UNIVERSITY OF COPENHAGEN
DEPARTMENT OF MATHEMATICAL SCIENCES

Master Thesis

Simon Christoffer Ziersen

Inference In Population Genetics

Supervisor: Carsten Wiuf

Submitted on: 6th of August 2018

Abstract

The field of mathematical population genetics has been around since the early part of the 20'th century, but it was not until the early 1990's that computational statistical methods were available for likelihood inference on population genetics data. This paper derives the well-known infinite-sites-model describing DNA-sequence data along with sample probabilities as recursion equations on the states of the model. Two importance sampling schemes, derived by Grffiths-Tavaré and Stephens-Donnelly respectively, are then presented along with a new importance sampler. In the work by Stephens and Donnelly they show that their sampler outperforms the one given by Griffiths and Tavaré in efficiency, and the comparison with the new sampler is thus made with the method presented by Stephens and Donnelly. The new sampler takes into account additional information carried in the DNA-sequence data than does the method of Stephens and Donnelly, thus resulting in efficiency gain.

Contents

1 Introduction **4**

2 Kingmans' Coalescent **7**

2.1 The Wright-Fisher Population 7

2.2 The Ancestral Process 8

2.3 The n-coalescent 11

2.4 Mutations 13

3 Infinite-Sites-Model **17**

3.1 Number of Segregating Sites 18

3.2 Genealogical Trees 23

3.3 Rooted Tree Probabilities 30

3.4 Unrooted Tree Probabilities 35

4 Estimation of θ **39**

4.1 The Griffiths-Tavaré scheme 40

4.1.1 Point estimates 40

4.1.2 Likelihood Surfaces 43

4.2 Importance Sampling 45

4.2.1 Construction of Q_θ 48

4.2.2 New Proposal Distribution 51

5 Performance **54**

6 Conclusion **57**

References **59**

1 Introduction

Statistical analysis of population genetics dates back to the early 20'th century, with Ronald Fisher and Sewall Wright posing as the driving forces in the debate on the evolution of species, which led to the early description of a stochastic model describing the evolution in a population. Since then other models have been presented, and for a larger part of the 20'th century, the analysis of population genetics was mostly theoretical, drawing on difussion limits in stochastic processes. For a review of the historical advances in mathematical population genetics see [Ewens, 2012]. With the advancements in biological technology and computers, computational statistical methods were developed in the later part of the last century, where the methods for likelihood inference on genetics data was introduced in the early 1990's, and since then improved, both in efficiency and versatility. These methods will serve as the main focus in this paper, but as the number of models describing the genetic evolution is large, with varying assumptions taken in each model, one has to choose a specific model before attempting to construct any computational likelihood scheme for analysing genetic data.

When trying to describe the genealogy of a sample of a population, a strong tool was developed by [Kingman, 1982a], where the theory of the coalescent process was first developed. The process describes the ancestral relationship of a sample through its evolutionary history back to the time of most recent common ancestor (MRCA) of the sample. The process can be derived from populations following different evolutionary models, as long as they agree on a difussion limit. In this paper, the populations are assumed to evolve according to a Wright-Fisher model, assuming a haploid population (individuals in the population carry only a single set of chromosomes, i.e., every offspring has only one parent) and disregarding the effects of selection.

To model data given by DNA samples, the infinite-sites-model, formulated in [Ethier and Griffiths, 1987], is the model chosen to be the main focus in this article. The model is shown to follow the Kingman's coalescent process, where, when going back in time, evolutionary events consist of either a mutation to a gene in the sample or the coalescence of two genes. The model thus consists of a coalescent process, under

which coalescent events are exponentially distributed with a parameter depending on the number of individuals in the sample at a given time, and a mutation process described by a Poisson process depending on a mutation parameter θ.

The aim of this paper is to construct an estimation scheme allowing for full likelihood inference of θ in the infinite-sites-model. The formulation given in [Ethier and Griffiths, 1987] describes the states of the model as genealogical trees, from which sample probabilities can be calculated. The probabilities are given by a recursion equation on the states of model, but solving the equation for a realistic sample size prove to be infeasible, as the number of trees to consider becomes enormous. The first method to solve this problem was provided by Griffiths and Tavaré in 1994 and the application of their method to the infinite-sites-model is given in [Griffiths and Tavaré, 1995]. Their method utilizes the recursion formulation of the sample probabilities to construct a Markov chain starting in the sample and moving back in time until a common ancestor is reached. A simple lemma regarding Markov Chains then states that the sample probability can be calculated as the expectation of a functional on the Markov chain.

[Stephens and Donnelly, 2000] realised that the method provided by Griffiths and Tavaré was a case of importance sampling, which motivated their method for likelihood inference. By analysis of the theoretical optimal proposal distribution, they were able to construct an importance sampler which outperforms the Griffiths-Tavaré in efficiency. The method presented by [Stephens and Donnelly, 2000] was developed under the infinite-alleles-model, but the derivations used to construct the proposal distribution can not be generalised to the infinite-sites-model, because the state space becomes uncountably infinite. In turn, they develop an importance sampler for the infinite-sites-model, by suggesting a Markov chain as a proposal distribution, which samples evolutionary events by choosing a sequence uniformly at random to be involved in the most recent event.

The importance sampler proposed by [Stephens and Donnelly, 2000] only regards the information about the genetic types and their multiplicities. Besides the genetic types recorded in the data, the infinite-sites-model also contains information about the number of mutations carried by each allele and the position of each mutation. This observation motivates the development of a new importance sampler making use of the mutations carried by the alleles. In this paper we propose a sampler

analogous to the one given by [Stephens and Donnelly, 2000], by weighting the transitions probabilities in a Markov chain by the number of mutations carried by each allele. This amounts to a proposal distribution more efficient than that presented in [Stephens and Donnelly, 2000].

The structure of the paper is as follows: In section 2 we derive the coalescent process from a Wright-Fisher population.

Section 3 presents the infinite-sites-model, first analysing a summary statistic given by the number of segregating sites. The statistic leads to Wattersons estimator, and though unbiased, it it generally not sufficient. The infinite-sites-model is then described using the formulation in [Ethier and Griffiths, 1987], and the sample probabilities are then calculated as recursion equations for the probabilities of rooted and unrooted trees respectively.

Section 4 derives the Griffiths-Tavaré method for the infinite-sites-model, by first presenting the general idea, and then extending it to cover approximations of likelihood surfaces. After the derivation of the GT-scheme, the general ideas of importance sampling in the model are discussed which leads to the derivation of the SD-sampler. In the wake of the SD-sampler the new proposal distribution is presented together with a result on general Poisson processes, motivating the choice of the new sampler.

Section 5 presents a comparison of the SD and new proposal distributions on an example data set given in [Griffiths and Tavaré, 1995], where the true sample probability is given for different values of θ. While both proposals produce fairly accurate estimates of the sample probability, the new proposal is shown to be more efficient by lowering the standard errors to roughly half of those produced by the SD-sampler.

Section 6 closes the paper with a conclusion.

2 Kingmans' Coalescent

2.1 The Wright-Fisher Population

We begin by describing a simple model of population genetics, the Wright-Fisher model, which describes the evolution of a two-allele locus in a population of fixed size N. We start with the simplest case of the model, where we describe the model without either mutation of selection effects.

Assume that we have a population of size N and n, $n = 0, 1, ...$, non-overlapping generations, where each allele is of either type A or B and let X_i be the number of A-alleles at generation i. The model describes the distribution at generation $r+1$ by conditioning on the number A-alleles at generation r. The conditional distribution of X_{r+1} given X_r is then assumed binomial, such that given $X_r = i$ the distribution at generation $r+1$ is given by

$$p_{ij} = P(X_{r+1} = j | X_r = i) = \binom{N}{j} \pi_i^j (1 - \pi_i)^{N-j}, \tag{2.1.1}$$

where $\pi_i = i/N$ is the frequency of A-alleles. From this formulation, it is clear that the $(X_r)_{r=0,1,2,...}$ forms a Markov chain with transition matrix $(p_{ij})_{i,j=0,...,N}$, but the even though many things could be explored from this model, the characteristics of this Markov chain is not our main interest. The Wright-Fisher model is a forward-looking process that takes an initial distribution of alleles and models the evolution forward in time with the basic, but important assumption, that all alleles are equally likely to produce offspring. The forward looking structure of the model makes it hard to use from an inference view point, and hence the lack of interest in this specific model. Rather, we will use the basic assumptions of a population evolving under the Wright-Fisher model when developing models from which inference of model parameters will be possible.

One specific distinction between the Wright-Fisher model and the models to come, is the forward looking structure, which is unappealing when trying to infer model parameters, since it lends it self poorly to sample data. Because of this we seek to express the genealogy in a backwards-looking manner, in order to describe a model that is suitable for sample data. In figure 1 we see an example of a population developed under the Wright-Fisher model, where $N = 6$ and generation $r = 0, ..., 5$.

The crosses represent A-alleles and clear circles corresponds to B-alleles. Here we are able to examine the evolution of the population, but in reality we would only be able to observe the bottom row, $r = 5$, and hence we want to describe the evolution starting from a sample going backwards in time.

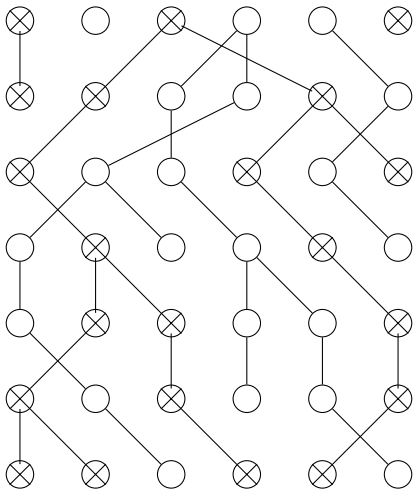

Figure 1: Example of the evolution in a Wright-Fisher population of size $N = 6$, where circles with crosses represent A-alleles and clear circles represent B-alleles. The top row represents generation 0 with the initial configuration $X_0 = 3$, and at generation 6, $X_6 = 4$.

2.2 The Ancestral Process

When trying to understand the genealogy of the Wright-Fisher model, the ancestral relationship of the individuals in the population plays an important role. Looking at figure 1 starting from the bottom row at generation $r = 5$, we see that individual 1 and 2, starting from the left, is connected in generation $r = 4$, since they share the same parent, and hence their most recent common ancestor, MRCA, is 1 generation ago. Likewise, individual 1, 2 and 3 share a common ancestor at generation 2, while individual 4 and 5 have their MRCA at generation 0. To capture this nature in our formulation it is convenient to change notation, so that the present is regarded as generation 0, and $r = 0, ..., s$ represents generation from the present and s generations into the past, and rather than describing the genealogy for the whole population we choose to look at a sample of size, say, k. In regards to figure 1 this corresponds to viewing the bottom row as generation 0 and the top row as generation 5, and letting the figure describe a sample of size k in a population of size N.

We now seek to describe the number of ancestors to the sample going backwards in time, and a key observation, which follows directly from the Wright-Fisher assumption, is that the individuals choose their parents independently at random from each generation to the next and independently of previous generations. Furthermore, we want describe the evolution in a large population as suppose to the fixed population size N and time measured in units of N generations, to describe the process in a continuous time frame.

Formally, Suppose we have a population of size N and take a sample of size k. Then

$$g_{kj} = \mathbb{P}\left(k \text{ individuals have } j \text{ distinct parents}\right) = N(N-1)\cdots(N-j+1)\mathcal{S}_k^{(j)}N^{-k},$$

where $N(N-1)\cdots(N-j+1)$ is the number of ways of assigning j distinct parents out of the population, $\mathcal{S}_k^{(j)}$ is a Stirling's number of second kind, which is the number of ways of partitioning a set of k elements into j non-empty sets, i.e.. the number of ways for k individuals to choose j different parents, and N^{-k} is the number of ways for k individuals to choose their parents. Define an ancestral process, $\{A_n^N(t),\ t = 0, 1, ...\}$, as the number of ancestors in a sample of size n in a population of size N at generation t. In figure 1 this gives $A_6^N(1) = 5$ and $A_6^N(4) = 3$. It is clear that $A_n^N(t)$ is a Markov chain with transition probabilities given by

$$\mathbb{P}\left(A_n^N(t+1) = j \mid A_n^N(t) = k\right) = g_{kj}.$$

For fixed n and $N \to \infty$ we see that

$$g_{k,k-1} = N(N-1)\cdots(N-k+2)\mathcal{S}_k^{(k-1)}N^{-k} = \binom{k}{2}N^{-1} + O(N^{-2}), \tag{2.2.1}$$

since $\mathcal{S}_k^{(k-1)} = \binom{k}{2}$. Furthermore

$$g_{kj} = O(N^{-2}), \quad \text{for } j < k-1, \tag{2.2.2}$$

and

$$g_{kk} = 1 - \binom{k}{2}N^{-1} + O(N^{-2}). \tag{2.2.3}$$

The last statement is harder to realize, so we will make a small proof.

proof of (2.2.3). The proof follows a simple induction: We show that the statement is true for g_{nn} with $n = 2, ..., k$.

Assume $n = 2$:

$$g_{nn} = g_{2,2} = \frac{N(N-1)}{N^2} = 1 - \frac{1}{N} + \frac{0}{N^2} = 1 - \binom{2}{2}N^{-1} + O(N^{-2}).$$

Now, assume that (2.2.3) is true for $n = 2, ..., k-1$, that is

$$g_{k-1,k-1} = \frac{N(N-1)\cdots(N-k+2)}{N^{k-1}} = 1 - \binom{k-1}{2}N^{-1} + O(N^{-2}).$$

Then for $n = k$ we have:

$$\begin{aligned}
g_{kk} &= N(N-1)\cdots(N-k+1)S_k^{(k)}N^{-k} \\
&= \frac{N(N-1)\cdots(N-k+2)(N-k+1)}{N^{k-1}N} \\
&= \left(1 - \binom{k-1}{2}N^{-1} + O(N^{-2})\right)\frac{N-k+1}{N} \\
&= \frac{N-k+1}{N} - \binom{k-1}{2}\frac{N-k+1}{N^2} + \frac{N-k+1}{N}O(N^{-2}) \\
&= 1 - \frac{k-1}{N} - \binom{k-1}{2}\frac{1}{N} + \binom{k-1}{2}\frac{k-1}{N^2} + O(N^{-2}) \\
&= 1 - \left((k-1) + \binom{k-1}{2}\right)N^{-1} + O(N^{-2}) \\
&= 1 - \left(\binom{k-1}{1} + \binom{k-1}{2}\right)N^{-1} + O(N^{-2}) \\
&= 1 - \binom{k}{2}N^{-1} + O(N^{-2}).
\end{aligned}$$

In the third equality we use the induction step, and in the last equality we use the recursive formula for binomial coefficients. □

Now, writing G_N for the transition matrix of $A_n^N(t)$ we see that

$$G_N = I + N^{-1}Q + O(N^{-2}),$$

where Q is a lower diagonal matrix with

$$q_{k,k-1} = \binom{k}{2}, \quad q_{kk} = -\binom{k}{2}.$$

If we let time be measured in units of N generations, such that $r = \lfloor tN \rfloor$, we have

$$G_N^{tN} = \left(I + N^{-1}Q + O(N^{-2})\right)^{\lfloor tN \rfloor} \to e^{Qt},$$

as $N \to \infty$. By elementary results on convergence of Markov chains combined with the Kolmogorov backward equations, this shows that the number of ancestors at generation tN is approximated by a Markov chain, $\{A_n(t),\ t \geq 0\}$, with infinitesimal generator Q, where the soujorn time in any state η, where $|\eta| = k$, is exponentially distributed with parameter $\binom{k}{2}$. The waiting time in a state of size k will be denoted T_k which is thus exponential with parameter $\binom{k}{2}$.

2.3 The n-coalescent

In the ancestral process described above, we track the genealogy of a sample by recording the number of individuals ancestral to the sample at a given time t. If we furthermore want to look into the more detailed history of which members of the sample that are descended from which ancestors at a given time, we turn to the $n - coalescent$ formulated by [Kingman, 1982b]. This is done by labelling the members of the sample with the set $\{1, 2, ..., n\}$ and partition them into equivalence classes according to which individuals share a common ancestor at a given time t. The individuals related to each other at a fixed time is denoted an equivalence relation by

$$i \sim j \iff \text{individuals } i \text{ and } j \text{ share a common ancestor at time } t.$$

Let $\mathcal{E}_n$ be the finite set of equivalence relations on $\{1, 2, ..., n\}$. For a fixed time t where there are k ancestors to the sample, such that $A_n(t) = k$, we partition the set $\{1, 2, ..., n\}$ into equivalence classes according to their equivalence relations. This gives the classes $E_j = \{i_{j1}, i_{j2}, ..., i_{jl_j}\}$ for $j = 1, ..., k$, since there are precisely k classes - one corresponding to each of the ancestors at time t, where $E_i \cap E_j = \varnothing,\ i \neq j$, and $\cup_{i=1}^{k} E_i = \{1, 2, ..., n\}$.

