SpectralTDF: transition densities of diffusion processes with time-varying selection parameters, mutation rates, and effective population sizes

Matthias Steinrücken [1,*], Ethan M. Jewett [2,*], and Yun S. Song [2,3,4,5,6]

[1]Department of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA, USA.

[2]Computer Science Division, [3]Department of Statistics, and [4]Department of Integrative Biology, University of California, Berkeley, CA, USA.

[5]Department of Mathematics and [6]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA.

[*]These authors contributed equally to this work.

ABSTRACT

In the Wright-Fisher diffusion, the transition density function (TDF) describes the time-evolution of the population-wide frequency of an allele. This function has several practical applications in population genetics, and computing it for biologically realistic scenarios with selection and demography is an important problem. We develop an efficient method for finding a spectral representation of the TDF for a general model where the effective population size, selection coefficients, and mutation parameters vary over time in a piecewise constant manner. The method, called spectralTDF, is available at `https://sourceforge.net/projects/spectraltdf/`.

1. Introduction

The transition density function (TDF) of the Wright-Fisher diffusion describes the time-evolution of the frequency of an allele (Ewens 2004). The TDF is useful for understanding the effects of demography, mutation, and selection on genetic variation, and it is a key component of a number of methods for inferring selection coefficients (Williamson *et al.* 2004, Bollback *et al.* 2008, Steinrücken *et al.* 2014), predicting allele fixation times (Waxman 2011), and computing population genetic statistics such as the site frequency spectrum (Živković *et al.* 2015).

Most existing approaches for computing the TDF assume either restrictive models of dominance (Kimura 1955; 1957) or selective neutrality (Shimakura 1977, Griffiths 1979, Vogl 2014a), or are computationally slow for selection strengths commonly observed in biological data (Barbour *et al.* 2000). However, Song and Steinrücken (2012), Steinrücken *et al.* (2013) recently developed a numerically stable and computationally efficient method for finding a spectral representation of the TDF for a general selection model in the case of constant parameters (population size, mutation rates, and selection coefficients). Despite the utility of this new approach, assuming that model parameters remain constant over time is often too restrictive for biological applications (Siepielski *et al.* 2009).

Živković *et al.* (2015) have extended the spectral method of Song and Steinrücken (2012) to handle piecewise-constant population size functions. However, their approach requires a restricted model of selection in which the fitness of a homozygote is twice that of a heterozygote (i.e., additive or genic selection). Furthermore, selection parameters are assumed to remain constant over time and the model does not allow for recurrent mutations.

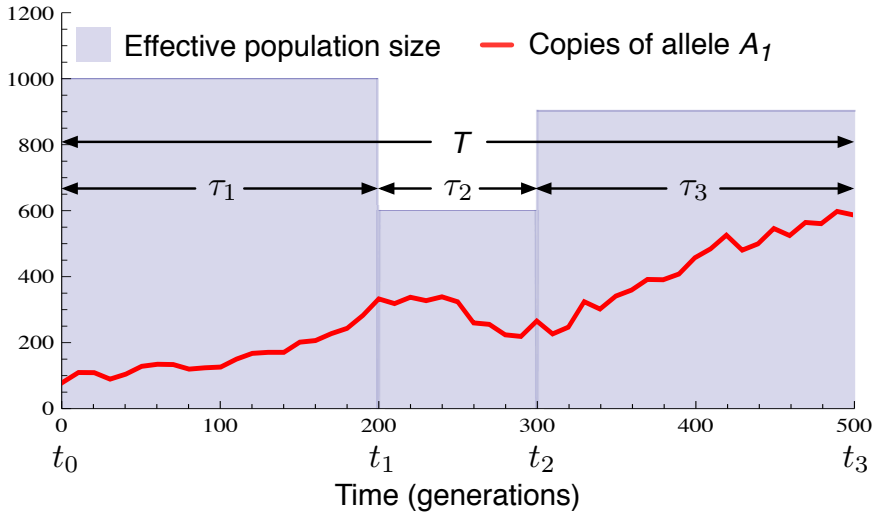

FIGURE 1. Diagram of the model. A population has constant size in each of K epochs ($N_1 = 1000$, $N_2 = 600$, $N_3 = 900$). An allele, A_1, at a locus of interest evolves over time, subject to pressures of mutation and selection that are constant within each epoch.

