Mathematical Biosciences 272 (2016) 100–112

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Exact Markov chain and approximate diffusion solution for haploid genetic drift with one-way mutation

Ola Hössjer[a,*], Peder A. Tyvand[b], Touvia Miloh[c]

[a] Department of Mathematics, Div. of Mathematical Statistics, Stockholm University, Stockholm SE 106 91, Sweden
[b] Department of Mathematical Sciences and Technology, Norwegian University of Life Sciences, Ås 1432, Norway
[c] Faculty of Engineering, Tel Aviv University, Tel Aviv 69978, Israel

ARTICLE INFO

Article history:
Received 11 July 2015
Revised 13 November 2015
Accepted 4 December 2015
Available online 24 December 2015

MSC:
92D10
92D25
60G15

Keywords:
Diffusion
Jordan decomposition
Markov chain
Mutation
Quasi-fixation
Wright–Fisher model

ABSTRACT

The classical Kimura solution of the diffusion equation is investigated for a haploid random mating (Wright–Fisher) model, with one-way mutations and initial-value specified by the founder population. The validity of the transient diffusion solution is checked by exact Markov chain computations, using a Jordan decomposition of the transition matrix. The conclusion is that the one-way diffusion model mostly works well, although the rate of convergence depends on the initial allele frequency and the mutation rate. The diffusion approximation is poor for mutation rates so low that the non-fixation boundary is regular. When this happens we perturb the diffusion solution around the non-fixation boundary and obtain a more accurate approximation that takes quasi-fixation of the mutant allele into account. The main application is to quantify how fast a specific genetic variant of the infinite alleles model is lost. We also discuss extensions of the quasi-fixation approach to other models with small mutation rates.

1. Introduction

Diffusion theory is an important approach to population genetics. It was first introduced by Fisher [17] and carried further by Wright [42]. A breakthrough for diffusion models came with the pioneering papers of Kimura [26,27]. In particular, Kimura gave an explicit solution of the diffusion equation for the classical two-allele and haploid Wright–Fisher model ([17,41]) without mutations, in terms of an infinite series that is parametrized by time $\tau \geq 0$ and the frequency $0 \leq x \leq 1$ of one the two alleles. Crow and Kimura [7] solved the diffusion problem for the two-allele model with immigration from an outside population, and for multi-allele models with a small amount of selection. Chapter 8 of [8] summarizes further diffusion solutions for the two allele Wright–Fisher model, for instance when mutations between the alleles are possible in one direction or both, referred to as one-way or two-way mutations. For a more recent account of the use of diffusion theory in population genetics, we refer to Chapter 10 of [11], Chapters 4 and 5 of [15], and Chapter 7 of [9]. The review article [22] contains different representations of the diffusion solution for a very wide class of Wright–Fisher models. This includes series expansions with Jacobi polynomial eigenfunctions that generalize the ones in [8], and series expansions in terms of a dual coalescence process ([10,20,36]).

For the two-allele model, it is convenient to treat the interior domain ($0 < x < 1$) and the boundary points ($x \in \{0, 1\}$) of the diffusion solution in a unified manner. McKane and Waxman [31] confirmed Kimura's formula for the Wright–Fisher model without mutations, and complemented it by introducing Dirac singularities at each boundary, a solution that is also implicit in [28] and in Sections 8.4 and 8.8 of [8]. These singularities conserve the total probability and possess a memory, as they integrate in time the outgoing flux through the boundaries of the open domain of the compositional x. A continuity equation for the probability flux at the boundary takes care of the conservation of total probability.

The diffusion approximation represents a similarity solution, where all population sizes evolve similarly, provided the time variable is scaled properly. This similarity is not exactly valid, as there are deviations from the diffusion limit for all finite population sizes. Therefore, the diffusion model is only asymptotically valid in the limit of large population sizes. This discrepancy becomes important for boundary points of mutant alleles, that have two-way

* Corresponding author. Tel.: +46 70 672 12 18.
E-mail addresses: ola@math.su.se (O. Hössjer), peder.tyvand@nmbu.no (P.A. Tyvand), miloh@eng.tau.ac.il (T. Miloh).

http://dx.doi.org/10.1016/j.mbs.2015.12.006
0025-5564/© 2015 Elsevier Inc. All rights reserved.

interaction with the interior domain. Since the diffusion solution has no singular delta function at such a point, boundary conditions are not needed for conservation of probability. But with a relatively small mutation rate, the probability density of the diffusion solution may go to to infinity around the boundary point, but in an integrable manner over the open interval $0 < x < 1$. In such cases it is much more subtle to deduce from the diffusion model itself, what the limit of infinity means for the underlying exact discrete time stochastic process for a finite population. An infinite probability density near a boundary point of a mutant allele means that exact probabilities for the genetic composition of the population, converges to zero at a slower rate.