Now, define the set of equivalence relations on $\{1, 2, ..., n\}$ by $\mathcal{E}_n$, and denote the discrete time process $\mathcal{R}_s$ on $\mathcal{E}_n$, for $s = 0, 1, 2, ...$ generations, which describes the equivalence classes at a given time. Note that $\mathcal{R}_0 = \Delta = \{(i, i), i = 1, 2, ..., n\}$, which is the state where nobody is related to anyone else, and that $\mathcal{R}_s \subseteq \mathcal{R}_{s+1}$. From simple construction of $\mathcal{R}_s$ it is noted that $\mathcal{R}_s$ forms a Markov chain on the discrete set of equivalence relations $\mathcal{E}_n$ with transition probabilities

$$P(\mathcal{R}_{s+1} = \eta | \mathcal{R}_s = \xi) = p_{\xi\eta},$$

where ξ and η are equivalence classes. To calculate these we start by observing that $p_{\xi\eta} = 0$ if $\xi \not\subseteq \eta$. From here the arguments follow those used to calculate the characteristics of the ancestral process.

We divide the event $(\mathcal{R}_{s+1} = \eta | \mathcal{R}_s = \xi)$ into three scenarios. The first being that η is obtained from ξ by collapsing more than two of the equivalence classes, i.e. $|\xi| < |\eta| + 1$, where $|\xi|$ and $|\eta$ is the number of equivalence classes in each state. Due to (2.2.2) the probability of this event is of order $O(N^{-2})$, since it is the probability that the number of ancestors drop by more than 1, which is given by (2.2.2), times a combinatorial factor determining which of the individuals to coalesce, not depending on N.

The second scenario to consider is that $\eta = \xi$. This happens when the number of ancestors is unchanged and the probability is then given by (2.2.3).

The third scenario is that η is obtained from ξ by collapsing exactly two equivalence classes, which we denote $\xi \prec \eta$, that is $|\xi| = |\eta| + 1$. The probability of this event is the probability that the number of ancestors drop by exactly one times a combinatorial factor determining the number of ways of choosing two equivalence classes. The probability of obtaining $|\xi| - 1$ ancestors is given by (2.2.1) and the combinatorial factor is a binomial coefficient depending on the size of ξ. To summarize, let $|\xi| = k$. Then:

$$p_{\xi\eta} = \begin{cases} P(|\eta| = k-1 \mid |\xi| = k) \cdot \binom{k}{2}^{-1} & \xi \prec \eta \\ P(|\eta| = k-1 \mid |\xi| = k) & \xi = \eta,\ |\xi| = k \\ O(N^{-2}) & otherwise \end{cases}$$

$$= \begin{cases} \binom{k}{2} N^{-1} \binom{k}{2}^{-1} + O(N^{-2}) & \xi \prec \eta \\ 1 - \binom{k}{2} N^{-1} + O(N^{-2}) & \xi = \eta,\ |\xi| = k \\ O(N^{-2}) & otherwise \end{cases}$$

In the same way as with the ancestral process we let P_N be the transition matrix of the chain $\mathcal{R}_s$, thus

$$P_N = I + QN^{-1} + O(N^{-2}),$$

where Q, this time, is given as a lower diagonal matrix with

$$q_{k,k-1} = 1, \quad q_{kk} = -\binom{k}{2}.$$

If we let t be measured in units of N generations, $r = \lfloor tN \rfloor$, we see that

$$P_N^{\lfloor tN \rfloor} = \left(I + QN^{-1} + O(N^{-2})\right)^{\lfloor tN \rfloor} \to e^{Qt},$$

as $N \to \infty$. Thus the process $\mathcal{R}_{tN}$ converges in distribution to the continuous time Markov process, R_t, with infinitesimal generator Q, as $N \to \infty$. The process R_t is called the $n-coalescent$. This leads to the formal definition:

Definition 2.1. *The n-coalescent,* $\{R_t, t \geq 0\}$ *is a continuous-time Markov chain with state space* $\mathcal{E}_n$, $n \in \mathbb{N}$, *and transition rates* $\{q_{\xi\eta},\ \xi, \eta \in \mathcal{E}_n\}$ *given by*

$$q_{\xi\eta} = \begin{cases} 1 & \xi \prec \eta, \quad \xi \neq \eta \\ -\binom{k}{2} & \xi = \eta,\ |\xi| = k \\ 0 & otherwise \end{cases},$$

with

$$C_0 = \Delta = \{(i,i), i = 1, 2, ..., n\}$$

where Δ *is the identity relation where "nobody is related to anyone else".*

The coalescent process given in the definition will be the cornerstone in the rest of the paper. In the sections to come, we define mutational processes to superimpose on the coalescent along with a model describing the genealogy given a data set of DNA sequences. In every case the aim is to adapt the new formulations to the coalescent process to describe the genealogical structure of a sample. The simple formulation of definition 2.1 makes it easy to keep track of the ancestral relationship between genes in a sample when exploring more complex models for the genealogy, as long as they can be shown to agree with the coalescent process. Although the mention of the coalescent will be subtle in the models to come, it serves as a driving force ensuring a structure of the ancestral history that is easy to manage however we extend our models.

2.4 Mutations

So far we have seen how to track the genealogy in a sample from a large population derived under the Wright-Fisher model. In order to further enhance our grasp on the genealogy we introduce mutations to the model. In the Wright-Fisher model, suppose

that at each generation there is a probability for an allele to mutate to another type, that is from A to B or from B to A. Denote $\mu_A > 0$ the probability that an A-allele mutate to a B-allele and $\mu_B > 0$ the probability of a B-allele to mutate to an A-allele. In the transition probabilities defined in (2.1.1) this amounts to a change in the π_i's such that

$$\pi_i = \frac{i}{N}(1-\mu_A) + \left(1 - \frac{i}{N}\right)\mu_B.$$

Then, when the population is large we assume that the mutation probabilities satisfy

$$\lim_{N\to\infty} 2N\mu_A = \theta_A, \quad \lim_{N\to\infty} 2N\mu_B = \theta_B,$$

so that mutation rates are of order $O(N^{-1})$, and we define the total mutation rate

$$\theta = \theta_A + \theta_B.$$

This derivation covers mutations in the Wright-Fisher model, and though it can be used for a number of interesting things, for example to calculate the stationary distribution of A-alleles in a large sample, we will not explore the consequences of mutations in a two-allele model further. Rather, we will note that this formulation extend itself to allowing multiple types of alleles. It is easily extended to k types, when mutations occur at rate $\theta/2$, the derivation of which we will see below, and the probability of a mutation resulting in allele A_i is given by π_i, such that $\theta_i = \theta\pi_i,\ i = 1, ..., k$, following the notation above, but with $i = \{1, ...k\}$ instead of $i = \{A, B\}$. This model is called the $k - allele - model$ in the literature.

The question is now how to proceed when allowing for infinitely many types of alleles. To see this, we suppose that an individual has the probability μ of mutating in a single generation creating a never-before-seen type. A way of doing this is by labelling the types by uniform random variables, such that every time a mutation occurs, its type is drawn uniformly at random from $(0, 1)$.
We assume that the probability of a mutation in an individual is independent of earlier mutations in that individual and of mutation carried out by other members of the population. Furthermore assume that

$$\lim_{N\to\infty} 2\mu N = \theta,$$

such that θ is the scaled probability of mutation in a large population and the probability of a mutation is of order $O(N^{-1})$. To calculate the distribution of mutations in a sample when N is large, take a sample of n individuals at generation $r = 0$ in a backwards-looking model as in the ancestral process. Define the stopping time $\tau = \inf\{s \in \mathbb{N} \mid t_i(s) \neq t_i(0)\}$, where $t_i(s)$ denotes the type of individual i at generation s, as the last occurrence of a mutation in individual i. If we rescale time in units of N generations we see that

$$P(\tau \geq s) = (1-\mu)^s = \left(1 - \frac{\theta}{2N}\right)^{\lfloor tN \rfloor} \to e^{t\theta/2}, \qquad N \to \infty,$$

so the time to mutation along a single linage in the sample is exponentially distributed with parameter $\frac{\theta}{2}$, which shows that mutations in a single lineage arrive as points of a Poisson process with parameter $\frac{t\theta}{2}$, when N is large. Because mutations are independent between individuals this means that the waiting time, T_n^{mut}, until a mutation happens in one of the lineages of the sample is exponentially distributed with parameter $n\frac{\theta}{2}$. This is seen by letting $X_1, ..., X_n$ be n i.i.d. random variables with $X_i \sim Exp(\frac{\theta}{2})$. Then

$$P(T_n^{\text{mut}} \geq t) = P(\min\{X_1, ..., X_n\} \geq t) = \prod_{i=1}^{n} P(X_i \geq t) = e^{-t\sum_{i=1}^{n}\theta/2} = e^{-tn\theta/2}.$$

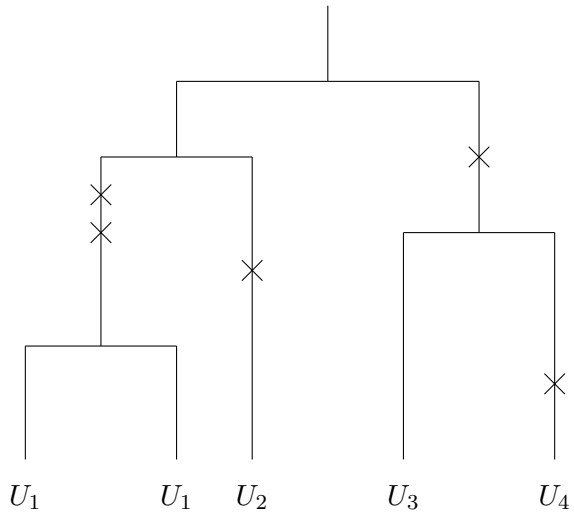

Figure 2: Example of the ancestral tree from a sample of size $n = 5$.

We are now able to formulate the genealogy of a sample derived from a Wright-Fisher population (or any neutral model that has the same large scale ancestral

relationship as the ancestral process) by the coalescent process on which we superimpose a Poisson process, with arrival times T_n^{mut}, describing the mutations along the coalescent tree. In figure 2 we see an example of such a tree. Here we have a sample of size $n = 5$ with 4 different types denoted (U_1, U_2, U_3, U_4). The crosses represent mutation, and at each mutation the type of the individual is drawn uniformly at random from $(0, 1)$, resulting in $(U_1, U_2, U_3, U_4) \in (0, 1)^4$. The collapsing of two lines into one represent coalescent events. So far we have only explored models recording the allelic types of genes. When data is given by DNA sequences the allelic types of genes is recorded, but more information can be extracted from the data. One key observation is the segregating sites, describing the nucleotide positions at which the alleles in a sample differ. The next section explores the infinite-sites-model, which strives to explain the genealogy using the information contained in sequence data. This model will be the focus from this point on, but the coalescent and mutation process still serve as the underlying mechanisms describing the genealogy.

3 Infinite-Sites-Model

We start this section with an example of a real DNA sequence data set. Table 1 is from [Griffiths and Tavaré, 1994a] and was first published in [Ward et al., 1991]. Briefly, the data represents sequences of nucleotides from 63 individuals where every sequence is of length 360, corresponding to positions $16,024 - 16,383$ in the human reference sequence [Griffiths and Tavaré, 1994a]. The first five columns consist of purine nucleotides (A, G), and the last 13 consist pyrimidine nucleotides (C, T), which gives the data a binary form since there are only two possible nucleotides at each site. The data is listed only with the distinct sequences and their multiplicities are found in the rightmost column. Furthermore, only the sites which are distinct between some sequences in the data are shown. These are called segregating sites and they form the base for analysis of DNA sequences through the infinite-sites-model first describe by [Watterson, 1975]. As the name suggest the model assumes that there are infinite sites in a given sequence, and thus, that mutations can only occur ones in any given sequence. This assumption leads to the attractive feature, that the number of segregating sites in a sample is equal to the number of mutations since the MRCA.

To determine the position of a mutation for a given sequence we need a reference point to conclude whether the nucleotide at a given site is a mutation or not, i.e., if it is ancestral to the sample or not. In a given sample such as table 1, label the sites in each sequence by 0 or 1, where 0 is given to the most frequent nucleotide at a given site and 1 is given to the rest. Assume that the nucleotides ancestral to the sample are the ones denoted 0. Then the mutations since the MRCA are the sites, for a given sequence, denoted 1.

To see how this formulation of mutational positions fits into the coalescent process described in the earlier section, consider again the example given in figure 2. The mutations are described by changing the binary states of a nucleotide a given site from 0 to 1. Thus for convenience, label the mutations (the crosses in the figure) from the top by site1, site2 etc. Then the sequence labelled U_4 experience a change in site 1 from 0 to 1, the sequences U_1 experience a change in site 2 and so on. Hence,

the sequence in the figure can be expressed in form of infinite-sites:

$$\begin{aligned} U_1 &= 01100,\ n_{U_1} = 2 \\ U_2 &= 00010,\ n_{U_2} = 1 \\ U_3 &= 10000,\ n_{U_3} = 1 \\ U_4 &= 10001,\ n_{U_4} = 1 \end{aligned} \tag{3.0.1}$$

where $n_{U_i},\ i = 1, 2, 3, 4$ are the multiplicities of the sequences in the sample. We note how the sequences are now on the form of table 1, which allows us to analyse sequence data via the coalescent process. For a complete derivation of the infinite-sites-model (we shall the derivation of another formulation of the infinite-sites-model later) see [Watterson, 1975], where the model is shown to apply to, among other population models, the Wright-Fisher model, which then gives rise to the coalescent formulation, as earlier shown.

3.1 Number of Segregating Sites

For data on the form in table 1, we look at number segregating sites. Since mutations are assumed to occur no more than once at each site, the number of segregating sites, S_n, is in fact the number of mutations since MRCA. From the section describing mutations we saw that mutations in a sample of n genes follow a Poisson process with parameter $tn\frac{\theta}{2}$, where the waiting time T_n^{mut}, describing the time to mutation in one of the n lineages, is exponentially distributed with parameter $n\frac{\theta}{2}$, where t is time in natural time scale. We note that the number of mutations in the sample is depending of the size of the sample at a given time, i.e. the number of ancestors at a given time, which is governed by the ancestral process, where the random variables $T_k \sim exp(\binom{k}{2})$ describe the waiting times to coalescent events. If we let $L_n = \sum_{i=2}^{n} jT_j$ be the length of the ancestral tree, then conditional on L_n, S_n follows a Poisson distribution with parameter $L_n\frac{\theta}{2}$. We shall see how S_n can be used to estimate the

Position	1 0 6	1 9 0	2 5 1	2 9 6	3 4 4	8 8	9 1	1 2 4	1 4 9	1 6 2	1 6 6	1 9 4	2 3 3	2 6 7	2 7 1	2 7 5	3 1 9	3 3 9	allele freqs.
Site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
allele																			
a	A	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	2
b	A	G	G	A	A	T	C	C	T	T	T	T	C	T	C	T	T	C	2
c	G	A	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	1
d	G	G	A	G	A	C	C	C	C	C	T	T	C	C	C	T	T	C	3
e	G	G	G	A	A	T	C	C	T	C	T	T	C	T	C	T	T	C	19
f	G	G	G	A	G	T	C	C	T	C	T	T	C	C	C	T	T	C	1
g	G	G	G	G	A	C	C	C	T	C	C	C	C	C	C	T	T	T	1
h	G	G	G	G	A	C	C	C	T	C	C	C	T	C	C	T	T	T	1
i	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	C	T	4
j	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	C	T	T	8
k	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	C	5
l	G	G	G	G	A	C	C	C	T	C	T	T	C	C	C	T	T	T	4
m	G	G	G	G	A	C	C	T	T	C	T	T	C	C	C	T	T	C	3
n	G	G	G	G	A	C	T	C	T	C	T	T	C	C	T	T	T	C	1

Table 1: Mitochondrial data presented in [Griffiths and Tavaré, 1994a]

mutation parameter θ. Observe that

$$\begin{aligned}\mathbb{E}(S_n) &= \mathbb{E}(\mathbb{E}(S_n|L_n)) \\ &= \mathbb{E}(L_n\theta/2) \\ &= \frac{\theta}{2}\sum_{i=2}^{n} i\mathbb{E}(T_i) \\ &= \theta\sum_{i=1}^{n-1}\frac{1}{i}.\end{aligned}$$

To calculate the variance of S_n, first define Y_j as the number of mutations occurring with j ancestors present in the ancestral tree. Then $S_n = \sum_{j=2}^{n} Y_j$, where the Y_j's are independent due to the independence of the waiting times T_j, $j = 2, ..., n$. The distribution of Y_j is given in the following lemma:

Lemma 3.1. *In a sample of size n let Y_j, $j = 2, ..., n$, be the number of mutations that occur when there are j ancestors in the sample. Then*

$$Y_j \sim Geo\left(\frac{j-1}{j-1+\theta}\right), \tag{3.1.1}$$

where $X \sim Geo(p)$, if $P(X = k) = (1-p)^k p$, for $k = 0, 1, ...$

Proof. As noted above, the number of mutations, when there are j ancestors, follow a Poission distribution with parameter $jt\frac{\theta}{2}$, with t being time in units of N generations, so

$$Y_j | T_j \sim Pois\left(jT_j\frac{\theta}{2}\right).$$

Now we are going to make three observations.

