Inference in Population Genetics Using Forward and Backward, Discrete and Continuous Time Processes

Juraj Bergman, Dominik Schrempf, Carolin Kosiol, Claus Vogl

PII: S0022-5193(17)30547-7
DOI: 10.1016/j.jtbi.2017.12.008
Reference: YJTBI 9290

To appear in: *Journal of Theoretical Biology*

Received date: 1 June 2017
Revised date: 23 November 2017
Accepted date: 8 December 2017

Please cite this article as: Juraj Bergman, Dominik Schrempf, Carolin Kosiol, Claus Vogl, Inference in Population Genetics Using Forward and Backward, Discrete and Continuous Time Processes, *Journal of Theoretical Biology* (2017), doi: 10.1016/j.jtbi.2017.12.008

Highlights

- Inference of population genetic parameters from a sample of sequences represented as site frequency spectra (SFS), using concepts akin to the forward-backward algorithm of hidden Markov models is described.
- Discrete transition matrices and continuous diffusion models of iterating the population allelic proportion, forward and backward in time, are used for calculating the marginal likelihood of the data for maximum likelihood inference of parameters.
- The method is demonstrated for simulated joint site frequency spectra (i.e., data from two or more populations) under different models of mutation and for different demographic scenarios.

Inference in Population Genetics Using Forward and Backward, Discrete and Continuous Time Processes

Juraj Bergman[a,b], Dominik Schrempf[a,b], Carolin Kosiol[a,d], Claus Vogl[c,*]

[a] *Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria*
[b] *Vienna Graduate School of Population Genetics, A-1210 Wien, Austria*
[c] *Institut für Tierzucht und Genetik, Vetmeduni Vienna, Veterinärplatz 1, A-1210 Wien, Austria*
[d] *Centre of Biological Diversity, School of Biology, University of St. Andrews, St Andrews KY16 9TH, UK*

Abstract

A central aim of population genetics is the inference of the evolutionary history of a population. To this end, the underlying process can be represented by a model of the evolution of allele frequencies parametrized by *e.g.*, the population size, mutation rates and selection coefficients. A large class of models use forward-in-time models, such as the discrete Wright-Fisher and Moran models and the continuous forward diffusion, to obtain distributions of population allele frequencies, conditional on an ancestral initial allele frequency distribution. Backward-in-time diffusion processes have been rarely used in the context of parameter inference. Here, we demonstrate how forward and backward diffusion processes can be combined to efficiently calculate the exact joint probability distribution of sample and population allele frequencies at all times in the past, for both discrete and continuous population genetics models. This procedure is analogous to the forward-backward algorithm of hidden Markov models. While the efficiency of discrete models is limited by the population size, for continuous models it suffices to expand the transition density in orthogonal polynomials of the order of the sample size to infer marginal likelihoods of population genetic parameters. Additionally, conditional allele trajectories and marginal likelihoods of samples from single populations or from multiple populations that split in the past can be obtained. The described approaches allow for efficient maximum likelihood inference of population genetic parameters in a wide variety of demographic scenarios.

Keywords: bi-allelic mutation-drift model, Markov chain, forward-backward algorithm, forward-backward diffusion, exact inference.

*Corresponding author
Email addresses: juraj.bergman@vetmeduni.ac.at (Juraj Bergman), dominik.schrempf@vetmeduni.ac.at (Dominik Schrempf), ck202@st-andrews.ac.uk (Carolin Kosiol), claus.vogl@vetmeduni.ac.at (Claus Vogl)

1. Introduction

Most basic population genetic models, *e.g.,* the Wright-Fisher and the Moran models as well as the forward and backward diffusion models, were introduced before molecular sequence data became available [reviewed in 10]. Thus, emphasis was on demonstrating processes over time and on qualitatively explaining observations, rather than on quantitative inference of population genetic forces given molecular data. Much later, coalescent theory [17] has been used both for demonstration of processes as well as for inference given a population sample [13, 38]. The coalescent reconstructs the genealogical history of a particular sample at a particular locus conditional on population genetic forces. However, the aim in statistical population genetics is usually the inference of evolutionary forces or of the evolutionary trajectory of allele proportions of the whole population.

Population genetic parameters have often been inferred from allele frequency data of a single locus sampled at multiple time-points in the past. Due to the short time-spans, mutation can usually be neglected, while selection is important. Bollback *et al.* [4] developed a method based on a forward diffusion model to infer the strength of selection acting on an allele. This method was later extended to additionally infer the age of the selected allele [22]. To calculate the likelihood of the observed trajectory, these methods rely on solving the diffusion equation using a numerical grid approach. On the other hand, Steinrücken *et al.* [31] use a system of orthogonal polynomials, *i.e.,* a spectral representation of the transition density [29], to analytically solve the diffusion equation and model the evolution of allele frequency. Recently, Schraiber *et al.* [27] developed a Bayesian approach that uses Markov chain Monte Carlo (MCMC) integration of allele frequency trajectories to provide estimates of population genetic parameters.

While the above-described methods deal with a single locus with data from multiple time-points, the focus of this study is to infer the demographic history and the population genetic forces acting on a whole population from present-day data. Specifically, we are interested in inference of population genetic parameters, such as the scaled mutation rate or mutation bias given data y from the present, $t = 0$, that consist of an alignment of M (haploid) sequences. Nucleotide data are assumed to be independently and identically drawn from a population across L freely recombining nucleotide sites. The sites are assumed to be neutral, *e.g.,* in short introns, or at least nearly-neutral, *e.g.,* fourfold degenerate sites, such that the data are informative about population demography and mutation processes. Because sites are assumed independent, they can be summarized as a site frequency spectrum (SFS), also called the allele frequency spectrum. The likelihood of the population sample y can be calculated given the present population allele frequency x_0 and a probability model of the sampling process. The distribution of x_0 is in turn given by a population genetic model parametrized to capture mutation or the demographic history. These population

genetic parameters can be inferred by first integrating over x_0 and subsequently maximizing the marginal likelihood of the data y by varying the model parameters; a strategy that may also be viewed as the empirical Bayes method [*e.g.,* 5]. Under the assumption of equilibrium and given a general mutation-drift model, this strategy leads to a beta-binomial likelihood, which can be maximized using an expectation-maximization algorithm [34]. Assuming that mutations are rare and arise only at fixed sites, *i.e.,* a boundary mutation model, it is possible to derive maximum likelihood estimators of the mutation rate and bias as well as the selection coefficient [35]. The estimator of the mutation rate in [35] is a variant of the well-know Ewens-Watterson θ [9, 39].

The assumption of equilibrium is often violated in natural populations and, therefore, within this framework, modelling allele frequency trajectories is necessary to accurately infer parameters from the observed SFS. Furthermore, even under equilibrium, maximum likelihood inference requires modelling of allele trajectories with data from two or more populations that split some time in the past, represented by a joint SFS (jSFS). Herein, we mostly focus on inference using the jSFS given the canonical model of two populations that split at some known or unknown time in the past, from which samples of sizes $M^{(1)}$ and $M^{(2)}$ are obtained at the present time. Inference using jSFS has been implemented in the well-known program $\partial a \partial i$ by Gutenkunst *et al.* [12]. It is widely used to infer migration rates, selection coefficients and split times given data from multiple populations using a numerical grid approach to solve the forward diffusion equation and model allele trajectories. An alternative approach was developed in Lukić *et al.* [21] and Lukić and Hey [20], where as in [29, 31], orthogonal polynomials are used to model allele frequency evolution. A similar, but discrete model of allele frequency evolution is presented in Jewett *et al.* [14].

All of these methods model the evolution of the allele frequency forward in time. However, backward models can also be used to model allele frequency trajectories and calculate the likelihood of the data y conditional on the population allele frequency x_t at earlier times ($t < 0$). Based on the Wright-Fisher model, Zhao *et al.* [46] provide an algorithm to calculate probabilities of intermediate states conditional on the starting and end states. This allows simulation of conditional trajectories. Schrempf *et al.* [28] use a Moran model in phylogenetic inference. The "pruning algorithm" [11] allows computation of the likelihood from the tips of a phylogenetic tree down to the root, *i.e.,* backward in time. For efficient inference of phylogenetic trees reversibility of the evolutionary process is generally assumed.

In this article, we demonstrate the usefulness of backward-in-time processes in parameter inference, while considering both discrete population genetics models and continuous diffusion. We also show parallels between discrete and continuous models. Combining the forward and backward processes, as with the forward-backward algorithm of hidden Markov models (HMM) [25], the probability distribution of population allele frequencies conditional on data $\Pr(x_t \mid y, \dots)$ can be inferred at time t in the past and the distribution of conditional trajectories can be simulated. We therefore use forward and backward processes to conveniently calculate probability distributions in time conditional

on a SFS or jSFS from the present. Furthermore, we introduce bi-allelic boundary mutation models, with mutations occurring only at fixed sites. Specifically, we present the solution to the boundary mutation-drift diffusion model, which underlies the infinite site or Poisson-random-fields models [16, 26] and is important in statistical inference in population genetics as a starting point to derive maximum likelihood estimators, such as the well-known Ewens-Watterson estimator of the scaled mutation rate [9, 39]. The Markov chains of the models under consideration have no absorbing states and therefore have stationary distributions. We do not always assume time-reversibility. For the discrete models, the transition matrix must be multiplied repeatedly to obtain the distribution of population allele frequencies forward and backward in time. As the size of the transition matrix depends on the population size N, multiplication becomes cumbersome if N is large. In the limit of large population sizes, the corresponding Kolmogorov forward and backward diffusion equations are obtained. Orthogonal polynomials provide a flexible and fast method to solve the diffusion equations and calculate marginal likelihoods for inference in population genetics. For most purposes, expansion of polynomials up to the order of the sample size M suffices to accurately infer the transition density. With two populations, it can be shown that the order of the expansion is between the minimum and the maximum of the two sample sizes, depending on the starting distribution. As this is usually much less than the population size, continuous diffusion models may be much more efficient for parameter inference in population genetics than equivalent discrete models.

2. Time-homogeneous discrete Markov chains

In this section we apply the forward-backward algorithm [25] to discrete population genetic models for inference given a SFS or a jSFS. To this end, we rephrase iteration using discrete population genetic models (Wright-Fisher or Moran) in the terminology of the forward-backward algorithm [*e.g.,* 25]. We mainly use matrix notation to emphasize the similarities between discrete iteration and the continuous models in Sections 3 and 7.1. For completeness and clarity, subsections include reviews of standard theory.

2.1. Assumptions

(i) Assume a haploid population of size N and a bi-allelic mutation model. The time-dependent frequency of allele one in the population at time t is denoted x_t ($0 \leq x_t \leq N$) and is assumed to evolve as a discrete, time-homogeneous Markov chain with a transition probability matrix $\mathbf{T}$, where $(\mathbf{T})_{ij} = \Pr(x_{t+1} = j \,|\, x_t = i)$ with $i, j \in \{0, \ldots, N\}$. $\mathbf{T}$ is an aperiodic, right stochastic matrix.

(ii) At a (possibly unknown) time $t = s$ ($s < 0$) in the past, a distribution of population allele proportions is given by $\boldsymbol{\rho}$ with entries $(\rho_i)_{i\in\{0,\ldots,N\}} = \Pr(x_s = i)$. In particular, $\boldsymbol{\rho}$ may be the stationary distribution $\boldsymbol{\pi} = (\pi_i)_{i\in\{0,\ldots,N\}}$ or may correspond to a joint distribution of some other data and the equilibrium allele frequency distribution.

(iii) The population evolves until the present time $t = 0$, when a sample of size M is drawn. We denote the sampled frequency of allele one as y ($0 \leq y \leq M$). The probability of observing y, *i.e.,* the likelihood, is $\Pr(y \,|\, M, x_0)$ (we may drop the dependency on M in the following) and will be defined according to the application.

For two populations, assumptions (ii) and (iii) are modified:

(ii) At a (possibly unknown) time $t = s$ ($s < 0$) in the past, x_s is drawn from a distribution of population allele proportions $\boldsymbol{\rho}$. The population separates immediately into two populations with the same initial allele frequency x_s.
(iii) The two populations evolve independently until the present time $t = 0$, when samples of sizes $M^{(1)}$ and $M^{(2)}$ are drawn from each population.

For discrete models, iteration is more efficient if the population size N is small. N can be decreased by increasing the mutation rate μ such that their product $\theta = N\mu$ remains constant. For moderate N, the error introduced by such scaling is small and converges to zero in the diffusion limit. Therefore, N can be set according to numerical convenience. Often, our data are from the present and we want to condition on the configuration of allele frequencies at earlier times.

2.2. The forward-backward algorithm

The forward-backward algorithm of hidden Markov models (HMMs) [*e.g.,* 25, 6, 37] is an efficient numerical method for calculating probabilities assuming a Markovian underlying process, where key variables, the "states", are assumed to be unknown, *i.e.,* "hidden". Intermediate results and the algorithm in general can readily be interpreted probabilistically. The algorithm's numerical efficiency is based on the simple, acyclic conditional dependence structure of the unknown variables, which allows for "dynamic programming". In our case, the possible values of the population allele frequency x_t correspond to the hidden states, while the probability distribution $\Pr(y|x_t = i)$ to the emission probabilities. With the Wright-Fisher or the Moran models, allele frequencies at the next time-point x_{t+1} depend only on the current ones, which conforms to a Markov process. Knowing the sample allele frequencies generally does not completely identify the population allele frequencies at any time-point; the exact state of the underlying variable remains "hidden".

