density $f^{(f,f)}_t(\cdot)$, each of the two free lineages is absorbed into each lineage of the trunk genealogy $\mathcal{A}_0(\mathbf{n})$ at rate 1, and the two free lineages coalesce at rate 2, so that the total rate is $2n + 2$. If a lineage is absorbed, the remaining lineage becomes a single free lineage at the time of absorption, and if the two lineages coalesce, the resulting lineage becomes a single free lineage at the time of coalescence. Thus, integrating over the time of the first event,

$$\begin{aligned} f^{(f,f)}_t(s_\ell) &= \int_t^\infty (2n+2)e^{-(2n+2)(t_\mathrm{e}-t)}\bigg[\frac{2}{2n+2}\delta_{t_\mathrm{e},t^{(\mathrm{c})}_\ell} f^{(f)}_{t_\mathrm{e}}(m) \\ &\qquad + \frac{n}{2n+2}\Big(\delta_{t_\mathrm{e},t^{(1)}_\ell}\frac{n_{h^{(1)}_\ell}}{n} f^{(f)}_{t_\mathrm{e}}(m^{(2)}_\ell) + \delta_{t_\mathrm{e},t^{(2)}_\ell}\frac{n_{h^{(2)}_\ell}}{n} f^{(f)}_{t_\mathrm{e}}(m^{(1)}_\ell)\Big)\bigg]dt_\mathrm{e} \\ &= \big[1-\delta_{t^{(\mathrm{c})}_\ell,\emptyset}\big]2e^{-(2n+2)(t^{(\mathrm{c})}_\ell-t)} f^{(f)}_{t^{(\mathrm{c})}_\ell}(m) \\ &\quad + \big[\mathbb{1}_{(t^{(1)}_\ell<t^{(2)}_\ell)}\big]e^{-(2n+2)(t^{(1)}_\ell-t)} n_{h^{(1)}_\ell} f^{(f)}_{t^{(1)}_\ell}(m^{(2)}_\ell) \\ &\quad + \big[\mathbb{1}_{(t^{(2)}_\ell<t^{(1)}_\ell)}\big]e^{-(2n+2)(t^{(2)}_\ell-t)} n_{h^{(2)}_\ell} f^{(f)}_{t^{(2)}_\ell}(m^{(1)}_\ell), \end{aligned} \tag{2.94}$$

for $s_\ell > t$. For the two-haplotype density $f^{(f,a)}_{b,t}(\cdot|m_{\ell-1})$, the anchored lineage is subject to recombination at rate ρ_b and the free lineage is subject to absorption into each lineage of the trunk genealogy and coalescence with the anchored lineage, at rates 1 and 2, respectively. The total rate of events is $\rho_b + n + 2$. If no event occurs prior to the absorption of the anchored lineage, the free lineage becomes a single free lineage at the time of absorption. Otherwise, if recombination occurs, the anchored lineage becomes free and there are two free lineages; if absorption or coalescence occurs, there remains a single anchored lineage at the time of the event. Integrating over these

possibilities and the time of the first event,

$$\begin{aligned}
f_{b,t}^{(f,a)}(s_\ell|m_{\ell-1}^{(2)}) &= e^{-(\rho_b+n+2)(t_{\ell-1}^{(2)}-t)}\delta_{m_\ell^{(2)},m_{\ell-1}^{(2)}} f_{\ell,t_{\ell-1}^{(2)}}^{(f)}(m_{\ell-1}^{(1)}) \\
&\quad + \int_t^{t_{\ell-1}^{(2)}} e^{-(\rho_b+n+2)(t_\mathrm{e}-t)}\Big[\rho_b f_{\ell,t_\mathrm{e}}^{(f,f)}(s_\ell) \\
&\qquad + 2\delta_{t_\mathrm{e},t_\ell^{(\mathrm{c})}} f_{b,t_\mathrm{e}}^{(a)}(m_\ell|m_{\ell-1}^{(2)}) + \delta_{t_\mathrm{e},t_\ell^{(1)}} n_{h_\ell^{(1)}} f_{b,t_\mathrm{e}}^{(a)}(m_\ell^{(2)}|m_{\ell-1}^{(2)})\Big] dt_\mathrm{e} \\
&= \big[1-\delta_{t_\ell^{(\mathrm{c})},\emptyset}\big] 2e^{-(\rho_b+n+2)(t_\ell^{(\mathrm{c})}-t)} f_{b,t_\ell^{(\mathrm{c})}}^{(a)}(m_\ell|m_{\ell-1}^{(2)}) \\
&\quad + \big[\mathbb{1}_{(t_\ell^{(1)}<t_\ell^{(2)})}\big] e^{-(\rho_b+n+2)(t_\ell^{(1)}-t)} n_{h_\ell^{(1)}} f_{b,t_\ell^{(1)}}^{(a)}(m_\ell^{(2)}|m_{\ell-1}^{(2)}) \\
&\quad + \big[\mathbb{1}_{(t_\ell^{(2)}<t_\ell^{(1)})}\big] e^{-(\rho_b+n+2)(t_{\ell-1}^{(2)}-t)} \delta_{m_\ell^{(2)},m_{\ell-1}^{(2)}} f_{\ell,t_{\ell-1}^{(2)}}^{(f)}(m_{\ell-1}^{(1)}) \\
&\quad + \int_t^{t_{\ell-1}^{(2)}} e^{-(\rho_b+n+2)(t_\mathrm{r}-t)} \rho_b f_{\ell,t_\mathrm{r}}^{(f,f)}(s_\ell) dt_\mathrm{r},
\end{aligned} \tag{2.95}$$

where $t_{\ell-1}^{(2)}, s_\ell > t$. The reasoning and outcome for $f_{b,t}^{(a,f)}(\cdot|m_{\ell-1}^{(1)})$ is identical, with all of the one-haplotype MCG labels reversed.

Finally, for the two-haplotype distribution $f_{b,t}^{(a,a)}(\cdot|s_{\ell-1})$, either coalescence or an absorption event occurs first within $s_{\ell-1}$. In each situation, recombination occurs on each lineage at rate ρ_b so that the total rate is $2\rho_b$. If recombination does not occur, the result is a single anchored lineage, and if it does occur, the result is a single anchored lineage and a single free lineage. Integrating over these possibilities and the time of the recombination event,

$$\begin{aligned}
f_{b,t}^{(a,a)}(s_\ell|s_{\ell-1}) &= \big[1-\delta_{t_{\ell-1}^{(\mathrm{c})},\emptyset}\big] e^{-2\rho_b(t_{\ell-1}^{(\mathrm{c})}-t)} \delta_{t_{\ell-1}^{(\mathrm{c})},t_\ell^{(\mathrm{c})}} f_{b,t_{\ell-1}^{(\mathrm{c})}}^{(a)}(m_\ell|m_{\ell-1}) \\
&\quad + \big[\mathbb{1}_{(t_{\ell-1}^{(1)}<t_{\ell-1}^{(2)})}\big] e^{-2\rho_b(t_{\ell-1}^{(1)}-t)} \delta_{t_{\ell-1}^{(1)},t_\ell^{(1)}} f_{b,t_{\ell-1}^{(1)}}^{(a)}(m_\ell^{(2)}|m_{\ell-1}^{(2)}) \\
&\quad + \big[\mathbb{1}_{(t_{\ell-1}^{(2)}<t_{\ell-1}^{(1)})}\big] e^{-2\rho_b(t_{\ell-1}^{(2)}-t)} \delta_{t_{\ell-1}^{(2)},t_\ell^{(2)}} f_{b,t_{\ell-1}^{(2)}}^{(a)}(m_\ell^{(1)}|m_{\ell-1}^{(1)}) \\
&\quad + \int_t^{s_{\ell-1}} e^{-2\rho_b(t_\mathrm{r}-t)} \rho_b \Big(f_{b,t_\mathrm{r}}^{(f,a)}(s_\ell|m_{\ell-1}^{(2)}) + f_{b,t_\mathrm{r}}^{(a,f)}(s_\ell|m_{\ell-1}^{(2)}) \Big) dt_\mathrm{r},
\end{aligned} \tag{2.96}$$

where $s_\ell, s_{\ell-1} < t$. Though we don't reproduce the work here, it is practically straightforward to obtain closed-form expressions for the two-haplotype marginal density $\zeta(\cdot) = f_0^{(f,f)}(\cdot)$ and the two-haplotype transition density $\phi_b(\cdot|s_{\ell-1}) = f_{b,0}^{(a,a)}(\cdot|s_{\ell-1})$ by direct substitution of the relevant expressions into (2.94) and (2.96). In Appendix B.2, we provide a proof that the two-haplotype transition density $\phi_b(\cdot|s_{\ell-1})$ satisfies detailed balance with respect to the two-haplotype marginal density $\zeta(\cdot)$, analogous to the one-haplotype case described in Section 2.3.2.

Finally, we consider the emission densities. We first consider the case that the two haplotypes have been sampled separately, so that the two observed alleles $(a_1, a_2) \in A_\ell \times A_\ell$ at locus $\ell \in L$ are *ordered*. For convenience, we define the density $f_\ell(\cdot|a', t)$ associated with the mutation process at locus ℓ for time $t \in \mathbb{R}_{\geq 0}$, and starting with allele $a' \in A_\ell$,

$$f_\ell(a|a', t) = e^{-\theta_\ell t} \sum_{m=0}^{\infty} \frac{(\theta_\ell t)^m}{m!} \Big[\big(\Phi^{(\ell)}\big)^m \Big]_{a',a}. \tag{2.97}$$

Letting $S_\ell = s_\ell \in \mathcal{S}$, if $t_\ell^{(\mathrm{c})} = \emptyset$, the two observed alleles are entirely independent, and if $t_\ell^{(\mathrm{c})} \neq \emptyset$ we may partition with respect to the unknown common allele at the time of coalescence. Note that this latter operation is a very simple application of Felsenstein's algorithm (Felsenstein, 1981). Therefore,

$$\begin{aligned}\xi_\ell((a_1,a_2)|s_\ell) = \big[\delta_{t_\ell^{(\mathrm{c})},\emptyset}\big] f_\ell(a_1|h_\ell^{(1)}[\ell], t_\ell^{(1)}) f_\ell(a_2|h_\ell^{(2)}[\ell], t_\ell^{(2)}) \\ + \big[1-\delta_{t_\ell^{(\mathrm{c})},\emptyset}\big] \sum_{a\in A_\ell} f_\ell(a|h_\ell[\ell], t_\ell) f_\ell(a_1|a, t_\ell^{(\mathrm{c})}) f_\ell(a_2|a, t_\ell^{(\mathrm{c})}).\end{aligned} \tag{2.98}$$

In many cases, the two haplotypes are not sampled separately, so that the alleles at each locus are not ordered; this type of data is often referred to as *unphased*. For example, in the two locus case, the observed data may be the alleles $a_1, a_2 \in A_1$ at locus 1, and $b_1, b_2 \in A_2$ at locus 2, but without knowledge as to whether the haplotypes are $(a_1,b_1),(a_2,b_2) \in \mathcal{H}$ or $(a_1,b_2),(a_2,b_1) \in \mathcal{H}$. Denoting the alleles of the unphased data by $\{a_1,a_2\}$ and summing over the possible orderings, which are *a priori* equally likely,

$$\xi_\ell(\{a_1,a_2\}|s_\ell) = \frac{1}{2}\Big(\xi_\ell((a_1,a_2)|s_\ell) + \xi_\ell((a_2,a_1)|s_\ell)\Big). \tag{2.99}$$

Thus, the CSP $\hat{\pi}_{\mathrm{SMC}}(h_1,h_2|\mathbf{n})$ can be computed using the forward recursion detailed in Section 2.3.1 can be applied, substituting in the initial, transition, and (phased or unphased) emission densities derived in this section. This is in contrast to the recursion for $\hat{\pi}_{\mathrm{PS}}$ described in Sections 2.1.2 and 2.2.2, which is not immediately applicable to unphased data; in fact, the most straightforward way to apply these recursions is to sum over each possible phasing, of which there are an exponential number, further reducing the efficiency of the recursion.

Finally, note that the general technique described in this section could, in principle, be extended to more than two haplotypes, and to structured populations and other demographic scenarios of the type illustrated in Section 2.3.3. In practice, however, without further approximation, we anticipate that the requisite algebra and even the ultimate closed-form solutions would become overly-complex for more than a very modest number of haplotypes.

2.3.5 Relationships among approximate CSDs

Throughout this chapter, we have stated and proved several relationships between $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{SMC}}$ and previously proposed CSDs, such as $\hat{\pi}_{\mathrm{SD}}$ and $\hat{\pi}_{\mathrm{FD}}$. In this section, these relationships are revisited and summarized. Begin by recalling that, in the absence of recombination, and for a single conditionally sampled haplotype $\hat{\pi}_{\mathrm{FD}} = \hat{\pi}_{\mathrm{SD}} = \hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{SMC}}$. The first equality is by construction, as described in Section 1.4.2, the second equality stated in Proposition 2.5, and the final equality is by construction, as described in Section 2.3. While $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{SD}}$ are not defined for more than one conditionally sampled haplotype, the final equality $\hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{SMC}}$ holds for an arbitrary number of conditionally sampled haplotypes. Finally, in the special case of a one-locus PIM model, for which recombination is not applicable, Proposition 2.3 proves that the CSDs are correct.

We next consider the case when $\rho_b = \rho$ for all $b \in B$ and the limit $\rho \to \infty$. We have seen that $\hat{\pi}_{\mathrm{FD}}$, $\hat{\pi}_{\mathrm{PS}}$, and $\hat{\pi}_{\mathrm{SMC}}$ all have the same limiting decomposition into a product of one-locus CSDs, demonstrated in Section 1.4.2, Section 2.1.2, and Section 2.3.2, respectively. As stated above, the one-locus CSDs are also identical, and so in this limit $\hat{\pi}_{\mathrm{FD}} = \hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{SMC}}$. Moreover, for a PIM model, the one-locus CSDs are correct, and therefore each of the resulting multiple-locus CSDs are also correct.

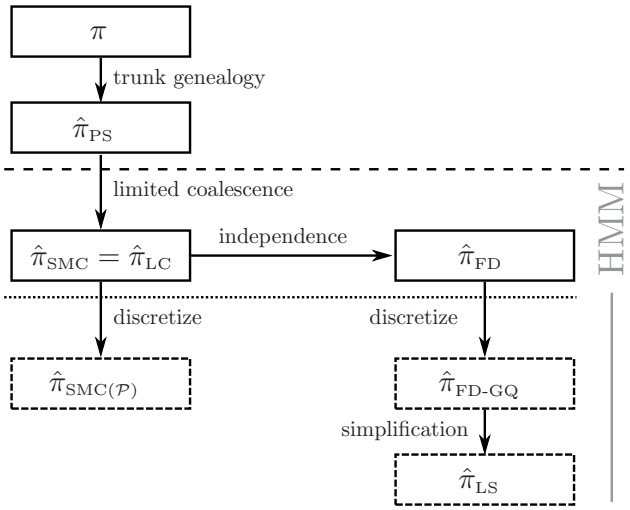


Figure 2.7. Illustration of the relationship between various CSDs. The CSD at the head of each arrow can be seen as an approximation to the CSD at the tail. Each arrow is also annotated with a (short) description of this approximation. The CSDs below the dashed line can be cast as an HMM: those above the dotted line have a continuous and infinite state space, while those below (including the discretized version of $\hat{\pi}_{\mathrm{SMC}}$, denoted $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$, to be described in Section 3.2 and the Gaussian quadrature discretized version of $\hat{\pi}_{\mathrm{FD}}$, which we denote $\hat{\pi}_{\mathrm{FD\text{-}GQ}}$) have a finite and discrete state space and are therefore amenable to simple dynamic programming algorithms. For more thorough descriptions of each approximation, see the main text. The equality $\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{LC}}$ has only been proved in the setting of a single conditionally sampled haplotype.

Finally, we consider the more general case, when the recombination rate is not restricted. As described in Section 2.3.1, $\hat{\pi}_{\mathrm{SMC}}$ is an approximation to $\hat{\pi}_{\mathrm{PS}}$ based on a sequentially Markov interpretation of the MCGs. Similarly, we have shown in Theorem 2.14 that, for a single conditionally sampled haplotype, $\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{NC}}$, where $\hat{\pi}_{\mathrm{NC}}$ is a modification to the conditional coalescent for which coalescence events are disallowed. More generally, we have conjectured that for multiple conditionally sampled haplotypes, $\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{LC}}$, where $\hat{\pi}_{\mathrm{LC}}$ is a modification to the conditional coalescent for which coalescence events between lineages with non-overlapping ancestral loci are disallowed.

In order to understand the relationship between $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{SMC}}$, we express $\hat{\pi}_{\mathrm{FD}}$ in an HMM framework similar to $\hat{\pi}_{\mathrm{SMC}}$. Let $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$ with $|\mathbf{n}| = n$, and recall from Section 1.4.2 that $\hat{\pi}_{\mathrm{FD}}$ extends $\hat{\pi}_{\mathrm{SD}}$ by introducing a recombination event at each breakpoint with probability $\rho_b/(n + \rho_b)$. Recombination events split the haplotype into intervals, and each interval is then sampled independently using $\hat{\pi}_{\mathrm{SD}}$; each interval is characterized by a haplotype chosen uniformly at random from $\mathbf{n}$, and a time chosen according to an exponential distribution with rate n. For locus $\ell \in L$, denote by (T_ℓ, H_ℓ) the random time and haplotype associated with the interval to which the locus belongs. Because the recombination events are independent, the sequence of random states is Markov, with marginal density

$$\zeta^{(\mathrm{FD})}(t_\ell, h_\ell) = n_{h_\ell} e^{-n t_\ell}, \tag{2.100}$$

and, letting $b = (\ell - 1, \ell) \in B$, transition density

$$\phi_b^{(\mathrm{FD})}(t_\ell, h_\ell | t_{\ell-1}, h_{\ell-1}) = \frac{n}{n + \rho_b} \cdot \delta_{t_{\ell-1}, t_\ell} \delta_{h_{\ell-1}, h_\ell} + \frac{\rho_b}{n + \rho_b} \cdot \frac{n_{h_\ell}}{n} e^{-t_\ell}, \tag{2.101}$$

where the first term and second terms in the transition densities correspond to no recombination and recombination, respectively, at breakpoint $b \in B$. Finally, conditioned on $(T_\ell, H_\ell) = (t_\ell, h_\ell)$, the allele at locus ℓ is independently sampled by mutating allele $h_\ell[\ell]$ a random number m_ℓ times, where m_ℓ is Poisson-distributed with mean $\theta_\ell t_\ell / n$. The emission density is therefore

$$\xi_\ell^{(\mathrm{FD})}(a|t_\ell, h_\ell) = e^{-\theta_\ell t_\ell} \sum_{m=0}^{\infty} \frac{\left(\theta_\ell t_\ell\right)^m}{m!} \cdot \left[\left(\Phi^{(\ell)}\right)^m\right]_{h_\ell[\ell],a}. \tag{2.102}$$

Comparing these equations to (2.73),(2.74), and (2.75), respectively, the HMM formulation of $\hat{\pi}_{\mathrm{FD}}$ is identical to $\hat{\pi}_{\mathrm{SMC}}$ with the exception of the transition density. Relative to the transition density associated with $\hat{\pi}_{\mathrm{SMC}}$, the transition density for $\hat{\pi}_{\mathrm{FD}}$ makes two assumptions: first, the probability of recombination is independent of $t_{\ell-1}$; and second, conditioned on recombination at $b = (\ell - 1, \ell) \in B$, the distribution of T_ℓ is independent of $t_{\ell-1}$. In the context of the trunk-conditional coalescent process, both of these independence assumptions are false, and we therefore expect that $\hat{\pi}_{\mathrm{SMC}}$ is a better approximation to the true CSD than $\hat{\pi}_{\mathrm{FD}}$.

In order to develop practicable algorithms for evaluating the CSP associated with $\hat{\pi}_{\mathrm{SMC}}$ and $\hat{\pi}_{\mathrm{FD}}$, it is necessary to discretize the continuous state space. The discretization procedure for $\hat{\pi}_{\mathrm{SMC}}$ will be considered in detail in Section 3.2. As discussed in Section 1.4.2, Fearnhead and Donnelly (2001) use Gaussian quadrature to discretize $\hat{\pi}_{\mathrm{FD}}$. Finally, recall from Section 1.4.3 that the CSD $\hat{\pi}_{\mathrm{LS}}$ is a simplification to a discretized version of $\hat{\pi}_{\mathrm{FD}}$. The relationships between these CSDs is summarized in Figure 2.7.

Chapter 3

Algorithms & Implementation

In the previous chapter, we introduced several techniques for obtaining an approximate conditional sampling distribution (CSD) for the coalescent with recombination. We discussed these techniques in the context of obtaining an approximate CSD that is both highly accurate and efficiently computable, the latter mandated by the large and growing repository of genetic and genomic data. In this chapter, we quantify the computational efficiency of evaluating the conditional sampling probability (CSP) associated with each CSD, providing several concrete algorithms and the associated asymptotic time complexities.

We demonstrate that explicit evaluation of the CSP associated with $\hat{\pi}_{\text{PS}}$, resulting from direct application of the diffusion-generator approximation, or equivalently the trunk-conditional coalescent, has computational complexity super-exponential in the number of loci, and is therefore computationally intractable for even modestly sized samples (Paul and Song, 2010). The CSD $\hat{\pi}_{\text{SMC}}$, resulting from the sequentially Markov approximation, can be approximated by a discrete-space HMM, and the associated CSP evaluated with computational complexity *linear* in the number of loci (Paul et al., 2011). Finally, taking advantage of the particular form of the forward and backward recursions in the context of the CSP computation, it is possible to obtain an algorithm that is, in practice, substantially faster than that obtained using ordinary HMM methodology (Paul and Song, 2012).

3.1 Computing $\hat{\pi}_{\text{PS}}$

We begin by considering computation of the CSP in the multiple-locus setting of Section 2.1.2. Letting $\mathbf{c} = (c_g)_{g \in \mathcal{G}}$ and $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$, recall that the recursion (2.12) for $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$ is given by

$$\begin{aligned}\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n}) = \frac{1}{\mathcal{N}} \sum_{g \in \mathcal{G}} c_g \Bigg\{ & \Bigg(\sum_{h \in \mathcal{H}: h \not\perp g} n_h \Bigg) \hat{\pi}_{\text{PS}}(\mathbf{c} - \mathbf{e}_g | \mathbf{n}) \\ & + \sum_{g' \in \mathcal{G}: g' \not\perp g} (c_{g'} - \delta_{g,g'}) \hat{\pi}_{\text{PS}}(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{C}(g,g')} | \mathbf{n}) \\ & + \sum_{\ell \in L(g)} \theta_\ell \sum_{a \in A_\ell} \Phi^{(\ell)}_{a, g[\ell]} \hat{\pi}_{\text{PS}}(\mathbf{c} - \mathbf{e}_g + \mathbf{e}_{\mathcal{M}^a_\ell(g)} | \mathbf{n}) \\ & + \sum_{b \in B(g)} \rho_b \hat{\pi}_{\text{PS}}(\mathbf{c} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}^-_b(h)} + \mathbf{e}_{\mathcal{R}^+_b(h))} | \mathbf{n}) \Bigg\},\end{aligned} \tag{3.1}$$

where $\mathcal{N} = \sum_{g\in\mathcal{G}} c_g\big(c + n - 1 + \sum_{\ell\in L(g)} \theta_\ell + \sum_{b\in B(g)} \rho_b\big)$. In this general setting, there is no known closed-form solution for this recursion. The procedure for exact computation of $\hat{\pi}_{\text{PS}}(\mathbf{c}|\mathbf{n})$ is therefore repeated application of the recursion (3.1), which yields a set of coupled linear equations. As described in Section 2.1.2, each variable in the resulting set of equations has form $\hat{\pi}_{\text{PS}}(\mathbf{c}'|\mathbf{n})$, and letting $L(\mathbf{c}')$ be the total number of specified loci in $\mathbf{c}'$, $L(\mathbf{c}') \leq L(\mathbf{c})$. As a result, the set of coupled linear equations is finite and can be numerically solved. We have generally found that iterative procedures, such as the Gauss-Seidel method, perform well.

Regardless of the specific numerical technique used, computational complexity is lower-bounded by the number of coupled equations in the system. Even in the case of a single conditionally sampled haplotype, the following proposition assures us that there is a very large number of such equations.

Theorem 3.1. *Let $\eta \in \mathcal{H}$ and $\mathbf{c} = \mathbf{e}_\eta$ and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ with $|\mathbf{n}| = n$. Suppose that the number of alleles at each locus $\ell \in L$ is given by $|A_\ell| = s$. Then for $|L| = k$ loci, the number of equations $Q(k, s)$ generated by repeated application of (3.1) is given by*

$$Q(k,s) = \sum_{j=0}^{k} \binom{k}{j} B_j s^j \geq B_{k+1}, \tag{3.2}$$

where B_j is the j-th Bell number (Sloane, 1998). The second inequality is strict for $s > 1$.

Proof. Each variable present in the set of equations is of the form $\hat{\pi}_{\text{PS}}(\mathbf{c}'|\mathbf{n})$, where $\mathbf{c}'$ has $L(\mathbf{c}')$ specified loci, and $0 \leq L(\mathbf{c}') \leq k$. For a given value $|L(\mathbf{c}')| = j$, there are $\binom{k}{j}$ unique sets of specified loci, and each of the j specified loci can have any of the $|A_\ell| = s$ alleles. Finally, the specified loci can be partitioned into j arbitrary haplotypes, and the number of such partitions is given by the j-th Bell number. These considerations yield the first equality. The inequality follows from the recursive identity on Bell numbers, $B_{k+1} = \sum_{j=0}^{k} \binom{k}{j} B_j$, and is therefore strict when $s > 1$. □

If we further assume a PIM model, the mutation term in CSP recursion (2.20) is simplified, and the following corollary holds,

Corollary 3.2. *In the same setting as Proposition 3.1, and given a PIM model, the number of equations $Q_{PIM}(k)$ generated by repeated application of (2.20) is given by*

$$Q_{PIM}(k) = \sum_{j=0}^{k} \binom{k}{j} B_j = B_{k+1}. \tag{3.3}$$

Proof. In contrast to the general finite-alleles case given above, each locus can have only the allele specified in haplotype η, as mutation produces an unspecified allele. Thus, the combinatorial factor associated with per-locus polymorphism is removed from (3.2), resulting in the first equality. The second equality is by the same recursive identity on the Bell numbers. □

Because the Bell numbers $\{B_j\}$ grow super-exponentially with j, the number of variables in the system of linear equations also grow super-exponentially, even for a PIM model. Thus, direct computation of the CSP by generation and solution of the system of equations is computationally practicable for only small numbers (less than $k \approx 10$) of loci. Note that we have only counted the number of variables in the system of linear equations. In practice, this serves as a lower bound for the computational complexity of generating and solving the equations; moreover, such solutions are prone to numerical instability due to the very small probabilities involved. We next consider two additional approximations that provide some level of algorithmic scalability.

3.1.1 Limiting coalescence

Recall from Section 2.2.2, that by appropriately limiting coalescence events, the recursive expressions for computing the CSP can be simplified. For a single haplotype $\eta \in \mathcal{G}$, the CSDs associated with limiting and disallowing coalescence, denoted by $\hat{\pi}_{\text{LC}}$ and $\hat{\pi}_{\text{NC}}$ respectively, coincide. Letting $\eta \in \mathcal{G}$ and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ with $|\mathbf{n}| = n$, and assuming a PIM model, (2.59) yields the following recursion for the CSP $\hat{\pi}_{\text{LC}}(\mathbf{e}_\eta|\mathbf{n}) = \hat{\pi}_{\text{NC}}(\mathbf{e}_\eta|\mathbf{n})$,

$$\begin{aligned}\hat{\pi}_{\text{NC}}(\mathbf{e}_\eta|\mathbf{n}) = \frac{1}{\mathcal{N}}\Bigg\{ &\sum_{h\in\mathcal{H}:h\curlywedge\eta} n_h + \sum_{\ell\in L(\eta)} \theta_\ell \Phi^{(\ell)}_{\eta[\ell]} \hat{\pi}_{\text{NC}}(\mathbf{e}_{\mathcal{M}_\ell(\eta)}|\mathbf{n}) \\ &+ \sum_{b\in B(\eta)} \rho_b \hat{\pi}_{\text{NC}}(\mathbf{e}_{\mathcal{R}_b^-(\eta)}|\mathbf{n})\hat{\pi}_{\text{NC}}(\mathbf{e}_{\mathcal{R}_b^+(\eta))}|\mathbf{n})\Bigg\},\end{aligned} \tag{3.4}$$

where $\mathcal{N} = n + \sum_{\ell\in L(\eta)} \theta_\ell + \sum_{b\in B(\eta)} \rho_b$. As described in Section 2.2.2, the recursion (3.4) is proper, and so the CSP $\hat{\pi}_{\text{NC}}(\mathbf{e}_\eta|\mathbf{n})$ can be evaluated using dynamic programming or memoization, rather than constructing and numerically or algebraically solving a system of coupled linear equations. We can determine the computational complexity of such a solution by counting the number of states that must be enumerated, and considering the associated complexity of computing each such value,

Theorem 3.3. *Let $\eta \in \mathcal{H}$ and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ with $|\mathbf{n}| = n$. Then for $|L| = k$ loci, and assuming a PIM model, the number of states $Q_{\text{NC-PIM}}(k)$ that must be enumerated in a simple dynamic programming solution of (3.4) is given by*

$$Q_{\text{NC-PIM}}(k) = 2^k, \tag{3.5}$$

and the asymptotic time complexity of the associated dynamic program is given by $O(nk \cdot 2^k)$.

Proof. Each variable enumerated has form $\hat{\pi}_{\text{NC}}(\mathbf{e}_{\eta'}|\mathbf{n})$ for some $\eta' \in \mathcal{G}$. Considering only whether the allele at each locus within η' is specified or unspecified, there are 2^k such haplotypes. Moreover, because we have assumed a PIM model, each locus $\ell \in L$ with specified allele must have the allele $\eta[\ell]$, as mutation yields an unspecified allele. Thus, the number of states is given by (3.5).

The time complexity of evaluating $\hat{\pi}_{\text{NC}}(\mathbf{e}_{\eta'}|\mathbf{n})$ within the dynamic program, assuming that the $\hat{\pi}_{\text{NC}}$ terms on the right-hand-side have been evaluated, is dominated by the first term, which can be trivially evaluated with asymptotic time complexity $O(nk)$. The remaining two terms can then be evaluated with time complexity $O(k) \subset O(nk)$, providing the second result. □

Though the number of enumerated states, and therefore the computational complexity, is still exponential in the number of loci k, this represents a substantial improvement over evaluating the recursion for $\hat{\pi}_{\text{PS}}$, which requires constructing and solving a coupled system of linear equations with the number of equations super-exponential in k. In practice, however, it is still only possible to extend this solution to $k \approx 20$.

3.1.2 Limiting mutations

We next examine the form of the recursion (3.4) associated with $\hat{\pi}_{\text{NC}}$, with the objective of finding a sensible polynomial-time approximation. Observe that it is necessary to consider a state for each mutational configuration of the k loci; as described for a PIM model, there are 2^k such

configurations, accounting for the exponential computational complexity obtained above. This remains true even in the absence of recombination (when $\rho_b = 0$, for all $b \in B$), indicating the complexity is primarily due to the mutation process. Though it is unreasonable to entirely disallow mutation, as we have done for coalescence, it is possible to artificially limit the number of mutational configurations that are explicitly considered.

Formally, let $\eta \in \mathcal{G}$ and $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$ with $|\mathbf{n}| = n$. Let $\hat{\pi}_{\text{Alt}}$ be an arbitrary alternative CSD, and denote by $\hat{\pi}_{\text{NC-A}}$ the CSD with associated CSP recursion,

$$\begin{aligned}\hat{\pi}_{\text{NC-A}}(\mathbf{e}_\eta|\mathbf{n}) = \frac{1}{\mathcal{N}}\Bigg\{\Bigg(\sum_{h\in\mathcal{H}:h\lambda\eta} n_h\Bigg) + \sum_{\ell\in L(\eta)}\theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]}\hat{\pi}_{\text{Alt}}(\mathbf{e}_{\mathcal{M}^\eta_\ell(a)}|\mathbf{n}) \\ + \sum_{b\in B(\eta)} \rho_b \hat{\pi}_{\text{NC-A}}(\mathbf{e}_{\mathcal{R}^+_b(\eta)})\hat{\pi}_{\text{NC-A}}(\mathbf{e}_{\mathcal{R}^+_b(\eta)}|\mathbf{n})\Bigg\},\end{aligned} \tag{3.6}$$

where $\mathcal{N} = n + \sum_{\ell\in L(\eta)}\theta_\ell + \sum_{b\in B(\eta)}\rho_b$. This is precisely the recursion (3.4) for $\hat{\pi}_{\text{NC}}$, limited to states that have not yet mutated. Genealogically, this corresponds to applying the process associated with $\hat{\pi}_{\text{NC}}$ to lineages that have not mutated; if a mutation does occur on a lineage, the process associated with $\hat{\pi}_{\text{Alt}}$ is applied, backward in time, thereafter. Observe that in the limit $\sum_{\ell\in L}\theta_\ell \to 0$, regardless of the alternative CSD $\hat{\pi}_{\text{Alt}}$ used, $\hat{\pi}_{\text{NC-A}}(\mathbf{c}|\mathbf{n}) \to \hat{\pi}_{\text{NC}}(\mathbf{c}|\mathbf{n})$.

By choosing $\hat{\pi}_{\text{Alt}}$ to be a CSD for which the CSP can be evaluated efficiently, the resulting approximate CSP $\hat{\pi}_{\text{NC-A}}(\mathbf{e}_\eta|\mathbf{n})$ can be evaluated efficiently. Assuming $|A_\ell| = s$ for all $\ell \in L$ and that the number of loci is $|L| = k$, then $O(s \cdot k^3)$ CSPs associated with $\hat{\pi}_{\text{Alt}}$ must be evaluated. Empirically, we have found that good results are obtained by setting $\hat{\pi}_{\text{Alt}} = \hat{\pi}_{\text{SMC}(\mathcal{P})}$, where $\hat{\pi}_{\text{SMC}(\mathcal{P})}$ is the discretized version of $\hat{\pi}_{\text{SMC}}$ to be discussed in Section 3.2. As we shall demonstrate, the computational complexity of evaluating the CSP $\hat{\pi}_{\text{NC-A}}(\mathbf{e}_\eta|\mathbf{n})$, setting $\hat{\pi}_{\text{Alt}} = \hat{\pi}_{\text{SMC}(\mathcal{P})}$, is *polynomial* in the number of loci, a dramatic improvement over the previously described exponential-complexity algorithms. Nonetheless, the technique can only practically be extended to $k \approx 500$, impeding application to modern genomic data.

3.2 Computing $\hat{\pi}_{\text{SMC}}$

We have demonstrated in the previous section that computing the CSP associated with the CSD $\hat{\pi}_{\text{PS}}$ is computationally challenging. Though some progress was made by considering genealogical approximations, such as limiting coalescence, application to genomic-scale datasets remains impracticable, even when conditionally sampling a single haplotype. In this section, we consider the sequentially Markov CSD $\hat{\pi}_{\text{SMC}}$ discussed in Section 2.3, and describe an algorithm for the evaluating the associated CSP that is *linear* in the number of loci. Provided our earlier observation that $\hat{\pi}_{\text{SMC}}$ is equivalent to $\hat{\pi}_{\text{LC}}$, this result is remarkable.

Recall that the CSD $\hat{\pi}_{\text{SMC}}$ is naturally cast as an HMM, where the hidden state at each locus $\ell \in L$ represented by the marginal conditional genealogy (MCG), denoted $s_\ell \in \mathcal{S}$, and the corresponding observed state is the collection of alleles at the locus ℓ of conditionally sampled haplotypes. Because the state space $\mathcal{S}$ of MCGs is continuous, however, the dynamic programming algorithms associated with the classical HMM forward and backward recursions are not applicable. However, by discretizing the continuous component of $\mathcal{S}$, we are once again able to obtain a dynamic programming algorithm, resulting in an approximate algorithm for computing the CSP associated with $\hat{\pi}_{\text{SMC}}$ that is linear in the number of loci.

3.2.1 Single-deme, one-haplotype

Let $\eta \in \mathcal{H}$ and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ with $|\mathbf{n}| = n$, and consider computing the CSP $\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n})$. Recall from Section 2.3.2 that in the single-deme setting for a single conditionally sampled haplotype, the MCG at locus $\ell \in L$ is given by $s_\ell = (t_\ell, h_\ell) \in \mathcal{S} = \mathbb{R}_{\geq 0} \times \mathcal{H}$, where t_ℓ is the absorption time and h_ℓ is haplotype associated with the absorption lineage. The initial, transition, and emission densities are given by (2.73), (2.74), and (2.75), respectively.

Transforming time

Recall that marginal absorption time T_ℓ at each locus $\ell \in L$ is exponentially distributed with parameter n. In order to use the same discretization for all n, we follow Stephens and Donnelly (2000) and Fearnhead and Donnelly (2001), and transform the absorption time to a more natural scale in which the marginal absorption time is independent of n. Define the transformed MCG at locus $\ell \in L$ by $\tilde{s}_\ell = (\tilde{t}_\ell, h_\ell)$ where $\tilde{t}_\ell = nt_\ell$. Applying this transformation to the initial, transition, and emission densities yields the following transformed densities,

$$\zeta(\tilde{s}_\ell) = \frac{n_{h_\ell}}{n} e^{-\tilde{t}_\ell}, \tag{3.7}$$

$$\phi_b(\tilde{s}_\ell|\tilde{s}_{\ell-1}) = e^{-\frac{\rho_b}{n}\tilde{t}_{\ell-1}}\delta_{\tilde{s}_{\ell-1},\tilde{s}_\ell} + \frac{n_{h_\ell}}{n}\int_0^{\tilde{t}_{\ell-1}\wedge\tilde{t}_\ell} \frac{\rho_b}{n} e^{-\frac{\rho_b}{n}t_{\text{r}}} e^{-(\tilde{t}_\ell - t_{\text{r}})} dt_{\text{r}}, \tag{3.8}$$

and

$$\xi_\ell(h[\ell]|\tilde{s}_\ell) = e^{-\frac{\theta_\ell}{n}\tilde{t}_\ell} \sum_{k=0}^{\infty} \frac{(\frac{\theta_\ell}{n}\tilde{t}_\ell)^k}{k!} \big(\Phi^{(\ell)}\big)^k_{h_\ell[\ell],h[\ell]}. \tag{3.9}$$

As desired, using this time-rescaled model, the marginal absorption time at each locus is exponentially distributed with parameter 1. Because this distribution is independent of n and the coalescent model parameters $\{\rho_\ell\}$ and $\{\theta_\ell\}$, we expect that a single discretization of the transformed absorption time is appropriate for a wide range of haplotype configurations and parameter values. Using these time-transformed states, we thus re-write the CSP $\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n})$

$$\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n}) = \int_{\mathcal{S}} f_k(\tilde{s}_k) d\tilde{s}_k, \tag{3.10}$$

where the density $f_\ell(\cdot)$ is given by

$$f_\ell(\tilde{s}_\ell) = \xi_\ell(\eta[\ell]|\tilde{s}_\ell) \cdot \int_{\mathcal{S}} \phi_{(\ell-1,\ell)}(\tilde{s}_\ell|\tilde{s}_{\ell-1}) f_{\ell-1}(\tilde{s}_{\ell-1})\, d\tilde{s}_{\ell-1}, \tag{3.11}$$

for $1 < \ell \leq k$, with base case

$$f_1(\tilde{s}_1) = \xi_\ell(\eta[1]|\tilde{s}_1) \cdot \zeta(\tilde{s}_1). \tag{3.12}$$

Recall that, as described in Section 2.3.1, the forward densities $f_\ell(\cdot)$, and also the initial, transition, and emission densities, generally depend on both the conditionally sampled haplotype η and the previously sampled configuration $\mathbf{n}$ (and also the model parameters). In order to simplify notation, we have suppressed this dependence.

Discretizing time

Our next objective is to discretize the absorption time $\tilde{t} \in \mathbb{R}_{\geq 0}$. Let $0 = \tau_0 < \tau_1 < \cdots < \tau_m = \infty$ be a finite strictly increasing sequence in $\mathbb{R}_{\geq 0} \cup \{\infty\}$ so that $\mathcal{P} = \{[\tau_{j-1}, \tau_j)\}_{j=1,\ldots,m}$ is a finite partition of $\mathbb{R}_{\geq 0}$, into $|\mathcal{P}| = m$ intervals. This general partition will serve as the requisite discretization for absorption time; later in this section we provide some guidance on specific choices for the partition $\mathcal{P}$. The discretized space of MCGs is denoted by $\ddot{\mathcal{S}} = \mathcal{P} \times \mathcal{H}$, and the discretized MCG at locus $\ell \in L$ is denoted by $\ddot{s}_\ell = (p_\ell, h_\ell) \in \ddot{\mathcal{S}}$, where $p_\ell \in \mathcal{P}$ is the time interval in which absorption occurs, and $h_\ell \in \mathcal{H}$ is the absorption haplotype.

Towards formulating a $\mathcal{P}$-discretized version of the dynamics exhibited by the transformed HMM, we define the following $\mathcal{P}$-discretized version of the marginal, transition, and emission densities; overloading our present notation, we denote these densities by $\zeta(\ddot{s}_\ell)$, $\phi_b(\ddot{s}_\ell|\ddot{s}_{\ell-1})$, and $\xi_\ell(a|\ddot{s}_\ell)$, respectively. The discretized marginal density is obtained by integrating the transformed marginal density over the unknown transformed absorption time $\tilde{T}_\ell \in p_\ell$,

$$\zeta(\ddot{s}_\ell) = \int_{p_\ell} \zeta(\tilde{t}_\ell, h_\ell) d\tilde{t}_\ell = \frac{n_{h_\ell}}{n} \cdot x(p_\ell), \tag{3.13}$$

where $x(p) = \int_p e^{-\tilde{t}} d\tilde{t}$. The discretized transition density is similarly obtained by integrating the transformed transition density over the unknown absorption time $\tilde{T}_\ell \in p_\ell$, and partitioning with respect to the unknown absorption time $\tilde{T}_{\ell-1} \in p_{\ell-1}$. The latter is necessary because the discretized transition density is formally conditioned on the event $\{\tilde{T}_{\ell-1} \in p_{\ell-1}\}$ rather than $\{\tilde{T}_{\ell-1} = \tilde{t}_{\ell-1}\}$. Thus, making use of the $p_{\ell-1}$-truncated marginal distribution on the MCG at locus $\ell - 1$,

$$\begin{aligned} \phi_b(\ddot{s}_\ell|\ddot{s}_{\ell-1}) &= \frac{1}{\zeta(\ddot{s}_{\ell-1})} \int_{p_\ell} \int_{p_{\ell-1}} \phi_b(\tilde{t}_\ell, h_\ell|\tilde{t}_{\ell-1}, h_{\ell-1}) \zeta(\tilde{t}_{\ell-1}, h_{\ell-1}) d\tilde{t}_{\ell-1} d\tilde{t}_\ell \\ &= y_b(p_{\ell-1}) \cdot \delta_{\ddot{s}_{\ell-1}, \ddot{s}_\ell} + z_b(p_\ell|p_{\ell-1}) \cdot \frac{n_{h_\ell}}{n}, \end{aligned} \tag{3.14}$$

with analytic expressions for $y_b(\cdot)$ and $z_b(\cdot|\cdot)$ provided in Appendix C.1. Finally, the discretized emission density is obtained by integrating the transformed emission density over the unknown transformed absorption time $\tilde{T}_\ell \in p_\ell$, which is necessary because the discretized emission density is formally conditioned on the event $\{\tilde{T}_\ell \in p_\ell\}$ rather than $\{\tilde{T}_\ell = \tilde{t}_\ell\}$. As before, making use of the $p_{\ell-1}$-truncated marginal distribution on the MCG at locus $\ell - 1$,

$$\xi_\ell(a|\ddot{s}_\ell) = \frac{1}{\zeta(\ddot{s}_{\ell-1})} \int_{p_\ell} \xi_\ell(a|\tilde{t}_\ell, h_\ell) \zeta(\tilde{t}_{\ell-1}, h_{\ell-1}) d\tilde{t}_\ell = \sum_{k=0}^{\infty} v_\ell^{(k)}(p_\ell) \cdot \frac{(\theta_\ell/n)^k}{k!} \big(\Phi^{(\ell)}\big)^k_{h_\ell[\ell], a}, \tag{3.15}$$

with an analytic expression for $v_\ell^{(k)}(\cdot)$ provided in Appendix C.1. Note that we have not introduced any additional approximation in computing the discretized marginal, transition, and emission densities; the computation of these densities follows from elementary probability theory.

