A Simple Method for Finding Explicit Analytic Transition Densities of Diffusion Processes with General Diploid Selection

Yun S. Song*,†,1 and Matthias Steinrücken*
*Department of Statistics, University of California, Berkeley, California 94720 and †Computer Science Division, University of California, Berkeley, California 94720

ABSTRACT The transition density function of the Wright–Fisher diffusion describes the evolution of population-wide allele frequencies over time. This function has important practical applications in population genetics, but finding an explicit formula under a general diploid selection model has remained a difficult open problem. In this article, we develop a new computational method to tackle this classic problem. Specifically, our method explicitly finds the eigenvalues and eigenfunctions of the diffusion generator associated with the Wright–Fisher diffusion with recurrent mutation and arbitrary diploid selection, thus allowing one to obtain an accurate spectral representation of the transition density function. Simplicity is one of the appealing features of our approach. Although our derivation involves somewhat advanced mathematical concepts, the resulting algorithm is quite simple and efficient, only involving standard linear algebra. Furthermore, unlike previous approaches based on perturbation, which is applicable only when the population-scaled selection coefficient is small, our method is nonperturbative and is valid for a broad range of parameter values. As a by-product of our work, we obtain the rate of convergence to the stationary distribution under mutation–selection balance.

DIFFUSION processes, which are continuous-time Markov processes with almost surely continuous sample paths, have been successfully applied in various population genetic analyses in the past. Examples include finding the stationary distribution of allele frequencies and approximating fixation times and probabilities (see Karlin and Taylor 1981; Ewens 2004; Durrett 2008 for other applications of diffusion processes). This success is largely due to the fact that the diffusion approximation captures the key features of an evolutionary model while ignoring unimportant details, thereby arriving at a simpler process that facilitates computation. However, when a reasonably complex model of evolution is considered, one is faced with unwieldy equations even under the diffusion approximation. In particular, for Wright–Fisher diffusions with general diploid selection, finding an explicit analytic transition density function, which characterizes the evolution of population-wide allele frequencies over time, has remained a challenging open problem. The diffusion theory allows one to write down a partial differential equation (PDE) satisfied by the transition density, but solving the PDE analytically has proved to be difficult.

The transition density has several practical applications, including the following: Recently, there has been growing interest in analyzing samples taken from the same or related populations at different time points. For example, such data arise from experimental evolution of model organisms in the laboratory (*e.g.*, bacteria, see Lenski 2011), from viral/phage populations (Shankarappa *et al.* 1999; Wichman *et al.* 1999), or from ancient DNA (Hummel *et al.* 2005); see also Bollback *et al.* (2008) and references therein. In particular, the recent sequencing of Neanderthal (Green *et al.* 2010) and Denisova (Reich *et al.* 2010) genomes should provide new opportunities for studying the evolution of allele frequencies over time, possibly under the influence of natural selection. In applying the diffusion process to study the evolution of the populations underlying such samples, it is important to find the transition density accurately. Bollback *et al.* (2008) analyzed samples from multiple time points by using a hidden Markov model in which the hidden states are the population-wide allele frequencies. To approximate the evolution of the allele frequencies, they applied

doi: 10.1534/genetics.111.136929
Manuscript received November 16, 2011; accepted for publication December 20, 2011
[1]Corresponding author: University of California, EECS Department, 683 Soda Hall #1776, Berkeley, CA 94720-1776. E-mail: yss@cs.berkeley.edu

finite difference methods to obtain approximate numerical solutions to the PDE satisfied by the transition density. Finite difference methods were also employed by Gutenkunst *et al.* (2009) to obtain numerical approximations of the transition density in Wright–Fisher diffusions with population substructure, which the authors applied to develop a useful tool for demographic inference. When employing such numerical methods, however, one needs to exercise caution in choosing appropriate discretization grid points. Which discretization is appropriate may depend strongly on the parameters (*e.g.*, the selection coefficient) of the model, and it is difficult to predict *a priori* whether a particular discretization will produce accurate solutions. Also, in a numerical approach, note that the PDE needs to be solved afresh if the initial or the final frequency is changed. It would be useful to have a solution that is analytic in those variables.

Since the transition density of the Wright–Fisher diffusion with selection has practical applications, finding an explicit formula has significant merit and several researchers have considered the problem. As detailed later in the text, the so-called spectral representation of the transition density can be found if the eigenvalues and eigenfunctions of the diffusion generator are known. Indeed, this is the approach taken by Kimura (1955a, 1957), first for a diallelic model with genic selection (*i.e.*, the case with the dominance parameter $h = 1/2$, as described shortly) and later for the case of complete dominant selection (*i.e.*, with $h = 1$), assuming no recurrent mutation in both cases. More precisely, Kimura proposed perturbation expansions of the eigenvalues and eigenfunctions in powers of the population-scaled selection coefficient σ (defined more precisely later). Although this method is valid for $\sigma < 1$, the expansions fail to converge for σ substantially >1 [say, $\sigma > 10$, which is not so unusual for adaptive alleles (Eyre-Walker and Keightley 2007)]. Furthermore, the perturbation expansion scheme described in Kimura's work is not entirely transparent.

For a *neutral* parent-independent mutation model, an explicit spectral representation of the transition density for the one-locus Wright–Fisher diffusion has been known for some time (Shimakura 1977; Griffiths 1979). Griffiths and Li (1983) and Tavaré (1984) showed that this spectral representation can be interpreted in terms of a stochastic process dual to the diffusion. The time dependency of the transition density is solely given through the probability distribution of this dual process (see Griffiths and Spanò 2010 for an overview). Barbour *et al.* (2000) extended this duality approach to include a general selection model, but the transition rates of the dual process depend on the moments of the stationary distribution, and under selection these moments are difficult to compute (Donnelly *et al.* 2001). Hence, while being of theoretical interest, their method does not readily lead to efficient computation of the transition density.

In this article, we develop a new, simple computational method with which to find analytic transition density functions of diallelic Wright–Fisher diffusions under recurrent mutation and arbitrary diploid selection. In contrast to the aforementioned mathematical work based on duality, our method explicitly finds the eigenvalues and eigenfunctions of the diffusion generator associated with the diffusion, thus leading to an explicit spectral representation of the transition density function. Specifically, the eigenfunctions are found as a series of orthogonal functions. Although somewhat advanced mathematical concepts are needed to derive the necessary system of equations, the resulting algorithm is quite simple to describe and easy to implement, involving only standard linear algebra. Furthermore, unlike previous approaches (Kimura 1955a, 1957) based on perturbation, which is applicable only when the population-scaled selection coefficient σ is small, our method is nonperturbative and is valid for a broad range of parameter values, including large values of σ and an arbitrary dominance parameter h. As an application of our work, we obtain the rate of convergence to the stationary distribution under mutation–selection balance.

The rest of this article is organized as follows. We begin with a brief review of the Wright–Fisher diffusion and describe the notion of spectral representation. Orthogonal polynomials, which we extensively employ in our work, are also introduced. Then, we illustrate the key ideas behind our method in the simple case of genic selection and no recurrent mutation. Afterward, we apply our method to the general case of arbitrary diploid selection and recurrent mutation and show how the results for the no-mutation case can be recovered as a special case. We then assess the performance of our method and end with discussions on possible applications and extensions.

Background

In this section, we review useful facts about diffusion processes. In particular, we highlight some key properties satisfied by backward generators of one-dimensional diffusions. We also introduce the relevant orthogonal polynomials that we utilize in our method.

Wright–Fisher diffusions

We consider a Wright–Fisher diffusion process with two alleles, denoted A_0 and A_1. The population-wide frequency of A_1 is denoted by x; hence, the frequency of A_0 is $1 - x$. The genotype fitness scheme considered in this article is as follows:

$$\begin{array}{rccc} \text{Genotype :} & A_0/A_0 & A_0/A_1 & A_1/A_1 \\ \text{Relative fitness :} & 1 & 1+2\,h\,s & 1+2\,s \end{array}$$

We refer to the case with the dominance parameter $h = 1/2$ as *genic* selection. The population-scaled selection coefficient is defined as $\sigma = 2Ns$, where N corresponds to the diploid population size, which is assumed to remain constant over time. The rate of mutation from A_0 to A_1 is given by $\alpha = 4Nu_{01}$ and from A_1 to A_0 by $\beta = 4Nu_{10}$, where u_{01} (respectively, u_{10}) denotes the per-generation probability of mutation from A_0 to A_1 (respectively, from A_1 to A_0).

Note that the genotype fitness scheme introduced above does not include the case in which the homozygotes have a relative fitness of 1 and the heterozygote has a relative fitness unequal to 1. However, by choosing s close to zero and h large, we can mimic such a scheme in our framework. More generally, it is straightforward to apply the technique developed in this article to a selection scheme in which the heterozygote has relative fitness $1 + s_1$ and the homozygote A_1/A_1 has relative fitness $1 + s_2$. However, to conform to the convention widely adopted in the literature, we use the above-mentioned parameterization of relative fitnesses.

Throughout, we use f to denote a twice continuously differentiable bounded function over [0,1]. The backward generator $\mathscr{L}$ of a one-dimensional diffusion process on [0,1] with diffusion coefficient $\nu^2(x)$ and drift coefficient $\mu(x)$ acts on f as

$$\mathscr{L}f(x) = \frac{1}{2}\nu^2(x)\frac{\partial^2}{\partial x^2}\{f\ (x)\} + \mu(x)\frac{\partial}{\partial x}\{f\ (x)\}.$$

In the Wright–Fisher diffusion, $\nu^2(x) = x(1 - x)$. The contribution to $\mu\ (x)$ from selection is

$$2\sigma x\ (1-x)\ [x + h\ (1-2\,x)],$$

while the contribution from recurrent mutation is

$$\tfrac{1}{2}\,[\alpha\ (1-x) - \beta\ x].$$

See Ewens (2004, Chap. 5.1) for a more detailed description.