2.3. Forward in time

We introduce the row vector $\mathbf{f}_t$ with entries $(\mathbf{f}_t)_i = \Pr(x_t = i \,|\, \boldsymbol{\rho})$, where $i \in \{0, \ldots, N\}$, and $\mathbf{f}_s = \boldsymbol{\rho}$, *i.e.,* the vector of initial probabilities of states, and define recursively:

$$\mathbf{f}_{t+1} = \mathbf{f}_t \mathbf{T} \quad (s \leq t < 0). \tag{1}$$

Thus, $\mathbf{f}_t$ can be interpreted as the probability of the allele frequency at time t conditional on the ancestral state $\boldsymbol{\rho}$, $\mathbf{f}_t = \Pr(x_t \,|\, \boldsymbol{\rho})$. This corresponds to the forward method in the forward-backward algorithm in the theory of HMMs [*e.g.,*

25, 37]. Let $\mathbf{b}'_0$ be a column vector (the prime $'$ depicts matrix transposition) corresponding to the conditional of the sampling process, such that $(\mathbf{b}_0)_i = \Pr(y \,|\, x_0 = i)$ with $i \in \{0, \ldots, N\}$. The marginal likelihood then is

$$\Pr(y \,|\, \boldsymbol{\rho}) = \boldsymbol{\rho}\mathbf{T}^{|s|}\mathbf{b}'_0. \tag{2}$$

2.4. Backward in time

Using a strategy as with the backward method in the theory of HMM [25, 37], we set

$$\mathbf{b}'_t = \mathbf{T}\mathbf{b}'_{t+1} \quad (s \leq t < 0)\,, \tag{3}$$

which can also be written as

$$(\mathbf{b}_t)_i = \Pr(y \,|\, x_t = i) = \sum_j \Pr(x_{t+1} = j \,|\, x_t = i) \Pr(y \,|\, x_{t+1} = j). \tag{4}$$

From the definition of $\mathbf{b}_t$, it follows that we condition on x_t. The recursion moves the conditioning to ever earlier times. The marginal likelihood (2) may also be obtained as follows:

$$\begin{aligned} \Pr(y \,|\, \boldsymbol{\rho}) &= \boldsymbol{\rho}\left[\mathbf{T}^{|s|}\mathbf{b}'_0\right] \\ &= \boldsymbol{\rho}\mathbf{b}'_s \\ &= \sum_i \rho_i \Pr(y \,|\, x_s = i). \end{aligned} \tag{5}$$

2.5. Constant marginal distribution and adjointness

Considering the sampling probability, we can choose any arbitrary t such that

$$\Pr(y \,|\, \boldsymbol{\rho}) = \mathbf{f}_t\mathbf{b}'_t = \sum_i \Pr(x_t = i \,|\, \boldsymbol{\rho}) \Pr(y \,|\, x_t = i) = \langle \mathbf{f}_t, \mathbf{b}_t \rangle, \tag{6}$$

holds, where $\langle \cdot, \cdot \rangle$ denotes an inner product. It follows that the forward and backward transition matrices, *i.e.,* $\mathbf{T}$ and its transpose $\mathbf{T}'$, are adjoint since

$$\begin{aligned} \Pr(y \,|\, \boldsymbol{\rho}) &= \Pr(y \,|\, \boldsymbol{\rho}) \\ (\mathbf{f}_t\mathbf{T})\mathbf{b}'_{t+1} &= \mathbf{f}_t(\mathbf{T}\mathbf{b}'_{t+1}) \\ \langle \mathbf{f}_t\mathbf{T}, \mathbf{b}_{t+1} \rangle &= \langle \mathbf{f}_t, \mathbf{b}_{t+1}\mathbf{T}' \rangle. \end{aligned} \tag{7}$$

This adjoint relationship allows movement forward and backward in time.

2.6. Joint and conditional distribution

The probability of $x_t = i$ and y conditional on the starting distribution $\boldsymbol{\rho}$ is

$$\Pr(x_t = i, y \,|\, \boldsymbol{\rho}) = (\mathbf{f}_t)_i(\mathbf{b}_t)_i\,. \tag{8}$$

Furthermore, the probability of $x_t = i$ conditional on the data and the starting distribution is

$$\Pr(x_t = i \,|\, y, \boldsymbol{\rho}) = \frac{(\mathbf{f}_t)_i(\mathbf{b}_t)_i}{\mathbf{f}_t\mathbf{b}'_t}\,. \tag{9}$$

This allows calculation of the distribution of population allele frequencies conditional on the data and an initial condition at any time.

2.7. Sampling from conditional trajectories

It is possible to simulate trajectories given the initial distribution $\boldsymbol{\rho}$ at time s and the likelihood at time $t = 0$. Note that Zhao *et al.* [46] provide a similar algorithm based on the Wright-Fisher model to simulate trajectories of population allele proportions conditional on the starting and end states. In contrast, we start with a sample at time $t = s$ from the conditional probabilities (9). Given the state at time $t - 1$ the probability of the state at time t is

$$\Pr(x_t = j \,|\, x_{t-1} = i, y) = \frac{(\mathbf{T})_{ij}(\mathbf{b}_t)_j}{(\mathbf{b}_{t-1})_i}\,, \tag{10}$$

which can be used to obtain a sample trajectory. Although the probability distribution of trajectories depends on $\boldsymbol{\rho}$, the transition at a given time t (10) does not contain $\boldsymbol{\rho}$ since it is a Markov process.

2.8. Left and right eigenvectors, stationary distribution

Let $\boldsymbol{\pi} = (\pi_i)_{i\in\{0,\ldots,N\}}$ be the stationary distribution of $\mathbf{T}$, if it exists. $\boldsymbol{\pi}$ is the left eigenvector associated with the largest eigenvalue (equal to one) [10, p. 87]

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{T}. \tag{11}$$

All entries of $\boldsymbol{\pi}$ are strictly greater than zero because the transition matrix was assumed to be irreducible and $\sum \pi_i = 1$. Thus the entries of $\boldsymbol{\pi}$ can be interpreted as probabilities. Since the rows of $\mathbf{T}$ sum to one, it is obvious that a column vector of all ones $\mathbf{1}'$ is the right eigenvector associated with the unit eigenvalue. In our context, this means that iterating forward in time will converge to a vector proportional to $\boldsymbol{\pi}$ and iterating backward in time to a vector proportional to $\mathbf{1}'$. Thus, every state is equally likely when $s \to -\infty$ and we have no information about the initial distribution of states, because the process has already reached equilibrium.

2.9. Reversibility

Define the diagonal matrix $\boldsymbol{\Pi}$ with the entries π_i on the main diagonal. Since irreducible Markov chains with finite state space have stationary distributions with only strictly positive entries, $\boldsymbol{\Pi}$ is invertible with $\boldsymbol{\Pi}^{-1}$ being a diagonal matrix with entries $1/\pi_i$. Set

$$\mathbf{T}^* = \boldsymbol{\Pi}\mathbf{T}\boldsymbol{\Pi}^{-1}\,. \tag{12}$$

The Markov chain is reversible, if $\mathbf{T}^* = \mathbf{T}'$, because then

$$\begin{aligned} \mathbf{T}' &= \boldsymbol{\Pi}\mathbf{T}\boldsymbol{\Pi}^{-1} \\ \mathbf{T}'\boldsymbol{\Pi} &= \boldsymbol{\Pi}\mathbf{T}\,, \end{aligned} \tag{13}$$

which corresponds to the condition of detailed balance.

If reversibility holds, we can separate $\mathbf{f}_t$ into a product of a time dependent row vector $\mathbf{g}_t$ and the stationary distribution matrix $\mathbf{\Pi}$

$$\mathbf{f}_t = \mathbf{g}_t \mathbf{\Pi}. \tag{14}$$

Under reversibility, we have forward in time

$$\begin{aligned} \mathbf{g}_{t+1}\mathbf{\Pi} &= \mathbf{g}_t \mathbf{\Pi}\mathbf{T} \\ \mathbf{g}_{t+1} &= \mathbf{g}_t \mathbf{\Pi}\mathbf{T}\mathbf{\Pi}^{-1} \\ \mathbf{g}_{t+1} &= \mathbf{g}_t \mathbf{T}' . \end{aligned} \tag{15}$$

We may interpret $\mathbf{g}_t$ as a "projected likelihood" that, when multiplied with the stationary distribution, gives the joint distribution $\mathbf{f}_t$. Note that with the decomposition (14), the likelihood becomes

$$\Pr(y \,|\, \boldsymbol{\rho}) = \mathbf{g}_t \mathbf{\Pi} \mathbf{b}_t' \quad \text{for all } t. \tag{16}$$

The adjoint relationship (7) can be modified analogously, to result in the self-adjoint relationship

$$\begin{aligned} \Pr(y \,|\, \boldsymbol{\rho}) &= \Pr(y \,|\, \boldsymbol{\rho}) \\ (\mathbf{g}_t \mathbf{\Pi}\mathbf{T})\mathbf{b}_{t+1}' &= \mathbf{g}_t (\mathbf{T}' \mathbf{\Pi} \mathbf{b}_{t+1}') \\ \langle \mathbf{g}_t \mathbf{\Pi}\mathbf{T}, \mathbf{b}_{t+1} \rangle &= \langle \mathbf{g}_t, \mathbf{b}_{t+1} \mathbf{\Pi}\mathbf{T} \rangle. \end{aligned} \tag{17}$$

2.10. Example: Conditional probabilities under irreversible mutation

As a particular realization of a discrete process consider a bi-allelic model, where alleles can be labeled either as ancestral (zero) or derived (one). Mutation rates are assumed to be small (at most one mutation is segregating per site) and occur only at the boundary zero. When a derived allele is fixed, it immediately becomes ancestral. This process is a variant of the infinite sites model [16], but differs in that it allows for a stationary distribution at a particular site. Using diffusion theory, Evans *et al.* [8] provide an analysis based on moments of the allele proportions of a similar model with mutations from only one boundary, assuming changing population sizes, *i.e.,* not assuming equilibrium. Zivkovic *et al.* [48] extend the analysis to include selection.

The transition matrix $\mathbf{T}$ is defined as follows. Given a time-homogeneous mutation rate μ, transition probabilities at the boundary zero are

$$\begin{cases} \Pr(x_{t+1} = 0 \,|\, x_t = 0) & = 1 - \mu/(1 - \theta H_{N-1}) \\ \Pr(x_{t+1} = 1 \,|\, x_t = 0) & = \mu/(1 - \theta H_{N-1}), \end{cases} \tag{18}$$

where $\theta = N\mu$ and the harmonic number $H_{N-1} = \sum_{i=1}^{N-1} 1/i$. With this definition, we consider the Moran model where with each time-step (note that with the Moran model N time-steps correspond to one generation with the Wright-Fisher model), one individual sampled at random has one offspring that replaces

one other random individual. Within the polymorphic region, random drift is the only force affecting allele frequencies, such that for $2 \leq i \leq N-2$

$$\begin{cases} \Pr(x_{t+1} = i-1 \,|\, x_t = i) & = \frac{1}{N^2} i(N-i) \\ \Pr(x_{t+1} = i \,|\, x_t = i) & = 1 - \frac{1}{N^2} 2i(N-i) \\ \Pr(x_{t+1} = i+1 \,|\, x_t = i) & = \frac{1}{N^2} i(N-i)\,. \end{cases} \tag{19}$$

For $i = N-1$, drift may lead to fixation of the derived allele, which then becomes the ancestral allele, *i.e.,*

$$\begin{cases} \Pr(x_{t+1} = N-2 \,|\, x_t = N-1) & = \frac{1}{N^2} (N-1) \\ \Pr(x_{t+1} = N-1 \,|\, x_t = N-1) & = 1 - \frac{1}{N^2} 2(N-1) \\ \Pr(x_{t+1} = 0 \,|\, x_t = N-1) & = \frac{1}{N^2} (N-1)\,. \end{cases} \tag{20}$$

The state $i = N$ is never reached and is left out of the state space. The system is not in detailed balance, as probability mass moves from state $i = N-1$ to state $i = 0$, but not in the reverse direction.

The stationary distribution is

$$\boldsymbol{\pi}(x) = \begin{cases} \Pr(x = 0) & = 1 - \theta H_{N-1} \\ \Pr(x = i)_{i \in \{1,\ldots,N-1\}} & = \theta / i\,, \end{cases} \tag{21}$$

as can be ascertained by substitution.

Note that the proportion of polymorphism in equilibrium is θH_{N-1}. This equilibrium proportion corresponds to the Ewens-Watterson estimator θ_W [9, 39], which was derived using the infinite site model [16]. In formula (18), the mutation probability per time-step μ is weighted by the inverse of the probability of being at the boundary $1 - \theta H_{N-1}$, which ensures that the average probability of mutations per time-step is constant, irrespective of N. This in turn assures correspondence to the infinite site model.