Here, we present the first method for computing the TDF under arbitrary models of dominance and recurrent mutation while allowing selection parameters, mutation rates, and effective population sizes to change over time in a piecewise constant manner.

2. Model and approach

We consider a biallelic locus with two alleles, A_0 and A_1, evolving in a single panmictic population. In the corresponding Wright-Fisher diffusion, X_t denotes the frequency of allele A_1 at time t, measured continuously in units of generations. We assume that either X_0 is given or the distribution of X_0 is specified. The effective population size, mutation rates, and selection parameters are assumed to be constant within each of K disjoint epochs. As illustrated in Figure 1, the kth epoch has effective size N_k (diploid individuals) and duration τ_k. Epoch boundaries are denoted by $t_0, t_1, \ldots, t_K$, with $t_k = \sum_{i=1}^{k} \tau_i$.

Within the kth epoch, the per-generation probability that a copy of allele A_0 mutates to allele A_1 is a_k, and the per-generation probability that a copy of allele A_1 mutates to allele A_0 is b_k. In addition, selection acts in such a way that the relative fitness of an individual carrying i copies of allele A_1 is $1 + s_{ki}$ $(i = 1, 2)$.

Within each epoch, k, a spectral representation of the TDF, $p_k(t; x, y)$, can be obtained by employing the framework of Song and Steinrücken (2012). The challenge in computing the TDF for the full model with K epochs lies in knitting together the expressions for the densities $p_k(t; x, y)$ across the different epochs. We first review the derivations of Song and Steinrücken (2012) and Steinrücken *et al.* (2014) of the TDF in a single epoch of constant size. We then discuss our efficient polynomial interpolation method for knitting together the TDF across epochs of different constant sizes.

2.1. **The TDF and its generalization in a single epoch of constant size.** We wish to compute the TDF in the kth epoch, where the density is defined by $p_k(t;x,y)dy = \mathbb{P}(y \le X_{t_{k-1}+t} < y + dy | X_{t_{k-1}} = x)$, for $t \in [t_{k-1}, t_k)$. We are also interested in the generalization, $\phi_k(t;y) = \int_0^1 p_k(t;z,y)\rho_k(z)dz$, which extends the TDF to the case of a general initial density ρ_k rather than a point mass at x.

In an epoch of constant size N_k, let $\alpha_k = 4N_k a_k$, $\beta_k = 4N_k b_k$, $\sigma_{k,1} = N_k s_{k,1}$, and $\sigma_{k,2} = N_k s_{k,2}$ denote population-size scaled versions of the per-generation mutation parameters (a_k, b_k) and selection coefficients $(s_{k,1}, s_{k,2})$. The Kolmogorov backward operator $\mathcal{L}_k$ for the epoch is the second-order linear differential operator given by

$$\mathcal{L}_k = \frac{1}{2}\xi^2(x)\frac{\partial^2}{\partial x^2} + \mu(x)\frac{\partial}{\partial x}. \tag{1}$$

In Equation (1), the quantity

$$\xi^2(x) = x(1-x) \tag{2}$$

captures the contribution from genetic drift and the term

$$\mu(x) = \frac{1}{2}[\alpha_k - (\alpha_k + \beta_k)x] + 2x(1-x)[\sigma_{k,1}(1-2x) + \sigma_{k,2}x] \tag{3}$$

captures the contribution from recurrent mutation and selection. The TDF is the solution of the Kolmogorov backward equation

$$\frac{\partial p_k(t;x,y)}{\partial t} = \frac{N_{\text{ref}}}{N_k}\mathcal{L}_k p_k(t;x,y) \tag{4}$$

satisfying specified boundary conditions [see Song and Steinrücken (2012), p. 119, for a discussion of the boundary conditions]. We measure time in units of $2N_{\text{ref}}$ generations in all epochs, where N_{ref} is the size of a fixed reference population. For ease of interpretation, we choose $N_{\text{ref}} = 1/2$ so that time is measured in units of generations in all epochs.