Tyvand and Thorvaldsen [39] showed that the mutation free Wright–Fisher model has quite a slow convergence rate to the diffusion solution close the two absorbing boundaries $x = 0, 1$. The fact that the Wright–Fisher model involves many neighbor interactions slows down its convergence to diffusion near boundaries. This is because the Markov chain may hit the boundary in one single generation from sample points close to a boundary, whereas the diffusion model predicts a longer time for this to happen. It is therefore important to focus on the behavior near the boundaries, in qualitative and quantitative evaluations of a diffusion model.

In the present work we focus on the two allele Wright–Fisher model with one-way mutations, which is of interest in the context of the infinite alleles model ([29]), with a new allele created after each mutation. One boundary point is then absorbing, having one-way interaction with the interior domain, whereas the other is non-absorbing, with two-way interaction. We demonstrate that the diffusion solution approximates the exact Markov chain close to both boundaries, but the accuracy is higher close to the non-fixation boundary. The accuracy at the absorbing boundary point is also high, at least for large time points τ, but at the non-absorbing boundary, it is highly dependent on the mutation rate. Following the terminology of [15], we distinguish between whether a non-absorbing boundary point is regular (accessible but not absorbing) or entrance (neither accessible nor absorbing). It turns out that the diffusion solution at the non-absorbing boundary is very close to the Markov chain solution if the mutation rate is so large that the boundary point is entrance and the diffusion density in the interior domain is bounded around it. But the accuracy of the diffusion solution at the non-absorbing boundary is very poor if the mutation rate is low and the boundary point is regular, so that the diffusion density around it tends to infinity. When this happens we perturb the diffusion approximation around the non-fixation boundary, treating the mutant allele as quasi-fixed, with a strictly positive quasi fixation probability that tends to zero at an exponential rate.

We also use a Jordan decomposition of the transition matrix to write the exact Markov chain probabilities as a series, and motivate that each term of it converges to the analogous term of the diffusion solution, at a rate inversely proportional to the population size.

The paper is organized as follows. The Wright–Fisher stochastic process with one-way mutations is defined in Section 2, and its link to the continuous time diffusion approach is outlined in Section 3. In Section 4 we review properties of the analytical diffusion equation and present its solution in a new form, as a preparation for the numerical comparison with the exact Markov chain in Section 5. A summary with final conclusions is provided in Section 6, and some of the mathematics is collected in the appendix.

2. The Wright–Fisher stochastic process

We consider the simple Wright–Fisher model of discrete non-overlapping generations of diploid monoecious individuals in random mating. The random mating is assumed to depend only on the present population, independently of all preceding generations. The stochastic process is a Markov chain whose transition probabilities follow a binomial distribution. Ewens [13] calculated the exact Markov chain at the haploid gamete level. Tyvand [38] presented a more general Markov model on the diploid genotype level, and added numerical verification by Monte Carlo simulations.

We consider one locus with two alleles A and a. The generation number is t, and the number of individuals n is constant for all generations. The mating takes place on the gamete level, provided the probability of self-fertilization is $1/n$. The haploid gamete pool consists of $2n$ members, from which we pick at random with replacement. Random mutation is assumed to take place in the gamete pool between two consecutive generations. The probability of mutation (mutation rate) from gamete A to gamete a during one generation is denoted by μ, while the mutation rate from gamete a to gamete A is taken to be identically zero.

A parental population has $n^{(t)}$ gametes of type A and $2n - n^{(t)}$ gametes of type a in generation t. The different possible compositions are numbered by the running index

$$i = n^{(t)}, \ 0 \le i \le 2n. \tag{1}$$

The Markov chain models dynamics of the frequency $n^{(t)}$ of the A allele, with state space $\{0, 1, \ldots, 2n\}$, transition matrix

$$M = (M_{ij})_{i,j=0}^{2n} \tag{2}$$

and transition probabilities

$$\begin{aligned} M_{ij} &= P\Big(n^{(t+1)} = j | n^{(t)} = i\Big) \\ &= \frac{(2n)!}{j!(2n-j)!}\left(\frac{(1-\mu)i}{2n}\right)^{j}\left(\frac{2n-(1-\mu)i}{2n}\right)^{2n-j}. \end{aligned} \tag{3}$$

It is convenient to introduce the probability distribution vector

$$V^{(t)} = (V_0^{(t)}, \ldots, V_{2n}^{(t)}) \tag{4}$$

over all population compositions at generation t, and use the Chapman–Kolmogorov equation to infer

$$V^{(t)} = V^{(0)} M^t. \tag{5}$$

We choose one specific founder population, with a fixed number $n^{(0)}$ and $2n - n_1^{(0)}$ of A and a alleles respectively. This corresponds to an initial vector $V^{(0)}$ with only one nonzero component;

$$V_i^{(0)} = \begin{cases} 1, & i = n^{(0)}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