Self-adjointness and the spectrum of a generator

Let $L^2([0,1],\rho)$ denote the space of real-valued functions on [0,1] that are square integrable with respect to some real positive density $\rho(x)$. We refer to ρ as the weight function. Define the inner product $\langle\cdot,\cdot\rangle_\rho$ as

$$\langle f, g\rangle_\rho = \int_0^1 f(x)g(x)\rho(x)\mathrm{d}x, \qquad (1)$$

for $f,\ g \in L^2([0,1],\ \rho)$.

For a diffusion process with diffusion coefficient $\nu^2(x)$ and drift coefficient $\mu(x)$, the *scale* density $\xi(x)$ is defined as

$$\xi(x) = \exp\left[-\int_{x_0}^{x}\frac{2\ \mu\ (z)}{\nu^2\ (z)}\,d\,z\right], \qquad (2)$$

and the *speed* density $\pi(x)$ is defined as

$$\pi\ (x) = \frac{\gamma}{\nu^2\ (x)\ \xi\ (x)}, \qquad (3)$$

where γ is some positive constant and x_0 is an arbitrary state in [0,1]. For the results derived in this article it is crucial to establish that $\mathscr{L}$ is *self-adjoint* with respect to π. To this end, let $f,\ g \in L^2([0,1],\ \pi)$ satisfy appropriate boundary conditions relevant to the boundary behavior of the corresponding diffusion. The diffusions considered in this article exhibit exit, regular reflecting, or entrance boundaries. If 0 is an exit boundary, then the appropriate boundary condition is $\mathrm{Lim}_{x\downarrow 0}\ f(x)\ =\ 0$. If 0 is either a regular reflecting or an entrance boundary, the appropriate boundary condition is $\lim_{x\downarrow 0}(1/\xi(x))(df\,(x)/dx) = 0$. Similar boundary conditions apply as $x\uparrow 1$. See Durrett (2008) or Ewens (2004) for more details. For the diffusions considered in this article, their corresponding boundary conditions and integration by parts imply

$$\langle\mathscr{L}\,f, g\rangle_\pi\ =\ \langle f, \mathscr{L}\,g\rangle_\pi,$$

thus establishing that $\mathscr{L}$ is self-adjoint.

The key property (known as the spectral theorem) that we utilize in our work is the following: Suppose B and B' are eigenfunctions of $\mathscr{L}$ that satisfy the requisite boundary conditions of the diffusion process. If their eigenvalues Λ and Λ' are distinct, then the self-adjointness of $\mathscr{L}$ (*i.e.*, $\langle B, \mathscr{L}\,B'\rangle_\pi = \langle\mathscr{L}\,B, B'\rangle_\pi$) implies $\langle B, B'\rangle_\pi = 0$. Hence, eigenfunctions of $\mathscr{L}$ with distinct eigenvalues are orthogonal with respect to the weight function $\pi(x)$.

That $\mathscr{L}$ is a self-adjoint negative semidefinite differential operator implies that its eigenvalues are all real and nonpositive. Furthermore, for many boundary conditions, including the ones considered in this article, solutions of $\mathscr{L}\ B\ (x) = -\Lambda\ B\ (x)$ satisfying the requisite boundary conditions exist for countably many distinct values of Λ. Thus, for the diffusion processes considered in this article, there is a unique sequence

$$0 \le \Lambda_0 < \Lambda_1 < \Lambda_2 < \cdots,$$

with $\Lambda_n\ \to\ \infty$ as $n\ \to\ \infty$ (Karlin and Taylor 1981, Chap. 15.13). These eigenvalues $\{-\Lambda_n\}_{n=0}^{\infty}$ are called the "spectrum" of $\mathscr{L}$, and it can be shown that their associated eigenfunctions $\{B_n\ (x)\}_{n=0}^{\infty}$, which satisfy

$$\mathscr{L}\ B_n\ (x) = -\ \Lambda_n\ B_n\ (x),$$

form a basis of $L^2([0,1],\ \pi)$.

Spectral representation of the transition density

For any subset $S \subset [0,1]$, the transition density function of a diffusion process is the function p: $\mathbb{R}_{\ge 0} \times [0,1] \times [0,1] \to \mathbb{R}_{\ge 0}$ such that

$$\mathbb{P}[X_t \in S\,|\,X_0 = x] = \int_S p(t;x,y)dy.$$

The transition density $p(t;\ x,\ y)$ satisfies the Kolmogorov backward equation

$$\begin{aligned}\frac{\partial p\ (t;x,y)}{\partial t} &= \mathscr{L}\ p\ (t;x,y)\\ &= \frac{1}{2}\ \nu^2\ (x)\ \frac{\partial^2}{\partial x^2}\ \{p\ (t;x,y)\}\\ &\quad + \mu\ (x)\ \frac{\partial}{\partial x}\ \{p\ (t;x,y)\},\end{aligned}$$

and the appropriate boundary conditions, see Karlin and Taylor (1981, Chap. 15.5). Here, the differential operator $\mathscr{L}$ is the backward generator of the diffusion and it acts on x.

Let $\{B_n(x)\}$ be the eigenfunctions of $\mathscr{L}$ that satisfy the proper boundary conditions of the diffusion process. Further, let $-\Lambda_n$ denote the eigenvalue of $B_n(x)$. Then, $\phi_n\,(t,x) = e^{-\Lambda_n\,t}\,B_n\,(x)$ satisfies the partial differential equation

$$\frac{\partial\phi_n\,(t,x)}{\partial t} = \mathscr{L}\,\phi_n\,(t,x), \tag{4}$$

and the requisite boundary conditions. Furthermore, since $\mathscr{L}$ is a linear differential operator, a linear combination of $e^{-\Lambda_n t}B_n(x)$ is also a solution to (4). The spectral representation of $p(t;\,x,\,y)$ is given by

$$p\,(t;x,y) = \sum_{n=0}^{\infty} c_n\,(y)\,e^{-\Lambda_n\,t}\,B_n\,(x),$$

where the coefficients $c_n(y)$ depend on y and are set to satisfy the initial condition. For $p(0;\,x,\,y) = \delta(x - y)$, the Dirac-delta distribution, we obtain

$$p(t;x,y) = \sum_{n=0}^{\infty} e^{-\Lambda_n\,t}\,\pi\,(y)\frac{B_n(x)B_n(y)}{\langle B_n, B_n\rangle_\pi}, \tag{5}$$

where π is the speed density defined in (3) and $\langle\cdot,\cdot\rangle_\pi$ is the inner product defined in (1). See Karlin and Taylor (1981, Chap. 15.13) for further details and examples.

In summary, the transition density function of a diffusion process can be determined if the eigenvalues and the eigenfunctions of $\mathscr{L}$ are known. The orthogonal polynomials described in the following two subsections are such eigenfunctions for certain neutral Wright–Fisher diffusion processes, and we make extensive use of them in our work to solve the eigenvalue problem in the presence of selection.

In practice, we do not need to sum over infinitely many terms in (5). Since $\Lambda_n \to \infty$ as $n \to \infty$, the exponential term $e^{-\Lambda_n\,t}$ will be negligibly small for n sufficiently large. Hence, we can obtain accurate approximations of $p(t;\,x,\,y)$ for $t > 0$ by summing over n from 0 to some reasonable finite cutoff. In *Empirical transition densities and stationary distributions* and *Rate of convergence to the stationary distribution* we provide explicit examples illustrating this property.

Jacobi polynomials

An excellent treatise on orthogonal polynomials can be found in Szegö (1939) and a concise collection of related formulas can be found in Abramowitz and Stegun (1965, Chap. 22). Here, we briefly review some key facts about a particular type of classical orthogonal polynomials.

For $z \in [-1,1]$, the Jacobi polynomials $P_n^{(a,b)}\,(z)$ satisfy the differential equation

$$\begin{aligned}(1 - z^2)\frac{d^2f\,(z)}{dz^2} + [b - a - (a + b + 2)z]\;\frac{df\,(z)}{dz}\\ + n\,(n + a + b + 1)\,f\,(z) = 0.\end{aligned} \tag{6}$$

For fixed $a,\,b > -1$, $\{P_n^{(a,b)}\,(z)\}$ form an orthogonal system with respect to the weight function $(1 - z)^a\,(1 + z)^b$ on the interval $[-1,1]$. Since the domain and the parameters of $P_n^{(a,b)}\,(z)$ are not suitable for our purpose, we define the following modified Jacobi polynomials, for $x \in [0,1]$ and $a,\,b > 0$:

$$R_n^{(a,b)}\,(x) = P_n^{(b-1,a-1)}\,(2x - 1).$$

Griffiths and Spanò (2010) use a slightly different, although related, convention.

For $x \in [0,1]$, the modified Jacobi polynomials $R_n^{(a,b)}\,(x)$ satisfy the differential equation

$$\begin{aligned}x\,(1 - x)\;\frac{d^2\,f\,(x)}{dx^2} + [a - (a + b)\,x]\;\frac{df\,(x)}{dx}\\ + n\,(n + a + b - 1)\,f\,(x) = 0,\end{aligned} \tag{7}$$

which follows immediately from (6). For fixed $a,\,b > 0$, $\{R_n^{(a,b)}\,(x)\}$ is an orthogonal system with respect to the weight function $x^{a-1}(1 - x)^{b-1}$ on $[0,1]$. More precisely,

$$\int_0^1 R_n^{(a,b)}\,(x)\,R_m^{(a,b)}(x)\,x^{a-1}(1{-}x)^{b-1}\,dx = \delta_{n,m}\Delta_n\,(a,b), \tag{8}$$

where $\delta_{n,m}$ denotes the Kronecker delta and the coefficient $\Delta_n(a,b)$ is defined as

$$\Delta_n\,(a,b) = \frac{\Gamma(n + a)\Gamma(n + b)}{(2\,n + a + b - 1)\,\Gamma\,(n + a + b - 1)\,\Gamma\,(n + 1)}. \tag{9}$$

Furthermore, $\{R_n^{(a,b)}\,(x)\}$ form a complete basis of the Hilbert space $L^2([0,1],\,x^{a-1}(1 - x)^{b-1})$.