1. Assume that $Y \sim Pois(\lambda)$. Then for $s \in [0, 1]$
$$\mathbb{E}s^Y = \sum_{k=0}^{\infty} s^k \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(s\lambda)^k}{k!} = e^{-\lambda} e^{s\lambda} = e^{-\lambda(1-s)}.$$
2. Let $u \in \mathbb{R}$ and $Z \sim exp(\lambda)$. Then
$$\mathbb{E}e^{-uZ} = \lambda \int_0^{\infty} e^{-uz} e^{-\lambda z} dz = \lambda \int_0^{\infty} e^{-z(u+\lambda)} dz = \frac{\lambda}{u+\lambda}.$$
3. Assume that $X \sim Geo(p)$ such that $P(X = k) = (1-p)^k p$, for $k \in \mathbb{N}_0$. Then for $s \in [0, 1]$
$$\mathbb{E}s^X = \sum_{k=0}^{\infty} s^k p(1-p)^k = p \sum_{k=0}^{\infty} (s(1-p))^k = \frac{p}{1-s(1-p)} = \frac{p}{p+(1-p)(1-s)}.$$

Now we proceed with Y_j. Observe

$$\begin{aligned}
\mathbb{E}s^{Y_j} &= \mathbb{E}(\mathbb{E}(s^{Y_j} | T_j)) \\
&= \mathbb{E}\left(\exp\left(-j\frac{\theta}{2}T_j(1-s)\right)\right) \\
&= \frac{j(j-1)/2}{j(j-1)/2 + (1-s)j\theta/2} \\
&= \frac{j-1}{j-1+\theta(1-s)} \\
&= \frac{p}{p+(1-p)(1-s)},
\end{aligned}$$

where $p = \frac{j-1}{j-1+\theta}$. In the second equality we used the result derived 1. and in the third equality we used that derived in 2. $\mathbb{E}s^{Y_j}$ is now on the form derived in 3., which shows that $Y_j \sim Geo(p)$. □

Note that since $S_n = \sum_{j=2}^{n} Y_j$, the distribution of S_n is given by the convolution

$$\overset{n-1}{\underset{j=1}{\ast}} Geo\left(\frac{j}{j+\theta}\right). \tag{3.1.2}$$

By lemma 3.1 it is easy calculate the variance of S_n. Since the Y_j's are independent and the variance in a geometric distribution is known to be $\frac{1-p}{p^2}$, the variance of S_n is given by

$$\mathrm{Var}(S_n) = \sum_{j=2}^{n} \mathrm{Var}(Y_j) = \sum_{j=2}^{n} \frac{1-\frac{j-1}{j-1+\theta}}{\left(\frac{j-1}{j-1+\theta}\right)^2} = \sum_{j=2}^{n} \frac{\theta(j-1)+\theta^2}{(j-1)^2} = \theta \sum_{j=1}^{n-1} \frac{1}{j} + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2}.$$

The mean of S_n suggest the simple estimator for θ, first discovered by [Watterson, 1975],

$$\theta_W = \frac{S_n}{\sum_{j=1}^{n-1} \frac{1}{j}}.$$

The mean of the estimator is given by

$$\mathbb{E}\theta_W = \frac{\mathbb{E}S_n}{\sum_{j=1}^{n-1} \frac{1}{j}} = \theta,$$

and it follows that θ_W is an unbiased estimator for θ, with

$$\mathrm{Var}(\theta_W) = \mathrm{Var}(S_n)\left(\sum_{j=1}^{n-1} \frac{1}{j}\right)^{-2} = \left(\theta \sum_{j=1}^{n-1} \frac{1}{j} + \theta^2 \sum_{j=1}^{n-1} \frac{1}{j^2}\right)\left(\sum_{j=1}^{n-1} \frac{1}{j}\right)^{-2}.$$

The harmonic series is of order $\log(n)$, and thus $\mathrm{Var}(\theta_W)$ is of order $O(\log(n)^{-1})$, from which it follows that $\mathrm{Var}(\theta_W) \to 0$ as $n \to \infty$, showing that θ_W is a consistent estimator of θ. While the Wattersons estimator is unbiased and consistent, it is not in general sufficient. In [RoyChoudhury and Wakeley, 2010] they show that S_n is sufficient in the finite-sites-model when the number of sites tends to infinity, but as our focus is the infinite-sites-model this case will not be treated. Instead we will see how θ_W performs in general.

Lemma 3.1 gives the distribution of the number of mutations in the genealogy, and if we could observe the number of mutations in every state of the genealogy, lemma 3.1

could be used to calculate the likelihood and then, by classical statistical tools, used in estimating θ. This is of course more information than one could hope to obtain in reality, but it serves as a standard from which to compare summary statistics.

From 3.1 we see that the likelihood for a sample of n individuals is

$$\begin{aligned} L_n(\theta) &= \prod_{j=2}^{n} \left(\frac{\theta}{\theta+j-1}\right)^{Y_j} \left(\frac{j-1}{\theta+j-1}\right) \\ &= \theta^{\sum_{j=2}^{n} Y_j} \prod_{j=2}^{n} (\theta+j-1)^{-(Y_j+1)} (j-1) \\ &= \theta^{S_n} (n-1)! \prod_{j=2}^{n} (\theta+j-1)^{-(Y_j+1)}. \end{aligned}$$

The log-likelihood along with its first order derivative is

$$\begin{aligned} \ell_n(\theta) = \log L_n(\theta) &= S_n \log(\theta) + \log((n-1)!) - \sum_{j=2}^{n} \log(\theta+j-1)(Y_j), \\ \frac{\partial \ell_n}{\partial \theta} &= \frac{S_n}{\theta} - \sum_{j=2}^{n} \frac{Y_j+1}{\theta+j-1}, \end{aligned}$$

and hence the maximum likelihood estimator satisfies the equation

$$\theta = \frac{S_n}{\sum_{j=2}^{n} \frac{Y_j+1}{\theta+j-1}}.$$

Observe that

$$\begin{aligned} \frac{\partial^2 \ell_n}{\partial \theta^2} &= -\frac{S_n}{\theta^2} + \sum_{j=2}^{n} \frac{Y_j+1}{(\theta+j-1)^2}, \\ -\mathbb{E}\left(\frac{\partial^2 \ell_n}{\partial \theta^2}\right) &= \frac{\theta \sum_{j=2}^{n} \frac{1}{j-1}}{\theta^2} - \sum_{j=2}^{n} \frac{\mathbb{E}(Y_j+1)}{(\theta+j-1)^2} \\ &= \frac{1}{\theta} \sum_{j=2}^{n} \frac{1}{j-1} - \sum_{j=2}^{n} \left(\frac{\theta}{j-1}+1\right) (\theta+j-1)^{-2} \\ &= \frac{1}{\theta} \sum_{j=2}^{n} \frac{1}{j-1} - \sum_{j=2}^{n} ((j-1)(\theta+j-1))^{-1} \\ &= \frac{1}{\theta} \sum_{j=1}^{n-1} \frac{1}{j} + \frac{\theta}{j(\theta+j)} \\ &= \frac{1}{\theta} \sum_{j=1}^{n-1} \frac{1}{j+\theta}. \end{aligned}$$

This gives that the variance of any unbiased estimator θ_U satisfy

$$\frac{\theta}{\sum_{j=1}^{n-1} \frac{1}{j+\theta}} \leq \mathrm{Var}(\theta_U),$$

by the Cramér-Rao lower bound, where the left hand side is also the asymptotic variance of the MLE. For a fixed θ we see that

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\theta_{MLE})}{\mathrm{Var}(\theta_W)} = \lim_{n\to\infty} \frac{\left(\sum_{j=1}^{n-1} \frac{1}{j}\right)^2}{\left(\sum_{j=1}^{n-1} \frac{1}{j+\theta}\right)\left(\sum_{j=1}^{n-1} \frac{1}{j} + \theta \sum_{j=1}^{n-1} \frac{1}{j^2}\right)} = 1.$$

Certainly $\mathrm{Var}(\theta_{MLE}) \leq \mathrm{Var}(\theta_W)$, but the result above shows that θ_W is asymptotically optimal. Consider now the case where θ is large. Then for fixed n

$$\lim_{\theta\to\infty} \frac{\mathrm{Var}(\theta_{MLE})}{\mathrm{Var}(\theta_W)} = \frac{\left(\sum_{j=1}^{n-1} \frac{1}{j}\right)^2}{(n-1)\sum_{j=1}^{n-1} \frac{1}{j^2}},$$

shows a decrease in efficiency for large values of θ. This indicates that there is value in using maximum likelihood estimation when inferring θ. As mentioned above, we will not in reality obtain data equivalent to Y_j, but the result serves as a motivator for the development of maximum likelihood schemes, which will be the focus in the sections to come. It is worth noting that θ_W is asymptotically optimal, and when n is sufficiently large it might serve as a more practical tool than computationally intensive likelihood schemes.

3.2 Genealogical Trees

From (3.0.1) we saw how a coalescent tree can be condensed into data on infinite-sites-form. In reality it is the other way around - for a given data set on the infinite-sites-form we will try to describe it using the infinite-sites-model. Hence, it is the data that gives rise to the coalescent tree in practice. The coalescent tree in figure 2 is just one tree consistent with the data as there are several ways of ordering the sequences that gives rise to the same data. The question is now, how to extract information about the genealogy from the data. The first observation we make, is that under the infinite-sites-model, mutations can only occur once at every site. This means that at every site there is a an ancestral nucleotide of type A or G for the purine nucleotides and T or C for the pyrimidine. If we label the sites 0 or 1 and let 0 denote the ancestral site, then mutations in the sequences can be read directly from the data as the sites containing a 1 for a given sequence. A way of representing a data matrix of 0's and 1's is by a *rooted tree*. The rooted tree assumes an ancestral type for each site, and each mutational site is then represented by a vertex on the

tree, labelled by the position of the site in the data. The individuals are then given by a sequence of mutations separating them from the common ancestor. The rooted tree can be thought of as a coalescent tree stripped of its time topology.

A way of finding a rooted tree from a data matrix of 0's and 1's is given in [Gusfield, 1991] and goes as follows

- Remove duplicated columns in the data matrix.
- View each column as a binary number and sort the matrix by ordering the columns in ascending order from left to right.
- Label vertices in the tree by the column label and construct paths from the leaves (being the alleles) to the root by reading the columns from right to left, where 1 occurs.

As an example, observe again the data given in table 1. By labelling the most frequent base 0 and the others 1, we can use the algorithm to find a rooted tree corresponding to the data. The rooted tree produced is given in table 2 and every allele is now given as a sequence of its mutations since the MRCA with the left most column describing the most recent mutations. Every sequence is given a 0 as the earliest "mutation", but this is just to indicate the compliance with the common ancestor when stripping a sequence of its mutations. Figure 3 gives an illustration of the rooted tree produced by the algorithm, and it is noticed that the rooted tree indeed corresponds to a coalescent tree without the time topology. This is a key observation as it allows us to express the probability of a rooted tree - the sample probability - through a recursion on the earlier stages on the tree. The recursions will form the basis for likelihood inference, as we shall see in the later sections.

Notice that in figure 3 sequence k corresponds to the ancestral type given by the root, as a consequence of choosing the most frequent bases in table 1 to be ancestral. In reality the ancestral type is not known, and the rooted tree presented in figure 3 is just one rooted tree compatible with the data. All that can be deduced from the data is the number of segregating sites between each sequence, which leads to the formulation of *unrooted trees*.

Data on the form presented in table 1 is equivalent to an unrooted trees, where

Allele							Allele freqs.
a	1	6	4	14	0		2
b	10	1	6	4	14	0	8
c	2	18	0				1
d	9	3	0				3
e	6	4	14	0			19
f	5	6	4	14	0		1
g	12	11	18	0			1
h	13	12	11	18	0		1
i	17	16	18	0			4
j	16	18	0				3
k	0						5
l	18	0					4
m	8	0					3
n	7	15	0				1

Table 2: Rooted tree corresponding to data given in table 1.

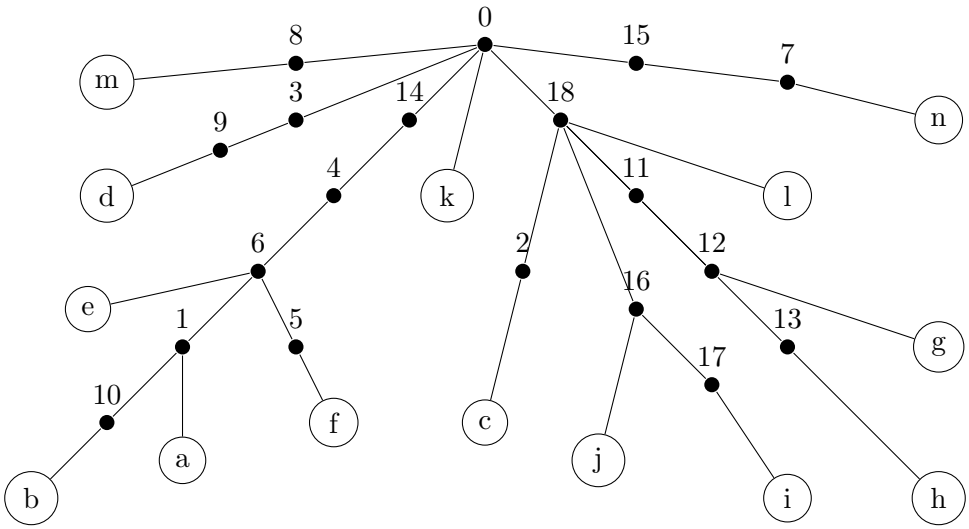

Figure 3: Rooted tree corresponding to table 2.

the vertices represent distinct lineages in the data and mutations are placed on the edges between vertices. An unrooted tree can be constructed from any rooted tree by reorganizing the tree by letting the leaves in the rooted tree be represented by

vertices and connecting them by edges labelled by the mutations dividing each lineage. Figure 4 shows the unrooted tree corresponding to the rooted tree in figure 2. The root in figure 2 has the type given by sequence k in the data set, since sequence k consists of only the most frequent base types, but if the root is not in the data, it will not be in the unrooted tree. Furthermore, every rooted tree corresponding to the data can be found from the unrooted tree, by placing the root at a vertex or between mutations appearing on the same edge. This observation will become useful when calculating sample probabilities, since it turns out that the probability of obtaining the unrooted tree equivalent to the data is just the sum of the probabilities of the corresponding rooted trees, but more on that later. If S is the number of segregating sites in the data, then there are $S+1$ rooted trees corresponding to the unrooted tree [Griffiths and Tavaré, 1995].

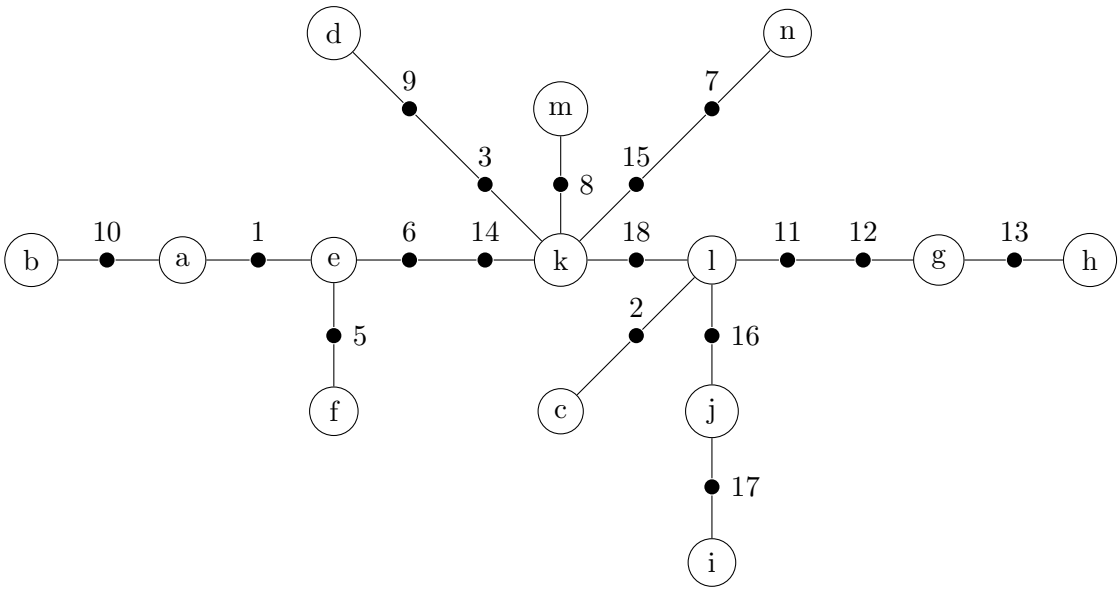

Figure 4: Unrooted tree corresponding to the data in table 1

In order to establish a framework that allows for likelihood inference we first need to calculate sample probabilities that take into account the whole genealogical history. To do this we turn to the formulation of rooted trees given in [Ethier and Griffiths, 1987].