3. Markov chain linked to diffusion theory

The diffusion model is a similarity solution that represents the stochastic process asymptotically for large population sizes. For comparing the Wright–Fisher process with its continuous diffusion model, we reformulate it in terms of diffusion variables. The first independent diffusion variable is the compositional coordinate x of a parental population, defined from (1) as

$$x = \frac{i}{2n}. \tag{7}$$

The variable x for an offspring population is defined similarly. The second independent diffusion variable is the stretched time coordinate τ. It can be defined in different ways depending on which properties of the Wright–Fisher model that are of interest. Papangelou [33] obtained large deviation results for rapid allele frequency changes over time intervals of length $o(n)$, conditionally on that the number of A-alleles at each one of the end points of the time interval is known. This requires $\tau \gg t/n$, but here we will follow [41,42] and study the unconditional behaviour of the Markov chain, with

$$\tau = \frac{t}{2n}. \tag{8}$$

The dependent diffusion variable is the probability density $v(x, \tau)$, which, for each τ gives the probability density function over all the different population compositions. The probability that the population in the one-dimensional x space has a composition belonging to the compositional interval $[x, x+dx]$ is by definition given by $v(x, \tau)dx$, taking the diffusion approach as representative for the stochastic process.

A diffusion model takes $v(x, \tau)$ to be a continuous function for $0 < x < 1$. The appropriate definition of $v_n(x, \tau)$ for the discrete Markov model with one-way mutation is

$$v_n(x, \tau) = 2nV_i^{(t)}, \tag{9}$$

with x and i related as in (7). This definition may approximate the continuous diffusion solution by a properly scaled histogram that is obtained from the internal points of the Markov chain. An appropriate choice of histogram is by means of a right Riemann sum based on the $2n$ sample points that remain when we exclude the boundary that we locate at $x = 0$ when a is fixed and A is lost in the whole population. The other boundary point $x = 1$ is not excluded, since it behaves like all internal points. The diffusion solution may possibly go to infinity at $x = 1$, but in a continuous and integrable manner. At the fixation boundary $x = 0$ the diffusion solution will possess a Dirac singularity.

The probability distribution obeys the normalizing constraint of conserved total probability. With only one-way mutations, normalization is not achieved by integrating $v(x, \tau)$ over the open domain $0 < x < 1$, because of the Dirac singularity at $x = 0$. We can express the conservation of probability in diffusion theory by the integral

$$\int_{-\infty}^{\infty} v(x, \tau)dx = 1, \tag{10}$$

where $v(x, \tau)$ is defined as identically zero for $x < 0$ and $x > 1$.

The founder population $t = 0$ has a fixed number $x = p$ which is

given by the initial frequency

$$p = \frac{n_1^{(0)}}{2n} \tag{11}$$

of the A gamete, and the corresponding initial frequency of the a gamete is $1 - p$. In the sequel, we disregard the case $p = 0$ because it means complete fixation in gamete a at all times, and thus assume $0 < p \le 1$. The diffusion representation of a founder population is the initial condition

$$v(x, 0) = \delta(x - p), \tag{12}$$

for the continuous variable $v(x, \tau)$, where $\delta(x)$ is Dirac's delta function.

4. Analytical solution of the diffusion equation

In this section we review properties of the diffusion solution $v(x, \tau)$ of the Wright–Fisher model, as a preparation for the numerical and analytical results of the next section. Using the Kolmogorov forward or Fokker–Planck equation, it follows that

$$\frac{\partial v(x,t)}{\partial t} = -\frac{\partial}{\partial x}[M(x)v(x,t)] + \frac{1}{4n}\frac{\partial^2}{\partial x^2}[x(1-x)v(x,t)]. \tag{13}$$

Mutations are represented by the drift term with the variable coefficient $M(x)$. For a model with one-way mutations, it is defined by

$$M(x) = -\mu x, \tag{14}$$

see for instance [31]. We rescale the mutation rate μ by

$$\mu^* = 2n\mu \tag{15}$$

and use the stretched time variable in (8). The rescaled diffusion equation will be

$$x(1-x)\frac{\partial^2 v}{\partial x^2} + [2 - (3+\beta)x]\frac{\partial v}{\partial x} - (1+\beta)v = 2\frac{\partial v}{\partial \tau}, \tag{16}$$

where we have introduced the new mutation parameter β defined by

$$\beta = 1 - 2\mu^* = 1 - 4n\mu. \tag{17}$$

Since $\mu \ge 0$, we have $\beta \le 1$, where $\beta = 1$ represents the case of no mutation, and $\beta = 0$ is a model which conditionally on non-fixation of the a allele behaves very similar to a mutation free Wright–Fisher model. It follows from Section 5.5 of [15] that $x = 1$ is a regular boundary point for $0 < \beta < 1$ and an entrance boundary point for $\beta \le 0$.