Assume a hypergeometric likelihood of y, conditional on N, $x_0 = i$, and the sample size $M \leq N$

$$\Pr(y \,|\, N, x_0 = i, M) = \frac{\binom{i}{y}\binom{N-i}{M-y}}{\binom{N}{M}}\,, \tag{22}$$

where $0 \leq y \leq M$ and $0 \leq i \leq (N-1)$. In equilibrium, the joint distribution is obtained by multiplying the stationary distribution with the likelihood. Summing out the population allele frequency x_0, the marginal distribution is obtained

$$\Pr(y \,|\, M) = \begin{cases} \Pr(y = 0 \,|\, M) & = 1 - \theta H_{M-1} \\ \Pr(y = i \,|\, M)_{i \in \{1,\ldots,M-1\}} & = \theta / i\,. \end{cases} \tag{23}$$

It follows that the expected heterozygosity, *i.e.,* the probability of obtaining one derived allele and one ancestral allele in a sample of size $M = 2$ is θ.

As an example of a demographic scenario (Fig. 1A), consider a population with a stationary allele frequency distribution (21) defined by the ancestral mutation rate μ_a at some time s in the past; *i.e.*, $\boldsymbol{\rho} = \boldsymbol{\pi_a}$. Furthermore, assume an instantaneous increase in the mutation rate μ between generations s and $s+1$. As $\theta = N\mu$, this mimicks an expansion of the population size, without the inconvenience of having to change the dimension of the transition matrix. From then on, the population is out of equilibrium and evolving with a new current mutation rate $\mu_c > \mu_a$. At the present time ($t = 0$), we sample M haplotypes from the population. Assume that the ancestral state of the sampled haplotypes can be determined without error. Thus, a polarized SFS may be constructed. The transition matrix $\mathbf{T}$ and its transpose $\mathbf{T}'$ can be calculated conditional on μ_c. Assume hypergeometric sampling. The conditional probabilities of allelic states $\Pr(x_t \,|\, y, \boldsymbol{\rho})$, for any time $s \leq t \leq 0$, in a site frequency spectrum of size M can then be calculated (Fig. 2).

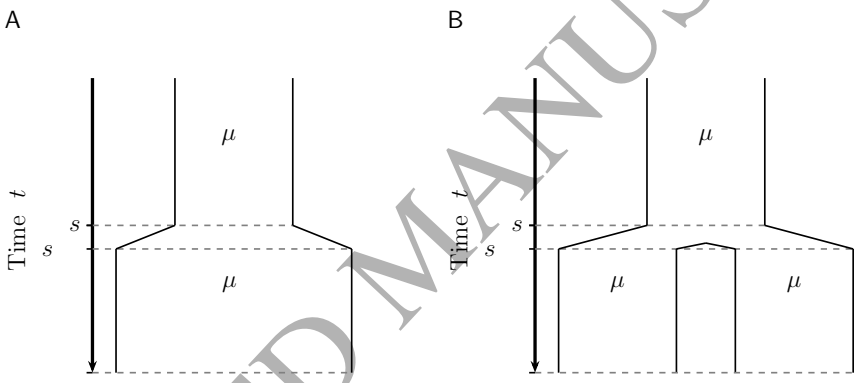

Figure 1: Demographic scenarios. A) Population expansion. B) Population split.

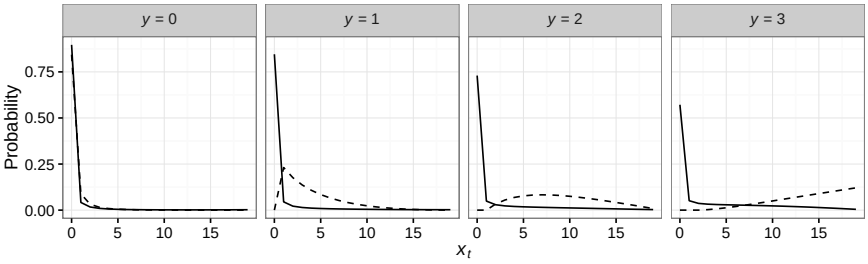

Figure 2: Conditional probabilities of allelic states in a site frequency spectrum of size $M = 3$. The solid lines represent the conditional probabilities of an allelic state x_t given y, at $t = s$, while the dashed lines represent the probabilities at $t = 0$. The parameters were set to $\mu_a = 0.05$, $\mu_c = 0.1$, $s = -200$ and $N = 20$.

2.11. Example: Joint site frequency spectrum under reversible mutation

As another realization of a discrete process consider a bi-allelic mutation-drift decoupled Moran model [2, 7] with haploid population size N, mutation rate towards zero μ_0 and mutation rate towards one μ_1 ($\mu = \mu_0 + \mu_1$). We introduce the parameters $\alpha = \mu_1/\mu$ ($0 \leq \alpha \leq 1$) and $\beta = 1 - \alpha = \mu_0/\mu$ which are the mutation biases towards allele one and zero, respectively. Let i ($0 \leq i \leq N$) be the frequency of allele one. Then, the tri-diagonal transition rate matrix $\mathbf{T}$ depends on N, μ and α

$$\begin{cases} \Pr(x_{t+1} = i-1 \,|\, x_t = i) & = \frac{i(N-i)}{N^2} + \beta\mu\frac{i}{N} \\ \Pr(x_{t+1} = i \,|\, x_t = i) & = 1 - \frac{2i(N-i)}{N^2} + \beta\mu\frac{i}{N} + \alpha\mu\frac{N-i}{N} \\ \Pr(x_{t+1} = i+1 \,|\, x_t = i) & = \frac{i(N-i)}{N^2} + \alpha\mu\frac{N-i}{N}\,. \end{cases} \tag{24}$$

The stationary distribution of x is a beta-binomial

$$\Pr(x = i \,|\, N, \alpha, \theta) = \binom{N}{i} \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} \frac{\Gamma(i + \alpha\theta)\Gamma(N - i + \beta\theta)}{\Gamma(N + \theta)}\,, \tag{25}$$

which can be verified by substitution into the equations of detailed balance (25). As above, hypergeometric sampling at time $t = 0$ is assumed. Assuming equilibrium, the marginal likelihood of a single sample of size M is again a beta-binomial, with M replacing N [34].

Consider an ancestral population with the stationary allele frequency distribution (25). The ancestral population splits into two at some time s in the past (Fig. 1B). For simplicity, no change in the mutation, the drift parameter, and the size in both populations is assumed. A jSFS is simulated from both populations (Table 1) at $t = 0$. The likelihood of the split time s calculated given the simulated jSFS (Figure 3A) has a single maximum close to the true value of $t = -40$.

It may be instructive to calculate some marginal and conditional probabilities with this example. We set for the likelihood of the second population, *i.e.,* the conditional distribution of the data given the allele frequencies in the second population at time $t = 0$, $\mathbf{b}_0^{(2)} = \Pr(y^{(2)} \,|\, x_0^{(2)})$. We then iterate backward within the second population until $t = s$ to obtain the joint probability of the second sample $y^{(2)}$ and the ith allele frequency $x_s = i$ at time $t = s$:

$$\Pr(x_s = i, y^{(2)} \,|\, \boldsymbol{\rho}) = \boldsymbol{\rho}_i (\mathbf{b}_s^{(2)})_i\,. \tag{26}$$

Note that, on the left side of the above equation, we drop the superscript to indicate the population for x_s, because time $t = s$ is just before the split into the two descendant populations. Without information from the second population, we would set the starting distribution of the first population $\mathbf{f}_s^{(1)}$ to the prior probability of the allele frequencies at time $t = s$, *i.e.,* $\mathbf{f}_s^{(1)} = \boldsymbol{\rho}$. With information on the second population, we instead start at time $t = s$ from the joint probability (26) and set $\mathbf{f}_s^{(1)*} = \Pr(x_s, y^{(2)} \,|\, \boldsymbol{\rho})$. As before, we iterate forward to obtain $\mathbf{f}_t^{(1)*}$ within the first population; we can interpret $\mathbf{f}_t^{(1)*}$ as the joint

probability of the allele frequency in the first population and the data of the second population: $\mathbf{f}_t^{(1)*} = \Pr(x_t^{(1)}, y^{(2)} \,|\, \boldsymbol{\rho})$. Setting now for the likelihood of the first population $\mathbf{b}_0^{(1)} = \Pr(y^{(1)} \,|\, x_0^{(1)})$ and iterating backward within the first population until t, we obtain the probability of the allele frequency of the first population at t, conditional on data from both the first and second population as well as on the prior distribution $\boldsymbol{\rho}$ as:

$$\Pr(x_t^{(1)} = i \,|\, y^{(1)}, y^{(2)}, \boldsymbol{\rho}) = \frac{(\mathbf{f}_t^{(1)*})_i (\mathbf{b}_t^{(1)})_i}{\mathbf{f}_t^{(1)*} \mathbf{b}_t^{(1)}} \,. \tag{27}$$

Figure 3B gives the conditional probability $\Pr(x_t \,|\, y^{(1)}, y^{(2)}, \boldsymbol{\rho})$ for one site class of the jSFS determined by $y^{(1)}$ and $y^{(2)}$ which denote the polymorphism levels of the specific class for populations one and two, respectively; *e.g.,* the site class determined by $y^{(1)} = 1$ and $y^{(2)} = 2$ contains all sites with one derived allele in population one and two derived alleles in population two.

Table 1: A jSFS simulated with a discrete Moran model with parameters $L = 10^5$, $M^{(1)} = M^{(2)} = 3$, $\alpha = 2/3$, $\theta = 0.1$, $s = -40$ and $N = 20$.

y	0	1	2	3
0	29037	1315	436	185
1	1276	688	539	432
2	446	529	662	1524
3	202	507	1430	60792

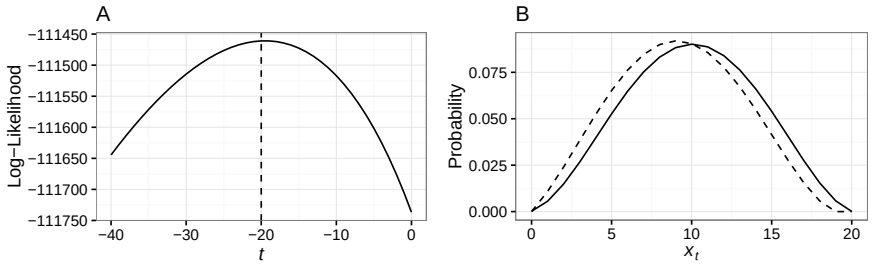

Figure 3: A) The log-likelihood of the split time s, given a jSFS (Table 1). The dashed line indicates the true split time. B) The conditional probability of the allelic state x_t given $y^{(1)} = 1$ and $y^{(2)} = 2$, at $t = s$ (solid line) and $t = 0$ (dashed line).

2.12. Summary: discrete Markov chains

With standard discrete population genetic models, *e.g.,* the Wright-Fisher or the Moran models, iteration of discrete Markov chains forward in time corresponds to the forward algorithm and backward in time to the backward algorithm of the forward-backward algorithm [25]. With such algorithms, it is

straightforward to calculate exact likelihoods given SFS and jSFS from the present. Some standard population genetic mutation models are reversible, others are not. In contrast to phylogenetic applications [11, 28], reversibility of the Markov chain does not simplify calculations considerably; in both cases, iteration of an $(N+1)\times(N+1)$ transition matrix is needed.

3. Forward and backward diffusion equations

In this section, we provide theory for the continuous analogs of the discrete forward and backward transition probabilities both for reversible and irreversible Markov processes and illustrate with examples. We derive the forward and backward diffusion equations from the discrete general mutation-drift Moran model using only the definitions of the first and second symmetric derivative (Appendix 7.1).

With the forward and backward diffusion operators

$$\begin{aligned}\mathcal{L} &= -\frac{\partial}{\partial x}P(x) + \frac{\partial^2}{\partial x^2}Q(x)\\ \mathcal{L}^* &= P(x)\frac{\partial}{\partial x} + Q(x)\frac{\partial^2}{\partial x^2}\,,\end{aligned} \tag{28}$$

the forward and backward diffusion equations are written as

$$\begin{aligned}\frac{\partial}{\partial \tau}\phi(x\,|\,\tau,\rho) &= \mathcal{L}\phi(x\,|\,\tau,\rho)\\ -\frac{\partial}{\partial \tau}\psi(y\,|\,x,\tau) &= \mathcal{L}^*\psi(y\,|\,x,\tau)\,,\end{aligned} \tag{29}$$

where τ is the continuous-time analog of t, and ρ is the initial condition of the countinuous allele frequency x. The functions $\phi(x\,|\,\tau,\rho)$ and $\psi(y\,|\,x,\tau)$ are transition density functions of the forward an backward diffusion, respectively. Obviously, these functions must be twice differentiable in the open interval $(0,1)$. The operators $\mathcal{L}$ and $\mathcal{L}^*$ together with the boundary conditions correspond to the forward transition matrix $\mathbf{T}$ and its transpose $\mathbf{T}'$, respectively.