Suppose we have a sample of n sequences, and let $\mathbf{x}_i$ be the i'th sequence, where $\mathbf{x}_i = (x_{i0}, x_{i1}, x_{i2}, ...)$, $x_{ik} \in E = [0,1]^{\mathbb{Z}_+}$, $i \in \{1, ..., n\}, k \in \mathbb{N}$. Then $x_{i0}, x_{i1}, x_{i2}, ...$ is the

sites at which mutations have occurred for the i'th gene, where x_{i0} is the most recent mutation to individual i. This is achieved by taking $\mathbf{x}_i = (x_{i0}, x_{i1}, x_{i2}, ...)$ at generation r, looking forward in time. In generation $r+1$, an individual chooses a parent uniformly at random among the individuals in generation r, and if the parent is of type $\mathbf{x}_i$, the offspring is of type $\mathbf{x}_i$ with probability $(1-\mu)$, not undergoing mutation. The offspring is of type $(Y, \mathbf{x}_i)$ with probability μ, where Y is a uniformly distributed random variable on the unit interval. At any generation take a sample of n genes. The state of the model is $(\mathbf{x}_1, ..., \mathbf{x}_n)$, and in [Ethier and Griffiths, 1987], theorem 2.1, they show that the model is governed by a Markov chain on $\mathcal{P}(E)$, the set of Borel probability measures on E. Furthermore they show that this Markov chain converges in distribution to a Markov process indeed describing the coalescent process given in definition 2.1 with the mutation process superimposed on it. Measure-theoretic concerns are to be acknowledged, since the state space of the model has become uncountably infinite making derivations analogous to those in section 2 insufficient. We will not go more into detail about the measure-theoretic justifications of the model, but refer the interested reader to [Ethier and Griffiths, 1987] for a rigorous mathematical explanation. From here on we assume theorem 2.1, without further mention, thus assuming that the new formulation indeed follows the coalescent process.
To see, intuitively, that this is essentially the same model as previously described, consider three sequences from the [Watterson, 1975] formulation,

$$(..., 1, 1, 0, 0, 0, 0, 0, ...),$$
$$(..., 1, 0, 1, 1, 1, 0, 0, ...),$$
$$(..., 1, 0, 1, 0, 0, 1, 0, ...).$$

and three sequnces from the formulation of [Ethier and Griffiths, 1987],

$$(y_1, x_0, x_1, ...),$$
$$(y_2, y_3, y_4, x_0, x_1, ...),$$
$$(y_2, y_5, x_0, x_1, ...),$$

where $y_1, y_2, y_3, y_4, y_5, x_0, x_1$ are all distinct. Figure 5 illustrates how the new representation of the infinite-sites-data is equivalent to that earlier described.

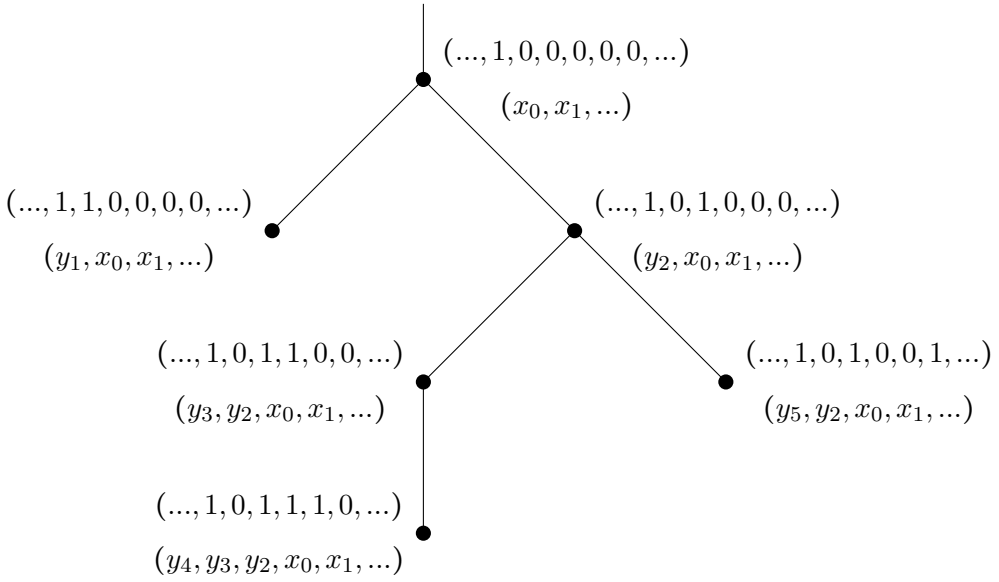

Figure 5: Figure from [Ethier and Griffiths, 1987].

Following the derivation in [Ethier and Griffiths, 1987], we say that a sample $(\mathbf{x}_1, ..., \mathbf{x}_n)$ is a tree if the following three conditions are met:

The coordinates $x_{ij},\ j = 0, 1, ...,$ of $\mathbf{x}_i$ are distinct for fixed $i \in \{1, ..., n\}$, (3.2.1)

if $i, i^{'} \in \{1, ..., n\},\ j, j^{'} \in \mathbb{Z}_+$ and $x_{ij} = x_{i^{'}j^{'}},$ then $x_{i,j+l} = x_{i^{'},j^{'}+l},$ for $l = 0, 1, ...,$ (3.2.2)

there exist $j_1, ..., j_n \in \mathbb{Z}_+$ such that $x_{1j_1} = \cdots = x_{nj_n}$. (3.2.3)

Condition (3.2.1) means that mutations never occur more than once at the same site, (3.2.2) means that if two genes have have ancestors who share their most recent mutation at the same site, then the ancestors are of the same type and (3.2.3) means that the n genes in the sample share a common ancestor.

Define $\mathcal{J}_n = \{(\mathbf{x}_1, ..., \mathbf{x}_n) \in E^n \mid (\mathbf{x}_1, ..., \mathbf{x}_n)$ is a tree$\}$ and define the equivalence-relations $\sim$ and $\approx$ by

$$(\mathbf{x}_1, ..., \mathbf{x}_n) \sim (\mathbf{y}_1, ..., \mathbf{y}_n)$$

if there exist a bijection $\zeta : [0, 1] \to [0, 1]$, such that $y_{ij} = \zeta(x_{ij}),\ i = 1, .., n,\ j = 0, 1, 2, ...,$ and

$$(\mathbf{x}_1, ..., \mathbf{x}_n) \approx (\mathbf{y}_1, ..., \mathbf{y}_n)$$

if there exist a bijection $\zeta : [0,1] \to [0,1]$ and a permutation σ of $(1,..,n)$ such that $y_{\sigma(i)j} = \zeta(x_{ij})$. Then the quotient set $\mathcal{J}_n/\sim$ corresponds to the labelled trees, while $\mathcal{J}_n/\approx$ corresponds to the unlabelled and hence unordered trees. In [Ethier and Griffiths, 1987] it is stated that the distributions for all $T \in \mathcal{J}_n/\approx$ is yet to be discovered (though the article is from 1987, the author of this paper know of no such result), but this is not insurmountable as we will develop sample distributions for the labelled trees from which distributions of the unlabelled trees are found through combinatorics on the labelling. This is due to the following observation made in [Ethier and Griffiths, 1987]: given $T \in \mathcal{J}_n/\sim$ let $[T]$ denote the equivalence class in $\mathcal{J}_n/\approx$ with $T \subseteq [T]$. Let $c(T)$ be the number of equivalence classes $T' \in \mathcal{J}_n/\sim$ where $T' \subseteq T$. If $p(T)$ is the probability of obtaining $T \in \mathcal{J}_n/\sim$ then

$$p([T]) = c(T)p(T). \tag{3.2.4}$$

Thus it will suffice to consider trees $T \in \mathcal{J}_n/\sim$ since the distributions of the labelled and unlabelled trees, respectively, are proportional given a combinatoric factor on the number of labelled trees contained in the unlabelled. This is convenient, since in reality we do not know the labelling of trees, and the labelling provided in a data set, such as table 1, is arbitrarily given.

In a sample $(\mathbf{x}_1, ..., \mathbf{x}_n)$, multiple of the sequences might be identical and to comprise this information denote the multiplicities of the sample by $\mathbf{n} \in \mathbb{N}^d$ if there are d distinct sequences, such that $\mathbf{n} = n_1 + \cdots + n_d$. Define $\Phi_n : E^d \to E^n$ by

$$\Phi_n(\mathbf{x}_1, ..., \mathbf{x}_d) = (\mathbf{x}_1, ..., \mathbf{x}_1, ..., \mathbf{x}_d, ..., \mathbf{x}_d)$$

where $\mathbf{x}_k$ appears n_k times in the sample. We see that Φ_n maps $\mathcal{J}_d/\sim$ onto $\mathcal{J}_n/\sim$ and if we let $p(T, \mathbf{n})$ be the probability of obtaining $T \in \mathcal{J}_n/\sim$ with multiplicities $\mathbf{n}$, it is clear that $p(T, \mathbf{n})$ is the probability that the sample forms a tree of class $\Phi_n(T)$. On this note define

$$(\mathcal{J}_d/\sim)_0 = \{T \in \mathcal{J}_d/\sim \mid \mathbf{x}_1, ..., \mathbf{x}_d \text{ are distinct for all } (\mathbf{x}_1, ..., \mathbf{x}_d) \in T\},$$

and let

$$\mathcal{J}^* = \bigcup_{d=1}^{\infty} \left[(\mathcal{J}_d/\sim)_0 \times \mathbb{N}^d \right].$$

Thus, $\mathcal{J}^*$ is the set of all labelled ordered trees with multiplicities, and it will form the basis for our derivation of sample probabilities. Before we jump to the calculation of $p(T, \mathbf{n})$, we shall see that the new formulation is essentially equivalent to the one given by [Watterson, 1975]. Let

$$\mathcal{J}_{s,n} = \{(\mathbf{x}_1, ..., \mathbf{x}_n) \in \mathcal{J}_n \mid (\mathbf{x}_1, ..., \mathbf{x}_n) \text{ has s segregating sites}\},$$

and observe the following:

Theorem 3.2. *In a sample of $n \in \mathbb{N}$ genes, take $T \in \mathcal{J}_{s,n}$. Then*

$$p(s, n) = \left[\bigstar_{j=1}^{n-1} Geo\left(\frac{j}{j+\theta}\right)\right](\{s\}) \tag{3.2.5}$$

where $p(s, n)$ is probability of obtaining T.

Proof. The theorem is a restatement of a part of corollary 4.6 in [Ethier and Griffiths, 1987] from which the proof follows. Essentially the proof in 4.6 goes by showing that the probability generating function for $p(s, n)$ is on the form

$$g(\xi, n) = \sum_{s=0}^{\infty} p(s, n)\xi^s = \prod_{j=1}^{n-1} \frac{j}{j + \theta(1-\xi)}$$

since

$$\mathbb{E}\xi^{S_n} = \prod_{j=2}^{n} \mathbb{E}\xi^{Y_j} = \prod_{j=1}^{n-1} \frac{j}{j + \theta(1-\xi)}$$

by the calculations in the the proof of lemma 3.1, which shows that $T \overset{D}{=} S_n$, where

$$S_n \sim \left[\bigstar_{j=1}^{n-1} Geo\left(\frac{j}{j+\theta}\right)\right]$$

by (3.1.2). □

Theorem 3.2 shows that the new formulation is indeed equivalent to the one presented earlier, so it is sensible to use the new formulation to describe infinite sites data and, as is the primary target earlier motivated, to construct sample probabilities to infer θ through the likelihood.

3.3 Rooted Tree Probabilities

Let $p(T, \mathbf{n})$ be the probability of obtaining $(T, \mathbf{n}) \in \mathcal{J}^*$. Thus, $p(T, \mathbf{n})$ is the probability of obtaining a given ordered sample of d distinct sequences with multiplicities $\mathbf{n}$.

The following theorem was established in [Ethier and Griffiths, 1987], but the proof presented here is quite different and utilizes the coalescent structure of the graph T.

Theorem 3.3. *For $(T, \boldsymbol{n}) \in \mathcal{J}^*$ let $p(T, \boldsymbol{n})$ be the probability of obtaining T with multiplicity $\boldsymbol{n}$. Define the shift operator $\mathscr{F} : E \to E$ by*

$$\mathscr{F}(x_0, x_1, ...) = (x_1, x_2, ...)$$

and let $\mathscr{F}_k : E^d \to E^d$ for $k = 1, ..., d$ by

$$\mathscr{F}_k(\boldsymbol{x}_1, ..., \boldsymbol{x}_d) = (\boldsymbol{x}_1, ..., \boldsymbol{x}_{k-1}, \mathscr{F}\boldsymbol{x}_k, ..., \boldsymbol{x}_d).$$

Let $\mathscr{R} : E^d \to E^{d-1}$ for $k = 1, ..., d$ and $d \geq 2$ by

$$\mathscr{R}_k(\boldsymbol{x}_1, ..., \boldsymbol{x}_d) = (\boldsymbol{x}_1, ..., \boldsymbol{x}_{k-1}, \boldsymbol{x}_{k+1}, ..., \boldsymbol{x}_d).$$

Then $p(T, \boldsymbol{n})$ satisfy the system of linear equations

$$\begin{aligned} n(n-1+\theta)p(T,\boldsymbol{n}) = &\sum_{k:n_k\geq 2} n_k(n_k-1)p(T,\boldsymbol{n}-e_k) \\ &+\theta \sum_{\substack{k:n_k=1 \\ x_{k0}\ distinct \\ \mathscr{F}\boldsymbol{x}_k\neq\boldsymbol{x}_j,\ \forall j}} p(\mathscr{F}_k T,\boldsymbol{n}) \\ &+\theta \sum_{\substack{k:n_k=1 \\ x_{k0}\ distinct}} \sum_{j:\mathscr{F}\boldsymbol{x}_k=\boldsymbol{x}_j} p\left(\mathscr{R}_k T, \mathscr{R}_k(\boldsymbol{n}+e_j)\right), \end{aligned} \tag{3.3.1}$$

with boundary condition $p(T_1, 1) = 1$. Here 'x_{k0} distinct' means $x_{k0} \neq x_{ij}$ for all $(\boldsymbol{x}_1, ..., \boldsymbol{x}_d)$ with $(k, 0) \neq (i, j)$.

Before proving theorem 3.3, we need a lemma to calculate the probability of the first event back in time.