In order to analyze (16), the interior $0 < x < 1$ and the fixation boundary $x = 0$ need separate treatment, because of the Dirac singularity at 0. In the appendix we prove that

$$\begin{aligned} v(x,\tau) &= v(x,\tau;p,\beta) \\ &= \Pi_0(t)\delta(x) \\ &\quad + p(1-p)^\beta \sum_{m=0}^{\infty} (2m+2-\beta)(m+1-\beta)(m+1) \\ &\qquad \cdot {}_2F_1(\beta-m, m+2; 2; p) \\ &\qquad \cdot {}_2F_1(\beta-m, m+2; 2; x)e^{-\lambda_m\tau/2} \\ &= \Pi_0(t)\delta(x) + \sum_{m=0}^{\infty} r_m(p;\beta)l_m(x;\beta)e^{-\lambda_m\tau/2}, \end{aligned} \tag{18}$$

is a solution of (16) when $\beta < 1$ and $0 < p \le 1$, that is equivalent to the one presented in Section 8.5 of [8], see in particular formulas 8.5.17 and 8.5.19. An advantage of (18) is its simpler parametrization in terms of hypergeometric functions ${}_2F_1(a, b; c; x)$, which we will use below to explain the behaviour of the diffusion solution as $\beta \to 1^-$. The number

$$\lambda_m = \lambda_m(\beta) = (m+1)(m+1-\beta) \tag{19}$$

is the m:th eigenvalue of a differential operator corresponding to minus the left hand side of (16). It follows from (56) in the appendix that (18) is well defined as $p \to 1$ for all values of β, and in particular that the right eigenfunction

$$r_m(p;\beta) = (2m+2-\beta)p(1-p)^\beta {}_2F_1(\beta-m, m+2; 2; p) \tag{20}$$

is a polynomial in p of degree $m + 1$, closely related to a Jacobi polynomial. The left eigenfunction

$$l_m(x;\beta) = (m+1)(m+1-\beta){}_2F_1(\beta-m, m+2; 2; x) \tag{21}$$

is typically not a polynomial in x though, unless β is an integer. The normalizing factor $(m+1)(m+1-\beta)$ guarantees that l_m integrates to 1 for all m. The fixation probability

$$\begin{aligned} \Pi_0(\tau) &= \Pi_0(\tau;p,\beta) \\ &= 1 - p(1-p)^\beta \sum_{m=0}^{\infty} (2m+2-\beta)e^{-\lambda_m\tau/2} \\ &\qquad \cdot {}_2F_1(\beta-m, m+2; 2; p) \\ &= 1 - \sum_{m=0}^{\infty} r_m(p)e^{-\lambda_m\tau/2} \end{aligned} \tag{22}$$

is represented as a coefficient of a Dirac singularity at the $x = 0$ boundary in the diffusion approximation.

Any choice of initial distribution (12) will ultimately lead to $\Pi_0(\infty) = 1$, provided $\mu > 0$. Therefore the only steady-state solution is that of full fixation in gamete a

$$v(x,\infty) = \Pi_0(\infty)\delta(x) = \delta(x). \tag{23}$$

Consequently, the uniform steady state with extinction of gamete A is independent of both the initial condition (founder population) and the magnitude of the mutation rate $\mu > 0$.

There are several interesting features of (18). First, we motivate in the appendix that for fixed τ, the diffusion solution can be written as

$$v(x,\tau) = \Pi_0(\tau)\delta(x) + C(x,\tau)\cdot(1-\beta)(1-x)^{-\beta}, \tag{24}$$

where $C(x, \tau)$ is a bounded function, with an expansion

$$C(x,\tau) = (2-\beta)pe^{-(1-\beta)\tau/2} + o\left(e^{-(1-\beta)\tau/2}\right) \text{ as } \tau \to \infty, \tag{25}$$

first noted by Wright [41], and equivalent to formula 8.5.18 of [8]. This implies that v is integrable in x for each $\beta < 1$, but yet unbounded at $x = 1$ whenever this boundary point is regular ($0 < \beta < 1$), with a pole of order β.

Second, it is instructive to investigate the limit of (18) or (24) when mutations are removed. The size of the pole of v at the right boundary gets increasingly large and the leading eigenvalue vanishes ($\lambda_0 \to 0$) when $\beta \to 1^-$. But v must be transient between the two boundaries for the mutation free model because of genetic drift. And since the $m = 0$ term of (18) vanishes in the interior and becomes non-transient ($\lambda_0 \to 0$) when $\beta \to 1^-$, it must be removed and replaced by a point mass at the $x = 1$ boundary when $\beta = 1$. This implies that the limit of (18) and (22) as $\beta \to 1^-$ differs from the classical diffusion solution