3.1. Forward and backward in time

As in the discrete case, consider the situation when the distribution of the continuous allelic proportion x at time $\tau = s$ is given by $\rho(x)$. Setting $\phi(x\,|\,\tau = s) = \rho(x)$, $\phi(x\,|\,\tau = 0,\rho)$ can be calculated using the forward diffusion equation (29). Assume again a discrete sample of size M with a frequency of y alleles of type one at time $\tau = 0$. In the backward time direction, $\psi(y\,|\,x,\tau = 0) = \Pr(y\,|\,x,\tau = 0,M)$, which corresponds to a binomial likelihood as the allelic proportion is now assumed to be continuous. Note that a binomial likelihood corresponds to a polynomial of order of the sample size M and is thus finite. With the backward diffusion equation (29), the conditioning on x may

be moved backward in time. The marginal likelihood of y may be obtained by integration over the product of the forward and backward functions

$$\Pr(y \,|\, \rho) = \int_0^1 \phi(x \,|\, \tau, \rho)\psi(y \,|\, x, \tau)\, dx \qquad \text{for } s \le \tau \le 0, \tag{30}$$

analogously to equation (6). As with the discrete case, we require the marginal likelihood to be constant irrespective of time. Furthermore, for any marginal likelihood of a discrete random variable $0 \le \Pr(y \,|\, \rho) \le 1$ must hold. This constrains the boundary conditions.

As $\Pr(y \,|\, \rho)$ is independent of time τ, its derivative with respect to time τ must be 0. Exchanging the order of differentiation and integration and applying the product rule to $\Pr(y \,|\, \rho)$, we have

$$\begin{gathered} \frac{\partial}{\partial \tau} \Pr(y \,|\, \rho) = 0 \\ \int_0^1 \left[\frac{\partial}{\partial \tau}\phi(x \,|\, \tau, \rho)\right] \psi(y \,|\, x, \tau)\, dx + \int_0^1 \phi(x \,|\, \tau, \rho)\left[\frac{\partial}{\partial \tau}\psi(y \,|\, x, \tau)\right] dx = 0\,. \end{gathered} \tag{31}$$

Substituting the right sides of the forward and backward diffusion equations (29) for the time derivatives, we have the adjoint relationship

$$\begin{gathered} \int_0^1 \left[\mathcal{L}\,\phi(x \,|\, \tau, \rho)\right] \psi(y \,|\, \tau)\, dx = \int_0^1 \phi(x \,|\, \tau, \rho)\left[\mathcal{L}^* \psi(y \,|\, x, \tau)\right] dx \\ \langle \mathcal{L}\,\phi(x \,|\, \tau, \rho), \psi(y \,|\, x, \tau)\rangle = \langle \phi(x \,|\, \tau, \rho), \mathcal{L}^* \psi(y \,|\, x, \tau)\rangle. \end{gathered} \tag{32}$$

The adjoint relationship (32) requires the boundary condition (84) to hold (Appendix 7.2). At each time-point, any change to the marginal likelihood from applying the forward operator $\mathcal{L}$ to the forward function $\phi(x \,|\, \tau, \rho)$ is exactly matched by a change from applying the backward operator $\mathcal{L}^*$ to the backward function $\psi(y \,|\, x, \tau)$. As in the discrete case, the adjoint relationship allows movement forward and backward in time.

3.2. Self-Adjointness and Reversibility

In this section, we deal with reversible Markov processes. Introduce the weight or speed function [*e.g.,* 10, 29]

$$w(x) = \frac{1}{Q(x)} e^{\int_0^x \frac{P(z)}{Q(z)}\, dz}\,. \tag{33}$$

Substituting $w(x)g(x, \tau, \rho)$ for $\phi(x \,|\, \tau, \rho)$, the boundary condition (84) becomes (Appendix 7.2)

$$w(x)Q(x)\left(g(x, \tau, \rho)\frac{d}{dx}\psi(y \,|\, x, \tau) - \psi(y \,|\, x, \tau)\frac{d}{dx}g(x, \tau, \rho)\right)\Bigg|_0^1 = 0\,. \tag{34}$$

Since $w(x)Q(x)$ may be infinite at the boundary, $\psi(y \,|\, x, \tau)$ and $g(x, \tau, \rho)$ need to be finite.

Assume $w(x) > 0$ for $x \in]0,1[$, and substitute $w(x)g(x,\tau,\rho)$ for $\phi(x\,|\,\tau,\rho)$ into the general forward equation (29)

$$\begin{aligned}
\frac{\partial}{\partial\tau}w(x)g(x,\tau,\rho) &= -\frac{\partial}{\partial x}P(x)w(x)g(x,\tau,\rho) + \frac{\partial^2}{\partial x^2}Q(x)w(x)g(x,\tau,\rho) \\
w(x)\frac{\partial}{\partial\tau}g(x,\tau,\rho) &= P(x)w(x)\frac{\partial}{\partial x}g(x,\tau,\rho) + Q(x)w(x)\frac{\partial^2}{\partial x^2}g(x,\tau,\rho) \qquad (35) \\
\frac{\partial}{\partial\tau}g(x,\tau,\rho) &= P(x)\frac{\partial}{\partial x}g(x,\tau,\rho) + Q(x)\frac{\partial^2}{\partial x^2}g(x,\tau,\rho)\,.
\end{aligned}$$

Note that the last line is identical to the backward equation (29), with the exception of the reversed sign to the left. Note that, nevertheless, $\phi(x\,|\,\tau,\rho)$ may be infinite. If the stationary distribution $\pi(x)$ exists, it is proportional to $w(x)$. From substituting $\pi(x)g(x,\tau,\rho)$ for $\phi(x\,|\,\tau,\rho)$ into the marginal likelihood (30), it follows that g and ϕ are square integrable with respect to the weight function $\pi(x) \propto w(x)$ [29]. The Markov process is then self-adjoint and reversible and the relationship between the forward operator $\mathcal{L}$ and its adjoint $\mathcal{L}^*$ may be written compactly

$$\mathcal{L}^* = \frac{1}{\pi(x)}\left[\mathcal{L}\pi(x)\right], \qquad (36)$$

similar to the reversed transition matrix (eq. 12) or to the condition of detailed balance (eq. 13) in the discrete case.

3.3. Joint and conditional distributions

The function corresponding to the joint distribution of the allelic proportion x and the sample allele frequency y in the discrete case (8) at time τ ($s \le \tau \le 0$) is

$$j(x,y\,|\,\tau) = \phi(x\,|\,\tau,\rho)\psi(y\,|\,x,\tau)\,. \qquad (37)$$

For the conditional distribution of the allelic proportion x given the sample allele frequency y, corresponding to eq. (9) in the discrete case, $j(x,y\,|\,\tau)$ must be divided by the marginal likelihood (30)

$$p(x\,|\,\tau,\rho,y) = \frac{j(x,y\,|\,\tau)}{\Pr(y\,|\,\rho)}\,. \qquad (38)$$

3.4. General mutation and drift and orthogonal polynomials

The diffusion operators in this section are as in (28), with $P(x) = \theta(\alpha - x)$ and $Q(x) = x(1-x)$. In population genetics, $Q(x)$ is generally half the genetic variance with the bi-allelic Moran model (see also Appendix 7.1). In the context we consider, the backward function $\psi(y\,|\,x,\tau)$ at time $\tau = 0$ is a binomial likelihood, *i.e.,* a polynomial of the degree of the sample size M. Without selection, the backward function remains a polynomial with degree M for $s \le \tau \le 0$.

With the general bi-allelic mutation-drift model, Song and Steinrücken [29] already demonstrated self-adjointness and showed how to use modified Jacobi

polynomials to obtain a solution. For the general mutation-drift model, the weight function $w(x,\alpha,\theta) = x^{\alpha\theta-1}(1-x)^{\beta\theta-1}$ is proportional to the stationary distribution

$$\pi(x) = \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)}\, x^{\alpha\theta-1}(1-x)^{\beta\theta-1}\,. \tag{39}$$

Since $Q(x) = x(1-x)$, the boundary condition (34) holds if, at both boundaries $x = 0$ and $x = 1$, $x(1-x)w(x) = 0$ and $\psi(y\,|\,x,\tau)$ and $g(x,\tau,\rho)$ are finite. Since $x(1-x)w(x) = x^{\alpha\theta}(1-x)^{\beta\theta}$ is zero at both boundaries for the non-degenerate case of $\theta > 0$ and $0 < \alpha < 1$, the boundary condition (34) holds if $\frac{\partial}{\partial x}\big(g(x,\tau,\rho)\psi(y\,|\,x,\tau)\big)$ is finite at the boundaries, which can be assumed for population genetic applications.

The (modified) Jacobi polynomials (compare formula 22.3.2 in Abramowitz and Stegun [1])

$$R_n^{(\alpha,\theta)}(x) = \sum_{l=0}^{n}(-1)^l \frac{\Gamma(n-1+l+\theta)\Gamma(n+\alpha\theta)}{\Gamma(n-1+\theta)\Gamma(l+\alpha\theta)l!(n-l)!}x^l \tag{40}$$

are eigenvectors of the backward operator

$$-\lambda_n R_n^{(\alpha,\theta)}(x) = \mathcal{L}^* R_n^{(\alpha,\theta)}(x)\,, \tag{41}$$

with eigenvalues

$$\lambda_n = n(n+\theta-1)\,. \tag{42}$$

The corresponding eigenfunctions of the forward operator are $w(x)R_n^{(\alpha,\theta)}(x)$ with identical eigenvalues.

Since a binomial distribution with sample size M corresponds to a polynomial of order M, the likelihood can be represented by an expansion with coefficients $c_n(y)$ into the modified Jacobi polynomials up to order M. Note that a change in the effective population size (population demography), or equivalently in the scaled mutation rate from θ_a to θ_c needs to be accommodated with a change in the base from $R_n^{(\alpha,\theta_a)}(x)$ to $R_n^{(\alpha,\theta_c)}(x)$.

The orthogonality relationship of the modified Jacobi polynomials is

$$\int_0^1 R_n^{(\alpha,\theta)}(x)R_m^{(\alpha,\theta)}(x)\,w(x)\,dx = \delta_{n,m}\Delta_n^{(\alpha,\theta)}\,, \tag{43}$$

where $\delta_{n,m}$ is the Kronecker delta, and

$$\Delta_n^{(\alpha,\theta)} = \frac{\Gamma(n+\alpha\theta)\Gamma(n+\beta\theta)}{(2n+\theta-1)\Gamma(n+\theta-1)\Gamma(n+1)}\,. \tag{44}$$

Let $c_n(y)$ be the coefficients of the expansion of the likelihood into the modified Jacobi polynomials, which breaks off at $n = M$. Then the solution to the backward equation can be written as

$$\psi(y\,|\,x,\tau) = \sum_{n=0}^{M} c_n(y)R_n^{(\alpha,\theta)}(x)e^{-\lambda_n\tau}\,, \tag{45}$$

with $\psi(y \mid x, \tau = 0) = \Pr(y \mid M, x)$ corresponding to the likelihood.

Let ρ_n be the coefficients of the expansion of the starting distribution $\rho(x)$ at time $\tau = s$. The solution to the forward equation can then be represented as

$$\phi(x \mid \tau, \rho) = w(x) \sum_{n=0}^{\infty} \rho_n R_n^{(\alpha,\theta)}(x) e^{-\lambda_n (s-\tau)} \,. \tag{46}$$

The orthogonality relationship can be used to simplify the marginal likelihood

$$\begin{aligned} \Pr(y \mid \rho) &= \int_0^1 \phi(x \mid \tau, \rho) \psi(y \mid x, \tau) \, dx \\ &= \int_0^1 \sum_{n=0}^{M} \rho_n c_n(y) w(x) \big[R_n^{(\alpha,\theta)}(x)\big]^2 e^{-\lambda_n \tau} e^{-\lambda_n (s-\tau)} \, dx \\ &= \sum_{n=0}^{M} \rho_n c_n(y) \Delta_n^{(\alpha,\theta)} e^{-\lambda_n s}. \end{aligned} \tag{47}$$

Because of the orthogonality relation (43), the calculation of the marginal likelihood (47) requires an expansion in eigenfunctions up to order M, where M is the minimum of the forward-in-time expansion of $\rho(x)$, say M_f, and the backward-in-time expansion of $\Pr(y|x, \tau = 0)$, say M_b. Therefore, for calculating the joint distribution (37) and thus also the conditional (38), an expansion up to order $M_f \times M_b$ is needed.