For $n \geq 1$, it can be shown that $R_n^{(a,b)}\,(x)$ satisfies the recurrence relation

$$\begin{aligned}x\,R_n^{(a,b)}\,(x) = {}&\frac{(n + a - 1)\,(n + b - 1)}{(2\,n + a + b - 1)\,(2\,n + a + b - 2)}\;R_{n-1}^{(a,b)}\,(x)\\ &+ \left[\frac{1}{2} - \frac{b^2 - a^2 - 2\,(b - a)}{2\,(2\,n + a + b)\,(2\,n + a + b - 2)}\right]R_n^{(a,b)}\,(x)\\ &+ \frac{(n + 1)\,(n + a + b - 1)}{(2\,n + a + b)\,(2\,n + a + b - 1)}\;R_{n+1}^{(a,b)}\,(x),\end{aligned} \tag{10}$$

while, for $n = 0$,

$$x\,R_0^{(a,b)}\,(x) = \frac{a}{a + b}\;R_0^{(a,b)}(x) + \frac{1}{a + b}\;R_1^{(a,b)}\,(x). \tag{11}$$

Also, note that $R_0^{(a,b)}\,(x)\equiv 1$. The above recurrence relations plays an important role in our work.

Gegenbauer polynomials

The classical Gegenbauer polynomials are a special case of the classical Jacobi polynomials, namely $P_n^{(1,1)}\ (2x-1)$. In our work, we define $G_n(x)$ as

$$G_n\ (x) = -x\ (1-x)\ P_n^{(1,1)}\ (2x-1) = -x\ (1-x)\ R_n^{(2,2)}\ (x)$$

and refer to them as *modified* Gegenbauer polynomials. The minus sign will prove convenient later. Using (7), it can be shown that $G_n(x)$ satisfies the differential equation

$$x\ (1-x)\ \frac{d^2 f\ (x)}{dx^2} + (n+2)\ (n+1)\, f\ (x) = 0. \qquad (12)$$

Further, $\{G_n(x)\}$ form an orthogonal system of polynomials with respect to the weight function $x^{-1}(1-x)^{-1}$:

$$\int_0^1 G_n\ (x)\ G_m\ (x)\ x^{-1}\ (1-x)^{-1}\ dx = \delta_{n,m}\ \frac{n+1}{(n+2)\ (2n+3)}.$$

Using the completeness of the Jacobi polynomials, it can be shown that $\{G_n(x)\}$ form a complete basis of $L^2([0,1], x^{-1}(1-x)^{-1})$.

For $n \geq 1$, $G_n(x)$ satisfies the recurrence relation

$$\begin{aligned} xG_n\ (x) = {} & \frac{n+1}{2\ (2n+3)}\ G_{n-1}\ (x) + \frac{1}{2}\ G_n\ (x) \\ & + \frac{(n+1)\ (n+3)}{2(n+2)(2n+3)}\ G_{n+1}\ (x), \end{aligned} \qquad (13)$$

while, for $n = 0$,

$$x\ G_0\ (x) = \frac{1}{2}\ G_0\ (x) + \frac{1}{4}\ G_1\ (x).$$

These relations follow from (10) and (11). Furthermore, we have $G_0(x) \equiv -x(1-x)$.

Diffusions with Genic Selection and No Mutation

As described earlier, to obtain the spectral representation of $p\ (t;\ x,\ y)$, we need to solve the eigenvalue problem for the diffusion generator. In this section, we illustrate the key ideas underlying our method by considering the simple case of no mutation and genic selection ($h = 1/2$), in which case the involved algebra simplifies significantly. Incidentally, the genic selection case has been considered by many other researchers in the past; for example, see Kimura (1955a, 1957), Etheridge and Griffiths (2009), and Griffiths (2003). The modified Gegenbauer polynomials introduced above will play an important role in this section. The case with both recurrent mutation and general diploid selection (*i.e.*, h not necessarily equal to 1/2) is addressed in the next section.

Description of the main idea

Let $\mathscr{L}_0$ denote the diffusion part of the backward generator:

$$\mathscr{L}_0 f\ (x) = \frac{1}{2}\ x\ (1-x)\ \frac{\partial^2}{\partial x^2}\ \{f\ (x)\}. \qquad (14)$$

As is well known (Kimura 1955a,b, 1957; Karlin and Taylor 1981), it follows from Equation 12 that the modified Gegenbauer polynomials $G_n(x)$ are eigenfunctions of $\mathscr{L}_0$,

$$\mathscr{L}_0 G_n\ (x) = -\lambda_n\ G_n\ (x),$$

where

$$\lambda_n = \binom{n+2}{2}.$$

With genic selection, the full backward generator is

$$\mathscr{L} f\ (x) = \frac{1}{2}\ x\ (1-x)\ \frac{\partial^2}{\partial x^2}\ \{f\ (x)\} + \sigma\ x\ (1-x)\ \frac{\partial}{\partial x}\ \{f\ (x)\}. \qquad (15)$$

The speed density corresponding to this diffusion process is

$$\pi\ (x) = \frac{e^{2\,\sigma\, x}}{x\ (1-x)}, \qquad (16)$$

where we used $x_0 = 0$ and $\gamma = 1$ in (2) and (3), respectively.

Our goal is to find the eigenfunctions $B_n(x)$ and the associated eigenvalues $-\Lambda_n$ of the full generator $\mathscr{L}$:

$$\mathscr{L}\, B_n\ (x) = -\Lambda_n\ B_n\ (x). \qquad (17)$$

As discussed in *Background*, $\mathscr{L}$ is self-adjoint with respect to the weight function $\pi(x)$, which implies that its eigenfunctions $B_n(x)$ and $B_m(x)$, for $n \neq m$, are orthogonal with respect to $\pi(x)$; *i.e.*,

$$\int_0^1 B_n\ (x)\ B_m\ (x)\ \pi\ (x)\ dx \propto \delta_{n,m},$$

where $\pi(x)$ is shown in (16). In addition to the eigenfunctions, there may exist other sets of functions that are orthogonal with respect to the same weight function $\pi(y)$. For example, consider

$$H_n\ (x) = e^{-\sigma\, x}\ G_n\ (x). \qquad (18)$$

We can verify that $H_n(x)$ and $H_m(x)$, for $n \neq m$, are orthogonal with respect to the weight function $\pi(x)$:

$$\begin{aligned} \textstyle\int_0^1 H_n\ (x)\ H_m\ (x)\ \pi\ (x)\ \mathrm{d}x &= \textstyle\int_0^1 G_n\ (x)\ G_n\ (x)\ x^{-1}\ (1-x)^{-1}\ dx \\ &= \delta_{n,m}\ \frac{n+1}{(n+2)\ (2n+3)}. \end{aligned}$$

However, by directly applying $\mathscr{L}$ to $H_n(x)$, one can check that $H_n(x)$ are not eigenfunctions of $\mathscr{L}$. But, since both $\{H_n(x)\}$ and $\{B_n(x)\}$ are orthogonal with respect to the same weight function $\pi(x)$, and $\{H_n(x)\}$ form a basis of $L^2([0,1],\pi)$, we can represent $B_n(x)$ as a linear combination of $H_m(x)$,

$$B_n\ (x) = \sum_{m=0}^{\infty} u_{n,m}\ H_m\ (x), \tag{19}$$

where $u_{n,m}$ are constants to be determined. In the absence of mutation, states 0 and 1 are absorbing states (exit boundaries), so, as discussed in *Background*, $B_n(x)$ must satisfy the boundary conditions $\lim_{x\downarrow 0} B_n(x) = \lim_{x\uparrow 1} B_n(x) = 0$. Indeed, our proposed eigenfunctions satisfy those conditions since $H_m(0) = H_m(1) = 0$ for all $m \geq 0$.

Now, one can show

$$\begin{aligned} \mathscr{L} H_n\ (x) &= e^{-\sigma\, x}\ [\mathscr{L}_0\ G_n\ (x) - Q\ (x;\sigma)\ G_n\ (x)] \\ &= -\, e^{-\sigma\, x}\ [\lambda_n\ G_n\ (x) + Q\ (x;\sigma)\ G_n\ (x)], \end{aligned} \tag{20}$$

where

$$Q\ (x;\sigma) = \frac{1}{2}\,\sigma^2\, x\ (1-x). \tag{21}$$

For small σ, Kimura (1955a) employed an equation similar to (20) to obtain perturbation expansions in powers of σ for the eigenvalues and the eigenfunctions of the forward diffusion generator. Here, we proceed along a different avenue. The key difference is that our approach is nonperturbative and that it is valid for all parameter values.