Lemma 3.4. *For a sample of size n let $T_n^f = \{coalescent, mutation\}$ be a Bernoulli variable recording the first event back in time in the coalescent process. Then*

$$P(T_n^f = x) = \begin{cases} \frac{n-1}{n-1+\theta} & x = coalescent \\ \frac{\theta}{n-1+\theta} & x = mutation \end{cases}.$$

Proof. Let T_n^{mut} denote the time to mutation, and let T_n^{coal} denote the time to coalescent. By the definition of the coalescent and mutation process

$$T_n^{mut} \sim Exp\left(\frac{n\theta}{2}\right), \quad T_n^{coal} \sim Exp\left(\binom{n}{2}\right).$$

Let A and B be two independent exponentially distributed random variables with parameters μ and λ respectively. Then

$$P(A < B) = \frac{\mu}{\mu + \lambda}, \quad P(B < A) = \frac{\lambda}{\mu + \lambda}.$$

Applying this to T_n^{mut} and T_n^{coal} gives the result. □

proof of theorem 3.3. Below, we assume lemma 3.4 without further mention. Devide $p(T, \mathbf{n})$ into

$$p(T, \mathbf{n}) = P(T, \mathbf{n} | T_n^f = \text{coal}) P(T_n^f = \text{coal}) + P(T, \mathbf{n} | T_n^f = \text{mut}) P(T_n^f = \text{mut}). \quad (3.3.2)$$

For the left term we have

$$\begin{aligned}
&P(T, \mathbf{n} \big| T_n^f = \text{coal}) P(T_n^f = \text{coal}) \\
&= \frac{n-1}{n-1+\theta} \sum_{k:n_k \geq 2} P\left((T, \mathbf{n} - e_k), (T, \mathbf{n}) \mid T_n^f = \text{coal}\right) \\
&= \frac{n-1}{n-1+\theta} \sum_{k:n_k \geq 2} P\left((T, \mathbf{n}) \mid (T, \mathbf{n} - e_k), T_n^f = \text{coal}\right) P\left((T, \mathbf{n} - e_k) \mid T_n^f = \text{coal}\right) \\
&= \frac{n-1}{n-1+\theta} \sum_{k:n_k \geq 2} P\left((T, \mathbf{n}) \mid (T, \mathbf{n} - e_k), T_n^f = \text{coal}\right) P\left(T, \mathbf{n} - e_k\right) \\
&= \frac{1}{n(n-1+\theta)} \sum_{k:n_k \geq 2} n_k(n_k - 1) P\left(T, \mathbf{n} - e_k\right)
\end{aligned}$$

where the third equality follows from the Markov property, and the fourth follows from the fact that

$$P\left((T, \mathbf{n}) \mid (T, \mathbf{n} - e_k), T_n^f = \text{coal}\right) = \frac{n_k - 1}{n-1} \frac{n_k}{n}.$$

This is seen as the event forward in time of obtaining $(T, \mathbf{n})$ from $(T, \mathbf{n} - e_k)$ by adding a duplicated sequence to the tree $(T, \mathbf{n} - e_k)$. Starting with $(T, \mathbf{n} - e_k)$, we choose a sequence uniformly at random among the $n-1$ sequences and add a duplicate of that sequence to the tree. There are then $n_k - 1$ sequences that, by adding one of them to the tree, would result in $(T, \mathbf{n})$. There are n possible "places" for the duplicated sequence to fall, but only n_k of them would give the ordering of $(T, \mathbf{n})$.

Now, consider the right hand term in (3.3.2). Here $P((T,\mathbf{n}) \mid T_n^f = \text{mut})$ is the probability of obtaining $(T,\mathbf{n})$ from the tree at time $T_n^f + \epsilon$. It is possible to split the event of obtaining $(T,\mathbf{n})$ by a mutation into two disjoint events.

The first being that one of the unique sequences is chosen to mutate. The tree before mutation must then be of type $(\mathscr{F}_k T,\mathbf{n})$ if the k'th sequences is chosen to mutate. There are $\#\{k : n_k = 1\}$ unique sequences to choose from, but only those where $x_{k0} \neq x_{ij}, (k,0) \neq (i,j)$, meaning that the most recent mutation of sequence k is not shared by other sequences, and $\mathscr{F}\mathbf{x}_k \neq \mathbf{x}_j \ \forall j$, meaning that $\mathbf{x}_k$ is still unique when stripped of its most recent mutation, would result in $(T,\mathbf{n})$.

The second event resulting in $(T,\mathbf{n})$ by mutation is if a duplicated sequence undergoes a mutation. The tree before mutation must then be of type $(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j))$, if the k'th sequence is chosen to mutate and if it was of type $\mathbf{x}_j$ before mutation. Again, there are $\#\{k : n_k = 1\}$ possible unique sequences from $(T,\mathbf{n})$ that could have mutated with $x_{k0} \neq x_{ij}, (k,0) \neq (i,j)$, where for each k, there are $\#\{j : \mathscr{F}\mathbf{x}_k = \mathbf{x}_j\}$ possible sequences where a mutation would result in $(T,\mathbf{n})$.

In each of the two cases, every time we choose a sequence to mutate, that sequence is chosen uniformly at random among the n sequences. The considerations about the two cases of mutation combined with the calculations of the left hand term of (3.3.2) then gives

$$\begin{aligned} p(T,\mathbf{n}) &= \frac{1}{n(n-1+\theta)} \sum_{k:n_k\geq 2} n_k(n_k-1)P\left(T,\mathbf{n}-e_k\right) + P(T,\mathbf{n}|T_n^f = \text{mut})P(T_n^f = \text{mut}) \\ &= \frac{1}{n(n-1+\theta)} \sum_{k:n_k\geq 2} n_k(n_k-1)P\left(T,\mathbf{n}-e_k\right) \\ &\quad + \frac{\theta}{n-1+\theta}\left(\sum_{\substack{k:n_k=1 \\ x_{k0}\text{ distinct} \\ \mathscr{F}\mathbf{x}_k\neq\mathbf{x}_j,\ \forall j}} \frac{1}{n} p(\mathscr{F}_k T,\mathbf{x}) + \theta \sum_{\substack{k:n_k=1 \\ x_{k0}\text{ distinct}}} \sum_{j:\mathscr{F}\mathbf{x}_k=\mathbf{x}_j} \frac{1}{n} p\left(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j)\right)\right). \end{aligned} \tag{3.3.3}$$

Multiplying with $n(n-1+\theta)$ gives the result. □

For the purpose of simulating recursions on the form of (3.3.1), it is more convenient to consider the probability $p^0(T, \mathbf{n})$ given by

$$p^0(T, \mathbf{n}) = \frac{n!}{n_1! \cdots n_d!} p(T, \mathbf{n}). \tag{3.3.4}$$

which is the probability of a labelled $(T, \mathbf{n})$ with the ordering removed. This is due to the earlier observation, that when presented with data, we do not know the ordering or labelling of the tree. The recursion for $p^0(T, \mathbf{n})$ is given in the following corollary.

Corollary 3.5. *Let* $(T, \boldsymbol{n}) \in \mathcal{J}^*$. *then*

$$\begin{aligned} n(n-1+\theta)p^0(T, \boldsymbol{n}) = &\sum_{k:n_k \geq 2} n(n_k - 1)p^0(T, \boldsymbol{n} - e_k) \\ &+ \theta \sum_{\substack{k:n_k=1 \\ x_{k0} \text{ distinct} \\ \mathscr{F}\boldsymbol{x}_k \neq \boldsymbol{x}_j, \ \forall j}} p^0(\mathscr{F}_k T, \boldsymbol{x}) \\ &+ \theta \sum_{\substack{k:n_k=1 \\ x_{k0} \text{ distinct}}} \sum_{j:\mathscr{F}\boldsymbol{x}_k = \boldsymbol{x}_j} (n_j + 1)p^0\left(\mathscr{R}_k T, \mathscr{R}_k(\boldsymbol{n} + e_j)\right), \end{aligned} \tag{3.3.5}$$

with boundary condition $p^0(T_1, 1) = 1$.

Proof. Observe that

$$\begin{aligned} p(T, \mathbf{n}) &= \frac{n_1! \cdots n_d!}{n!} p^0(T, \mathbf{n}) \\ p(T, \mathbf{n} - e_k) &= \frac{n_1! n_{k-1}! (n_k - 1)! n_{k+1} \cdots n_d!}{(n-1)!} p^0(T, \mathbf{n} - e_k) \\ p(\mathscr{F}_k T, \mathbf{n}) &= \frac{n_1! \cdots n_d!}{n!} p^0(\mathscr{F}_k T, \mathbf{n}) \\ p\left(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n} + e_j)\right) &= \frac{n_1! \cdots (n_j + 1)! \cdots (n_k - 1)! \cdots n_d!}{n!} p^0\left(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n} + e_j)\right) \end{aligned} \tag{3.3.6}$$

plugging this into (3.3.1) and multiplying by $\frac{n!}{n_1! \cdots n_d!}$, theorem 3.3 gives the result, since in the last term of (3.3.5)

$$n_k = 1 \ \Rightarrow \ n_k! = (n_k - 1)! = 1.$$

□

In (3.2.4) we saw that when considering an unlabelled tree, it suffices to consider the corresponding labelled tree. lemma 3.5 shows how to calculate the sample distribution of a labelled tree when removing the ordering of the labels, by utilizing the

recursion formula given in theorem 3.3. Below we extend this result to cover the distribution of unlabelled, unordered trees.

Let P_d denote the set of permutations of $(1, ..., d)$. Then for $T \in \mathcal{J}_d/\sim$, $\mathbf{n} \in \mathbb{N}^d$, and $\sigma \in P_d$, define

$$T_\sigma = \left\{(\mathbf{x}_{\sigma(1)}, ..., \mathbf{x}_{\sigma(d)}) : \ (\mathbf{x}_1, ..., \mathbf{x}_1) \in T\right\}$$

and

$$\mathbf{n}_\sigma = (n_{\sigma(1)}, ..., n_{\sigma(d)}).$$

For $(T, \mathbf{n}) \in \mathcal{J}*$ let

$$a(T, \mathbf{n}) = \#\{\sigma \in P_d : \ T_\sigma = T, \ \mathbf{n}_\sigma = \mathbf{n}\},$$

and

$$a(\mathbf{n}) = \#\{\sigma \in P_d : \ \mathbf{n}_{\sigma(d)} = \mathbf{n}\}.$$

If $p^*(T, \mathbf{n})$ denotes the probability of the corresponding unlabelled tree, then it is given by

$$p^*(T, \mathbf{n}) = \frac{1}{a(T, \mathbf{n})} p^0(T, \mathbf{n}),$$

as shown in [Ethier and Griffiths, 1987].

3.4 Unrooted Tree Probabilities

So far we have seen how to establish sample probabilities for genealogical trees through the work of [Ethier and Griffiths, 1987], where the derivation was carried out assuming that the trees are rooted. We will extend this formulation to cover the unrooted trees as well and establish sample distributions analogous to $p(T, \mathbf{n})$, $p^0(T, \mathbf{n})$ and $p^*(T, \mathbf{n})$.

Following the formulation in [Griffiths and Tavaré, 1995], a labelled unrooted tree is defined by its vertex set V and its set of edges $\mathbf{Q}$ along with the number of mutations on each edges, described by $m_{ij}, i, j \in V$, where m_{ij} is the number of mutations dividing vertices i and j, i.e., m_{ij} is the number of segregating sites between i and j. As previously $\mathbf{n}$ denotes the multiplicities of the sequences. An unrooted labelled genealogy is then described by $(\mathbf{Q}, \mathbf{n})$. As with the rooted trees, we want a formulation

describing the equivalence relations $\sim_u$ and $\approx_u$ of the trees. For a tree $T = (\mathbf{x}_1, ..., \mathbf{x}_d)$ let $C(T)$ be a $d \times d$ matrix defined by

$$C_{ab} = \left|\{x_{ai} \in \mathbf{x}_a,\ x_{ai} \notin \mathbf{x}_b\} \cup \{x_{bj} \in \mathbf{x}_b,\ x_{bj} \notin \mathbf{x}_a\}\right|, \quad i, j \in \mathbb{N},\ a, b \in V.$$

Thus $C(T)$ is a matrix describing the number of segregating sites between vertices in the tree. For two trees, T_1 and T_2, let $T_1 \sim_u T_2$ if $C(T_1) = C(T_2)$, that is, two labelled unrooted trees are equivalent if hey have the same number of mutations between their vertices. As with $\approx$ define $\approx_u$ such that $T_1 \approx_u T_2$ if there exist a permutation $\sigma \in P_d$ where $C(T_{1\sigma}) = C(T_2)$. As with $\sim_u$, equivalence of unlabelled unrooted trees is determined by the number of differences between sequences in the trees.

The following observation from [Griffiths and Tavaré, 1995] is the key observation on probabilities of unrooted trees. It describes a link between the rooted and unrooted trees allowing both for the derivation of sample probabilities for unrooted trees, and for the calculation of these probabilities by calculation of the rooted trees corresponding to the unrooted:
Let $p(\mathbf{Q}, \mathbf{n})$ be the probability of an unrooted tree. Suppose that ξ is an equivalence class under $\sim_u$ with pairwise-difference-of-mutations matrix C_0. Then ξ is the union of equivalence classes of trees T under $\sim_u$ with $C(T) = C_0$. This gives that

$$p(\mathbf{Q}, \mathbf{n}) = \sum_{T:C(T)=C_0} p(T, \mathbf{n}) \tag{3.4.1}$$

for distinct T. Then $p(\mathbf{Q}, \mathbf{n})$ satisfy a recursion similar to (3.3.1):

Corollary 3.6. *Let $p(\boldsymbol{Q}, \boldsymbol{n})$ be the probability of obtaining a labelled unrooted tree under $\sim_u$. Then*

$$\begin{aligned} n(n-1+\theta)p(\boldsymbol{Q}, \boldsymbol{n}) = &\sum_{k:n_k \geq 2} n_k(n_k-1)p(\boldsymbol{Q}, \boldsymbol{n} - e_j) \\ &+ \theta \sum_{\substack{k:n_k=1,\ |k|=1 \\ k \to j,\ m_{kj}>1}} p(\boldsymbol{Q} - e_{kj}, \boldsymbol{n}) \\ &+ \theta \sum_{\substack{k:n_k=1,\ |k|=1 \\ k \to j,\ m_{kj}=1}} p(\boldsymbol{Q} - e_{kj}, \boldsymbol{n} + e_j - e_k), \end{aligned} \tag{3.4.2}$$

where $|k| = 1$ means that the degree of vertex k is 1, that is, vertex k is a leaf. $k \to j$ means that vertex k and j is connected with an edge. The boundary conditions for

$n = 2$ are

$$p((0), 2e_1) = \frac{1}{1+\theta},$$

and

$$p((m), e_1 + e_2) = \left(\frac{\theta}{1+\theta}\right)^m \frac{1}{1+\theta},\ m = 1, 2, ...$$

Proof. Theorem 3.3 combined with (3.4.1) gives

$$\begin{aligned} n(n-1+\theta)p(Q,\mathbf{n}) = &\sum_{k:n_k\geq 2}\sum_{T:C(T)=C_0} n_k(n_k-1)p(T,\mathbf{n}-e_k) \\ &+\theta \sum_{\substack{k:n_k=1 \\ x_{k0}\text{ distinct} \\ \mathscr{F}\mathbf{x}_k\neq\mathbf{x}_j,\ \forall j}}\sum_{T:C(T)=C_0} p(\mathscr{F}_k T,\mathbf{n}) \\ &+\theta \sum_{\substack{k:n_k=1 \\ x_{k0}\text{ distinct}}}\sum_{j:\mathscr{F}\mathbf{x}_k=\mathbf{x}_j}\sum_{T:C(T)=C_0} p\left(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j)\right), \end{aligned} \tag{3.4.3}$$

The first term in (3.4.2) follows immediately from the first term in (3.4.3) given (3.4.1).

The second term follows by realising that the summation over x_{k0} distinct in the rooted trees corresponds to summing over the edges with multiplicity 1 which are leafs in the unrooted tree, i.e., $k : n_k = 1,\ |k| = 1$. The summation over $\mathscr{F}\mathbf{x}_k \neq \mathbf{x}_j, \forall j$ corresponds to the edges connecting k and j, where if you remove a mutation, there are still at least one mutation separating k and j, that is, $k \to j,\ m_{kj} > 1$. Furthermore, the rooted tree $\mathscr{F}_n T$, when summing over $\mathscr{F}\mathbf{x}_k \neq \mathbf{x}_j, \forall j$, corresponds to deleting a mutation on the edge connecting k and j, that is $(\mathbf{Q} - e_{kj}, \mathbf{n})$. The second term in (3.4.3) combined with (3.4.1) then gives the second term in (3.4.2).

The third term follows similar to the second, by realising that $j : \mathscr{F}\mathbf{x}_k = \mathbf{x}_j$ corresponds to the edges connecting j to k, where removing a mutation along an edge would result in zero mutations on that edge, i.e., $k \to j,\ m_{kj} = 1$. Lastly, $(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n} + e_j))$, when summing over $j : \mathscr{F}\mathbf{x}_k = \mathbf{x}_j$, corresponds to deleting a mutation on the edge between j and k, i.e. $(\mathbf{Q} - e_{kj}, \mathbf{n} - e_k + e_j)$.

To see that the system of equations satisfy the boundary conditions, consider first the event $((0), 2e_1)$. This is the event that in a sample of size 2 both sequences are of the same type. This happens if first event was a coalescent event, and using lemma 3.4, this happens with probability $\frac{1}{1+\theta}$. Next consider the case where there are a total of m mutations in the sample. This event happens when the first m evolutionary events (going back in time) is mutation events, followed by a coalescent event. By lemma 3.4 this happen with probability

$$\left(\frac{1}{1+\theta}\right)^m \left(\frac{1}{1+\theta}\right).$$

□

Relationships between $p(Q, \mathbf{n})$, $p^0(Q, \mathbf{n})$ and $p^*(Q, \mathbf{n})$ defined analogous to the probability of the rooted trees holds for the unrooted trees as well, but when constructing likelihood schemes for the infinite-sites-model, the most important observation from this sections is that of (3.4.1), which shows that it suffices to construct methods for rooted trees, and then summing the estimated sample probabilities of the rooted trees to get an estimate of the probability of the unrooted tree, since (3.4.1) also holds by replacing p with p^0 [Griffiths and Tavaré, 1995].