We next wish to write the key HMM forward recursion for the discretized space of MCGs. We thus define the discretized forward density $f_\ell(\ddot{s}_\ell)$:

$$f_\ell(\ddot{s}_\ell) = \int_{p_\ell} f_\ell(\tilde{t}_\ell, h_\ell) d\tilde{t}_\ell. \tag{3.16}$$

Unfortunately, we cannot directly obtain a recursion for the discretized forward density $f_\ell(\ddot{s}_\ell)$ via the recursion (3.11) for the transformed forward density $f_\ell(\tilde{t}_\ell, h_\ell)$. We therefore make an additional

approximation, that the transformed *transition* and *emission* densities, $\phi_b(\cdot|\tilde{t}, h)$ and $\xi_\ell(\cdot|\tilde{t}, h)$, respectively, depend on the interval $p \in \mathcal{P}$, such that $\tilde{t} \in p$, but not on the actual transformed time $\tilde{t}$. Formally, letting $p \in \mathcal{P}$, then for all $\tilde{t} \in p$, we approximate

$$\phi_b(\cdot|\tilde{t}, h) \approx \phi_b(\cdot|p, h), \text{ and} \tag{3.17}$$

$$\xi_\ell(\cdot|\tilde{t}, h) \approx \xi_\ell(\cdot|p, h). \tag{3.18}$$

Observe that, under the assumption of well-behaved transition and emission densities, these approximations can be made arbitrarily accurate by using increasingly refined partitions $\mathcal{P}$ of $\mathbb{R}_{\geq 0}$. Thus, using the recursive definition (3.11) of the transformed density $f_\ell(\tilde{t}_\ell, h_\ell)$, and applying the approximations, (3.17) and (3.18),

$$\begin{aligned} f_\ell(\ddot{s}_\ell) &= \int_{p_\ell} f_\ell(\tilde{t}_\ell, h_\ell) d\tilde{t}_\ell \\ &= \int_{p_\ell} \xi_\ell(\eta[\ell]|\tilde{t}_\ell, h_\ell) \cdot \int_{\mathcal{S}} \phi_{(\ell-1,\ell)}(\tilde{t}_\ell, h_\ell|\tilde{t}_{\ell-1}, h_{\ell-1}) f_{\ell-1}(\tilde{t}_{\ell-1}, h_{\ell-1})\, d\tilde{s}_{\ell-1} d\tilde{t}_\ell \\ &\approx \xi_\ell(\eta[\ell]|\ddot{s}_\ell) \cdot \sum_{\ddot{s}_{\ell-1} \in \ddot{\mathcal{S}}} \phi_{(\ell-1,\ell)}(\ddot{s}_\ell|\ddot{s}_{\ell-1}) f_{\ell-1}(\ddot{s}_{\ell-1}) \end{aligned} \tag{3.19}$$

With the support of this approximate discretized forward recursion, we can thus write

$$\hat{\pi}_{\text{SMC}(\mathcal{P})}(\mathbf{e}_\eta|\mathbf{n}) = \sum_{\ddot{s}_k \in \ddot{\mathcal{S}}} F_k(\ddot{s}_k) \approx \sum_{\ddot{s}_k \in \ddot{\mathcal{S}}} f_k(\ddot{s}_k) = \hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n}), \tag{3.20}$$

where the discretized forward density is defined

$$F_\ell(\ddot{s}_\ell) = \xi_\ell(\eta[\ell]|\ddot{s}_\ell) \cdot \sum_{\ddot{s}_{\ell-1} \in \ddot{\mathcal{S}}} \phi_{(\ell-1,\ell)}(\ddot{s}_\ell|\ddot{s}_{\ell-1}) F_{\ell-1}(\ddot{s}_{\ell-1}), \tag{3.21}$$

with base case

$$F_1(\ddot{s}_1) = \xi_1(\eta[1]|\ddot{s}_1) \cdot \zeta(\ddot{s}_1). \tag{3.22}$$

In summary, equations (3.20), (3.21), and (3.22) provide the requisite $\mathcal{P}$-discretized recursions necessary to use the classical forward algorithm for HMMs. Observe that the fact that the Markov property holds on the discretized state space $\ddot{\mathcal{S}} = \mathcal{P} \times \mathcal{H}$ follows from the assumptions (3.17) and (3.18) (Rosenblatt, 1959). In fact, the relevant discretized forward recursions may alternatively be obtained by *assuming* that the Markov property holds on $\ddot{\mathcal{S}}$ and writing down the relevant transition and emission probabilities with the interpretations given above. In the remainder of this section, we examine some general properties of the discretized dynamics, and also provide one method for choosing a discretization $\mathcal{P}$. The computational complexity of evaluating $\hat{\pi}_{\text{SMC}(\mathcal{P})}(\mathbf{e}_\eta|\mathbf{n})$ is examined in Section 3.3.

Properties of the discretization

Recall the *detailed-balance condition* (2.77) associated with the marginal and transition densities for $\hat{\pi}_{\text{SMC}}$. Using expressions for the discretized marginal and transition densities, (3.13) and (3.14), along with the non-discretized detailed balance condition (2.77), it is possible to verify that

$$\phi_b(\ddot{s}_\ell|\ddot{s}_{\ell-1})\zeta(\ddot{s}_{\ell-1}) = \phi_b(\ddot{s}_{\ell-1}|\ddot{s}_\ell)\zeta(\ddot{s}_\ell). \tag{3.23}$$

Thus, the discretized marginal and transition densities satisfy an analogous detailed balance condition. As discussed in Section 2.3.2, the stated Markov process is therefore reversible, and the discretized marginal distribution is stationary under the given transition dynamics. Because we start the Markov process using the discretized marginal distribution, this property ensures that the CSP computation for $\hat{\pi}_{\text{SMC}(\mathcal{P})}$ will yield the same result regardless of whether we proceed from left to right, as in (3.20), or from right to left, for any discretization $\mathcal{P}$.

Furthermore, recall the *locus-skipping property* (2.79) associated with $\hat{\pi}_{\text{SMC}}$. Using the expression for the discretized transition density (3.14) along with the non-discretized locus-skipping property (2.79), it is possible to show that an analogous property approximately holds for $\hat{\pi}_{\text{SMC}(\mathcal{P})}$,

$$\sum_{\ddot{s}_\ell \in \ddot{\mathcal{S}}} \phi_{(\ell-1,\ell)}(\ddot{s}_\ell | \ddot{s}_{\ell-1}) \cdot \phi_{(\ell,\ell+1)}(\ddot{s}_{\ell+1} | \ddot{s}_\ell) \approx \phi_{(\ell-1,\ell+1)}(\ddot{s}_{\ell+1} | \ddot{s}_{\ell-1}), \tag{3.24}$$

where the non-equality is a direct consequence of (3.17). As indicated in Section 2.3.2, this approximation is useful in scenarios when data is missing (i.e. $\eta[\ell]$ is unknown for one or more $\ell \in L$), as it reduces the computational complexity of the dynamic program. Again, this approximation holds for any discretization $\mathcal{P}$, and the approximation error will decrease for more refined partitions.

Discretization choice

Finally, we discuss a method for choosing a discretization $\mathcal{P}$ of the absorption time. Recalling that the marginal transformed absorption time is exponentially distributed with parameter 1, let $\{(w^{(j)}, t^{(j)})\}_{j=1,\dots,m}$ be the m-point Gaussian quadrature associated with the function $f(t) = e^{-t}$ (Abramowitz and Stegun, 1972, Section 25.4.45). Set $\tau_0 = 0$, and for each value $j = 1, \dots, m$, set τ_j such that

$$\int_{\tau_{j-1}}^{\tau_j} e^{-t} dt = w^{(j)}. \tag{3.25}$$

Note that $\sum_{j=1}^m w^{(j)} = 1$, and therefore $\tau_m = \infty$, and the points $0 = \tau_0 < \cdots < \tau_m = \infty$ determine a partition $\mathcal{P} = \{[\tau_{j-1}, \tau_j)\}_{j=1,\dots,m}$ of $\mathbb{R}_{\geq 0}$. This partition may then be used to compute $\hat{\pi}_{\text{SMC}(\mathcal{P})}$; we shall henceforth write $\hat{\pi}_{\text{SMC}(d)}$ for the d-point Gaussian quadrature-discretized version of $\hat{\pi}_{\text{SMC}}$.

The use of Gaussian quadrature evokes the work of Stephens and Donnelly (2000) and Fearnhead and Donnelly (2001). Although the method we employ is related, it is different in that we do not use the quadrature directly (for example, the values of the quadrature points $\{t^{(j)}\}$ are never used explicitly); rather we use the Gaussian quadrature as a reasonable way of choosing a partition $\mathcal{P}$. We briefly note that we experimented with other methods of discretization, including using the Gaussian quadrature points and weights as in Stephens and Donnelly (2000), but these techniques failed to satisfy the detailed-balance condition, and did not produce superior results.

3.2.2 Multiple-deme, one-haplotype

Suppose $\mathcal{D}$ is a finite set of demes; let $\eta \in \mathcal{H}$, $\alpha \in \mathcal{D}$ and $\mathbf{n} = (n_{d,h})_{d \in \mathcal{D}, h \in \mathcal{H}}$ with $|\mathbf{n}| = n$, and consider computing the CSP $\hat{\pi}_{\text{SMC}}(\mathbf{e}_{\alpha,\eta} | \mathbf{n})$. Recall from Section 2.3.3 that the MCG at locus $\ell \in L$ is given by $s_\ell \in \mathcal{S} = \mathbb{R}_{\geq 0} \times \mathcal{H} \times \mathcal{Q}$, where $\mathcal{Q}$ is the space of full migrational histories. Though this is the correct space of MCGs to consider for the sequentially Markov CSD $\hat{\pi}_{\text{SMC}}$, the infinite dimensionality of the space $\mathcal{Q}$ presents practical computational difficulties.

For this reason, we proposed the CSD $\hat{\pi}_{\text{SMC-ADO}}$. At locus $\ell \in L$, the MCG associated with $\hat{\pi}_{\text{SMC-ADO}}$ is given by $\hat{s}_\ell = (t_\ell, h_\ell, d_\ell) \in \hat{\mathcal{S}}$, where $\hat{\mathcal{S}} = \mathbb{R}_{\geq 0} \times \mathcal{H} \times \mathcal{D}$; the values t_ℓ and h_ℓ are

the absorption time and haplotype, and d_ℓ is the absorption deme. As described in Section 2.3.3, restricting the state space to the absorption deme instead of the full migrational history introduces a non-Markov dependence into the sequence of MCGs; it is nonetheless possible to approximate the sequence by a Markov process, with marginal and transition densities given by (2.88), (2.90), and common emission density (2.75).

Because the space of MCGs $\hat{\mathcal{S}}$ associated with the CSD $\hat{\pi}_{\text{SMC-ADO}}$ is of finite-dimension, we proceed with developing a practicable algorithm for approximating the CSP $\hat{\pi}_{\text{SMC-ADO}}(\mathbf{e}_{\alpha,\eta}|\mathbf{n})$. As in the previous section for a single-deme, it remains necessary to discretize the continuous component of the state space $\hat{\mathcal{S}}$ associated with absorption time; however, unlike the single-deme setting, the marginal density (2.88) does not have a natural time re-scaling, such that the transformed density does not depend on $\mathbf{n}$, and so we do not attempt to re-scale time. We note at the outset, however, that this implies that the eventual choice of discretization must be sensitive both to the configuration $\mathbf{n}$ and to the parameters associated with the migration model.

Discretizing time

As previously, let $0 = \tau_0 < \tau_1 < \cdots < \tau_m = \infty$ be a finite strictly increasing sequence in $\mathbb{R}_{\geq 0} \cup \{\infty\}$ so that $\mathcal{P} = \{[\tau_{j-1}, \tau_j)\}_{j=1,\ldots,m}$ is a partition of $\mathbb{R}_{\geq 0}$ into $|\mathcal{P}| = m$ intervals. The discretized space of MCGs is denoted by $\ddot{\mathcal{S}} = \mathcal{P} \times \mathcal{H} \times \mathcal{D}$ and the MCG at locus $\ell \in L$ is denoted by $\ddot{s}_\ell = (p_\ell, h_\ell, d_\ell) \in \ddot{\mathcal{S}}$, where $p_\ell \in \mathcal{P}$ is the time interval in which absorption occurs, and $h_\ell \in \mathcal{H}$ and $d_\ell \in \mathcal{D}$ are the absorption haplotype and deme, respectively.

The $\mathcal{P}$-discretized marginal, transition, and emission densities are computed using the same basic probability theory described in Section 3.2.1. In particular, for the discretized marginal density, we obtain

$$\zeta(\ddot{s}_\ell) = \int_{p_\ell} \zeta(t_\ell, h_\ell, d_\ell) dt_\ell = x(p_\ell, d_\ell) \cdot \frac{n_{d_\ell, h_\ell}}{n_{d_\ell}}, \tag{3.26}$$

where, recalling that the matrix Z governs the absorption process,

$$x(p, d) = \int_p \left(Z e^{Zt}\right)_{r_\alpha, a_d} dt. \tag{3.27}$$

Similarly, for the discretized transition density, we obtain

$$\begin{aligned} &\phi_b(\ddot{s}_\ell | \ddot{s}_{\ell-1}) \\ &\quad = \frac{1}{\zeta(\ddot{s}_{\ell-1})} \int_{p_\ell} \int_{p_{\ell-1}} \phi_b(t_\ell, h_\ell, d_\ell | t_{\ell-1}, h_{\ell-1}, d_{\ell-1}) \zeta(t_{\ell-1}, h_{\ell-1}, d_{\ell-1}) dt_{\ell-1} dt_\ell \\ &\quad = y_b(p_{\ell-1}, d_{\ell-1}) \delta_{\ddot{s}_{\ell-1}, \ddot{s}_\ell} + z_b(p_\ell, d_\ell | p_{\ell-1}, d_{\ell-1}) \cdot \frac{n_{d_\ell, h_\ell}}{n_{d_\ell}}, \end{aligned} \tag{3.28}$$

where explicit expressions of $y_b(\cdot)$ and $z_b(\cdot|\cdot)$ are provided in Appendix C.2. Finally, for the discretized emission density, we obtain

$$\xi_\ell(a | \ddot{s}_\ell) = \frac{1}{\zeta(\ddot{s}_\ell)} \int_{p_\ell} \xi_\ell(a | t_\ell, h_\ell) \zeta(t_\ell, h_\ell, d_\ell)) dt_\ell, \tag{3.29}$$

and we again provide a more explicit form of this quantity in Appendix C.2. Note that the emission probability (2.75) in the continuous case is only dependent on the time of absorption and the allele

$h_\ell[\ell]$ of the absorption haplotype. The discretized analog (3.29) on the other hand also depends on the deme that the absorption haplotype resides in. This is due to the fact that the latter emission probability is conditioned on absorption in a particular deme at any time in the discretized interval; because the rate of absorption depends on the deme, the distribution on absorption time, and hence the emission probability, must also depend on the deme.

As in Section 3.2.1, in order to write the HMM forward recursion for the discretized space of the MCGs, we make an additional approximation. Formally, letting $p \in \mathcal{P}$, then for all $t \in p$, we approximate

$$\phi_b(\cdot|t,h,d) \approx \phi_b(\cdot|p,h,d), \text{ and} \tag{3.30}$$

$$\xi_\ell(\cdot|t,h,d) \approx \xi_\ell(\cdot|p,h,d). \tag{3.31}$$

These approximations in conjunction with the discretized marginal, transition, and emission densities provided above yields a discretized forward recursion that approximates the CSP $\hat{\pi}_{\text{SMC-ADO}}(\mathbf{e}_{\alpha,\eta}|\mathbf{n})$. As before, approximating $F_\ell(\ddot{s}_\ell) \approx \int_{p_\ell} f_\ell(t_\ell, h_\ell, d_\ell)dt_\ell$, we obtain the discretized approximation

$$\hat{\pi}_{\text{SMC}(\mathcal{P})}(\mathbf{e}_{\alpha,\eta}|\mathbf{n}) = \sum_{\ddot{s}_k \in \ddot{\mathcal{S}}} F_k(\ddot{s}_\ell) \approx \int_{\hat{\mathcal{S}}} f_k(\hat{s}_k)d\hat{s}_k = \hat{\pi}_{\text{SMC-ADO}}(\mathbf{e}_{\alpha,\eta}|\mathbf{n}) \approx \hat{\pi}_{\text{SMC}}(\mathbf{e}_{\alpha,\eta}|\mathbf{n}), \tag{3.32}$$

where the first approximate equality is due to the discretization, (3.30) and (3.31), and the second approximate equality is due to the restriction of the full migrational history to the deme in which absorption occurs. The discretized forward density $F_k(\ddot{s}_\ell)$ is defined as in (3.21) and (3.22). As before, the $\mathcal{P}$-discretized recursions enable the classical forward algorithm for HMMs to be used to evaluate the CSP $\hat{\pi}_{\text{SMC}(\mathcal{P})}(\mathbf{e}_{\alpha,\eta}|\mathbf{n})$.

Discretization choice

Recall that in Section 3.2.1, for a single deme, the transformed absorption time is marginally distributed as an exponential random variable with rate parameter 1; it was therefore natural to use Gaussian quadrature associated with the function $f(t) = e^{-t}$ to obtain the discretization intervals. In the present setting, for a structured population including migration, there is no such natural time transformation or related evident choice for the discretization intervals. In practice, we have obtained reasonable and numerically stable results by using a logarithmic discretization. To produce a discretization $\mathcal{P}$ comprising $|\mathcal{P}| = m$ intervals, we set

$$\tau_j = -\frac{1}{r}\log\left(\frac{m-j}{m}\right), \tag{3.33}$$

where r is the harmonic mean of the absorption rates in each deme $r = \left(\prod_{d\in\mathcal{D}} \kappa_d^{-1} n_d\right)^{1/q}$. Observe that $0 = \tau_0 < \cdots < \tau_m = \infty$, and so the resulting discretization $\mathcal{P}$ is well-defined.

3.2.3 Backward algorithm and marginal decoding

In addition to the general HMM forward recursion described in Section 2.3.1, there exists a corresponding backward recursion (Cappé et al., 2005). Letting $\mathbf{c}$ and $\mathbf{n}$ be haplotype configurations, the CSP $\hat{\pi}_{\text{SMC}}(\mathbf{c}|\mathbf{n})$ can be expressed in terms of the backward recursion,

$$\hat{\pi}_{\text{SMC}}(\mathbf{c}|\mathbf{n}) = \int_{\mathcal{S}} \xi_\ell^{(\mathbf{n})}(\mathbf{c}[1]|s_1) \cdot e_k^{(\mathbf{c},\mathbf{n})}(s_1)\, ds_1, \tag{3.34}$$

where $e_\ell^{(\mathbf{c},\mathbf{n})}(\cdot)$ is defined (for $1 \leq \ell < k$) by

$$e_\ell^{(\mathbf{c},\mathbf{n})}(s_\ell) = \int_{\mathcal{S}} \xi_{\ell+1}^{(\mathbf{n})}(\mathbf{c}[\ell+1]|s_{\ell+1}) \cdot \phi_{(\ell,\ell+1)}^{(\mathbf{n})}(s_{\ell+1}|s_\ell) \cdot e_{\ell+1}^{(\mathbf{c},\mathbf{n})}(s_{\ell+1})\, ds_{\ell+1}, \tag{3.35}$$

with base case

$$e_k^{(\mathbf{c},\mathbf{n})}(s_k) = 1. \tag{3.36}$$

As for the forward recursion, the MCG state space $\mathcal{S}$ is continuous, and explicit evaluation of the integrals is not generally possible. Fortunately, the preceding work in Sections 3.2.1 and 3.2.2 on discretizing the MCG state space is directly applicable; recall that $\ddot{\mathcal{S}}$ is the discretized space of MCGs. For a single conditionally sampled haplotype $\eta \in \mathcal{H}$, then analogous to equations (3.20),(3.21) and (3.22), it is possible to compute $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta|\mathbf{n}) \approx \hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta|\mathbf{n})$,

$$\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta|\mathbf{n}) = \sum_{\ddot{s}_1 \in \ddot{\mathcal{S}}} \xi_1(\eta[1]|\ddot{s}_1) E_1(\ddot{s}_1), \tag{3.37}$$

where the discretized backward density is defined, for $1 \leq \ell < k$

$$E_\ell(\ddot{s}_\ell) = \sum_{\ddot{s}_{\ell+1} \in \ddot{\mathcal{S}}} \xi_{\ell+1}(\eta[\ell+1]|\ddot{s}_{\ell+1}) \phi_{(\ell,\ell+1)}(\ddot{s}_{\ell+1}|\ddot{s}_\ell) E_{\ell+1}(\ddot{s}_{\ell+1}), \tag{3.38}$$

with base case

$$E_k(\ddot{s}_k) = 1. \tag{3.39}$$

Much as the discretized forward recursion, the discretized backward recursion can be used, in conjunction with dynamic programming, to evaluate the CSP $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta|\mathbf{n})$ with computational complexity linear in the number of loci.

Finally, we consider *marginal decoding* on the discretized HMM associated with the CSD $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$. In this context, marginal decoding provides the posterior distribution for the random MCG $\ddot{S}_\ell$ at an arbitrary locus $\ell \in L$; as we discuss in Chapter 4, this distribution is useful in several applications of the CSD. General HMM methodology (Cappé et al., 2005) stipulates that the posterior probability $p_\ell(\ddot{s}_\ell|\mathbf{c},\mathbf{n})$ of the MCG $\ddot{s}_\ell \in \ddot{\mathcal{S}}$ is given by

$$p_\ell(\ddot{s}_\ell|\mathbf{c},\mathbf{n}) = \frac{F_\ell(\ddot{s}_\ell) E_\ell(\ddot{s}_\ell)}{\sum_{\ddot{s} \in \ddot{\mathcal{S}}} F_\ell(\ddot{s}) E_\ell(\ddot{s})}, \tag{3.40}$$

where $F_\ell(\ddot{s}_\ell)$ and $E_\ell(\ddot{s}_\ell)$ are the forward and backward values associated with the forward and backward recursions. Thus, by computing and caching the forward and backward values at each locus $\ell \in L$, and for each relevant $\ddot{s}_\ell \in \ddot{\mathcal{S}}$, marginal decoding at an arbitrary locus can be efficiently realized. In Section 3.3.3, we re-visit the problem of marginal decoding, and demonstrate that it is possible to substantially reduce the associated time and space complexity.

3.3 Computing $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$ efficiently

In the previous section, we described a $\mathcal{P}$-discretized approximation $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$ to the CSD $\hat{\pi}_{\mathrm{SMC}}$, and derived discretized versions of the marginal, transition, and emission densities. The CSP associated with the discretized approximation can be efficiently computed using the forward algorithm for

HMMs. The computational complexity of the algorithm is *linear* in the number of loci, representing a fundamental improvement over exact algorithms associated with $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{SMC}}$, for which the computational complexity of the best known exact algorithms are super-exponential and exponential in the number of loci, respectively. As we shall demonstrate, however, the constants associated with the forward algorithm remain large, thus making it difficult or impossible to directly apply the algorithm to genomic-scale data.

In this section, we examine the forward algorithm in detail, and propose two related optimizations that help to overcome this limitation. Consider sampling a large number of sequences from a population. If the sampled sequences are very long, it is likely that nearly all of them will be unique. However, for most relatively short regions, the number of unique subsequences will be reduced due to the effects of linkage disequilibrium, or alternatively, finite recombination rates between loci. This intuition forms the basis of the first optimization, which locally reduces the complexity of the forward algorithm, thereby improving efficiency. The collection of locally unique subsequences on which this optimization depends are formalized as a partition $\mathcal{C}$ of the sampled sequences; we characterize the optimal partition given the sampled sequences, and provide a fast algorithm for approximating this optimum.

A second common feature of the sampled sequences is an abundance of non-polymorphic sites. These sites are informative – for example, a local over-abundance of non-polymorphic sites indicates a recent common ancestor, which in turn indicates a low propensity for recombination – and should be included in the analysis. Leveraging the fact that non-polymorphic sites do not differentiate the sequences, we show that it is possible to reduce the complexity of the forward algorithm at non-polymorphic sites. Note that this is different from simply ignoring non-polymorphic sites; instead, we propose algorithmic improvements to efficiently incorporate non-polymorphic sites into the analysis.

In formally describing the optimizations, we restrict attention to the CSD $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$ in the multiple-locus and single-deme setting, described in Section 3.2.1. Recall that the discretized MCG at locus $\ell \in L$ is denoted by $\ddot{s}_\ell = (p_\ell, h_\ell) \in \ddot{\mathcal{S}} = \mathcal{P} \times \mathcal{H}$, where $\mathcal{P}$ is a partition of $\mathbb{R}_{\geq 0}$ into m intervals, and $p_\ell \in \mathcal{P}$ is the absorption interval of the MCG. Letting $\eta \in \mathcal{H}$ and $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$, the CSP $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta | \mathbf{n})$ can be expressed

$$\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta | \mathbf{n}) = \sum_{\ddot{s}_\ell \in \ddot{\mathcal{S}}_{\mathbf{n}}} F_k(\ddot{s}_k), \tag{3.41}$$

where the discretized forward density is defined, for $1 \leq \ell \leq k$,

$$F_\ell(\ddot{s}_\ell) = \xi_\ell(\eta[\ell] | \ddot{s}_\ell) \cdot \sum_{\ddot{s}_{\ell-1} \in \ddot{\mathcal{S}}_{\mathbf{n}}} \phi_{(\ell-1,\ell)}(\ddot{s}_\ell | \ddot{s}_{\ell-1}) F_{\ell-1}(\ddot{s}_{\ell-1}), \tag{3.42}$$

with base case

$$F_0(\ddot{s}_0) = \zeta(\ddot{s}_0), \tag{3.43}$$

where the $\mathcal{P}$-discretized marginal, transition, and emission densities are given by (3.13), (3.14), and (3.15), respectively. Observe that we have restricted the summations to the space $\ddot{\mathcal{S}}_{\mathbf{n}} = \mathcal{P} \times \mathcal{H}_{\mathbf{n}} \subset \ddot{\mathcal{S}}$, where $\mathcal{H}_{\mathbf{n}} = \{h \in \mathcal{H} : n_h > 0\}$ is the space of haplotypes with positive multiplicity in $\mathbf{n}$. It can be verified that $F_\ell(\cdot, h) = 0$ for all $h \in \mathcal{H}$ such that $n_h = 0$, and therefore this modification does not affect the computation. Hereafter, we write that $|\mathcal{H}_{\mathbf{n}}| = n_{\mathrm{u}}$, so that $|\ddot{\mathcal{S}}_{\mathbf{n}}| = m n_{\mathrm{u}}$. In order to regularize the forward recursion, we have also extended it to a fictitious 0-th locus; it can be verified

Algorithm 1 Compute $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta|\mathbf{n})$ using the ordinary forward algorithm

1: **for all** $\ddot{s}_0 \in \ddot{\mathcal{S}}_{\mathbf{n}}$ **do**
2: Compute $F_0(\ddot{s}_0)$ by (3.43)
3: **end for**
4: **for** $\ell = 1 \to k$ **do**
5: **for all** $\ddot{s}_\ell \in \ddot{\mathcal{S}}_{\mathbf{n}}$ **do**
6: Compute $F_\ell(\ddot{s}_\ell)$ using (3.42)
7: **end for**
8: **end for**
9: Compute $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta|\mathbf{n})$ using (3.41)

that, given an arbitrary value of $\rho_{(0,1)}$ to be used for computing the transition density $\phi_{(0,1)}(\ddot{s}_1|\ddot{s}_0)$, this modification does not affect the computation.

As a point of reference, we begin with the most basic forward algorithm (Cappé et al., 2005), provided in Algorithm 1. Within the critical loop, lines 4–8, all of the required quantities on the right hand side of the recursion (3.42) for $F_\ell(\ddot{s}_\ell)$ have been computed by the previous iteration, or in the initialization, lines 1–3; the forward algorithm is therefore a dynamic programming solution to the recursion for the forward variable $F_\ell(\cdot)$. The time complexity of line 6 is $O(mn_{\mathrm{u}})$, of lines 5–7 is $O((mn_{\mathrm{u}})^2)$, and of lines 4–8 is $O(k(mn_{\mathrm{u}})^2)$. The initialization, lines 1–3, and termination, line 9 both have time complexity $O(mn_{\mathrm{u}})$, and so the overall time complexity is $O(k(mn_{\mathrm{u}})^2)$.

In the remainder of this section, we demonstrate how the ideas above can be used to refine the discretized forward recursion, and in turn to construct more efficient dynamic programs. We present these refinements in the context of two sufficient conditions, and later revisit the sufficient conditions to show that the optimizations are applicable, either in whole or in part, to alternative CSDs, such as those of Fearnhead and Donnelly (2001) and Li and Stephens (2003), or to more complex demographic scenarios, such as structured populations with migration.

As a measure of real-world performance, asymptotic complexity analyses often leave much to be desired. Consider, for example, a sample in which 1 out of every 1000 sites is polymorphic. Letting $k = |L|$ be the total number of sites, and $k_{\mathrm{p}} \leq k$ the number of polymorphic sites, then formally $O(k) = O(k_{\mathrm{p}})$. Nevertheless, for the present purposes, we would like to distinguish between an algorithm that operates on each of the k sites and an algorithm that operates only on the k_{p} polymorphic sites, as the latter will be some 1000× faster; we thus write the complexities for the two algorithms as $O(k)$ and $O(k_{\mathrm{p}})$, respectively.

3.3.1 Improving efficiency via the transition distribution

Consider the marginal and transition distributions on MCGs, with densities $\zeta(\cdot)$ and $\phi_b(\cdot)$, defined in (3.13) and (3.14), respectively. In the marginal distribution, the absorption haplotype is independent of the absorption interval, and uniformly distributed; conditioned on recombination, the same is true for the transition distribution. We therefore, observe the following property,

Property 1. *The initial and transition densities, $\zeta(\cdot)$ and $\phi_b(\cdot)$, take the following functional form*

$$\zeta(\ddot{s}_\ell) = x(p_\ell) \cdot \frac{n_{h_\ell}}{n},$$
$$\phi_b(\ddot{s}_\ell|\ddot{s}_{\ell-1}) = y_b(p_{\ell-1}) \cdot \delta_{\ddot{s}_{\ell-1},\ddot{s}_\ell} + z_b(p_\ell|p_{\ell-1}) \cdot \frac{n_{h_\ell}}{n},$$

where $x(\cdot)$, $y_b(\cdot)$, $z_b(\cdot|\cdot)$ are known analytic expressions.

Algorithm 2 Compute $\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n})$ using a forward-type algorithm improved by considering Property 1

1: **for all** $p_0 \in \mathcal{P}$ **do**
2: Compute $F_0(p_0, h_0)$ by (3.48), $\forall h_0 \in \mathcal{H}_\mathbf{n}$
3: Compute $Q_0(p_0)$ using (3.44)
4: **end for**
5: **for** $\ell = 1 \to k$ **do**
6: **for all** $p_\ell \in \mathcal{P}$ **do**
7: Compute $U_{\ell-1}(p_\ell)$ using (3.45)
8: Compute $F_\ell(p_\ell, h_\ell)$ using (3.47), $\forall h_\ell \in \mathcal{H}_\mathbf{n}$
9: Compute $Q_\ell(p_\ell)$ using (3.44)
10: **end for**
11: **end for**
12: Compute $\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n})$ using (3.46)

For the CSD $\hat{\pi}_{\text{SMC}(\mathcal{P})}$, the analytic expressions for $x(\cdot)$, $y_b(\cdot)$, $z_b(\cdot|\cdot)$ are given in Appendix C.1. Using Property 1 in conjunction with definitions,

$$Q_\ell(p_\ell) = \sum_{h_\ell \in \mathcal{H}_\mathbf{n}} F_\ell(p_\ell, h_\ell), \text{and} \tag{3.44}$$

$$U_\ell(p_{\ell+1}) = \sum_{p_\ell \in \mathcal{P}} z_{(\ell,\ell+1)}(p_{\ell+1}|p_\ell) Q_\ell(p_\ell), \tag{3.45}$$

we can express (3.41), (3.42), and (3.43) as

$$\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n}) = \sum_{p_k \in \mathcal{P}} Q_k(p_k), \tag{3.46}$$

where, for $1 < \ell \leq k$,

$$F_\ell(\ddot{s}_\ell) = \xi_\ell(\eta[\ell]|\ddot{s}_\ell)\left[y_{(\ell-1,\ell)}(p_\ell) F_{\ell-1}(\ddot{s}_\ell) + \frac{n_{h_\ell}}{n} U_{\ell-1}(p_\ell)\right], \tag{3.47}$$

with base case

$$F_0(\ddot{s}_0) = x(p_0) \cdot \frac{n_{h_0}}{n}. \tag{3.48}$$

Making use of these refined recursions directly, the dynamic program in Algorithm 2 can be used to compute $\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n})$. The time complexity of lines 7, 8, and 9 are $O(m)$, $O(n_\text{u})$, and $O(n_\text{u})$, respectively, and the time complexity of lines 6–10 is therefore $O(m(m+n_\text{u}))$. As a result, the time complexity for lines 5–11, and for the algorithm as a whole, is $O(km(m+n_\text{u}))$.

Algorithm 2 represents a substantial improvement over the quadratic dependence on n_u in the ordinary forward algorithm for HMMs, given in Algorithm 1. The key improvement is that the quantity $U_{\ell-1}(p_\ell)$ is reused in computing each value of $F_\ell(p_\ell, h_\ell)$, which is made possible by the independence described in Property 1. This simple optimization has been generally adopted (Fearnhead and Donnelly, 2001; Li and Stephens, 2003; Paul et al., 2011), and serves as a baseline for improvement.

3.3.2 Improving efficiency via the emission distribution

The MCG at locus $\ell \in L$, representing the hidden state of the HMM, is denoted by a tuple $\ddot{s}_\ell = (p_\ell, h_\ell)$. However, the emission distribution, with density $\xi_\ell(\cdot|\ddot{s}_\ell)$ defined by (3.15), associated

with the observed allele $\eta[\ell]$ depends on the absorption haplotype $h_\ell \in \mathcal{H}_{\mathbf{n}}$ only through the allele $h_\ell[\ell] \in A_\ell$. As a result,

Property 2. *Consider a subset $\mathcal{B} \subset \mathcal{H}_{\mathbf{n}}$ such that there exists an allele a with $h[\ell] = a$ for all $h \in \mathcal{B}$. Then, for each absorption interval $p_\ell \in \mathcal{P}$, the emission distribution $\xi_\ell(\cdot|p_\ell, h_\ell)$ is identical for all $h_\ell \in \mathcal{B}$. We indicate this fact by writing $\xi_\ell(\cdot|p_\ell, h_\ell) = \xi_\ell(\cdot|p_\ell, \mathcal{B})$ for all $h_\ell \in \mathcal{B}$.*

With this in mind, define a *partition* $\mathcal{C}$ of the haplotype configuration $\mathbf{n}$ to be a collection of blocks of the form $(\mathcal{B}, \ell_{\mathrm{s}}, \ell_{\mathrm{e}})$, where $\mathcal{B} \subset \mathcal{H}_{\mathbf{n}}$ and $1 \le \ell_{\mathrm{s}} \le \ell_{\mathrm{e}} \le k$, such that

- For every $(\mathcal{B}, \ell_{\mathrm{s}}, \ell_{\mathrm{e}}) \in \mathcal{C}$, there exists a sub-haplotype x such that $h[\ell_{\mathrm{s}} : \ell_{\mathrm{e}}] = x$ for all $h \in \mathcal{B}$.
- For every haplotype $h \in \mathcal{H}_{\mathbf{n}}$ and $1 \le \ell \le k$, there exists *exactly* one $(\mathcal{B}, \ell_{\mathrm{s}}, \ell_{\mathrm{e}}) \in \mathcal{C}$ with $h \in \mathcal{B}$ and $\ell_{\mathrm{s}} \le \ell \le \ell_{\mathrm{e}}$.

For a given locus $\ell \in L$, a configuration partition $\mathcal{C}$ induces a natural partition of the haplotypes $\mathcal{H}_{\mathbf{n}}$, denoted by $\mathcal{C}_\ell$, and Property 2 applies to each $\mathcal{B} \in \mathcal{C}_\ell$. In the next sections, we present new forward recursions and dynamic programming algorithms valid for an arbitrary partition $\mathcal{C}$.

The computational complexity of these algorithms will depend on $\mathcal{C}$ through two functions, namely $\Psi(\mathcal{C})$ and $\Omega(\mathcal{C})$, defined as follows: For locus ℓ, define $\psi_\ell(\mathcal{C}) = |\mathcal{C}_\ell|$, the number of blocks in $\mathcal{C}_\ell$, and define $\omega_\ell(\mathcal{C})$ to be the total number of haplotypes in blocks of the configuration partition *ending* at locus ℓ. Then,

$$\Psi(\mathcal{C}) = \sum_{\ell=1}^{k} \psi_\ell(\mathcal{C}) = \sum_{\ell=1}^{k} |\mathcal{C}_\ell|,$$
$$\Omega(\mathcal{C}) = \sum_{\ell=1}^{k} \omega_\ell(\mathcal{C}) = \sum_{(\mathcal{B},\ell_{\mathrm{s}},\ell_{\mathrm{e}})\in\mathcal{C}} |\mathcal{B}|.$$

In some cases, we are primarily concerned with polymorphic loci, and so we define $\Psi_{\mathrm{p}}(\mathcal{C})$ to be the analog of $\Psi(\mathcal{C})$ summed over *only* polymorphic loci.

Finally, we define the trivial partition C_{T} for haplotype configuration $\mathbf{n}$ as the partition containing a single block $(\{h\}, 1, k)$ for each $h \in \mathcal{H}_{\mathbf{n}}$. Note that $\Psi(C_{\mathrm{T}}) = k \cdot n_{\mathrm{u}}$ and $\Omega(C_{\mathrm{T}}) = n_{\mathrm{u}}$. See Figure 3.1 for an illustration of both C_{T} and two non-trivial configuration partitions.

A general configuration partition

Let $\mathcal{C}$ be a configuration partition of $\mathbf{n}$. Begin by defining

$$Q_\ell(p_\ell, \mathcal{B}) = \sum_{h_\ell \in \mathcal{B}} F_\ell(p_\ell, h_\ell), \tag{3.49}$$

so that $Q_\ell(p_\ell) = \sum_{\mathcal{B}\in\mathcal{C}_\ell} Q_\ell(p_\ell, \mathcal{B})$. Now suppose $(\mathcal{B}, \ell_{\mathrm{s}}, \ell_{\mathrm{e}}) \in \mathcal{C}$. Applying Definition (3.49) and Property 2 to (3.47), then for $\ell_{\mathrm{s}} \le \ell \le \ell_{\mathrm{e}}$,

$$Q_\ell(p_\ell, \mathcal{B}) = \xi_\ell(\eta[\ell]|p, \mathcal{B}) \left[y_{(\ell-1,\ell)}(p) Q_{\ell-1}(p, \mathcal{B}) + \frac{n_{\mathcal{B}}}{n} U_{\ell-1}(p) \right], \tag{3.50}$$

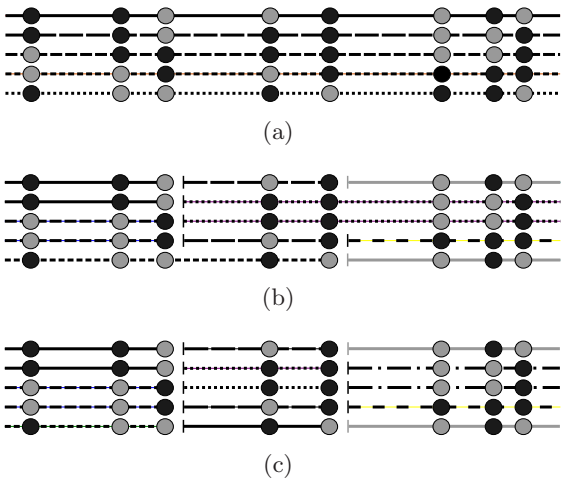


Figure 3.1. Illustration of three alternative configuration partitions. Each row represents a haplotype, with white and black circles representing the allele at each of 8 polymorphic loci. The color of haplotype indicates the block to which it belongs. (a) The trivial configuration partition C_{T}; $\Psi_{\mathrm{p}}(C_{\mathrm{T}}) = 40$ and $\Omega(C_{\mathrm{T}}) = 5$. (b) A non-trivial configuration partition, C; $\Psi_{\mathrm{p}}(C) = 24$ and $\Omega(C) = 12$. (c) The configuration partition C_s found by the algorithm described in Section 3.3.2 for $s = 3$; $\Psi_{\mathrm{p}}(C_s) = 24$ and $\Omega(C_s) = 15$.

where we have defined $n_{\mathcal{B}} = \sum_{h \in \mathcal{B}} n_h$. Similarly, by induction, and making use of (3.47) and (3.50), it is possible to show that, for $\ell_{\mathrm{s}} \leq \ell \leq \ell_{\mathrm{e}}$ and $h_\ell \in \mathcal{B}$,

$$F_\ell(p_\ell, h_\ell) = T_\ell(p_\ell, \mathcal{B}) \cdot F_{\ell_{\mathrm{s}}-1}(p_\ell, h_\ell) + \frac{n_{h_\ell}}{n_{\mathcal{B}}} \Big(Q_\ell(p_\ell, \mathcal{B}) - T_\ell(p_\ell, \mathcal{B}) Q_{\ell_{\mathrm{s}}-1}(p_\ell, \mathcal{B}) \Big), \tag{3.51}$$

where $T_\ell(p_\ell, \mathcal{B}) = \prod_{\ell'=\ell_{\mathrm{s}}}^{\ell} \xi_{\ell'}(\eta[\ell'] | p_\ell, \mathcal{B}) \cdot y_{(\ell-1,\ell)}(p_\ell)$, and solves the recursion,

$$T_\ell(p_\ell, \mathcal{B}) = \xi_\ell(\eta[\ell] | p_\ell, \mathcal{B}) \cdot y_{(\ell-1,\ell)}(p_\ell) \cdot T_{\ell-1}(p_\ell, \mathcal{B}), \tag{3.52}$$

for $\ell_{\mathrm{s}} \leq \ell \leq \ell_{\mathrm{e}}$, with base case $T_{\ell_{\mathrm{s}}-1}(p, \mathcal{B}) = 1$.

For each block $(\mathcal{B}, \ell_{\mathrm{s}}, \ell_{\mathrm{e}}) \in \mathcal{C}$, we take advantage of (3.50) and (3.52) to directly compute $Q_\ell(p, \mathcal{B})$ and $T_\ell(p, \mathcal{B})$ for each value of $p_\ell \in \mathcal{P}$, at every locus $\ell_{\mathrm{s}} \leq \ell \leq \ell_{\mathrm{e}}$. At the end of the block, when $\ell = \ell_{\mathrm{e}}$, the finer-grain values $F_\ell(p_\ell, h_\ell)$ are computed for each $p_\ell \in \mathcal{P}$ and $h_\ell \in \mathcal{B}$ using (3.51), and subsequently used to compute initial values for blocks beginning at locus $\ell + 1$. The associated dynamic program to compute the CSP $\hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta | \mathbf{n})$ is given in Algorithm 3. Observe that Algorithm 2 is a special case of this algorithm for $\mathcal{C} = C_{\mathrm{T}}$.

Within Algorithm 3, the time complexity of line 8 is $O(m)$; of line 9 is $O(\psi_\ell(\mathcal{C}))$; and of lines 10 and 11 is $O(\omega_\ell(\mathcal{C}))$. Thus, the time complexity of lines 7–12, and is $O(m(m + \psi_\ell(\mathcal{C}) + \omega_\ell(\mathcal{C})))$, and the time complexity of the entire algorithm is $O(km^2 + m(\Psi(\mathcal{C}) + \Omega(\mathcal{C})))$. Thus, if it is possible to obtain a configuration partition $\mathcal{C}$ for $\mathbf{n}$ such that $\Psi(\mathcal{C}) + \Omega(\mathcal{C})$ is substantially less than $\Psi(C_{\mathrm{T}}) + \Omega(C_{\mathrm{T}}) = kn_{\mathrm{u}} + n_{\mathrm{u}}$, our new algorithm may be considerably faster than Algorithm 2; constructing such a configuration partition is the subject of Section 3.3.2.

The absence of polymorphism

In many reasonable evolutionary scenarios, a great many loci will not be polymorphic. Accommodating such loci in the analysis is important and can be done efficiently making use of Property 2.

Algorithm 3 Compute $\hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta|\mathbf{n})$ using a forward-type algorithm improved by considering Properties 1 and 2, for a configuration partition $\mathcal{C}$

for all $p_0 \in \mathcal{P}$ **do**
  Compute $F_0(p_0, h)$ using (3.48), $\forall h_0 \in \mathcal{H}_{\mathbf{n}}$
  Compute $Q_0(p_0, \mathcal{B})$ using (3.49) and $T_0(p_0, \mathcal{B}) = 1$, $\forall(\mathcal{B}, 1, \ell_{\mathrm{e}}) \in \mathcal{C}$
  Compute $Q_0(p_0)$ using (3.49)
end for
for $\ell = 1 \to k$ **do**
  for all $p_\ell \in \mathcal{P}$ **do**
    Compute $U_{\ell-1}(p_\ell)$ using (3.45)
    Compute $Q_\ell(p_\ell, \mathcal{B})$ and $T_\ell(p_\ell, \mathcal{B})$ using (3.50) and (3.52), $\forall(\mathcal{B}, \ell_{\mathrm{s}}, \ell_{\mathrm{e}}) \in \mathcal{C}$ such that $\ell_{\mathrm{s}} \leq \ell \leq \ell_{\mathrm{e}}$; compute $Q_\ell(p_\ell)$ using (3.49)
    Compute $F_\ell(p_\ell, h_\ell)$ using (3.51), $\forall h_\ell \in \mathcal{B}$ and $\forall(\mathcal{B}, \ell_{\mathrm{s}}, \ell) \in \mathcal{C}$
    Compute $Q_\ell(p_\ell, \mathcal{B})$ using (3.49) and $T_\ell(p_\ell, \mathcal{B}) = 1$, $\forall(\mathcal{B}, \ell + 1, \ell_{\mathrm{e}}) \in \mathcal{C}$
  end for
end for
Compute $\hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta|\mathbf{n})$ using (3.46)

In particular, for a non-polymorphic locus ℓ, Property 2 applies to the trivial set $\mathcal{B}_0 = \mathcal{H}_{\mathbf{n}}$, and therefore the emission distribution can be written $\xi_\ell(\cdot|p, \mathcal{B}_0) = \xi_\ell(\cdot|p)$; moreover, $Q_\ell(p) = Q_\ell(p, \mathcal{B}_0)$.