Song and Steinrücken (2012) derived a formula for the TDF by obtaining a solution of the backward Equation (4) in the form of the infinite series

$$p_k(t;x,y) = \sum_{n=0}^{\infty} d_{k,n}(t,y)B_{k,n}(x) = \sum_{n=0}^{\infty} e^{-\lambda_{k,n}t/(2N_k)}\frac{\pi_k(y)B_{k,n}(y)B_{k,n}(x)}{\langle B_{k,n}, B_{k,n}\rangle_{\pi_k}}, \tag{5}$$

where $\{B_{k,n}(x)\}_{n=0}^{\infty}$ is the set of eigenfunctions of $\mathcal{L}_k$ with associated eigenvalues $\{\lambda_{k,n}\}_{n=0}^{\infty}$ (Section 2.2.1) and the function $\pi_k(y)$ is given by

$$\pi_k(y) = e^{\bar{\sigma}(y)}y^{\alpha_k - 1}(1-y)^{\beta_k - 1}, \tag{6}$$

where $\bar{\sigma}_k(y) = 4\sigma_{k,1}y(1-y) + 2\sigma_{k,2}y^2$. The inner product $\langle f, g\rangle_\omega$ with respect to a weight function $\omega(x)$ in Equation (5) is defined for two functions f and g on an interval $[a,b]$ by

$$\langle f, g\rangle_\omega = \int_a^b f(x)g(x)\omega(x)dx. \tag{7}$$

In Equation (5), the inner product $\langle \cdot, \cdot \rangle_{\pi_k}$ is taken over the interval $[0,1]$ with respect to $\pi_k(y)$. Equation (5) can be thought of as a function in either the initial frequency, x, or the final frequency, y.

2.2. **The generalized TDF.** The generalization $\phi_k(t;y) = \langle p_k(t;\cdot,y),\rho_k\rangle$ of Equation (5) to the case in which the initial probability density is given by $\rho_k(x)$ is easily obtained by noting that, viewed as an expansion in the basis functions $\{B_{k,n}(x)\}_{n=0}^{\infty}$, the partial sum $p_k^{(M)}(t;x,y) = \sum_{n=0}^{M} d_{k,n}(t,y)B_{k,n}(x)$ converges strongly to $p_k(t;x,y)$, from which it follows that

$$\begin{aligned}
|\langle p_k^{(M)}(t;\cdot,y),\rho_k\rangle - \langle p_k(t;\cdot,y),\rho_k\rangle| &= |\langle p_k^{(M)}(t;\cdot,y) - p_k(t;\cdot,y),\rho_k\rangle| \\
&= |\langle p_k^{(M)}(t;\cdot,y) - p_k(t;\cdot,y),\rho_k/\pi_k\rangle_{\pi_k}| \\
&\leq \|p_k^{(M)}(t;\cdot,y) - p_k(t;\cdot,y)\|_{\pi_k}\|\rho_k\|_{1/\pi_k} \to 0 \text{ as } M\to\infty,
\end{aligned} \tag{8}$$

for $\rho_k(x)$ satisfying $\|\rho_k\|_{1/\pi_k} < \infty$. Thus, we have

$$\begin{aligned}
\phi_k(t;y) &= \lim_{M\to\infty} \langle p_k^{(M)}(t;\cdot,y),\rho_k\rangle \\
&= \lim_{M\to\infty} \int_0^1 \rho_k(x) \sum_{n=0}^{M} e^{-\lambda_{k,n}t/(2N_k)} \frac{\pi_k(y)B_{k,n}(y)B_{k,n}(x)}{\langle B_{k,n},B_{k,n}\rangle_{\pi_k}} dx \\
&= \lim_{M\to\infty} \sum_{n=0}^{M} \left[\frac{\int_0^1 \rho_k(x)B_{k,n}(x)dx}{\langle B_{k,n},B_{k,n}\rangle_{\pi_k}}\right] e^{-\lambda_{k,n}t/(2N_k)}\pi_k(y)B_{k,n}(y) \\
&= \sum_{n=0}^{\infty} c_{k,n} e^{-\lambda_{k,n}t/(2N_k)}\pi_k(y)B_{k,n}(y),
\end{aligned} \tag{9}$$

where

$$c_{k,n} = \frac{\int_0^1 \rho_k(x)B_{k,n}(x)dx}{\langle B_{k,n},B_{k,n}\rangle_{\pi_k}} = \frac{\int_0^1 \rho_k(x)\pi_{k,n}(x)B_{k,n}(x)\frac{1}{\pi_k(x)}dx}{\langle \pi_k B_{k,n},\pi_k B_{k,n}\rangle_{1/\pi_k}} = \frac{\langle \rho_k,\pi_k B_{k,n}\rangle_{1/\pi_k}}{\|\pi_k B_{k,n}\|_{1/\pi_k}}. \tag{10}$$

Determining the generalization, $\phi_k(t;y)$, for any initial condition $\rho_k \in L^2([0,1],1/\pi_k)$ thus amounts to projecting ρ_k onto the functions $\{\pi_k B_{k,n}\}_{n=0}^{\infty}$ and plugging these coefficients into Equation (9) (Steinrücken *et al.* 2014, Vogl 2014b).