$$\begin{aligned} v(x,\tau;p,1) = {}& \Pi_0(\tau;p,1)\delta(x) + \Pi_1(\tau;p,1)\delta(x-1) \\ & + p(1-p)\sum_{m=1}^{\infty}(2m+1)m(m+1) \\ & \cdot {}_2F_1(1-m,m+2;2;p) \\ & \cdot {}_2F_1(1-m,m+2;2;x)e^{-m(m+1)\tau/2}, \end{aligned} \tag{26}$$

of the mutation free Wright–Fisher model, with absorption probabilities

$$\begin{aligned} \Pi_1(\tau;p,1) &= p + p(1-p)\sum_{m=1}^{\infty}(2m+1)(-1)^m e^{-m(m+1)\tau/2} \\ & \quad \cdot {}_2F_1(1-m,m+2;2;p), \\ \Pi_0(\tau;p,1) &= \Pi_1(\tau;1-p,1) \\ &= 1-p+p(1-p)\sum_{m=1}^{\infty}(2m+1)(-1)^m e^{-m(m+1)\tau/2} \\ & \quad \cdot {}_2F_1(1-m,m+2;2;1-p) \end{aligned} \tag{27}$$

at the two boundaries. For the mutation free model, the density in (26) for $0 < x < 1$ first appeared in [27], and the fixation probability (27) in [28]. See also formulas 8.4.3 and 8.8.3.4 of [8], and [31]. The most obvious difference between the mutation free and one-way mutation models, is that a second point mass at $x = 1$ is added to the steady state solution (23) when mutations are removed, so that

$$\begin{aligned} v(x,\infty;p,1) &= \Pi_0(\infty)\delta(x) + \Pi_1(\infty)\delta(x-1) \\ &= (1-p)\delta(x) + p\delta(x-1). \end{aligned} \tag{28}$$

For a fixed time point $\tau > 0$, and any $0 < x < 1$, we also have that

$$\begin{aligned} v(x,\tau;p,\beta) &\to v(x,\tau;p,1) \\ \Pi_0(\tau;p,\beta) &\to \Pi_0(\tau;p,1) \\ 0 = \Pi_1(\tau;p,\beta) &\not\to \Pi_1(\tau;p,1) \text{ as } \beta \to 1^-. \end{aligned} \tag{29}$$

While the first and third equations of (29) follow easily from (18) and (26)–(28), the second one is less obvious, see the appendix for a proof. Since the total probability mass (10) is conserved over time, a consequence of (29) is that although the integrand of the non-fixation probability

$$\int_{0^+}^{1^-} v(x,\tau;p,\beta)dx \overset{\beta\to 1^-}{\to} \int_{0^+}^{1^-} v(x,\tau;p,1)dx + \Pi_1(\tau;p,1) \tag{30}$$

converges pointwise for any fixed $0 < x < 1$, a part $\Pi_1(\tau;\, p,\, 1)$ of the integral still escapes to the right hand boundary as mutations are removed.

Third, the quasi equilibrium distribution of the diffusion solution is the conditional distribution

$$v_{quasi}(x;\beta) = \lim_{\tau\to\infty}\frac{v(x,\tau;p,\beta)}{\int_{0^+}^{1^-} v(y,\tau;p,\beta)dy} \tag{31}$$

of the frequency of the A allele conditionally on non-fixation, in the limit of large time points. It follows from (18), our normalization of the left eigenfunctions in (21), and formula (56) of the appendix, that

$$\begin{aligned} & v_{quasi}(x;\beta) \\ & = \begin{cases} l_0(x;\beta) = (1-\beta)\,{}_2F_1(\beta,2;2;x) = (1-\beta)(1-x)^{-\beta}, & \beta < 1, \\ {}_2F_1(0,3;2;x) = 1, & \beta = 1. \end{cases} \end{aligned} \tag{32}$$

We may also derive the upper part of (32) directly from (24) and (25). The interesting conclusion of (32) is that v_{quasi} is uniform for $\beta \in \{0, 1\}$, skewed to the left for $\beta < 0$ and skewed to the right for $0 < \beta < 1$. Again, the discontinuity at $\beta = 1$ is due to that part of the quasi equilibrium distribution "gets lost" at the right boundary as $\beta \to 1^-$.

Fourth, the rate of fixation is determined by (twice) the smallest positive eigenvalue, i.e.

$$\lambda_{min} = \begin{cases} \lambda_0(\beta) = 1-\beta = 4nu, & \beta < 1, \\ \lambda_1(1) = 2, & \beta = 1. \end{cases} \tag{33}$$

When $\beta = 1$, fixation of either allele can occur through genetic drift. With rescaled time (15), this occurs at a rate $\lambda_{min}/2 = 1$ that is independent of population size. When $\beta < 1$, only the a allele can be fixed, with a rate $\lambda_{min}/2 = \mu^*$ proportional to the mutation probability u. This rate is lower than 1 when $0 < \beta < 1$, but since genetic drift works in both directions, it may cause the mutant allele A to dominate the population temporarily, so that the rate of final fixation is still smaller than the genetic drift rate of 1.