Suppose consecutive loci $\ell^*_{\mathrm{s}}, \ldots, \ell^*_{\mathrm{e}}$ are not polymorphic. Rewriting equations (3.50) and (3.51) for block $(\mathcal{B}_0, \ell^*_{\mathrm{s}}, \ell^*_{\mathrm{e}})$ yields, for $\ell^*_{\mathrm{s}} \leq \ell \leq \ell^*_{\mathrm{e}}$,

$$Q_\ell(p_\ell) = \xi_\ell(\eta[\ell]|p_\ell) \cdot \Big[y_{(\ell-1,\ell)}(p_\ell) Q_{\ell-1}(p_\ell) + U_{\ell-1}(p_\ell) \Big], \tag{3.53}$$

and, for $\ell^*_{\mathrm{s}} \leq \ell \leq \ell^*_{\mathrm{e}}$ and $h_\ell \in \mathcal{B}_0 = \mathcal{H}_{\mathbf{n}}$,

$$F_\ell(p_\ell, h_\ell) = T_\ell(p_\ell) \cdot F_{\ell^*_{\mathrm{s}}-1}(p_\ell, h_\ell) + \frac{n_{h_\ell}}{n} \Big(Q_\ell(p_\ell) - T_\ell(p_\ell) Q_{\ell^*_{\mathrm{s}}-1}(p_\ell) \Big), \tag{3.54}$$

where $T_\ell(p_\ell) = \prod_{\ell'=\ell^*_{\mathrm{s}}}^{\ell} \xi_{\ell'}(\eta[\ell']|p_\ell) \cdot y_{(\ell-1,\ell)}(p_\ell)$ and solves the recursion

$$T_\ell(p_\ell) = \xi_\ell(\eta[\ell]|p_\ell) \cdot y_{(\ell-1,\ell)}(p_\ell) \cdot T_{\ell-1}(p_\ell), \tag{3.55}$$

for $\ell^*_{\mathrm{s}} \leq \ell \leq \ell^*_{\mathrm{e}}$, with base case $T_{\ell^*_{\mathrm{s}}-1}(p_\ell) = 1$.

Now let $\mathcal{C}$ be a configuration partition with $(\mathcal{B}, \ell_{\mathrm{s}}, \ell_{\mathrm{e}}) \in \mathcal{C}$. Suppose that there is a stretch of non-polymorphic loci $\ell^*_{\mathrm{s}}, \ldots, \ell^*_{\mathrm{e}}$, and that $\ell_{\mathrm{s}} \leq \ell^*_{\mathrm{s}} \leq \ell^*_{\mathrm{e}} \leq \ell_{\mathrm{e}}$. Applying definition (3.49) to (3.54), yields, for $\ell^*_{\mathrm{s}} \leq \ell \leq \ell^*_{\mathrm{e}}$,

$$Q_\ell(p_\ell, \mathcal{B}) = T_\ell(p_\ell) Q_{\ell^*_{\mathrm{s}}-1}(p_\ell, \mathcal{B}) + \frac{n_{\mathcal{B}}}{n} \Big[Q_\ell(p_\ell) - T_\ell(p_\ell) Q_{\ell^*_{\mathrm{s}}-1}(p_\ell) \Big]. \tag{3.56}$$

Similarly considering the definition of $T_\ell(p_\ell, \mathcal{B})$ along with (3.55),

$$T_\ell(p_\ell, \mathcal{B}) = T_\ell(p_\ell) \cdot T_{\ell^*_{\mathrm{s}}-1}(p_\ell, \mathcal{B}). \tag{3.57}$$

Algorithm 3 can be modified to accommodate such stretches of non-polymorphic loci as a special case, making use of (3.53) and (3.55) to directly compute the values of $Q_\ell(p_\ell)$ and $T_\ell(p_\ell)$ for each $p_\ell \in \mathcal{P}$, and at each non-polymorphic locus ℓ. If we then assume (without loss of generality) that each $(\mathcal{B}, \ell_{\mathrm{s}}, \ell_{\mathrm{e}}) \in \mathcal{C}$ has ℓ_{e} at a polymorphic locus, then at the final non-polymorphic locus, for

Algorithm 4 Computation of $\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n})$ improved by considering Properties 1 and 2, and a special case for non-polymorphic loci, for a configuration partition $\mathcal{C}$ such that $\forall(\mathcal{B}, \ell_\text{s}, \ell_\text{e}) \in \mathcal{C}$, ℓ_e is polymorphic

1: Algorithm 3, lines 1–5; and set $T_0(p_0) = 1\ \forall p_0 \in \mathcal{P}$ and $\ell_\text{s}^* = 1$
2: **for** $\ell = 1 \to k$ **do**
3: **for all** $p_\ell \in \mathcal{P}$ **do**
4: **if** locus ℓ is polymorphic **then**
5: **if** locus $\ell - 1$ is not polymorphic **then**
6: Compute $Q_{\ell-1}(p_\ell, \mathcal{B})$ and $T_{\ell-1}(p_\ell, \mathcal{B})$ using (3.56) and (3.57), $\forall(\mathcal{B}, \ell_\text{s}, \ell_\text{e}) \in \mathcal{C}$ such that $\ell_\text{s} \le \ell \le \ell_\text{e}$
7: **end if**
8: Algorithm 3, lines 8–11
9: Set $T_\ell(p_\ell) = 1$ and $\ell_\text{s}^* = \ell + 1$
10: **else**
11: Compute $U_{\ell-1}(p_\ell)$, $Q_\ell(p_\ell)$, and $T_\ell(p_\ell)$ using (3.45), (3.53), and (3.55)
12: **end if**
13: **end for**
14: **end for**
15: Compute $\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n})$ using (3.46)

which $\ell = \ell_\text{e}^*$, (3.56) and (3.57) may be used to compute $Q_\ell(p_\ell, \mathcal{B})$ and $T_\ell(p_\ell, \mathcal{B})$ for each $p_\ell \in \mathcal{P}$ and $\mathcal{B} \in \mathcal{C}_\ell$. This modification is detailed in Algorithm 4.

Within Algorithm 4, the time complexity of lines 6 and 9 is $O(1)$, of line 8 is $O(m+\psi_\ell(\mathcal{C})+\omega_\ell(\mathcal{C}))$, and of line 11 is $O(m)$. As a result, the time complexity of lines 2 – 14, and of the entire algorithm, is $O(km^2 + m(\Psi_\text{p}(\mathcal{C}) + \Omega(\mathcal{C})))$. Relative to Algorithm 3, less computation needs to be done for non-polymorphic loci; thus, in the typical case of many non-polymorphic loci, this dynamic program will have a decreased running time. For $\mathcal{C} = C_\text{T}$, the time complexity is $O(km^2 + k_\text{p} m n_\text{u})$.

An optimization for non-polymorphic loci

The key recursions (3.53) and (3.55) for non-polymorphic loci can be written in matrix form. Consider an ordering $\mathcal{P} = \{p^{(1)}, \ldots, p^{(m)}\}$, and define the quantities:

- The m-dimensional column vectors $\mathcal{Q}_\ell$ and $\mathcal{T}_\ell$, with the i-th entry given by $Q_\ell(p^{(i)})$ and $T_\ell(p^{(i)})$, respectively.
- The $(m \times m)$-dimensional diagonal matrices $\mathcal{E}_\ell$ and $\mathcal{Y}_\ell$, with the (i,i)-th entry given by $\xi_\ell(\eta[\ell]|p^{(i)})$ and $y_{(\ell-1,\ell)}(p^{(i)})$, respectively.
- The $(m \times m)$-dimensional matrix $\mathcal{Z}_\ell$, with the (i,j)-th entry given by $z_{(\ell-1,\ell)}(p^{(i)}|p^{(j)})$.

Then (3.53) and (3.55) can be written in the following matrix forms,

$$\mathcal{Q}_\ell = \mathcal{E}_\ell(\mathcal{Y}_\ell + \mathcal{Z}_\ell)\mathcal{Q}_{\ell-1} \tag{3.58}$$

$$\mathcal{T}_\ell = \mathcal{E}_\ell \mathcal{Y}_\ell \mathcal{T}_{\ell-1}, \tag{3.59}$$

Now, suppose that the mutation model is symmetric and the mutation rate constant for all loci. Then $\mathcal{E}_\ell = \mathcal{E}$, for all non-polymorphic loci $\ell \in L$. Similarly if the recombination rate between each pair of loci is constant, then $\mathcal{Y}_\ell = \mathcal{Y}$ and $\mathcal{Z}_\ell = \mathcal{Z}$ for all non-polymorphic loci $\ell \in L$. With these

Algorithm 5 Computation of $\hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta|\mathbf{n})$ improved by considering Properties 1 and 2, and a special *optimized* case for non-polymorphic loci, for a configuration partition $\mathcal{C}$ such that $\forall(\mathcal{B}, \ell_{\mathrm{s}}, \ell_{\mathrm{e}}) \in \mathcal{C}$, ℓ_{e} is polymorphic

```
1: Algorithm 4, line 1
2: for polymorphic ℓ ∈ {1 → k} do
3:     for all p_ℓ ∈ P do
4:         if locus ℓ − 1 is not polymorphic then
5:             Compute Q_{ℓ−1}(p_ℓ) and T_{ℓ−1}(p_ℓ) using (3.60)
6:         end if
7:         Algorithm 4, lines 5–9
8:     end for
9: end for
10: Compute π̂_SMC(e_η|n) using (3.46)
```

assumptions, for $\ell^*_{\mathrm{s}} \leq \ell \leq \ell^*_{\mathrm{e}}$,

$$\mathcal{Q}_\ell = \mathcal{E}(\mathcal{Y} + \mathcal{Z})\mathcal{Q}_{\ell-1} = (\mathcal{E}(\mathcal{Y} + \mathcal{Z}))^{\ell-\ell^*_{\mathrm{s}}+1}\mathcal{Q}_{\ell^*_{\mathrm{s}}-1}, \tag{3.60}$$

$$\mathcal{T}_\ell = \mathcal{E}\mathcal{Y}\mathcal{T}_{\ell-1} = (\mathcal{E}\mathcal{Y})^{\ell-\ell^*_{\mathrm{s}}+1}\mathcal{T}_{\ell^*_{\mathrm{s}}-1}, \tag{3.61}$$

and the values of $(\mathcal{E}(\mathcal{Y}+\mathcal{Z}))^r$ and $(\mathcal{E}\mathcal{Y})^r$ can be pre-computed (either by eigenvalue decomposition or repeated multiplication) for a relevant range of r-values. Using this technique for explicitly computing only the necessary values of $Q_\ell(p)$ and $T_\ell(p)$, stretches of non-polymorphic loci can be *analytically* skipped.

The modified dynamic program associated with this optimization is given in Algorithm 5. The time complexity of line 5 is $O(m)$, and of line 7 $O(m+\psi_\ell(\mathcal{C})+\omega_\ell(\mathcal{C}))$. Thus, the time complexity of lines 2–9, and for the entire algorithm, is $O(k_{\mathrm{p}}m^2 + m(\Psi_{\mathrm{p}}(\mathcal{C}) + \Omega(\mathcal{C})))$. This refinement once again reduces the computation required for non-polymorphic loci, and so we might expect substantial improvements in performance over Algorithms 3 and 4. For the choice $\mathcal{C} = C_{\mathrm{T}}$, the time complexity is $O(k_{\mathrm{p}}m(m + n_{\mathrm{u}}))$.

Note that the assumptions necessary for Algorithm 5, namely a symmetric mutation model and uniform mutation and recombination rates, can be relaxed, but at the expense of additional pre-computation. For example, given non-uniform, but locally similar recombination rates, pre-computation might be performed for each of several rates; each stretch of non-polymorphic loci could then use the pre-computed values associated with the closest recombination rate.

A fast algorithm for configuration partitions

Thus far, we have assumed that a configuration partition $\mathcal{C}$ was specified, and showed that, for Algorithms 3–5, the time complexity depends on $\mathcal{C}$ through the functions $\Psi(\mathcal{C})$ (or $\Psi_{\mathrm{p}}(\mathcal{C})$) and $\Omega(\mathcal{C})$, and more particularly their sum. These complexity results are summarized in Table 3.1, for both a general configuration partition $\mathcal{C}$ and assuming the trivial configuration partition $\mathcal{C} = C_{\mathrm{T}}$. It is intuitively clear that a configuration partition minimizing Ω will naturally maximize Ψ (as in C_{T}), and vice versa; minimizing a convex combination of these quantities is therefore difficult. In this section, we propose a fast and simple parameterized algorithm for constructing reasonably good configuration partitions.

Given a configuration $\mathbf{n}$, the algorithm proceeds sequentially over the loci: Initially, set $\ell_{\mathrm{s}} = 1$. Given ℓ_{s}, find the largest polymorphic locus ℓ_{e} such that $\ell_{\mathrm{s}} \leq \ell_{\mathrm{e}} \leq k$, and the number of unique

	$\mathcal{C} = C_{\mathrm{T}}$	General $\mathcal{C}$
Algorithm 3	$O(km \cdot (m + n_{\mathrm{u}}))$	$O(km^2 + m \cdot (\Psi(\mathcal{C}) + \Omega(\mathcal{C})))$
Algorithm 4	$O(km^2 + k_{\mathrm{p}} m n_{\mathrm{u}})$	$O(km^2 + m \cdot (\Psi_{\mathrm{p}}(\mathcal{C}) + \Omega(\mathcal{C})))$
Algorithm 5	$O(k_{\mathrm{p}} m \cdot (m + n_{\mathrm{u}}))$	$O(k_{\mathrm{p}} m^2 + m \cdot (\Psi_{\mathrm{p}}(\mathcal{C}) + \Omega(\mathcal{C})))$

Table 3.1. A summary of the optimized algorithms for computing $\hat{\pi}_{\mathrm{SMC}(m)}$, along with their asymptotic time complexities, for both a general configuration partition $\mathcal{C}$ and assuming the trivial configuration partition $\mathcal{C} = C_{\mathrm{T}}$. As described in the text, Algorithm 3 with $\mathcal{C} = C_{\mathrm{T}}$ is formally equivalent to Algorithm 2

sub-haplotypes between loci ℓ_{s} and ℓ_{e} is at most some threshold parameter s. Then, for each unique sub-haplotype x between ℓ_{s} and ℓ_{e}, group all $h \in \mathcal{H}_{\mathbf{n}}$ such that $h[\ell_{\mathrm{s}} : \ell_{\mathrm{e}}] = x$ into the same block $\mathcal{B}$ and add $(\mathcal{B}, \ell_{\mathrm{s}}, \ell_{\mathrm{e}})$ to the configuration partition. Set $\ell_{\mathrm{s}} = \ell_{\mathrm{e}} + 1$ and repeat until the final locus k is reached. An example configuration partition resulting from this algorithm is shown in Figure 3.1(c).

Applying this procedure to configuration $\mathbf{n}$ with threshold parameter s results in a configuration partition which we denote C_s. Observe that for $s = |\mathcal{H}_{\mathbf{n}}|$, we obtain $C_s = C_{\mathrm{T}}$, which minimizes Ω. On the other hand, for $s = 2$ (in the bi-allelic case), the algorithm produces a configuration partition that minimizes Ψ_{p}. Using intermediate values of s should then produce the intermediate configuration partitions that are of interest.

A plot of $\Psi_{\mathrm{p}}(C_s)$ and $\Omega(C_s)$ for several values of s is given in Figure 3.2(a) for a particular haplotype configuration $\mathbf{n}$, which was generated using coalescent simulation. As anticipated, there is an inverse relationship between $\Psi_{\mathrm{p}}(C_s)$ and $\Omega(C_s)$. In order to gauge the effect of different combinations of Ψ_{p} and Ω on the running time, the CSP $\hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta|\mathbf{n})$ was computed for each of the configuration partition C_s for several values of s, and the associated time recorded; the results are plotted in Figure 3.2(b). As our intuition suggested, the running time depends substantially on the choice of $\mathcal{C}$, and, in accordance with the asymptotic time complexity results, depends linearly on both Ψ_{p} and Ω.

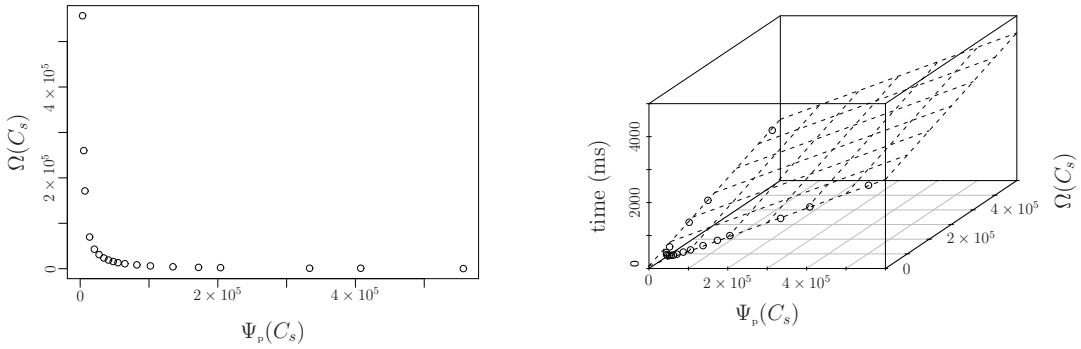
By fitting a linear model to the data, we can deduce the constants associated with Ψ_{p} and Ω, which the asymptotic results alone cannot provide. Though the particular values for these constants will depend on the implementation and hardware, their ratio should be relatively robust to these details, and therefore informative for choosing an optimal trade-off between Ψ_{p} and Ω. We have found that the constant associated with Ψ_{p} is approximately 1.5 times that associated with Ω, suggesting that running time is minimized for a choice of $\mathcal{C}$ that minimizes $1.5 \cdot \Psi_{\mathrm{p}}(\mathcal{C}) + \Omega(\mathcal{C})$. Further, making use of the above algorithm, we define

$$s^* = \operatorname*{argmin}_s \Big\{ 1.5 \cdot \Psi_{\mathrm{p}}\big(C_s\big) + \Omega\big(C_s\big) \Big\},$$

and $C^* = C_s$. In practice the value s^* is found using binary search, and determining C^* is very fast. This definition will be used frequently in Chapter 4, as C^* (and the analogous result for Algorithm 2, using Ψ in place of Ψ_{p}) provides a good, though not necessarily optimal, choice for $\mathcal{C}$.

3.3.3 Backward algorithm and marginal decoding

We have thus far considered optimizations and algorithms for evaluating the *forward* recursion associated with the HMM formulation of the CSP $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$. Recall from Section 3.2.3 that there is also a *backward* recursion (3.38) associated with the same HMM. We here state, but do not explicitly demonstrate, that for each optimized forward recursion, there exists an analogous optimized



(a) (b)

Figure 3.2. The relationship of $\Psi_{\mathrm{p}}(C_s)$, $\Omega(C_s)$, and running time for several values of $s \in (2, \ldots, 500)$ and a particular configuration $\mathbf{n}$. The configuration $\mathbf{n}$ was generated using coalescent simulation for 500 individuals, each having 100000 bi-allelic loci, using population-scaled mutation rate $\theta = 0.005$ per locus and population-scaled recombination rate $\rho = 0.001$ per breakpoint, and resulting in $k_{\mathrm{p}} = 1724$ polymorphic loci and $n_{\mathrm{u}} = 324$ unique haplotypes. (a) Plot of the values of $\Psi_{\mathrm{p}}(C_s)$ and $\Omega(C_s)$ for each value of s, demonstrating the tradeoff between small Ψ_{p} (small s values), and small Ω (large s values). (b) Plot including the empirically observed running time of Algorithm 5 used to compute $\hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta|\mathbf{n})$, for arbitrary $\eta \in \mathcal{H}_{\mathbf{n}}$. As predicted by the asymptotic time complexity results, running time appears to depend linearly on both Ψ_{p} and Ω values, and fitting a linear model indicates the constant associated with Ψ_{p} is approximately 1.5 times greater than the constant associated with Ω.

backward recursion. Consequently, Algorithms 1–5 each have a backward analogue, with identical time and space complexity.

Recall from Section 3.2.3 that marginal decoding can be efficiently realized by pre-computing and storing both the forward and backward values, $F_\ell(\ddot{s}_\ell)$ and $E_\ell(\ddot{s}_\ell)$, at every $\ddot{s}_\ell \in \ddot{\mathcal{S}}_{\mathbf{n}}$ and for every $\ell \in L$. Using Algorithm 2, this can be accomplished with time complexity $O(km(m + n_{\mathrm{u}}))$ and space complexity $O(kmn_{\mathrm{u}})$. Following this pre-computation, marginal decoding at an arbitrary locus can be accomplished by directly applying (3.40) with associated time complexity $O(mn_{\mathrm{u}})$.

Using the optimized recursions and dynamic programming algorithms, it is possible to compute and store substantially less. Consider, for example, Algorithm 4 and the recursions of Section 3.3.2. Suppose each *computed* value of $F_\ell(\ddot{s}_\ell)$ (that is, for each $\ell \in L$ with $(\mathcal{B}, \ell_{\mathrm{s}}, \ell) \in \mathcal{C}$ and for $\ddot{s}_\ell \in \mathcal{P} \times \mathcal{B}$) is cached, and similarly, each computed value of $Q_\ell(p_\ell, \mathcal{B})$, $T_\ell(p_\ell, \mathcal{B})$, $Q_\ell(p_\ell)$, and $T_\ell(p_\ell)$ is cached. Then by using only cached values, $F_\ell(\ddot{s}_\ell)$ can be computed for all $\ddot{s}_\ell \in \ddot{\mathcal{S}}_{\mathbf{n}}$ with time complexity $O(mn_{\mathrm{u}})$ using (3.51), in conjunction with (3.56) and (3.57). Combined with the analogous backward computations, a marginal decoding can accomplished at locus $\ell \in L$ by applying (3.40) with time complexity $O(mn_{\mathrm{u}})$.

By using Algorithm 4 in place of Algorithm 2, the required pre-computation can realized with time complexity $O(km^2 + m(\Psi_{\mathrm{p}}(\mathcal{C}) + \Omega(\mathcal{C})))$ and space complexity $O(km + m(\Psi_{\mathrm{p}}(\mathcal{C}) + \Omega(\mathcal{C}))$, representing a substantial and practically beneficial improvement over the baseline algorithm. The same techniques can be applied using Algorithm 3 and Algorithm 5 in place of Algorithm 4. Moreover, if a particular application requires only a coarse-grained marginal decoding, consisting of a probability distribution over discretized time and the sets comprising the partition $\mathcal{C}_\ell$, the computations can again be simplified. Using these general principles, many posterior inference tasks can be car-

ried out more efficiently, with respect to both time and space complexity, than by using the most general HMM methodology.

3.3.4 Applicability to related CSDs

Though we have developed optimized algorithms for the CSD $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$ provided a single conditionally sampled haplotype in the absence of population structure, it is natural to question whether similar optimizations are applicable to related CSDs, such as those proposed by Fearnhead and Donnelly (2001) and by Li and Stephens (2003). In Sections 3.3.1 and 3.3.2, we have provided two sufficient conditions: Property 1, which stipulates that, upon recombination, a new haplotype is chosen independently and uniformly at random; and Property 2, which stipulates that the emission distribution depends only on the allele at the current locus of the hidden haplotype. It is straightforward to verify that both $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$ do satisfy both of these properties, and so the optimizations described are immediately applicable. Moreover, stronger forms of Property 1 hold for both $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$, enabling additional optimization. Though we have not empirically analyzed the resulting algorithms, asymptotic complexity results suggest that the improvements in efficiency will be qualitatively comparable to those obtained for $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$.

It is also interesting to consider variants of $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$ for more complex demographic scenarios such as a structured population including migration, as described in Section 3.2.2. Observe that in this setting, Property 1 is not satisfied, as the haplotypes are only sampled independently *within the current deme*, and the optimizations are therefore not applicable. Nonetheless, a relaxed version of Property 1 incorporating the population structure, is satisfied, along with Property 2, and we conjecture that analogous optimizations are possible. The outcome is similar if $\hat{\pi}_{\mathrm{SMC}}$ is extended to conditionally sampling two haplotypes, as described in Section 2.3.4. More generally, we anticipate the properties akin to Property 1 and Property 2 can be used to derive similar optimizations for a broad class of population genetic HMMs.

Chapter 4

Results & Applications

In the past decade, the conditional sampling distribution (CSD) has found a wide range of applications in population genetics. In part, this is due to the fact that many general statistical procedures requiring the joint analysis of many individuals can be naturally rephrased in terms of one or more CSDs. Moreover, the CSD is intuitively appealing, and, as demonstrated in the previous chapters, well-suited to efficient approximation. In this chapter, we describe and extend several frequently used CSD-based statistical methods, and also empirically assess both the relative accuracy and computational efficiency of our proposed approximate CSDs.

Methods employing the CSD can be roughly partitioned into several overlapping categories. One such category is parametric inference based on the sampling probability, or likelihood, of a sample. For small samples, the sampling probability can be computed directly using CSD-based importance sampling (Stephens and Donnelly, 2000; Fearnhead and Donnelly, 2001; De Iorio and Griffiths, 2004a,b; Griffiths et al., 2008); for larger samples, importance sampling can be used in conjunction with composite methods (Hudson, 2001; Fearnhead and Donnelly, 2002). Alternatively, the sampling probability can be approximated directly using a decomposition into approximate conditional sampling probabilities; this technique is referred to as the product of approximate conditionals (PAC) method (Li and Stephens, 2003). In conjunction with expectation-maximization, and Markov chain Monte Carlo, these methods have been fruitfully used for the estimation of fine-scale recombination rates (Li and Stephens, 2003; Crawford et al., 2004; McVean et al., 2004; Fearnhead and Smith, 2005), gene conversion parameters (Gay et al., 2007; Yin et al., 2009), and population demography (Davison et al., 2009).

It is also possible to directly employ the genealogical interpretation of the CSD. In particular, provided a CSD that can be cast as an HMM, such as the sequentially Markov CSD $\hat{\pi}_{\mathrm{SMC}}$ described in Section 2.3, the hidden states can be inferred and used directly. This technique has been used for admixture inference (Sundquist et al., 2008; Price et al., 2009; Wegmann et al., 2011), for which genomic segments corresponding to ancestral populations are identified, for inference of colonization history and structure (Hellenthal et al., 2008; Lawson et al., 2012), and within a pseudo-Gibbs framework for statistically phasing genotype data into haplotype data and imputing missing data (Stephens and Scheet, 2005; Li and Abecasis, 2006; Marchini et al., 2007; Howie et al., 2009; Li et al., 2010). We remark that the latter methods can also be used for multi-sample genotype calling and phasing for next-generation sequence data (Nielsen et al., 2011).

In all such applications, the fidelity with which the surrogate CSD $\hat{\pi}$ approximates the true CSD π directly impacts the quality of the resulting inference. Similarly, because the methods

generally rely on iterative Monte Carlo or expectation maximization techniques, with nearly all of the running time expended on CSD computation, the surrogate CSD $\hat{\pi}$ must be computationally efficient. We remark that many of the above techniques require several hours, or in some cases days, to produce a result, even for relatively modest non-genomic datasets (Howie et al., 2009); consequently, the choice of CSD is often made on the basis of efficiency, and at the expense of accuracy (Li and Stephens, 2003; Stephens and Scheet, 2005; Scheet and Stephens, 2006; Browning and Browning, 2007).

The remainder of this chapter is organized as follows. We first empirically assess both the relative accuracy and computational efficiency of our proposed CSDs; we find that our CSDs are generally more accurate, and using the algorithms and optimizations described in Sections 3.2 and 3.3, more computationally efficient than previously-proposed CSDs. We next describe and extend two commonly-used CSD-based computational kernels, importance sampling and the PAC method, and evaluate their performance using the CSDs developed herein. Finally, we propose a novel CSD-based method for efficiently sampling the marginal genealogy at an given locus from an approximate posterior distribution; this method is directly applicable for techniques requiring ancestral inference, including the identification of regions that are identical by descent, and the identification of risk-increasing polymorphisms in case-control association studies.

4.1 Empirical Accuracy and Timing

In this section, we empirically investigate the accuracy of our proposed approximate CSDs, and compare the results with the accuracy of the frequently-used CSDs $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$, described in Sections 1.4.2 and 1.4.3. We are specifically interested in the CSDs $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{SMC}}$, described in Chapter 2, as well as the $\mathcal{P}$-discretized approximation $\hat{\pi}_{\text{SMC}(\mathcal{P})}$, described in Section 3.2. We also empirically investigate the running time associated with CSP computation, particularly provided the algorithmic optimizations introduced in Section 3.3.

Directly assessing the accuracy of the CSDs requires evaluating the CSP associated with the true CSD π. In order to compute this quantity, we rely on importance sampling, a Monte Carlo technique described in Section 4.2, to estimate the ordered sampling probabilities comprising the definition (1.61) of the CSP. Even within this Monte Carlo framework, the size of samples that can be analyzed is modest, limited in practice to $n \leq 10$ haplotypes and $k \leq 10$ loci. Consequently, in order to understand the behavior of the approximate CSDs for larger samples, it is necessary to use successive approximations to the CSD π. We remark that although interpretation is confounded when using an approximate CSD in place of π, it remains possible to obtain useful information about the relationship of the various CSDs.

4.1.1 Data simulation

For simplicity, we consider a 2-allele model with $\Phi^{(\ell)} = \Phi_0 = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$, $\theta_\ell = \theta$ for $\ell \in L$, and $\rho_b = \rho$ for $b \in B$. We also assume that there is no population structure, and consequently no migrational process. With the objective of sampling a k-locus n-haplotype configuration $\mathbf{n}$, we propose the following distinct coalescent-based methodologies, parameterized by θ_0 and ρ_0.

M1: Directly sample the k-locus n-haplotype configuration $\mathbf{n}$, using the coalescent with recombination, setting $\theta = \theta_0$ and $\rho = \rho_0$.

M2: Set $k_0 \gg k$, and sample a k_0-locus n-haplotype configuration $\mathbf{n}_0$, using the coalescent with recombination, setting $\theta = \theta_0$ and $\rho = \rho_0$. Restrict the configuration to the central k *polymorphic* loci, recording their positions, to form the k-locus n-haplotype configuration $\mathbf{n}$.

The first methodology (M1) simulates genomic data; consequently, setting θ_0 and ρ_0 to biologically-motivated values, most of the loci in the sampled configuration $\mathbf{n}$ will be non-polymorphic, reflecting the common biological observation. In contrast, the second methodology (M2) produces a simulated configuration wherein all of the loci of $\mathbf{n}$ are polymorphic by construction. The latter is useful for producing non-trivial haplotype configurations with a small number of loci, comparable to the SNP data commonly used for population genetic analyses.

Provided a k-locus n-haplotype configuration $\mathbf{n}$, we sample a k-locus n-haplotype conditional configuration $C = (\mathbf{e}_\eta, \mathbf{n} - \mathbf{e}_\eta)$ by selecting a single haplotype η from $\mathbf{n}$ uniformly at random. For notational convenience, we define the CSP $\hat{\pi}(C) = \hat{\pi}(\mathbf{e}_\eta | \mathbf{n} - \mathbf{e}_\eta)$. For a dataset C simulated using method M1, we evaluate $\hat{\pi}(C)$ using the true parameter values, $\theta_\ell = \theta_0$ and $\rho_b = \rho_0$ for all $\ell \in L$ and $b \in B$. For a dataset C simulated using method M2, we evaluate $\hat{\pi}(C)$ using parameter values $\theta_\ell = \theta_0$ for all $\ell \in L$, and $\rho_b = \rho_0 \cdot d_b$, where d_b is the distance, in loci, between the associated polymorphic sites. Observe that in in the latter case, the resulting CSP is computed for a model that is inequivalent to that which produced the data; nonetheless, the operation is well-defined, and frequently used in practice.

4.1.2 Accuracy

We evaluate the accuracy of a CSD $\hat{\pi}$ relative to a reference CSD π_0 using the expected absolute log-ratio (ALR) error,

$$\mathrm{ALRErr}_{k,n}(\hat{\pi} | \pi_0) \approx \frac{1}{N} \sum_{i=1}^{N} \left| \log_{10} \left(\frac{\hat{\pi}(C^{(i)})}{\pi_0(C^{(i)})} \right) \right|, \tag{4.1}$$

where N denotes the number of simulated data sets and $C^{(i)}$ is a k-locus n-haplotype conditional configuration sampled using one of the methods indicated above. For example, if $\mathrm{ALRErr}_{k,n}(\hat{\pi} | \pi_0) = 1$, the CSP obtained using $\hat{\pi}$ differs from that obtained by π_0 by a factor of 10, on average, for a randomly sampled k-locus n-haplotype conditional configuration.

Experiment 1: High mutation and recombination rate

For the first experiment, conditional haplotype configurations were simulated using method M1, setting $\theta_0 = 1$ and $\rho_0 = 4$. Biologically, $\theta_0 = 1$ corresponds to a very high mutation rate; though such rates can occur in retroviruses (McVean et al., 2002), our primary objective in this experiment is directly assessing the accuracy of CSDs $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{SMC}}$ for a small number $k \leq 10$ of loci and a small number $n \leq 20$ of haplotypes. The CSP for $\hat{\pi}_{\mathrm{PS}}$ is evaluated directly using the recursion, and the CSP for $\hat{\pi}_{\mathrm{SMC}}$ is evaluated using the identity $\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{NC}}$ and the recursion for $\hat{\pi}_{\mathrm{NC}}$. The true CSD π is used as the reference, and the associated CSP estimated using importance sampling.

We examine the accuracy $\mathrm{ALRErr}_{k,n}(\cdot | \pi)$ as function of the number of loci k and the number of haplotypes n, for each of the CSDs $\hat{\pi}_{\mathrm{PS}}$, $\hat{\pi}_{\mathrm{SMC}}$, $\hat{\pi}_{\mathrm{FD}}$, and $\hat{\pi}_{\mathrm{LS}}$. The results are plotted in Figure 4.1(a) and Figure 4.1(b), respectively. In this setting, the accuracy of the approximate CSDs $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{SMC}}$ are nearly identical, and considerably better than both $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$. We remark that these results

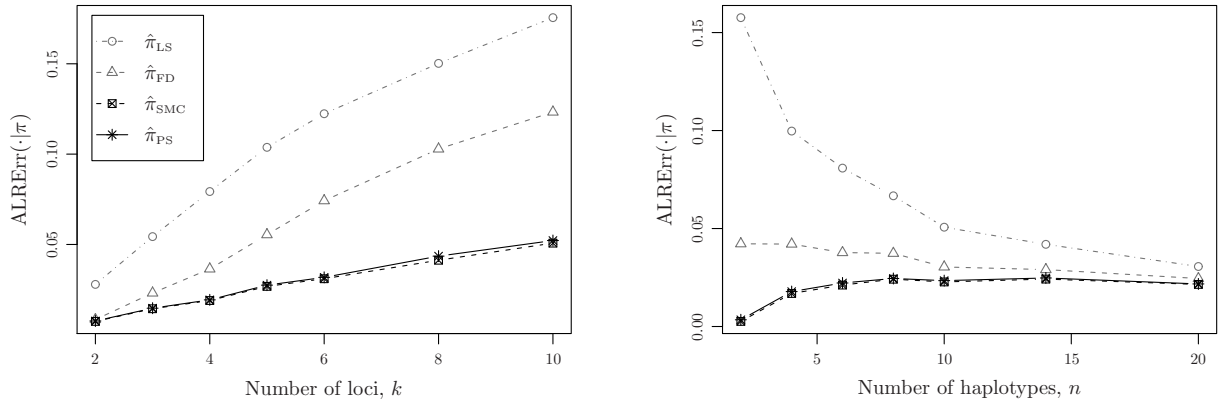


Figure 4.1. Absolute log-ratio (ALR) error (4.1) for data simulated using method M1 with $\theta_0 = 1$ and $\rho_0 = 4$. The error $\text{ALRErr}_{k,n}(\cdot|\pi)$ is evaluated as a function of the number of loci k and the number of haplotypes n for the approximate CSDs $\hat{\pi}_{\text{PS}}$, $\hat{\pi}_{\text{SMC}}$, $\hat{\pi}_{\text{FD}}$, and $\hat{\pi}_{\text{LS}}$, relative to the true CSD π. The accuracies of $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{SMC}}$ are comparable, and considerably better than both $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$. For each datapoint, $N = 250$ conditional configurations were simulated, (a) $2 \leq k \leq 10$, $n = 6$. (b) $k = 4$, $2 \leq n \leq 20$.

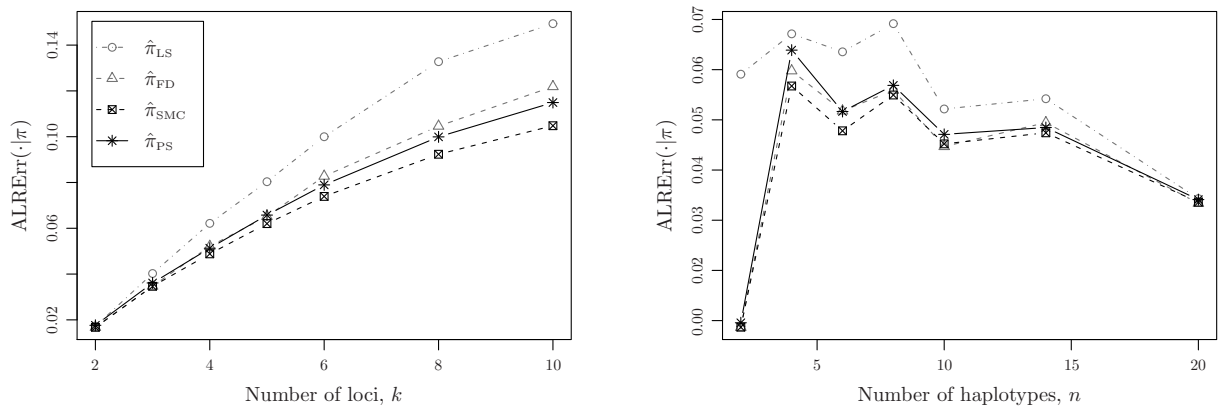
are obtained by averaging over $N = 250$ configurations, and do not imply that the CSPs produced by $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{SMC}}$ are always more accurate than those produced by $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$.

All of the approximate CSDs become less accurate as the number of loci increases. Importantly, however, the improvement in accuracy observed for CSDs $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{SMC}}$, relative to $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$, is amplified for larger numbers of loci; this result may have significant consequence at a genomic scale, in which many thousands of segregating loci are considered. In contrast, the accuracy of the CSDs converge as the number of haplotypes n increases. Recall from Section 1.4 that in the limit $n \to \infty$ the true CSD is described by a sample taken uniformly at random from the previously-observed haplotypes; all of the approximate CSDs we consider exhibit the correct behavior in this limit, accounting for their convergence with one another. As the number of haplotypes decreases, $\hat{\pi}_{\text{LS}}$ becomes less accurate, while $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{SMC}}$ become more accurate, providing further evidence that the true CSD is modeled more accurately by our proposed CSDs.

Experiment 2: Biologically realistic SNP data

For the second experiment, conditional haplotype configurations were simulated using method M2, setting $\theta_0 = 0.01$ and $\rho_0 = 0.1$. Biologically, $\theta_0 = 0.01$ is a moderate mutation rate, so that the sampled configurations represent realistic SNP data. As before, we assess the accuracy of CSDs $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{SMC}}$ for a small number $k \leq 10$ of loci and a small number $n \leq 20$ of haplotypes, using the true CSD π as the reference.

As in the previous experiment, we examine the accuracy $\text{ALRErr}_{k,n}(\cdot|\pi)$ as function of the number of loci k and the number of haplotypes n, for each of the CSDs $\hat{\pi}_{\text{PS}}$, $\hat{\pi}_{\text{SMC}}$, $\hat{\pi}_{\text{FD}}$, and $\hat{\pi}_{\text{LS}}$. The results are plotted in Figure 4.2(a) and Figure 4.2(b), respectively. The approximate CSDs $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{SMC}}$ are, on average, more accurate than the approximate CSDs $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$. The differences in accuracy, however, are less pronounced than in the previous experiment; quantifying the precise



(a) (b)

Figure 4.2. Absolute log-ratio (ALR) error (4.1) for data simulated using method M2 with $\theta_0 = 0.01$ and $\rho_0 = 0.1$. The error $\mathrm{ALRErr}_{k,n}(\cdot|\pi)$ is evaluated as a function of the number of loci k and the number of haplotypes n for the approximate CSDs $\hat{\pi}_{\mathrm{PS}}$, $\hat{\pi}_{\mathrm{SMC}}$, $\hat{\pi}_{\mathrm{FD}}$, and $\hat{\pi}_{\mathrm{LS}}$, relative to the true CSD π. Compared to Figure 4.1, the differences in accuracy are less pronounced, but still $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{SMC}}$ show an improvement relative to $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$. For each datapoint, $N = 250$ conditional configurations were simulated, (a) $2 \leq k \leq 10$, $n = 6$. (b) $k = 4$, $2 \leq n \leq 20$.

cause and degree of this effect remains an open problem, and requires further theoretical and empirical investigation.

In concordance with the previous experiment, all of the CSDs become less accurate as the number of loci increases. Observe that $\hat{\pi}_{\mathrm{SMC}}$ is more accurate than $\hat{\pi}_{\mathrm{PS}}$, a surprising result because the CSD $\hat{\pi}_{\mathrm{SMC}}$ is itself an approximation of $\hat{\pi}_{\mathrm{PS}}$. Preliminary investigation (data not shown) suggests that this effect is *local*, and does not persist for larger numbers of loci k; once again, this hypothesis requires further investigation. Finally, as the number of haplotypes in the conditional configuration increases, the accuracy of the different CSDs converge; for small numbers of haplotypes $\hat{\pi}_{\mathrm{LS}}$ is less accurate than $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{SMC}}$, though the difference is once again less pronounced.

Experiment 3: The effect of discretization

In the third experiment, we investigate the effect of discretization on accuracy, particularly as the number of loci k increases. Denote by $\hat{\pi}_{\mathrm{SMC}(m)}$ the CSD resulting from the discretization $\mathcal{P}$ comprising $|\mathcal{P}| = m$ intervals, produced using the Gaussian quadrature method described in Section 3.2. For comparison, we include the CSDs $\hat{\pi}_{\mathrm{FD}}$ and $\hat{\pi}_{\mathrm{LS}}$, and the CSD $\hat{\pi}_{\text{NC-A}}$, described in Section 3.1.2, setting $\hat{\pi}_{\mathrm{Alt}} = \hat{\pi}_{\mathrm{SMC}(16)}$. Requisite conditional haplotype configurations were simulated using method M2, setting $\theta_0 = 0.01$ and $\rho_0 = 0.05$.

For $k > 10$ loci, it is computationally impracticable to estimate the CSP associated with the true CSD π; it is similarly difficult to directly evaluate the CSP associated with the CSD $\hat{\pi}_{\mathrm{PS}}$. We therefore use $\hat{\pi}_{\mathrm{SMC}}$ as the reference CSD, evaluating the CSP using the identity $\hat{\pi}_{\mathrm{SMC}} = \hat{\pi}_{\mathrm{NC}}$ and the recursion for $\hat{\pi}_{\mathrm{NC}}$. We examine the accuracy $\mathrm{ALRErr}_{k,n}(\cdot|\hat{\pi}_{\mathrm{SMC}})$ as a function of the number of loci, for $n = 10$ haplotypes and $k \leq 20$ loci. The results are plotted in Figure 4.3. Observe that $\hat{\pi}_{\mathrm{SMC}(m)}$ approximates $\hat{\pi}_{\mathrm{SMC}}$ closely, with the fidelity of the approximation increasing with the number of intervals m in the discretization. The approximation $\hat{\pi}_{\text{NC-A}}$ is indistinguishable from

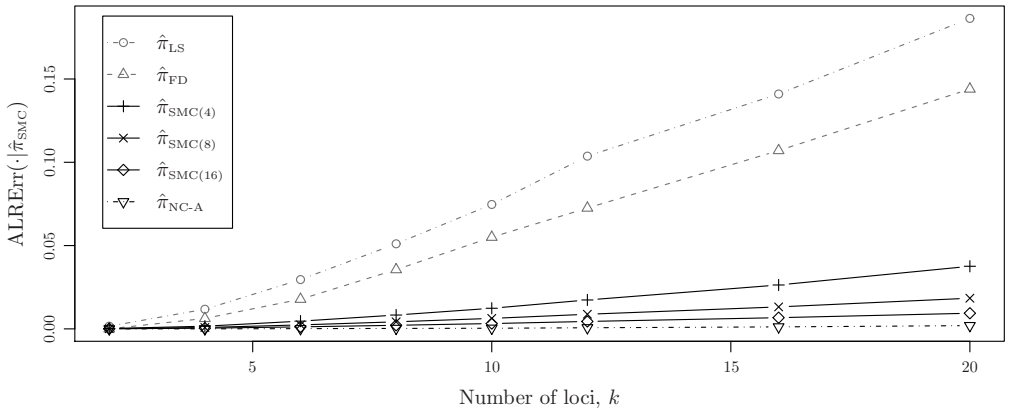


Figure 4.3. Absolute log-ratio (ALR) error (4.1) for data simulated using method M2 with $\theta_0 = 0.01$ and $\rho_0 = 0.05$. The error $\text{ALRErr}_{k,n}(\cdot|\hat{\pi}_{\text{SMC}})$ is evaluated for $n = 10$ haplotypes, and as a function of the number of loci k for the approximate CSDs $\hat{\pi}_{\text{SMC}(m)}$, $\hat{\pi}_{\text{NC-A}}$, $\hat{\pi}_{\text{FD}}$, and $\hat{\pi}_{\text{LS}}$, relative to $\hat{\pi}_{\text{SMC}}$. The CSD $\hat{\pi}_{\text{SMC}(m)}$ approximates $\hat{\pi}_{\text{SMC}}$ very well, and produces more accurate result than $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$. For each datapoint, $N = 250$ conditional configurations were simulated.

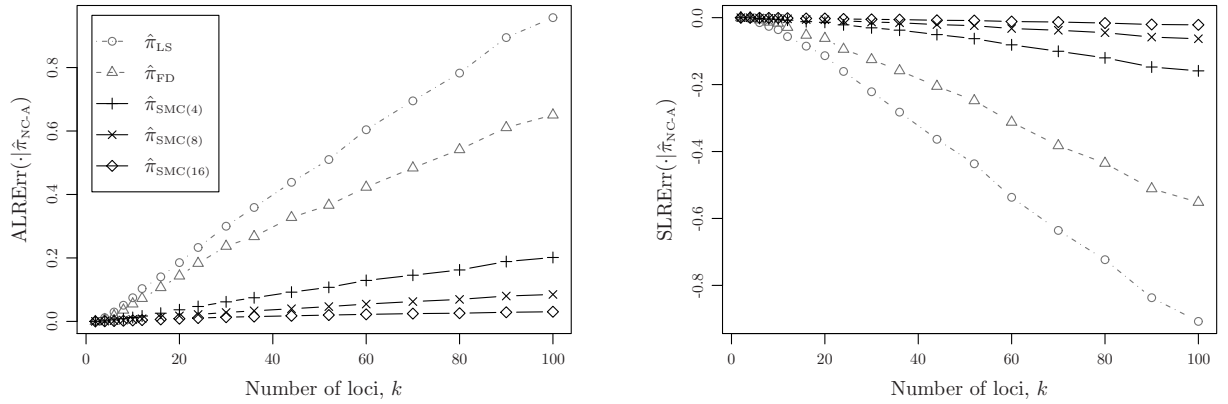
$\hat{\pi}_{\text{SMC}}$. Moreover, as k increases, both $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$ continue to diverge from $\hat{\pi}_{\text{SMC}}$, suggesting the disparity in accuracy, directly observed in the previous experiments, increases for larger value of k. We tentatively conclude that, even for small values of m, the CSD $\hat{\pi}_{\text{SMC}(m)}$ is substantially more accurate than both $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$.