Using (20) together with (17) and (19), we obtain

$$\sum_{m=0}^{\infty} u_{n,m}\ [\lambda_m + Q\ (x;\sigma)]\ G_m\ (x) = \Lambda_n\ \sum_{m=0}^{\infty} u_{n,m}\ G_m\ (x). \tag{22}$$

Now, for $m \geq 0$, (13) can be used to show

$$Q(x;\sigma) G_m(x) = a_m^{(-2)} G_{m-2}(x) + a_m^{(0)} G_m(x) + a_m^{(+2)} G_{m+2}(x),$$

where

$$\begin{aligned} a_m^{(-2)} &= -\sigma^2\ \frac{1}{8}\ \frac{m\ (m+1)}{(2m+1)\ (2m+3)}\ 1_{\{m\geq 2\}}, \\ a_m^{(0)} &= +\sigma^2\ \frac{1}{4}\ \frac{(m+1)\ (m+2)}{(2m+1)\ (2m+5)}, \\ a_m^{(+2)} &= -\,\sigma^2\ \frac{1}{8}\ \frac{(m+1)\ (m+4)}{(2m+3)\ (2m+5)}. \end{aligned} \tag{23}$$

In the first line of (23), $1_{\{Y\}}$ denotes an indicator function that is equal to 1 if statement Y is true or 0 otherwise. For a nonnegative integer k, multiplying (22) by $G_k(x)$ and integrating over [0,1] with respect to the weight function $x^{-1}(1-x)^{-1}$ yields

$$\lambda_k\ u_{n,k} + a_{k+2}^{(-2)}\ u_{n,k+2} + a_k^{(0)}\ u_{n,k} + a_{k-2}^{(+2)}\ u_{n,k-2} = \Lambda_n\ u_{n,k}, \tag{24}$$

where we define $a_{-2}^{(+2)} = a_{-1}^{(+2)} = 0$. Note that (24) specifies a linear system of equations with $u_{n,0}, u_{n,1}, u_{n,2}, \ldots$, as variables.

Algorithm 1 (genic selection)

The eigenvalues and eigenfunctions of the backward generator $\mathscr{L}$ (15) for the genic selection case can be obtained as follows. In matrix form, (24) can be written as

$$\begin{pmatrix} \lambda_0 + a_0^{(0)} & 0 & a_2^{(-2)} & 0 & 0 & \cdots \\ 0 & \lambda_1 + a_1^{(0)} & 0 & a_3^{(-2)} & 0 & \cdots \\ a_0^{(+2)} & 0 & \lambda_2 + a_2^{(0)} & 0 & a_4^{(-2)} & \cdots \\ 0 & a_1^{(+2)} & 0 & \lambda_3 + a_3^{(0)} & 0 & \cdots \\ 0 & 0 & a_2^{(+2)} & 0 & \lambda_4 + a_4^{(0)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} u_{n,0} \\ u_{n,1} \\ u_{n,2} \\ u_{n,3} \\ u_{n,4} \\ \vdots \end{pmatrix} = \Lambda_n \begin{pmatrix} u_{n,0} \\ u_{n,1} \\ u_{n,2} \\ u_{n,3} \\ u_{n,4} \\ \vdots \end{pmatrix}. \tag{25}$$

Let M denote the infinite-dimensional matrix on the left hand side of (25). The key fact is that the eigenvalues Λ_n of M correspond to the eigenvalues of $\mathscr{L}$ (up to a sign), and the associated eigenvectors $\boldsymbol{u}_n = (u_{n,0}, u_{n,1}, u_{n,2}, \ldots)$ of M determine the eigenfunctions of $\mathscr{L}$ via (19). Now, we consider a sequence of approximations by truncating (25). For a positive integer D, we let $M^{[D]}$ be an D-by-D matrix obtained by taking the first D rows and the first D columns of M, and let $\boldsymbol{u}_n^{(D)} = (u_{n,0}^{[D]}, u_{n,1}^{[D]}, \ldots, u_{n,D-1}^{[D]})$. Then, we approximate (25) by

$$M^{[D]}\ \boldsymbol{u}_n^{[D]} = \Lambda_n^{[D]}\ \boldsymbol{u}_n^{[D]}$$

and solve this finite-dimensional linear system to obtain eigenvalues $\Lambda_n^{[D]}$ and eigenvectors $\mathbf{u}_n^{[D]}$. This linear algebra problem can be easily solved using standard software packages such as Matlab, Mathematica, or the freely available LAPACK library (http://www.netlib.org/lapack/). We show in *Empirical Results* that $\Lambda_n^{[D]}$ and $u_{n,m}^{[D]}$ converge very rapidly as the truncation level D increases.

The eigenvectors $\mathbf{u}_n^{[D]}$ come in two types: Either $u_{n,m}^{[D]} = 0$ for all m even or $u_{n,m}^{[D]} = 0$ for all m odd. In fact, the linear system $M^{[D]}\ \mathbf{u}_n^{[D]} = \Lambda_n^{[D]}\ \mathbf{u}_n^{[D]}$ can be decomposed into two subsystems, one involving only the even rows and even columns of $M^{[D]}$ acting on $u_{n,m}$ for m odd, and the other involving only the odd rows and odd columns of $M^{[D]}$ acting on $u_{n,m}$ for m even. Hence, the eigenvalues and eigenvectors of $M^{[D]}$, for $D = 2D'$, can be determined by solving two D'-dimensional linear systems.

In the case of genic selection with no recurrent mutation, the eigenfunctions $B_n(x)$ of the backward generator $\mathscr{L}$ are also known as the oblate spheroidal functions in mathematical physics, and they have received considerable amounts of attention previously (*e.g.*, see Stratton *et al.* 1941). Note that the algorithm presented in this section provides an efficient way to evaluate these functions, a problem that remained difficult in the past.

Due to the exponential weighting factors in the speed density (16) and in the basis functions (18) for the eigenfunction expansion, evaluation of the transition density for large selection coefficients involves combining quantities with substantially different orders of magnitude. Thus, to obtain accurate numerical values of the transition density under strong selection, the coefficients $\mathbf{u}_n^{[D]}$ must be determined with high precision.

Diffusions with General Diploid Selection and Recurrent Mutation

In this section, we generalize the method developed in the previous section by incorporating recurrent mutation and general diploid selection into the diffusion process. The same overall strategy described above applies here as well. The main computational differences are that general diploid selection leads to more involved algebra and that, to handle recurrent mutation, we need to deal with general Jacobi polynomials instead of the modified Gegenbauer polynomials.

Neutral diffusion with recurrent mutation

For a neutral diallelic model with recurrent mutation, the backward generator $\mathscr{L}_0$ is given by

$$\mathscr{L}_0 f(x) = \frac{1}{2}\, x\,(1-x)\, \frac{\partial^2}{\partial x^2}\, \{f\,(x)\} + \frac{1}{2}\, [\alpha\,(1-x) - \beta\, x]\, \frac{\partial}{\partial x}\, \{f\,(x)\}. \tag{26}$$

See *Background* for the definitions of α and β. By appropriately choosing the constants x_0 and γ in (2) and (3), the speed density corresponding to this diffusion can be defined as

$$\pi_0\,(x) = x^{\alpha-1}\,(1-x)^{\beta-1}, \tag{27}$$

which is the unnormalized Beta distribution. It can be shown (see Karlin and Taylor 1981, Chap. 15.13, or compare with Equation 7) that the Jacobi polynomials $R_n^{(\alpha,\beta)}\,(x)$ are eigenfunctions of the backward generator $\mathscr{L}_0$ with eigenvalues $-\lambda_n^{(\alpha,\beta)}$, where

$$\lambda_n^{(\alpha,\beta)} = \frac{1}{2}\, n\,(n+\alpha+\beta-1). \tag{28}$$

Furthermore, the Jacobi polynomials $R_n^{(\alpha,\beta)}(x)$ form an orthogonal system with respect to the weight function $\pi_0(x)$. Under recurrent mutation, the diffusion exhibits either regular or entrance boundaries (*e.g.*, see Karlin and Taylor 1981, Chap. 15.6, Example 8). The respective conditions given in *Background* for $x = 0$ and $x = 1$ imply that the eigenfunctions $\varphi(x)$ of $\mathscr{L}_0$ need to satisfy

$$\lim_{x\downarrow 0} x^{\alpha}\, \frac{\partial}{\partial x}\, \{\phi\,(x)\} = 0 \quad \text{and} \quad \lim_{x\uparrow 1} (1-x)^{\beta}\, \frac{\partial}{\partial x}\, \{\phi\,(x)\} = 0,$$

and the modified Jacobi polynomials $R_n^{(\alpha,\beta)}(x)$ obey these conditions.

Adding general diploid selection

The backward generator of the diffusion process with recurrent mutation and general diploid selection is

$$\mathscr{L} f\,(x) = \mathscr{L}_0 f\,(x) + 2\,\sigma\, x\,(1-x)\,[x + h\,(1-2\,x)]\, \frac{\partial}{\partial x}\, \{f\,(x)\}, \tag{29}$$

where $\mathscr{L}_0 f(x)$ is the selectively neutral part shown in (26). With appropriate constants x_0 and γ in (2) and (3), the speed density for this diffusion can be defined as

$$\pi\,(x) = e^{\bar{\sigma}\,(x)}\, \pi_0\,(x), \tag{30}$$

where $\pi_0(x)$ is given in (27) and $\bar{\sigma}(x)$ is the mean fitness function given by

$$\bar{\sigma}\,(x) = 2\,h\,\sigma\cdot 2\,(1-x)\,x + 2\,\sigma\cdot x^2 = 2\,\sigma\, x\,[x + 2\,h\,(1-x)], \tag{31}$$

which simplifies to the linear function $2\sigma x$ for $h = 1/2$. The discussion in *Background* implies that the full backward generator $\mathscr{L}$ is self-adjoint with respect to the weight function $\pi(x)$, and that its eigenfunctions $\{B_n(x)\}$ form an orthogonal system with respect to the same weight function. Now, if we define $K_n(x)$ as

$$K_n\,(x) = e^{-\bar{\sigma}\,(x)/2}\, R_n^{(\alpha,\beta)}\,(x), \tag{32}$$

then (8) implies that $\{K_n(x)\}$ is a complete system of orthogonal functions with respect to the weight function $\pi(x)$. However, by applying the generator $\mathscr{L}$ to $K_n(x)$, one can show that $K_n(x)$ is not an eigenfunction of $\mathscr{L}$. Rather, we obtain

$$\begin{aligned} \mathscr{L}\, K_n\,(x) &= e^{-\bar{\sigma}\,(x)/2} \left\{\mathscr{L}_0\, R_n^{(\alpha,\beta)}\,(x) - Q\,(x;\alpha,\beta,\sigma,h)\, R_n^{(\alpha,\beta)}\,(x)\right\} \\ &= -\,e^{-\bar{\sigma}\,(x)/2} \left[\lambda_n^{(\alpha,\beta)}\, R_n^{(\alpha,\beta)}\,(x) + Q\,(x;\alpha,\beta,\sigma,h)\, R_n^{(\alpha,\beta)}\,(x)\right], \end{aligned}$$

where $Q(x;\alpha,\beta,\sigma,h)$ is the following degree-4 polynomial in x:

$$\begin{aligned} Q\,(x;\alpha,\beta,\sigma,h) = \sigma\, &\{h\,\alpha + [1+\alpha-(2+3\,\alpha+\beta)\,h]\,x - (1+\alpha+\beta)\,(1-2\,h)\,x^2\} \\ &+ 2\,\sigma^2\, x\,(1-x)\,(h + x - 2\,hx)^2. \end{aligned} \tag{33}$$

For no recurrent mutation ($\alpha = \beta = 0$) we get $(1 - x)x\sigma\,[1 - 2h + 2(h + x - 2hx)^2\sigma]$, and for $h = 1/2$ (genic selection), (33) reduces to a degree-2 polynomial: $\frac{1}{2}\,\{\sigma\,[-\beta\, x + \alpha\,(1-x)] + \sigma^2\, x\,(1-x)\}$ In the case of just drift and genic selection, we obtain $\frac{1}{2}\,\sigma^2\, x\,(1-x)$ as in (21).