5. Comparisons with the exact Markov chain

In this section we investigate numerically for which parameters of the Wright–Fisher model the diffusion solution $v(\,\cdot,\, \tau)$ in (18) provides a good approximation of the exact Markov chain distribution $V^{(2n\tau)}$ in (5). The diffusion approximation represents a similarity solution that is supposed to be asymptotically valid for very large population sizes n, and several general results are available. Shimakura [35] gives convergence results for a class of multiple alleles models with mutation, migration and selection, and Ethier and Kurz [11] provide a mathematical theory for Markov chain convergence towards diffusion processes.

Let us first compare T_n, the Markov chain's rescaled time until the a allele gets fixed, with the corresponding fixation time T of the diffusion process. Theorem 10.2.4 of [11] implies convergence

$$\begin{aligned} P(T_n \le \tau) &= V_0^{([2n\tau])} \\ &= \Pi_{n0}(\tau;p,\beta) \\ &\to \Pi_0(\tau;p,\beta) \\ &= P(T \le \tau) \end{aligned} \tag{34}$$

as $n \to \infty$ of the fixation probability up to time $\tau > 0$, for any $0 < p \le 1$ and $\beta \le 1$. It is shown in the appendix that (34) implies

$$E(T_n) \to E(T) \text{ as } n \to \infty, \tag{35}$$

and an explicit formula for $E(T)$ appears on Page 96 of [15]. The accuracy of (34) and (35) was investigated numerically by Ewens [13] in a slightly different context - the haploid and mutation free Wright–Fisher model, with or without selection, and Kimura [30] studied the accuracy of (35) for diploid Wright–Fisher models with one-way mutations and various selection schemes.

In the interior domain, we use Theorem 10.1.1 of [11] to deduce weak convergence of the Markov chain distribution towards the diffusion solution for each $\tau > 0$. Because of (34), we can phrase this solely in terms of the rescaled probabilities $\nu_n(x,\tau) = \nu_n(x,\tau;p,\beta)$ in (9), as

$$\frac{1}{2n}\sum_{i=1}^{[2nx]}\nu_n(\frac{i}{2n},\tau) \to \int_{0^+}^{x}\nu(y,\tau)dy$$

when $n \to \infty$ for all $0 < x \le 1$. Theorem 1 of [12] provides a rate of convergence

$$\left|[\Pi_{n0}(\tau)-\Pi_0(\tau)]f(0)+\frac{1}{2n}\sum_{i=1}^{2n}\nu_n\left(\frac{i}{2n},\tau\right)f\left(\frac{i}{2n}\right) - \int_{0^+}^{1}\nu(y,\tau)f(y)dy\right| \le \frac{B(f,\beta)}{2n} \qquad (36)$$

for expected values of functions f of the Markov chain that have six continuous derivatives on [0, 1], and a constant $B(f,\beta)$ that is independent of τ and p. Ewens [14] looked at diffusion approximations of the quasi equilibrium distribution

$$\nu_{n,quasi}(x,\beta) = \lim_{\tau\to\infty}\frac{\nu_n(x,\tau)}{\frac{1}{2n}\sum_{i=1}^{2n}\nu_n(\frac{i}{2n},\tau)}$$

that improve upon ν_{quasi} in (32). In more detail, he found a function $b_{quasi}(x;\beta)$ for which

$$\nu_{n,quasi}(x;\beta) = \nu_{quasi}(x;\beta) + \frac{b_{quasi}(x;\beta)}{2n} + o(n^{-1}) \qquad (37)$$

for any $0 < x = i/(2n) < 1$. Formulas (36) and (37) suggest that the diffusion solution $v(x,\tau)$ approximates the rescaled probability $\nu_n(x,\tau)$ at an accuracy inversely proportional to population size. We formulate this as a local limit result

$$\nu_n(x,\tau) = \nu(x,\tau) + \frac{b(x,\tau)}{2n} + o(n^{-1}), \qquad (38)$$

for some function $b(x,\tau) = b(x,\tau;p,\beta)$, whenever $0 < x = i/(2n) < 1$ and $\tau > 0$. In order to motivate (38), we compare the series expansion (18) of the diffusion solution, with one for the Markov chain. To this end, we first rewrite (18) as

$$\nu(x,\tau) = \sum_{m=0}^{\infty} h_m(x;p)\exp(-\lambda_m\tau/2), \qquad (39)$$

for $x > 0$, with $h_m(x;p) = h_m(x;p,\beta) = r_m(p;\beta)l_m(x;\beta)$ the product of the m:th right and left eigenfunctions. For the Markov chain we use a Jordan decomposition