For $k > 20$ loci, it becomes computationally impracticable to evaluate the CSP associated with $\hat{\pi}_{\text{SMC}}$. In Figure 4.4, we observed that the CSD $\hat{\pi}_{\text{NC-A}}$ is nearly indistinguishable from $\hat{\pi}_{\text{SMC}}$; we therefore use $\hat{\pi}_{\text{NC-A}}$ as the reference CSD. Once again, we examine the accuracy $\text{ALRErr}_{k,n}(\cdot|\hat{\pi}_{\text{NC-A}})$, and the analogously-defined signed log-ratio (SLR) error $\text{SLRErr}_{k,n}(\cdot|\hat{\pi}_{\text{NC-A}})$ as a function of the number of loci, for $n = 10$ haplotypes and $k \leq 100$ loci. The results are plotted in Figures 4.4(a) and 4.4(b). The trends observed in Figure 4.4 are recapitulated in 4.4(a), suggesting that they continue to hold for substantially larger values of k. Interestingly, 4.4(b) shows that $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$ produce CSPs that are, on average, smaller than $\hat{\pi}_{\text{NC-A}}$ (and $\hat{\pi}_{\text{SMC}}$); for example, $\hat{\pi}_{\text{LS}}$ produces values that are, on average, a factor of 10 smaller than $\hat{\pi}_{\text{NC-A}}$ for $k = 100$. In conjunction with our conclusion that $\hat{\pi}_{\text{SMC}}$ is more accurate than $\hat{\pi}_{\text{LS}}$ and $\hat{\pi}_{\text{FD}}$, this suggests a similar systematic error with respect to the true CSD.

4.1.3 Timing

We next empirically investigate the running time required to evaluate each of the CSPs. The CSDs $\hat{\pi}_{\text{SMC}}$ and $\hat{\pi}_{\text{NC-A}}$ are computed using the algorithms provided in Section 3.1. For the moment, we restrict attention to computing $\hat{\pi}_{\text{SMC}(m)}$ using Algorithm 2, the baseline algorithm described in Section 3.3.1; $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$ are computed using the analogous dynamic programming algorithms provided in Fearnhead and Donnelly (2001) and Li and Stephens (2003) and the associated released software. In Table 4.1, we present the timing results for conditional configurations generated using simulation method M2, setting $\theta_0 = 0.01$ and $\rho_0 = 0.05$, with $n = 10$ haplotypes and $k \leq 100$ loci.

Looking across each row, it is evident that the running time under $\hat{\pi}_{\text{SMC}(m)}$, $\hat{\pi}_{\text{FD}}$, and $\hat{\pi}_{\text{LS}}$ depends linearly on the number of loci k, matching the asymptotic time complexity. Similarly, the running



(a) (b)

Figure 4.4. Log-ratio error for data simulated using method M2 with $\theta_0 = 0.01$ and $\rho_0 = 0.05$. The error is evaluated for $n = 10$ haplotypes, and as a function of the number of loci k for the approximate CSDs $\hat{\pi}_{\mathrm{SMC}(m)}$, $\hat{\pi}_{\mathrm{FD}}$, and $\hat{\pi}_{\mathrm{LS}}$, relative to $\hat{\pi}_{\mathrm{NC\text{-}A}}$. The improvement in accuracy of $\hat{\pi}_{\mathrm{SMC}(m)}$ over $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$ is amplified as the number of loci k increases; moreover, both $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$ produce significantly smaller values than $\hat{\pi}_{\mathrm{NC\text{-}A}}$ (and $\hat{\pi}_{\mathrm{SMC}}$). For each datapoint, $N = 250$ conditional configurations were simulated. (a) The absolute log-ratio error $\mathrm{ALRErr}_{k,n}(\cdot|\hat{\pi}_{\mathrm{NC\text{-}A}})$. (b) The signed log-ratio error $\mathrm{SLRErr}_{k,n}(\cdot|\hat{\pi}_{\mathrm{NC\text{-}A}})$.

time under $\hat{\pi}_{\mathrm{NC\text{-}A}}$ is well-matched by the theoretical cubic dependence on k. Comparing $\hat{\pi}_{\mathrm{SMC}(m)}$, $\hat{\pi}_{\mathrm{FD}}$, and $\hat{\pi}_{\mathrm{LS}}$, observe that the running time for $\hat{\pi}_{\mathrm{SMC}(4)}$ is approximately a factor of 10 slower than $\hat{\pi}_{\mathrm{LS}}$, and approximately a factor of 2 slower than $\hat{\pi}_{\mathrm{FD}}$. Similarly, $\hat{\pi}_{\mathrm{SMC}(8)}$ is approximately a factor of 20 and 4 slower than $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$, respectively; and $\hat{\pi}_{\mathrm{SMC}(16)}$ is approximately a factor of 40 and 8 slower than $\hat{\pi}_{\mathrm{LS}}$ and $\hat{\pi}_{\mathrm{FD}}$, respectively. Importantly, these factors are *constant* in the number of loci k. Also note that the time required to compute the CSD for $\hat{\pi}_{\mathrm{SMC}(m)}$ appears to depend linearly, rather than quadratically, on the number of discretization intervals m for the values considered.

Finally, we assess the speed-up obtained by using the optimized algorithms for computing $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$ described in Section 3.3. Recall that our optimizations are realized in Algorithms 3–5, each of which relies on a partition $\mathcal{C}$ of the haplotype configuration $\mathbf{n}$. We have characterized optimal such partitions, and proposed a simple and fast method for constructing good partitions $\mathcal{C} = C^*$. For the sake of comparison, we also consider the trivial partition $\mathcal{C} = C_{\mathrm{T}}$. Relative to Algorithm 3, Algorithms 4 and 5 represent successive improvements in efficiency for non-polymorphic loci. Finally, recall that setting $\mathcal{C} = C_{\mathrm{T}}$ in Algorithm 3 is equivalent to Algorithm 2, applied above.

The optimized algorithms, along with their asymptotic time complexities, are summarized in Table 3.1. For a fixed number of haplotypes n, and assuming coarse homogeneity across the genome, the running times of each of these algorithms is asymptotically linear in the number of loci. We are interested in determining the constants associated with this linear behavior for each algorithm. Note, however, that for the cases when $\mathcal{C} = C_{\mathrm{T}}$, the time complexities do not depend on n directly, but rather the number of unique haplotypes n_{u}. For a particular value of n, the quantity n_{u} will increase with the number of loci under consideration until $n_{\mathrm{u}} = n$; only at this point do the running times become linear in the number of loci. A similar argument can be made for a more general configuration partition $\mathcal{C}$. In order to attain and analyze the linear behavior for the modestly-sized

Method	Complexity	Number of Loci			
		$k = 10$	$k = 20$	$k = 60$	$k = 100$
$\hat{\pi}_{\text{SMC}} = \hat{\pi}_{\text{NC}}$	$O(c^k \cdot n)$	6.4×10^0	4.8×10^4	NA	NA
$\hat{\pi}_{\text{NC-A}}$	$O(k^3 \cdot n)$	2.9×10^0	2.3×10^1	5.6×10^2	2.5×10^3
$\hat{\pi}_{\text{SMC(16)}}$	$O(k \cdot (nm + m^2))$	1.0×10^{-1}	2.1×10^{-1}	6.1×10^{-1}	1.0×10^0
$\hat{\pi}_{\text{SMC(8)}}$	$O(k \cdot (nm + m^2))$	4.6×10^{-2}	9.6×10^{-2}	3.0×10^{-1}	4.7×10^{-1}
$\hat{\pi}_{\text{SMC(4)}}$	$O(k \cdot (nm + m^2))$	2.3×10^{-2}	5.1×10^{-2}	1.6×10^{-1}	2.8×10^{-1}
$\hat{\pi}_{\text{FD}}$	$O(k \cdot n)$	1.1×10^{-2}	2.7×10^{-2}	7.7×10^{-2}	1.3×10^{-1}
$\hat{\pi}_{\text{LS}}$	$O(k \cdot n)$	2.1×10^{-3}	4.6×10^{-3}	1.5×10^{-2}	2.5×10^{-2}

Table 4.1. Asymptotic time complexity and empirically observed average running time. The second column shows asymptotic time complexity (with the value c indicating an unknown constant) and the last four columns show empirically observed average running time (in milliseconds) required to compute the CSP under various CSDs, for $n = 10$ and the number of loci k as specified within the table; "NA" indicates that the computation could not be completed within a reasonable amount of time. Results were obtained on a single core of a MacPro with dual quad-core 3.0GHz Xeon CPUs.

Method	Implementation	Number of Haplotypes		
		$n = 100$	$n = 2000$	$n = 5000$
$\hat{\pi}_{\text{SMC(16)}}$	Algorithm 3, $\mathcal{C} = C_{\text{T}}$	45 (1.0×)	870 (1.0×)	2153 (1.0×)
$\hat{\pi}_{\text{SMC(16)}}$	Algorithm 4, $\mathcal{C} = C_{\text{T}}$	3.5 (13×)	21 (41×)	54 (40×)
$\hat{\pi}_{\text{SMC(16)}}$	Algorithm 5, $\mathcal{C} = C_{\text{T}}$	0.63 (71×)	18 (48×)	49 (44×)
$\hat{\pi}_{\text{SMC(16)}}$	Algorithm 3, $\mathcal{C} = C^*$	3.8 (12×)	7.8 (110×)	10.3 (208×)
$\hat{\pi}_{\text{SMC(16)}}$	Algorithm 4, $\mathcal{C} = C^*$	3.0 (15×)	3.5 (250×)	3.9 (546×)
$\hat{\pi}_{\text{SMC(16)}}$	Algorithm 5, $\mathcal{C} = C^*$	0.14 (320×)	0.68 (1300×)	1.17 (1845×)
$\hat{\pi}_{\text{FD}}$	Fearnhead and Donnelly (2001)	7.47 (6×)	149 (6×)	367 (6×)
$\hat{\pi}_{\text{LS}}$	Li and Stephens (2003)	1.96 (23×)	39.4 (23×)	96.5 (23×)

Table 4.2. A summary of several key statistics from Figure 4.5. The table indicates the time (in seconds) required to compute the CSP $\hat{\pi}_{\text{SMC}}(\alpha|\mathbf{n})$ for $|\mathbf{n}| = n$, per 1×10^5 loci. The speed-up versus Algorithm 3 with $\mathcal{C} = C_{\text{T}}$, equivalent to the commonly used Algorithm 2, is given in parentheses. See Tables 3.1 and 4.1 for the asymptotic time complexities associated with each algorithm.

configurations that are considered, we formally interpret even non-unique haplotypes to be unique, thereby forcing $n_{\text{u}} = n$.

We use simulation method M1, with $\theta_0 = 0.005$ and $\rho_0 = 0.001$ to produce haplotype configurations with $k = 2 \times 10^5$ loci and $n \leq 5000$ haplotypes, for each of several values of n. We compute the partitions C_{T} and C^*, and subsequently record the running time of each algorithm in computing $\hat{\pi}_{\text{SMC}(m)}(\mathbf{e}_\eta|\mathbf{n})$, for a haplotype η chosen from $\mathbf{n}$ uniformly at random. Throughout, we use a time discretization consisting of $m = 16$ intervals. The running times are plotted, on a logarithmic scale, as a function of n in Figures 4.5(a) and 4.5(b), for $\mathcal{C} = C_{\text{T}}$ and $\mathcal{C} = C^*$, respectively. For comparison, we also include the running times for the CSD $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$, computed as before, using the dynamic programming algorithms provided in Fearnhead and Donnelly (2001) and Li and Stephens (2003) and the associated released software.

From Figure 4.5(a), for which $\mathcal{C} = C_{\text{T}}$, it is clear that our refinements for non-polymorphic loci have practical benefits, as Algorithms 4 and 5 perform substantially better than Algorithm 3, and also better than the standard implementation of $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$. The asymptotic results summarized

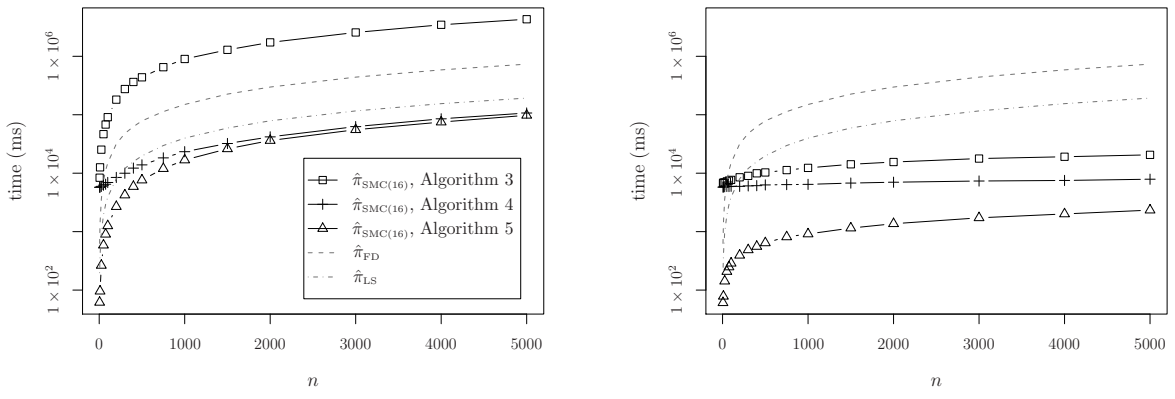


Figure 4.5. Log-scaled plots of the running time (in milliseconds) required to compute $\hat{\pi}_{\mathrm{SMC}(16)}(\mathbf{e}_\eta|\mathbf{n})$ for $\mathbf{n}$ with 2×10^5 loci and $|\mathbf{n}| = n$, as a function n, for each of Algorithms 3–5. The algorithms used to compute $\hat{\pi}_{\mathrm{FD}}$ (Fearnhead and Donnelly, 2001) and $\hat{\pi}_{\mathrm{LS}}$ (Li and Stephens, 2003) are analogous to Algorithm 3 with $\mathcal{C} = C_{\mathrm{T}}$. Configurations were generated using coalescent simulation as described in the text, and results obtained on a single core of a MacPro with dual quad-core 3.0GHz Xeon CPUs. (a) $\mathcal{C} = C_{\mathrm{T}}$, the trivial configuration partition. (b) $\mathcal{C} = C^*$, the configuration partition described in Section 3.3.2.

in Table 3.1 suggest the running time of Algorithm 5 is a factor of k/k_{p} faster than Algorithm 3. This factor is roughly reflected in the logarithmic plot of Figure 4.5(a) as a vertical shift, with deviations occurring because k_{p} increases (slowly) with n. Similarly, as n increases, the asymptotic results indicate that computation is dominated by the $O(k_{\mathrm{p}}mn_{\mathrm{u}})$ term for both Algorithms 4 and 5; this is reflected in Figure 4.5(a) by a near identity in running times for these algorithms for larger values of n.

Comparing Figure 4.5(b) to Figure 4.5(a), the benefits of taking $\mathcal{C} = C^*$ can be observed. For each algorithm, this optimization improves performance substantially, particularly as the number of haplotypes n increases. Given the results for Algorithm 4 in particular, it is clear that the key quantity $\Psi_{\mathrm{p}}(\mathcal{C}) + \Omega(\mathcal{C})$, taken from Table 3.1, increases more slowly with n for $\mathcal{C} = C^*$ than for $\mathcal{C} = C_{\mathrm{T}}$. Finally, as in the previous case, the asymptotic results for general $\mathcal{C}$ indicate that computation is dominated by the $O(m(\Psi_{\mathrm{p}}(\mathcal{C}) + \Omega(\mathcal{C})))$ term for both Algorithms 4 and 5; the associated convergence of running times appears to be occurring in Figure 4.5(b), though more slowly than in Figure 4.5(a); thus, Algorithm 5 is a practically useful alternative to Algorithm 4, even for larger values of n.

Though general trends are clear from Figure 4.5, the logarithmic scale makes it difficult to appreciate the magnitude of the effects of the optimizations. As mentioned earlier, assuming rough homogeneity over the genome, the computation time increases linearly with the number of loci. In Table 4.2, we summarize the constant associated with this linear behavior as the time required to process 10^5 loci, along with the speed-up relative to the baseline, Algorithm 2 for $\hat{\pi}_{\mathrm{SMC}(16)}$. Observe that Algorithm 4, with $\mathcal{C} = C^*$, which can be applied in complete generality, provides a speed-up of 15×, 250×, 546× for sample sizes $n = 100$, $n = 2000$, and $n = 5000$, respectively; and in most cases, Algorithm 5 can be applied, which increases these speed-ups to $320\times, 1300\times$, and $1845\times$, respectively. Importantly, the speed-up increases with the number of haplotypes n; moreover, even

for modest values of n, the optimized algorithms provide a substantial speed-up relative to standard implementations of $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$.

4.2 Importance Sampling

In this section, we return to the problem of computing the sampling probability associated with a haplotype configuration. Because exact evaluation of the sampling probability is generally impracticable, we consider a Monte Carlo method, importance sampling (Liu, 2008). Compared to naive Monte Carlo, importance sampling (IS) seeks to minimize the variance of the estimator by judicious choice of a *proposal distribution*. In the context of computing the sampling probability, Stephens and Donnelly (2000) showed that the optimal such proposal distribution can be expressed in terms of the true CSD; using an approximate surrogate CSD then results in a sub-optimal, but still reasonable, proposal distribution. We introduce the practical CSD-based approach to IS in the presence of recombination described by Fearnhead and Donnelly (2001), and propose two optimizations to improve efficiency.

4.2.1 IS Motivation

Let $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$ be a haplotype configuration. As described in Chapter 1, the ordered sampling probability $q(\mathbf{n})$ can be exactly evaluated by constructing and either numerically or algebraically solving a finite set of coupled linear equations. However, the number of equations in the system grows super-exponentially with the number of loci and the number of haplotypes of the configuration $\mathbf{n}$, limiting the practical applicability of this method to configurations with $k \leq 5$ loci and $n \leq 5$ haplotypes. Thus, in order to evaluate $q(\mathbf{n})$ for larger haplotype configurations, we appeal to Monte Carlo methods. Let $\hat{n}$ be the untyped configuration associated with $\mathbf{n}$, and recall from Section 1.3 that a typed history $\mathcal{F}$ for $\hat{n}$ is given by

$$\mathcal{F} = (v_0, e_1, v_1, \ldots, e_\tau, v_\tau), \tag{4.2}$$

where v_i is the *typed* configuration after the i-th genealogical event e_i, and the untyped configuration associated with v_0 is $\hat{n}$. Moreover, such a typed history can be sampled directly using the coalescent process, and we denote the corresponding density by $p(\cdot|\hat{n})$. We can therefore partition with respect to the typed history to obtain the following expression for $q(\mathbf{n})$,

$$q(\mathbf{n}) = \int p(\mathbf{n}|\mathcal{F}) p(\mathcal{F}|\hat{n}) d\mathcal{F} \approx \frac{1}{M} \sum_{j=1}^{M} p(\mathbf{n}|\mathcal{F}^{(j)}), \tag{4.3}$$

where $p(\mathbf{n}|\mathcal{F}) = 1$ if $v_0 = \mathbf{n}$ and 0 otherwise. The latter Monte Carlo approximation then assumes that the typed histories $\{\mathcal{F}^{(j)}\}_{j=1,\ldots,M}$ are sampled independently from the coalescent process, with density $p(\cdot|\hat{n})$. In practice, even for modestly-sized configuration $\mathbf{n}$, the probability that $p(\mathbf{n}|\mathcal{F}^{(j)}) = 1$ for a randomly sampled history $\mathcal{F}^{(j)}$ is very small, and in order to obtain an estimator with acceptably low variance, the number of sampled histories M must be impracticably large.

IS attempts to improve the Monte Carlo estimator by biasing the sampled histories toward regions of high probability. Formally, introduce an alternative *proposal distribution* on histories, with associated density $q(\cdot|\hat{n})$, and with support including $\{\mathcal{F} : p(\mathbf{n}|\mathcal{F}) > 0\}$. Then (4.3) can be

expressed,

$$q(\mathbf{n}) = \int p(\mathbf{n}|\mathcal{F})\frac{p(\mathcal{F}|\hat{n})}{q(\mathcal{F}|\hat{n})}q(\mathcal{F}|\hat{n})d\mathcal{F} \approx \frac{1}{M}\sum_{j=1}^{M}\underbrace{p(\mathbf{n}|\mathcal{F}^{(j)})\frac{p(\mathcal{F}^{(j)}|\hat{n})}{q(\mathcal{F}^{(j)}|\hat{n})}}_{w^{(j)}} = \frac{1}{M}\sum_{j=1}^{M} w^{(j)}, \tag{4.4}$$

where the typed histories $\{\mathcal{F}^{(j)}\}_{j=1,\ldots,M}$ are sampled independently from the proposal distribution, and $\{w^{(j)}\}_{j=1,\ldots,M}$ are the associated *importance weights*. Note that the proposal distribution $q(\cdot|\hat{n})$ may explicitly depend on the configuration $\mathbf{n}$.

4.2.2 Optimal proposal distribution

An optimal proposal distribution minimizes the variance of the resulting estimator (4.4), or equivalently, the variance of the importance weights $\{w^{(j)}\}_{j=1,\ldots,M}$. Setting the proposal distribution equal to the *posterior* distribution on typed histories, with density $p(\cdot|\mathbf{n})$ immediately yields, for the importance weight $w^{(j)}$,

$$w^{(j)} = p(\mathbf{n}|\mathcal{F}^{(j)})\frac{p(\mathcal{F}^{(j)}|\hat{n})}{q(\mathcal{F}^{(j)}|\hat{n})} = p(\mathbf{n}|\mathcal{F}^{(j)})\frac{p(\mathcal{F}^{(j)}|\hat{n})}{p(\mathcal{F}^{(j)}|\mathbf{n})} = p(\mathbf{n}|\hat{n}) = q(\mathbf{n}), \tag{4.5}$$

where the penultimate equality is by Bayes' Law, and the final equality by definition. Because the resulting importance weight does not depend on the sampled history $\mathcal{F}^{(j)}$, the variance of the importance weights is 0, and a single sample is required to determine the ordered sampling probability $q(\mathbf{n})$. Thus, the optimal proposal distribution is precisely the posterior distribution.

Though obtaining the posterior distribution and density is generally as difficult as the problem of evaluating the sampling probability, Stephens and Donnelly (2000) observe that the posterior sequence of events and typed configurations $\{(e_i, v_i)\}_{i=1,\ldots,\tau}$ is Markov backward in time, and the posterior density therefore admits the decomposition,

$$p(\mathcal{F}|\mathbf{n}) = p(e_1, v_1|v_0)p(e_2, v_2|v_1)\cdots p(e_\tau, v_\tau|v_{\tau-1}) = \prod_{i=1}^{\tau} p(e_i, v_i|v_{i-1}), \tag{4.6}$$

where $v_0 = \mathbf{n}$. The stated Markov property is evident from the construction of Section 1.3.1, and in particular the graphical model representation of Figure 1.4. Moreover, using the same construction, in conjunction with Bayes' Law, it is possible to derive the following expression for the Markov posterior transition density,

$$p(e_i, v_i|v_{i-1}) = p(v_{i-1}|e_i, v_i)p(e_i|u_{i-1}) \cdot \frac{q(v_i)}{q(v_{i-1})}. \tag{4.7}$$

Recall that the first two factors of the final expression are specified directly by the genealogical process, and are explicitly provided for the coalescent with recombination in Section 1.3.2. Moreover, recalling the definition (1.61) of the CSP, the ratio of ordered sampling probabilities can generally be written as a ratio of CSPs. For the genealogical process described in Section 1.3.2,

$$\frac{q(v_i)}{q(v_{i-1})} = \begin{cases} \frac{1}{\pi(\mathbf{e}_h|v_{i-1}-\mathbf{e}_h)}, & \text{for } v_i = v_{i-1} - \mathbf{e}_h, \\ \frac{\pi(\mathbf{e}_{\mathcal{M}^a_\ell(h)}|v_{i-1}-\mathbf{e}_h)}{\pi(\mathbf{e}_h|v_{i-1}-\mathbf{e}_h)}, & \text{for } v_i = v_{i-1} - \mathbf{e}_h + \mathbf{e}_{\mathcal{M}^a_\ell(h)}, \\ \frac{\pi(\mathbf{e}_{\mathcal{R}_b(h,h')}+\mathbf{e}_{\mathcal{R}_b(h',h)}|v_{i-1}-\mathbf{e}_h)}{\pi(\mathbf{e}_h|v_{i-1}-\mathbf{e}_h)}, & \text{for } v_i = v_{i-1} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h)}. \end{cases} \tag{4.8}$$

The Markov property of the posterior distribution on histories suggests sampling the history starting in the present, with $v_0 = \mathbf{n}$, and proceeding backward in time. At the i-th step, the pair (e_i, v_i) is sampled conditional on v_{i-1}, and this process is iterated until a single haplotype $|v_i| = 1$ remains. Though this optimal method is not realizable, as we can not generally evaluate the true CSP, in the following section we describe the approximations necessary to obtain a practical proposal distribution and IS procedure.

4.2.3 Practical importance sampling

Letting $\hat{\pi}$ be an approximate CSD, and substituting the associated CSP into (4.8) immediately yields a practicable proposal distribution. Before proceeding, however, we revisit the general IS formulation. Motivated by the optimal proposal distribution, we hereafter consider proposal distributions that exhibit the corresponding Markov property,

$$q(\mathcal{F}|\hat{n}) = \prod_{i=1}^{\tau} q(e_i, v_i|v_{i-1}), \tag{4.9}$$

where $v_0 = \mathbf{n}$, and

$$q(e_i, v_i|v_{i-1}) \propto p(v_{i-1}|e_i, v_i)p(e_i|u_{i-1}) \cdot \frac{\hat{q}(v_i)}{\hat{q}(v_{i-1})}. \tag{4.10}$$

Observe that we have replaced the ratio of ordered sampling probabilities with a ratio of *approximate* ordered sampling probabilities, to be computed using an approximate CSD; the proportionality results from this approximation. By construction, any history $\mathcal{F}$ obtained from such a distribution has $p(\mathbf{n}|\mathcal{F}) = 1$. Moreover, the density $p(\cdot|\hat{n})$ associated with the prior distribution of histories can be similarly decomposed using the Markov construction of the coalescent,

$$p(\mathcal{F}|\hat{n}) = \left[\prod_{i=1}^{\tau} p(e_i|u_{i-1})\right]\left[\prod_{i=1}^{\tau} p(v_{i-1}|e_i, v_i)\right]p(v_\tau) = p(v_\tau)\prod_{i=1}^{\tau} p(e_i|u_{i-1})p(v_{i-1}|e_i, v_i), \tag{4.11}$$

where $u_0 = \hat{n}$ and u_i is the untyped configuration associated with v_i. As a consequence of (4.9),(4.10), and (4.11), the importance weight w associated with the history $\mathcal{F}$ can be written

$$w = \frac{p(\mathcal{F}|\hat{n})}{q(\mathcal{F}|\hat{n})} = p(v_\tau)\prod_{i=1}^{\tau} \underbrace{c_{i-1} \cdot \frac{\hat{q}(v_{i-1})}{\hat{q}(v_i)}}_{w_i} = p(v_\tau) \cdot \prod_{i=1}^{\tau} w_i, \tag{4.12}$$

where c_i is the constant of proportionality associated with v_{i-1} in (4.10). Thus, as the history $\mathcal{F}$ is sampled, starting in the present and proceeding backward in time, the corresponding importance weight w can be multiplicatively updated. This formulation is an example of *sequential importance sampling* (SIS), for which both the sample and importance weight are constructed sequentially (Liu, 2008); we remark that Jenkins (2012) has explored advanced SIS techniques, including resampling, for coalescent models.

Finally, we consider the space of histories from which $\mathcal{F}$ is sampled. Recall the *reduced* coalescent with recombination, introduced in Section 1.3.2, for which the genealogical history of non-ancestral loci is not explicitly constructed; in sampling a haplotype configuration, such non-ancestral loci can then be left unspecified. Using the reduced model, the space of histories is dramatically reduced,

$e_i \in \mathcal{E}(u_{i-1})$	Lineage(s)	v_i	$p(v_{i-1}\|e_i, v_i)$ $\times p(e_i\|u_{i-1})$	$\hat{q}(v_i)/\hat{q}(v_{i-1})$
Coalescence I	$g \in \mathcal{G}$	$v_{i-1} - \mathbf{e}_g$	$\frac{2}{\mathcal{N}}$	$\frac{1}{\hat{\pi}(\mathbf{e}_g\|v_{i-1}-\mathbf{e}_g)}$
Coalescence II	$g, g' \in \mathcal{G} : g \nless g'$	$v_{i-1} - \mathbf{e}_g - \mathbf{e}_{g'}$ $+ \mathbf{e}_{\mathcal{C}(g,g')}$	$\frac{2}{\mathcal{N}}$	$\frac{\hat{\pi}(\mathbf{e}_{\mathcal{C}(g,g')}\|v_{i-1}-\mathbf{e}_g-\mathbf{e}_{g'})}{\hat{\pi}(\mathbf{e}_g+\mathbf{e}_{g'}\|v_{i-1}-\mathbf{e}_g-\mathbf{e}_{g'})}$
Mutation, $\ell \in L$	$g \in \mathcal{G} : \ell \in L(g)$	$v_{i-1} - \mathbf{e}_g$ $+ \mathbf{e}_{\mathcal{M}^a_\ell(g)}$	$\Phi^{(\ell)}_{a,g[\ell]} \cdot \frac{\theta_\ell}{\mathcal{N}}$	$\frac{\hat{\pi}(\mathbf{e}_{\mathcal{M}^a_\ell(g)}\|v_{i-1}-\mathbf{e}_g)}{\hat{\pi}(\mathbf{e}_g\|v_{i-1}-\mathbf{e}_g)}$
Recombination, $b \in B$	$g \in \mathcal{G} : b \in B(g)$	$v_{i-1} - \mathbf{e}_g$ $+ \mathbf{e}_{\mathcal{R}^-_b(g)} + \mathbf{e}_{\mathcal{R}^+_b(g)}$	$\frac{\rho_b}{\mathcal{N}}$	$\frac{\hat{\pi}(\mathbf{e}_{\mathcal{R}^-_b(g)}+\mathbf{e}_{\mathcal{R}^+_b(g)}\|v_{i-1}-\mathbf{e}_g)}{\hat{\pi}(\mathbf{e}_g\|v_{i-1}-\mathbf{e}_g)}$

Table 4.3. Specification of the proposal transition density for each event. Let v_{i-1} be a typed haplotype configuration with associated untyped configuration u_{i-1}. The support of the proposal transition density $q(\cdot|v_{i-1})$ is all pairs (e_i, v_i) such that $e_i \in \mathcal{E}(u_{i-1})$ and $v_{i-1} \in \mathcal{V}(v_i, e_i)$. Each pair is specified in the table, along with explicit forms computing the unnormalized proposal transition probability (4.10) and the incremental importance weight (4.12). Setting $v_{i-1} = \mathbf{n}'$, the normalization constant is given by $\mathcal{N} = \sum_{g\in\mathcal{G}} n'_g\big(n' - 1 + \sum_{\ell\in L(g)} \theta_\ell + \sum_{b\in B(g)} \rho_b\big)$.

providing a considerable improvement in importance sampling efficiency. Specific values associated with both the proposal transition distribution (4.10) and the incremental importance weight (4.12) for the *reduced* coalescent with recombination are tabulated in Table 4.3.

Thus, letting $\hat{\pi}$ be an arbitrary approximate CSD, the expressions in Table 4.3 completely specify the IS procedure. Observe that there is no direct method for sampling from proposal distribution. Instead, it is necessary, at the i-th step, to enumerate all event-configuration pairs (e_i, v_i) in the support of the proposal transition distribution, compute the proposal transition probability for each pair, normalize the resulting probabilities, and finally sample a pair at random according to the normalized probabilities. As the number of event-configuration pairs is large, the selection process represents a substantial computational burden, and Fearnhead and Donnelly (2001) propose the following two-step approximation. First, select a labelled partially-specified haplotype from $v_{i-1} = \mathbf{n}'$ using the *prior* distribution; a haplotype with type $\eta \in \mathcal{G}$ is chosen with probability

$$p(\eta|\mathbf{n}') = \frac{n' - 1 + \sum_{\ell\in L(\eta)} \theta_\ell + \sum_{b\in B(\eta)} \rho_b}{\sum_{g\in\mathcal{G}} n'_g\left(n' - 1 + \sum_{\ell\in L(g)} \theta_\ell + \sum_{b\in B(g)} \rho_b\right)}. \tag{4.13}$$

Following selection of the labelled haplotype, an event-configuration pair is selected conditional on the event incorporating the selected labelled haplotype. The full proposal transition probability is then the product of the haplotype proposal probability and the conditional event-configuration proposal probability; the corresponding incremental importance weight is given by the quotient of the incremental prior and the appropriately normalized two-step proposal transition probability.

Additionally, explicit evaluation of the ratio of CSPs associated with the second class of coalescence events (Coalescence II, in Table 4.3) requires computing the CSP for two conditionally sampled haplotypes. Recalling that $\hat{\pi}_{\text{FD}}$ is only defined for a single conditionally sampled haplotype,

Fearnhead and Donnelly (2001) suggest approximating the ratio as follows,

$$\frac{\hat{\pi}(\mathbf{e}_{\mathcal{C}(g,g')}|v_{i-1}-\mathbf{e}_g-\mathbf{e}_{g'})}{\hat{\pi}(\mathbf{e}_g+\mathbf{e}_{g'}|v_{i-1}-\mathbf{e}_g-\mathbf{e}_{g'})} \approx \frac{\hat{\pi}(\mathbf{e}_{\mathcal{C}(g,g')}|v_{i-1}-\mathbf{e}_g-\mathbf{e}_{g'})}{\hat{\pi}(\mathbf{e}_g|v_{i-1}-\mathbf{e}_g)\hat{\pi}(\mathbf{e}_{g'}|v_{i-1}-\mathbf{e}_g-\mathbf{e}_{g'})}. \tag{4.14}$$

The ratio of CSPs associated with a recombination event can be similarly approximated. Observe, however, that for computationally-tractable CSDs $\hat{\pi}$ making the sequentially Markov assumption, including $\hat{\pi}_{\text{SMC}} = \hat{\pi}_{\text{LC}} = \hat{\pi}_{\text{NC}}$, $\hat{\pi}_{\text{FD}}$, and $\hat{\pi}_{\text{LS}}$, the following identity holds,

$$\frac{\hat{\pi}(\mathbf{e}_{\mathcal{R}_b^-(g)}+\mathbf{e}_{\mathcal{R}_b^+(g)}|v_{i-1}-\mathbf{e}_g)}{\hat{\pi}(\mathbf{e}_g|v_{i-1}-\mathbf{e}_g)} = \frac{\hat{\pi}(\mathbf{e}_{\mathcal{R}_b^-(g)}|v_{i-1}-\mathbf{e}_g)\hat{\pi}(\mathbf{e}_{\mathcal{R}_b^+(g)}|v_{i-1}-\mathbf{e}_g)}{\hat{\pi}(\mathbf{e}_g|v_{i-1}-\mathbf{e}_g)}. \tag{4.15}$$

In conjunction with the two-step proposal transition probability described above, using these expressions provides an efficiently computable proposal transition distribution. Finally, we remark that while we have described the proposal transition distribution in terms of atomic events specifying individual labeled lineages, a practical implementation should aggregate events of the same type, rather than explicitly enumerating them; for example, provided the haplotype chosen in the first step is of type η, the proposal probability of coalescence with any of the $n_\eta - 1$ remaining haplotypes of type η can be computed at once.

4.2.4 Parent independent mutation

Recall from Section 2.2.2 that, provided a PIM model and a mutation event at locus $\ell \in L$, locus ℓ is non-ancestral in the haplotype ancestral to the mutation event. This observation yields a further-reduced recursion (1.26) for the ordered sampling probability $q(\mathbf{n})$, and can also be used to reduce the space of histories for IS. Before describing this improvement, we demonstrate that even a non-PIM model can be *decomposed* into a PIM component and non-PIM component; consider a mutation model with scaled mutation rate θ and stochastic mutation matrix Φ, and define

$$\phi = \sum_{a\in A} \phi_a, \text{ where } \phi_a = \min_{a'\in A} \Phi_{a',a}. \tag{4.16}$$

Further defining the PIM mutation model

$$\theta_{\text{PIM}} = \theta \cdot \phi, \quad \Phi_{\text{PIM}} = (\phi_a/\phi)_{a\in A}, \tag{4.17}$$

and the non-PIM mutation model

$$\theta_{\text{non-PIM}} = \theta \cdot (1-\phi), \quad \Phi_{\text{non-PIM}} = \big((\Phi_{a,a'} - \phi_{a'})/(1-\phi)\big)_{a,a'\in A}, \tag{4.18}$$

it can be verified that the two mutation models jointly produce the same sampling distribution as the original model. Observe that, provided a stochastic mutation matrix Φ that exhibits PIM, the resulting decomposition is trivial, as $\phi = 1$.

In the context of exact CSP evaluation using a recursive expression, such a decomposition of the mutation model has no computational benefit. However, in the context of IS, for which individual histories are constructed, the decomposed mutation model provides a mechanism for sampling histories with reduced complexity with high probability. In particular, we consider two alternative classes of mutation events, one for each of the mutation models in the decomposition; the row associated with the mutation event in Table 4.3 can thus, in complete generality, be replaced by

$e_i \in \mathcal{E}(u_{i-1})$	Lineage(s)	v_i	$p(v_{i-1}\|e_i, v_i)$ $\times p(e_i\|u_{i-1})$	$\hat{q}(v_i)/\hat{q}(v_{i-1})$
Mutation I, $\ell \in L$	$g \in \mathcal{G} : \ell \in L(g)$	$v_{i-1} - \mathbf{e}_g$ $+ \mathbf{e}_{\mathcal{M}_\ell(g)}$	$\frac{\phi_a^{(\ell)}}{\phi^{(\ell)}} \cdot \frac{\theta_\ell \cdot \phi^{(\ell)}}{\mathcal{N}}$	$\frac{\hat{\pi}(\mathbf{e}_{\mathcal{M}_\ell(g)}\|v_{i-1}-\mathbf{e}_g)}{\hat{\pi}(\mathbf{e}_g\|v_{i-1}-\mathbf{e}_g)}$
Mutation II, $\ell \in L$	$g \in \mathcal{G} : \ell \in L(g)$	$v_{i-1} - \mathbf{e}_g$ $+ \mathbf{e}_{\mathcal{M}_\ell^a(g)}$	$\frac{\Phi_{a,g[\ell]}^{(\ell)} - \phi_{g[\ell]}^{(\ell)}}{1-\phi^{(\ell)}} \cdot \frac{\theta_\ell(1-\phi^{(\ell)})}{\mathcal{N}}$	$\frac{\hat{\pi}(\mathbf{e}_{\mathcal{M}_\ell^a(g)}\|v_{i-1}-\mathbf{e}_g)}{\hat{\pi}(\mathbf{e}_g\|v_{i-1}-\mathbf{e}_g)}$

Table 4.4. Modification of the proposal transition densities in Table 4.3 to incorporate two classes of mutation events. As described in the text, the general mutation process can be decomposed into a PIM and non-PIM process; Mutation I events correspond to the PIM process, and Mutation II events to the non-PIM process. The normalization constant $\mathcal{N}$ is identical to that provided in Table 4.3.

the two rows in Table 4.4, resulting in a modified IS procedure. Importantly, the ratio $\hat{q}(v_i)/\hat{q}(v_{i-1})$ is larger for a PIM mutation event (Mutation I) than for a corresponding non-PIM mutation event (Mutation II). Consequently, provided a sufficiently large value of ϕ, many proposed mutation events are PIM, providing a reduction in complexity of the associated history, and a corresponding reduction in the variance of the importance weights.

Finally, recall that it is generally possible to alter the mutation model while retaining the same sampling distribution. Provided the mutation model described above, then for any value of c such that $c \geq \sum_{a' \in A} \Phi_{a,a'}$ for all $a \in A$, the following c-parameterized model produces an identical sampling distribution,

$$\theta_c = c\theta, \quad \Phi_c = (\Phi'_{a,a'})_{a,a' \in A} \text{ where } \Phi'_{a,a'} = \begin{cases} 1 - \frac{1}{c}\sum_{a' \in A} \Phi_{a,a'}, & \text{if } a = a' \\ \frac{1}{c}\Phi_{a,a'}, & \text{otherwise.} \end{cases} \tag{4.19}$$

Thus, the value c can be chosen to maximize the value ϕ associated with PIM mutation in the decomposed model. We have been unable to determine an analytic expression for such a maximizing c, but the value can be obtained using straightforward numerical techniques. We remark, however, that altering the value of θ can adversely affect the efficiency of the IS procedure. Similarly, using a decomposed mutation model requires computing the proposal transition probability associated with additional events, also affecting the efficiency of the IS procedure. In practice, the latter effect is diminished by the algorithmic optimization described below.

4.2.5 Algorithmic optimization

We next consider the computation required to sample each event-configuration pair. Recall that, employing the two-step transition proposal distribution described in Section 4.2.3, a haplotype is first sampled from the prior distribution, and an event-configuration pair is then sampled by enumerating event-configuration pairs incorporating the haplotype, and computing each proposal transition probability. Provided that the sampled haplotype is of type $\eta \in \mathcal{G}$, and assuming that the number of alleles is given by $|A_\ell| = s$ for all $\ell \in L$, the number of PIM and non-PIM mutation events is given by $|L(\eta)| \cdot (s+1)$, and the number of recombination events is given by $|B(\eta)|$. The number of mutation and recombination events is therefore $O(k)$, where k is the number of loci.

Moreover, as indicated in Tables 4.3 and 4.4, computing the proposal transition probability associated with each mutation event requires computing the CSP $\hat{\pi}(\mathbf{e}_{\mathcal{M}_\ell^a(\eta)}|v_{i-1} - \mathbf{e}_\eta)$ or

$\hat{\pi}(\mathbf{e}_{\mathcal{M}_\ell(\eta)}|v_{i-1}-\mathbf{e}_\eta)$, and using (4.15), computing the proposal transition probability associated with each recombination event requires computing the CSPs $\hat{\pi}(\mathbf{e}_{\mathcal{R}_b^-(\eta)}|v_{i-1}-\mathbf{e}_\eta)$ and $\hat{\pi}(\mathbf{e}_{\mathcal{R}_b^+(\eta)}|v_{i-1}-\mathbf{e}_\eta)$. Consequently, using a sequentially-Markov CSD, such as $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$ or $\hat{\pi}_{\mathrm{FD}}$, with time complexity linear in the number of loci k, the overall time complexity associated computing proposal transition probabilities for mutation and recombination events is $O(k^2)$. In practice, this accounts for a substantial proportion of the overall computation.

However, because the conditionally sampled haplotype in each of the requisite CSPs is derived from $\eta \in \mathcal{G}$, there is opportunity to re-use computation. Assuming $\hat{\pi} = \hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$, consider computing and storing the forward and backward values associated with the CSP $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta|\mathbf{n}-\mathbf{e}_\eta)$, which can be accomplished in $O(k)$ time. Then, using properties of the HMM formulation of $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$,

$$\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_{\mathcal{M}_\ell^a(\eta)}|v_{i-1}-\mathbf{e}_\eta) = \sum_{\ddot{s}_\ell \in \ddot{\mathcal{S}}} F_\ell(\ddot{s}_\ell) E_\ell(\ddot{s}_\ell) \cdot \frac{\xi_\ell(a|\ddot{s}_\ell)}{\xi_\ell(\eta[\ell]|\ddot{s}_\ell)}, \tag{4.20}$$

where F_ℓ and E_ℓ are the forward and backward probabilities and ξ_ℓ is the emission density, all at locus $\ell \in L$. Similarly, for recombination, assuming $b = (\ell, \ell+1) \in B$,

$$\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_{\mathcal{R}_b^-(\eta)}|v_{i-1}-\mathbf{e}_\eta) = \sum_{\ddot{s}_\ell \in \ddot{\mathcal{S}}} F_\ell(\ddot{s}_\ell), \tag{4.21}$$

$$\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_{\mathcal{R}_b^+(\eta)}|v_{i-1}-\mathbf{e}_\eta) = \sum_{\ddot{s}_\ell \in \ddot{\mathcal{S}}} \zeta(\ddot{s}_\ell) \cdot E_{\ell-1}(\ddot{s}_\ell), \tag{4.22}$$

where ζ is the marginal density. Thus, each such computation can be accomplished with time complexity $O(|\ddot{\mathcal{S}}|)$, and critically, this is constant in the number of loci. Thus, by pre-computing the forward and backward values for $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta|\mathbf{n}-\mathbf{e}_\eta)$, and using the above method to compute each of the relevant CSPs, the overall time complexity for computing the proposal transition probabilities for mutation and recombination events is $O(k)$.

As will be demonstrated, this optimization confers a practical benefit, and increases the size of samples to which IS can be applied. We also note that the method can, in principle, be used with the algorithmic optimizations for computing $\hat{\pi}_{\mathrm{SMC}}$ detailed in Section 3.3, though typical IS applications involve few haplotypes and few non-polymorphic loci, limiting their utility. Finally, we remark that the CSPs associated with coalescence events, which involve a second haplotype, cannot generally be evaluated using the pre-computed forward and backward values for $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta|\mathbf{n}-\mathbf{e}_\eta)$; we thus leave further optimization and improvement of the IS procedure as a future research direction.