Again, $\{B_n(x)\}$ and $\{K_n(x)\}$ are orthogonal with respect to the same weight function $\pi(x)$, and $\{K_n(x)\}$ form a basis of $L^2([0,1],\pi)$, where π is defined in (30). Hence, we pose a representation for the eigenfunctions of the form

$$B_n\,(x) = \sum_{m=0}^{\infty} w_{n,m}\, K_m\,(x) = \sum_{m=0}^{\infty} w_{n,m}\, e^{-\bar{\sigma}\,(x)/2}\, R_m^{(\alpha,\beta)}\,(x), \tag{34}$$

where $w_{n,m}$ are constants to be determined. It can be checked that $K_m\,(x) = e^{-\bar{\sigma}\,(x)/2}\, R_m^{(\alpha,\beta)}\,(x)$, for all $m \geq 0$, satisfies the proper regular reflecting or entrance boundary conditions, and hence so does $B_n(x)$.

Now, $\mathscr{L}B_n(x) = -\Lambda_n B_n(x)$ implies the following algebraic equation:

$$\sum_{m=0}^{\infty} w_{n,m} \left[\lambda_m^{(\alpha,\beta)} + Q\ (x;\alpha,\beta,\sigma,h)\right] R_m^{(\alpha,\beta)}\ (x) = \Lambda_n \sum_{m=0}^{\infty} w_{n,m}\ R_m^{(\alpha,\beta)}\ (x). \tag{35}$$

Using (10), we can represent $Q\ (x;\alpha,\beta,\sigma,h)\ R_n^{(\alpha,\beta)}\ (x)$ as a finite linear combination of $R_j^{(\alpha,\beta)}\ (x)$:

$$\begin{aligned} Q\ (x;\alpha,\beta,\sigma,h)\ R_m^{(\alpha,\beta)}\ (x) = b_m^{(-4)}\ R_{m-4}^{(\alpha,\beta)}\ (x) + b_m^{(-3)}\ R_{m-3}^{(\alpha,\beta)}\ (x) \\ + \cdots + b_m^{(+3)}\ R_{m+3}^{(\alpha,\beta)}\ (x) + b_m^{(+4)}\ R_{m+4}^{(\alpha,\beta)}\ (x), \end{aligned} \tag{36}$$

where the coefficients $b_m^{(i)}$ are constants that depend on m, α, β, σ, and h.

For a nonnegative integer k, multiplying (35) by $R_k^{(\alpha,\beta)}\ (x)$ and integrating over [0,1] with respect to the weight function $\pi_0(x)$ yields

$$\begin{aligned} \lambda_k^{(\alpha,\beta)}\ w_{n,k} + b_{k+4}^{(-4)}\ w_{n,k+4} + b_{k+3}^{(-3)}\ w_{n,k+3} + \cdots + b_{k-3}^{(+3)}\ w_{n,k-3} \\ + b_{k-4}^{(+4)}\ w_{n,k-4} = \Lambda_n\ w_{n,k}, \end{aligned} \tag{37}$$

where we define $b_j^{(i)} = 0$ if $j < 0$.

Algorithm 2 (recurrent mutation and general diploid selection)

We can now describe our algorithm for finding the eigenvalues and eigenfunctions of the backward generator $\mathscr{L}$ defined in (29) for the case with recurrent mutation and general diploid selection. From (37), we arrive at a linear system $M\ \mathbf{w}_n = \Lambda_n\ \mathbf{w}_n$, where $\mathbf{w}_n = (w_{n,0}, w_{n,1}, w_{n,2}, \ldots)$ is an infinite-dimensional vector of variables and M is an infinite-dimensional matrix given by

$$M = \begin{pmatrix} \lambda_0^{(\alpha,\beta)} + b_0^{(0)} & b_1^{(-1)} & b_2^{(-2)} & b_3^{(-3)} & b_4^{(-4)} & 0 & 0 & \cdots \\ b_0^{(+1)} & \lambda_1^{(\alpha,\beta)} + b_1^{(0)} & b_2^{(-1)} & b_3^{(-2)} & b_4^{(-3)} & b_5^{(-4)} & 0 & \cdots \\ b_0^{(+2)} & b_1^{(+1)} & \lambda_2^{(\alpha,\beta)} + b_2^{(0)} & b_3^{(-1)} & b_4^{(-2)} & b_5^{(-3)} & b_6^{(-4)} & \cdots \\ b_0^{(+3)} & b_1^{(+2)} & b_2^{(+1)} & \lambda_3^{(\alpha,\beta)} + b_3^{(0)} & b_4^{(-1)} & b_5^{(-2)} & b_6^{(-3)} & \cdots \\ b_0^{(+4)} & b_1^{(+3)} & b_2^{(+2)} & b_3^{(+1)} & \lambda_4^{(\alpha,\beta)} + b_4^{(0)} & b_5^{(-1)} & b_6^{(-2)} & \cdots \\ 0 & b_1^{(+4)} & b_2^{(+3)} & b_3^{(+2)} & b_4^{(+1)} & \lambda_5^{(\alpha,\beta)} + b_5^{(0)} & b_6^{(-1)} & \cdots \\ 0 & 0 & b_2^{(+4)} & b_3^{(+3)} & b_4^{(+2)} & b_5^{(1)} & \lambda_6^{(\alpha,\beta)} + b_6^{(0)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Closed-form formulas for $b_m^{(i)}$ can be found easily using symbolic computation software such as Mathematica. In the Appendix, we provide a dynamic programming algorithm for computing $b_m^{(i)}$ which is useful for implementation in an imperative programming language such as C/C++. If $h = 1/2$, $b_m^{(-4)} = b_m^{(-3)} = b_m^{(+3)} = b_m^{(+4)} = 0$ for all $m \geq 0$, and therefore only the innermost five diagonals of M will be nonzero. As in Algorithm 1, we approximate $M\mathbf{w}_n = \Lambda_n\ \mathbf{w}_n$ by a finite-dimensional truncated linear system

$$M^{[D]}\ \mathbf{w}_n^{[D]} = \Lambda_n^{[D]}\ \mathbf{w}_n^{[D]},$$

where $\mathbf{w}_n^{(D)} = (w_{n,0}^{[D]}, w_{n,1}^{[D]}, \ldots, w_{n,D-1}^{[D]})$ and $M^{[D]}$ is the submatrix of M consisting of its first D rows and D columns. This finite-dimensional linear system can be easily solved to obtain the eigenvalues $\Lambda_n^{[D]}$ and the eigenvectors $\mathbf{w}_n^{[D]}$ of $M^{[D]}$. We show in *Empirical Results* that $\Lambda_n^{[D]}$ and $w_{n,m}^{[D]}$ converge very rapidly as the truncation level D increases.

Note that the same cautionary remark mentioned at the end of Algorithm 1 also applies here.

Special case: No recurrent mutation

Let $\mathscr{L}_0$ denote the diffusion generator defined in (14), which can be obtained from (26) by setting $\alpha = \beta = 0$. Since $R_0^{(\alpha,\beta)}\ (x) \equiv 1$, it satisfies $\mathscr{L}_0 B = -\lambda B$ with $\lambda = 0$. However, for $\alpha = \beta = 0$, the boundaries are exit boundaries, and therefore $R_0^{(\alpha,\beta)}\ (x)$ does not satisfy the requisite boundary conditions. Furthermore, $R_1^{(\alpha,\beta)}\ (x) = \beta x - \alpha\ (1-x) \rightarrow 0$ as α, $\beta \rightarrow 0$, so it is not of interest. In contrast, for $n \geq 0$, $R_{n+2}^{(\alpha,\beta)}\ (x)$ converges to $G_n(x)$ as α, $\beta \rightarrow 0$, and, as established in *Background*, $G_n(x)$ satisfies $\mathscr{L}_0 B = -\lambda B$ with $\lambda = \lambda_{n+2}^{(0,0)}$ and satisfies the requisite boundary conditions for $\alpha = \beta = 0$. In summary, as α, $\beta \rightarrow 0$, the first two modified Jacobi polynomials become irrelevant, while the rest converge to the appropriate eigenfunctions of $\mathscr{L}_0$. These facts have been noticed before (*e.g.*, see Griffiths and Spanò 2010), and they allow us to embed the model with no recurrent mutation conveniently into the model with recurrent mutation as described below.