3.4.1. Example: two splitting populations and binomial likelihoods

Here, we apply the theory to a model with two populations and binomial likelihoods; *i.e.,* a jSFS analogous to the second example in the discrete case (subsection 2.11). The initial distribution $\rho(x)$ is assumed to be the equilibrium distribution. Only the first eigenfunction is necessary to expand the equilibrium distribution; *i.e.,* $\rho_0 = \frac{1}{\Delta_0^{(\alpha,\theta)}}$ while $\rho_{n\geq 1} = 0$. In equilibrium, the marginal likelihood of a single-population sample of size M assuming mutation-drift equilibrium with parameters α and θ is a beta-binomial, as in the discrete case (25),

$$\begin{aligned} \Pr(y \mid M, \alpha, \theta) &= \int_0^1 \Pr(y \mid M, x) \pi(x, \alpha, \theta) \, dx \\ &= \int_0^1 \binom{M}{y} \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} \, x^{\alpha\theta + y - 1} (1-x)^{\beta\theta + M - y - 1} \, dx \\ &= \binom{M}{y} \frac{\Gamma(\theta)}{\Gamma(\alpha\theta)\Gamma(\beta\theta)} \frac{\Gamma(y + \alpha\theta)\Gamma(M - y + \beta\theta)}{\Gamma(M + \theta)} \,. \end{aligned} \tag{48}$$

It follows from the orthogonality relation that only the first term in the expansion $n = 0$ contributes to the marginal likelihood, *i.e.,* the inner product

$$\begin{aligned}\Pr(y \mid M, \alpha, \theta) &= \int_0^1 c_0(y) R_0^{(\alpha,\theta)}(x) \pi(x, \alpha, \theta)\, dx \\ &= \int_0^1 c_0(y) R_0^{(\alpha,\theta)}(x) \frac{1}{\Delta_0^{(\alpha,\theta)}} R_0^{(\alpha,\theta)}(x) x^{\alpha\theta-1}(1-x)^{\beta\theta-1}\, dx \\ &= c_0(y)\,.\end{aligned} \tag{49}$$

For two populations with sample sizes $M^{(1)}$ and $M^{(2)}$, the respective likelihoods $\Pr(y^{(1)} \mid M^{(1)})$ and $\Pr(y^{(2)} \mid M^{(2)})$ are similarly expanded into the modified Jacobi polynomials with coefficients $c_n(y^{(1)})$ and $c_m(y^{(2)})$. At time τ back in the past, we have

$$\Pr(y^{(1)} \mid x, M^{(1)}, \alpha, \theta, \tau) = \sum_{n=0}^{M^{(1)}} c_n(y^{(1)}) R_n^{(\alpha,\theta)}(x) e^{-\lambda_n \tau} \tag{50}$$

and similarly for the second population. If the two populations join at time $\tau = s$ in the past, when the population is assumed to be in mutation-drift equilibrium, the marginal likelihood is

$$\begin{aligned}\Pr(y^{(1)}, y^{(2)} \mid M^{(1)}, M^{(2)}, \alpha, \theta, \tau = s) &= \sum_{n=0}^{M^{(1)}} \sum_{m=0}^{M^{(2)}} \int_0^1 c_n(y^{(1)}) R_n^{(\alpha,\theta)}(x) e^{-\lambda_n s} \\ &\quad \times c_m(y^{(2)}) R_m^{(\alpha,\theta)}(x) \pi(x, \alpha, \theta) e^{-\lambda_m s}\, dx \\ &= \sum_{n=0}^{M} \int_0^1 c_n(y^{(1)}) c_n(y^{(2)}) \left[R_n^{(\alpha,\theta)}(x)\right]^2 \pi(x, \alpha, \theta) e^{-2\lambda_n s}\, dx \\ &= \sum_{n=0}^{M} \frac{c_n(y^{(1)}) c_n(y^{(2)}) \Delta_n^{(\alpha,\theta)} e^{-2\lambda_n s}}{\Delta_0^{(\alpha,\theta)}}\,,\end{aligned} \tag{51}$$

where $M = \min(M^{(1)}, M^{(2)})$, since higher order terms contribute zero weight to the inner product.

A joint site frequency spectrum is drawn (Table 2) at the present time $\tau = 0$. Given the jSFS, the likelihood of the population split time is readily calculated (Figure 4). The jSFSs in Tables 1 and 2 are similar because scaled mutation rates and biases under which they are simulated are identical; for the discrete model, the population size is set to 20 instead of approaching infinity as in the continuous model, which, together with sampling variation, explains the slight differences.

3.4.2. Summary: bi-allelic general mutation-drift diffusion

Assuming a bi-allelic general mutation-drift model, forward and backward diffusion equations and continuous analogs to the discrete forward and backward

Table 2: A jSFS simulated with a continuous diffusion model with parameters $L = 10^5$, $M^{(1)} = M^{(2)} = 3$, $\alpha = 2/3$, $\theta = 0.1$, and $s = -0.1$.

y	0	1	2	3
0	28877	1447	494	231
1	1448	570	491	557
2	497	516	543	1491
3	253	521	1506	60558

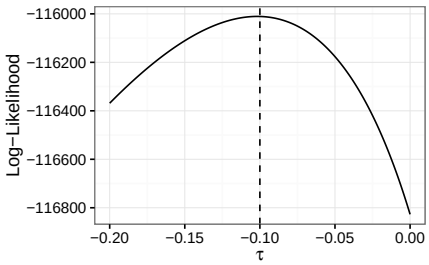

Figure 4: The log-likelihood of the split time s, given a jSFS (Table 2). The dashed line indicates the true split time.

algorithms, as well as the forward-backward algorithm, are derived. As with the discrete models, it is straightforward to calculate exact likelihoods given a SFS or a jSFS from the present. With the bi-allelic general mutation-drift model a self-adjoint system results. Modified Jacobi polynomials $R_n^{(\alpha,\theta)}(x)$ provide a convenient base for calculations, both forward and backward in time. In the discrete case, iteration of an $(N+1) \times (N+1)$ transition matrix is needed to evolve the allelic proportion; in the continuous case, only polynomials up to the sample size M are needed with mutation-drift models. As $M \ll N$, this may lead to considerably increased efficiency. A change in the effective population size (population demography), or equivalently in the scaled mutation rate needs to be accommodated with a change in the base of the orthogonal polynomials as in Steinrücken *et al.* [32].

4. Boundary mutation-drift model

In this section we deal with irreversible Markov processes. If mutation rates are small relative to drift, polymorphism in a sample of moderate size originates from a single mutation. We can therefore assume that mutations originate exclusively from sites fixed for allele zero or one, *i.e.,* from the boundaries. Such models are particularly important for statistical inference in population genetics [*e.g.,* 9, 39, 12] and it is therefore worthwhile to provide solutions to the corresponding diffusion equations. As a solution to the forward and backward diffusion equations we present a system of orthogonal eigenfunctions. Through-

out the presentation, we emphasize the similarities with previous approaches. While the solution to the forward diffusion is mainly a review, the backward direction and the overall concepts are new.

4.1. Pure drift model

We start with the pure drift model and clarify basic concepts. The forward and backward diffusion operators are

$$\begin{aligned} \mathcal{L} &= \frac{\partial^2}{\partial x^2} Q(x) \\ \mathcal{L}^* &= Q(x) \frac{\partial^2}{\partial x^2} \,. \end{aligned} \tag{52}$$

For the pure drift model, the adjoint relationship between the forward and backward operators holds as long as the boundary condition (84) with $Q = x(1-x)$ holds within the unit interval

$$0 = \Big(x(1-x)\phi\psi' - (x(1-x)\phi)'\psi \Big) \Big|_0^1 . \tag{53}$$

Following Kimura [15], most population geneticists implicitly or explicitly require at both boundaries $\psi(y \,|\, x, \tau)$ and $x(1-x)\phi(x \,|\, \tau, \rho)$ to be zero [see also 10, 29]. With these assumptions, modified Gegenbauer polynomials $U_n(x) = -\frac{2}{n} C_{n-2}^{(3/2)}(2x-1)$ ($C_k^\nu(z)$ are the Gegenbauer polynomials as defined in [1]) are eigenfunctions of the forward diffusion equation with eigenvalues $\lambda_n = n(n-1)$ for $n \geq 2$. Furthermore $x(1-x)U_n(x)$ are eigenfunctions of the backward equation with identical eigenvalues. The forward and backward operators are then self-adjoint with the weight function $w(x) = x^{-1}(1-x)^{-1}$ [10, 29]. Note that without mutation no stationary distribution exists. The orthogonality relation of $U_n(x)$ is

$$\int_0^1 U_n(x) U_m(x) w(x) \, dx = \, \delta_{n,m} \Delta_n \,, \tag{54}$$

with

$$\Delta_n = \frac{n-1}{(2n-1)n}. \tag{55}$$

However, these assumptions are too restrictive; polynomials of zeroth and first degree, 1 and x, cannot be represented by $x(1-x)U_n(x)$, but both are eigenfunctions of the pure drift backward equation with eigenvalues $\lambda_0 = \lambda_1 = 0$. Importantly, assuming a binomial likelihood, these eigenfunctions are needed when representing monomorphic samples. To address this issue, Tran *et al.* [33] add 1 and x to the eigenfunctions of the backward equation. The two new backward eigenfunctions require augmenting the forward eigenfunctions with point masses at the boundaries that counterbalance the probability mass in the interior. Additionally, point masses at the boundaries, independent of those associated with the forward eigenfunctions, need to be introduced [33].

Independently from Tran *et al.* [33], we derived a boundary mutation-drift model forward in time from probabilistic population genetic considerations [35] with eigenfunctions proportional to those in Tran *et al.* [33]. Our approach is similar to that presented in McKane and Waxman [23] and Waxman [40]. Furthermore, we showed that the forward eigenfunctions can be derived from those of the general mutation-drift model, *i.e.,* from Jacobi polynomials times the stationary beta distribution (or the proportional weight function $w(x, \alpha, \theta)$), by expanding into a Taylor series in θ and keeping terms up to order zero [36, Appendix A.1]. Therefore, in the context of pure drift, the set of eigenfunctions, which provide the solution to the forward diffusion equation, can then be represented in relation to Jacobi polynomials $R_n^{(\alpha,\theta)}$ as

$$\begin{cases} F_0^{(\alpha,0)}(x) & = \lim_{\theta\to 0} \pi(x,\alpha,\theta) = \beta\delta(x) + \alpha\delta(x-1) \\ F_1^{(\alpha,0)}(x) & = \lim_{\theta\to 0} w(x,\alpha,\theta) R_1^{(\alpha,\theta)} = -\delta(x) + \delta(x-1) \\ F_{n\geq 2}^{(\alpha,0)}(x) & = \lim_{\theta\to 0} w(x,\alpha,\theta) R_n^{(\alpha,\theta)} = -\frac{(-1)^n}{n}\delta(x) + U_n(x) - \frac{1}{n}\delta(x-1)\,, \end{cases} \tag{56}$$

where $\delta(x)$ is the Dirac delta functional. Note that eigenfunctions are only defined up to a proportionality constant. The associated eigenvalues are

$$\begin{cases} \lambda_0 & = 0 \\ \lambda_1 & = \lim_{\theta\to 0} \theta = 0 \\ \lambda_{n\geq 2} & = n(n-1)\,. \end{cases} \tag{57}$$

Similarly, the backward eigenfunctions can be derived by expanding the modified Jacobi polynomials into a Taylor series in θ and keeping terms up to order zero.

$$\begin{cases} B_0^{(\alpha,0)}(x) & = R_0^{(\alpha,\theta)} = 1 \\ B_1^{(\alpha,0)}(x) & = \frac{1}{\theta} R_1^{(\alpha,\theta)} = x - \alpha \\ B_{n\geq 2}^{(\alpha,0)}(x) & = \lim_{\theta\to 0} R_n^{(\alpha,\theta)} = x(1-x)U_n(x)\,. \end{cases} \tag{58}$$

The eigenvalues correspond to those forward in time in eq. (57). The mutation bias α may obtain any value between zero and one. If α is set to zero, the backward eigenfunctions correspond to those of Tran *et al.* [33].

The orthogonality relation is

$$\int_0^1 F_n^{(\alpha,0)}(x) B_m^{(\alpha,0)}(x)\, dx = \delta_{n,m}\Delta_n\,, \tag{59}$$

with $\Delta_0 = \Delta_1 = 1$ and Δ_n as in (55). However, note that

$$\int_0^1 B_n^{(\alpha,0)}(x) B_m^{(\alpha,0)}(x) w(x)\, dx = \delta_{n,m}\Delta_n \tag{60}$$

only holds for pairs $m, n \geq 2$ and the pair $m = 0$ and $n = 1$, but not for the pairs $m = 0$ (or $m = 1$) and $n \geq 2$; and similarly for the forward eigenfunctions $F_n(x)$.

The forward function is then set to

$$\phi(x \mid \tau, \rho) = \sum_{n=0}^{\infty} \rho_n F_n^{(\alpha,0)}(x) e^{-\lambda_n (s-\tau)} \tag{61}$$

and the backward function to

$$\psi(y \mid x, \tau) = \sum_{m=0}^{\infty} c_n(y) B_n^{(\alpha,0)}(x) e^{-\lambda_n \tau} \,. \tag{62}$$

The marginal and joint distribution can now be defined as above. The time derivative of the marginal likelihood (31) of the eigenfunctions with $n = 0$ and $n = 1$ is zero, because the respective eigenvalues are zero. For $n \geq 2$, the backward expansion contains only the terms $x(1-x)U_n(x)$ as does $w(x)$ times the forward expansion, $w(x)F_{n\geq 2}^{(\alpha,0)}(x) = x(1-x)U_{n\geq 2}(x)$. Indeed the eigenfunctions with $n \geq 2$ correspond to those usually considered [15, 29]. As backward and forward functions are thus zero at both boundaries, the boundary condition (53) is met. It is also straightforward to show for $n = 0$ and $n = 1$ that condition (32) holds, because the integrals on both sides are always zero.