2.2.1. *Computing the eigenfunctions* $\{B_{k,n}(y)\}_{n=0}^{\infty}$. Before discussing how to extend Equations (5) and (9) to the case of a population with piecewise constant parameters, which is the goal of this article, we first review the derivation of the eigenfunctions $\{B_{k,n}(y)\}_{n=0}^{\infty}$ derived by Steinrücken *et al.* (2014).

Steinrücken *et al.* (2014) showed that the eigenfunctions $\{B_{k,n}(y)\}_{n=0}^{\infty}$ can be expressed as

$$B_{k,n}(y) = \sum_{m=0}^{\infty} w_{k,n,m} e^{-\bar{\sigma}_k(y)} R_m^{(\alpha_k,\beta_k)}(y), \tag{11}$$

where $R_m^{(\alpha,\beta)}(y) = p_n^{(\beta-1,\alpha-1)}(2y-1)$, and $p_n^{(a,b)}(y)$ is the nth classical Jacobi polynomial. The vector $\mathbf{w}_{k,n} = (w_{k,n,0}, w_{k,n,1},\ldots)$ is the left eigenvector of the matrix

$$\mathbf{M}_k := -\left(\mathbf{\Lambda}^{(\alpha_k,\beta_k)} + \sum_{\ell=0}^{4} q_{k,\ell}\mathbf{G}_k^{\ell}\right) \tag{12}$$

corresponding to the nth eigenvalue $\lambda_{k,n}$. In Equation (12), $\mathbf{\Lambda}^{(\alpha,\beta)} = \mathrm{diag}(\lambda_0^{(\alpha,\beta)}, \lambda_1^{(\alpha,\beta)}, \ldots)$ is the diagonal matrix with elements given by $\lambda_n^{(\alpha,\beta)} = \frac{1}{2}n(n+\alpha+\beta-1)$ and $q_{k,\ell}$ and $\mathbf{G}_k^\ell$ are given by

$$
\begin{aligned}
q_{k,0} &= \alpha_k \sigma_{k,1}, \\
q_{k,1} &= -(2+3\alpha_k+\beta_k-2\sigma_{k,1})\sigma_{k,1} + (1+\alpha_k)\sigma_{k,2}, \\
q_{k,2} &= -10\sigma_{k,1}^2 - (1+\alpha_k+\beta_k)\sigma_{k,2} + (2+2\alpha_k+2\beta_k+4\sigma_{k,2})\sigma_{k,1}, \\
q_{k,3} &= 16\sigma_{k,1}^2 - 12\sigma_{k,1}\sigma_{k,2} + 2\sigma_{k,2}^2, \\
q_{k,4} &= -2(\sigma_{k,2}-2\sigma_{k,1})^2,
\end{aligned}
\tag{13}
$$

and

$$
[\mathbf{G}_k]_{n,m} = \begin{cases} \frac{(n+\alpha_k-1)(n+\beta_k-1)}{(2n+\alpha_k+\beta_k-1)(2n+\alpha_k+\beta_k-2)}, & \text{if } m = n-1 \text{ and } n > 0, \\ \frac{1}{2} - \frac{\beta_k^2-\alpha_k^2-2(\beta_k-\alpha_k)}{2(2n+\alpha_k+\beta_k)(2n+\alpha_k+\beta_k-2)}, & \text{if } m = n \text{ and } n \geq 0, \\ \frac{(n+1)(n+\alpha_k+\beta_k-1)}{2(2n+\alpha_k+\beta_k)(2n+\alpha_k+\beta_k-1)}, & \text{if } m = n+1 \text{ and } n \geq 0, \\ 0, & \text{otherwise.} \end{cases}
\tag{14}
$$