If $\alpha = \beta = 0$, let Λ_n and $\mathbf{w}_n = (w_{n,0}, w_{n,1}, \ldots)$, respectively, denote the eigenvalues and eigenvectors of M', the submatrix of M obtained by omitting the first two rows and the first two columns. Defining $K_m\ (x) := e^{\overline{\sigma}}\ (x)/2\ G_m\ (x)$ (instead of Equation 32) yields the eigenvalues Λ_n and the eigenfunctions $B_n(x)$ for the backward diffusion generator with a general diploid selection model but no recurrent mutation. Indeed, under genic selection ($h = 1/2$) and $\alpha = \beta = 0$, one can show that $b_{m+2}^{(-4)} = b_{m+2}^{(-3)} = b_{m+2}^{(-1)} = b_{m+2}^{(+1)} = b_{m+2}^{(+3)} = b_{m+2}^{(+4)} = 0$, while $b_{m+2}^{(-2)} = a_m^{(-2)}$, $b_{m+2}^{(0)} = a_m^{(0)}$, and $b_{m+2}^{(+2)} = a_m^{(+2)}$, where $a_j^{(i)}$ are defined in (23). The cases $\alpha > 0$, $\beta = 0$ and $\alpha = 0$, $\beta > 0$ can be treated along similar lines.

Empirical Results

In this section we study the convergence behavior of the eigenvalues and eigenvectors of the submatrix $M^{[D]}$ as its dimension D increases. Further, we show how the spectral representation of the transition density can be employed to characterize the convergence rate of the diffusion to stationarity, *i.e.*, mutation–selection equilibrium.

Convergence of the eigenvalues and eigenfunctions

As the dimension D of the submatrix $M^{[D]}$ increases, we generally observe rapid convergence of the eigenvalues $\Lambda_n^{[D]}$ and the entries $w_{n,m}^{[D]}$ of the eigenvectors, for fixed $n,m < D$. For example, the convergence behavior of $\Lambda_0^{[D]}, \Lambda_5^{[D]}, \Lambda_{10}^{[D]}$ is shown in Figure 1, A and B, for $\sigma = 10$ and $\sigma = 100$, respectively, with mutation parameters $\alpha = 0.01$, $\beta = 0.01$ and the

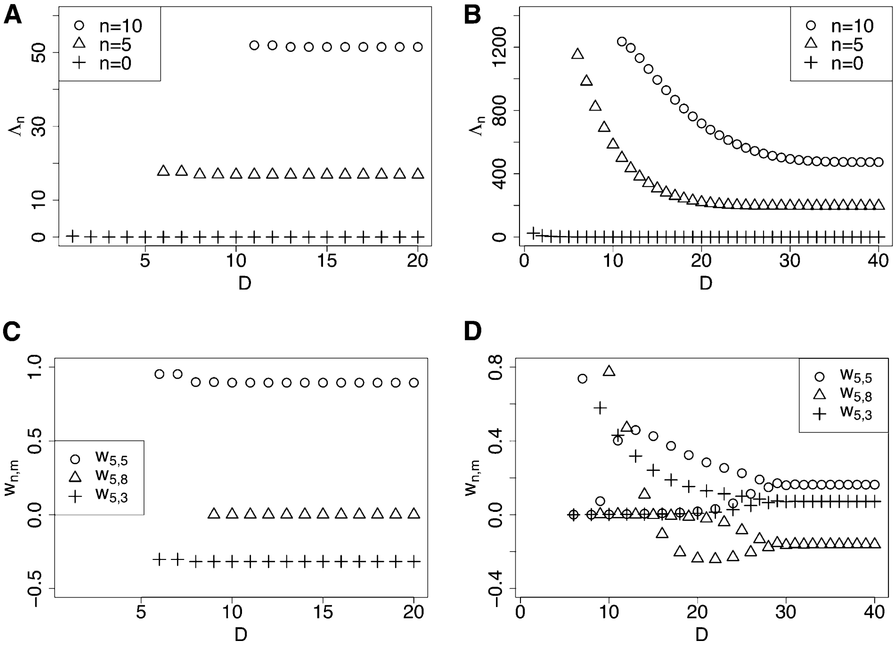

Figure 1 Convergence of the eigenvalues $\Lambda_n^{[D]}$ and coefficients $w_{n,m}^{[D]}$ with increasing truncation level D. The mutation rates are set to $\alpha = \beta = 0.01$ and the dominance parameter $h = 0.5$. (A) $\Lambda_0^{[D]}, \Lambda_5^{[D]}, \Lambda_{10}^{[D]}$ for $\sigma = 10$. (B) $\Lambda_0^{[D]}, \Lambda_5^{[D]}, \Lambda_{10}^{[D]}$ for $\sigma = 100$. (C) $w_{5,3}^{[D]}, w_{5,5}^{[D]}, w_{5,8}^{[D]}$ for $\sigma = 10$. (D) $w_{5,3}^{[D]}, w_{5,5}^{[D]}, w_{5,8}^{[D]}$ for $\sigma = 100$.

dominance parameter $h = 1/2$. Figure 1, A and B, illustrates that, for both $\sigma = 10$ and $\sigma = 100$, $\Lambda_0^{[D]}$ converges rapidly to 0 as D increases, consistent with our expectation (see below). The figures illustrate that in general $\Lambda_n^{[D]}$ converges rapidly for a wide range of σ values, but that the convergence rate slows down as σ increases. For α and β in biologically relevant ranges (say, 10^{-3} to 10^{-1}), we generally observe that changing the mutational parameters does not affect the convergence behavior significantly. Also, the dominance parameter h has little influence on the convergence rate provided that $0 \leq h \leq 1$.

The typical convergence behavior of the eigenvector entries $w_{n,m}^{[D]}$ is illustrated in Figure 1, C and D, for $\sigma = 10$ and $\sigma = 100$, respectively. As Figure 1C shows, the rate of convergence is very fast for small σ. For large σ, as in Figure 1D, $w_{n,m}^{[D]}$ may fluctuate for small values of D, but they stabilize rapidly as D increases. In general, we observe that the convergence of $\Lambda_n^{[D]}$ and that of $w_{n,m}^{[D]}$ are roughly synchronized; *i.e.*, for a fixed n, Λ_n and $w_{n,m}^{[D]}$ stabilize near similar values of D.

Figure 2 shows the dependence of $\Lambda_n^{[D]}$ on σ, h, and n. Observe that $\Lambda_n^{[D]}$ increases rapidly as n increases, which implies that using a finite number of terms in the spectral representation of the transition density should yield an accurate approximation of the true transition density. Increasing σ or choosing h significantly different from 0.5 (the genic selection case) shifts the entire spectrum upward, but in all cases we observe that $\Lambda_n^{[D]}$ increases rapidly with n.

Empirical transition densities and stationary distributions

For given mutation and selection parameters, the eigenvalues $-\Lambda_n$ and the eigenfunctions $B_n(x)$ found by our method can be used to obtain the transition density via the spectral representation (5), for arbitrary $t > 0$ and $x, y \in [0,1]$. This representation includes the stationary density, which admits a more explicit analytic form. To this end, note that a diffusion generator $\mathscr{L}$ maps constant functions to zero. In the case with recurrent mutation, we have either regular reflecting or entrance boundaries, and constant functions actually satisfy the requisite boundary conditions. Hence, constant functions are valid eigenfunctions of $\mathscr{L}$ with eigenvalue zero. That is, $\Lambda_0 = 0$ and $B_0(x) = C$, where C is some constant, for all $x \in [0,1]$. Thus, the density of the stationary measure is given by

$$\lim_{t\to\infty} p\,(t;x,y) = \pi\,(y)\,\frac{B_0\,(x)\,B_0\,(y)}{\langle B_0, B_0\rangle_\pi} = \pi\,(y)\,\frac{C^2}{\langle C, C\rangle_\pi}$$
$$= \frac{\pi\,(y)}{\int_0^1 \pi\,(z)\,\mathrm{d}\,z},$$

where $\pi\,(y) = e^{\bar\sigma\,(y)}\,\pi_0\,(y)$ is the speed density defined in (30). The integral in the denominator (which corresponds to a normalization constant for the stationary density) can be solved efficiently using our approach: Since B_0 is a constant function, we can express the integral as

$$\int_0^1 \pi\,(z)\,\mathrm{d}\,z = \frac{\langle B_0, B_0\rangle_\pi}{B_0\,(1)\,B_0\,(1)}.$$

Then, using the representation

$$B_0\,(x) = \sum_{m=0}^{\infty} w_{0,m}\,e^{-\bar\sigma\,(x)/2}\,R_m^{(\alpha,\beta)}\,(x),$$

and the facts $\bar\sigma\,(1) = 2\,\sigma$ [*cf.*, (31)] and $R_n^{(\alpha,\beta)}\,(1) = \Gamma\,(n+\beta)/[\Gamma\,(n+1)\,\Gamma\,(\beta)]$, we obtain

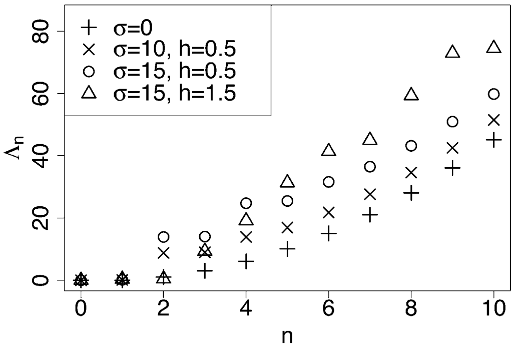

Figure 2 Magnitude of the eigenvalues $\{-\Lambda_n\}$ of the diffusion generator $\mathscr{L}$ for $\alpha = 0.01$, $\beta = 0.01$, and various values of the selection coefficient σ and the dominance parameter h. The truncation level D was set to 400. Note that Λ_n gets larger with increasing n, a general trend that holds for other parameter settings. Also, Λ_n increases when selection gets stronger. For $\sigma = 0$, note that $\Lambda_n = \lambda_n^{(\alpha,\beta)}$, defined in (28).