4.2.6 Empirical results

The convergence of the above IS framework for a particular haplotype configuration is often assessed using the *effective sample size* (ESS), defined

$$\mathrm{ESS} = N \cdot \frac{\mu^2}{\mu^2 + \sigma^2} \approx N \cdot \frac{\hat{\mu}^2}{\hat{\mu}^2 + \hat{\sigma}^2}, \tag{4.23}$$

where N is the number of samples drawn from the proposal distribution, and μ and σ^2 are the mean and variance of the corresponding importance weights. Observe that, although the mean μ of the importance weights is the sampling probability for the haplotype configuration, and does not depend on the particular proposal distribution, the variance σ^2 does depend on the proposal

distribution. Importantly, ESS increases monotonically with decreasing variance σ^2, and is therefore a natural measure for comparing the efficiency of proposal distributions; in particular, the optimal proposal distribution has $\sigma^2 = 0$ and $\text{ESS} = N$.

Because the true mean μ and variance σ^2 are unknown, and the ESS is approximated using the sample mean $\hat{\mu}$ and sample variance $\hat{\sigma}^2$. In practice, this approximation makes the ESS difficult to use for assessing convergence, as both the sample mean and variance themselves are random; even for modestly-sized haplotype configurations, we have found that these quantities, particularly the sample variance, converge very slowly, often substantially changing after hundreds of thousands of samples, representing hours or days of runtime. Unfortunately, we are unaware of a resolution to this problem, and our recourse is to use the largest practicable value of N.

Hereafter assuming that N is chosen large enough to obtain an adequate estimate of the true mean μ and variance σ^2, and therefore the true ESS, we are interested in using the ESS to compare the efficiency of the proposal distributions described above. In order to incorporate the computational efficiency of the proposal distribution, we consider the ESS per unit of time,

$$\text{ESSRate} = \frac{\text{ESS}}{t} = \frac{\mu^2}{\mu^2 + \sigma^2} \cdot \frac{N}{t}, \tag{4.24}$$

where t is the time (in seconds) required to draw the N samples from the proposal distribution. The first factor, which we refer to as the per-sample effective sampling rate, depends only on the statistical properties of the proposal distribution. The second factor, which we refer to as the per-second sampling rate, depends on both the statistical properties and the computational efficiency of the proposal distribution. Both of these quantities are useful for understanding the overall efficiency of an IS procedure.

Using the ESS framework, we compare the efficiency of the following IS methodologies: the procedure (FD) introduced by Fearnhead and Donnelly (2001) and described in Section 4.2.3, for both $\hat{\pi} = \hat{\pi}_{\text{FD}}$ and $\hat{\pi} = \hat{\pi}_{\text{SMC(4)}}$; the PIM procedure (PIM) described in Section 4.2.4 for $\hat{\pi} = \hat{\pi}_{\text{SMC(4)}}$; and the optimized PIM procedure (PIM-Optimized) described in Section 4.2.5 for $\hat{\pi} = \hat{\pi}_{\text{SMC(4)}}$. We simulate data under the coalescent with recombination for a single panmictic population, setting $\theta_\ell = \theta = 1$ for all $\ell \in L$ and $\rho_b = \rho = 1$ for all $b \in B$. For each value of $k \in \{4, 8, 12, 16, 20\}$, 25 k-locus 10-haplotype configurations were generated. This simulation procedure is analogous to method M1, described in Section 4.1.1. Using each of the above IS methodologies, we computed the sampling probability associated with each haplotype configuration, stopping when $\text{ESS} \geq 10000$ or $N = 100000$. We then computed the per-sample effective sampling rate, the per-second sampling rate, and the overall ESSRate, and averaged the results across haplotype configurations. The results are presented in Figure 4.6.

We begin by considering the log-scaled effective sampling rate (ESR), presented in Figure 4.6(a). Observe that the ESR decreases exponentially with with the number of loci, illustrating one reason that IS sampling, at least in its present form, does not scale beyond small haplotype configurations. Within this general trend, the PIM and PIM-Optimized procedures are nearly indistinguishable, the expected result as the difference between the procedures is purely algorithmic and does not affect the relevant distributions. Moreover, as predicted, these procedures perform considerably better than FD; as described above, this is due to the fact that the space of explored genealogies is markedly reduced for the former, reducing the complexity of the problem. Finally, note that within the FD procedure, using $\hat{\pi} = \hat{\pi}_{\text{SMC(4)}}$ in place of $\hat{\pi} = \hat{\pi}_{\text{FD}}$ does produce an improvement, and this improvement increases with the number of loci. This is in concordance with our earlier finding in Section 4.1, indicating that $\hat{\pi}_{\text{SMC(4)}}$ is more accurate than $\hat{\pi}_{\text{FD}}$.

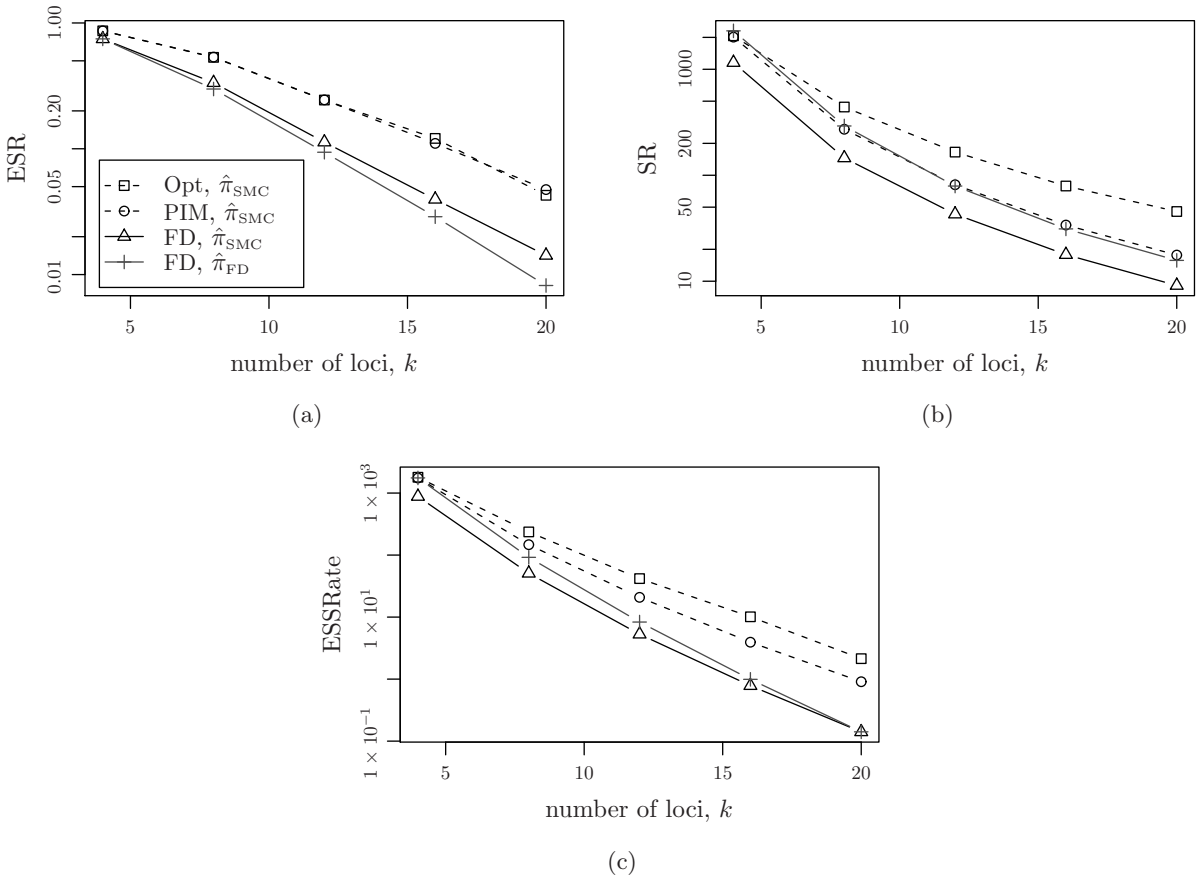


Figure 4.6. Empirically observed average effective sampling rate (ESR), sampling rate (SR), and effective sampling size rate (ESSRate) for several importance sampling procedures, as a function of the number of loci, k. The importance sampling procedures labeled FD correspond to the basic procedure introduced by Fearnhead and Donnelly (2001), setting $\hat{\pi} = \hat{\pi}_{\mathrm{SMC}(4)}$ and $\hat{\pi} = \hat{\pi}_{\mathrm{FD}}$. The procedures labeled PIM and Opt correspond to the improvements/optimizations described in Section 4.2.4 and Section 4.2.5. For each value of $k \in \{4, 8, 12, 16, 20\}$, $N = 25$ 10-haplotype k-locus haplotype configurations were generated using coalescent simulation with $\theta = 1$ and $\rho = 1$. (a) The effective sampling rate. (b) The sampling rate. (c) The effective sample size rate, computed as the product of the the effective sampling rate and the sampling rate.

We next consider the log-scaled sampling rate (SR), presented in Figure 4.6(b). Once again, the sampling rate decreases rapidly, though sub-exponentially, with the number of loci. In contrast to the ESR, it is clear that the PIM-optimized procedure performs better than the simple PIM procedure, due to the algorithmic improvement. The simple PIM procedure also performs better than the FD procedure for $\hat{\pi} = \hat{\pi}_{\mathrm{SMC}(4)}$; once again, this is due to the reduced complexity of each sampled genealogy. Observe that, within the FD procedure, using $\hat{\pi} = \hat{\pi}_{\mathrm{SMC}(4)}$ in place of $\hat{\pi} = \hat{\pi}_{\mathrm{FD}}$ reduces performance due to the increased computational complexity of $\hat{\pi}_{\mathrm{SMC}(4)}$ relative to $\hat{\pi}_{\mathrm{FD}}$; critically, however, the performance is reduced by a constant factor, independent of the number of loci.

Finally, in Figure 4.6(c), we consider the log-scaled ESSRate, the product of the ESR and SR. As expected, the PIM-optimized procedure is the best, providing a 2× improvement in IS efficiency

over the PIM procedure and a 15× improvement over the FD procedure for $k = 20$ loci. We anticipate that this improvement will continue to grow with the number of loci. Within the FD procedure, observe that using $\hat{\pi} = \hat{\pi}_{\text{SMC(4)}}$ in place of $\hat{\pi} = \hat{\pi}_{\text{FD}}$ reduces overall efficiency; however, this effect is reduced as the number of loci increases, and we anticipate that for $k > 20$ loci, $\hat{\pi}_{\text{SMC(4)}}$ will produce a more efficient IS procedure. In conclusion, we remark that, though we have produced some improvements in overall IS efficiency using both statistical and algorithmic improvements, IS remains impracticable for all but very small haplotype configurations.

4.3 Approximate Likelihood Methods

In the previous section, we described the use of importance sampling to approximate the probability, or likelihood, of a haplotype configuration in the multiple-locus, single-deme setting. Though we were able to improve the efficiency of importance sampling by incorporating parent independent mutation and a judicious implementation, the procedure remains impracticable for even modestly sized samples. In this section, we describe several approximate likelihood frameworks, for which the computational complexity scales linearly with the size of the sample.

We note at the outset that the use of approximate likelihood methods in population genetics is already an established research area. Hudson (2001) and Fearnhead and Donnelly (2002) considered composite likelihoods formed by considering products over pairs and small sets of loci, respectively. The former provides the foundation for the estimation of fine-scale recombination rates (McVean et al., 2004; Chan et al., 2012), and the latter provides the foundation for the estimation of recombination hotspots (Fearnhead and Smith, 2005). Explicitly related to the CSD, Li and Stephens (2003) proposed a decomposition of the sampling probability into a product of approximate CSPs, referred to as the product of approximate conditionals (PAC) likelihood. Incorporated into both Bayesian and frequentist frameworks, PAC likelihoods have been used to infer recombination rates (Li and Stephens, 2003), gene conversion parameters (Gay et al., 2007; Yin et al., 2009), and population demography (Davison et al., 2009; Sheehan et al., 2012).

Though the PAC likelihood was introduced concomitantly with the approximate CSD $\hat{\pi}_{\text{LS}}$ (Li and Stephens, 2003), it can be evaluated using any approximate CSD, including $\hat{\pi}_{\text{SMC}(\mathcal{P})}$. Importantly, for all known approximate CSDs, the PAC likelihood depends on the ordering of the approximate CSPs. In order to reduce this dependence, Li and Stephens suggest defining the PAC likelihood as the arithmetic mean over a small number of randomly-chosen orderings. In Section 4.3.1, we provide a more explicit description of the PAC likelihood, and also introduce two alternative composite likelihoods that do not depend on CSP ordering. In Sections 4.3.2 and 4.3.3, we make use of these approximate likelihoods in an ML framework to estimate migration and recombination rates. We remark that these example applications are primarily intended to demonstrate that the CSD $\hat{\pi}_{\text{SMC}(\mathcal{P})}$ can be used for estimation in an approximate likelihood framework, and also to evaluate the effect of using the alternative composite likelihoods.

4.3.1 Composite and approximate likelihoods

Let $\mathcal{D}$ be a finite set of demes, and $\mathbf{n} = \mathbf{e}_{d^{(1)},h^{(1)}} + \cdots + \mathbf{e}_{d^{(n)},h^{(n)}}$ be a structured haplotype configuration, where $d^{(i)} \in \mathcal{D}$ and $h^{(i)} \in \mathcal{H}$ for $1 \leq i \leq n$. Recalling that $q(\mathbf{n})$ is the ordered sampling probability of the configuration under a population genetic model, by repeated application of the

definition (1.61) of the CSP,

$$q(\mathbf{n}) = \prod_{i=1}^{n} \pi\Big(\mathbf{e}_{d^{(i)},h^{(i)}} \Big| \mathbf{n} - \sum_{j=1}^{i} \mathbf{e}_{d^{(j)},h^{(j)}}\Big), \tag{4.25}$$

where π is the exact CSD for the population genetic model. Because the density $q(\cdot)$ is exchangeable, the prescribed haplotype ordering does not affect the result, and we therefore obtain an identical result for $\mathbf{n} = \mathbf{e}_{d^{(\sigma(1))},h^{(\sigma(1))}} + \cdots + \mathbf{e}_{d^{(\sigma(n))},h^{(\sigma(n))}}$, where σ is an arbitrary permutation on $\{1,\ldots,n\}$.

As it is not generally possible to evaluate the CSP associated with the exact CSD π, Li and Stephens (2003) suggest replacing the exact CSD with an approximate CSD $\hat{\pi}$ for which the requisite CSPs can be efficiently evaluated,

$$q(\mathbf{n}) \approx \prod_{i=1}^{n} \hat{\pi}\Big(\mathbf{e}_{d^{(i)},h^{(i)}} \Big| \mathbf{n} - \sum_{j=1}^{i} \mathbf{e}_{d^{(j)},h^{(j)}}\Big). \tag{4.26}$$

Provided an approximate CSP $\hat{\pi}$, the exchangeability property described above no longer holds, and the approximate likelihood generally depends on the specific ordering of haplotypes in the configuration $\mathbf{n}$. In practice, we have found that even for moderately-size samples, the approximate likelihood can fluctuate by many orders of magnitude depending on the ordering. We also remark that the extent to which the approximate likelihood varies with the ordering depends on the choice of CSD $\hat{\pi}$; those CSDs that are more accurate, as described in Section 4.1, generally produce narrower ranges of approximate likelihoods (data not shown).

In order to reduce the dependence this estimate on the ordering chosen, Li and Stephens suggest taking the arithmetic mean over approximately 20 randomly selected orderings. Thus, letting Σ be a set of randomly-selected permutations on $\{1,\ldots,n\}$, with $|\Sigma| = 20$, the PAC likelihood is defined

$$\hat{q}_{\mathrm{PAC}}(\mathbf{n}) = \frac{1}{|\Sigma|} \sum_{\sigma\in\Sigma} \prod_{i=1}^{n} \hat{\pi}\Big(\mathbf{e}_{d^{(\sigma(i))},h^{(\sigma(i))}} \Big| \mathbf{n} - \sum_{j=1}^{i} \mathbf{e}_{d^{(\sigma(i))},h^{(\sigma(j))}}\Big). \tag{4.27}$$

In the context of likelihood-based estimation, it is critical to select a single permutation set Σ, and define the approximate likelihood $\hat{q}_{\mathrm{PAC}}(\cdot)$ with respect to that permutation set.

Inspired by the locus-wise composite likelihoods mentioned above, we also consider two haplotype-wise composite methods. The first of these, the *leave-one-out composite likelihood* (LCL), formulates the likelihood as a product of CSPs, each the result of sampling a single haplotype conditioned on the remaining haplotypes. We take the n-th root in order to provide the interpretation of the LCL as the geometric mean the n leave-one-out CSPs,

$$\hat{q}_{\mathrm{LCL}}(\mathbf{n}) \propto \Big[\prod_{i=1}^{n} \hat{\pi}(\mathbf{e}_{d^{(i)},h^{(i)}} | \mathbf{n} - \mathbf{e}_{d^{(i)},h^{(i)}})\Big]^{1/n}. \tag{4.28}$$

We have used proportionality rather than equality to reflect that the composite likelihood does not directly approximate the true likelihood, but rather serves as a proxy for the purposes of inference, for which it is only necessary to know (or approximate) the likelihood up to a constant or proportionality.

The second haplotype-wise composite method, the *pairwise composite likelihood* (PCL), formulates the likelihood as a product of pairwise CSPs, each the result of sampling a single haplotype

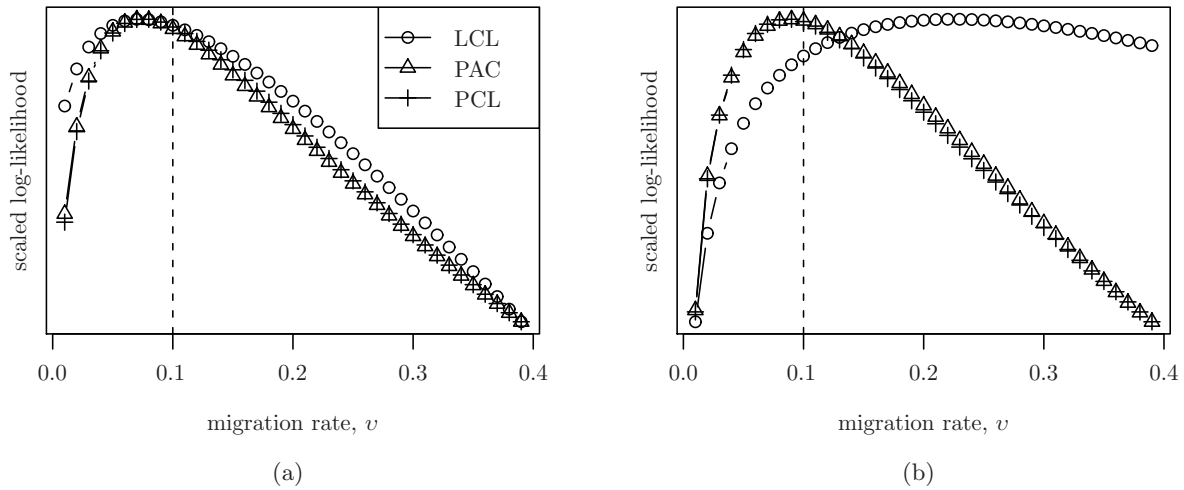


Figure 4.7. Re-scaled log likelihood surfaces for two haplotype configurations (generated for $\upsilon_0 = 0.10$, indicated by a vertical line in the plots), and for each of the three approximate likelihood formulations (LCL, PAC, PCL) described in the text, setting $\hat{\pi} = \hat{\pi}_{\text{SMC}(\mathcal{P})}$ and provided the true values of θ and ρ. (a) A case for which all of the likelihood surfaces are similar (b) A case for which the LCL likelihood surface is substantially different than the likelihood surfaces for PAC and PCL

conditioned on a single alternative haplotype. As before, we take the (n^2)-th root in order to provide the interpretation of the PCL as the geometric mean of the n^2 pairwise CSPs,

$$\hat{q}_{\text{PCL}}(\mathbf{n}) \propto \left[\prod_{i=1}^{n} \prod_{j=1}^{n} \hat{\pi}(\mathbf{e}_{d^{(i)},h^{(i)}} | \mathbf{e}_{d^{(i)},h^{(j)}}) \right]^{1/n^2} \tag{4.29}$$

Unlike the PAC-based likelihood, neither the LCL nor PCL composite likelihoods depend on the prescribed haplotype ordering, and so it is unnecessary to define the likelihood with respect to a particular permutation set.

4.3.2 Estimation of migration rates

To demonstrate the utility of our approximate CSD $\hat{\pi}_{\text{SMC}(\mathcal{P})}$, we consider the problem of estimating migration rates for data simulated under the coalescent with recombination and migration. Assume a structured population with two demes, $\mathcal{D} = \{1, 2\}$, and set the population proportion within each deme $\kappa_1 = \kappa_2 = 0.5$ and the migration rates $\upsilon_{12} = \upsilon_{21} = \upsilon$. We use a 2-allele model, setting $\theta_\ell = \theta = 5 \times 10^{-2}$ for all $\ell \in L$, and $\rho_b = \rho = 5 \times 10^{-2}$ for all $b \in B$. For each value of $\upsilon = \upsilon_0 \in \{0.01, 0.10, 1.00, 10.0\}$, 100 haplotype configurations with $n_1 = n_2 = 10$ haplotypes in each of the two demes and $k = 10^4$ loci were generated. This simulation procedure is analogous to method M1, described in Section 4.1.1

Observe that the per-individual mutation and recombination rates are both approximately $10^4 \cdot 5 \times 10^{-2} = 5 \times 10^2$. In humans, for which average per-base mutation and recombination rates are on the order of 10^{-3}, these values correspond to a genomic sequence on the order of 500kb. We thus reason that the simulated haplotypes are representative of a relatively longer genomic sequence that has been "compressed", for reasons of computational efficiency, into 10^4 loci. Further, we chose

the range of migration rates to be concordant with recent estimates in humans (Gutenkunst et al., 2009; Gravel et al., 2011), as well as Drosophila (Wang and Hey, 2010).

For each of the three approximate likelihood formulations described above, $\hat{q}_{\text{LCL}}$, $\hat{q}_{\text{PAC}}$, and $\hat{q}_{\text{PCL}}$, we set $\hat{\pi} = \hat{\pi}_{\text{SMC}(\mathcal{P})}$ with discretization $\mathcal{P}$ chosen using the logarithmic procedure detailed in Section 3.2.2 for $|\mathcal{P}| = 8$, and consider the approximate likelihood surface for the parameter υ, fixing the values of θ and ρ to the true values used for simulation. Figure 4.7 shows the likelihood surfaces for two example configurations (generated as described above) for data simulated using parameter $\upsilon_0 = 0.10$. Perhaps most importantly, the likelihood surfaces appear to be unimodal and otherwise well-behaved. In Figure 4.7(a), the likelihood curves are quite similar to one another, and the maximum likelihood occurs near the true parameter. This is not generally true, however, as evidenced by Figure 4.7(b), for which the likelihood surface for the LCL method is substantially different than that of PAC and PCL.

We next consider the behavior of the maximum likelihood estimate (MLE) under each of the likelihood approximations. For each simulated dataset, we compute, using golden section search, the MLE migration rate $\hat{\upsilon}$. For each MLE, we then evaluate $\log_2(\hat{\upsilon}/\upsilon_0)$, where υ_0 is the true migration rate used to generate the dataset. Using the transformed MLE, results for different values of υ_0 are directly comparable; a correct estimate of the migration rate produces a value of 0, and under- and overestimation by a factor of two produce values of -1 and 1, respectively. Box plots for the transformed MLE under each likelihood approximation and for each true migration rate $\upsilon_0 \in \{0.01, 0.10, 1.00, 10.0\}$ are presented in Figure 4.8.

Observe that the LCL-based MLE performs poorly for $\upsilon_0 = 0.01$ (see Figure 4.8(a)), consistently underestimating the true value; this may be because the final haplotype to be sampled is generally very similar to previously sampled haplotypes within the deme, obviating the need for migration events within the conditional genealogy. Intuitively, this effect should be diminished when the data are produced using larger migration rates, which does appears to be the case (see Figures 4.8(b)–4.8(d)). On the other hand, the PCL-based MLE performs poorly for $\upsilon_0 = 10.0$, again consistently underestimating the true value. This may be because, for large migration rates, there simply is not enough information in a pairwise analysis of the haplotypes to determine the true rate; intuitively, this effect should be diminished when the data are produced using smaller migration rates, relative to the rate of recombination. This is indeed the case, and in fact, for smaller migration rates, the PCL-based MLE is well-correlated with the PAC-based MLE (data not shown).

The PAC-based MLE appears not to suffer at either of these extremes. We speculate that this is because PAC incorporates both pairwise and higher-order terms, making it less susceptible to the problems we observe with the LCL- and PCL-based MLEs; we remark that Li and Stephens (2003) came to a similar conclusion for recombination rates. Perhaps most importantly, the PAC-based estimation is quite accurate, demonstrating that, using the CSD $\hat{\pi}_{\text{SMC}(\mathcal{P})}$, it is possible to obtain excellent estimates of the migration rate.

4.3.3 Estimation of recombination rates

Motivated by our results for estimating migration rates, we next consider the problem of estimating recombination rates in a single panmictic population. As before, we assume a 2-allele model, setting $\theta_\ell = \theta = 5 \times 10^{-2}$ for all $\ell \in L$, and $\rho_b = \rho$ for all $b \in B$. For each value of the recombination rate $\rho = \rho_0 \in \{0.005, 0.0.01, 0.05, 0.10\}$, 100 haplotype configurations of $n = 20$ haplotypes and $k = 10^4$ loci were simulated. As described above, we reason that the simulated haplotypes are

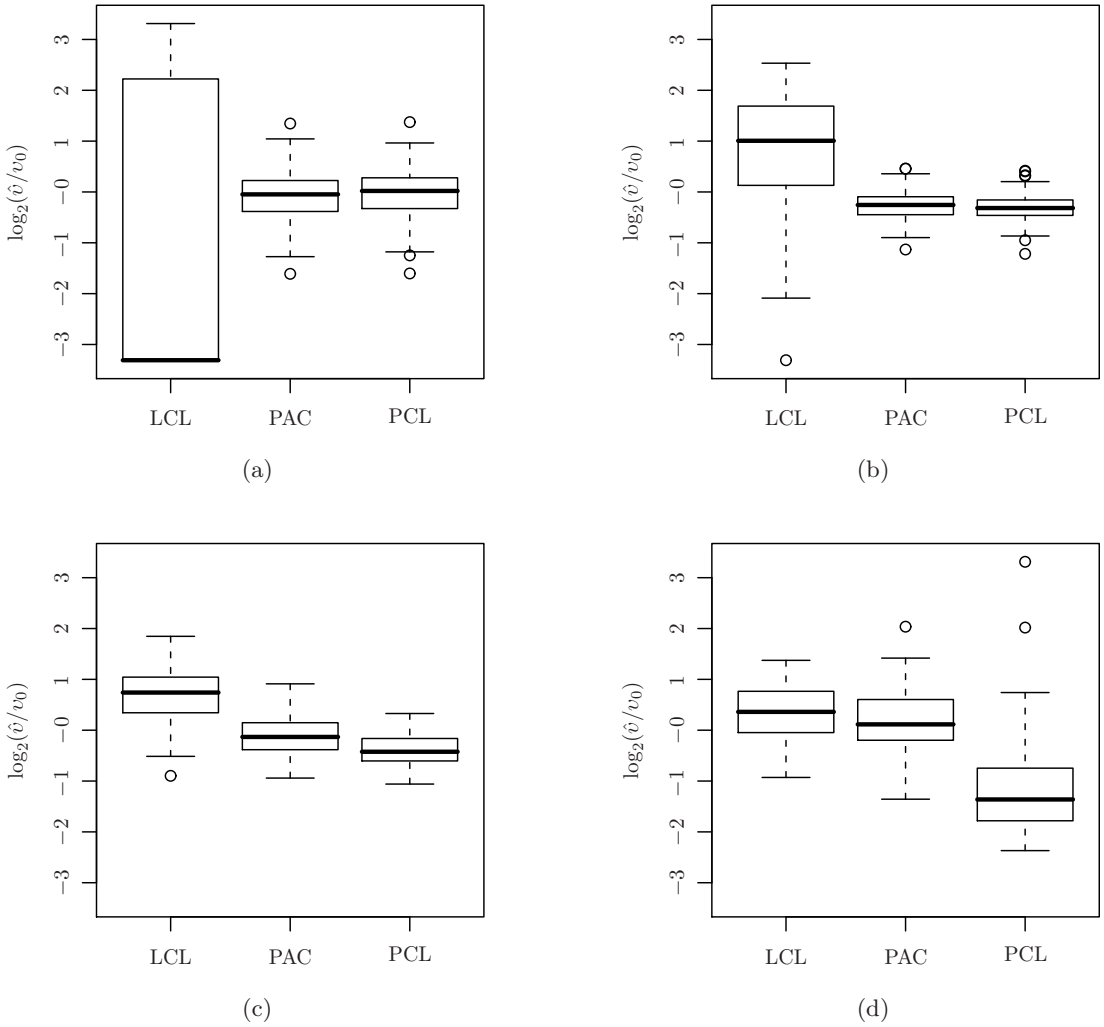


Figure 4.8. Box plots (produced using the software package R, and including outliers) for the quantity $\log_2(\hat{\upsilon}/\upsilon_0)$ over 100 samples, where υ_0 is the migration rate used for simulation, and $\hat{\upsilon}$ is the ML migration rate for each of the three approximate likelihood formulations (LCL, PAC, PCL) described in the text, setting $\hat{\pi} = \hat{\pi}_{\mathrm{SMC}(\mathcal{P})}$ and provided the true values of θ and ρ. The value $\hat{\upsilon}$ is computed using golden section search in the interval $(\upsilon_0 \cdot 10^{-1}, \upsilon_0 \cdot 10)$. (a) $\upsilon_0 = 0.01$ (b) $\upsilon_0 = 0.10$ (c) $\upsilon_0 = 1.00$ (d) $\upsilon_0 = 10.0$. Note that the median of the LCL estimator in (a) lies on the lower bound of the interval, and therefore at least half of the estimates reach this bound and are likely smaller.

representative of a relatively longer genomic sequence that has been "compressed", for reasons of computational efficiency, into 10^4 loci.

For each of the three approximate likelihood formulations described above, $\hat{q}_{\text{LCL}}$, $\hat{q}_{\text{PAC}}$, and $\hat{q}_{\text{PCL}}$, we set $\hat{\pi} = \hat{\pi}_{\text{SMC}(\mathcal{P})}$ with discretization $\mathcal{P}$ chosen using the Gaussian quadrature procedure detailed in Section 3.2.1 for $|\mathcal{P}| = 8$, and fix θ to the true value used for simulation. For each simulated dataset, we compute, using golden section search, the MLE recombination rate $\hat{\rho}$ and $\log_2(\hat{\rho}/\rho_0)$, where ρ_0 is the true recombination rate used to generate the dataset. As before, using the transformed MLE, results for different values of ρ_0 are directly comparable. Box plots for the transformed MLE under each likelihood approximation and for each true migration rate $\rho_0 \in \{0.005, 0.0.01, 0.05, 0.10\}$ are presented in Figure 4.9.

In contrast to the ML estimation of migration, both the LCL and PAC estimates of recombination rate are generally biased upward. As demonstrated in Figures 4.9(a)–4.9(d), the bias is maximal for the smallest value of ρ_0, and decreases for larger values of ρ_0. In order to understand the source of this bias, we have investigated the LCL estimator in detail. We observe that when ρ_0 is small and few recombinations occur, the resulting likelihood surfaces for the CSPs comprising $\hat{q}_{\text{LCL}}$ are markedly heterogeneous, both in their absolute value, and in their component-wise MLE; the resulting composite likelihood surface is therefore sensitive to the precise balance of the component CSPs. In general, the balance produces an upward bias, but the effect is mediated as ρ_0 becomes larger, and the likelihood surfaces for the component CSPs more homogeneous. Provided the correlation between the LCL and PAC estimates (data not shown), we anticipate that a similar effect occurs for the PAC estimate. In contrast, the PCL estimate of recombination is biased downward; moreover, the bias is minimal for the smallest value of ρ_0, and increases for larger values of ρ_0. A possible explanation is that there is are too few polymorphic sites in a pairwise analysis to provide support for a high recombination rate; intuitively, this effect should be diminished for recombination rates that are smaller relative to the mutation rate. A similar explanation was posed previously for the downward biased migration rate estimate using the PCL approximate likelihood.

Finally, we remark that although the results of approximate likelihood based estimation of recombination rate are difficult to interpret, they are not entirely defective. In all cases, the median estimate is within a factor of 4 of the truth, and the distribution of estimates is narrow, suggesting the potential for an empirically-driven correction similar to that proposed by Li and Stephens (2003). Moreover, this type of result is not exclusive to $\hat{\pi}_{\text{SMC}}$. Setting $\hat{\pi} = \hat{\pi}_{\text{FD}}$, we obtained similar results, and Li and Stephens (2003) also report biased estimates of ρ in some settings. Because the PAC likelihood is used extensively for parameter estimation, we believe that it would be useful to carry out a comprehensive study on the bias and variance of the MLE, for a wider variety of parameter settings and choices for the approximate CSD $\hat{\pi}$.

4.4 Pseudo-Posterior Sampling

In the previous section, we demonstrated that it is possible to approximate the probability, or likelihood, of a haplotype configuration as a product of approximate CSPs. Critically, because the CSP can evaluated efficiently for large class of approximate CSDs, including $\hat{\pi}_{\text{SMC}(\mathcal{P})}$, the resulting likelihoods can be used for computationally efficient statistical inference, for example to estimate model parameters. In contrast, known methods for exact or consistent likelihood computation, including numerically solving the recursion for sampling probability (as in Section 1.2.2) and Monte Carlo methods such as importance sampling (as in Section 4.2), are computationally impracticable.

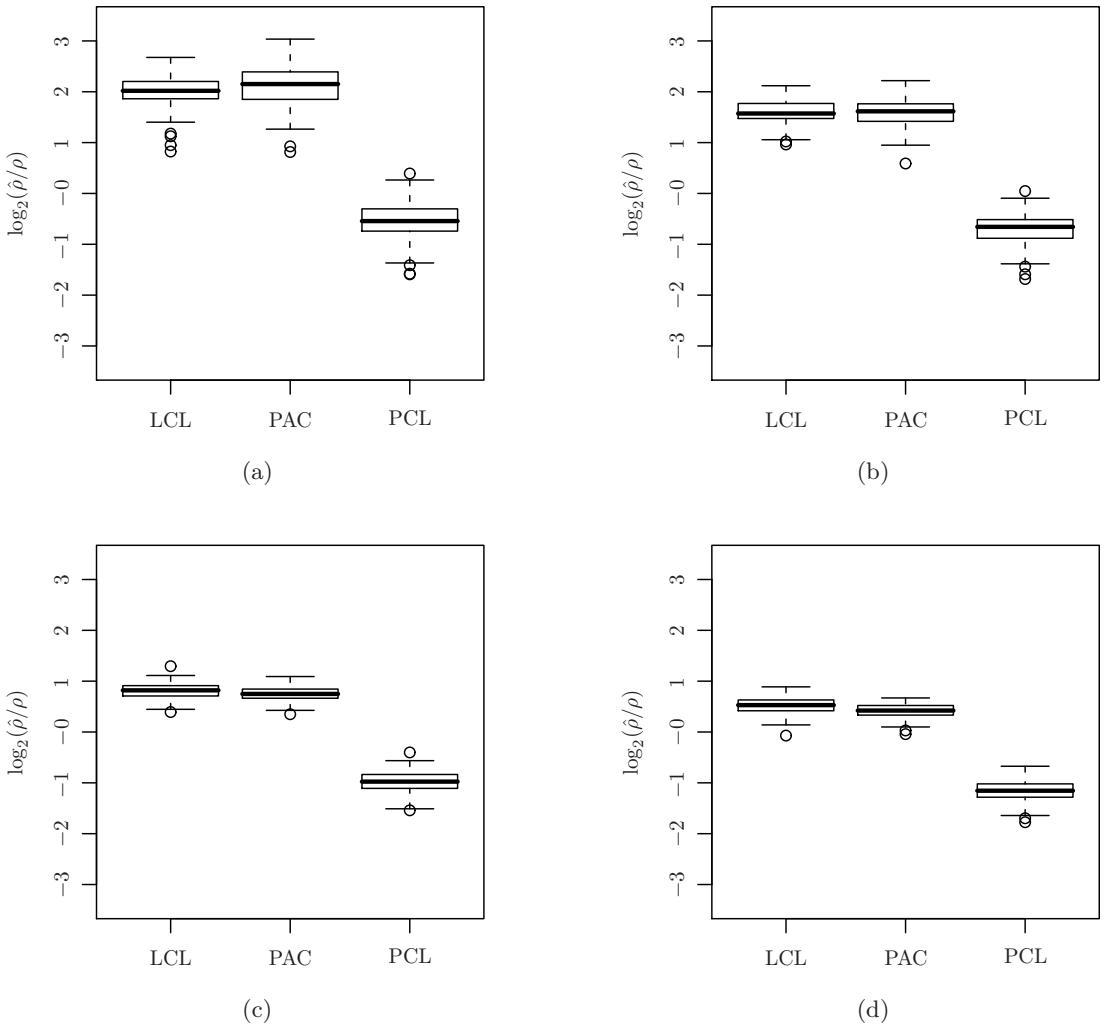


Figure 4.9. Box plots (produced using the software package R, and including outliers) for the quantity $\log_2(\hat{\rho}/\rho_0)$ over 100 samples, where ρ_0 is the migration rate used for simulation, and $\hat{\rho}$ is the ML recombination rate for each of the three approximate likelihood formulations (LCL, PAC, PCL) described in the text, setting $\hat{\pi} = \hat{\pi}_{\mathrm{SMC}(8)}$ and provided the true value of θ. The value $\hat{\rho}$ is computed using golden section search in the interval $(\rho_0 \cdot 10^{-1}, \rho_0 \cdot 10)$. (a) $\rho_0 = 0.005$ (b) $\rho_0 = 0.01$ (c) $\rho_0 = 0.05$ (d) $\rho_0 = 0.10$.

We next consider the problem of ancestral inference: provided a population genetic sample, we may wish to infer whether a mutation occurred more than once at a polymorphic locus, the ancestry of an admixed group of individuals at a particular locus, or the degree of relatedness within and between groups of individuals. Such questions are naturally addressed by explicitly invoking the genealogy relating the individuals in the sample, and not simply a likelihood. Because the true genealogy is not typically known, a theoretically well-motivated procedure is to integrate over the posterior distribution of genealogies, assuming the appropriate coalescent-based prior distribution. Much as in computing the sample likelihood, though it is possible to sample from the true posterior distribution using Monte Carlo methods, known techniques are computationally impracticable for even modestly-sized samples.

Recall that a genealogy relating individuals in a sample induces, at each locus, a marginal tree. In this section, we propose two related CSD-based pseudo-posterior distributions on the marginal tree at a specified locus, conditioned on the observed sample; notably, the observed sample includes information at all loci, which impacts inference of the marginal tree at the specified locus. Though a posterior distribution on the marginal tree at one locus is less beneficial than the posterior distribution on full genealogies (or, similarly, the joint posterior distribution on the collection of marginal trees at all loci), it is sufficient for many questions of interest, including the examples given above. Importantly, the marginal trees sampled from the pseudo-posteriors include time information, but do not explicitly include mutation events; the latter can be efficiently incorporated using, for example, Felsenstein's algorithm (Felsenstein, 1981).

The central idea in constructing the pseudo-posterior distributions is to make direct use of the marginal conditional genealogies (MCGs) associated with the genealogical interpretation of $\hat{\pi}_{\text{SMC}}$. By interpreting an absorption event within the MCG as a coalescence event, we infer coalescence events within the marginal tree. The primary complication with this approach is then integrating coalescence events across the MCGs associated with several CSDs. We address this issue by constructing a posterior process for each MCG, and then combining these processes into a single posterior process for the marginal tree. Letting $\mathbf{n}$ be a haplotype configuration, and specifying an arbitrary locus $\ell \in L$, the two pseudo-posterior distributions on marginal trees are then formed by considering different combinations of the MCG posterior processes:

Pairwise: The CSPs $\{\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{e}_{\eta'})\}$ for each pair $\eta, \eta' \in \mathcal{H}_{\mathbf{n}}$ result in an MCG posterior process at locus ℓ for each pair of haplotypes. The pairwise MCG posterior processes are transformed into a posterior coalescence process for each pair of lineages in the tree, and these processes are then combined to produce a posterior process on marginal trees.

Leave-one-out: The CSPs $\{\hat{\pi}_{\text{SMC}}(\mathbf{e}_\eta|\mathbf{n}-\mathbf{e}_\eta)\}$ for each $\eta \in \mathcal{H}_{\mathbf{n}}$ result in a directed MCG posterior process at locus ℓ for each haplotype. The directed MCG posterior processes are transformed into a posterior coalescence process for each pair of lineages in the tree, and these processes are then combined to produce a posterior process on marginal trees.

In Section 4.4.1, the problem is introduced formally, and relevant notation described. In Section 4.4.2, the construction of the MCG posterior process is discussed, and in Sections 4.4.3 and 4.4.4, the pairwise and leave-one-out methodologies are described in detail, respectively. We remark at the outset that a guiding principle in our development of pseudo-posterior distributions is that, in the absence of data, the pseudo-posterior distributions should reduce to the known prior distribution on marginal trees, given by Kingman's coalescent (Kingman, 1982a).

4.4.1 Sampling marginal trees

Let $\mathbf{n} = \mathbf{e}_{h^{(1)}} + \cdots + \mathbf{e}_{h^{(n)}}$ be a haplotype configuration. Towards sampling a marginal tree, we define a *lineage set* $\mathcal{L}$ as a partition of $\{1, \ldots, n\}$, representing the state of the tree at a particular time, where each lineage $\mu \in \mathcal{L}$ is the set of haplotypes subtended by the lineage. The initial lineage set $\mathcal{L}^{(0)}$ contains a single lineage associated with each of the n haplotypes, and the lineage set $\mathcal{L}^{(r)}$ contains each of the $n - r$ lineages after r coalescence events. A marginal tree $\mathcal{T}$ is then specified by a sequence of coalescence events $\mathcal{T} = (E^{(1)}, \ldots, E^{(n-1)})$, where $E^{(r)}$ is the r-th coalescence event. The coalescence event $E^{(r)}$ comprises a coalescence time and a pair of distinct coalescing lineages $\mu, \nu \in \mathcal{L}^{(r-1)}$, and produces the lineage set $\mathcal{L}^{(r)}$ by joining the lineages μ and ν into a single lineage.

We first consider sampling a marginal tree $\mathcal{T}$ under the prior coalescent process, namely Kingman's coalescent. Suppose that r coalescence events have already been sampled, so that the current set of lineages is $\mathcal{L}^{(r)}$, with $|\mathcal{L}^{(r)}| = n - r$. Recall that for the prior coalescent process, each pair of distinct lineages $\mu, \nu \in \mathcal{L}^{(r)}$ coalesce at rate 1 so that the total rate is $\binom{n-r}{2}$. The process transitions when the first pair of lineages coalesce; the time and pair of lineages then determine the event $E^{(r+1)}$ and the lineage set $\mathcal{L}^{(r+1)}$. This procedure is iterated until the final event $E^{(n-1)}$ has been determined, thus completing the sampled marginal genealogy $\mathcal{T}$.

Similarly, consider sampling a marginal tree $\mathcal{T}$ under the posterior process, conditioned on the observed haplotype configuration $\mathbf{n}$. Again, suppose that r coalescence events have already been sampled, so that the current set of lineages is $\mathcal{L}^{(r)}$. Then for each pair of distinct lineages $\mu, \nu \in \mathcal{L}^{(r)}$, denote by $\sigma^{(r)}_{\mu\nu}(t)$ the time-heterogenous *posterior* rate of coalescence between lineages μ and ν at time $t \in \mathbb{R}_{\geq 0}$. The rate $\sigma^{(r)}_{\mu\nu}(t)$ generally depends on the configuration $\mathbf{n}$ and the previous r coalescence events; for simplicity, we suppress this dependence in our notation. Entirely analogous to the prior process described above, the posterior coalescent process transitions when the first pair of lineages coalesce, determining the coalescence event $E^{(r+1)}$ and the lineage set $\mathcal{L}^{(r+1)}$, and this procedure is iterated until the final event $E^{(n-1)}$ has been determined, thus completing the sampled marginal genealogy $\mathcal{T}$.

In contrast to the prior process, the posterior coalescent process is not time-homogeneous, as the posterior rates depend on the time t, and so it is necessary to consider the absolute time in sampling the tree $\mathcal{T}$. In practice, it is necessary to discretize the absolute time into a finite set of intervals, denoted $\mathcal{P}$, so that for all $t \in p \in \mathcal{P}$, $\sigma^{(r)}_{\mu\nu}(t) = \sigma^{(r)}_{\mu\nu}(p)$. Nonetheless, observe that by setting $\sigma^{(r)}_{\mu\nu}(p) = 1$ for all $\mu, \nu \in \mathcal{L}^{(r)}$ and all $p \in \mathcal{P}$, the posterior process is reduced to the prior process. In the following sections, we describe two methods for approximating the discretized posterior rates of coalescence; these approximations can then be used in the present framework to sample trees from a pseudo-posterior.