The eigenfunctions computed using Equation (11) can then be plugged into Equations (5) and (9), yielding the series expansions of $\phi_k(t;y)$ and $p_k(t;x,y)$. In practice, we truncate the summations in Equations (5) and (9) at some large integer N, yielding

$$
\phi_k(t;y) \approx \sum_{n=0}^{N} c_{k,n} e^{-\lambda_{k,n} t/(2N_k)} \pi_k(y) B_{k,n}(y)
\tag{15}
$$

and

$$
p_k(t;x,y) \approx \sum_{n=0}^{N} e^{-\lambda_{k,n} t/(2N_k)} \frac{\pi_k(y) B_{k,n}(y) B_{k,n}(x)}{\langle B_{k,n}, B_{k,n} \rangle_{\pi_k}}.
\tag{16}
$$

We also truncate the summation in Equation (11) at some large integer $M \geq N$, yielding

$$
B_{k,n}(y) \approx \sum_{m=0}^{M} w_{k,n,m} e^{-\bar{\sigma}_k(y)/2} R_m^{(\alpha_k,\beta_k)}(y).
\tag{17}
$$

The eigenvalues, $\lambda_{k,n}$, and eigenvectors, $w_{k,n}$, of the matrix $\mathbf{M}_k$ in Equation (12) are also computed by truncating the matrix $\mathbf{M}_k$ to dimension $D \times D$, for some large integer $D \geq M$.

2.3. **The TDF in a population with piecewise-constant parameters.** The goal of this work is to extend the results of Song and Steinrücken (2012) to populations with piecewise constant parameters. The challenge in computing the TDF for such piecewise constant populations lies in knitting together the expressions for the densities $\phi_k(t;y)$ across the different epochs. This knitting procedure can be accomplished by taking the density at the end of epoch k as the initial condition for the density in epoch $k+1$ (i.e., $\rho_{k+1}(y) = \phi_k(t_k - t_{k-1}; y)$) by transforming the time-propagated coefficients $\{c_{k,n} e^{-\lambda_{k,n}(t_k - t_{k-1})/(2N_k)}\}_{n=0}^{\infty}$ at the end of epoch k into the unpropagated coefficients $\{c_{k+1,n}\}_{n=0}^{\infty}$ at the beginning of epoch $k+1$.

Here, we focus on the extension of the generalization $\phi_k(t;y)$ to multiple epochs, rather than generalizing $p_k(t;x,y)$ itself. The TDF, $p_k(t;x,y)$, along with generalized transition densities for

other initial distributions $\rho_k(x)$ are obtained as special cases of $\phi_k(t;y)$ in Section 3 by fitting the initial values of the coefficients $\{c_{1,n}\}_{n=0}^{\infty}$ to different initial conditions, $\rho_k(x)$.

2.4. **Transforming coefficients across epochs.** Plugging Equations (16) and (17) into the condition $\rho_{k+1}(y) = \phi_k(t_k - t_{k-1}; y)$ and dividing both sides by $\pi_{k+1}(y)e^{-\bar{\sigma}_{k+1}(y)/2}$ gives

$$\sum_{n=0}^{N} c_{k+1,n} \sum_{m=0}^{M} w_{k+1,n,m} R_m^{(\alpha_{k+1},\beta_{k+1})}(y)$$

$$= \frac{\pi_k(y)e^{-\bar{\sigma}_k(y)/2}}{\pi_{k+1}(y)e^{-\bar{\sigma}_{k+1}(y)/2}} \sum_{n=0}^{N} c_{k,n} e^{-\lambda_{k,n}(t_k - t_{k-1})/(2N_k)} \sum_{m=0}^{M} w_{k,n,m} R_m^{(\alpha_k,\beta_k)}(y). \tag{18}$$

Because the left-hand side of Equation (18) is a polynomial of degree M, it is determined by $M+1$ points. Therefore, we can determine the coefficients on the left-hand side by evaluating both sides of Equation (18) at a set of points $\mathbf{y} = \{y_0, \ldots, y_M\}$. We choose the set $\mathbf{y}$ to be the Chebyshev nodes because they minimize Runge's phenomenon (Epperson 1987).