$$\int_0^1 \pi(z)\,\mathrm{d}z = \frac{\sum_{m=0}^{\infty}\left(w_{0,m}\right)^2 \left\langle R_m^{(\alpha,\beta)}, R_m^{(\alpha,\beta)}\right\rangle_{\pi_0}}{e^{-\bar{\sigma}(1)}\left[\sum_{k=0}^{\infty} w_{0,k} R_k^{(\alpha,\beta)}(1)\right]^2} = \frac{\sum_{m=0}^{\infty}\left(w_{0,m}\right)^2 \Delta_m(\alpha,\beta)}{e^{-2\sigma}\left[\sum_{k=0}^{\infty} w_{0,k} \dfrac{\Gamma(k+\beta)}{\Gamma(k+1)\,\Gamma(\beta)}\right]^2}, \qquad (38)$$

where $\Delta_m(\alpha,\beta)$ is the combinatorial coefficient defined in (9). Thus, the integral can be evaluated purely algebraically. For a fixed n, $w_{n,m} \to 0$ as $m \to \infty$, so we can obtain an accurate approximation of (38) by truncating the infinite sums and by computing $w_{0,m}$ using the method described in this article. In special cases, the integral $\int_0^1 \pi(z)\,\mathrm{d}z$ can be evaluated numerically using other methods (*e.g.*, see Wakeley and Sargsyan 2009), but, for general σ and h, standard numerical integration techniques do not seem to provide accurate answers.

Figure 3 shows some examples of the time evolution of the transition density function, with the $t = \infty$ case corresponding to the stationary distribution. Specifically, three different types of selection schemes are illustrated:

1. Illustrated in Figure 3A are the densities for *strong positive selection* ($\sigma = 100$, $h = 0.5$), when starting with a small initial frequency of $x = 0.0005$. As expected, for small t there is still some probability mass near 0, but already a substantial amount has moved to 1. At stationarity, the mass is concentrated at the boundaries, with the concentration near 1 being far more pronounced than that near 0.
2. Figure 3B shows the dynamics of *balancing selection* ($\sigma = 0.01$, $h = 10000$), starting from initial frequency $x = 0.0005$. As time evolves, the mass gets shifted from the boundary at 0 to an intermediate frequency of $y = 0.5$, where a large fraction of probability mass resides at stationarity.
3. In Figure 3C, the allele A_1 exhibits *weakly deleterious selection* ($\sigma = -1$, $h = 0.5$), with the initial frequency being $x = 0.5$. Initially most of the probability mass is concentrated around frequency $y = 0.5$. As the density evolves with time, it spreads out over the interval, and the peak of the density moves to lower frequencies. At stationarity, most of the mass is concentrated around the boundary at 0.

Rate of convergence to the stationary distribution

The spectral representation also allows us to obtain the rate of convergence to the stationary density. The difference $d(t; x, y)$ between the transition density and the stationary density is given by

$$d(t;x,y) := p(t;x,y) - \frac{\pi(y)}{\int_0^1 \pi(z)\,\mathrm{d}z} = \sum_{n=1}^{\infty} e^{-\Lambda_n t}\, \pi(y)\, \frac{B_n(x)\,B_n(y)}{\langle B_n, B_n\rangle_\pi}.$$

Define $\|f\|_{1/\pi} = \sqrt{\langle f, f\rangle_{1/\pi}}$. Then, by orthogonality of the eigenfunctions, we obtain

$$\|d(t;x,\cdot)\|_{1/\pi}^2 = \sum_{n=1}^{\infty} e^{-2\Lambda_n t} \frac{\left[B_n(x)\right]^2}{\langle B_n, B_n\rangle_\pi} = \sum_{n=1}^{\infty} e^{-2\Lambda_n t} \frac{e^{-\bar{\sigma}(x)}\left[\sum_{k=0}^{\infty} w_{n,k} R_k^{(\alpha,\beta)}(x)\right]^2}{\sum_{m=0}^{\infty}\left(w_{n,m}\right)^2 \Delta_m(\alpha,\beta)}, \qquad (39)$$

which can be approximated by truncating the infinite sums. Figure 4 shows the dependence of $\|d(t;x,\cdot)\|_{1/\pi}^2$ on time t, for $\alpha = 0.01$, $\beta = 0.01$, $h = 0.5$, $\sigma \in \{1,10,100\}$, and initial frequency $x = 0.0005$. As expected, the distance to the stationary distribution decreases over time, and the rate of convergence is faster for larger σ. We note that the spectral representation can also be readily employed to study convergence rates measured by other metrics such as the total variation distance or relative entropy.

Discussion

In this article, we developed a simple method for finding the eigenvalues and eigenfunctions of the diffusion generator associated with the Wright–Fisher diffusion with recurrent mutation and general diploid selection. As described in *Background*, these eigenvalues and eigenfunctions can be used to construct a spectral representation (5) of the transition density. Since the eigenvalues $-\Lambda_n$ tend to $-\infty$ as $n \to \infty$, and the contribution of the nth eigenfunction to the transition density is proportional to $e^{-\Lambda_n t}$, we can truncate the series (5) at some appropriate level and obtain a highly accurate approximation of the transition density. The mathematical derivation of our work invokes the theory of self-adjoint operators and orthogonal functions, but the resulting algorithm involves only standard linear algebra, which is straightforward to

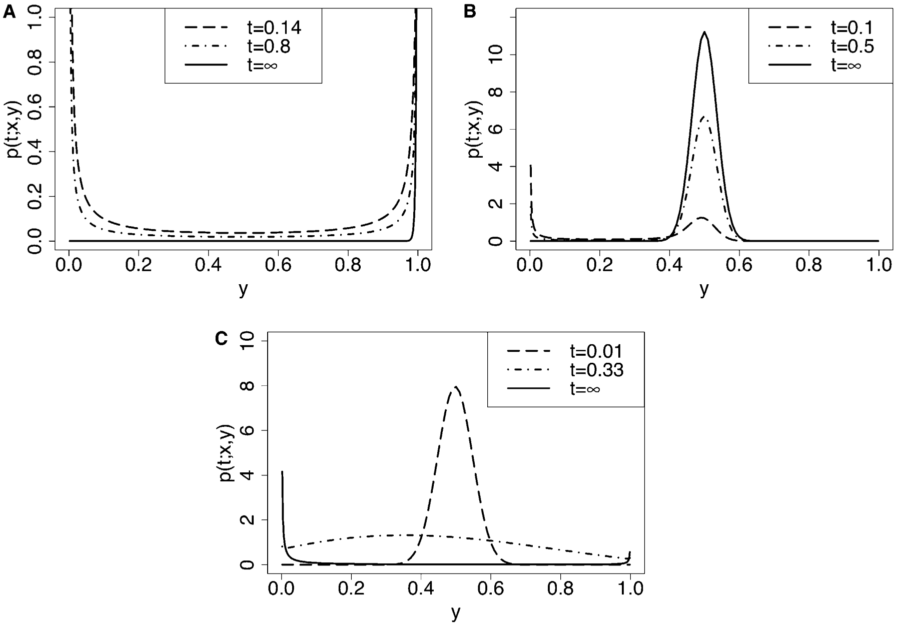

Figure 3 The transition density $p(t;x,y)$ as a function of y. Various times, selection parameters, and initial frequencies were considered. The mutation rates were set to $\alpha = \beta = 0.01$ in all examples. The $t = \infty$ case corresponds to the stationary distribution. A truncation level of $D = 1000$ was used in the computation, and Equations 5 and 34 were approximated by summing over $0 \leq n \leq 300$ and $0 \leq m \leq 500$. (A) Strong positive selection: $\sigma = 100$, $h = 0.5$, $x = 0.0005$. (B) Balancing selection: $\sigma = 0.01$, $h = 10000$, $x = 0.0005$. (C) Weakly deleterious selection: $\sigma = -1$, $h = 0.5$, $x = 0.5$.

implement. For a given set of parameters, computing the first 500 eigenvalues and eigenfunctions using our method takes only a few seconds in Mathematica.

An accurate transition density enables one to estimate the parameters of Wright–Fisher diffusions, perhaps most interestingly the selection parameters. As mentioned in the Introduction, Bollback *et al.* (2008) suggested a hidden Markov model framework for estimating the selection coefficient σ by analyzing samples taken from multiple time points. The analytic transition density obtained from our method can be incorporated into that framework and thereby ameliorate potential numerical problems that may arise from trying to solve the Kolmogorov equation using discretization. Furthermore, our approach can be applied to devise an algebraic method for computing the sampling probability at stationarity under a general selection model.

There are several interesting extensions of our work to explore. It is known (Shimakura 1977; Griffiths 1979; Griffiths and Spanò 2010) that multivariate Jacobi polynomials, orthogonal with respect to the Dirichlet distribution, are eigenfunctions of multiallelic diffusions under parent-independent mutation models. We believe that the technique developed in this article can be extended to find the spectral representation of the transition density of a multiallelic diffusion with parent-independent mutation and general diploid selection.

For a neutral diallelic Wright–Fisher model with subdivided population structure, Lukić *et al.* (2011) recently obtained numerical approximations of the transition density by using a certain class of orthogonal polynomials. We remark that the orthogonal polynomials used in that approach are not eigenfunctions of the diffusion generator. Further, the system of ordinary differential equations (ODEs) satisfied by the coefficients of the basis functions does not admit a simple solution, so Lukić *et al.* (2011) employed a finite difference method with which to solve the ODEs numerically. Note that their method does not provide a proper spectral representation of the transition density, since it does not find the eigenvalues and eigenfunctions of the diffusion generator. It might be possible to extend the technique developed in this article to obtain a spectral representation of the transition density in the case with subdivided population structure and general diploid selection.