4.4.2 MCG posterior process

Suppose we wish to sample a single haplotype conditional on the previously-observed configuration $\mathbf{n} = \mathbf{e}_{h^{(1)}} + \cdots + \mathbf{e}_{h^{(n)}}$ using the CSD $\hat{\pi}_{\mathrm{SMC}}$. Recall from Section 2.3.2 that in the single-deme setting, the random MCG at an arbitrary locus is denoted by a pair $S = (T, H)$, where T denotes the absorption time, and H the absorption haplotype. The marginal distribution on S is described by a genealogical process wherein the lineage associated with conditionally sampled haplotype is absorbed into each of the n haplotypes at homogenous rate 1. Letting $s = (t, h) \in \mathcal{S} = \mathbb{R}_{\geq 0} \times \mathcal{H}$, the density $\zeta(s)$ is given by (2.73). As above, we discretize the continuous component of $\mathcal{S}$ associated with absorption time into the set of intervals $\mathcal{P}$, and consider absorption into a specific, labeled

haplotype $h^{(i)}$ where $i \in \{1, \ldots, n\}$. Letting $(p, i) \in \mathcal{P} \times \{1, \ldots, n\}$, we deduce from the genealogical process the density,

$$\zeta(p, i) = \int_p e^{-nt} dt. \tag{4.30}$$

Conversely, given an arbitrary density $f(\cdot)$ over the space $\mathcal{P} \times \{1, \ldots, n\}$ of discretized, labeled MCGs, it is possible to construct a marginal genealogical process inducing this density. Critically, because the density $f(\cdot)$ is over the discretized space of MCGs, the rates associated with the genealogical process are constant within each interval $p \in \mathcal{P}$, but are not generally constant between intervals. Before proceeding, it is convenient to define several functions associated with $f(\cdot)$; for $p \in \mathcal{P}$ and $i \in \{1, \ldots, n\}$,

$$f(p) = \sum_{i=1}^{n} f(p, i), \quad \hat{f}(p) = \frac{f(p)}{\sum_{p' \geq p} f(p')}, \quad \hat{f}(p, i) = \frac{f(p, i)}{\sum_{p' \geq p} f(p')}. \tag{4.31}$$

These functions correspond to the total probability of being absorbed in interval p, the total probability of being absorbed in interval p conditioned on not being absorbed prior to p, and the probability of being absorbed into the labeled haplotype $\mathbf{e}_{h^{(i)}}$ in interval p conditioned on not being absorbed prior to p. Denote by $\lambda(p)$ the total rate of absorption during the time interval p. Then using the theory of continuous-time Markov processes,

$$\lambda(p) = \begin{cases} -\frac{1}{|p|} \log\left(1 - \hat{f}(p)\right), & \text{for } p < p_{\mathrm{F}}, \\ n, & \text{for } p = p_{\mathrm{F}}, \end{cases} \tag{4.32}$$

where p_{F} is the final (infinite) discretization interval. The rate of absorption in the final interval p_{F} cannot be deduced from the density $f(\cdot)$; we have thus chosen to set the total rate in this interval equal to the total prior rate, n. Further denoting by $\lambda_i(p)$ the rate of absorption into the lineages associated with labeled haplotype $\mathbf{e}_{h^{(i)}}$ during the time interval $p \in \mathcal{P}$,

$$\lambda_i(p) = \frac{\hat{f}(p, i)}{\hat{f}(p)} \cdot \lambda(p) = \begin{cases} -\frac{\hat{f}(p,i)}{\hat{f}(p)} \cdot \frac{1}{|p|} \log\left(1 - \hat{f}(p)\right), & \text{for } p < p_{\mathrm{F}}, \\ \frac{\hat{f}(p,i)}{\hat{f}(p)} \cdot n, & \text{for } p = p_{\mathrm{F}}. \end{cases} \tag{4.33}$$

Using (4.33) it can easily be verified that setting $f = \zeta$, defined in (4.30), yields the correct homogenous prior absorption rate, $\lambda_i(p) = 1$, for all $i \in \{1, \ldots, n\}$ and $p \in \mathcal{P}$.

Now, let $\eta \in \mathcal{H}$, and consider computing $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_\eta | \mathbf{n})$. As described in Section 3.3.3, using marginal decoding, it is possible to compute an approximate posterior density $\vartheta(\cdot)$ on the space $\mathcal{P} \times \{1, \ldots, n\}$ for a particular locus $\ell \in L$. Setting $f = \vartheta$ and using (4.33), it is thus possible to deduce the approximate absorption rates associated with a *posterior* marginal genealogical process at locus ℓ. Unlike the prior genealogical process, the rates $\{\lambda_i(p)\}_{p \in \mathcal{P}}$ associated with the posterior process are not generally time-homogenous.

We will also be interested in computing posterior rates associated with a lineage set $\mathcal{L}$. Letting $f(\cdot)$ be a density on the space $\mathcal{P} \times \mathcal{L}$, it is possible to compute, using an equation entirely analogous to (4.32) the total rate of absorption $\lambda(p)$ during time interval $p \in \mathcal{P}$; as before, the rate of absorption in the final discretization interval must be independently specified. Similarly, using an equation entirely analogous to (4.33), it is possible to calculate the rate of absorption $\lambda_\mu(p)$ into lineage $\mu \in \mathcal{L}$ during time interval p. The precise methodology for constructing the density $f(\cdot)$ is described in the following sections.

4.4.3 Pairwise pseudo-posterior

Consider sampling a marginal coalescent tree $\mathcal{T}$ at locus $\ell \in L$ from the pseudo-posterior conditioned on configuration $\mathbf{n}$. Given that the first r coalescence events have been sampled, the current set of lineages is denoted by $\mathcal{L}^{(r)}$, and the objective is sample the $(r+1)$-th coalescence event $E^{(r+1)}$, comprising a time and a pair of distinct lineages $\mu, \nu \in \mathcal{L}^{(r)}$. The process for sampling this event is determined by the non-homogeneous posterior rates of coalescence between each such pair of lineages $\{\sigma^{(r)}_{\mu\nu}(p)\}_{p\in\mathcal{P}}$. In this section, we describe how to approximate these rates by appropriately combining the posterior distributions on MCGs for *pairs* of labeled haplotypes.

Let $h^{(i)}$ and $h^{(j)}$ be distinct labeled haplotypes of the configuration $\mathbf{n}$. As described in Section 3.3.3, posterior decoding for the CSP $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_{h^{(i)}}|\mathbf{e}_{h^{(j)}})$ provides a posterior distribution on MCGs at locus ℓ. Denote the corresponding density by $\vartheta_{ij}(\cdot)$, so that $\vartheta_{ij}(p)$ is the probability of the lineage associated with haplotype $h^{(i)}$ being absorbed into the trunk lineage associated with haplotype $h^{(j)}$ during the time interval p. These densities, computed for each pair of labeled haplotypes, $h^{(i)}$ and $h^{(j)}$, form the building blocks of the posterior lineage rates. We assume a symmetric mutation model, so that the density $\vartheta_{ij}(\cdot)$ is invariant with respect to the ordering of i and j.

In order to provide some intuition, consider first approximating the posterior lineage coalescence rates $\{\sigma^{(0)}_{\mu\nu}(p)\}_{p\in\mathcal{P}}$ when no coalescence events have occurred. Each lineage in $\mu \in \mathcal{L}^{(0)}$ is a singleton, so that $\mu = \{i\}$ for some $1 \le i \le n$. For an arbitrary pair of distinct lineages $\mu, \nu \in \mathcal{L}^{(0)}$, and assuming without loss of generality that $\mu = \{i\}$ and $\nu = \{j\}$, we set $f(p) = \vartheta_{ij}(p)$ for all $p \in \mathcal{P}$, and use (4.32) to obtain the associated rates $\{\lambda(p)\}_{p\in\mathcal{P}}$, setting $\lambda(p_{\mathrm{F}}) = 1$. By the symmetry stated above, these rates are independent of the ordering of μ and ν, and so we set, for all $p \in \mathcal{P}$,

$$\sigma^{(0)}_{\mu\nu}(p) = \lambda(p). \tag{4.34}$$

These rates are produced for each unordered pair of lineages $\mu, \nu \in \mathcal{L}^{(0)}$, and together provide a pseudo-posterior distribution for the first coalescence event $E^{(1)}$.

We next consider the more general case, after r coalescences have occurred, and the current set of lineages is given by $\mathcal{L}^{(r)}$. For an arbitrary pair of distinct lineages $\mu, \nu \in \mathcal{L}^{(r)}$, recall that $\mu, \nu \subset \{1, \ldots, n\}$ and $\mu \cap \nu = \emptyset$. As in the initial case, when $r = 0$, we define a density $f(\cdot)$ on the space $\mathcal{P}$, this time by combining the pairwise densities $\vartheta_{ij}(\cdot)$ for each $i \in \mu$ and $j \in \nu$. Note that there are a variety of ways to combine these densities; in the absence of a strong theoretical foundation, we choose a technique that is intuitively straightforward. For each pair of haplotypes, $h^{(i)}$ and $h^{(j)}$ with $i \in \mu$ and $j \in \nu$, we envision an *ongoing* posterior MCG process, associated with the CSP computation $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_{h^{(i)}}|\mathbf{e}_{h^{(j)}})$, and the rate within a particular time interval $p \in \mathcal{P}$ is determined by joining these processes together. We thus compute the rate $\sigma^{(r)}_{\mu\nu}(p)$ using the following procedure:

1. Condition each posterior MCG distribution on known information. For $h^{(i)}$ and $h^{(j)}$ with $i \in \mu$ and $j \in \nu$, the known information is that absorption has not occurred prior to the interval $p \in \mathcal{P}$. The probability of absorption during the interval $p \in \mathcal{P}$, conditioned on absorption not having occurred is then
$$\hat{\vartheta}_{ij}(p) = \frac{\vartheta_{ij}(p)}{\sum_{p' \ge p} \vartheta_{ij}(p')}. \tag{4.35}$$

2. Directly define the probability $\hat{f}(p)$ as the arithmetic mean of the associated MCG probabil-

ities for all pairs $h^{(i)}$ and $h^{(j)}$ with $i \in \mu$ and $j \in \nu$,

$$\hat{f}(p) = \frac{1}{|\mu||\nu|} \sum_{i \in \mu} \sum_{j \in \nu} \hat{\vartheta}_{ij}(p). \tag{4.36}$$

3. Finally, substituting the derived value of $\hat{f}(p)$ into (4.32) yields the rate $\lambda(p)$, with $\lambda(p_{\mathrm{F}}) = 1$. As before, this rate is independent of the ordering of lineages μ and ν, and so we set

$$\sigma^{(r)}_{\mu\nu}(p) = \lambda(p). \tag{4.37}$$

As before, such posterior lineage rates can be produced for each unordered pair of lineages $\mu, \nu \in \mathcal{L}^{(r)}$, and together provide a pseudo-posterior distribution for the $(r+1)$-th coalescence event $E^{(r+1)}$. Setting $r = 0$, this procedure is equivalent to the procedure described above for determining the first coalescence event. Moreover, when no data are provided, it is evident that $f(p) = \vartheta_{ij}(p) = \int_p e^{-t} dt$ for all $i \in \mu$ and $j \in \nu$, and therefore $\sigma^{(r)}_{\mu\nu}(p) = \lambda(p) = 1$ for all $\mu, \nu \in \mathcal{L}^{(r)}$, yielding the prior process on trees, as desired.

Thus, in order to compute the pairwise pseudo-posterior for an arbitrary locus $\ell \in L$, it is necessary to compute the marginal decoding at locus ℓ associated with the CSP $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_{h^{(i)}} | \mathbf{e}_{h^{(j)}})$ for each unordered pair of distinct haplotypes $h^{(i)}$ and $h^{(j)}$. Given these densities, the above formulation, which involves only elementary arithmetic, can be used to efficiently sample trees from the pseudo-posterior. Importantly, computing the marginal decoding for multiple loci can also be done efficiently by storing additional forward and backward values, particularly using the algorithms described in Section 3.3; marginal trees can then be sampled at each of these loci without the overhead of re-computing each of the pairwise CSPs.

Finally, let $\mu, \nu \in \mathcal{L}^{(r-1)}$ and suppose $\mu, \nu \in \mathcal{L}^{(r)}$ so that neither lineage μ nor ν was involved in the r-th coalescence event $E^{(r)}$. Then by the above description, $\sigma^{(r-1)}_{\mu\nu}(p) = \sigma^{(r)}_{\mu\nu}(p)$ for all $p \in \mathcal{P}$. Similarly, let $\mu_1, \mu_2, \nu \in \mathcal{L}^{(r-1)}$ and suppose that lineages μ_1 and μ_2 are chosen to coalesce into lineage μ in the r-th coalescence event $E^{(r)}$. Then applying the given definitions, for $p < p_{\mathrm{F}} \in \mathcal{P}$,

$$\begin{aligned}
\sigma^{(r)}_{\mu\nu}(p) &= -\frac{1}{|p|} \log \left(1 - \frac{1}{|\mu||\nu|} \sum_{i \in \mu} \sum_{j \in \nu} \hat{\vartheta}_{ij}(p) \right) \\
&= -\frac{1}{|p|} \log \left(1 - \frac{1}{|\mu_1| + |\mu_2|} \Big(\frac{1}{\nu} \sum_{i \in \mu_1} \sum_{j \in \nu} \hat{\vartheta}_{ij}(p) + \frac{1}{\nu} \sum_{i \in \mu_2} \sum_{j \in \nu} \hat{\vartheta}_{ij}(p) \Big) \right) \\
&= -\frac{1}{|p|} \log \left(\frac{1}{|\mu_1| + |\mu_2|} \Big(|\mu_1| \exp(|p| \cdot \sigma^{(r-1)}_{\mu_1\nu}(p)) + |\mu_2| \exp(|p| \cdot \sigma^{(r-1)}_{\mu_2\nu}(p)) \Big) \right).
\end{aligned} \tag{4.38}$$

The rates $\{\sigma^{(r)}_{\mu\nu}(p)\}_{\mu,\nu \in \mathcal{L}^{(r)}}$ can therefore be written in terms of the rates $\{\sigma^{(r-1)}_{\mu\nu}(p)\}_{\mu,\nu \in \mathcal{L}^{(r-1)}}$, and the rates for $r = 0$ are immediate from the marginal decodings. Aside from providing an optimization for sampling from the pairwise pseudo-posterior, this formulation bears some resemblance to the venerable UPGMA algorithm (Sokal and Michener, 1958), used for the construction of ultrametric binary trees (e.g. marginal coalescent trees) given pairwise distances. We might therefore think of the pairwise pseudo-posterior as a stochastic interpretation of the UPGMA algorithm.

4.4.4 Leave-one-out pseudo-posterior

The pairwise pseudo-posterior described in the previous section is straightforward to describe and implement. By construction, however, the posterior rates $\{\sigma^{(r)}_{\mu\nu}(p)\}_{p \in \mathcal{P}}$ for lineages $\mu, \nu \in \mathcal{L}^{(r)}$

are derived by considering only pairs of labeled haplotypes, $h^{(i)}$ and $h^{(j)}$ with $i \in \mu$ and $j \in \nu$. In principle, it should be possible to provide a more accurate pseudo-posterior by considering larger sets of haplotypes, thereby capturing more complex interactions. In this section, we employ the MCGs associated with the CSP $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(h^{(i)}|\mathbf{n}-\mathbf{e}_{h^{(i)}})$ for each labeled haplotype $h^{(i)}$, and estimate the posterior rates $\{\sigma^{(r)}_{\mu\nu}(p)\}_{p\in\mathcal{P}}$ by appropriately combining the MCG posterior distributions. Observe that each such posterior distribution thus involves all haplotypes of $\mathbf{n}$.

Let $h^{(i)}$ be a labeled haplotype within the configuration $\mathbf{n}$ and $\ell \in L$ a specified locus. As described in Section 3.3.3, posterior decoding for the CSP $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_{h^{(i)}}|\mathbf{n}-\mathbf{e}_{h^{(i)}})$ provides a posterior distribution on MCGs at locus ℓ. Denote the corresponding density by $\vartheta_i(\cdot)$, so that $\vartheta_i(p,j)$ is the probability of the lineage associated with haplotype $h^{(i)}$ being absorbed into the trunk lineage associated with haplotype $h^{(j)}$ during the time interval p. These densities, computed for each labeled haplotypes $h^{(i)}$ form the building blocks of the posterior lineage rates.

In order to provide some intuition, consider first approximating the posterior lineage coalescence rates $\{\sigma^{(0)}_{\mu\nu}(p)\}_{p\in\mathcal{P}}$ when no coalescence events have occurred. As before, for an arbitrary pair of distinct lineages $\mu,\nu \in \mathcal{L}^{(0)}$, we may assume without loss of generality that $\mu=\{i\}$ and $\nu=\{j\}$. Then set $f(p,j)=\vartheta_i(p,j)$ and use (4.33) to obtain the associated rates $\{\lambda_j(p)\}_{p\in\mathcal{P}}$. We then define the *directed* lineage coalescence rate $\sigma^{(0)}_{\mu\to\nu}(p)=\lambda_j(p)$ for all $p\in\mathcal{P}$, setting $\lambda(p_{\mathrm{F}})=n-1$. Reversing the indices, we similarly obtain an expression for the directed lineage coalescence rate $\sigma^{(0)}_{\nu\to\mu}(p)$, and finally write, for all $p\in\mathcal{P}$,

$$\sigma^{(0)}_{\mu\nu}(p)=\frac{1}{2}\Big(\sigma^{(0)}_{\mu\to\nu}(p)+\sigma^{(0)}_{\nu\to\mu}(p)\Big). \tag{4.39}$$

Note that we have used an arithmetic mean over the directed lineage coalescence rates. These rates are produced for each unordered pair of lineages $\mu,\nu\in\mathcal{L}^{(0)}$, and together provide a pseudo-posterior distribution for the first coalescence event $E^{(1)}$.

We next consider the more general case, after r coalescences have occurred, and the current set of lineages is given by $\mathcal{L}^{(r)}$. As before, for an arbitrary pair of distinct lineages $\mu,\nu\in\mathcal{L}^{(r)}$, recall that $\mu,\nu\subset\{1,\ldots,n\}$ and $\mu\cap\nu=\emptyset$. As in the initial case, when $r=0$, we define a density $f(\cdot)$ on the space $\mathcal{P}\times\{1,\ldots,n\}$ this time by combining the MCG densities $\vartheta_i(\cdot)$ for each $i\in\mu$ and $j\in\nu$, to determine the directed coalescence rates $\sigma^{(r)}_{\mu\to\nu}(p)$ and $\sigma^{(r)}_{\mu\to\nu}(p)$, respectively; the arithmetic mean of the directed rates is then used to determine the undirected coalescence rate $\sigma^{(r)}_{\mu\nu}(p)$. As before, there are a variety of ways to combine MCG densities, and in the absence of a strong theoretical foundation, we proceed using a technique analogous to the pairwise method described above. For each haplotype $i\in\mu$, we envision an ongoing posterior MCG process, associated with the CSP computation $\hat{\pi}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_{h^{(i)}}|\mathbf{n}-\mathbf{e}_{h^{(i)}})$, and in order to compute the coalescence rate $\sigma^{(r)}_{\mu\to\nu}(p)$, we use the following procedure:

1. Condition each posterior MCG distribution on known information. For $h^{(i)}$ with $i\in\mu$, the known information is the previous coalescence events $(E^{(1)},\ldots,E^{(r)})$ and that absorption has not occurred prior to the current interval p. The probability of absorption into the lineage associated with haplotype $h^{(j)}$ with $j\neq i$ during the interval $p\in\mathcal{P}$, conditioned on known information is then

$$\hat{\vartheta}_i(p,j|E^{(1)},\ldots,E^{(r)})=\frac{\vartheta_i(p,j|E^{(1)},\ldots,E^{(r)})}{\sum_{p'\geq p}\vartheta_i(p',j|E^{(1)},\ldots,E^{(r)})}, \tag{4.40}$$

   where $\hat{\vartheta}_i(\cdot|E^{(1)},\ldots,E^{(r)})$ is the density associated with conditioning on the known coalescence events, and will be discussed in greater detail below.

2. Directly define the probability $\hat{f}(p, j)$ as the arithmetic mean of the associated MCG probabilities for all $h^{(i)}$ with $i \in \mu$.

$$\hat{f}(p, j) = \frac{1}{|\mu|} \sum_{i \in \mu} \hat{\vartheta}_i(p, j | (E^{(1)}, \ldots, E^{(r)})), \tag{4.41}$$

and the probabilities $\hat{f}(p, \nu')$ for each $\nu' \in \mathcal{L}^{(r)} \backslash \{\mu\}$, by summing $\hat{f}(p, j)$ over all $j \in \nu'$,

$$\hat{f}(p, \nu') = \sum_{j \in \nu'} \hat{f}(p, j), \tag{4.42}$$

3. Finally, substituting the derived values of $\hat{f}(p, \nu')$ into (4.33), with $\lambda(p_{\mathrm{F}}) = n - r - 1$, yields the rate $\lambda_\nu(p)$. We then set the directed coalescence rate

$$\sigma^{(r)}_{\mu \to \nu}(p) = \lambda_\nu(p) \tag{4.43}$$

Taking the arithmetic mean of $\sigma^{(r)}_{\mu \to \nu}(p)$ and the similarly deduced $\sigma^{(r)}_{\nu \to \mu}(p)$ then yields the undirected coalescence rate $\sigma^{(r)}_{\mu\nu}(p)$. Such posterior lineage rates can be produced for each unordered pair of lineages $\mu, \nu \in \mathcal{L}^{(r)}$, and together provide a pseudo-posterior distribution for the $(r+1)$-th coalescence event $E^{(r+1)}$. As before, setting $r = 0$ this procedure is equivalent to the procedure described above for determining the first coalescence event. We next describe a method for approximating the conditional MCG density $\vartheta_i(\cdot | E^{(1)}, \ldots, E^{(r)})$.

Conditional MCG density

Recall from the above description that the conditional MCG density $\vartheta_i(\cdot | E^{(1)}, \ldots, E^{(r)})$ is used in order to sample the next coalescence event, $E^{(r+1)}$. Intuitively, then, this density is associated with the MCG process for the next *absorption* event, conditioned the previous coalescence events $(E^{(1)}, \ldots, E^{(r)})$, where each coalescence event $E^{(u)}$ comprises a time $t_u \in \mathbb{R}_{\geq 0}$ and two distinct lineages $\mu_u, \nu_u \in \mathcal{L}^{(u)}$. Starting with the unconditional MCG density, the following mathematically imprecise adjustments are necessary: for each coalescence event $E^{(u)}$ with $1 \leq u \leq r$,

- If $i \in \mu_u$, then in the context of the CSP, the lineage associated with $h^{(i)}$ has been absorbed into the trunk lineages associated with each haplotype $h^{(j)}$ for $j \in \nu_u$ at time t_u. In order to consider the *next* absorption event, it is necessary to disallow absorption prior to the coalescence time t_u, and further to disallow absorption after time t_u into the trunk lineage associated with $h^{(j)}$, for all $j \in \nu_u$. The situation is reversed if $i \in \nu_u$.
- If $i \notin \mu_u$ and $i \notin \nu_u$, then in the context of the CSP, the trunk lineages associated with haplotypes $h^{(i')}$ for $i' \in \mu_u$ and $h^{(j')}$ for $j' \in \nu_u$ should be identified after time t_u.

An appealing and straightforward way to mathematically realize these adjustments is to directly modify the unconditional MCG density, $\vartheta_i(\cdot)$. Note that the events $(E^{(1)}, \ldots, E^{(r)})$ determine the lineage set $\mathcal{L}^{(r)}$, and let $\mu \in \mathcal{L}^{(r)}$ such that $i \in \mu$, and let $p_{\mathrm{c}} \in \mathcal{P}$ such that $t_r \in p_{\mathrm{c}}$. Then from the conditioned MCG density,

$$\vartheta_i(p, j | E^{(1)}, \ldots, E^{(r)}) \propto \begin{cases} 0, & \text{if } p < p_{\mathrm{c}} \text{ or } j \in \mu, \\ \vartheta_i(p, j)/|\nu|, & \text{if } p \geq p_{\mathrm{c}} \text{ and } j \in \nu \in \mathcal{L}^{(r)}, \nu \neq \mu. \end{cases} \tag{4.44}$$

Observe that this formulation appears to have the intended qualitative effects: absorption is formally disallowed where it should be and the probabilities for haplotypes associated with coalescence events not involving $h^{(i)}$ have been appropriately adjusted so that, when summed, give the arithmetic mean probability.

There remain two problems with this mathematical formulation. The first can be observed by considering to the case with no observed data, wherein the total rate of absorption during each time interval should be $n-r-1$ for sampling the $(r+1)$-th coalescence event. As described, the above mathematical formulation yields a total rate of $n-1$, and genealogies sampled using this formulation will have coalescence times that are, on average, too small. This problem can be fixed by considering the alternative CSP $\hat{\pi}^{(r)}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_{h^{(i)}}|\mathbf{n}-\mathbf{e}_{h^{(i)}})$, for which the prior rate of absorption into each trunk lineage is $(n-r-1)/(n-1)$, and so the total rate of absorption is $n-r-1$. We denote the posterior decoding associated with this CSP by $\vartheta^{(r)}_i(p,j)$, and replace $\vartheta_i(p,j)$ with this density on the right-hand side of (4.44).

A second, more subtle, problem is related to the range of coalescence events. The formulation provided in the first line of (4.44) is equivalent to conditioning on the absorption haplotype not being $h^{(j)}$ for any $j \in \mu$ and the absorption interval p being greater than or equal to p_{c}, but only at locus ℓ. In principle, at either locus $\ell-1$ or $\ell+1$, the absorption haplotype may be $h^{(j)}$ for some $j \in \mu$ and the absorption interval p may be less than p_{c}; these possibilities erroneously affect the MCG density $\vartheta^{(r)}_i(p,j)$ at locus ℓ. Suppose that we associate a range $(\ell^{(u)}_{\mathrm{s}}, \ell^{(u)}_{\mathrm{e}})$ with each coalescence event $E^{(u)}$, such that $\ell^{(u)}_{\mathrm{s}} \leq \ell \leq \ell^{(u)}_{\mathrm{e}}$. Then for a coalescence event $E^{(u)}$ with $i \in \mu_u$, it is possible, using an efficient local update to the CSP computation $\hat{\pi}^{(r)}_{\mathrm{SMC}(\mathcal{P})}(\mathbf{e}_{h^{(i)}}|\mathbf{n}-\mathbf{e}_{h^{(i)}})$, to condition on absorption not occurring prior to interval t_u and the absorption haplotype not being $h^{(j)}$ after time t_u for the *entire range* $(\ell^{(u)}_{\mathrm{s}}, \ell^{(u)}_{\mathrm{e}})$. Again, the resulting density on MCGs at locus ℓ can then be substituted into (4.44). The range of a coalescence event can be efficiently sampled by a simple modification to the forward and backward algorithms, once the time lineages associated with the coalescence event have been sampled.

Finally, we comment that the techniques we have proposed to compute the requisite density $\vartheta_i(p,j|E^{(1)},\ldots,E^{(r)})$ are complex, and incorporate several *ad hoc* decisions. The requirement that the overall sampling method reduce to the prior when no data is observed gives some guidance, but still affords many choices. Thus, we believe it is worthwhile to seek out alternatives to this formulation which are, at a minimum, more intuitively appealing and mathematically concise.

4.4.5 Evaluating the pseudo-posterior

We conclude this section by remarking that a key remaining research element in the construction of the pseudo-posterior is a framework for evaluation. In this context, evaluation is challenging for two reasons: first, it is difficult to obtain or sample from the true posterior distribution; and second, the posterior distributions are over trees, which are mathematically complex objects. For the former, it is possible to obtain samples from the true posterior distribution using Monte Carlo methods, such as importance sampling, but this methodology is only practicable for small data sets. Alternatively, for data simulated under the coalescent process, the true marginal tree is known, and so it is possible to compare to pseudo-posterior distribution to the true marginal tree; unfortunately, we have found that, in practice, the posterior distributions under consideration are relatively diffuse, and so it is difficult to draw strong conclusions in this way.

Moreover, comparing distributions on trees is challenging in its own right, particularly because the trees under consideration have continuous-valued lengths. We have considered several lower-

dimensional statistics on trees, such as the time to most recent common ancestor (TMRCA), the partitions induced by the tree at various time points, and the simple tree topology obtained by disregarding branch lengths; for the latter two, we have made use of the existing literature on metrics (Simovici and Jaroszewicz, 2006) on partitions and tree topologies. Using these statistics, it is possible to compute the average distance over a posterior distribution of marginal trees to the true marginal tree and, in conjunction with the Wasserstein (or earth mover) distance (Rueshendorff, 1998), to compare distributions on tree statistics. Although preliminary results using these techniques is promising, there remains considerable research to be done.

Chapter 5

Discussion & Future Work

For much of the history of population genetics, there has been a paucity of genetic data from which to draw concrete conclusions about the mechanisms and natural history of evolution. With the emergence of high-throughput sequencing in the past decade, however, such genetic and genomic data is being produced at an ever-increasing rate. Though evolutionary models, such as the Wright-Fisher diffusion and the coalescent, are a cornerstone of population genetic theory, statistical inference under these models remains a challenging computational problem. To cope with the recent profusion of data, modern population genetic methods must therefore realize a balance between computational efficiency and fidelity to these underlying evolutionary models. A promising class of such methods employ the conditional sampling distribution (CSD).

In this thesis, we have undertaken a theoretical and algorithmic investigation of the CSD for coalescent models including recombination, and made several contributions to this expanding field, including a family of principled CSDs that are both more accurate and more computationally efficient than previously-proposed CSDs. We have also refined and extended two well-known applications of the CSD, and introduced a novel procedure for sampling marginal genealogies from an approximate posterior distribution. In this chapter, we briefly review these contributions, discuss them in the context of both previous and current research in the field, and propose several future research directions.

The CSD $\hat{\pi}_{\mathbf{PS}}$

The motivation for much of our research is the seminal work of Stephens and Donnelly (2000) and De Iorio and Griffiths (2004a,b). The CSD was first introduced in the context of population genetics by the former, and the latter proposed the *diffusion-generator approximation*, by which a one-locus CSD can be algebraically derived directly from the Wright-Fisher diffusion dual to the coalescent model. Importantly, for the special case of a parent independent mutation (PIM) model, the resulting CSD is equal to the true CSD, providing evidence that the approximation is reasonable. The diffusion-generator approximation has been extended to two loci, separated by recombination, by Griffiths et al. (2008); however, the ensuing derivation of the CSD relies on an additional approximation, is limited to PIM models, and cannot be generalized beyond two loci.

In Section 2.1, we described a complete generalization of the diffusion-generator approximation to an arbitrary finite-sites finite-alleles model (Paul and Song, 2010). The ensuing CSD derivation does not require additional approximations, and the resulting CSD, which we denote $\hat{\pi}_{\mathrm{PS}}$, accommodates an arbitrary number of conditionally sampled haplotypes. The generalized

diffusion-generator technique can, in principle, be used to derive an approximate CSD for an arbitrary time-homogeneous coalescent model. To illustrate this point, we have derived variants of the CSD $\hat{\pi}_{\text{PS}}$ for the coalescent with recombination, both with and without population structure and migration, and parameterized by an arbitrary mutation model. For a single locus, the CSD $\hat{\pi}_{\text{PS}}$ is equivalent to the CSD of De Iorio and Griffiths (2004a,b); for two or more loci, however, $\hat{\pi}_{\text{PS}}$ is distinct from all previously-proposed CSDs, including $\hat{\pi}_{\text{FD}}$ (Fearnhead and Donnelly, 2001), $\hat{\pi}_{\text{LS}}$ (Li and Stephens, 2003), and $\hat{\pi}_{\text{GJS}}$ (Griffiths et al., 2008).

In parallel with the generalization of the diffusion-generator approximation, we have introduced an intuitive genealogical process for the CSD $\hat{\pi}_{\text{PS}}$, the *trunk-conditional coalescent*, described in Section 2.2. Provided a collection of previously sampled haplotypes, the trunk-conditional coalescent produces a *conditional genealogy* relating an untyped collection of conditionally sampled haplotypes to each other and the previously sampled haplotypes. A central feature of the trunk-conditional coalescent is the assumption that the unknown genealogy for the collection of previously sampled haplotypes is the *trunk genealogy*, within which haplotypes do not mutate, recombine, or coalesce (see Figure 2.1 for an illustration); lineages of the conditional genealogy are then *absorbed* into the lineages of the trunk. In order to compensate for the trunk genealogy assumption, the rate of non-absorption events within the conditional genealogy are doubled relative to the analogous coalescent process. It is remarkable that this simple genealogical process produces the same CSD, $\hat{\pi}_{\text{PS}}$, as the diffusion-generator approximation.

In contrast to the diffusion-generator approximation, the trunk-conditional coalescent admits a natural extension to time-inhomogeneous population models including variable population size and sub-population splits and merges. Consider, for example, a single panmictic population: time-inhomogeneous population size is incorporated by assuming that the relative population size t time units in the past is given by $\kappa(t)$; the rates of both coalescence and absorption are then *scaled* by the factor $(\kappa(t))^{-1}$. In conjunction with the methods introduced in Section 2.2.3 for incorporating population structure and migration, it is thus possible to obtain a generalization of $\hat{\pi}_{\text{PS}}$ for an arbitrary time-inhomogeneous structured population model, including migration, variable population size, and sub-population splits and mergers. We remark that, although the trunk-conditional coalescent for $\hat{\pi}_{\text{PS}}$ remains well-specified for such time-inhomogeneous models, the methodology introduced in Section 1.3.1 for deriving an explicit recursion for the conditional sampling probability (CSP) is no longer applicable. It is possible, in principle, to extend the recursive framework to explicitly incorporate time, but exact solutions can no longer be obtained. We further discuss such extensions in the context of the sequentially Markov CSD below.

The trunk-conditional coalescent also exposes potential problems with the CSD $\hat{\pi}_{\text{PS}}$. For example, recall from Section 2.2.2 that, upon absorption of a lineage into the trunk genealogy, the allelic type of the absorption haplotype is propagated forward on the lineage; in order to account for the absence of mutations on the trunk lineage, the mutation rate is doubled. Thus, at locus $\ell \in L$, provided the allelic type of the absorption haplotype is $a_1 \in A_\ell$ and the absorption time is t, the probability of conditionally sampling an allele of type $a_2 \in A_\ell$ is given by

$$\xi_\ell(a_2|t, a_1) = \left[e^{t\theta_\ell(\Phi^{(\ell)}-I)}\right]_{a_1,a_2}. \tag{5.1}$$

Recalling the unconditional coalescent, described in Section 1.3.2, a natural mutation process would be to choose an allelic type $a \in A_\ell$ at the time of absorption from the stationary probability conditioned on the allelic type of the absorption haplotype t time units later, and propagate this type forward on the lineage in the conditional genealogy, all at a non-doubled rate. Denoting by

$\phi_\ell(\cdot)$ the stationary density of the mutation process, the associated probability is given by

$$\xi'_\ell(a_2|t,a_1) = \sum_{a\in A_\ell} \frac{\phi_\ell(a)}{\phi_\ell(a_1)} \left[e^{t\frac{\theta_\ell}{2}(\Phi^{(\ell)}-I)}\right]_{a,a_1} \left[e^{t\frac{\theta_\ell}{2}(\Phi^{(\ell)}-I)}\right]_{a,a_2}. \tag{5.2}$$

In general, $\xi_\ell(a_2|t,a_1) \neq \xi'_\ell(a_2|t,a_1)$. However, if the mutation model specified by $\Phi^{(\ell)}$ is *reversible*,

$$\begin{aligned}\xi'_\ell(a_2|t,a_1) &= \sum_{a\in A_\ell} \frac{\phi_\ell(a)}{\phi_\ell(a_1)} \left[e^{t\frac{\theta_\ell}{2}(\Phi^{(\ell)}-I)}\right]_{a,a_1} \left[e^{t\frac{\theta_\ell}{2}(\Phi^{(\ell)}-I)}\right]_{a,a_2} \\ &= \sum_{a\in A_\ell} \left[e^{t\frac{\theta_\ell}{2}(\Phi^{(\ell)}-I)}\right]_{a_1,a} \left[e^{t\frac{\theta_\ell}{2}(\Phi^{(\ell)}-I)}\right]_{a,a_2} = \left[e^{t\theta_\ell(\Phi^{(\ell)}-I)}\right]_{a_1,a_2} = \xi_\ell(a_2|t,a_1),\end{aligned} \tag{5.3}$$

where the second equality follows immediately from reversibility, and the third equality by the Chapman-Kolmogorov equation. Moreover, a large class of reasonable mutation models are reversible, including all 2-locus models, parent independent mutation (PIM) models, and models that are symmetric in the sense that $\Phi_{a_1,a_2} = \Phi_{a_2,a_1}$ for all $a_1, a_2 \in A_\ell$.

A more pressing problem is evident when population structure and migration are incorporated into the trunk-conditional coalescent. As described in Section 2.2.3, the rates of migration in the conditional genealogy are doubled to account for the absence of migration in the trunk genealogy. However, it is not clear whether such a rate-doubled process can be reconciled with the unconditional coalescent with migration, described in Section 1.3.3, which permits all lineages to migrate. For example, consider a biologically plausible model of two demes $\mathcal{D} = \{1, 2\}$, for which migration from deme $1 \in \mathcal{D}$ to deme $2 \in \mathcal{D}$ occurs at a high rate, and in the reverse direction at a low rate: a haplotype sampled in deme 2 must migrate, backward in time, to deme 1 to be absorbed into a haplotype previously sampled from deme 1, a low-probability event. The trunk-conditional coalescent thus discards the high-probability event that absorption actually occurs in deme 2, following migration of the previously sampled haplotype. Though such problems are avoided by selecting deme-symmetric models of population structure, it remains an open problem to extend the trunk-conditional coalescent to gracefully cope with biologically relevant non-symmetric models.

Finally, because the trunk genealogy is time-homogeneous, the rate of absorption into the trunk is constant. Recalling that the trunk genealogy acts as a surrogate for the true unknown genealogy relating the previously sampled individuals, the assumed constant rate of absorption may introduce inaccuracy. Sheehan et al. (2012) suggest retaining the essential form of the trunk genealogy, but altering the rate of absorption in accordance with Kingman's coalescent (Kingman, 1982a). Provided n previously sampled individuals, denote by $A_n(t)$ the prior distribution on the number of lineages ancestral to the n individuals at time t; the total rate of absorption at time t is then taken to be the expected value of $A_n(t)$. Because this expectation, and therefore the absorption rate, is monotonically decreasing in t, the resulting variation on the trunk genealogy is referred to as the *wedding cake genealogy*. Importantly, by adopting such a modified trunk genealogy, the one-locus PIM model, known to be exact for $\hat{\pi}_{\text{PS}}$ in a single panmictic population, is altered, and therefore degraded. More generally, we advise prudence in making *ad hoc* alterations to the trunk-conditional coalescent, as the consequences may be unpredictable and far-reaching.

The CSD $\hat{\pi}_{\text{SMC}}$

As described in Sections 2.1 and 2.2, the CSP associated with the CSD $\hat{\pi}_{\text{PS}}$ is subject to a recursive expression related to the recursive expression for the unconditional sampling probability. In prin-

ciple, explicit evaluation of the CSP is possible by repeated application of the recursive expression, which results in a finite system of coupled linear equations that can be algebraically or numerically solved. We showed in in Section 3.1, however, the number of equations in the system grows super-exponentially with the number of loci, restricting practical application of this method. By making suitable genealogical simplifications to the trunk-conditional coalescent, however, it is possible to obtain approximations to $\hat{\pi}_{\text{PS}}$ with desirable computational properties.

Inspired by the work of Wiuf and Hein (1999) and McVean et al. (2004), we have considered a *sequentially Markov* approximation to the trunk-conditional coalescent (Paul et al., 2011; Steinrücken et al., 2012), described in Section 2.3. At each locus, a conditional genealogy induces a *marginal conditional genealogy* (MCG), relating the conditionally sampled haplotypes to each and to the previously sampled haplotypes at the locus under consideration; due to the process of recombination, the MCGs may be different at distinct loci. The central idea is then to construct a Markov approximation for the sequence of random MCGs. The resulting sequentially Markov CSD is denoted $\hat{\pi}_{\text{SMC}}$, and is provably equivalent to a trunk-conditional coalescent model for which a certain class of coalescence events are disallowed. Importantly, $\hat{\pi}_{\text{SMC}}$ can be cast as a hidden Markov model (HMM), wherein the hidden state at each locus is the MCG at the locus, and the observed state is the associated allelic configuration for the conditionally sampled haplotypes. To illustrate the construction of $\hat{\pi}_{\text{SMC}}$, in Sections 2.3.2–2.3.4 we have derived the requisite HMM densities for the coalescent with recombination, both with and without population structure.

In general, the space of MCGs for the CSD $\hat{\pi}_{\text{SMC}}$ is continuous-valued; consequently, standard HMM methodologies, which require a finite hidden state space, are not immediately applicable. In Section 3.2, we describe a procedure for *discretizing* the continuous space of MCGs into a finite space for a single conditionally sampled haplotype; by increasing the granularity of the discretization, the CSD $\hat{\pi}_{\text{SMC}}$ can be approximated to an arbitrary degree of accuracy. Thus, using standard HMM methodologies, the discretized form of $\hat{\pi}_{\text{SMC}}$ admits efficient computation; for example, evaluating the CSP has time complexity linear in the number of loci, a dramatic improvement over the exponential or super-exponential time complexities associated with $\hat{\pi}_{\text{PS}}$. Moreover, as described in Section 3.3, by specializing the HMM methodologies to the specific densities associated with $\hat{\pi}_{\text{SMC}}$ for a single panmictic population, we obtain optimized algorithms. These optimizations take advantage of structural features common to large genomic samples, including linkage disequilibrium and an abundance of non-polymorphic loci.

The CSD $\hat{\pi}_{\text{FD}}$ (Fearnhead and Donnelly, 2001) can also be cast as an HMM, and directly compared to $\hat{\pi}_{\text{SMC}}$ in the case of a single panmictic population. As described in Section 2.3.5, it is thus possible to interpret $\hat{\pi}_{\text{FD}}$ as a sequentially Markov approximation to $\hat{\pi}_{\text{PS}}$, implicitly requiring two additional approximations: first, the probability of recombination between loci $\ell - 1$ and ℓ is independent of the MCG at the locus $\ell - 1$; and second, conditioned on a recombination event occurring, the distribution of the MCG the locus ℓ is independent of the MCG at the locus $\ell - 1$. In the context of both the unconditional coalescent and the trunk-conditional coalescent process, both of these independence assumptions are fallacious, providing an explanation for the empirically observed deterioration in accuracy relative to $\hat{\pi}_{\text{SMC}}$. The CSD $\hat{\pi}_{\text{LS}}$ (Li and Stephens, 2003) can similarly be interpreted as a sequentially Markov approximation to $\hat{\pi}_{\text{PS}}$, requiring additional approximations that improve computational efficiency, but at further expense of accuracy.

The sequentially Markov approximation can also be applied to the time-inhomogeneous forms of the trunk-conditional coalescent described above. Importantly, it remains possible to construct the key densities associated with the HMM formulation of $\hat{\pi}_{\text{SMC}}$; by further discretizing the con-

tinuous space of MCGs, a form of $\hat{\pi}_{\text{SMC}}$ amenable to efficient evaluation is obtained. Sheehan et al. (2012) have applied this procedure to obtain a CSD for a single panmictic population with time-inhomogeneous size; research on obtaining a generalized form of $\hat{\pi}_{\text{SMC}}$ incorporating multiple populations with migration, and time-inhomogeneity, including sub-population splits and mergers, is also presently underway. We anticipate that the latter will be generally more accurate than previously-proposed CSDs (Price et al., 2009; Hellenthal et al., 2008; Lawson et al., 2012) deriving from $\hat{\pi}_{\text{LS}}$. The algorithmic optimizations described above are not immediately applicable to forms of $\hat{\pi}_{\text{SMC}}$ for coalescent models incorporating complex demography, and a secondary future research direction is the development and application of related optimizations.

When conditionally sampling more than one haplotype, the concrete inference procedures described herein are no longer immediately applicable. In this more general setting, the MCG state space is tree-like. Though it is, in principle, possible to discretize the state space and proceed with inference using the resulting finite space of discretized MCGs, the space grows rapidly with the number of conditionally sampled haplotypes and number of intervals in the discretization; the resulting discrete HMM is thus no longer amenable to efficient computation. There are other possibilities for obtaining a computationally practicable approximation to $\hat{\pi}_{\text{SMC}}$, for example the use of Monte Carlo algorithms such as importance sampling or Markov chain Monte Carlo. Exploring these possibilities is an exciting future research direction. We remark that, in the absence of previously sampled haplotypes, the CSD $\hat{\pi}_{\text{SMC}}$ is identical to the sequentially Markov coalescent, and we believe that recent research (Hobolth et al., 2007; Dutheil et al., 2009; Li and Durbin, 2011) in this area may foster efficient approximations for $\hat{\pi}_{\text{SMC}}$, and vice versa.

Applications

In Section 4.1, we empirically investigated the accuracy and computational efficiency of our proposed CSDs. In general, our CSDs, including $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{SMC}}$, are more accurate than previously proposed CSDs, such as $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$. Importantly, the improvement in accuracy is amplified for increasing numbers of loci. Moreover, using our optimized algorithms for the discretized form of $\hat{\pi}_{\text{SMC}}$, we have demonstrated a substantial computational speed-up relative to standard algorithms used for $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$. Consequently, $\hat{\pi}_{\text{SMC}}$ is a promising candidate for a wide range of CSD-based applications, including those enumerated at the beginning of Chapter 4; we anticipate that, relative to previously proposed CSDs, $\hat{\pi}_{\text{SMC}}$ will produce more accurate results for such applications.

We have explicitly demonstrated the utility of our work in the context of several CSD-based methods. Importance sampling (IS), introduced in the context of the coalescent by Stephens and Donnelly (2000) is one such method, used for both estimation of the sampling probability and ancestral inference. In Section 4.2, we adapted the IS technique introduced by Fearnhead and Donnelly (2001) to use $\hat{\pi}_{\text{SMC}}$, and also proposed two extensions that dramatically improve the efficiency. Interestingly, using $\hat{\pi}_{\text{SMC}}$ in place of $\hat{\pi}_{\text{FD}}$ produces only a minimal improvement in efficiency; we hypothesize that inherent inaccuracy in the IS technique may be overwhelming the improvements in accuracy of $\hat{\pi}_{\text{SMC}}$, and regard further interpretation and improvement as an interesting future research direction. A second well-established application of the CSD is approximate likelihood-based inference of model parameters, particularly using the product of approximate conditionals (PAC) approximate likelihood (Li and Stephens, 2003). In Section 4.3, we use $\hat{\pi}_{\text{SMC}}$, both within the PAC framework and two other composite likelihood frameworks, to estimate migration and recombination rates. We obtain promising results, though estimation of the recombination rate is generally

biased; interpreting and correcting this bias, either in the approximate likelihood framework or in the CSD itself, is another interesting research direction.