4 Estimation of θ

In this section we describe methods for likelihood inference of θ. In the previous section we saw how to calculate sample probabilities for rooted and unrooted genealogical trees. One could in principle use these recurrence equations to calculate the likelihood for different values of θ and obtain maximum likelihood estimates, but direct calculation of the sample probabilities are not practical for even small yet realistic sample sizes, and for larger, though still realistic, samples it becomes unfeasible. Instead of settling for Watterson's estimator θ_W when encountering samples of vast size, we look to likelihood schemes that overcome the very problem of considering all possible trees from the sample. A way to motivate this is to consider a simple Monte Carlo scheme, where start with a common ancestor in accordance with the state space E and split it into two lines both with the type of the ancestor. From here, evolve the tree according to the model for a given value of θ until there are n individuals in the sample. Continue this procedure until a pre-specified stopping criteria is met, and take the average of the number of runs it takes to produce the particular sample you are interested in. The procedure serves as a way of calculating the likelihood in a single point, and to get an estimate of θ one would have to repeat this procedure for several values of θ. The reason that this is a naive method comes from the fact that in order to produce point estimates for the likelihood, one would have to repeat this method an impossible amount of times (in the range $10^6 - 10^{100}$, as they point out in [Stephens and Donnelly, 2000]).

Though this naive Monte Carlo scheme is indeed not feasible in any situation of interest, it still provides an idea of how to construct a way of calculating the likelihood without the need for considering the different possible trees corresponding to the sample data. One way to overcome the impossibilities of simulating the sample paths needed for the naive estimator is to develop a simulation technique that reduces the amount of simulations needed to produce likelihood estimates. The First method developed in this context is by [Griffiths and Tavaré, 1994b]. Their method utilizes the sample probabilities derived in section 3, and it has since been extended, by them selves and others, to cover different kind of models and inference problems. Since the development in [Griffiths and Tavaré, 1994b], [Stephens and Donnelly, 2000] pro-

vided an importance sampling scheme utilizing results on the theoretical optimal proposal distribution. It is these Monte Carlo schemes that will be the focus of this section.

4.1 The Griffiths-Tavaré scheme

4.1.1 Point estimates

Griffiths and Tavaré developed a technique for likelihood inference using an elementary result about Markov chains given by the following lemma:

Lemma 4.1. *Let $\{X_k,\ k \geq 0\}$ be a Markov chain with state space S and transition matrix P. Let A be a set of states for which the hitting time*

$$\eta = \inf\{k \geq 0 :\ X_k \in A\}$$

is almost surely finite starting from any state $x \in T = S \setminus A$. Let $f \geq 0$ be a function on S, and define

$$u_x(f) = \mathbb{E}_x \prod_{k=0}^{\eta} f(X_k) \tag{4.1.1}$$

for all $X_0 = x \in S$, so that

$$u_x(f) = f(x),\ x \in A.$$

Then for all $x \in T$

$$u_x(f) = f(x) \sum_{y \in S} p_{xy} u_y(f). \tag{4.1.2}$$

Proof.

$$\begin{aligned}
u_x(f) &= \mathbb{E}_x \prod_{k=0}^{\eta} f(X_k) \\
&= \mathbb{E}_x \left[f(X_0) \prod_{k=1}^{\eta} f(X_k) \right] \\
&= f(x) \mathbb{E}_x \left[\mathbb{E}_x \left(\prod_{k=1}^{\eta} f(X_k) | X_1 \right) \right] \\
&= f(x) \mathbb{E}_x \left[\mathbb{E}_{x_1} \left(\prod_{k=1}^{\eta} f(X_k) \right) \right] \\
&= f(x) \mathbb{E} u_{x_1}(f) \\
&= f(x) \sum_{y \in S} p_{xy} u_y(f),
\end{aligned}$$

where in the fourth line we have used the Markov property.

□

This result suggest that if we wants to solve equations like that on the right of (4.1.2), we could simulate the Markov chain in question until it hits the absorbing state A at time η and compute the value $\prod_{k=0}^{\eta} f(X_k)$. Then repeat this for a large number of times and take the average of the results.

To use this in the means of likelihood inference for the infinite-sites-model, we seek to adjust the recursion equations in section 3 so they fit into lemma 4.1 for an appropriate Markov chain. In this way we overcome the difficulties of calculating the probabilities for every tree describing the sample, by simulating trajectories of the appropriate Markov chains instead. Let

$$f(T,\mathbf{n}) = \sum_{k=1}^{d} \frac{n_k - 1}{(n+\theta-1)} + \frac{\theta m}{n(n+\theta-1)}$$

where,

$$m = |\{k : n_k = 1,\ x_{k0} \text{ distinct},\ \mathscr{F}x_k \neq x_j \forall j\}|.$$

Then (3.3.5) can be recast in the form

$$\begin{aligned} p^0(T,\mathbf{n}) = f(T,\mathbf{n}) &\sum_{k:n_k\geq 2} \frac{n_k-1}{(n+\theta-1)f(T,\mathbf{n})} p^0(T,\mathbf{n}-e_k) \\ &+ f(T,\mathbf{n}) \sum_{\substack{k:n_k=1 \\ x_{k0}\text{ distinct} \\ \mathscr{F}\mathbf{x}_k\neq\mathbf{x}_j,\ \forall j}} \frac{\theta}{n(n+\theta-1)f(T,\mathbf{n})} p^0(\mathscr{F}_k T,\mathbf{n}) \\ &+ f(T,\mathbf{n}) \sum_{\substack{k:n_k=1 \\ x_{k0}\text{ distinct}}} \sum_{j:\mathscr{F}\mathbf{x}_k=\mathbf{x}_j} \frac{\theta(n_j+1)}{n(n+\theta-1)f(T,\mathbf{n})} p^0\left(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j)\right). \end{aligned} \tag{4.1.3}$$

The appropriate Markov chain, $\{X_l,\ l = 0,1,...\}$, to consider has a tree state space according to $(T,\mathbf{n})$ for different states of the tree from the sample to the MRCA, and transition probabilities

$$\begin{aligned} &(T,\mathbf{n}) \longrightarrow \\ &(T,\mathbf{n}-e_k),\ \text{with probability } \lambda_k,\ \text{if } n_k \geq 2 \\ &(\mathscr{F}_k T,\mathbf{n}),\ \text{with probability } \mu_k,\ \text{if } n_k = 1,\ x_{k0}\ \textit{distinct},\ \mathscr{F}x_k \neq x_j\ \forall j \\ &(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j)),\ \text{with probability } \rho_{kj},\ \text{if } n_k = 1,\ x_{k0}\ \textit{distinct},\ \exists j : \mathscr{F}x_k = x_j \end{aligned} \tag{4.1.4}$$

where

$$\lambda_k = \frac{n_k-1}{(n+\theta-1)f(T-\mathbf{n})}, \quad \mu_k = \frac{\theta}{n(n+\theta-1)f(T-\mathbf{n})}, \quad \rho_{kj} = \frac{\theta(n_j+1)}{n(n+\theta-1)f(T-\mathbf{n})}.$$

Now $p^0(T, \mathbf{n})$ fits into lemma 4.1 with the Markov chain defined above, and it is then possible to calculate the sample probability by simulating X starting from $X_0 = (T, \mathbf{n})$ until it has two sequences $(x_{10}, ..., x_{1i})$ and $(x_{20}, ..., x_{2j})$ where $x_{1i} = x_{2j}$ representing the tree T^2. This corresponds to the root of the tree, where the corresponding labelled ordered tree has probability

$$\begin{aligned}
p(T^2) &= P(\mathbf{x}_1 \text{ has i mutations, } \mathbf{x}_2 \text{ has j mutations}) \\
&= \int_0^\infty P(\mathbf{x}_1 \text{ has i mutations, } \mathbf{x}_2 \text{ has j mutations} | T_2 = t) e^{-t} dt \\
&= \int_0^\infty \frac{(t\theta)^i}{2^i i!} e^{-t\theta/2} \frac{(t\theta)^j}{2^j j!} e^{-t\theta/2} e^{-t} dt \\
&= \frac{\theta^{i+j}}{2^{i+j} i! j!} \frac{(i+j)!}{(1+\theta)^{i+j+1}} \int_0^\infty \frac{(1+\theta)^{i+j+1}}{(i+j)!} t^{(i+j+1)-1} e^{-(1+\theta)t} dt \\
&= \binom{i+j}{j} \left(\frac{\theta}{2(1+\theta)} \right)^{i+j} \frac{1}{1+\theta}.
\end{aligned}$$

The first equality is due to independence of mutations along separate branches, the second comes from $T_2 \sim exp(1)$, where T_2 is the time to coalescent when there are 2 individuals in the sample, the third comes from the fact that the number of mutations along a branch, conditional on the length of the branch, is Poisson distributed with parameter $\frac{t\theta}{2}$, and in the fifth we recognize the integrand as a gamma density $\Gamma(i + j, 1 + \theta)$. To obtain $p^0(T^2)$ we divide by a multinomial factor, and since there are only two sequences this is given by

$$p^0(T^2) = (2 - \delta_{i+j,0}) p(T^2) = (2 - \delta_{i+j,0}) \binom{i+j}{j} \left(\frac{\theta}{2(1+\theta)} \right)^{i+j} \frac{1}{1+\theta}.$$

Combining this with the derivation of X, lemma 4.1 gives that

$$p^0(T, \mathbf{n}) = \mathbb{E}_{(T,\mathbf{n})} \left[\prod_{l=0}^{\eta-1} f(T(l), \mathbf{n}(l)) \right] p^0(T^2) \tag{4.1.5}$$

where $X(l) = (T(l), \mathbf{n}(l))$ is the tree at time l. Using this formulation, an estimate of $p^0(T, \mathbf{n})$ can be calculated by simulating trajectories of X until $\eta - 1$. Repeat this procedure a given number of times and the average of $\prod_{l=0}^{\eta-1} f(T(l), \mathbf{n(l)})$ is then an unbiased estimator of $p^0(t, \mathbf{n})$. An estimate of $p^*(T, \mathbf{n})$ can then be found by dividing with the value $a(T, \mathbf{n})$.

This technique can be used to calculate likelihood estimates for a given value of

θ. In terms of likelihood inference, this means that one would have to calculate estimates for $p^*(T, \mathbf{n})$ for a grid of θ-values, and examine the curve of the function. In practise this can be quite time consuming, especially if the θ in question is of dimensions higher than 1, and we will extend the ideas described in this section to a more general approach of estimating likelihood surfaces.

4.1.2 Likelihood Surfaces

Observe the following generalization of lemma 4.1:

Lemma 4.2. *Let $\{X_k,\ k \geq 0\}$ be a Markov chain with state space S and transition matrix P. Let A be a set of states for which the hitting time*

$$\eta = \inf\{k \geq 0 :\ X_k \in A\}$$

is almost surely finite starting from any state $x \in T = S \setminus A$. Let $h \geq 0$ be a given function on A, let $f \geq 0$ be a function on $S \times S$ and define

$$u_x(f) = \mathbb{E}_x h(X_\eta) \prod_{k=0}^{\eta-1} f(X_k, X_{k+1}) \tag{4.1.6}$$

for all $X_0 = x \in S$, so that

$$u_x(f) = h(x),\ x \in A.$$

Then for all $x \in T$

$$u_x(f) = \sum_{y \in S} f(x,y) p_{xy} u_y(f). \tag{4.1.7}$$

Proof. The proof follow that of lemma 4.1;

$$\begin{aligned} u_x(f) &= \mathbb{E}_x h(X_\eta) \prod_{k=0}^{\eta-1} f(X_k, X_{k+1}) \\ &= E_x \left[f(x, x_1) \mathbb{E}_{x_1} h(X_\eta) \prod_{k=1}^{\eta-1} f(X_k, X_{k+1}) \right] \\ &= E_x f(x, x_1) u_{x_1}(f) \\ &= \sum_{y \in S} f(x,y) p_{xy} u_y(f) \end{aligned}$$

□

To use this lemma for estimating likelihood surfaces we apply the general ideas of importance sampling. Let $q_\theta(x)$ be the probability of a given sample $x \in \mathcal{J}^*$ under

θ. In this case

$$q_\theta(x) = \sum_{y\in S} f_\theta(x) p_\theta(x,y) q_\theta(y),$$

where $S = \mathcal{J}^*$. Denote

$$\Theta_0 = \{\theta_0 : \ f_\theta(x)p_\theta(x,y) > 0 \Rightarrow p_{\theta_0}(x,y) > 0 \ \forall\theta\}.$$

Then for $\theta_0 \in \Theta_0$

$$q_\theta(x) = \sum_{y\in S} f_\theta(x) \frac{p_\theta(x,y)}{p_{\theta_0}(x,y)} p_{\theta_0}(x,y) q_\theta(y),$$

and if we let

$$f_{\theta_0,\theta}(x,y) = f_\theta(x)\frac{p_\theta(x,y)}{p_{\theta_0}(x,y)}$$

then

$$q_\theta(x) = \sum_{y\in S} f_{\theta_0,\theta}(x,y) p_{\theta_0}(x,y) q_\theta(y).$$

Lemma 4.2 then states that

$$q_\theta(x) = \mathbb{E}_x q_\theta(X_\eta) \prod_{k=0}^{\eta-1} f_{\theta_0,\theta}(X_k, X_{k+1}),$$

but here the distribution of the Markov chain is no longer governed by the transition matrix P_θ but of P_{θ_0} instead, and hence, the expectation should be regarded as the expectation under the θ_0.

This procedure is known from standard importance sampling, and it allows us to calculate the value of $q_\theta(x)$ for all θ by realisations of a single Markov chain governed by the transition matrix P_{θ_0}. Thus, the likelihood surfaces $p^0(T,\mathbf{n};\theta)$ is given by

$$p^0_\theta(T,\mathbf{n}) = p^0_\theta(T^2)\mathbb{E}^{\theta_0}_{(T,\mathbf{n})} \prod_{k=0}^{\eta-1} h\left((T(k),\mathbf{n}(k)),(T(k+1),\mathbf{n}(k+1))\right) \tag{4.1.8}$$

where

$$h((T,\mathbf{n}),(T,\mathbf{n}-e_k)) = f_{\theta_0}(T,\mathbf{n})\frac{n+\theta_0-1}{n+\theta-1}$$

and

$$h((T,\mathbf{n}),(T',\mathbf{n}')) = f_{\theta_0}(T,\mathbf{n})\frac{\theta(n+\theta_0-1)}{\theta_0(n+\theta-1)},$$

where the last expression holds for the two types of transitions given in (4.1.4), $(T,\mathbf{n}) \to (\mathscr{F}_k T,\mathbf{n})$ and $(T,\mathbf{n}) \to (\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j)$. The estimation scheme now goes as follows:

step 1: Simulate a Markov chain staring from $(T, \mathbf{n})$ with transitions as given in (4.1.4) with $\theta = \theta_0$ until the tree has only two lineages left.

step 2: Repeat step 1 a pre-specified number of times.

step 3: For a grid of θ's calculate the value of (4.1.8) to obtain a likelihood surface for θ.

4.2 Importance Sampling

This section explores the methods of likelihood estimation through importance sampling, which is in large part due to [Felsenstein et al., 1999] and [Stephens and Donnelly, 2000], the former describing the Griffiths-Tavaré scheme as an importance sampling trough histories of events in the genealogy - this will be elaborated through the rest of the section, and the latter exploiting this observation by developing an importance sampler through a thorough investigation on a Markov chain constructed on the histories. It is the methods of [Stephens and Donnelly, 2000] that will be the main focus of this section.

The coalescent history $\mathcal{H} = \{H_k : k = 0, ..., -m\}$ is defined as the embedded events in the Markov process describing the coalescent and mutation events. $\mathcal{H}$ thus corresponds to states visited by the Markov process starting from the MRCA at H_{-m} and ending at $H_0 \in E$. In our case E is the state space of the infinite-sites-model, and H_0 corresponds to a tree $(T, \mathbf{n})$. The definition of the histories is not reserved for the infinite-sites-model, and in fact, Stephens and Donnelly derived their importance sampler assuming that the state space E is countable, which is not the case for the infinite-sites-model. Nevertheless, their method is still viable for cases of uncountable state spaces, as it has a great general appeal.