4.2. Mutation-drift model

Following Vogl and Bergman [36], we introduce recurrent mutations into the pure drift model by setting the eigenvalue $\lambda_1 = \theta$. We consider the case where $0 < \theta \ll 1$, such that mutations occur at a low rate and thus, do not affect the allele frequency dynamics of the polymorphic classes; these classes are governed exclusively by genetic drift and therefore, eigenfunctions with $n \geq 2$ remain as in the pure drift model. We may thus distinguish between two classes of sites with distinct spatial and temporal differences: the slowly evolving boundaries, where the rate of evolution depends on θ, and the fast evolving polymorphic classes governed by genetic drift [*e.g.,* 42, 36]. Furthermore, we may think of the boundary mutation-drift model as a first order Taylor series expansion in the scaled mutation rate θ of the general mutation-drift model.

Note that, with the discrete boundary mutation model, we scaled the mutation rate such that, independent of the population size N, the heterozygosity in a sample of size two is equal to θ for the model with mutations from a single boundary (compare the term $\mu/(1-\theta\sum_{i=1}^{N-1}\frac{1}{i})$ in (18)), or $2\alpha\beta\theta$ for the model with mutations from both boundaries. With the transition to continuous diffusion, $N \to \infty$ and thus $\theta\sum_{i=1}^{N-1}\frac{1}{i}$ will grow logarithmically without bound. Mutations are therefore modeled from the boundary zero at a rate $\alpha\theta b_0(\tau)$, where $\alpha\theta$ is the mutation rate towards allele one and $b_0(\tau)$ corresponds to the probability mass already at boundary zero plus the probability mass to arrive there quickly by drift, and similarly at the boundary one. The system is thus not in detailed balance and therefore not reversible.

Forward expansion. With mutations from the boundaries and forward in time, Vogl and Bergman [36] use the same augmented forward eigenfunctions as with

pure drift (56) to model the spatial part of the eigensystem. With pure drift, the temporal parts of the eigenfunctions ($e^{-\lambda_n(s-\tau)}$) with $n \geq 2$ fulfill homogeneous differential equations, *i.e.,* are decreasing exponentially from starting values at rates $\lambda_n = n(n-1)$, while the first two eigenfunctions with $n = 0$ and $n = 1$ do not change with time. With the boundary mutation model, the temporal part $T_n(\tau)$ corresponds to a system of linear differential equations: homogeneous for $n = 0$ and $n = 1$ with eigenvalues $\lambda_0 = 0$ and $\lambda_1 = \theta$, and inhomogenous for $n \geq 2$ with eigenvalues $\lambda_n = n(n-1)$:

$$\begin{cases} \frac{d}{d\tau}T_0(\tau) & = 0 \\ \frac{d}{d\tau}T_1(\tau) & = -\theta T_1(\tau) \\ \frac{d}{d\tau}T_{n\geq 2}(\tau) & = -\lambda_n T_n(\tau) + \vartheta E_n T_0(\tau) + \theta O_n T_1(\tau)\,, \end{cases} \tag{63}$$

with

$$\begin{aligned} \vartheta &= \alpha\beta\theta\,, \\ E_n &= -(n-1)\frac{((-1)^n+1)}{\Delta_n}\,, \\ O_n &= -(n-1)\frac{(-1)^n\alpha-\beta}{\Delta_n}\,, \end{aligned} \tag{64}$$

where $\beta = (1-\alpha)$ and Δ_n as in (55).

The forward system can be diagonalized by setting

$$\begin{cases} F_0^{(\alpha,\theta)}(x) & = F_0^{(\alpha,0)}(x) + \vartheta\sum_{n=2}^{\infty}\frac{E_n}{\lambda_n}F_n^{(\alpha,0)}(x) \\ F_1^{(\alpha,\theta)}(x) & = F_1^{(\alpha,0)}(x) + \theta\sum_{n=2}^{\infty}\frac{O_n}{\lambda_n}F_n^{(\alpha,0)}(x) \\ F_{n\geq 2}^{(\alpha,\theta)}(x) & = F_n^{(\alpha,0)}(x)\,, \end{cases} \tag{65}$$

where the polynomials with base $(\alpha, 0)$ on the right hand side of the equations are as in (56). The temporal parts of the system are then $\frac{d}{d\tau}T_n(\tau) = -\lambda_n T_n(\tau)$ for all n.

With increasing N, the stationary distribution converges to the following function [35, 36]

$$\pi(x,\alpha,\theta) = F_0^{(\alpha,\theta)}(x) = \lim_{N\to\infty}\begin{cases} \beta - \vartheta\int_{\frac{1}{N}}^{\frac{N-1}{N}}\frac{1}{x}\,dx & \text{if } 0 \leq x < 1/N \\ \vartheta\frac{1}{x(1-x)} & \text{if } 1/N \leq x \leq 1-1/N \\ \alpha - \vartheta\int_{\frac{1}{N}}^{\frac{N-1}{N}}\frac{1}{1-x}\,dx & \text{if } 1-1/N < x \leq 1\,. \end{cases} \tag{66}$$

This function integrates to unity, but has singularities at the boundaries, which makes it difficult to interpret probabilistically. Moments about zero up to an order $m = M_{\max}$ may be defined meaningfully, by multiplying $\pi(x,\alpha,\theta)$ with

x^m and integrating. We have

$$\begin{aligned}\int_0^1 \pi(x)x^m\,dx &= \alpha - \vartheta\int_0^1 \frac{1-x^{m-1}}{1-x}\,dx \\ &= \alpha - \vartheta H_{m-1}\,,\end{aligned} \tag{67}$$

where H_{m-1} is the harmonic number. As this same relationship must also hold for the moments about boundary one, $\min(\alpha,\beta)/\vartheta < H_{m-1}$, which leads to $M_{\max} \approx e^{\min(\alpha,\beta)/\vartheta}$. Note that a monomorphic sample from a binomial distribution, with sample size M, leads to terms x^M or $(1-x)^M$, which correspond to the moments about zero and one. Thus the sample size needs to be restricted to $M \approx e^{\min(\alpha,\beta)/\vartheta}$ to avoid negative values for probabilities. Since the boundary mutation model generally requires $\theta < 0.1$ [35], this constraint on M should not pose practical problems.

Note that the same issue occurs with the closely related Ewens-Watterson estimator $\hat{\theta}_W$ of molecular diversity [9, 39]. With the assumptions used for deriving $\hat{\theta}_W$, the probability of obtaining a monomorphic sample of size M is $1 - \theta\sum_{i=1}^{M-1}\frac{1}{i}$. It is therefore necessary to restrict the sample size below $M_{\max} \approx e^{1/\theta}$.

Backward expansion. The backward system of differential equations with eigenfunctions $B_n^{(\alpha,\theta)}(x)$ is the transpose of the forward system (65). It can also be diagonalized by setting

$$\begin{cases} B_0^{(\alpha,\theta)}(x) &= B_0^{(\alpha,0)}(x) = 1 \\ B_1^{(\alpha,\theta)}(x) &= B_1^{(\alpha,0)}(x) = x - \alpha \\ B_{n\geq 2}^{(\alpha,\theta)}(x) &= B_n^{(\alpha,0)}(x) - \vartheta\frac{E_n\Delta_n}{\lambda_n}B_0^{(\alpha,0)}(x) - \theta\frac{B_n\Delta_n}{\lambda_n}B_1^{(\alpha,0)}(x)\,. \end{cases} \tag{68}$$

It can be verified that the forward and backward eigenfunctions fulfil the orthogonality relation (59) with $\Delta_0 = \Delta_1 = 1$ and Δ_n as in (55). In particular, for $n = 0$ and $m \geq 2$, we have

$$\begin{aligned}\int_0^1 F_0^{(\alpha,\theta)}(x)B_m^{(\alpha,\theta)}(x)\,dx &= \int_0^1 \left(F_0^{(\alpha,0)}(x) + \vartheta\sum_{n=2}^{\infty}\frac{E_n}{\lambda_n}F_n^{(\alpha,0)}(x)\right) \\ &\quad\times\left(B_m^{(\alpha,0)}(x) - \vartheta\frac{E_m\Delta_m}{\lambda_m}B_0^{(\alpha,0)}(x) - \theta\frac{O_m\Delta_m}{\lambda_m}B_1^{(\alpha,0)}(x)\right)dx \\ &= \vartheta\frac{E_m}{\lambda_m}\Delta_m - \vartheta\frac{E_m\Delta_m}{\lambda_m}\Delta_0 = 0\,,\end{aligned} \tag{69}$$

and similarly for $m = 1$ and $n \geq 2$.

Furthermore, we have, as before, the forward function

$$\phi(x\,|\,\tau,\rho) = \sum_{n=0}^{\infty}\rho_n F_n^{(\alpha,\theta)}(x)T_n(\tau)\,, \tag{70}$$

and the backward function

$$\psi(y \,|\, x, \tau) = \sum_{n=0}^{\infty} c_n(y) B_n^{(\alpha,\theta)}(x) T_n(\tau)\,. \tag{71}$$

The backward function and the marginal distribution, as long as $M < M_{\max} \approx e^{\min(\alpha,\beta)/\vartheta}$, can be interpreted probabilistically as with the general mutation-drift or the pure drift model. As the forward function may attain negative values, expanding it beyond the sample size M has little meaning.

4.2.1. Example: one change in the mutation parameters

We present the version of the boundary mutation model with the inhomogeneous linear differential equations, *i.e.,* with the eigenfunctions $F_0^{(\alpha,0)}$ and $B_n^{(\alpha,0)}$. With this choice, a change in the effective population size (population demography), or equivalently in the scaled mutation rate does not necessitate a change in the base. Assume a population in equilibrium at $\tau = s$ with mutation parameters θ_a and α_a, such that the initial distribution is $\rho(x) = \pi(x \,|\, \theta_a, \alpha_a)$. The scaled mutation parameters then changes immediately to θ and α, respectively, and remain constant thereafter. Expanding the stationary distribution at time $\tau = s$ into the forward eigenfunctions $F_n^{(\alpha,0)}(x)$ results in

$$\begin{aligned}\phi(x \,|\, \tau = s) = F_0^{(\alpha,0)}(x) &+ (\alpha_a - \alpha) e^{-\theta\tau} F_1^{(\alpha,0)}(x) \\ &+ \sum_{n=2}^{\infty} \Big(E_n(\vartheta + (\vartheta_a - \vartheta) e^{-\lambda_n (s-\tau)} \\ &+ (\alpha_a - \alpha)\theta O_n (e^{-\theta(s-\tau)} - e^{-\lambda_n(s-\tau)}) \Big) F_n^{(\alpha,0)}(x)\,.\end{aligned} \tag{72}$$

With a sample of size M with y alleles of the first type at time $\tau = 0$, the binomial likelihood can be expanded into the backward eigenfunctions with

$$\psi(y \,|\, x, \tau = 0) = \sum_{n=0}^{M} c_n(y) B_n^{(\alpha,0)}(x) \tag{73}$$

The marginal likelihood, calculated at time $\tau = 0$, is

$$\begin{aligned}\Pr(y) = \int_0^1 \phi(x \,|\, \tau = 0, \rho) \psi(y \,|\, x, \tau = 0)\, dx &= \Big[c_0(y) \cdot 1 \Big] + \Big[c_1(y)(\alpha_a - \alpha) e^{-\theta s} \cdot 1 \Big] \\ &+ \Big[\sum_{n=2}^{M} c_n(y) \Big(E_n(\vartheta + (\vartheta_a - \vartheta) e^{-\lambda_n s}) \\ &+ (\alpha_a - \alpha)\theta O_n (e^{-\theta s} - e^{-\lambda_n s}) \Big) \cdot \Delta_n \Big]\,,\end{aligned} \tag{74}$$

where the terms in the successive square brackets come from the terms in the expansion with $n = 0$, $n = 1$, and $2 \le n \le M$, respectively, while all terms with

$n > M$ are zero. Within the square brackets, the terms before the dot are the time-dependent functions of the forward expansion. The same marginal likelihood is also obtained by using the backward eigenfunctions $B_n^{(\alpha,0)}$, multiplying with the stationary distribution at $\tau = s$, and integrating:

$$\begin{aligned}\Pr(y) &= \int_0^1 \psi(y \mid x, \tau = s)\pi(x, \alpha_a, \theta_a)\, dx \\ &= \left[\left(c_0(y) + \vartheta \sum_{n=2}^{M} c_n(y) E_n \Delta_n (1 - e^{-n(n-1)s})\right) \cdot 1\right] \\ &\quad + \left[\left(c_1(y) e^{-\theta s} + \theta \sum_{n=2}^{M} c_n(y) E_n \Delta_n (e^{-\theta s} - e^{-\lambda_n s})\right) \cdot (\alpha_a - \alpha) \cdot 1\right] \\ &\quad + \left[\sum_{n=2}^{M} c_n(y) e^{-\lambda_n s} \cdot \vartheta_a E_n \Delta_n\right].\end{aligned} \tag{75}$$

Within the square brackets, the terms before the dot are the time-dependent functions of the backward expansion. The two different versions of the marginal likelihoods evaluated at $\tau = 0$ and $\tau = s$ are identical.