Finally, in Section 4.4, we have proposed two novel CSD-based methods for efficiently sampling the marginal genealogy at a particular locus from an approximate posterior distribution. These methods rely on the CSD $\hat{\pi}_{\text{SMC}}$, and the central idea is to directly interpret the posterior distribution on MCGs as a posterior rate of coalescence events. By appropriately combining these posterior rates, it is possible to construct a pseudo-posterior process for marginal genealogies that is analogous to the coalescent prior process. Preliminary results are promising, and fully developing and evaluating the pseudo-posterior process are exciting future research directions. We believe that the pseudo-posterior can be fruitfully used in a variety of application contexts, particularly for questions of ancestral inference, including quantifying identity by descent along the genome, and within case-control association studies for identifying disease correlated polymorphism.

Bibliography

Abramowitz, M. and Stegun, I. A., editors. 1972. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications Inc., New York.

Bhaskar, A. and Song, Y. S. 2012. Closed-form asymptotic sampling distributions under the coalescent with recombination for an arbitrary number of loci. *Advances in Applied Probability*, **44,** 391–407.

Bhaskar, A., Kamm, J. A., and Song, Y. S. 2012. Approximate sampling formulae for general finite-alleles models of mutation. *Advances in Applied Probability*, **44,** 408–428.

Browning, B. L. and Browning, S. R. 2007. Rapid and accurate haplotype phasing and missing data inference for whole genome association studies using localized haplotype clustering. *Am. J. Hum. Genet.*, **81,**(5) 1084–1097.

Cappé, O., Moulines, E., and Ryden, T. 2005. *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer.

Chan, A., Jenkins, P., and Song, Y. S. 2012. Genome-wide fine-scale recombination rate variation in drosophila melanogaster. *PLoS Genet,* in press.

Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A., and Stephens, M. 2004. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat. Genet.*, **36,** 700–706.

Davison, D., Pritchard, J. K., and Coop, G. 2009. An approximate likelihood for genetic data under a model with recombination and population splitting. *Theor. Popul. Biol.*, **75,**(4) 331–345.

De Iorio, M. and Griffiths, R. C. 2004a. Importance sampling on coalescent histories. I. *Adv. in Appl. Probab.*, **36,**(2) 417–433.

De Iorio, M. and Griffiths, R. C. 2004b. Importance sampling on coalescent histories. II: Subdivided population models. *Adv. in Appl. Probab.*, **36,**(2) 434–454.

Donnelly, P. 1986. Dual processes in population genetics. In *Stochastic spatial processes (Heidelberg, 1984)*, volume 1212 of *Lecture Notes in Math.*, pages 94–105. Springer, Berlin.

Dutheil, J. Y., Ganapathy, G., Hobolth, A., Mailund, T., Uoyenoyama, M. K., and Schierup, M. H. 2009. Ancestral population genomics: the coalescent hidden markov model approach. *Genetics*, **183,** 259–274.

Ewens, W. J. 2004. *Mathematical population genetics. I*, volume 27 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, second edition. ISBN 0-387-20191-2. Theoretical introduction.

Fearnhead, P. and Donnelly, P. 2001. Estimating recombination rates from population genetic data. *Genetics*, **159,** 1299–1318.

Fearnhead, P. and Donnelly, P. 2002. Approximate likelihood methods for estimating local recombination rates. *J. Royal Statist. Soc. B*, **64,** 657–680.

Fearnhead, P. and Smith, N. G. 2005. A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *Am. J. Hum. Genet.*, **77,** 781–794.

Felsenstein, J. 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17,**(6) 368–376.

Gay, J., Myers, S. R., and McVean, G. A. T. 2007. Estimating meiotic gene conversion rates from population genetic data. *Genetics*, **177,** 881–894.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., Project, T. . G., and Bustamante, C. D. 2011. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*.

Griffiths, R. C. and Tavaré, S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **344,** 403–410.

Griffiths, R. C., Jenkins, P. A., and Song, Y. S. 2008. Importance sampling and the two-locus model with subdivided population structure. *Adv. in Appl. Probab.*, **40,**(2) 473–500.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. 10 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet*, **5,**(10) e1000695.

Hellenthal, G., Auton, A., and Falush, D. 2008. Inferring human colonization history using a copying model. *PLoS Genet.*, **4,**(5) e1000078.

Hobolth, A., Christensen, O. F., Mailund, T., and Schierup, M. H. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genet*, **3,**(2) e7.

Howie, B. N., Donnelly, P., and Marchini, J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, **5,**(6) e1000529.

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.*, **23,**(2) 183–201.

Hudson, R. R. 2001. Two-locus sampling distributions and their application. *Genetics*, **159,** 1805–1817.

Jenkins, P. A. 2012. Stopping-time resampling and population genetic inference under coalescent models. *Stat. Appl. Genet. Mol. Biol.*, **11,**(1) Article 9.

Jenkins, P. A. and Song, Y. S. 2009. Closed-form two-locus sampling distributions: accuracy and universality. *Genetics*, **183,** 1087–1103.

Jenkins, P. A. and Song, Y. S. 2010. An asymptotic sampling formula for the coalescent with recombination. *Ann. Appl. Probab.*, **20,**(3) 1005–1028.

Jenkins, P. A. and Song, Y. S. 2012. Padè approximants and exact two-locus sampling distributions. *Annals of Applied Probability*, **22,** 576–607.

Karlin, S. and Taylor, H. M. 1981. *A second course in stochastic processes*. Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York. ISBN 0-12-398650-8.

Kingman, J. F. C. 1982a. The coalescent. *Stochastic Process. Appl.*, **13,**(3) 235–248.

Kingman, J. F. C. 1982b. On the genealogy of large populations. *J. Appl. Probab.*, **19A,** 27–43.

Lawson, D., Hellenthal, G., Myers, S., and Falush, D. 2012. Inference of population structure using dense haplotype data. *PLoS Genetics*, **8,**(1) e1002453.

Li, H. and Durbin, R. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, **475,** 493–496.

Li, N. and Stephens, M. 2003. Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics*, **165,** 2213–2233.

Li, Y. and Abecasis, G. R. 2006. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.*, **S79,** 2290.

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. 2010. Mach: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, **34,** 816–834.

Liu, J. S. 2008. *Monte Carlo Strategies in Scientific Computing*. Springer.

Marchini, J., Howie, B., Myers, S. R., McVean, G. A. T., and Donnelly, P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39,**(7) 906–13.

Marjoram, P. and Wall, J. D. 2006. Fast "coalescent" simulation. *BMC Genet.*, **7,** 16.

McVean, G., Awadalla, P., and Fearnhead, P. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. **160,** 1231–1241.

McVean, G. A. T., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. 2004. The fine-scale structure of recombination rate variation in the human genome. *Science*, **304,** 581–584.

McVean, G. A. and Cardin, N. J. 2005. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **360,** 1387–93.

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. 2011. Genotype and snp calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12,** 443–451.

Notohara, M. 1990. The coalescent and the genealogical process in geographically structured population. *J. Math. Biol.*, **29,**(1) 59–75.

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

Paul, J. S. and Song, Y. S. 2012. Blockwise HMM computation for large-scale population genomic inference. *Bioinformatics*, **28,** 2008–2015.

Paul, J. S., Steinrücken, M., and Song, Y. S. 2011. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics*, **187,** 1115–1128.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D., and Myers, S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.*, **5,**(6) e1000519.

Rosenblatt, M. 1959. Functions of a markov process that are markovian. *J. Math. Mech.*, **8,** 585–596.

Rueshendorff, L. 1998. Wasserstein metric. In Hazewinkel, M., editor, *Encyclopedia of Mathematics*. Springer.

Scheet, P. and Stephens, M. 2006. A fast and flexible method for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78,**(4) 629–644.

Sheehan, S., Harris, K., and Song, Y. S. 2012. Estimating variable effective population sizes from multiple genomes: A sequentially markov conditional sampling distribution approach. *in preparation*.

Simovici, D. A. and Jaroszewicz, S. 2006. A new metric splitting criterion for decision trees. *International Journal of Parallel, Emergent and Distributed Systems*, **21,** 239–256.

Sloane, N. 1998. Bell numbers. In Hazewinkel, M., editor, *Encyclopedia of Mathematics*. Springer.

Sokal, R. R. and Michener, C. D. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, **28,** 1409–1438.

Steinrücken, M., Paul, J. S., and Song, Y. S. 2012. An efficient conditional sampling distribution for structured populations exchanging migrants. *Theoretical Population Biology,* in press.

Stephens, M. and Donnelly, P. 2000. Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **62,**(4) 605–655.

Stephens, M. and Scheet, P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.*, **76,**(3) 449–462.

Sundquist, A., Fratkin, E., Do, C. B., and Batzoglou, S. 2008. Effect of genetic divergence in identifying ancestral origin using hapaa. *Genome Research*, **18,** 676–682.

Wang, Y. and Hey, J. 2010. Estimating divergence parameters with small samples from a large number of loci. *Genetics*, **184,**(2) 363–379.

Wegmann, D., Kessner, D. E., Veeramah, K. R., Mathias, R. A., Nicolae, D. L., Yanek, L. R., Sun, Y. V., Torgerson, D. G., Rafaels, N., Mosley, T., Becker, L. C., Ruczinski, I., Beaty, T. H., Kardia, S. L. R., Meyers, D. A., Barnes, K. C., Becker, D. M., Freimer, N. B., and Novembre, J. 2011. Recombination rates in admixed individuals identified by ancestry-based inference. *Nat. Genet.*, **43,** 847–853.

Wiuf, C. and Hein, J. 1999. Recombination as a point process along sequences. *Theor. Pop. Biol.*, **55,** 248–259.

Wright, S. 1949. Adaptation and selection. In Jepson, G. L., Mayr, E., and Simpson, G. G., editors, *Genetics, Paleontology and Evolution*, pages 365–389. Princeton University Press.

Yin, J., Jordan, M. I., and Song, Y. S. 2009. Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. *Bioinformatics*, **25,**(12) i231–i239.

Appendix A

Table of Common Notation

Fully-specified haplotypes (Section 1.1)	
L	The set of loci, $L = \{1, \ldots, k\}$, where k is the number of loci.
B	The set of breakpoints, $B = \{(1,2), \ldots, (k-1,k)\}$.
A_ℓ	The set of alleles at locus $\ell \in L$.
$\mathcal{H}$	The space of fully-specified haplotypes, $\mathcal{H} = A_1 \times \cdots \times A_k$.
$h[\ell] \in A_\ell$	The allele at locus $\ell \in L$ of $h \in \mathcal{H}$. More generally, $h[\ell_{\text{s}} : \ell_{\text{e}}]$ is the sub-haplotype for the loci ℓ, $\ell_{\text{s}} \leq \ell \leq \ell_{\text{e}}$.
$\mathcal{M}^a_\ell(h) \in \mathcal{H}$	For $\ell \in L$, the haplotype derived form $h \in \mathcal{H}$ by setting $h[\ell] = a$.
$\mathcal{R}_b(h, h') \in \mathcal{H}$	For $b = (\ell, \ell+1) \in B$, the haplotype derived by joining sub-haplotype $h[1:\ell]$ with sub-haplotype $h'[\ell+1:k]$.
Partially-specified haplotypes (Section 1.1)	
$\bullet$	An unspecified allele.
$\mathcal{G}$	The space of partially-specified haplotypes, $\mathcal{G} = (A_1 \cup \{\bullet\}) \times \cdots \times (A_k \cup \{\bullet\})$.
$L(g)$	Given $g \in \mathcal{G}$, the set of specified loci in g.
$B(g)$	Given $g \in \mathcal{G}$, the set of recombination breakpoints between the left- and right-most specified loci.
$g \curlywedge g'$	Given $g, g' \in \mathcal{G}$, a binary relation indicating compatibility.
$\mathcal{C}(g, g') \in \mathcal{G}$	Given $g, g' \in \mathcal{G}$ with $g \curlywedge g'$, the haplotype derived from g and g' by merging the two haplotypes, as defined in (1.1).
$\mathcal{M}_\ell(g) \in \mathcal{G}$	Given $\ell \in L(g)$, the haplotype derived from $g \in \mathcal{G}$ by setting $g[\ell] = \bullet$.
$\mathcal{R}^-_b(g) \in \mathcal{G}$	Given $b \in B(g)$, the haplotype derived from $g \in \mathcal{G}$ by joining the sub-haplotype $g[1,\ell]$ with the complementary sub-haplotype of unspecified alleles. The reverse construction is used for $\mathcal{R}^+_b(g) \in \mathcal{G}$.
Haplotype configurations (Sections 1.1 and 1.3.1)	
$\mathbf{n} = (n_h)_{h \in \mathcal{H}}$	A fully-specified haplotype configuration comprising n_h haplotypes of type h. Similarly, a partially-specified configuration $\mathbf{n} = (n_g)_{g \in \mathcal{G}}$. We frequently assume an arbitrary ordering or labeling of the constituent haplotypes.

$\mathbf{n}[\ell]$	The one-locus configuration induced by haplotype configuration $\mathbf{n}$ at locus $\ell \in L$. More generally, $\mathbf{n}[\ell_{\mathrm{s}} : \ell_{\mathrm{e}}]$ is the configuration induced by $\mathbf{n}$ for the set of loci ℓ such that $\ell_{\mathrm{s}} \leq \ell \leq \ell_{\mathrm{e}}$.
$\hat{n}$	An untyped configuration comprising untyped (or place-holder) haplotypes, often including additional ancestral information. We frequently assume an arbitrary ordering or labeling of the constituent untyped haplotypes.
$\lvert\mathbf{n}\rvert, \lvert\hat{n}\rvert$	The number of haplotypes in a typed or untyped configuration.
	Structured haplotype configurations (Section 1.2.3)
$\mathcal{D}$	A finite set of demes, $\mathcal{D} = \{1, \ldots, q\}$.
$\mathbf{n} = (n_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$	A fully-specified structured haplotype configuration comprising $n_{d,h}$ haplotypes of type h in deme d. We frequently assume an arbitrary ordering or labeling of the constituent haplotypes within each deme.
$\mathbf{n}_d$	The haplotype configuration in deme $d \in \mathcal{D}$ induced by $\mathbf{n} = (n_{d,h})_{d\in\mathcal{D},h\in\mathcal{H}}$.
	Parameters (Sections 1.2.2 and 1.2.3)
θ_ℓ	The scaled mutation rate at locus $\ell \in L$.
$\Phi^{(\ell)}$	The $\lvert A_\ell\rvert \times \lvert A_\ell\rvert$-dimension stochastic matrix governing mutations.
ρ_b	The scaled recombination rate at breakpoint $b \in B$.
κ_d	The relative size of deme $d \in \mathcal{D}$, such that $\sum_{d\in\mathcal{D}} \kappa_d = 1$.
$\upsilon_{dd'}$	The scaled migration rate, backward in time, from deme $d \in \mathcal{D}$ to deme $d' \in \mathcal{D}$ with $d' \neq d$. We also write $\upsilon_d = \sum_{d'\in\mathcal{D}} \upsilon_{dd'}$.
	Genealogies (Sections 1.3.1, 1.3.2, and 2.2.1)
$\mathcal{A}_{\hat{n}}$	An untyped genealogy associated with untyped configuration $\hat{n}$. Similarly, $\mathcal{A}_{\mathbf{n}}$ is a typed genealogy associated with typed configuration $\mathbf{n}$.
$\mathcal{C}_{\hat{c}}$	An untyped conditional genealogy (including absorption events) associated with untyped configuration $\hat{c}$. Similarly, $\mathcal{C}_{\mathbf{c}}$ is a typed conditional genealogy associated with typed configuration $\mathbf{c}$.
$\mathcal{A}_0(\mathbf{n})$	The improper typed trunk genealogy (including no genealogical events) associated with typed configuration $\mathbf{n}$.
$\mathcal{A}_{\hat{n}}[\ell], \mathcal{C}_{\hat{c}}[\ell]$	The marginal genealogy and conditional genealogy, respectively, induced by $\mathcal{A}_{\hat{n}}$ and $\mathcal{C}_{\hat{c}}$ at locus ℓ. More generally, $\mathcal{A}_{\hat{n}}[\ell_{\mathrm{s}} : \ell_{\mathrm{e}}]$ and $\mathcal{C}_{\hat{c}}[\ell_{\mathrm{s}} : \ell_{\mathrm{e}}]$ are the marginal genealogies induced by the set of loci ℓ such that $\ell_{\mathrm{s}} \leq \ell \leq \ell_{\mathrm{e}}$.
	Genealogical processes (Section 1.3.1)
E_i, U_i, V_i	The i-th random genealogical event, backward in time, and the untyped and typed configurations after the i-th genealogical event. Note that a particular typed configuration $V_i = v$ entails the corresponding untyped configuration $U_i = u$.

$p(\cdot\|u)$	The density of events E_i conditioned on $U_{i-1} = u$. The support is given by $\mathcal{E}(u)$, and given $E_i = e \in \mathcal{E}(u)$, the untyped configuration $U_{i-1} = e(u)$ is uniquely determined.
$p(\cdot\|v, e)$	The density of typed configurations V_i conditioned on $V_{i+1} = v$ and $E_i = e$. The support is given by $\mathcal{V}(v, e)$.
	Sequentially Markov CSD (Sections 2.3.1 and 2.3.2)
$\mathcal{S}$	The space of marginal conditional genealogies (MCGs) associated with a particular trunk-conditional coalescent model.
S_ℓ	The random MCG at locus $\ell \in L$, without mutation events.
T_ℓ, H_ℓ	The random absorption time and haplotype associated with S_ℓ for a single conditionally sampled haplotype.
$\zeta^{(\mathbf{n})}(\cdot)$	The marginal density on the MCG S_ℓ for all $\ell \in L$.
$\phi_b^{(\mathbf{n})}(\cdot\|s_{\ell-1})$	The density on MCG S_ℓ conditioned on $S_{\ell-1} = s_{\ell-1}$, and provided $b = (\ell - 1, \ell) \in B$. Used as the transition density for $\hat{\pi}_{\text{SMC}}$.
$\xi_\ell^{(\mathbf{n})}(\cdot\|s_\ell)$	The density on emitted alleles at locus $\ell \in L$ conditioned on $S_\ell = s_\ell$.
	Discretization for $\hat{\pi}_{\text{SMC}}$ (Sections 3.2.1 and 3.3.2)
$\mathcal{P}$	A discretization of $\mathbb{R}_{\geq 0}$. Letting $0 = \tau_0 < \tau_1 < \cdots < \tau_m = \infty$ be a strictly increasing sequence, $\mathcal{P} = \{[\tau_{j-1}, \tau_j)\}_{j=1,\ldots,m}$.
$\ddot{\mathcal{S}}$	The space of *discretized* marginal conditional genealogies (MCGs) associated with a particular trunk-conditional coalescent model.
$\mathcal{C}$	A configuration partition $\mathcal{C} = \{(\mathcal{B}, \ell_{\text{s}}, \ell_{\text{e}})\}$ where $\mathcal{B} \subset \mathcal{H}$ and $1 \leq \ell_{\text{s}} \leq \ell_{\text{e}} \leq k$ such that each locus of each haplotype in a configuration $\mathbf{n}$ is represented in precisely one block $(\mathcal{B}, \ell_{\text{s}}, \ell_{\text{e}})$. $\mathcal{C} = C_{\text{T}}$ is the *trivial* configuration partition comprising a single block for each haplotype.
$\mathcal{C}_\ell$	The partition of haplotypes induced by $\mathcal{C}$ at a particular locus $\ell \in L$.
$\Psi(\mathcal{C})$	Given a configuration partition $\mathcal{C}$, $\Psi(\mathcal{C}) = \sum_{\ell \in L} \lvert\mathcal{C}_\ell\rvert$ is a summation of the size of the $\mathcal{C}$-induced haplotype partition at each locus. Similarly, $\Psi_{\text{p}}(\mathcal{C})$ is a summation over only polymorphic loci.
$\Omega(\mathcal{C})$	Given a configuration partition $\mathcal{C}$, $\Omega(\mathcal{C}) = \sum_{(\mathcal{B}, \ell_{\text{s}}, \ell_{\text{e}}) \in \mathcal{C}} \lvert\mathcal{B}\rvert$ is a summation of the number of haplotypes in each block of $\mathcal{C}$.

Appendix B

Longer Proofs

B.1 Proof of equivalence of $\hat{\pi}_{\mathbf{NC}}$ and $\hat{\pi}_{\mathbf{SMC}}$

Recalling the definition (2.71) of the forward probability $f_\ell^{(\mathbf{e}_\eta,\mathbf{n})}(s_\ell)$, we define the generalized forward probability $f_{\ell',\ell}(\eta, s_\ell)$, which describes the joint probability of observing loci $\ell' : \ell$ of η and $S_\ell = s_\ell$

$$f_{\ell',\ell}(\eta, s_\ell) = \xi_\ell(\mathbf{c}[\ell]|s_\ell) \cdot \int_{\mathcal{S}} \phi_{(\ell-1,\ell)}(s_\ell|s_{\ell-1}) \cdot f_{\ell',\ell-1}(\eta, s_{\ell-1})\, ds_{\ell-1}, \tag{B.1}$$

for $\ell' < \ell$, with base case

$$f_{\ell,\ell}(\eta, s_\ell) = \xi_\ell^{(\mathbf{n})}(\mathbf{c}[\ell]|s_\ell) \cdot \zeta(s_\ell), \tag{B.2}$$

where the marginal, transition, and emission densities are provided in (2.73), (2.74), and (2.75), respectively. Observe that $f_\ell^{(\mathbf{e}_\eta,\mathbf{n})}(s_\ell) = f_{1,\ell}(\eta, s_\ell)$. For notational convenience, we have suppressed dependence on $\mathbf{n}$ in the generalized forward density, and moved the dependence on η into the function. We now provide a more detailed proof of Theorem 2.14 from the main paper.

Proof of Theorem 2.14. We begin by showing inductively that, for $\ell, \ell' \in L$ with $\ell' \leq \ell$ and $s_\ell \in \mathcal{S}$, the probability $f_{\ell',\ell}(\eta, s_\ell)$ is equal to the probability $g_{\ell,\ell'}(\eta, s_\ell)$, defined by the following genealogical recursion [c.f. Griffiths and Tavaré (1994)],

$$\begin{aligned} g_{\ell',\ell}(\eta, s_\ell) = \int_{t_{\mathrm{e}}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)} t_{\mathrm{e}}} \Bigg[& n_{h_\ell} \delta_{\eta,h_\ell}^{(\ell':\ell)} \delta_{t_{\mathrm{e}},t_\ell} \\ & + \sum_{\lambda \in L(\ell':\ell)} \theta_\lambda \sum_{a \in A_\lambda} \Phi_{a,\eta[\lambda]}^{(\lambda)} g_{\ell',\ell}(\mathcal{M}_\lambda^a(\eta), s_\ell - t_{\mathrm{e}}) \\ & + \sum_{\beta \in B(\ell':\ell)} \rho_\beta \Bigg(\int_{s_{\ell_{\mathrm{s}}} \in \mathcal{S}} g_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}}) \Bigg) g_{\ell_{\mathrm{e}},\ell}(\eta, s_\ell - t_{\mathrm{e}}) \Bigg], \end{aligned} \tag{B.3}$$

where the $\mathcal{N}_{(\ell',\ell)}$ is the $(\ell' : \ell)$-restricted rate of events,

$$\mathcal{N}_{(\ell',\ell)} = n + \sum_{\lambda \in L(\ell':\ell)} \theta_\lambda + \sum_{\beta \in B(\ell':\ell)} \rho_\beta. \tag{B.4}$$

For notational convenience, we have adopted the following conventions: given MCG $s_\ell \in \mathcal{S}$ and $t \in \mathbb{R}_{\geq 0}$, we write $s_\ell - t = (t_\ell - t, h_\ell) \in \mathcal{S}$; similarly, we express the $(\ell' : \ell)$-restricted delta function

$\delta^{(\ell':\ell)}_{\eta,h_\ell} = \delta_{\eta[\ell':\ell],h_\ell[\ell':\ell]}$; finally, we set $\beta = (\ell_\mathrm{s}, \ell_\mathrm{e}) \in B$, and that $b = (\ell - 1, \ell) \in B$. Setting $\ell' = \ell$,

$$g_{\ell,\ell}(\eta, s_\ell) = \int_{t_\mathrm{e}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell,\ell)}t_\mathrm{e}} \left[n_{h_\ell} \delta^{(\ell)}_{\eta,h_\ell} \delta_{t_\mathrm{e},t_\ell} + \theta_\ell \sum_{a \in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} g_{\ell,\ell}(\mathcal{M}^a_\ell(\eta), s_\ell - t_\mathrm{e}) \right]. \tag{B.5}$$

Substituting $g_{\ell,\ell} = f_{\ell,\ell}$ on the right-hand side,

$$\begin{aligned}
&\int_{t_\mathrm{e}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell,\ell)}t_\mathrm{e}} \left[n_{h_\ell} \delta^{(\ell)}_{\eta,h_\ell} \delta_{t_\mathrm{e},t_\ell} + \theta_\ell \sum_{a \in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} f_{\ell,\ell}(\mathcal{M}^a_\ell(\eta), s_\ell - t_\mathrm{e}) \right] \\
&\quad = e^{-\mathcal{N}_{(\ell,\ell)}t_\ell} n_{h_\ell} \delta^{(\ell)}_{\eta,h_\ell} + \int_{t_\mathrm{e}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell,\ell)}t_\mathrm{e}} \theta_\ell \sum_{a \in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} \xi_\ell(a | s_\ell - t_\mathrm{e}) \zeta(s_\ell - t_\mathrm{e}) \\
&\quad = n_{h_\ell} e^{-\mathcal{N}_{(\ell,\ell)}t_\ell} \left(\delta^{(\ell)}_{\eta,h_\ell} + \sum_{m=0}^{\infty} \left(\sum_{a \in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} \left[(\Phi^{(\ell)})^m \right]_{h_\ell[\ell],a} \right) \int_{t_\mathrm{e}=0}^{t_\ell} \theta_\ell \frac{\left(\theta_\ell (t_\ell - t_\mathrm{e})\right)^m}{m!} \right) \\
&\quad = n_{h_\ell} e^{-\mathcal{N}_{(\ell,\ell)}t_\ell} \left(\delta^{(\ell)}_{\eta,h_\ell} + \sum_{m=0}^{\infty} \left[(\Phi^{(\ell)})^{m+1} \right]_{h_\ell[\ell],\eta[\ell]} \frac{\left(\theta_\ell t_\ell\right)^{m+1}}{(m+1)!} \right) \\
&\quad = \xi_\ell(s_\ell) \zeta(s_\ell) = f_{\ell,\ell}(\eta, s_\ell),
\end{aligned} \tag{B.6}$$

Thus, $f_{\ell,\ell}$ satisfies the recursion for $g_{\ell,\ell}$, and so we conclude that $f_{\ell,\ell} = g_{\ell,\ell}$. Inductively assuming that $f_{\ell',\ell} = g_{\ell',\ell}$ for all $\ell, \ell' \in L$ such that $0 \le \ell - \ell' < j$, let $\ell', \ell \in L$ such that $\ell - \ell' = j$. Substituting $g_{\ell',\ell} = f_{\ell',\ell}$ on the right-hand side of (B.3), we obtain

$$\begin{aligned}
\int_{t_\mathrm{e}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)}t_\mathrm{e}} \Bigg[& n_{h_\ell} \delta^{(\ell':\ell)}_{\eta,h_\ell} \delta_{t_\mathrm{e},t_\ell} \\
& + \sum_{\lambda \in L(\ell':\ell)} \theta_\lambda \sum_{a \in A_\lambda} \Phi^{(\lambda)}_{a,\eta[\lambda]} f_{\ell',\ell}(\mathcal{M}^\eta_\lambda(a), s_\ell - t_\mathrm{e}) \\
& + \sum_{\beta \in B(\ell':\ell)} \rho_\beta \left(\int_{s_{\ell_\mathrm{s}} \in \mathcal{S}} f_{\ell',\ell_\mathrm{s}}(\eta, s_{\ell_\mathrm{s}}) \right) f_{\ell_\mathrm{e},\ell}(\eta, s_\ell - t_\mathrm{e}) \Bigg].
\end{aligned} \tag{B.7}$$

We consider this expression one term at a time. Beginning with the first term:

$$\begin{aligned}
&\int_{t_\mathrm{e}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)}t_\mathrm{e}} n_{h_\ell} \delta^{(\ell':\ell)}_{\eta,h_\ell} \delta_{t_\mathrm{e},t_\ell} \\
&\qquad = \int_{s_{\ell-1} \in \mathcal{S}} \int_{t_\mathrm{e}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_\mathrm{e}} n_{h_{\ell-1}} \delta^{(\ell':\ell-1)}_{\eta,h_{\ell-1}} \delta_{t_\mathrm{e},t_{\ell-1}} \left[e^{-(\theta_\ell + \rho_b)t_\mathrm{e}} \delta^{(\ell)}_{\eta,h_\ell} \delta_{s_{\ell-1},s_\ell} \right].
\end{aligned} \tag{B.8}$$

Moving on to the second term of (B.7), expand using the definition (B.1) of $f_{\ell',\ell}$, and apply the

inductive hypothesis to replace the resulting $f_{\ell',\ell-1}$ terms with the corresponding $g_{\ell',\ell-1}$ terms:

$$\begin{aligned}
&\int_{t_{\mathrm{e}}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)}t_{\mathrm{e}}} \sum_{\lambda\in L(\ell':\ell)} \theta_\lambda \sum_{a\in A_\lambda} \Phi^{(\lambda)}_{a,\eta[\lambda]} f_{\ell',\ell}(\mathcal{M}^a_\lambda(\eta), s_\ell - t_{\mathrm{e}}) \\
&\quad = \int_{t_{\mathrm{e}}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)}t_{\mathrm{e}}} \sum_{\lambda\in L(\ell':\ell-1)} \theta_\lambda \sum_{a\in A_\lambda} \Phi^{(\lambda)}_{a,\eta[\lambda]} \\
&\qquad \times \xi_\ell(\eta[\ell]|s_\ell - t_{\mathrm{e}}) \int_{s_{\ell-1}\in\mathcal{S}} \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1}) g_{\ell',\ell-1}(\mathcal{M}^a_\lambda(\eta), s_{\ell-1}) ds_{\ell-1} dt_{\mathrm{e}} \\
&\quad + \int_0^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)}t_{\mathrm{e}}} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} \\
&\qquad \times \xi_\ell(a|s_\ell - t_{\mathrm{e}}) \int_{\mathcal{S}} \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1}) g_{\ell',\ell-1}(\eta, s_{\ell-1}).
\end{aligned} \tag{B.9}$$

Concentrating on the first sub-term of (B.9), making the substitution $t_{\ell-1} \to t_{\ell-1}+t_{\mathrm{e}}$, and changing the order of integration, we obtain

$$\begin{aligned}
&\int_{s_{\ell-1}\in\mathcal{S}} \int_{t_{\mathrm{e}}=0}^{t_\ell\wedge t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} \sum_{\lambda\in L(\ell':\ell-1)} \theta_\lambda \sum_{a\in A_\lambda} \Phi^{(\lambda)}_{a,\eta[\lambda]} g_{\ell',\ell-1}(\mathcal{M}^a_\lambda(\eta), s_{\ell-1} - t_{\mathrm{e}}) \\
&\qquad \times \left[e^{-\theta_\ell t_{\mathrm{e}}} \xi_\ell(\eta[\ell]|s_\ell - t_{\mathrm{e}}) \cdot e^{-\rho_b t_{\mathrm{e}}} \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1} - t_{\mathrm{e}}) \right].
\end{aligned} \tag{B.10}$$

Now concentrating on the second sub-term of (B.9) and expanding using definition (B.3) of $g_{\ell',\ell-1}$:

$$\begin{aligned}
&\int_{t_{\mathrm{e}}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)}t_{\mathrm{e}}} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} \xi_\ell(a|s_\ell - t_{\mathrm{e}}) \int_{s_{\ell-1}\in\mathcal{S}} \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1}) \\
&\qquad \times \int_{t_q=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_q} \Bigg[n_{h_{\ell-1}} \delta^{(\ell':\ell-1)}_{\eta,h_{\ell-1}} \delta_{t_q,t_{\ell-1}} \\
&\qquad\quad + \sum_{\lambda\in L(\ell':\ell-1)} \theta_\lambda \sum_{a\in A_\lambda} \Phi^{(\lambda)}_{a,\eta[\lambda]} g_{\ell',\ell-1}(\mathcal{M}^a_\lambda(\eta), s_{\ell-1} - t_q) \\
&\qquad\quad + \sum_{\beta\in B(\ell':\ell-1)} \rho_\beta \left(\int_{s_{\ell_{\mathrm{s}}}\in\mathcal{S}} g_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}}) \right) g_{\ell_{\mathrm{e}},\ell-1}(\eta, s_{\ell-1} - t_q) \Bigg] \\
&= \int_{s_{\ell-1}\in\mathcal{S}} \int_{t_q=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_q} \Bigg[n_{h_{\ell-1}} \delta^{(\ell':\ell-1)}_{\eta,h_{\ell-1}} \delta_{t_q,t_{\ell-1}} \\
&\qquad\quad + \sum_{\lambda\in L(\ell':\ell-1)} \theta_\lambda \sum_{a\in A_\lambda} \Phi^{(\lambda)}_{a,\eta[\lambda]} g_{\ell',\ell-1}(\mathcal{M}^a_\lambda(\eta), s_{\ell-1} - t_q) \\
&\qquad\quad + \sum_{\beta\in B(\ell':\ell-1)} \rho_\beta \left(\int_{s_{\ell_{\mathrm{s}}}\in\mathcal{S}} g_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}}) \right) g_{\ell_{\mathrm{e}},\ell-1}(\eta, s_{\ell-1} - t_q) \Bigg] \\
&\qquad \times \left[\int_{t_{\mathrm{e}}=0}^{t_q\wedge t_\ell} e^{-\theta_\ell t_{\mathrm{e}}} \theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} \xi_\ell(a|s_\ell - t_{\mathrm{e}}) \cdot e^{-\rho_b t_{\mathrm{e}}} \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1} - t_{\mathrm{e}}) \right],
\end{aligned} \tag{B.11}$$

with the equality obtained by making the substitutions $t_{\ell-1} \to t_{\ell-1} + t_{\mathrm{e}}$ and $t_q \to t_q + t_{\mathrm{e}}$ and then changing the order of integration. Finally, moving onto the third term of (B.7), expand using the definition of $f_{\ell',\ell-1}$, and apply the inductive hypothesis to replace the resulting $f_{\ell',\ell-1}$ terms with the corresponding $g_{\ell',\ell-1}$ terms:

$$\begin{aligned}
&\int_{t_{\mathrm{e}}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)}t_{\mathrm{e}}} \sum_{\beta\in B(\ell':\ell)} \rho_b\Big(\int_{s_{\ell_{\mathrm{s}}}\in\mathcal{S}} f_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}})\Big) f_{\ell_{\mathrm{e}},\ell}(\eta, s_\ell - t_{\mathrm{e}}) \\
&\quad= \int_{t_{\mathrm{e}}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)}t_{\mathrm{e}}} \sum_{\beta\in B(\ell':\ell-1)} \rho_\beta\Big(\int_{s_{\ell_{\mathrm{s}}}\in\mathcal{S}} g_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}})\Big) \\
&\qquad\times \xi_\ell(\eta[\ell]|s_\ell - t_{\mathrm{e}}) \int_{s_{\ell-1}\in\mathcal{S}} \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1}) g_{\ell_{\mathrm{e}},\ell-1}(\eta, s_{\ell-1}) \\
&\qquad+ \int_{t_{\mathrm{e}}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)}t_{\mathrm{e}}} \rho_b\Big(\int_{s_{\ell-1}\in\mathcal{S}} g_{\ell',\ell-1}(\eta, s_{\ell-1})\Big) \cdot g_\ell(\eta, s_\ell - t_{\mathrm{e}}).
\end{aligned} \tag{B.12}$$

Concentrating on the first sub-term of (B.12), making the substitution $t_{\ell-1} \to t_{\ell-1} + t_{\mathrm{e}}$, and changing the order of integration, we obtain:

$$\begin{aligned}
&\int_{s_{\ell-1}\in\mathcal{S}} \int_{t_{\mathrm{e}}=0}^{t_\ell \wedge t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} \sum_{\beta\in B(\ell':\ell-1)} \rho_\beta\Big(\int_{s_{\ell_{\mathrm{s}}}\in\mathcal{S}} g_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}})\Big) g_{\ell_{\mathrm{e}},\ell-1}(\eta, s_{\ell-1} - t_{\mathrm{e}}) \\
&\quad\times \Big[e^{-\theta_\ell t_{\mathrm{e}}} \xi_\ell(\eta[\ell]|s_\ell - t_{\mathrm{e}}) \cdot e^{-\rho_b t_{\mathrm{e}}} \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1} - t_{\mathrm{e}})\Big].
\end{aligned} \tag{B.13}$$

Now concentrating on the second sub-term of (B.12) and expanding using definition (B.3) of $g_{\ell',\ell-1}$:

$$\begin{aligned}
&\int_{t_{\mathrm{e}}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)}t_{\mathrm{e}}} \rho_b g_\ell(\eta, s_\ell - t_{\mathrm{e}}) \\
&\qquad\times \int_{s_{\ell-1}\in\mathcal{S}} \int_{t_q=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_q} \Big[n_{h_{\ell-1}} \delta_{\eta,h_{\ell-1}}^{(\ell':\ell-1)} \delta_{t_q,t_{\ell-1}} \\
&\qquad\quad+ \sum_{\lambda\in L(\ell':\ell-1)} \theta_\lambda \sum_{a\in A_\lambda} \Phi_{a,\eta[\lambda]}^{(\lambda)} g_{\ell',\ell-1}(\mathcal{M}_\lambda^a(\eta), s_{\ell-1} - t_q) \\
&\qquad\quad+ \sum_{\beta\in B(\ell':\ell-1)} \rho_\beta\Big(\int_{s_{\ell_{\mathrm{s}}}\in\mathcal{S}} g_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}})\Big) g_{\ell_{\mathrm{e}},\ell-1}(\eta, s_{\ell-1} - t_q)\Big] \\
&= \int_{s_{\ell-1}\in\mathcal{S}} \int_{t_q=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_q} \Big[n_{h_{\ell-1}} \delta_{\eta,h_{\ell-1}}^{(\ell':\ell-1)} \delta_{t_q,t_{\ell-1}} \\
&\qquad\quad+ \sum_{\lambda\in L(\ell':\ell-1)} \theta_\lambda \sum_{a\in A_\lambda} \Phi_{a,\eta[\lambda]}^{(\lambda)} g_{\ell',\ell-1}(\mathcal{M}_\lambda^a(\eta), s_{\ell-1} - t_q) \\
&\qquad\quad+ \sum_{\beta\in B(\ell':\ell-1)} \rho_\beta\Big(\int_{s_{\ell_{\mathrm{s}}}\in\mathcal{S}} g_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}})\Big) g_{\ell_{\mathrm{e}},\ell-1}(\eta, s_{\ell-1} - t_q)\Big] \\
&\qquad\times \Big[\int_{t_{\mathrm{e}}=0}^{t_q \wedge t_\ell} e^{-\theta_\ell t_{\mathrm{e}}} \xi_\ell(\eta[\ell]|s_\ell - t_{\mathrm{e}}) \cdot e^{-\rho_b t_{\mathrm{e}}} \rho_b n_{h_\ell} e^{-n(t_\ell - t_{\mathrm{e}})}\Big],
\end{aligned} \tag{B.14}$$

with the equality obtained by using the base definition (B.2) for $f_{\ell,\ell}$, making the substitutions $t_{\ell-1} \to t_{\ell-1} + t_{\mathrm{e}}$ and $t_q \to t_q + t_{\mathrm{e}}$, and changing the order of integration.

Having expanded each term of our key expression (B.7), aggregate common terms across the resulting sub-expressions. Collecting the $n_{h_{\ell-1}}\delta^{(\ell':\ell-1)}_{\eta,h_{\ell-1}}$ terms from (B.8),(B.11), and (B.14),

$$
\begin{aligned}
&\int_{s_{\ell-1}\in\mathcal{S}}\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} n_{h_{\ell-1}}\delta^{(\ell':\ell-1)}_{\eta,h_{\ell-1}}\delta_{t_{\mathrm{e}},t_{\ell-1}} \\
&\quad\times\Bigg[e^{-(\theta_\ell+\rho_b)t_{\mathrm{e}}}\delta^{(\ell)}_{\eta,h_\ell}\delta_{s_{\ell-1},s_\ell} \\
&\qquad+\int_{t_q=0}^{t_{\mathrm{e}}\wedge t_\ell} e^{-\theta_\ell t_q}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,\eta[\ell]}\xi_\ell(a|s_\ell-t_q)\cdot e^{-\rho_b t_q}\phi_b(s_\ell-t_q|s_{\ell-1}-t_q) \\
&\qquad+\int_{t_q=0}^{t_{\mathrm{e}}\wedge t_\ell} e^{-\theta_\ell t_q}\xi_\ell(\eta[\ell]|s_\ell-t_q)\cdot e^{-\rho_b t_q}\rho_b n_{h_\ell}e^{-n(t_\ell-t_q)}\Bigg] \\
&=\int_{s_{\ell-1}\in\mathcal{S}}\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} n_{h_{\ell-1}}\delta^{(\ell':\ell-1)}_{\eta,h_{\ell-1}}\delta_{t_{\mathrm{e}},t_{\ell-1}} \\
&\quad\times\Bigg[e^{-\rho_b t_{\ell-1}}\delta_{s_{\ell-1},s_\ell}\cdot\left(e^{-\theta_\ell t_\ell}\delta^{(\ell)}_{\eta,h_\ell}\right) \\
&\qquad+e^{-\rho_b t_{\ell-1}}\delta_{s_{\ell-1},s_\ell}\left(\int_{t_z=0}^{t_\ell} e^{-\theta_\ell t_z}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,\eta[\ell]}\xi_\ell(a|s_\ell-t_z)\right) \\
&\qquad+\int_{t_q=0}^{t_{\ell-1}\wedge t_\ell}\rho_b e^{-\rho_b t_q}n_{h_\ell}e^{-n(t_\ell-t_q)}\left(\int_{t_z=0}^{t_q} e^{-\theta_\ell t_z}\theta_\ell\sum_{a\in A_\ell}\Phi^{(\ell)}_{a,\eta[\ell]}\xi_\ell(a|s_\ell-t_z)\right) \\
&\qquad+\int_{t_q=0}^{t_{\ell-1}\wedge t_\ell}\rho_b e^{-\rho_b t_q}n_{h_\ell}e^{-n(t_\ell-t_q)}\left(e^{-\theta_\ell t_q}\xi_\ell(\eta[\ell]|s_\ell-t_q)\right)\Bigg] \\
&=\int_{s_{\ell-1}\in\mathcal{S}}\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} n_{h_{\ell-1}}\delta^{(\ell':\ell-1)}_{\eta,h_{\ell-1}}\delta_{t_{\mathrm{e}},t_{\ell-1}} \\
&\quad\times\xi_\ell(\eta[\ell]|s_\ell)\left[e^{-\rho_b t_{\ell-1}}\delta_{s_{\ell-1},s_\ell}+\int_{t_q=0}^{t_{\ell-1}\wedge t_\ell}\rho_b e^{-\rho_b t_q}n_{h_\ell}e^{-n(t_\ell-t_q)}\right] \\
&=\int_{s_{\ell-1}\in\mathcal{S}}\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} n_{h_{\ell-1}}\delta^{(\ell':\ell-1)}_{\eta,h_{\ell-1}}\delta_{t_{\mathrm{e}},t_{\ell-1}}\times\left[\xi_\ell(\eta[\ell]|s_\ell)\phi_b(s_\ell|s_{\ell-1})\right],
\end{aligned}
\tag{B.15}
$$

where the first equality is obtained by making use of the $\delta_{t_{\mathrm{e}},t_{\ell-1}}$ and $\delta_{s_{\ell-1},s_\ell}$ expressions and expanding $\phi_b(s_\ell-t_q|s_{\ell-1}-t_q)$ using (2.74) and exchanging integrals, the second equality is obtained by combining the first/second and third/fourth term and using a straightforward identity for $\xi_\ell(\eta[\ell]|s_\ell)$, and final equality by again making use of the (2.74). Similarly, collecting the $g_{\ell',\ell-1}(\mathcal{M}^a_\ell(\eta), s_{\ell-1}-t_q)$

terms from the resulting sub-expressions (B.10),(B.11), and (B.14),

$$
\begin{aligned}
&\int_{s_{\ell-1}\in\mathcal{S}}\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} \sum_{\lambda\in L(\ell':\ell-1)} \theta_\lambda \sum_{a\in A_\lambda} \Phi^{(\lambda)}_{a,\eta[\lambda]} g_{\ell',\ell-1}(\mathcal{M}^a_\lambda(\eta), s_{\ell-1}-t_{\mathrm{e}}) \\
&\quad\times \Bigg[\mathbb{I}_{(t_{\mathrm{e}}\le t_\ell)} e^{-\theta_\ell t_{\mathrm{e}}} \xi_\ell(\eta[\ell]|s_\ell - t_{\mathrm{e}}) \cdot e^{-\rho_b t_{\mathrm{e}}} \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1}-t_{\mathrm{e}}) \\
&\qquad + \int_{t_q=0}^{t_{\mathrm{e}}\wedge t_\ell} e^{-\theta_\ell t_q}\theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]} \xi_\ell(a|s_\ell - t_q)\cdot e^{-\rho_b t_q}\phi_b(s_\ell - t_q|s_{\ell-1}-t_q) \\
&\qquad + \int_{t_q=0}^{t_{\mathrm{e}}\wedge t_\ell} e^{-\theta_\ell t_q}\xi_\ell(\eta[\ell]|s_\ell - t_q)\cdot e^{-\rho_b t_q}\rho_b n_{h_\ell} e^{-n(t_\ell - t_q)}\Bigg] \\
&= \int_{s_{\ell-1}\in\mathcal{S}}\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} \sum_{\lambda\in L(\ell':\ell-1)} \theta_\lambda \sum_{a\in A_\lambda} \Phi^{(\lambda)}_{a,\eta[\ell]} g_{\ell',\ell-1}(\mathcal{M}^a_\lambda(\eta), s_{\ell-1}-t_{\mathrm{e}}) \\
&\quad\times \Bigg[\mathbb{I}_{(t_{\mathrm{e}}\le t_\ell)} e^{-\rho_b t_{\mathrm{e}}} \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1}-t_{\mathrm{e}}) \Big(e^{-\theta_\ell t_{\mathrm{e}}}\xi_\ell(\eta[\ell]|s_\ell - t_{\mathrm{e}})\Big) \\
&\qquad + \mathbb{I}_{(t_{\mathrm{e}}\le t_\ell)} e^{-\rho_b t_{\mathrm{e}}} \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1}-t_{\mathrm{e}}) \Big(\int_{t_z=0}^{t_{\mathrm{e}}} e^{-\theta_\ell t_z}\theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)} a,\eta[\ell]\xi_\ell(a|s_\ell - t_z)\Big) \qquad \text{(B.16)} \\
&\qquad + \int_{t_q=0}^{t_{\mathrm{e}}\wedge t_\ell} \rho_b e^{-\rho_b t_q} n_{h_\ell} e^{-n(t_\ell - t_q)} \Big(\int_{t_z=0}^{t_q} e^{-\theta_\ell t_z}\theta_\ell \sum_{a\in A_\ell} \Phi^{(\ell)}_{a,\eta[\ell]}\xi_\ell(a|s_\ell - t_z)\Big) \\
&\qquad + \int_{t_q=0}^{t_{\mathrm{e}}\wedge t_\ell} \rho_b e^{-\rho_b t_q} n_{h_\ell} e^{-n(t_\ell - t_q)} \Big(e^{-\theta_\ell t_q}\xi_\ell(\eta[\ell]|s_\ell - t_q)\Big)\Bigg] \\
&= \int_{s_{\ell-1}\in\mathcal{S}}\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} \sum_{\lambda\in L(\ell':\ell-1)} \theta_\lambda \sum_{a\in A_\lambda} \Phi^{(\lambda)}_{a,\eta[\ell]} g_{\ell',\ell-1}(\mathcal{M}^a_\ell(\eta), s_{\ell-1}-t_{\mathrm{e}}) \\
&\quad\times \xi_\ell(\eta[\ell]|s_\ell)\Bigg[\mathbb{I}_{(t_{\mathrm{e}}\le t_\ell)} e^{-\rho_b t_{\mathrm{e}}}\phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1}-t_{\mathrm{e}}) + \int_{t_q=0}^{t_{\mathrm{e}}\wedge t_\ell} \rho_b e^{-\rho_b t_q} n_{h_\ell} e^{-n(t_\ell - t_q)}\Bigg] \\
&= \int_{s_{\ell-1}\in\mathcal{S}}\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} \sum_{\lambda\in L(\ell':\ell-1)} \theta_\lambda \sum_{a\in A_\lambda} \Phi^{(\lambda)}_{a,\eta[\lambda]} g_{\ell',\ell-1}(\mathcal{M}^a_\ell(\eta), s_{\ell-1}-t_{\mathrm{e}}) \\
&\quad\times \Big[\xi_\ell(\eta[\ell]|s_\ell)\phi_b(s_\ell|s_{\ell-1})\Big],
\end{aligned}
$$

where the first equality is obtained by using the following expansion for $\phi_b(s_\ell - t_q|s_{\ell-1}-t_q)$,

$$
\begin{aligned}
\phi_b(s_\ell - t_q|s_{\ell-1}-t_q) = \mathbb{I}_{(t_{\mathrm{e}}\le t_\ell)} e^{-\rho_b(t_{\mathrm{e}}-t_q)}\cdot \phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1}-t_{\mathrm{e}}) \\
+ \int_{t_z=0}^{(t_{\mathrm{e}}\wedge t_\ell)-t_q} \rho_b e^{-\rho_b t_z} n_{h_\ell} e^{-n(t_\ell - t_q - t_z)},
\end{aligned}
$$

which can be verified in the present context, namely that $t_q \le t_{\mathrm{e}} \le t_{\ell-1}$ and $t_q \le t_\ell$. The second equality is obtained by combining the first/second and third/fourth term and using a straightforward identity for $\xi_\ell(\eta[\ell]|s_\ell)$, and the final equality by once again appealing to the above identity.