The idea is to construct a Markov chain that takes values in $\mathcal{H}$ with the transition probabilities $p_\theta(H_i, H_{i-1})$, where θ emphasize the dependence of the mutation rate in the transition probabilities. Then for a random sample $(T, \mathbf{n})$ the likelihood is given by

$$L(\theta) = P_\theta(T, \mathbf{n}) = \int P_\theta((T, \mathbf{n})|\mathcal{H}) P_\theta(\mathcal{H}) d\mathcal{H} \tag{4.2.1}$$

where

$$P_\theta((T,\mathbf{n})|\mathcal{H}) = P_\theta((T,\mathbf{n})|H_0) = \begin{cases} C_d & \text{if } (T,\mathbf{n}) \text{ is consistent with } H_0 \\ 0 & \text{otherwise} \end{cases}$$

where C_d is a combinatorial factor securing the ordering of $(T,\mathbf{n})$ if there are d distinct sequences in the sample, but as we shall see later, a clever choice of proposal distribution ensures that $P_\theta((T,\mathbf{n})|\mathcal{H}) = 1$ in every case. (4.2.1) suggest the naive estimator

$$L(\theta) \approx \frac{1}{R}\sum_{i=1}^{R} P_\theta((T,\mathbf{n})|\mathcal{H}_i),$$

as mentioned in the begging of the section, where R is the number of independent runs from the Markov chain governed by $p_\theta(H_i|H_{i-1})$. This becomes infeasible for most sample sizes since $P_\theta((T,\mathbf{n})|\mathcal{H})$ will be 0 for most i's. Instead we resort to importance sampling by choosing a proposal distribution $Q_\theta(\cdot)$ which has support $\{\mathcal{H} : P_\theta((T,\mathbf{n})|\mathcal{H}) > 0\}$. Then

$$\begin{aligned} L(\theta) &= \int P_\theta((T,\mathbf{n})|\mathcal{H})\frac{P_\theta(\mathcal{H})}{Q_\theta(\mathcal{H})}Q_\theta(\mathcal{H})d\mathcal{H} = \mathbb{E}_{Q_\theta(\mathcal{H})}P_\theta((T,\mathbf{n})|\mathcal{H})\frac{P_\theta(\mathcal{H})}{Q_\theta(\mathcal{H})} \\ &\approx \frac{1}{R}\sum_{i=1}^{R} P_\theta((T,\mathbf{n})|\mathcal{H}_i)\frac{P_\theta(\mathcal{H}_i)}{Q_\theta(\mathcal{H}_i)} = \frac{1}{R}\sum_{i=1}^{R} w_i, \end{aligned} \tag{4.2.2}$$

where $\mathcal{H}_i,\ i = 1,...,R$ is independent runs from $Q_\theta(\cdot)$, w_i is known as importance weights and $Q_\theta(\cdot)$ is the proposal distribution. This way of sampling circumvent the issue of having mostly zeros in the sum, if we can choose the proposal distribution in such a way, that $(T,\mathbf{n})$ is consistent with H_0 for most, or all, the runs of $\mathcal{H}$. The problem then, is how to choose an efficient proposal distribution, in the sense that it lowers the variance of the likelihood estimate. The optimal proposal distribution is given by $Q^*_\theta(\mathcal{H}) = P_\theta(\mathcal{H}|(T,\mathbf{n}))$ because

$$\begin{aligned} P_\theta((T,\mathbf{n})|\mathcal{H})\frac{P_\theta(\mathcal{H})}{Q^*_\theta(\mathcal{H})} &= P_\theta((T,\mathbf{n})|\mathcal{H})\frac{P_\theta(\mathcal{H})}{P_\theta(\mathcal{H}|(T,\mathbf{n}))} \\ &= \frac{P_\theta(\mathcal{H}|(T,\mathbf{n}))P_\theta(T,\mathbf{n})}{P_\theta(\mathcal{H})}\frac{P_\theta(\mathcal{H})}{P_\theta(\mathcal{H}|(T,\mathbf{n}))} = P_\theta(T,\mathbf{n}), \end{aligned}$$

which makes every term in the sum in (4.2.2) the same resulting in 0 variance in the likelihood estimate. Unfortunately $P_\theta(\mathcal{H}|(T,\mathbf{n})$ is generally unknown, and we have to resort to other ways of finding efficient proposal distributions. In [Stephens and Donnelly, 2000] they look at a class of proposal distribution by considering Markov chains that arises

by construction of histories $\{H_{-m}, ..., H_0\}$ backwards in time. The backwards transition probabilities is then characterized by $q_\theta(H_{i-1}|H_i)$ and they have support

$$\{H_{i-1} : p_\theta(H_i|H_{i-1}) > 0\}$$

for every i. It is then possible to sample from the proposal distribution by letting $H_0 = (T, \mathbf{n})$ and simulating the chain according to $q_\theta(H_{i-1}|H_i)$ until the MRCA is reached at time $-m$. By defining H_0 as the sample we secure that $\mathcal{H}$ is consistent with $(T, \mathbf{n})$ for every run of the chain and that the ordering of the sample is equal to that of H_0, that is when estimating $p^0(T, \mathbf{n})$ where the ordering is already accounted for, resulting in $P_\theta((T, \mathbf{n})|\mathcal{H}) = 1$ for every $\mathcal{H}$ sampled from $Q_\theta(\cdot)$. In that way the importance weights of (4.2.2) become

$$w_i = \frac{P_\theta(\mathcal{H}_i)}{Q_\theta(\mathcal{H}_i)}$$

when the proposal distribution is defined as above.

To derive a proposal distribution with the above stated characteristics, Stephens and Donnelly motivated their choice by theorem 1 in their article, which states the form on which the optimal proposal distribution is given. Here the proposal distribution is given by the transition probabilities

$$q^*_\theta(H_{i-1}|H_i) = p_\theta(H_{i-1}|H_i)\frac{p_\theta(H_{i-1})}{p_\theta(H_i)}$$

utilizing Bayes's theorem on the transition probabilities $p_\theta(\cdot|\cdot)$. The transition probabilities $p_\theta(\cdot|\cdot)$ are generally not known but Stephens and Donnelly were able to further characterize these probabilities, although still generally unknown. Theorem 1 in their article then inspires estimates of these unknown probabilities, which are then later proved to have certain favourable qualities. Theorem 1 is derived under the infinite-alleles-model, and the characterization of $p_\theta(\cdot)$ is given as the conditional probability of the type of the $(n+1)$'th sampled chromosome given a sample of A_n types. Unfortunately, in the infinite-sites-model the probability of a particular type is not well defined, so the same derivation is not possible. But, what we can take from the above considerations is the way of utilizing Bayes's theorem on the forward transition probabilities to construct backwards transition probabilities. Although $p_\theta(\cdot)$ is

unknown we shall see how to manipulate sample probabilities in order to obtain an estimate of $p_\theta(\cdot)$.

4.2.1 Construction of Q_θ

For the infinite-sites-model take a sample $(T, \mathbf{n}) \in \mathcal{J}^*$. In order to follow the strategy outlined above, consider the history of the sample $\mathcal{H} = \{H_i : \ i = -m, ..., 0\}$ given by the states that the model visits from the MRCA at time $-m$ to a sample at time H_0. The first objective is to define the Markov chain $P_\theta(\mathcal{H})$ according to the history. Recall from corollary 3.5 the sample probability of the unordered tree. The corollary is a recursion on the sample probabilities of the states that the model visits before reaching the MRCA, along with their respective forward transition probabilities, the derivation of which was given in the proof of theorem 3.3. These states correspond to the histories $H_{-m}, ..., H_0$, and thus, the transition probabilities for the Markov chain on $\mathcal{H}$ is given by corollary 3.5:

$$p_\theta(H_i|H_{i-1}) = \begin{cases} \frac{n_k - 1}{n-1+\theta} & H_{i-1} = (T, \mathbf{n} - e_k) \\ \frac{\theta}{n(n-1+\theta)} & H_{i-1} = (\mathscr{F}_k T, \mathbf{n}) \\ \frac{\theta(n_j+1)}{n(n-1+\theta)} & H_{i-1} = (\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n} + e_j)) \end{cases} . \tag{4.2.3}$$

The considerations in the subsection above then suggest choosing the transition probabilities for the proposal distribution by using Bayes's theorem on (4.2.3), but first we need to deal with the unknown probabilities $p_\theta(\cdot)$ arising when using Bayes. To do this, observe an analogous trick to that of (4.1.3) given in [De Iorio and Griffiths, 2004]: Let

$$n^\circ = \sum_{k:n_k \geq 2} n_k + \sum_{\substack{k:n_k=1 \\ x_{k0} \text{ distinct} \\ \mathscr{F}\mathbf{x}_k \neq \mathbf{x}_j, \ \forall j}} 1 + \sum_{\substack{k:n_k=1 \\ x_{k0} \text{ distinct}}} \sum_{j:\mathscr{F}\mathbf{x}_k = \mathbf{x}_j} 1 , \tag{4.2.4}$$

then by corollary 3.5

$$\begin{aligned}
&\frac{1}{n^\circ}\left(\sum_{k:n_k\geq 2} n_k + \sum_{\substack{k:n_k=1\\ x_{k0}\text{ distinct}\\ \mathscr{F}\mathbf{x}_k\neq\mathbf{x}_j,\ \forall j}} 1 + \sum_{\substack{k:n_k=1\\ x_{k0}\text{ distinct}}} \sum_{j:\mathscr{F}\mathbf{x}_k=\mathbf{x}_j} 1\right) p^0(T,\mathbf{n}) \\
&= \sum_{k:n_k\geq 2} \frac{n_k-1}{n+\theta-1} p^0(T,\mathbf{n}-e_k) \\
&+ \sum_{\substack{k:n_k=1\\ x_{k0}\text{ distinct}\\ \mathscr{F}\mathbf{x}_k\neq\mathbf{x}_j,\ \forall j}} \frac{\theta}{n(n+\theta-1)} p^0(\mathscr{F}_k T,\mathbf{n}) \\
&+ \sum_{\substack{k:n_k=1\\ x_{k0}\text{ distinct}}} \sum_{j:\mathscr{F}\mathbf{x}_k=\mathbf{x}_j} \frac{\theta(n_j+1)}{n(n+\theta-1)} p^0\left(\mathscr{R}_k T,\mathscr{R}_k(\mathbf{n}+e_j)\right).
\end{aligned} \tag{4.2.5}$$

In order to use Bayes on $p_\theta(H_i|H_{i-1})$ to obtain a proposal distribution equate, for each k, the term on the left hand side with the corresponding term on the right hand side. Let $H_i = (T,\mathbf{n})$ and consider the three cases

$$(i)\ H_{i-1} = (T,\mathbf{n}-e_k),$$
$$(ii)\ H_{i-1} = (\mathscr{F}_k T,\mathbf{n}),$$
$$(iii)\ H_{i-1} = (\mathscr{R}_k T,\mathscr{R}_k(\mathbf{n}+e_j)).$$

For (i) $H_{i-1} = (T,\mathbf{n}-e_k)$:

$$\frac{n_k}{n^\circ}\hat{p}(T,\mathbf{n}) = \frac{n_k-1}{n+\theta-1}\hat{p}(T,\mathbf{n}-e_k)$$

from which

$$\frac{\hat{p}(T,\mathbf{n}-e_k)}{\hat{p}(T,\mathbf{n})} = \frac{n_k}{n^\circ}\frac{n+\theta-1}{n_k-1}.$$

This gives the proposal transition probability

$$q_\theta(H_{i-1}|H_i) = p_\theta(H_{i-1}|H_i)\frac{\hat{p}_\theta(H_{i-1})}{\hat{p}_\theta(H_i)} = \frac{n_k-1}{n+\theta-1}\frac{n_k}{n^\circ}\frac{n+\theta-1}{n_k-1} = \frac{n_k}{n^\circ}.$$

For (ii) $(\mathscr{F}_k T,\mathbf{n})$:

$$\frac{1}{n^\circ}\hat{p}(T,\mathbf{n}) = \frac{\theta}{n(n+\theta-1)}\hat{p}(\mathscr{F}_k T,\mathbf{n}),$$

which implies

$$\frac{\hat{p}(\mathscr{F}_k T,\mathbf{n})}{\hat{p}(T,\mathbf{n})} = \frac{1}{n^\circ}\frac{n(n+\theta-1)}{\theta}.$$

This gives the proposal transition probability

$$q_\theta(H_{i-1}|H_i) = p_\theta(H_{i-1}|H_i)\frac{\hat{p}_\theta(H_{i-1})}{\hat{p}_\theta(H_i)} = \frac{\theta}{n(n+\theta-1)}\frac{1}{n^\circ}\frac{n(n+\theta-1)}{\theta} = \frac{1}{n^\circ}.$$

For (iii) $H_{i-1} = (\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j))$:

$$\frac{1}{n^\circ}\hat{p}(T,\mathbf{n}) = \frac{\theta(n_j+1)}{n(n+\theta-1)}\hat{p}(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j)),$$

which implies

$$\frac{\hat{p}(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j))}{\hat{p}(T,\mathbf{n})} = \frac{1}{n^\circ}\frac{n(n+\theta-1)}{\theta(n_j+1)}.$$

This gives the proposal transition probability

$$q_\theta(H_{i-1}|H_i) = p_\theta(H_{i-1}|H_i)\frac{\hat{p}_\theta(H_{i-1})}{\hat{p}_\theta(H_i)} = \frac{\theta(n_j+1)}{n(n+\theta-1)}\frac{1}{n^\circ}\frac{n(n+\theta-1)}{\theta(n_j+1)} = \frac{1}{n^\circ}.$$

We are now ready to state the proposal distribution $Q_\theta(\mathcal{H})$ through its backwards transition probabilities. For $H_i = (T,\mathbf{n})$:

$$q_\theta(H_{i-1}|H_i) = \begin{cases} \frac{n_k}{n^\circ} & \text{if } H_{i-1} = (T, \mathbf{n}-e_k) \\ \frac{1}{n^\circ} & \text{if } H_{i-1} = (\mathscr{F}_k T, \mathbf{n}) \\ \frac{1}{n^\circ} & \text{if } H_{i-1} = (\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j)) \end{cases}, \tag{4.2.6}$$

which is the same as the proposal distribution given in [Stephens and Donnelly, 2000] for the infinite-sites-model. This allows us to calculate the importance weights given in (4.2.2): For the i'th run of Markov chain $\{H_0,...,H_{-m}\}$ simulated from Q_θ we have

$$w_i = \frac{P_\theta(\mathcal{H})}{Q_\theta(\mathcal{H})} = \frac{p_\theta(H_0|H_{-1})\cdots p_\theta(H_{-(m-1)}|H_{-m})}{q_\theta(H_{-1}|H_0)\cdots q_\theta(H_{-m}|H_{-(m-1)})}p_\theta(H_{-m}) = p_\theta(H_{-m})\prod_{i=0}^{m}\frac{p_\theta(H_i|H_{i-1})}{q_\theta(H_{i-1}|H_i)}$$

where

$$\frac{p_\theta(H_i|H_{i-1})}{q_\theta(H_{i-1}|H_i)} = \begin{cases} \frac{n^\circ}{n_k}\frac{n_k-1}{n-1+\theta} & H_{i-1} = (T, \mathbf{n}-e_k) \\ \frac{n^\circ}{n}\frac{\theta}{n-1+\theta} & H_{i-1} = (\mathscr{F}_k T, \mathbf{n}) \\ \frac{n^\circ}{n}\frac{\theta}{n-1+\theta} & H_{i-1} = (\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n}+e_j)) \end{cases},$$

since the joint distribution for any Markov chain $\{X_0,...,X_n\}$ is given by

$$P(X_0 = x_0)P(X_1 = x_1|X_0 = x_0)\cdots P(X_n = x_n|X_{n-1} = x_{n-1}).$$

An important observation is that the proposal distribution Q_θ is independent of θ and, as with the Griffiths-Tavaré method, it is then possible to estimate the likelihood surface by simulating runs for a single chain and calculate the average of the

importance weights for different values of θ, but without the need for specifying a driving value θ_0, as was the case in the Griffiths-Tavaré scheme.

4.2.2 New Proposal Distribution

The importance sampler given by [Stephens and Donnelly, 2000] chooses a sequence uniform at random among those applicable in the latest evolutionary event. Once chosen, the type of the sequence uniquely specifies the next evolutionary event going back in time. The method thus regards every sequence with equal probability, only accounting for the type of the sequence. Though, in the infinite-sites-model we have more information in than just the types of sequences - the number of segregating sites and their positions - and here we propose a sampler drawing on the number of mutations for each sequence.