Evaluating Equation (18) at each of the points $\{y_0, \ldots, y_M\}$ and re-writing Equation (18) in matrix form gives

$$\mathbf{c}_{k+1}\mathbf{W}_{k+1}\mathbf{R}_{k+1}(\mathbf{y}) = \mathbf{c}_k\mathbf{E}_k(t_k - t_{k-1})\mathbf{W}_k\mathbf{R}_k(\mathbf{y})\mathbf{H}_{k,k+1}(\mathbf{y}), \tag{19}$$

where the quantities in Equation (19) are given by

$$[\mathbf{R}_k(\mathbf{y})]_{i,j} = R_i^{(\alpha_k,\beta_k)}(y_j), \tag{20}$$

$$[\mathbf{E}_k(\tau)]_{i,j} = \delta_{i,j} e^{-\lambda_{k,i}\tau/(2N_k)}, \tag{21}$$

$$[\mathbf{W}_k]_{i,j} = w_{k,i,j}, \tag{22}$$

$$[\mathbf{H}_{k,k+1}(\mathbf{y})]_{i,j} = \delta_{i,j} \frac{\pi_k(y_i)e^{-\bar{\sigma}_k(y_i)/2}}{\pi_{k+1}(y_i)e^{-\bar{\sigma}_{k+1}(y_i)/2}}, \tag{23}$$

and $\mathbf{c}_k = (c_{k,0}, c_{k,1}, \ldots)$. In Equations (21) and (23), $\delta_{i,j}$ is the Kronecker delta function satisfying $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$, otherwise. The coefficients $\mathbf{c}_{k+1}$ in epoch $k+1$ are obtained from the coefficients $\mathbf{c}_k$ in epoch k by solving Equation (19) for $\mathbf{c}_{k+1}$ using standard approaches for solving linear systems.

The generalization, $\phi_K(t;y)$, of the TDF at time $T = \sum_{k=1}^{K} \tau_k$ is evaluated by iteratively solving Equation (19) to obtain $\mathbf{c}_K$ in the final epoch, K, starting from a set of initial coefficients, $\mathbf{c}_1$. The final value of $\phi_K(t;y)$ at time T is then computed as

$$\phi_K(T;y) \approx \sum_{n=0}^{N} c_{K,n} e^{-\lambda_{K,n}(T - t_{K-1})/(2N_K)} \pi_K(y) B_{K,n}(y). \tag{24}$$

Equation (24) can be used to compute $\phi_K(t;y)$, at any time $t \in [0,T]$ by defining $T = t$, and choosing K to be the interval such that $t \in (t_{K-1}, t_K]$.

3. Initial conditions

By fitting the coefficients, $\{c_{k,n}\}_{n=0}^{\infty}$, in Equation (24) to different initial distributions of frequencies at time $t = 0$, we can obtain the multi-epoch TDF, along with generalizations of the multi-epoch

TDF to other initial distributions. Formulas for the starting coefficients $\mathbf{c}_1 = (c_{1,0}, c_{1,1}, \ldots)$ were presented in Steinrücken *et al.* (2014) for three different initial conditions: mutation drift balance, mutation selection balance, and the case of an initial starting frequency, x_0. For completeness, these formulas are presented again below.

3.1. **Initial frequency.** When the initial condition is a specified frequency, x_0, the initial density is $\rho_1(x) = \delta(x - x_0)$, where $\delta(\cdot)$ is the Dirac delta distribution. Steinrücken *et al.* (2014) showed that this choice of initial conditions gives rise to the transition density function $p_1(t; x_0, y)$, where the coefficients $\{c_{1,n}\}_{n=0}^{\infty}$ are given by

$$c_{1,n} = \frac{\langle \pi_1 B_{1,n}, \rho_k \rangle_{1/\pi_1}}{\langle \pi_1 B_{1,n}, \pi_1 B_{1,n} \rangle_{1/\pi_1}} = \frac{B_{1,n}(x_0)}{r_{1,n}}, \tag{25}$$

where

$$r_{1,n} = \sum_{m=0}^{\infty} w_{1,n,m}^2 d_m^{(\alpha_1,\beta_1)}, \tag{26}$$

and

$$d_m^{(\alpha,\beta)} = \frac{\Gamma(m+\alpha)\Gamma(m+\beta)}{(2m+\alpha+\beta-1)\Gamma(m+\alpha+\beta-1)\Gamma(m+1)}. \tag{27}$$