$$M = Q^{-1}DQ \qquad (40)$$

of its transition matrix (2). The matrix D is block diagonal in general, with its eigenvalues along the diagonal, see for instance [6] and [5]. Feller [16] proved that

$$d_k = \begin{cases} 1, & k = 0, \\ \left(1-\frac{1-\beta}{4n}\right)^k \prod_{l=1}^{k-1}(1-\frac{l}{2n}), & k = 1,\dots,2n, \end{cases} \qquad (41)$$

for the Wright–Fisher model with one-way mutations, with $\prod_{l=1}^{0}$ interpreted as 1 when $k=1$. Since the eigenvalues in (41) are all real valued and different, it follows that $D = \mathrm{diag}(d_0,\dots,d_{2n})$ is diagonal, so that the left eigenvectors $l_k = (l_{k0},\dots,l_{k,2n})$ and right eigenvectors $r_k = (r_{k0},\dots,r_{k,2n})'$ of M with eigenvalues d_k are the rows of Q and columns of Q^{-1}, i.e.

$$Q = \begin{pmatrix} l_0 \\ l_1 \\ \vdots \\ l_{2n} \end{pmatrix}, \quad Q^{-1} = \begin{pmatrix} r_0 & r_1 & \dots & r_{2n} \end{pmatrix}.$$

These rows and columns can be normalized after convenience as long as $l_k r_k = 1$. In particular, if $r_1 = (1,\dots,1)$ we choose the first left eigenvector $l_1 = (1,0,\dots,0)$ as the asymptotic distribution of M, corresponding to fixation of the a allele.

In the appendix we prove that the Jordan decomposition (40) implies a series expansion

$$\nu_n(x,\tau) = \sum_{m=0}^{2n-1} h_{nm}(x;p)\exp(-\lambda_{nm}\tau/2) \qquad (42)$$

of the renormalized absolute probability vector. For each term m, the quantities $\lambda_{nm} = -4n\log(d_{m+1})$ and $h_{nm}(x;p) = h_{nm}(x;p,\beta) = r_{m+1,2np}\cdot 2nl_{m+1,2nx}$ can be viewed as analogues of the diffusion solution's mth eigenvalue λ_m and corresponding product of eigenvectors $h_m(x;p)$, in (39). Formula (38) will follow if the latter quantities approximate the former at an accuracy inversely proportional to population size. Therefore, we assume functions $b_m = b_m(\tau,\beta)$ and $\tilde{b}_m(x;p) = \tilde{b}_m(x;p,\beta)$ exist, such that

$$\exp(-\lambda_{nm}\tau/2) = \exp(-\lambda_m\tau/2)\left[1+\frac{b_m}{2n}+o(n^{-1})\right] \qquad (43)$$

and

$$h_{nm}(x;p) = h_m(x;p) + \frac{\tilde{b}_m(x;p)}{2n} + o(n^{-1}). \qquad (44)$$

We then take the difference of (42) and (39), and use (43) and (44) to deduce that (38) holds, with

$$b(x,\tau) = \sum_{m=0}^{\infty}\left[b_m + \tilde{b}_m(x;p)\right]\exp(-\lambda_m\tau/2).$$

In the appendix we prove (43), and in order to motivate (44), we use that the right and left eigenvectors of Q may be normalized so that the two functions $p \to r_{m+1,2np}$ and $x \to 2nl_{m+1,2nx}$ approximate the right and left eigenfunctions $r_m(p)$ and $l_m(x)$ of the diffusion operator in (20) and (21). Using similar methods of proof as in Chapter 13 of [23] and in [4], it follows from (3) and moment properties of the binomial distribution that $r_{m+1,2np}$ is a polynomial in p of degree $m+1$, just as $r_m(p)$. Notice also that (44) is related to (37) when $m=1$, since the quasi equilibrium distribution of the Markov chain and diffusion solutions are proportional to $h_{n1}(x;p)$ and $h_1(x;p)$ respectively, as functions of x.

Table 1 gives numerical results for a scenario with a small mutation rate ($\beta = 0.5$), with both alleles starting at the same allele frequency ($p = 0.5$). It is seen that all four quantities $\nu_n(x,\tau)$, $\Pi_{n0}(\tau)$, λ_{nm} and $h_{nm}(x)$ converge rather quickly as n grows. The approximation errors of v and h_m are inversely proportional to population size, in agreement with (38) and (44), whereas $\Pi_{n0}(\tau)$ seems to be converging at a slower rate. We also investigated another scenario with a larger mutation rate ($\beta = -2$) where the A allele starts at a higher frequency ($p = 0.9$). All four quantities converge, although at a somewhat slower rate, and formulas (38) and (44) were confirmed in this setting as well (results not shown).