4.2.2. Summary: boundary mutation-drift diffusion

Assuming a bi-allelic boundary mutation-drift model, a system of orthogonal eigenfunctions is defined. As with Jacobi polynomials for the general mutation-drift model, these functions provide a convenient base for calculations. While some mathematical inconvenience compared to the modified Jacobi polynomials is encountered, changes in the (effective) population size (*i.e.,* θ) are easily accommodated, because the base of the polynomials need not be changed. As with the general mutation-drift model, efficiency is increased compared to the discrete models since only eigenfunction expansions up to order M instead of N are needed.

5. The order of the expansion

With bi-allelic diffusion models we naturally assumed a binomial likelihood. This likelihood function corresponds to a polynomial of the order of the sample size M. Both with the general mutation-drift model as with the boundary mutation-drift model only orthogonal polynomials up to the order of the sample size are needed when modeling the allele trajectory backward in time. We also note that a change in the base of the polynomials, because the scaled mutation parameters changed, does not change the order of the expansion.

Now consider two populations with sample sizes $M^{(1)}$ and $M^{(2)}$. Tracing back the allele frequency evolution to the split time requires a polynomial expansion of up to $\max(M^{(1)}, M^{(2)})$. Integrating over the population allelic proportion to obtain the marginal likelihood of the data at the split time then requires multiplication with the starting distribution, which can also be expanded into

orthogonal polynomials of order M_a. If the starting population is in equilibrium, then $M_a = 0$. If we first multiply the starting distribution with the backward orthogonal expansion of the smaller population, we obtain a forward expansion of order at least $M_a + \min(M^{(1)}, M^{(2)})$. Because of the orthogonality relation, when multiplying with the backward expansion of the second population, only polynomials of order up to the minimum of $M_a + \min(M^{(1)}, M^{(2)})$ and $\max(M^{(1)}, M^{(2)})$ are needed for obtaining the marginal likelihood. Thus the maximal expansion needed depends on the sample sizes and the starting distribution, but is always at least $\min(M^{(1)}, M^{(2)})$ and at most $\max(M^{(1)}, M^{(2)})$. Therefore, the required degree of the polynomial expansion is considerably less than previously thought necessary [21, 20]. Similar considerations also apply to more than two populations, where it can be shown that the required expansion to obtain the marginal likelihood is less than the sum of the sample sizes.

6. Discussion

Starting from bi-allelic mutation-drift models, we use forward and backward processes in discrete or continuous time to efficiently calculate probabilities of population allele proportions. Given a sample from a single population, *i.e.,* a SFS, or samples from more than one population, *i.e.,* a jSFS, from the present, this theory may be used to infer trajectories of population allele frequencies in the past. Integrating over the population allelic proportion, the marginal likelihood of the data may be used to infer population genetic parameters. The discrete-time algorithm is a variant of the forward-backward algorithm and thus makes use of dynamic programming. The continuous time algorithm uses orthogonal polynomials for even more convenient calculation. Furthermore, we introduce bi-allelic population genetic models that provide us with time-reversible and irreversible transition matrices or kernels. The irreversible models are related to the infinite site [16, 8] or Poisson-random-field models [26]. Both reversible and irreversible models have stationary distributions.

Previous diffusion-based methods for inference of population genetic parameters are generally based on modelling allelic proportion trajectories forward-in-time. Solutions to the forward diffusion equations are either approximated numerically [*e.g.,* 4, 12, 22] or are provided as functions of orthogonal polynomials [*e.g.,* 21, 20, 29, 31]. These methods can, in principle, accommodate many demographic scenarios while considering general selection and continuous migration. The complexity of these models in combination with the forward-in-time approach often results in complex likelihood functions. Herein, we demonstrate that combining forward- and backward-in-time approaches naturally leads to relatively simple likelihood functions for both discrete and continuous population genetics models (compare eqs. 16 and 30, respectively).

Discrete models involve repeated multiplications with a transition matrix of dimension $(N + 1) \times (N + 1)$, where N is the haploid population size. For biological reasons, N should be large to model the large (effective) population sizes usually encountered. For numerical reasons, N should be small, because

iteration of large matrices is time-consuming and numerical errors may accumulate. Mutation rates can be scaled to account for a reduction of N. Transition matrices may be diagonalized to speed up calculations. In any case, N must be at least as big as the sample size M to not lose information. A prior distribution must be assumed at some time in the past. If this distribution is taken as the stationary distribution of the transition matrix, calculations simplify. At the present time, a probability model of the sampling process, generally a hypergeometric likelihood, is assumed that is conditional on the sample size M. Zhao *et al.* [46] present a similar method that is also based on the iteration of a transition matrix (in their case, based on the Wright-Fisher model) and allows for conditioning on the beginning and end states of the chain. They derive the marginal distribution of states intermediate in the chain and simulate trajectories. Extending this method to distributions instead of states (in our case, the prior at the beginning and the likelihood at the end of the chain) requires additional considerations and diagonalizing the transition matrix seems necessary in all but the simplest cases.

With continuous diffusion models, the use of orthogonal polynomials is convenient. The degree of the polynomials need not be higher than the sample size M, while the population size is large, which usually fits biological reality. Thus, the diffusion approach is mostly preferable over the discrete approach.

Song and colleagues [29, 30, 31, 48] analyse self-adjoint continuous models, such as the general mutation-drift model herein. These authors usually take a Dirac delta function as starting condition instead of a prior distribution at $\tau = s$ (but see Supplemental Information, Section D in Steinrücken *et al.* [31]). Representation of a Dirac delta function requires an infinite expansion and modeling an arbitrary distribution as starting condition would require a further step (see Appendix 7.3). As these authors also consider selection, eigenfunctions with, in principle, infinite expansions are necessary in any case. A problem with their approach for pure drift models, however, is the restriction at the boundaries, which allows only polymorphic samples to be analyzed (see the subsection 4.1). Interestingly, Zhao *et al.* [45] also present a diffusion approach to calculate conditional trajectories that involves the product of solutions of the forward and backward equations. They consider a Dirac delta function as starting state and, additionally, also as a final state. Usually in population genetics, however, only a sample from the present is given, while the starting conditions are even less well defined. Applying this approach to real data thus requires integration over possible starting and final states, which adds another layer of complexity avoided with our approach. In contrast, Lukić and Hey [20] also use the equilibrium distribution as a starting condition as with the approach presented herein.

Generally, using a delta function as an initial condition requires an infinite expansion in orthogonal polynomials. Yet for calculating marginal likelihoods a much lower expansion is needed. Lukić and Hey [20], citing [26], set the degree of polynomial expansion to $(M-2)^K$, where M is the number of haplotypes sampled and K the number of populations. Yet we show that only an expansion between $\min(M^{(1)}, M^{(2)})$ and $\max(M^{(1)}, M^{(2)})$ is needed, where $M^{(1)}$ and

$M^{(2)}$ are the sample sizes in the two populations. With additional populations, the expansion needed is less than $\sum_{i=1}^{K} M_i$. Furthermore, these authors use Chebyshev polynomials, which are not orthogonal with respect to the forward and backward operators. This necessitates numerical integration of a linear system of differential equations to obtain the temporal part of the solution. With orthogonal polynomials, the corresponding system of differential equations is diagonal and thus much simpler.

An analysis also involving a coupled system of ordinary differential equations for the temporal evolution of moments [8, 47, 48] also provides solutions for the forward and backward diffusions. The basic model analyzed by these authors is the continuous version of the single-boundary mutation-drift model presented here, where ancestral and derived alleles are differentiated. Zivkovic and Stephan [47] also point out relations of the backward approach to coalescent theory. Recently, a diffusion framework of weak mutation and selection has been incorporated in the theoretical analysis of adaptive landscapes [42], a concept first formulated by Wright [41].

We note that many approaches above [8, 21, 20, 47, 48, 36] use boundary mutation models. Indeed, much of the statistics of population genetics is based on this model, *e.g.,* the important Ewens-Watterson θ [9, 39]. For this model, only the forward transition probabilities have been given so far [8, 21, 36]. For the first time, we give the backward system of orthogonal polynomials and their corresponding eigenvalues herein. The system of eigenfunctions of the pure drift model [33] follows as a special case. As explained above, the possibility to move backward simplifies inference.

The demographic scenarios presented here (Fig. 1) are common, *e.g.,* in natural populations of fruit flies of the *Drosophila* genus [*e.g.,* 19, 43, 24]. Additionally, the abundance of population data for *Drosophila* species makes them especially suitable for SFS and jSFS analysis under the described framework. Furthermore, the theory can be extended to more than two populations, *i.e.,* to phylogenetic inference. Our methods can also be adjusted to an experimental setting with samples from multiple time points, as *e.g.,* in evolve-and-resequence experiments [18]. Furthermore, a setting with multiple time-points also applies to the analysis of ancient DNA samples as noted by Steinrücken *et al.* [31].

Generally, the methods and models we present in this article are simple, yet allow for maximum marginal likelihood analysis of SFS and jSFS from splitting populations with mutation-drift or pure drift models, and for inference of evolutionary trajectories of population allele proportions conditional on data.

Acknowledgments

The authors thank Reinhard Bürger, Joachim Hermisson and other colleagues from the Faculty of Mathematics of the University of Vienna, all members of the Institute of Population Genetics at the University of Veterinary Medicine, Vienna, and Andreas Futschik (Johannes Kepler Universtiy, Linz). All authors were supported by the Austrian Science Fund (FWF): DK W1225-B20. DS and CK were partially funded by FWF-P24551-B25. CK has been

partially funded by the Vienna Science and Technology Fund (WWTF) through project MA16-061.

References

[1] Abramowitz, M. and Stegun, I., editors (1970). *Handbook of Mathematical Functions*. Dover, 9th ed. edition.

[2] Baake, E. and Bialowons, R. (2008). Ancestral processes with selection: branching and Moran models. *Banach center publications*, **80**, 33–52.

[3] Bayin, S. (2006). *Mathematical methods in science and engineering*. Wiley, N.Y.

[4] Bollback, J. P., York, T. L., and Nielsen, R. (2008). Estimation of $2N_e s$ from temporal allele frequency data. *Genetics*, **179**(1), 497–502.

[5] Carlin, B. and Louis, T. (2000). *Bayes and empirical Bayes methods*. Chapman and Hall, 2nd ed. edition.

[6] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis*. Cambridge University Press, Cambridge.

[7] Etheridge, A. and Griffiths, R. (2009). A coalescent dual process in a Moran model with genic selection. *Theoretical Population Biology*, **75**, 320–330.

[8] Evans, S., Shvets, Y., and Slatkin, M. (2007). Non-equilibrium theory of the allele frequency spectrum. *Theoretical Population Biology*, **71**, 109–119.

[9] Ewens, W. (1974). A note on the sampling theory for infinite alleles and infinite sites models. *Theoretical Population Biology*, **6**, 143–148.

[10] Ewens, W. (2004). *Mathematical Population Genetics*. Springer, N.Y., 2nd edition.

[11] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.

[12] Gutenkunst, R., Hernandez, R., Williamson, S., and Bustamante, C. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genetics*, **5**, e1000695.

[13] Hein, J., Schierup, M., and Wiuf, C. (2005). *Gene genealogies, variation, and evolution: a primer in coalescent theory.* Oxford University Press.

[14] Jewett, E. M., Steinrücken, M., and Song, Y. S. (2016). The effects of population size histories on estimates of selection coefficients from time-series genetic data. *Molecular biology and evolution*, **33**(11), 3002–3027.

[15] Kimura, M. (1955). Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA*, **41**, 144–150.

[16] Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**, 893–903.

[17] Kingman, J. (1982). On the genealogy of large populations. *Journal of Applied Probability*, **19A**, 27–43.

[18] Kofler, R. and Schlötterer, C. (2014). A guide for the design of evolve and resequencing studies. *Molecular Biology and Evolution*, **31**, 474–483.

[19] Li, H. and Stephan, W. (2006). Inferring the Demographic History and Rate of Adaptive Substitution in *Drosophila*. *PLOS Genetics*, **10**, e166.

[20] Lukić, S. and Hey, J. (2012). Demographic inference using spectral methods on SNP data, with an analysis of the human out-of-Africa expansion. *Genetics*, **192**(2), 619–639.

[21] Lukić, S., Hey, J., and Chen, K. (2011). Non-equilibrium allele frequency spectra via spectral methods. *Theoretical population biology*, **79**(4), 203–219.

[22] Malaspinas, A.-S., Malaspinas, O., Evans, S. N., and Slatkin, M. (2012). Estimating allele age and selection coefficient from time-serial data. *Genetics*, **192**(2), 599–607.

[23] McKane, A. and Waxman, D. (2007). Singular solutions of the diffusion equation of population genetics. *Journal of Theoretical Biology*, **247**, 849–858.

[24] Pool, J. E., Corbett-Detig, R. B., Sugino, R. P., Stevens, K. A., Cardeno, C. M., Crepeau, M. W., Duchen, P., Emerson, J. J., Saelao, P., Begun, D. J., and Langley, C. H. (2012). Population genomics of sub-saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLOS Genet*, **8**(12), e1003080.

[25] Rabiner, L. and Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP magazine*, **3**, 4–16.