3.2. **Mutation-selection balance.** Under the initial condition of mutation-selection balance, the initial density is the normalized stationary density

$$\pi_1(y)/C_{\pi_1}, \tag{28}$$

where $\pi_1(y)$ is given in Equation (6) and C_{π_1} is a normalizing constant defined such that $\int_0^\infty \pi_1(y)/C_{\pi_1} dy = 1$. Using this initial distribution, Steinrücken *et al.* (2014) showed that the initial coefficients are given by

$$c_{1,n} = \frac{B_{1,0}(0)}{r_{1,0}} \delta_{0,n}, \tag{29}$$

where $r_{1,0}$ is given by Equation (26). The numerator in Equation (29) is given by

$$B_{1,0}(0) = \sum_{m=0}^{\infty} (-1)^m w_{1,0,m} \frac{\Gamma(m+\alpha_1)}{\Gamma(m+1)\Gamma(\alpha_1)}. \tag{30}$$

3.3. **Mutation-drift balance.** Finally, under the initial condition of mutation-drift balance, the initial density is given by

$$\frac{y^{\alpha_1-1}(1-y)^{\beta_1-1}}{B(\alpha_1,\beta_1)}, \tag{31}$$

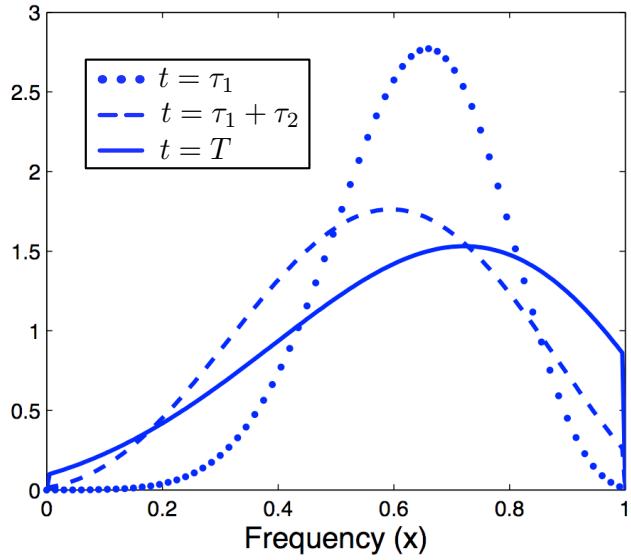

FIGURE 2. Plot of the TDF for the model shown in Figure 1 with the parameters specified in the example in Section 4, evaluated at the times t_1, t_2, and T.

where $B(\alpha, \beta)$ is the beta function. Steinrücken *et al.* (2014) showed that the initial coefficients in the case of mutation-drift balance are given by

$$c_{1,n} = \frac{1}{d_{1,n} B(\alpha_1, \beta_1) B^-_{1,0}(0)} \sum_{m=0}^{\infty} w_{1,n,m} w^-_{1,0,m} d_m^{(\alpha_1,\beta_1)}, \tag{32}$$

where the values $w^-_{k,n,m}$ are the entries of the nth left eigenvector of the matrix $\mathbf{M}_k$ with $\sigma_{1,1}$ and $\sigma_{1,2}$ replaced by $-\sigma_{1,1}$ and $-\sigma_{1,2}$, respectively, and $B^-_{1,0}(x)$ is the corresponding eigenfunction.

4. Implementation

Our algorithm has been implemented in JAVA. The inputs to the program are the effective population sizes (number of diploid individuals) $\boldsymbol{N} = (N_1, \ldots, N_K)$; epoch durations $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$; per-generation mutation rates $\boldsymbol{a} = (a_1, \ldots, a_K)$ and $\boldsymbol{b} = (b_1, \ldots, b_K)$; selection parameters $\boldsymbol{s}_1 = (s_{11}, \ldots, s_{K1})$ and $\boldsymbol{s}_2 = (s_{12}, \ldots, s_{K2})$; initial allele frequency X_0; and the time $t \in [0, T]$ at which the TDF will be evaluated. A plot of the TDF evaluated at each epoch boundary point ($t = \tau_1, \tau_1 + \tau_2$, and T) in Figure 1 is shown in Figure 2. The full command options are detailed in the user manual distributed with the software.

The approach described in Section 2.3 for knitting together transition densities across epochs, combined with the method of Song and Steinrücken (2012) for computing the eigenfunctions (Section 2.2.1), produces a computationally efficient method for computing $p_K(t; x, y)$ and $\phi_K(t; y)$. Table 1 shows the runtime of our method, SpectralTDF, for different numbers of epochs, and for different values of the parameters that control the precision of the eigenvalue computations.