In this article, we considered only one-locus Wright–Fisher diffusions. It is generally acknowledged that inference of evolutionary parameters, especially regarding selection, can be improved significantly by taking into account additional data at closely linked loci. Hence, it would be desirable to extend the approach described here to handle the dynamics of multilocus diffusions. However, our current technique relies on the fact that the eigenfunctions are known for the diffusion generator under neutrality. Therefore, to be able to apply our approach to multilocus diffusions with recombination and selection, one

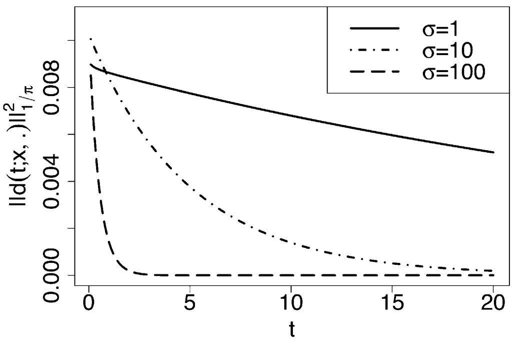

Figure 4 Convergence of the transition density to stationarity as time evolves, for initial frequency $x = 0.0005$. Deviation from the stationary density is measured by $\|d(t;x,\cdot)\|^2_{1/\pi}$, defined in (39). The mutation and selection parameters were set to $\alpha = 0.01$, $\beta = 0.01$, $h = 0.5$, and $\sigma \in \{1,10,100\}$. A truncation level of $D = 1000$ was used in the computation, and (39) was approximated by summing over $0 \leq n \leq 300$ and $0 \leq k, m \leq 500$.

would have to know the eigenfunctions in the neutral case. To our knowledge, no such eigenfunctions are known.

Since diffusion processes also arise in other disciplines (*e.g.*, physics and mathematical finance), several approaches have been proposed to obtain efficient approximations of the transition densities for diffusions more general than the Wright–Fisher diffusion (see Srensen 2004; Aït-Sahalia 2008, for example). It would be interesting to investigate whether one could borrow techniques from those fields to the population genetics applications mentioned above.

Finally, we note that Mano (2009) recently employed the representation of the transition density given by Kimura (1955a) and the moment duality used in Barbour *et al.* (2000) to investigate the dynamics of the number of lineages in the ancestral selection graph dual to the Wright–Fisher diffusion. The representation of the transition density found in this article can be employed to include recurrent mutation into that framework.

Acknowledgments

We thank two reviewers for helpful comments and suggestions. This research is supported in part by National Institutes of Health grants R00-GM080099 and R01-GM094402 to Y.S.S. and Deutsche Forschungsgemeinschaft Research Fellowship STE 2011/1-1 to M.S.

Literature Cited

Abramowitz, M., and I. A. Stegun (Editors), 1965 *Handbook of Mathematical Functions*. Dover, New York.

Aït-Sahalia, Y., 2008 Closed-form likelihood expansions for multivariate diffusions. Ann. Stat. 36(2): 906–937.

Barbour, A. D., S. N. Ethier, and R. C. Griffiths, 2000 A transition function expansion for a diffusion model with selection. Ann. Appl. Probab. 10: 123–162.

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

Donnelly, P., M. Nordborg, and P. Joyce, 2001 Likelihoods and simulation methods for a class of nonneutral population genetics models. Genetics 159: 853–867.

Durrett, R., 2008 *Probability Models for DNA Sequence Evolution*. Springer, New York.

Etheridge, A. M., and R. C. Griffiths, 2009 A coalescent dual process in a moran model with genic selection. Theor. Popul. Biol. 75(4): 320–330.

Ewens, W. J., 2004 *Mathematical Population Genetics: I. Theoretical Introduction*. Springer, New York.

Eyre-Walker, A., and P. Keightley, 2007 The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8(8): 610–618.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel *et al.*, 2010 A draft sequence of the neandertal genome. *Science* 328 (5979): 710–722.

Griffiths, R. C., 1979 A transition density expansion for a multi-allele diffusion model. Adv. Appl. Probab. 11: 310–325.

Griffiths, R. C., 2003 The frequency spectrum of a mutation, and its age, in a general diffusion model. Theor. Popul. Biol. 64(2): 241–251.

Griffiths, R. C., and W.-H. Li, 1983 Simulating allele frequencies in a population and the genetic differentiation of populations under mutation pressure. Theor. Popul. Biol. 23(1): 19–33.

Griffiths, R. C., and D. Spanò, 2010 Diffusion processes and coalescent trees, pp. 358–375 in *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*, Vol. 10, edited by N. H. Bingham, and C. M. Goldie. Cambridge University Press, Cambridge, UK.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, 2009 Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS Genet. 5(10): e1000695.

Hummel, S., D. Schmidt, B. Kremeyer, B. Herrmann, and M. Oppermann, 2005 Detection of the CCR5-Δ32 HIV resistance gene in bronze age skeletons. Genes Immun. 6(4): 371–374.

Karlin, S., and H. Taylor, 1981 *A Second Course in Stochastic Processes*. Academic Press, San Diego.

Kimura, M., 1955a Stochastic processes and distribution of gene frequencies under natural selection, pp. 33–53 in *Cold Spring Harbor Symposia on Quantitative Biology*, Vol. 20. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Kimura, M., 1955b Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA 41: 144–150.

Kimura, M., 1957 Some problems of stochastic processes in genetics. Ann. Math. Stat. 28(4): 882–901.

Lenski, R. E., 2011 The *E. coli* long-term experimental evolution project site. Available at: http://myxo.css.msu.edu/ecoli. Accessed: November, 2011.

Lukić, S., J. Hey, and K. Chen, 2011 Non-equilibrium allele frequency spectra via spectral methods. Theor. Popul. Biol. 79(4): 203–219.

Mano, S., 2009 Duality, ancestral and diffusion processes in models with selection. Theor. Popul. Biol. 75(2–3): 164–175.

Reich, D., R. E. Green, M. Kircher, J. Krause, N. Patterson *et al.*, 2010 Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468(7327): 1053–1060.

Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch *et al.*, 1999 Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J. Virol. 73(12): 10489–10502.

Shimakura, N., 1977 Equations différentielles provenant de la génétique des populations. Tohoku Math. J. 29: 287–318.

Stratton, J. A., P. M. Morse, L. J. Chu, and R. A. Hutner, 1941 *Eliptic Cylinder and Spheroidal Wave functions*. Wiley, New York.

Szegö, G., 1939 *Orthogonal Polynomials*, Ed. 4th. American Mathematical Society, Providence, RI.

Srensen, H. 2004 Parametric inference for diffusion processes observed at discrete points in time: a survey. Int. Statist. Rev. 73 (3): 337–354.

Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26: 119–164.

Wakeley, J., and O. Sargsyan, 2009 The conditional ancestral selection graph with strong balancing selection. Theor. Popul. Biol. 75(4): 355–364.

Wichman, H. A., M. R. Badgett, L. A. Scott, C. M. Boulianne, and J. J. Bull, 1999 Different trajectories of parallel evolution during viral adaptation. Science 285(5426): 422–424.

Communicating editor: L. M. Wahl

Appendix

Here, we describe the computation of the coefficients $b_m^{(i)}$ in Equation (36). Recall that the polynomial $Q(x;\,\alpha,\beta,\sigma,h)$ defined in (33) is of degree 4. Represent this polynomial as

$$Q\,(x;\alpha,\beta,\sigma,h)=\sum_{l=0}^{4}q_l x^l, \tag{40}$$

where q_l are coefficients that depend on α, β, σ, and h. As shown in (10) and (11), $x\,R_m^{(\alpha,\beta)}\,(x)$ satisfies a three-term recurrence relation of the form

$$\begin{aligned} x\,R_m^{(\alpha,\beta)}\,(x) = {} & g\,(m,m-1)\,R_{m-1}^{(\alpha,\beta)}\,(x)+g\,(m,m)\,R_m^{(\alpha,\beta)}\,(x)\\ & +g\,(m,m+1)\,R_{m+1}^{(\alpha,\beta)}\,(x), \end{aligned}$$

where $g(m,\,m-1)$, $g(m,\,m)$, $g(m,\,m+1)$ are coefficients that depend on m and α, β. Note that (11) implies $g(0,\,-1)=0$. Using the recurrence relation inductively gives

$$x^l R_m^{(\alpha,\beta)}\,(x)=\sum_{k=m-l}^{m+l}h\,(m,l,k)\,R_k^{(\alpha,\beta)}\,(x), \tag{41}$$

where

$$h\,(m,l,k):=\begin{cases}\delta_{m,k},\\ \quad \text{if } \ l=0,\\ 1_{\{|m-1-k|\le l-1\}}\,g\,(m,m-1)\,h\,(m-1,l-1,k)\\ \quad +1_{\{|m-k|\le l-1\}}\,g\,(m,m)\,h\,(m,l-1,k)\\ \quad +1_{\{|m+1-k|\le l-1\}}\,g\,(m,m+1)\,h\,(m+1,l-1,k),\\ \quad \text{if } \ l>0.\end{cases} \tag{42}$$

Now, (40) and (41) imply

$$Q\,(x;\alpha,\beta,\sigma,h)\,R_m^{(\alpha,\beta)}\,(x)=\sum_{l=0}^{4}q_l\sum_{k=m-l}^{m+l}h\,(m,l,k)\,R_k^{(\alpha,\beta)}\,(x)=\sum_{k=m-4}^{m+4}\left[\sum_{l=|k-m|}^{4}q_l\,h\,(m,l,k)\right]R_k^{(\alpha,\beta)}\,(x).$$

Thus, the coefficients $b_m^{(i)}$ in (36) are given by

$$b_m^{(i)}=\sum_{l=|i|}^{4}q_l\,h\,(m,l,k),$$

where $h(m,\,l,\,k)$ can be computed efficiently using the dynamic programming in (42).