Collecting the $g_{\ell_{\mathrm{e}},\ell-1}(\eta, s_{\ell-1}-t_q)$ terms from (B.13), (B.11), and (B.14):

$$\begin{aligned}
&\int_{s_{\ell-1}\in\mathcal{S}}\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} \sum_{\beta\in B(\ell':\ell-1)} \rho_\beta\Big(\int_{s_{\ell_{\mathrm{s}}}} g_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}})\Big) g_{\ell_{\mathrm{e}},\ell-1}(\eta, s_{\ell-1}-t_{\mathrm{e}}) \\
&\quad\times\Bigg[\mathbb{I}_{(t_{\mathrm{e}}\le t_\ell)} e^{-\theta_\ell t_{\mathrm{e}}}\xi_\ell(\eta[\ell]|s_\ell - t_{\mathrm{e}})\cdot e^{-\rho_b t_{\mathrm{e}}}\phi_b(s_\ell - t_{\mathrm{e}}|s_{\ell-1}-t_{\mathrm{e}}) \\
&\qquad+\int_{t_q=0}^{t_{\mathrm{e}}\wedge t_\ell} e^{-\theta_\ell t_q}\theta_\ell \sum_{a\in A_\ell}\Phi^{(\ell)}_{a,\eta[\ell]}\xi_\ell(a|s_\ell - t_q)\cdot e^{-\rho_b t_q}\phi_b(s_\ell - t_q|s_{\ell-1}-t_q) \\
&\qquad+\int_{t_q=0}^{t_{\mathrm{e}}\wedge t_\ell} e^{-\theta_\ell t_q}\xi_\ell(\eta[\ell]|s_\ell - t_q)\cdot e^{-\rho_b t_q}\rho_b n_{h_\ell} e^{-n(t_\ell - t_q)}\Bigg] \\
&=\int_{s_{\ell-1}\in\mathcal{S}}\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}} \sum_{\beta\in B(\ell':\ell-1)} \rho_\beta\Big(\int_{s_{\ell_{\mathrm{s}}}\in\mathcal{S}} g_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}})\Big) g_{\ell_{\mathrm{e}},\ell-1}(\eta, s_{\ell-1}-t_{\mathrm{e}}) \\
&\quad\times\Big[\xi_\ell(\eta[\ell]|s_\ell)\phi_b(s_\ell|s_{\ell-1})\Big].
\end{aligned} \tag{B.17}$$

Thus, combining equations (B.15),(B.16), and (B.17), we may re-write (B.7):

$$\begin{aligned}
&\xi_\ell(\eta[\ell]|s_\ell)\int_{s_{\ell-1}\in\mathcal{S}}\phi_b(s_\ell|s_{\ell-1})\cdot\int_{t_{\mathrm{e}}=0}^{t_{\ell-1}} e^{-\mathcal{N}_{(\ell',\ell-1)}t_{\mathrm{e}}}\Bigg[n_{h_{\ell-1}}\delta^{(\ell':\ell-1)}_{\eta,h_{\ell-1}}\delta_{t_{\mathrm{e}},t_{\ell-1}} \\
&\qquad+\sum_{\lambda\in L(\ell':\ell-1)}\theta_\lambda\sum_{a\in A_\lambda}\Phi^{(\lambda)}_{a,\eta[\lambda]} g_{\ell',\ell-1}(\mathcal{M}^a_\lambda(\eta), s_{\ell-1}-t_{\mathrm{e}}) \\
&\qquad+\sum_{\beta\in B(\ell':\ell-1)}\rho_\beta\Big(\int_{s_{\ell_{\mathrm{s}}}\in\mathcal{S}} g_{\ell',\ell_{\mathrm{s}}}(\eta, s_{\ell_{\mathrm{s}}})\Big) g_{\ell_{\mathrm{e}}:\ell-1}(\eta, s_{\ell-1}-t_{\mathrm{e}})\Bigg] \\
&=\xi_\ell(\eta[\ell]|s_\ell)\int_{s_{\ell-1}\in\mathcal{S}}\phi_b s_\ell|s_{\ell-1}) g_{\ell',\ell-1}(\eta, s_{\ell-1}) \\
&=f_{\ell',\ell}(\eta, s_\ell),
\end{aligned} \tag{B.18}$$

where the first equality is obtained by definition (B.3) for $g_{\ell',\ell-1}$, and the second equality by using the inductive hypothesis and definition (B.1). Thus, $f_{\ell',\ell}$ satisfies the recursion for $g_{\ell',\ell}$, and so we

conclude that $f_{\ell',\ell} = g_{\ell',\ell}$. Moreover,

$$\begin{aligned}
\int_{s_\ell \in \mathcal{S}} g_{\ell',\ell}(\eta, s_\ell) &= \int_{s_\ell \in \mathcal{S}} \int_{t_\mathrm{e}=0}^{t_\ell} e^{-\mathcal{N}_{(\ell',\ell)} t_\mathrm{e}} \Bigg[n_{h_\ell} \delta_{\eta,h_\ell}^{(\ell':\ell)} \delta_{t_\mathrm{e},t_\ell} \\
&\quad + \sum_{\lambda \in L(\ell':\ell)} \theta_\lambda \sum_{a \in A_\lambda} \Phi_{a,\eta[\lambda]}^{(\lambda)} g_{\ell',\ell}(\mathcal{M}_\lambda^a(\eta), s_\ell - t_\mathrm{e}) \\
&\quad + \sum_{\beta \in B(\ell':\ell)} \rho_\beta \Bigg(\int_{s_{\ell_\mathrm{s}} \in \mathcal{S}} g_{\ell',\ell_\mathrm{s}}(\eta, s_{\ell_\mathrm{s}}) \Bigg) g_{\ell_\mathrm{e},\ell}(\eta, s_\ell - t_\mathrm{e}) \Bigg] \\
&= \frac{1}{\mathcal{N}_{(\ell',\ell)}} \Bigg[\sum_{\substack{h \in \mathcal{H}: \\ h[\ell':\ell] = \eta[\ell':\ell]}} n_h \\
&\quad + \sum_{\lambda \in L(\eta[\ell':\ell])} \theta_\lambda \sum_{a \in A_\lambda} \Phi_{a,\eta[\lambda]}^{(\lambda)} \int_{s_\ell \in \mathcal{S}} g_{\ell',\ell}(\mathcal{M}_\lambda^a(\eta), s_\ell) \\
&\quad + \sum_{\beta \in B(\eta[\ell':\ell])} \rho_\beta \int_{s_{\ell_\mathrm{s}} \in \mathcal{S}} g_{\ell',\ell_\mathrm{s}}(\eta, s_{\ell_\mathrm{s}}) \int_{s_\ell \in \mathcal{S}} g_{\ell_\mathrm{e},\ell}(\eta, s_\ell) \Bigg],
\end{aligned} \tag{B.19}$$

where the first equality is by definition (B.3), and the second equality obtained by exchanging the integrals and making the substitution $t_\ell \to t_\ell - t_\mathrm{e}$. Thus, $\int_{s_\ell \in \mathcal{S}} g_{\ell',\ell}(\eta, s_\ell)$ satisfies the recursion (2.59) for $\hat{\pi}_{\mathrm{NC}}(\mathbf{e}_\eta[\ell', \ell] | \mathbf{n}[\ell', \ell])$ and we conclude that $\int_{s_\ell \in \mathcal{S}} g_{\ell',\ell}(\eta, s_\ell) = \hat{\pi}_{\mathrm{NC}}(\mathbf{e}_\eta[\ell' : \ell] | \mathbf{n}[\ell', \ell])$. Thus,

$$\hat{\pi}_{\mathrm{SMC}}(\mathbf{e}_\eta | \mathbf{n}) = \int_{s_k \in \mathcal{S}} f_{1,k}(\eta, s_k) = \int_{s_k \in \mathcal{S}} g_{1,k}(\eta, s_k) = \hat{\pi}_{\mathrm{NC}}(\mathbf{e}_\eta | \mathbf{n}), \tag{B.20}$$

thereby establishing the desired identity. □

B.2 Proof of detailed balance for two-haplotype $\hat{\pi}_{\mathrm{SMC}}$

We have shown that the Markov process associated with one-haplotype CSD $\hat{\pi}_{\mathrm{SMC}}$, governed by transition density $f_{b,0}^{(a)}$, satisfies detailed balance with respect to the marginal density $f_0^{(f)}$. We begin by stating and proving a general form of this result as a proposition, preceded by two minor lemmas. Recalling the definitions of Section 2.3.4,

Lemma B.1. *Let $m_\ell = (t_\ell, h_\ell) \in \mathcal{M}$. For $t, t' < t_\ell$,*

$$f_t^{(f)}(m_\ell) = e^{-n(t'-t)} f_{t'}^{(f)}(m_\ell).$$

As a consequence, letting $m_{\ell-1} = (t_{\ell-1}, h_{\ell-1}) \in \mathcal{M}$, then for $t, t' < t_{\ell-1}, t_\ell$,

$$f_t^{(f)}(m_{\ell-1}) \cdot f_{t'}^{(f)}(m_\ell) = f_t^{(f)}(m_\ell) \cdot f_{t'}^{(f)}(m_{\ell-1}).$$

Proof. Using expression (2.92) for $f_t^{(f)}$,

$$e^{-n(t'-t)} f_{t'}^{(f)}(m_\ell) = e^{-n(t'-t)} \cdot n_{h_\ell} e^{-n(t_\ell - t')} = n_{h_\ell} e^{-n(t_\ell - t)} = f_t^{(f)}(m_\ell).$$ □

Lemma B.2. *Let $m_{\ell-1}, m_\ell \in \mathcal{M}$. Then for $t < t_{\ell-1}, t_\ell, t'$ and $t' < t_\ell$,*

$$f_{b,t}^{(a)}(m_\ell | m_{\ell-1}) = \left[\mathbb{1}_{(t' < t_{\ell-1})} \right] \cdot e^{-\rho(t'-t)} \cdot f_{b,t'}^{(a)}(m_\ell | m_{\ell-1}) + \int_t^{t' \wedge t_{\ell-1}} \rho e^{-\rho(t_r - t)} \cdot f_{t_r}^{(f)}(m_\ell) dt_r.$$

Proof. Beginning with the right hand side, and using expression (2.93) to expand $f^{(a)}_{b,t'}(m_\ell|m_{\ell-1})$,

$$\begin{aligned}
&\left[\mathbb{1}_{(t'<t_{\ell-1})}\right] \cdot e^{-\rho(t'-t)} \cdot f^{(a)}_{b,t'}(m_\ell|m_{\ell-1}) + \int_t^{t'\wedge t_{\ell-1}} \rho e^{-\rho(t_{\text{r}}-t)} \cdot f^{(f)}_{t_{\text{r}}}(m_\ell) dt_{\text{r}} \\
&\quad = e^{-\rho(t_{\ell-1}-t)} \cdot \delta_{m_\ell, m_{\ell-1}} + \left[\mathbb{1}_{(t'<t_{\ell-1})}\right] \cdot \int_{t'}^{t_{\ell-1}\wedge t_\ell} \rho e^{-\rho(t_{\text{r}}-t)} \cdot n_{h_\ell} e^{-n(t_\ell - t_{\text{r}})} dt_{\text{r}} \\
&\qquad + \int_t^{t'\wedge t_{\ell-1}} \rho e^{-\rho(t_{\text{r}}-t)} \cdot f^{(f)}_{t_{\text{r}}}(m_\ell) dt_{\text{r}} \\
&\quad = e^{-\rho(t_{\ell-1}-t)} \cdot \delta_{m_\ell, m_{\ell-1}} + \int_t^{t_\ell \wedge t_{\ell-1}} \rho e^{-\rho(t_{\text{r}}-t)} \cdot f^{(f)}_{t_{\text{r}}}(m_\ell) dt_{\text{r}} \\
&\quad = f^{(a)}_{b,t}(m_\ell|m_{\ell-1}).
\end{aligned}$$

□

Proposition B.3. *Let $m_\ell, m_{\ell-1} \in \mathcal{M}$. Then for $t, t' < t_{\ell-1}, t_\ell$, the following detailed balance condition holds for the densities $f^{(a)}_{b,t}$ and $f^{(f)}_{t'}$:*

$$f^{(a)}_{b,t}(m_\ell|m_{\ell-1}) \cdot f^{(f)}_{t'}(m_{\ell-1}) = f^{(a)}_{b,t}(m_{\ell-1}|m_\ell) \cdot f^{(f)}_{t'}(m_\ell).$$

Proof. Using the expression (2.93) to expand $f^{(a)}_{b,t}(m_\ell|m_{\ell-1})$, and applying Lemma B.1,

$$\begin{aligned}
&f^{(a)}_{b,t}(m_\ell|m_{\ell-1}) \cdot f^{(f)}_{t'}(m_{\ell-1}) \\
&\quad = e^{-\rho(t_{\ell-1}-t)} \cdot \delta_{m_\ell, m_{\ell-1}} \cdot f^{(f)}_{t'}(m_{\ell-1}) + \left[\int_t^{t_{\ell-1}\wedge t_\ell} \rho e^{-\rho(t_{\text{r}}-t)} \cdot f^{(f)}_{t_{\text{r}}}(m_\ell) dt_{\text{r}}\right] \cdot f^{(f)}_{t'}(m_{\ell-1}) \\
&\quad = e^{-\rho(t_\ell - t)} \cdot \delta_{m_{\ell-1}, m_\ell} \cdot f^{(f)}_{t'}(m_\ell) + \left[\int_t^{t_\ell \wedge t_{\ell-1}} \rho e^{-\rho(t_{\text{r}}-t)} \cdot f^{(f)}_{t_{\text{r}}}(m_{\ell-1}) dt_{\text{r}}\right] \cdot f^{(f)}_{t'}(m_\ell) \\
&\quad = f^{(a)}_{b,t}(m_{\ell-1}|m_\ell) \cdot f^{(f)}_{t'}(m_\ell).
\end{aligned}$$

□

We now move on the analogous detailed balance result for the two-haplotype CSD $\hat{\pi}_{\text{SMC}}$. We begin by defining an auxiliary distribution, and using it to relate the previously defined distributions in a series of lemmas. The final lemma provides a condition that is analogous to Lemma B.1. The auxiliary distribution is associated with sampling the conditional genealogy s_ℓ conditioned on the marginal conditional genealogy $m^{(2)}_\ell$, and starting at time t. Denoting the density $f^{(1|2)}_t$,

$$\begin{aligned}
f^{(1|2)}_t(s_\ell|m^{(2)}_\ell) = &\left[1 - \delta_{t^{(\text{c})}_\ell, \emptyset}\right] \cdot 2e^{-(n+2)(t^{(\text{c})}_\ell - t)} \\
&+ \left[\mathbb{1}_{(t^{(1)}_\ell < t^{(2)}_\ell)}\right] \cdot e^{-(n+2)(t^{(1)}_\ell - t)} n_{h^{(1)}_\ell} \\
&+ \left[\mathbb{1}_{(t^{(1)}_\ell > t^{(2)}_\ell)}\right] \cdot e^{-(n+2)(t^{(2)}_\ell - t)} \cdot f^{(f)}_{t^{(2)}_\ell}(m^{(1)}_\ell)
\end{aligned} \tag{B.21}$$

Then we can immediately establish the following simple lemma

Lemma B.4. *Let $s_\ell \in \mathcal{S}$. Then for $t, t' < s_\ell$,*

$$f^{(1|2)}_t(m^{(1)}_\ell, t^{(c)}_\ell | m^{(2)}_\ell) = e^{-(n+2)(t'-t)} \cdot f^{(1|2)}_{t'}(m^{(1)}_\ell, t^{(c)}_\ell | m^{(2)}_\ell).$$

As a consequence, letting $s_{\ell-1} \in \mathcal{S}$, then for $t, t' < s_{\ell-1}, s_\ell$,

$$f^{(1|2)}_t(m^{(1)}_\ell, t^{(c)}_\ell | m^{(2)}_\ell) \cdot f^{(1|2)}_{t'}(m^{(1)}_{\ell-1}, t^{(c)}_{\ell-1} | m^{(2)}_{\ell-1}) = f^{(1|2)}_t(m^{(1)}_{\ell-1}, t^{(c)}_{\ell-1} | m^{(2)}_{\ell-1}) \cdot f^{(1|2)}_{t'}(m^{(1)}_\ell, t^{(c)}_\ell | m^{(2)}_\ell).$$

Proof. As in the proof of Lemma B.1, this is a simple algebraic identity. □

The next two lemmas provide two simple sampling relations. In order to (unconditionally) sample the MCG s_ℓ, it is sufficient to first (unconditionally) sample $m_\ell^{(2)}$, and then sample s_ℓ conditioned on $m_\ell^{(2)}$. Similarly, to sample the MCG s_ℓ conditioned on $m_{\ell-1}^{(2)}$, it is sufficient to first sample $m_\ell^{(2)}$ conditioned on $m_{\ell-1}^{(2)}$, and then sample the MCG s_ℓ conditioned on $m_\ell^{(2)}$.

Lemma B.5. *Let $s_\ell \in \mathcal{S}$. Then for $t < s_\ell$,*

$$f_t^{(f,f)}(s_\ell) = f_t^{(1|2)}(s_\ell|m_\ell^{(2)}) \cdot f_t^{(f)}(m_\ell^{(2)}).$$

Proof. Expanding factor $f_t^{(1|2)}(s_\ell|m_\ell^{(2)})$ using expression (B.21), and applying Lemma B.1,

$$\begin{aligned}
&f_t^{(1|2)}(s_\ell|m_\ell^{(2)}) \cdot f_t^{(f)}(m_\ell^{(2)}) \\
&\quad = \left[1 - \delta_{t_\ell^{(\mathrm{c})},\emptyset}\right] \cdot 2e^{-(2n+2)(t_\ell^{(\mathrm{c})} - t)} \cdot f_{t_\ell^{(\mathrm{c})}}^{(f)}(m_\ell) \\
&\qquad + \left[\mathbb{1}_{(t_\ell^{(1)} < t_\ell^{(2)})}\right] \cdot e^{-(2n+2)(t_\ell^{(1)} - t)} n_{h_\ell^{(1)}} \cdot f_{t_\ell^{(1)}}^{(f)}(m_\ell^{(2)}) \\
&\qquad + \left[\mathbb{1}_{(t_\ell^{(1)} > t_\ell^{(2)})}\right] \cdot e^{-(2n+2)(t_\ell^{(2)} - t)} n_{h_\ell^{(2)}} \cdot f_{t_\ell^{(2)}}^{(f)}(m_\ell^{(1)}) \\
&\quad = f_t^{(f,f)}(s_\ell),
\end{aligned}$$

where the final equality is by (2.94). □

Lemma B.6. *Let $s_\ell \in \mathcal{S}$ and $m_{\ell-1}^{(2)} \in \mathcal{M}$. Then for $t < s_\ell, t_{\ell-1}^{(2)}$,*

$$f_{b,t}^{(f,a)}(s_\ell|m_{\ell-1}^{(2)}) = f_t^{(1|2)}(s_\ell|m_\ell^{(2)}) \cdot f_{b,t}^{(a)}(m_\ell^{(2)}|m_{\ell-1}^{(2)}).$$

Proof. Expanding factors $f_t^{(1|2)}(s_\ell|m_\ell^{(2)})$ and $f_{b,t}^{(a)}(m_\ell^{(2)}|m_{\ell-1}^{(2)})$ using expression (B.21) and Lemma B.2, respectively, and recollecting those terms containing integrals,

$$\begin{aligned}
&f_t^{(1|2)}(s_\ell|m_\ell^{(2)}) \cdot f_{b,t}^{(a)}(m_\ell^{(2)}|m_{\ell-1}^{(2)}) \\
&\quad = \left[1 - \delta_{t_\ell^{(\mathrm{c})},\emptyset}\right] \cdot 2e^{-(n+2)(t_\ell^{(\mathrm{c})} - t)} \cdot \left(\mathbb{1}_{(t_\ell^{(\mathrm{c})} < t_{\ell-1}^{(2)})} \cdot e^{-\rho(t_\ell^{(\mathrm{c})} - t)} f_{b,t_\ell^{(\mathrm{c})}}^{(a)}(m_\ell|m_{\ell-1}^{(2)})\right) \\
&\qquad + \left[\mathbb{1}_{(t_\ell^{(1)} < t_\ell^{(2)})}\right] \cdot e^{-(n+2)(t_\ell^{(1)} - t)} n_{h_\ell^{(1)}} \cdot \left(\mathbb{1}_{(t_\ell^{(1)} < t_{\ell-1}^{(2)})} \cdot e^{-\rho(t_\ell^{(1)} - t)} f_{b,t_\ell^{(1)}}^{(a)}(m_\ell^{(2)}|m_{\ell-1}^{(2)})\right) \\
&\qquad + \left[\mathbb{1}_{(t_\ell^{(1)} > t_\ell^{(2)})}\right] \cdot e^{-(n+2)(t_\ell^{(2)} - t)} \cdot f_{t_\ell^{(2)}}^{(f)}(m_\ell^{(1)}) \cdot \left(e^{-\rho(t_{\ell-1}^{(2)} - t)} \delta_{m_{\ell-1}^{(2)}, m_\ell^{(2)}}\right) \\
&\qquad + f_t^{(1|2)}(s_\ell|m_\ell^{(2)}) \cdot \int_t^{t_{\ell-1}^{(2)} \wedge s_\ell} \rho e^{-\rho(t_\mathrm{r} - t)} \cdot f_{t_\mathrm{r}}^{(f)}(m_\ell^{(2)}) dt_\mathrm{r} \\
&\quad = \left[1 - \delta_{t_\ell^{(\mathrm{c})},\emptyset}\right] \cdot 2e^{-(\rho+n+2)(t_\ell^{(\mathrm{c})} - t)} \cdot f_{b,t_\ell^{(\mathrm{c})}}^{(a)}(m_\ell|m_{\ell-1}^{(2)}) \\
&\qquad + \left[\mathbb{1}_{(t_\ell^{(1)} < t_\ell^{(2)})}\right] \cdot e^{-(\rho+n+2)(t_\ell^{(1)} - t)} n_{h_\ell^{(1)}} \cdot f_{b,t_\ell^{(1)}}^{(a)}(m_\ell^{(2)}|m_{\ell-1}^{(2)}) \\
&\qquad + \left[\mathbb{1}_{(t_\ell^{(1)} > t_\ell^{(2)})}\right] \cdot e^{-(\rho+n+2)(t_\ell^{(2)} - t)} \delta_{m_{\ell-1}^{(2)}, m_\ell^{(2)}} \cdot f_{t_\ell^{(2)}}^{(f)}(m_\ell^{(1)}) \\
&\qquad + \int_t^{t_{\ell-1}^{(2)} \wedge s_\ell} \rho e^{-(\rho+n+2)(t_\mathrm{r} - t)} \cdot f_{t_\mathrm{r}}^{(f,f)}(s_\ell) dt_\mathrm{r} \\
&\quad = f_{b,t}^{(f,a)}(s_\ell|m_{\ell-1}^{(2)}),
\end{aligned}$$

where the penultimate equality is obtained by applying Lemmas B.4 and B.5 to the final term, and the final equality by (2.95) □

We establish the final key lemma before proving the main proposition.

Lemma B.7. *Let $s_{\ell-1}, s_\ell \in \mathcal{S}$. Then for $t, t' < s_{\ell-1}, s_\ell$,*

$$f^{(f,a)}_{b,t'}(s_\ell | m^{(2)}_{\ell-1}) \cdot f^{(f,f)}_t(s_{\ell-1}) = f^{(f,a)}_{b,t'}(s_{\ell-1} | m^{(2)}_\ell) \cdot f^{(f,f)}_t(s_\ell).$$

By symmetry, we may also conclude that

$$f^{(a,f)}_{b,t'}(s_\ell | m^{(2)}_{\ell-1}) \cdot f^{(f,f)}_t(s_{\ell-1}) = f^{(a,f)}_{b,t'}(s_{\ell-1} | m^{(2)}_\ell) \cdot f^{(f,f)}_t(s_\ell).$$

Proof. Using Lemmas B.5 and B.6, to expand $f^{(f,f)}_t(s_{\ell-1})$ and $f^{(f,a)}_{b,t'}(s_\ell | m^{(2)}_{\ell-1})$, respectively,

$$\begin{aligned}
&f^{(f,a)}_{b,t'}(s_\ell | m^{(2)}_{\ell-1}) \cdot f^{(f,f)}_t(s_{\ell-1}) \\
&\quad = \left[f^{(1|2)}_{t'}(s_\ell | m^{(2)}_\ell) \cdot f^{(a)}_{b,t'}(m^{(2)}_\ell | m^{(2)}_{\ell-1}) \right] \cdot \left[f^{(1|2)}_t(s_{\ell-1} | m^{(2)}_{\ell-1}) \cdot f^{(f)}_t(m^{(2)}_{\ell-1}) \right] \\
&\quad = \left[f^{(1|2)}_{t'}(s_\ell | m^{(2)}_\ell) \cdot f^{(1|2)}_t(s_{\ell-1} | m^{(2)}_{\ell-1}) \right] \cdot \left[f^{(a)}_{b,t'}(m^{(2)}_\ell | m^{(2)}_{\ell-1}) \cdot f^{(f)}_t(m^{(2)}_{\ell-1}) \right] \\
&\quad = \left[f^{(1|2)}_{t'}(s_{\ell-1} | m^{(2)}_{\ell-1}) \cdot f^{(1|2)}_t(s_\ell | m^{(2)}_\ell) \right] \cdot \left[f^{(a)}_{b,t'}(m^{(2)}_{\ell-1} | m^{(2)}_\ell) \cdot f^{(f)}_t(m^{(2)}_\ell) \right] \\
&\quad = \left[f^{(1|2)}_{t'}(s_{\ell-1} | m^{(2)}_{\ell-1}) \cdot f^{(a)}_{b,t'}(m^{(2)}_{\ell-1} | m^{(2)}_\ell) \right] \cdot \left[f^{(1|2)}_t(s_\ell | m^{(2)}_\ell) \cdot f^{(f)}_t(m^{(2)}_\ell) \right] \\
&\quad = f^{(f,a)}_{b,t'}(s_{\ell-1} | m^{(2)}_\ell) \cdot f^{(f,f)}_t(s_\ell),
\end{aligned}$$

where the second equality is obtained by rearranging factors, the third equality by applying Lemma B.4 and Proposition B.3, the fourth equality by rearranging factors, and the final equality by Lemma B.5 and Lemma B.6. □

Proposition B.8. *Let $s_\ell, s_{\ell-1} \in \mathcal{S}$. Then for $t, t' < s_{\ell-1}, s_\ell$, the following detailed balance condition holds for the densities $f^{(a,a)}_{b,t}$ and $f^{(f,f)}_{t'}$,*

$$f^{(a,a)}_{b,t}(s_\ell | s_{\ell-1}) \cdot f^{(f,f)}_{t'}(s_{\ell-1}) = f^{(a,a)}_{b,t}(s_{\ell-1} | s_\ell) \cdot f^{(f,f)}_{t'}(s_\ell).$$

This implies that $f^{(f,f)}_0$ is a stationary distribution for the Markov chain governed by transition density $f^{(a,a)}_{b,0}$.

Proof. Expanding factors $f^{(a,a)}_{b,t}(s_\ell|s_{\ell-1})$ and $f^{(f,f)}_{t'}(s_{\ell-1})$ using (2.96) and (2.94), respectively,

$$
\begin{aligned}
&f^{(a,a)}_{b,t}(s_\ell|s_{\ell-1})\cdot f^{(f,f)}_{t'}(s_{\ell-1})\\
&\quad=\left[\left[1-\delta_{t^{(\mathrm{c})}_{\ell-1},\emptyset}\right]e^{-2\rho(t^{(\mathrm{c})}_{\ell-1}-t)}\delta_{t^{(\mathrm{c})}_{\ell},t^{(\mathrm{c})}_{\ell-1}}\cdot f^{(a)}_{b,t^{(\mathrm{c})}_{\ell}}(m_\ell|m_{\ell-1})\right]\left[e^{-(n+1)\cdot(t^{(\mathrm{c})}_{\ell-1}-t')}\cdot f^{(f)}_{t^{(\mathrm{c})}_{\ell-1}}(m_{\ell-1})\right]\\
&\qquad+\left[\left[\mathbb{1}_{(t^{(1)}_{\ell-1}<t^{(2)}_{\ell-1})}\right]e^{-2\rho(t^{(1)}_{\ell-1}-t)}\delta_{m^{(1)}_{\ell},m^{(1)}_{\ell-1}}\cdot f^{(a)}_{b,t^{(1)}_{\ell}}(m^{(2)}_{\ell}|m^{(2)}_{\ell-1})\right]\left[n_{h^{(1)}_{\ell-1}}e^{-(n+1)\cdot(t^{(1)}_{\ell-1}-t')}\cdot f^{(f)}_{t^{(1)}_{\ell-1}}(m^{(2)}_{\ell-1})\right]\\
&\qquad+\left[\left[\mathbb{1}_{(t^{(1)}_{\ell-1}>t^{(2)}_{\ell-1})}\right]e^{-2\rho(t^{(2)}_{\ell-1}-t)}\delta_{m^{(2)}_{\ell},m^{(2)}_{\ell-1}}\cdot f^{(a)}_{b,t^{(2)}_{\ell}}(m^{(1)}_{\ell}|m^{(1)}_{\ell-1})\right]\left[n_{h^{(2)}_{\ell-1}}e^{-(n+1)\cdot(t^{(2)}_{\ell-1}-t')}\cdot f^{(f)}_{t^{(2)}_{\ell-1}}(m^{(1)}_{\ell-1})\right]\\
&\qquad+\int_t^{s_{\ell-1}\wedge s_\ell}\rho e^{-2\rho(t_{\mathrm{r}}-t)}\left[f^{(f,a)}_{b,t_{\mathrm{r}}}(s_\ell|m^{(2)}_{\ell-1})+f^{(a,f)}_{b,t_{\mathrm{r}}}(s_\ell|m^{(1)}_{\ell-1})\right]dt_{\mathrm{r}}\cdot f^{(f,f)}_{t'}(s_{\ell-1})\\
&\quad=\left[\left[1-\delta_{t^{(\mathrm{c})}_{\ell},\emptyset}\right]\cdot e^{-2\rho(t^{(\mathrm{c})}_{\ell}-t)}\delta_{t^{(\mathrm{c})}_{\ell-1},t^{(\mathrm{c})}_{\ell}}\cdot f^{(a)}_{b,t^{(\mathrm{c})}_{\ell-1}}(m_{\ell-1}|m_\ell)\right]\left[e^{-(n+1)\cdot(t^{(\mathrm{c})}_{\ell}-t')}\cdot f^{(f)}_{t^{(\mathrm{c})}_{\ell}}(m_\ell)\right]\\
&\qquad+\left[\left[\mathbb{1}_{(t^{(1)}_{\ell}<t^{(2)}_{\ell})}\right]\cdot e^{-2\rho(t^{(1)}_{\ell}-t)}\delta_{m^{(1)}_{\ell-1},m^{(1)}_{\ell}}\cdot f^{(a)}_{b,t^{(1)}_{\ell-1}}(m^{(2)}_{\ell-1}|m^{(2)}_{\ell})\right]\left[n_{h^{(1)}_{\ell}}e^{-(n+1)\cdot(t^{(1)}_{\ell}-t')}\cdot f^{(f)}_{t^{(1)}_{\ell}}(m^{(2)}_{\ell})\right]\\
&\qquad+\left[\left[\mathbb{1}_{(t^{(1)}_{\ell}>t^{(2)}_{\ell})}\right]\cdot e^{-2\rho(t^{(2)}_{\ell}-t)}\delta_{m^{(2)}_{\ell-1},m^{(2)}_{\ell}}\cdot f^{(a)}_{b,t^{(2)}_{\ell-1}}(m^{(1)}_{\ell-1}|m^{(1)}_{\ell})\right]\left[n_{h^{(2)}_{\ell}}e^{-(n+1)\cdot(t^{(2)}_{\ell}-t')}\cdot f^{(f)}_{t^{(2)}_{\ell}}(m^{(1)}_{\ell})\right]\\
&\qquad+\int_t^{s_\ell\wedge s_{\ell-1}}\rho e^{-2\rho(t_{\mathrm{r}}-t)}\left[f^{(f,a)}_{b,t_{\mathrm{r}}}(s_{\ell-1}|m^{(2)}_{\ell})+f^{(a,f)}_{b,t_{\mathrm{r}}}(s_{\ell-1}|m^{(1)}_{\ell})\right]dt_{\mathrm{r}}\cdot f^{(f,f)}_{t'}(s_\ell)\\
&\quad=f^{(a,a)}_{b,t}(s_{\ell-1}|s_\ell)\cdot f^{(f,f)}_{t'}(s_\ell),
\end{aligned}
$$

where the second equality is obtained by applying Proposition B.3 and Lemma B.7, and the final equality by (2.96) and (2.94). □

Appendix C

Analytic Forms

C.1 Single-deme, single-haplotype

Given a discretization $\mathcal{P}$ of $\mathbb{R}_{\geq 0}$, the discretized marginal, transition, and emission densities are defined by (3.13), (3.14), and (3.15), respectively. Critically, these densities can be written in terms of the quantities $x(p)$, $y_b(p)$, $z_b(p'|p)$ and $v_\ell^{(k)}(p)$, where $p, p' \in \mathcal{P}$. We now provide analytic expressions for each of these quantities, derived by evaluating the requisite integrals. Suppose that $p = [\tau_{i-1}, \tau_i)$ and $p' = [\tau_{j-1}, \tau_j)$. Then

$$x(p) = e^{-\tau_{i-1}} - e^{-\tau_i}, \tag{C.1}$$

and

$$y_b(p) = \frac{1}{x(p)} \frac{n}{\rho_b + n} (e^{-\frac{\rho_b+n}{n}\tau_{i-1}} - e^{-\frac{\rho_b+n}{n}\tau_i}). \tag{C.2}$$

For $\rho_b \neq n$,

$$z_b(p'|p) = \frac{1}{x(p)} \frac{\rho_b}{\rho_b - n} \cdot \begin{cases} x(p)\big(x(p') - \frac{n}{\rho_b}(e^{-\frac{\rho_b}{n}\tau_{j-1}} - e^{-\frac{\rho_b}{n}\tau_j})\big), & \text{if } j < i, \\ x(p')\big(x(p) - \frac{n}{\rho_b}(e^{-\frac{\rho_b}{n}\tau_{i-1}} - e^{-\frac{\rho_b}{n}\tau_i})\big), & \text{if } j > i, \\ x(p)\big(x(p) - \frac{n}{\rho_b}(e^{-\frac{\rho_b}{n}\tau_{i-1}} - e^{-\frac{\rho_b}{n}\tau_i})\big) & \\ \quad - \frac{\rho_b - n}{\rho_b}\frac{n}{\rho_b+n}(e^{-\frac{\rho_b+n}{n}\tau_{i-1}} - e^{-\frac{\rho_b+n}{n}\tau_i}) & \\ \quad - \frac{n}{\rho_b}(e^{-\tau_{i-1}}e^{-\frac{\rho_b}{n}\tau_i} - e^{-\tau_i}e^{-\frac{\rho_b}{n}\tau_{i-1}}), & \text{if } j = i, \end{cases} \tag{C.3}$$

and for $\rho_b = n$,

$$z_b(p'|p) = \frac{1}{x(p)} \cdot \begin{cases} x(p)\big(x(p') + (\tau_{j-1}e^{-\tau_{j-1}} - \tau_j e^{-\tau_j})\big), & \text{if } j < i, \\ x(p')\big(x(p) + (\tau_{i-1}e^{-\tau_{i-1}} - \tau_i e^{-\tau_i})\big), & \text{if } j > i, \\ x(p)\big(x(p) + (\tau_{i-1}e^{-\tau_{i-1}} - \tau_i e^{-\tau_i})\big) & \\ \quad -(\tau_{i-1} - \tau_i)e^{-(\tau_{i-1}+\tau_i)} - \frac{1}{2}(e^{-2\tau_{i-1}} - e^{-2\tau_i}), & \text{if } j = i. \end{cases} \tag{C.4}$$

Finally,

$$\begin{aligned} v_\ell^{(k)}(p) &= \frac{1}{x(p)} \sum_{j=0}^{k} \Big(\frac{n}{\theta_\ell + n}\Big)^{j+1} \frac{k!}{(k-j)!} [e^{-\frac{\theta_\ell+n}{n}\tau_{i-1}} \tau_{i-1}^{k-j} - e^{-\frac{\theta_\ell+n}{n}\tau_i} \tau_i^{k-j}] \\ &= \frac{n}{\theta_\ell + n} \left(v_\ell^{(k-1)}(p) \cdot k + \frac{e^{-\frac{\theta_\ell+n}{n}\tau_{i-1}} \tau_{i-1}^k - e^{-\frac{\theta_\ell+n}{n}\tau_i} \tau_i^k}{x(p)} \right). \end{aligned} \tag{C.5}$$

Note that recursive structure of $v_\ell^{(k)}(\cdot)$, in conjunction with the infinite sum in (3.15), suggest an efficient method for approximating the emission density by using partial sums.

C.2 Multiple-deme, single-haplotype

The matrix exponentials associated within the initial (2.88) and transition (2.90) densities associated with the CSD $\hat{\pi}_{\text{SMC-ADO}}$ can be approximated to arbitrary precision by using partial sums in the definition of the matrix exponential. However, to obtain the desired explicit analytic forms for the discretized marginal, transition, and emission densities, defined in (3.26), (3.28), and (3.29), respectively, we propose a different approach.

Suppose that the matrix Z is diagonalizable (which is true if and only if Υ is diagonalizable), then there exists a matrix $V = (v_1, \ldots, v_{2q})$, the columns of which are the eigenvectors v_i of Z, and a diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{2q})$, where λ_i are the eigenvalues of Z, such that $Z = V\Lambda V^{-1}$. Using this eigen-decomposition, the matrix exponential $(e^{tZ})_{i,j}$ can be expressed as $\sum_{k=1}^{2q} e^{t\lambda_k}(v_k w_k)_{i,j}$, where w_i is the i-th row of the matrix V^{-1}. Note that for a non-diagonalizable matrix, a similar eigen-decomposition can be obtained using generalized eigenvectors and the Jordan normal form, and similar, though more involved, explicit computations can be performed.

Recall that, as in the single-deme case, the discretized marginal and transition densities can be written in terms of the quantities $x(p,d)$, $y_b(p,d)$, and $z_b(p',d'|p,d)$ where $p, p' \in \mathcal{P}$ and $d, d' \in \mathcal{D}$. We now provide analytic expressions for each of these quantities, derived by using the spectral representation and evaluating the requisite integrals. Suppose that $p = [\tau_{i-1}, \tau_i)$ and $p' = [\tau_{j-1}, \tau_j)$. For convenience, define

$$\mathrm{I}_a^b(\lambda) = \int_{t=a}^{b} e^{\lambda t} dt = \begin{cases} \frac{1}{\lambda}(e^{\lambda b} - e^{\lambda a}), & \text{if } \lambda \neq 0, \\ b - a, & \text{if } \lambda = 0. \end{cases} \tag{C.6}$$

Then the quantities of interest can be expressed

$$x(p,d) = \sum_{k=1}^{2q} (v_k w_k)_{r_\alpha, a_d} \lambda_k \mathrm{I}_{\tau_{i-1}}^{\tau_i}(\lambda_k), \tag{C.7}$$

and

$$y_b(p,d) = \frac{1}{x(p,d)} \sum_{k=1}^{2q} (v_k w_k)_{r_\alpha, a_d} \lambda_k \mathrm{I}_{\tau_{i-1}}^{\tau_i}(\lambda_k - \rho_b), \tag{C.8}$$

and

$$\begin{aligned} z_b(p',d'|p,d) = \frac{\rho_b}{x(p,d)} \sum_{d_\mathrm{r} \in \mathcal{D}} \sum_{k=1}^{2q} \sum_{m=1}^{2q} \sum_{n=1}^{2q} & (v_k w_k)_{r_\alpha, r_{d_\mathrm{r}}} (v_m w_m)_{r_{d_\mathrm{r}}, a_d} (v_n w_n)_{r_{d_\mathrm{r}}, a_{d'}} \\ & \times \Big[e^{\lambda_m \tau_i} e^{\lambda_n \tau_j} \mathrm{I}_0^{\tau_i \wedge \tau_j}(\lambda_k - \lambda_m - \lambda_n - \rho) \\ & \quad - e^{\lambda_m \tau_i} e^{\lambda_n \tau_{j-1}} \mathrm{I}_0^{\tau_i \wedge \tau_{j-1}}(\lambda_k - \lambda_m - \lambda_n - \rho) \\ & \quad - e^{\lambda_m \tau_{i-1}} e^{\lambda_n \tau_j} \mathrm{I}_0^{\tau_{i-1} \wedge \tau_j}(\lambda_k - \lambda_m - \lambda_n - \rho) \\ & \quad + e^{\lambda_m \tau_{i-1}} e^{\lambda_n \tau_{j-1}} \mathrm{I}_0^{\tau_{i-1} \wedge \tau_{j-1}}(\lambda_k - \lambda_m - \lambda_n - \rho) \Big]. \end{aligned} \tag{C.9}$$

Finally, from (3.29) one can show that, letting $p = [\tau_{i-1}, \tau_i) \in \mathcal{P}$, $h \in \mathcal{H}$, and $d \in \mathcal{D}$,

$$\xi_\ell(\eta[\ell]|p,h,d) = \frac{1}{x(p,d)} \sum_{a \in A_\ell} \sum_{k=1}^{2q} (x_j y_j)_{h[\ell],\eta[\ell]} (v_k w_k)_{r_\alpha, a_d} \lambda_k \mathrm{I}_{\tau_{i-1}}^{\tau_i}(\lambda_k + \theta_\ell \omega_j - \theta_\ell) \tag{C.10}$$

where we have used the eigen-decomposition $\Phi^{(\ell)} = X\Omega X^{-1}$ of the mutation matrix. Here, $\Omega = \mathrm{diag}(\omega_1, \ldots, \omega_{|A_\ell|})$ is the diagonal matrix of eigenvalues, $X = (x_1, \ldots, x_{|A_\ell|})$ is the matrix which has the eigenvectors of the mutation matrix as columns, and y_j denotes the j-th row of X^{-1}.