High precision computations are sometimes required when selection coefficients are large and waiting times between sampling events are short. However, such high precision computations

Table 1. Runtime (in seconds) of the SpectralTDF algorithm for different numbers of epochs.

	Max	Number of Epochs									
Precision	Cutoff	1	2	3	4	5	6	7	8	9	10
60	150	35.8	43.9	51.1	54.4	60.3	67.3	71.9	78.6	83.4	90.7
80	250	108.3	156.6	198.0	223.7	260.3	300.5	339.0	365.4	404.1	443.3
100	350	271.5	444.9	563.6	775.2	928.9	896.7	994.7	1114.3	1241.0	1332.0

are often unnecessary. Table 1 shows that SpectralTDF can be used to compute the TDF in a population with ten epochs in under two minutes, and for scenarios requiring higher precision in under ten minutes.

5. Discussion

Our implementation provides a fast and numerically stable method for computing the TDF for a general model with piecewise-constant population sizes and a broad range of time-varying mutation and selection parameters. It also allows for a variety of initial conditions, including a specified initial frequency and stationary distributions under mutation-selection balance or mutation-drift balance.

The JAVA implementation is designed to be used either as a stand-alone application or in combination with other methods. For example, the code can be easily incorporated into the method of Steinrücken *et al.* (2014), allowing the inference of selection parameters from time series data sampled from populations with time-varying demographic and selection parameters. In general, the method we present provides a flexible and efficient tool for studying the evolution of allele frequencies over time under complex evolutionary scenarios.

Acknowledgement

We thank Daniel Živković and Anand Bhaskar for helpful discussions and collaboration on earlier related work. This research is supported in part by NIH grants R01-GM094402 (MS, YSS) and R01-GM109454 (EMJ), and by a Packard Fellowship for Science and Engineering (YSS).

References

Barbour, A., Ethier, S., and Griffiths, R. (2000). A transition function expansion for a diffusion model with selection. *Annals of Applied Probability*, pages 123–162.

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

Epperson, J. (1987). On the Runge example. *American Mathematical Monthly*, **94**(4), 329–341.

Ewens, W. (2004). *Mathematical Population Genetics: I, 2nd ed.* Springer.

Griffiths, R. (1979). A transition density expansion for a multi-allele diffusion model. *Advances in Applied Probability*, pages 310–325.

Kimura, M. (1955). Stochastic processes and distribution of gene frequencies under natural selection. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 20, pages 33–53. Cold Spring Harbor Laboratory Press.

Kimura, M. (1957). Some problems of stochastic processes in genetics. *The Annals of Mathematical Statistics*, pages 882–901.

Shimakura, N. (1977). Equations différentielles provenant de la génétique des populations. *Tohoku Mathematical Journal, Second Series*, **29**(2), 287–318.

Siepielski, A., DiBattista, J., and Carlson, S. (2009). Its about time: the temporal dynamics of phenotypic selection in the wild. *Ecology Letters*, **12**(11), 1261–1276.

Song, Y. S. and Steinrücken, M. (2012). A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics*, **190**(3), 1117–1129.

Steinrücken, M., Wang, Y., and Song, Y. S. (2013). An explicit transition density expansion for a multi-allelic Wright-Fisher diffusion with general diploid selection. *Theor. Popul. Biol.*, **83**, 1–14.

Steinrücken, M., Bhaskar, A., and Song, Y. S. (2014). A novel spectral method for inferring general diploid selection from time series genetic data. *Annals of Applied Statistics*, **8**(4), 2203–2222.

Vogl, C. (2014a). Biallelic mutation-drift diffusion in the limit of small scaled mutation rates. *arXiv*, page 1409.2299.

Vogl, C. (2014b). Computation of the likelihood in biallelic diffusion models using orthogonal polynomials. *Computation*, **2**, 199–220.

Živković, D., Steinrücken, M., Song, Y. S., and Stephan, W. (2015). Transition densities and sample frequency spectra of diffusion processes with selection and variable population size. *Genetics*, **doi:10.1534/genetics.115.175265**.

Waxman, D. (2011). A unified treatment of the probability of fixation when population size and the strength of selection change over time. *Genetics*, **188**, 907–913.

Williamson, S., Hernandez, R., Fledel-Alonl, A., Zhu, L., Nielsen, R., and Bustamante, C. (2004). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Nat. Acad. Sci.*, **102**(22), 7882–7887.