[26] Sawyer, S. and Hartl, D. (1992). Population genetics of polymorphism and divergence. *Genetics*, **132**, 1161–1176.

[27] Schraiber, J. G., Evans, S. N., and Slatkin, M. (2016). Bayesian inference of natural selection from allele frequency time series. *Genetics*, **203**(1), 493–511.

[28] Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., and Kosiol, C. (2016). Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology*, **407**, 362–370.

[29] Song, Y. and Steinrücken, M. (2012). A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics*, **190**, 1117–1129.

[30] Steinrücken, M., Wang, R., and Song, Y. (2013). An explicit transition density expansion for a multi-allelic WrightFisher diffusion with general diploid selection. *Theoretical Population Biology*, **83**, 1–14.

[31] Steinrücken, M., Bhaskar, A., and Song, Y. (2014). A novel method for inferring general diploid selection from time series genetic data. *Annals of Applied Statistics*, **8**, 2203–2222.

[32] Steinrücken, M., Jewett, E. M., and Song, Y. S. (2015). SpectralTDF: transition densities of diffusion processes with time-varying selection parameters, mutation rates and effective population sizes. *Bioinformatics*, **32**(5), 795–797.

[33] Tran, T., Hofrichter, J., and Jost, J. (2013). An introduction to the mathematical structure of the Wright-Fisher model of population genetics. *Theory in Biosciences*, **132**, 73–82.

[34] Vogl, C. (2014). Estimating the Scaled Mutation Rate and Mutation Bias with Site Frequency Data. *Theoretical Population Biology*, **98**, 19—27.

[35] Vogl, C. and Bergman, J. (2015). Inference of directional selection and mutation parameters assuming equilibrium. *Theoretical Population Biology*, **106**, 71–82.

[36] Vogl, C. and Bergman, J. (2016). Computation of the likelihood of joint site frequency spectra using orthogonal polynomials. *Computation*, **4**, 6.

[37] Vogl, C. and Futschik, A. (2010). Hidden markov models in biology. In O. Carugo and F. EIsenhaber, editors, *Biological Data Mining.*, Methods in Molecular Biology. Humana Press.

[38] Wakeley, J. (2009). *Coalescent theory, an Introduction*. Roberts and Co.

[39] Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

[40] Waxman, D. (2011). Comparison and content of the WrightFisher model of random genetic drift, the diffusion approximation, and an intermediate model. *Journal of Theoretical Biology*, **269**, 79–87.

[41] Wright, S. (1932). The roles of mutation, inbreeding, crossbreeding, and selection in evolution. *Proceedings of the sixth international congress of genetics*, **1**, 356–366.

[42] Xu, S., Jiao, S., Jiang, P., and Ao, P. (2014). Two-time-scale population evolution on a singular landscape. *Physical Review E*, **89**(1), 012724.

[43] Zeng, K. and Charlesworth, B. (2010). Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. *Journal of Molecular Evolution*, **183**, 651–662.

[44] Zhao, L., Lascoux, M., Overall, A., and Waxman, D. (2013a). The characteristic trajectory of a fixing allele: a consequence of fictitious selection that arises from conditioning. *Genetics*, **195**, 993–1006.

[45] Zhao, L., Yue, X., and Waxman, D. (2013b). Complete numerical solution of the diffusion equation of random genetic drift. *Genetics*, **194**, 419–426.

[46] Zhao, L., Yue, X., and Waxman, D. (2014). Exact solution of conditioned Wright-Fisher models. *Journal of Theoretical Biology*, **194**, 973–985.

[47] Zivkovic, D. and Stephan, W. (2011). Analytical results on the neutral non-equilibrium allele frequency spectrum based on diffusion theory. *Theoretical Population Biology*, **79**, 184–191.

[48] Zivkovic, D., Steinrücken, M., Song, Y., and Stephan, W. (2015). Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics*, **200**, 601–617.

7. Appendices

7.1. Derivation of the forward and backward diffusion equations from the decoupled general mutation-drift Moran model

In this appendix, we derive the forward and backward diffusion equation from the forward and backward transition probabilities of the decoupled Moran model with general mutation and drift and show the tight connections between the discrete and continuous models. Derivations are simpler than usual [10]; terms higher than the first derivative with respect to time and second derivative with respect to space do not occur.

Consider a focal bi-allelic site with the population frequency of allele one denoted by i ($1 \leq i \leq N-1$). With the transition probabilities of the decoupled Moran model (24), the frequency i may increase or decrease by one due to mutation or drift, or remain constant. Forward in time, the difference of the probability at frequency i per Moran step may be written as

$$\begin{aligned}
&\Pr(x_{t+1}=i) - \Pr(x_t=i) = \\
&\quad \frac{\alpha\theta}{N^2}\Big((N-i+1)\Pr(x_t=i-1) - (N-i)\Pr(x_t=i)\Big) \\
&\quad + \frac{\beta\theta}{N^2}\Big((i+1)\Pr(x_t=i+1) - i\Pr(x_t=i)\Big) \\
&\quad + \frac{1}{N^2}\Big((i-1)(N-i+1)\Pr(x_t=i-1) \\
&\quad + (i+1)(N-i-1)\Pr(x_t=i+1) - 2i(N-i)\Pr(x_t=i)\Big),
\end{aligned} \tag{76}$$

where the term within the first pair of square brackets corresponds to mutation towards allele one, the term within the second pair to mutation towards allele zero, and the term within the third pair to genetic drift.

To approximate the change in frequency as a process in continuous time and space, the quantities $\delta\tau = 1/N^2$ and $\delta x = 1/N$ are introduced. Furthermore, time is rescaled as $\tau = t\delta\tau$, the allele proportions as $x = i\delta x$, such that $\phi(x \,|\, \tau, \rho)\delta\tau\delta x = \Pr(x_t = i)$. Taking the limit $N \to \infty$, eq. (76) is rewritten as

$$\begin{aligned}
&\lim_{N\to\infty} \frac{\phi(x \,|\, \tau+\delta\tau, \rho) - \phi(x \,|\, \tau, \rho)}{\delta\tau} = \\
&\lim_{N\to\infty} \Bigg[\alpha\theta\Bigg(\frac{(1-x+\delta x)\phi(x-\delta x \,|\, \tau, \rho) - (1-x)\phi(x \,|\, \tau, \rho)}{\delta x}\Bigg) \\
&\quad + \beta\theta\Bigg(\frac{(x+\delta x)\phi(x+\delta x \,|\, \tau, \rho) - x\phi(x \,|\, \tau, \rho)}{\delta x}\Bigg) \\
&\quad + \Bigg(\frac{(x-\delta x)(1-x+\delta x)\phi(x-\delta x \,|\, \tau, \rho)}{\delta x^2} \\
&\quad + \frac{(x+\delta x)(1-x-\delta x)\phi(x+\delta x \,|\, \tau, \rho)}{\delta x^2} - \frac{2x(1-x)\phi(x \,|\, \tau, \rho)}{\delta x^2}\Bigg)\Bigg].
\end{aligned} \tag{77}$$

The term to the left of the equality sign of (77) corresponds to the definition of the first derivative with respect to time τ of $\phi(x \,|\, \tau, \rho)$; the terms with mutations correspond to the first derivatives with respect to x of $-(1-x)\phi(x \,|\, \tau, \rho)$ and $x\phi(x \,|\, \tau, \rho)$, respectively; the drift term corresponds to the definition of the second symmetric derivative with respect to x of $x(1-x)\phi(x \,|\, \tau, \rho)$. After minor rearrangements, the familiar form of the forward general mutation-drift diffusion equation is obtained

$$\frac{\partial}{\partial \tau}\phi(x \,|\, \tau, \rho) = -\frac{\partial}{\partial x}\theta(\alpha - x)\phi(x \,|\, \tau, \rho) + \frac{\partial^2}{\partial x^2}x(1-x)\phi(x \,|\, \tau, \rho). \tag{78}$$

Considering the Moran model backward in time (see Subsection 2.4), the change in frequency i back in time is determined by the transpose of the forward transition matrix (24) and can be written as

$$\begin{aligned}
&\Pr(y \,|\, x_t = i) - \Pr(y \,|\, x_{t+1} = i) = \\
&\quad \frac{\alpha\theta(N-i)}{N^2}\Big(\Pr(y \,|\, x_{t+1} = i+1) - \Pr(y \,|\, x_{t+1} = i)\Big) \\
&\quad + \frac{\beta\theta i}{N^2}\Big(\Pr(y \,|\, x_{t+1} = i-1) - \Pr(y \,|\, x_{t+1} = i)\Big) \\
&\quad + \frac{i(N-i)}{N^2}\Big(\Pr(y \,|\, x_{t+1} = i+1) + \Pr(y \,|\, x_{t+1} = i-1) \\
&\quad - 2\Pr(y \,|\, x_{t+1} = i)\Big).
\end{aligned} \tag{79}$$

After rescaling time and space, considering the limit $N \to \infty$, and setting $\psi(y \,|\, x, \tau) = \Pr(y \,|\, x_{t+1} = i)$, we get the backward diffusion equation

$$-\frac{\partial}{\partial \tau}\psi(y \,|\, x, \tau) = \theta(\alpha - x)\frac{\partial}{\partial x}\psi(y \,|\, x, \tau) + x(1-x)\frac{\partial^2}{\partial x^2}\psi(y \,|\, x, \tau). \tag{80}$$

The minus sign on the left side of the backward diffusion equation (80) may be unusual [compare 10], but necessary such that the time τ runs in the same direction in the forward and backward diffusion. Note that Zhao *et al.* [44] also use a pair of forward and backward diffusion equations with differing signs.

7.2. Boundary condition

In the following, we use the prime ($'$) to indicate the (partial) derivative with respect to x and leave away the terms in brackets for ϕ and ψ. Eq. (32) can then be written as

$$\int_0^1 \left[-(P\phi)' + (Q\phi)''\right]\psi \, dx = \int_0^1 \phi\left[P\psi' + Q\psi''\right] dx\,. \tag{81}$$

The first term on the right side is

$$\int_0^1 \phi P\psi' \, dx = \phi P\psi\big|_0^1 - \int_0^1 (\phi P)'\psi dx \tag{82}$$

and the second term on the right side is

$$\begin{aligned}\int_0^1 \phi Q\psi^{''}\,dx &= \phi Q\psi^{'}\big|_0^1 - \int_0^1 (Q\phi)^{'}\psi^{'}\,dx \\ &= \phi Q\psi^{'}\big|_0^1 - (\phi Q)^{'}\psi\big|_0^1 + \int_0^1 (Q\psi)^{''}\psi dx\,,\end{aligned} \tag{83}$$

Hence for eq. (81) to hold, we require the boundary condition

$$\left(\phi Q\psi^{'} - (\phi Q)^{'}\psi + \phi P\psi\right)\big|_0^1 = 0\,. \tag{84}$$

Using the weight function $w(x)$ defined in formula (33), this condition can be represented more compactly. The weight function fulfils

$$Pw = (wQ)^{'}\,. \tag{85}$$

Substitute $\phi(x\,|\,\tau) = w(x)g(x,\tau,\rho)$ into eq. (84) to obtain

$$\begin{aligned}0 &= \left(wQg\psi^{'} - (wQg)^{'}\psi + Pwg\psi\right)\big|_0^1 \\ &= \left(wQg\psi^{'} - ((wQ)^{'}g + wQg')\psi + Pwg\psi\right)\big|_0^1 \\ &= \left(wQg\psi^{'} - Pwg\psi - wQg'\psi + Pwg\psi\right)\big|_0^1 \\ &= wQ(g\psi^{'} - g'\psi)\big|_0^1\end{aligned} \tag{86}$$

Note that $w(x)Q(x) \propto 1/\xi(x)$ where $\xi(x)$ is the scale function defined in eq. (2) of Song and Steinrücken [29] and $g(x)$ and $\psi(x)$ correspond to $f(x)$ in Song and Steinrücken [29]. This condition obviously holds if, at both boundaries, either $w(x)Q(x) = 0$ while $(g\psi^{'} - g'\psi)$ is finite, or $(g\psi^{'} - g'\psi) = 0$ while $w(x)Q(x)$ is finite.

7.3. Propagator

Song and Steinrücken [29] analyze self-adjoint differential equations, with a Dirac delta function $\delta(x-p)$ as starting point at $\tau = s$. Denote the eigenfunctions of the diffusion equation with the backward operator $\mathcal{L}^*$ with $B_n(x)$. Eq. (5) of Song and Steinrücken [29] defines a "propagator" [3, chap. 19]

$$p(x\,|\,p,\tau) = \sum_{n=0}^{\infty} e^{-\lambda_i \tau}\pi(x)\frac{B_n(x)B_n(p)}{\langle B_n(x)B_n(p)\rangle_\pi} \tag{87}$$

as the solution of the diffusion equation with a starting state modeled by the Dirac Delta function $\delta(x-p)$. If the starting condition is not a particular state but, more usually, a distribution $\rho(p)$, the function

$$h(x\,|\,p,\rho,\tau) = \int_0^1 p(x\,|\,p,\tau)\rho(p)\,dp \tag{88}$$

solves the diffusion equation. From the orthogonality relation it is evident that, also with this indirect route, only an expansion of degree M is needed for calculating the marginal likelihood.