Consider again the trick used to obtain the proposal distribution in the section above. Now, let

$$n^\circ_{new} = \sum_{k:n_k \geq 2} n_k + \sum_{\substack{k:n_k=1 \\ x_{k0} \text{ distinct} \\ \mathscr{F}\mathbf{x}_k \neq \mathbf{x}_j,\ \forall j}} \tilde{S}_k + \sum_{\substack{k:n_k=1 \\ x_{k0} \text{ distinct}}} \sum_{j:\mathscr{F}\mathbf{x}_k=\mathbf{x}_j} \tilde{S}_k, \tag{4.2.7}$$

where $\tilde{S}_k$ is the number of segregating sites in the k'th sequence. Then by calculations analogous to the ones used in deriving the earlier stated proposal distribution, we define the new proposal distribution

$$q_\theta(H_{i-1}|H_i) = \begin{cases} \frac{n_k}{n^\circ_{new}} & \text{if } H_{i-1} = (T, \mathbf{n} - e_k) \\ \frac{\tilde{S}_k}{n^\circ_{new}} & \text{if } H_{i-1} = (\mathscr{F}_k T, \mathbf{n}) \\ \frac{\tilde{S}_k}{n^\circ_{new}} & \text{if } H_{i-1} = (\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n} + e_j)) \end{cases}. \tag{4.2.8}$$

with importance weights

$$\frac{p_\theta(H_i|H_{i-1})}{q_\theta(H_{i-1}|H_i)} = \begin{cases} \frac{n^\circ_{new}}{n_k} \frac{n_k - 1}{n-1+\theta} & H_{i-1} = (T, \mathbf{n} - e_k) \\ \frac{n^\circ_{new}}{n\tilde{S}_k} \frac{\theta}{n-1+\theta} & H_{i-1} = (\mathscr{F}_k T, \mathbf{n}) \\ \frac{n^\circ_{new}}{n\tilde{S}_k} \frac{\theta}{n-1+\theta} & H_{i-1} = (\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n} + e_j)) \end{cases}.$$

This distribution weights the probability of obtaining a given tree through mutation by the number of mutations in the sequences applicable to the most recent mutation event. Thus, when a mutation event occurs, the sequences with many mutations will

have a larger probability of being involved in the most recent evolutionary event than those with fewer mutations.

To motivate the new proposal distribution we take a look at the arrival times in a general Possion process.

Lemma 4.3. *Let $N(t)$ be a Poisson process with parameter λ and let $T_1 < T_2 < \cdots$ be the arrival times in the process. Then conditional on $N(t) = n$, $(T_1, ..., T_n)$ is equal in distribution to the order statistic of n uniformly distributed random variables on $[0, t]$.*

Proof. Fix $t_1, ..., t_n$ such that $0 < t_1 < \cdots < t_n < t$, and choose $\epsilon_1, ..., \epsilon_n$ so

$$0 < t_1 < t_1 + \epsilon_1 < t_2 < \cdots < t_n < t_n + \epsilon_n < t.$$

Then

$$\begin{aligned}
&P\left(\bigcap_{i=1}^{n}\{T_i \in (t_i, t_i + \epsilon_i)\}\Big|N(t) = n\right)\\
=&\frac{P\left(\bigcap_{i=1}^{n}\{N(t_{i-1} + \epsilon_{i-1}, t_i) = 0, N(t_i, t_i + \epsilon_i) = 1\} \cap \{N(t_n + \epsilon_n, t) = 0\}\right)}{P(N(t) = n)}\\
=&\frac{e^{-\lambda t}\prod_{i=1}^{n}\lambda\epsilon_i}{(\lambda t)^n e^{-\lambda t}/n!}\\
=&\frac{n!}{t^n}\prod_{i=1}^{n}\epsilon_i.
\end{aligned}$$

Hence,

$$\frac{P\left(\bigcap_{i=1}^{n}\{T_i \in (t_i, t_i + \epsilon_i)\}\Big|N(t) = n\right)}{\prod_{i=1}^{n}\epsilon_i} = \frac{n!}{t^n}.$$

Letting $\epsilon_i \to 0$ for all $i = 1, ..., n$ gives that the joint distribution of $T_1, ..., T_n$ conditional on $N(t) = n$ is

$$\frac{n!}{t^n}\mathbb{1}_{(0<t_1<\cdots<t_n<t)}$$

which is the distribution for the order statistic of n uniformly distributed random variables. □

Next, given that $N(t) = n$ we look at the distribution of the time to the last event in the process. When T_i, $i = 1, ..., n$ is the arrival times in the process, lemma 4.3 gives that the time to the last event in the process is distributed as $U_{(n)}$, describing the n'th order statistic of $U_1, ..., U_n$, where $U_i \sim \text{unif}[0, t]$. Hence

$$P(T_n \le s \mid N(t) = n) = P(U_{(n)} \le s) = P(\max\{U_1, ..., U_n\} \le s) = \left(\frac{s}{t}\right)^n$$

and for a given $\epsilon > 0$

$$P(|t - T_n| > \epsilon \mid N(t) = n) = P(t - T_n > \epsilon \mid N(t) = n)$$
$$= P(T_n \leq t - \epsilon \mid N(t) = n) = \left(\frac{t-\epsilon}{t}\right)^n \longrightarrow 0, \quad n \to \infty,$$

since $T_n < t$ for all n, showing that the time to the last event converges in probability to t as n tends to infinity.

The above considerations shows that the time to the last event in any Poisson process shrinks as the number of past events increases. Applying this thought to the mutation process describing the infinite-sites-model, it gives reason to think that, when choosing between two sequences to be involved in the most recent mutational event, the sequence with the highest number of segregating sites is chosen with larger probability. Of course, the considerations above does not apply in every case of the infinitely-sites-model, since two sequences do not follow two independent Poisson process, because they might share mutations up to the common ancestor for the entire genealogy. But, they have evolved independently since the MRCA of the two sequences in question, and by applying lemma 4.3 to the sequences from their "own" common ancestor, i.e. an ancestor not necessarily common to the entire sample, suggest choosing the sequence with the most mutations. Difficulties then arise when comparing more than two sequences. Some sequences may share mutations and some may not. Thus, when comparing multiple sequences, one should account for the structure of pairwise differences between the sequences, in order to achieve results like lemma4.3.

Nevertheless, we apply the new intuitive proposal distribution to see how it compares to the one given in [Stephens and Donnelly, 2000]. We note, as with the SD-proposal, the new proposal distribution does not depend on θ and therefore avoids the need of specifying a driving value θ_0.

5 Performance

To test the two proposal distributions, found in the earlier section, we use an example data set found in table 3. In [Griffiths and Tavaré, 1995] the data is listed together with the rooted tree given in figure 6, corresponding to placing the root in sequence 3. For this particular root the true probabilities $p^0(T, \mathbf{n})$ for different θ-values are computed, which makes it attractive for comparison of our importance samplers. In [Stephens and Donnelly, 2000] a comparison is made of their importance sampler and the Griffiths-Tavaré method. They show that their sampler decreases the variance substantially compared to the one given by Griffiths and Tavaré and therefore we will not make a similar comparison. Instead we choose only to compare the SD and new proposal distributions.

Sequences								Frequencies
a	0	0	1	0	0	0	1	3
b	0	0	0	0	0	0	1	4
c	0	0	0	0	0	0	0	4
d	1	0	0	1	0	0	0	11
e	1	0	0	0	0	0	0	1
f	0	1	0	0	0	0	0	2
g	0	0	0	0	1	0	1	2
h	0	0	0	0	1	1	1	3

Table 3: Example data from [Griffiths and Tavaré, 1995]

Table 4 provides the true probability $p^*(T, \mathbf{n})$ for different θ-values for the rooted tree presented in figure 6. The true probabilities are given by [Griffiths and Tavaré, 1995]. In this case, $p^*(T, \mathbf{n})$ happens to coincide with $p^0(T, \mathbf{n})$ because $a(T, \mathbf{n}) = 1$, which enables us to calculate estimates of the probability of the tree from a direct implementation of our importance samplers. Along with the true probabilities, the table also provides estimates of the probability for different θ-values obtained by 30,000 runs from the SD and new proposal distribution respectively. For each estimate the standard error is given in the parentheses.

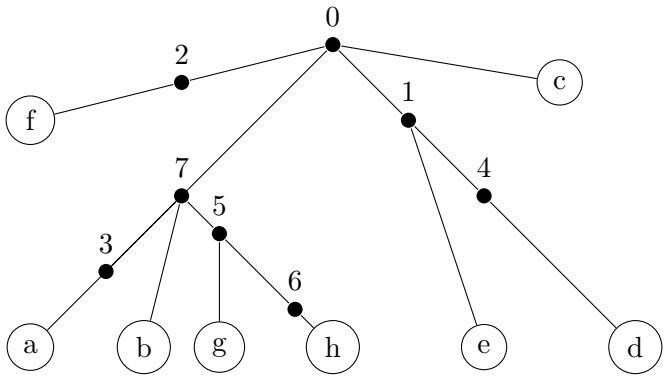

Figure 6: Rooted tree corresponding to table 3.

We see that both samplers provide decent estimates of the probability and that 30,000 runs indeed is sufficient for both samplers to provide accurate estiamtes. It was to be expected that the SD-proposal would perform well given the different performance tests given in [Stephens and Donnelly, 2000], but we see that the new proposal distribution provides estimates that are closer to the true probability with standard errors approximately half of those for the SD-proposal.

θ	Scale	True probability	Q^{SD}	Q^{new}
1.0	10^{-12}	4.77	4.85 (0.17)	4.72 (0.11)
2.0	10^{-11}	1.27	1.30 (0.51)	1.26 (0.32)
4.0	10^{-12}	3.33	3.46 (0.16)	3.29 (0.09)
6.0	10^{-13}	3.85	4.03 (0.21)	3.80 (0.11)
8.0	10^{-14}	4.12	4.35 (0.24)	4.05 (0.13)
10.0	10^{-15}	4.75	5.05 (0.30)	4.66 (0.15)

Table 4: Exact tree probabilities for the tree in figure 6 provided in [Griffiths and Tavaré, 1995] together with probability estimates given by the SD and new proposal distribution with standard errors in the parentheses.

To further investigate the efficiency of the proposal distributions, we examine figure 7 which was obtained by 30,000 runs from each importance sampler, with likelihood-values computed for a grid of θ = 0.1(0.1)6.0. The plot on the left was

produced by the SD-proposal distribution and the plot one the right was produced by the new proposal distribution. Here it is evident that the new proposal distribution decreases the variance of the estimated likelihood surface compared to the SD-proposal distribution.

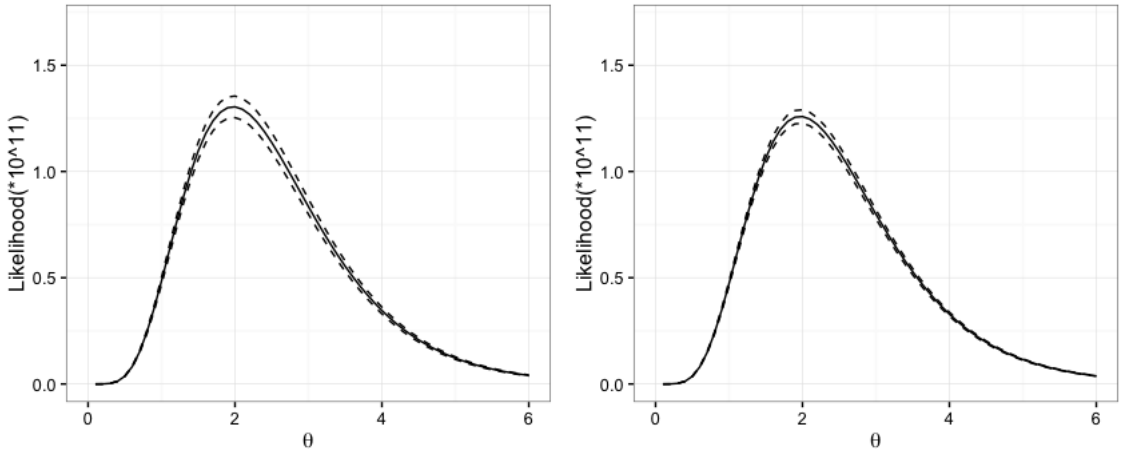

Figure 7: Comparison of likelihood curves (full line) and standard errors (dashed line) for the tree in fiugre 6, based on 30000 samples from the SD (left) and new (right) proposal distributions.

6 Conclusion

When making inference of the mutational parameter in genetic data described by the infinite-sites-model, several different approaches can be taken. In this paper we started out by briefly describing an example of an approach through a summary statistic, which in this case was the Watterson's estimator. The estimator utilizes distributional results about the number of segregating sites in a sample, but it is in general not sufficient, and the need for developing a technique allowing for full likelihood approximation is therefore needed. The first such technique was described by Griffiths and Tavaré which enables approximations of the probability of a sample through a clever scheme, using the recursion formulation of the sample probability to construct a Monte-Carlo simulation method, where simulations are taking from a Markov Chain starting in the sample and evolving back in time, ensuring that every simulation is in accordance with the sample, which permits approximations of the likelihood through simulations. Furthermore, their technique can be extended to facilitate approximations to the entire likelihood surface by simulations of single chain, which has huge computational benefits, as one does not need to run the estimation scheme for a large grid of θ-values. Rather, it is sufficient to simulate the chain for a single value θ_0, and evaluate the mean of a functional of the simulated chain for different values of θ.

The Griffiths-Tavaré scheme inspired the method developed by Stephens and Donnelly in [Stephens and Donnelly, 2000], where they note, that the GT-scheme can be formulated as in importance sampler. They then construct an importance sampler, by choosing a Markov chain starting in the sample as a proposal distribution, ensuring that every simulated chain is in accordance with the data, as was the case in the GT-scheme. Their method was developed under the infinite-alleles-model, and the derivation of their proposal distribution could not be analogously extended to the case of infinite sites, thus resulting in a proposal distribution choosing sequences uniformly at random to involved in the latest evolutionary event. Their method was shown in [Stephens and Donnelly, 2000] to outperform the GT-method, by decreasing the variance of the approximation of the likelihood curve.

By choosing sequences uniformly at random, the SD-sampler does not make full

use of the information contained in the data, and by constructing a proposal distribution analogous to the SD-proposal, by weighting the probabilities by the number of segregating sites in a sequence, we were able to decrease the variance of the sampler and provide more accurate estimates of the true probability. The new proposal distribution was motivated by a basic and intuitive result on Poisson processes, and the sampler itself is even more basic - not derived by any distributional results regarding the number of segregating sites. The new sampler, though simple, was shown to be more efficient than the SD-sampler, which goes to show that there is efficiency to be gained by considering the information actually contained in the data.

References

[De Iorio and Griffiths, 2004] De Iorio, M. and Griffiths, R. C. (2004). Importance sampling on coalescent histories. i. *Advances in Applied Probability*, 36(2):417–433.

[Ethier and Griffiths, 1987] Ethier, S. and Griffiths, R. (1987). The infinitely-many-sites model as a measure-valued diffusion. *The Annals of Probability*, pages 515–545.

[Ewens, 2012] Ewens, W. J. (2012). *Mathematical population genetics 1: theoretical introduction*, volume 27. Springer Science & Business Media.

[Felsenstein et al., 1999] Felsenstein, J., Kuhner, M. K., Yamato, J., and Beerli, P. (1999). Likelihoods on coalescents: a monte carlo sampling approach to inferring parameters from population samples of molecular data. *Lecture Notes-Monograph Series*, pages 163–185.

[Griffiths and Tavaré, 1995] Griffiths, R. and Tavaré, S. (1995). Unrooted genealogical tree probabilities in the infinitely-many-sites model. *Mathematical biosciences*, 127(1):77–98.

[Griffiths and Tavaré, 1994a] Griffiths, R. C. and Tavaré, S. (1994a). Ancestral inference in population genetics. *Statistical science*, pages 307–319.

[Griffiths and Tavaré, 1994b] Griffiths, R. C. and Tavaré, S. (1994b). Simulating probability distributions in the coalescent. *Theoretical Population Biology*, 46(2):131–159.

[Gusfield, 1991] Gusfield, D. (1991). Efficient algorithms for inferring evolutionary trees. *Networks*, 21(1):19–28.

[Kingman, 1982a] Kingman, J. F. (1982a). On the genealogy of large populations. *Journal of applied probability*, 19(A):27–43.

[Kingman, 1982b] Kingman, J. F. C. (1982b). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.

[RoyChoudhury and Wakeley, 2010] RoyChoudhury, A. and Wakeley, J. (2010). Sufficiency of the number of segregating sites in the limit under finite-sites mutation. *Theoretical population biology*, 78(2):118–122.

[Stephens and Donnelly, 2000] Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):605–635.

[Ward et al., 1991] Ward, R. H., Frazier, B. L., Dew-Jager, K., and Pääbo, S. (1991). Extensive mitochondrial diversity within a single amerindian tribe. *Proceedings of the National Academy of Sciences*, 88(19):8720–8724.

[Watterson, 1975] Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology*, 7(2):256–276.