In Table 2 we investigate how the mutation rate β, and the initial frequency p of the A allele affect accuracy of the approximate fixation probability (22) for $n = 50$ individuals. When β varies, we compare the diffusion solution with the exact Markov chain, under the assumption of a symmetric initial state, $p = 0.5$. When p varies, we first choose a relatively large mutation rate, represented

Table 1
Convergence of fixation probability $\Pi_{n0}(0)$, approximate diffusion function $v_n(x, \tau)$, transformed eigenvalue λ_{nm} and left eigenvector $h_{nm}(x)$ towards their diffusion ($n = \infty$) limits $\Pi_0(\tau)$, $v(x, \tau)$, λ_m and $h_m(x)$, when $\beta = 0.5$ and $p = 0.5$. For each quantity z_n ($= \Pi_n(0)$, $v_n(x, \tau)$ or $h_{nm}(x)$) of the Markov chain, the number $2n(z_n - z)$ in brackets refers to an approximation error of the corresponding diffusion quantity z ($= \Pi(0)$, $v(x, \tau)$ or $h_m(x)$), normalized by population size.

$\beta = 0.5,\ p = 0.5$					
τ	n	$\Pi_{n0}(0)$	$v_n(0.25, \tau)$	$v_n(0.5, \tau)$	$v_n(0.75)$
0.25	4	0.0395 (0.2618)	1.1582 (−0.2488)	1.5076 (−0.1899)	0.8999 (−0.2619)
	20	0.0131 (0.2506)	1.1831 (−0.2489)	1.5264 (−0.1976)	0.9253 (−0.2955)
	100	0.0083 (0.2880)	1.1880 (−0.2487)	1.5303 (−0.1989)	0.9312 (−0.3028)
	400	0.0072 (0.2832)	1.1890 (−0.2487)	1.5311 (−0.1991)	0.9323 (−0.3042)
	Diffusion	0.0068	1.1893	1.5313	0.9327
1	4	0.3427 (0.5339)	0.6331 (−0.3024)	0.6343 (−0.1495)	0.6121 (−0.3169)
	20	0.2938 (0.7136)	0.6635 (−0.2941)	0.6492 (−0.1515)	0.6439 (−0.3107)
	100	0.2804 (0.8976)	0.6694 (−0.2920)	0.6522 (−0.1518)	0.6501 (−0.3137)
	400	0.2773 (1.0558)	0.6705 (−0.2916)	0.6528 (−0.1518)	0.6513 (−0.3143)
	Diffusion	0.2759	0.6709	0.6530	0.6517
4	4	0.7438 (0.1705)	0.1478 (−0.1231)	0.1857 (−0.0957)	0.2532 (−0.1868)
	20	0.7277 (0.2108)	0.1600 (−0.1251)	0.1953 (−0.0967)	0.2721 (−0.1760)
	100	0.7237 (0.2499)	0.1625 (−0.1253)	0.1972 (−0.0967)	0.2756 (−0.1769)
	400	0.7228 (0.2832)	0.1630 (−0.1253)	0.1976 (−0.0968)	0.2763 (−0.1771)
	Diffusion	0.7225	0.1632	0.1977	0.2765
m	n	λ_{nm}	$h_{nm}(0.25)$	$h_{nm}(0.5)$	$h_{nm}(0.75)$
0	4	0.5080	0.4000 (−0.2639)	0.5072 (−0.1852)	0.6977 (−0.4184)
	20	0.5016	0.4264 (−0.2656)	0.5257 (−0.1862)	0.7404 (−0.3843)
	100	0.5003	0.4317 (−0.2656)	0.5294 (−0.1860)	0.7481 (−0.3861)
	400	0.5001	0.4327 (−0.2656)	0.5301 (−0.1860)	0.7495 (−0.3864)
	Diffusion	0.5000	0.4330	0.5303	0.7500
1	4	3.1525	1.6119 (0.3919)	1.1483 (0.8339)	0.3051 (0.4723)
	20	3.0286	1.5728 (0.3953)	1.0634 (0.7743)	0.2553 (0.3663)
	100	3.0056	1.5649 (0.3977)	1.0479 (0.7630)	0.2479 (0.3575)
	400	3.0014	1.5634 (0.3981)	1.0450 (0.7609)	0.2465 (0.3558)
	Diffusion	3.000	1.5629	1.0441	0.2461
2	4	8.2634	−0.6439 (0.0117)	0.1902 (−0.5293)	0.5051 (−0.5511)
	20	7.6336	−0.6475 (−0.0842)	0.2436 (−0.5098)	0.5596 (−0.5762)
	100	7.5261	−0.6459 (−0.1022)	0.2538 (−0.5059)	0.5711 (−0.5779)
	400	7.5065	−0.6455 (−0.1054)	0.2557 (−0.5052)	0.5733 (−0.5783)
	Diffusion	7.5000	−0.6454	0.2564	0.5740

by $\beta = -1$ (and hence $\mu = 0.005$). The diffusion solution is then in fairly good agreement with the tabulated results for the exact Markov chain. This is also true in the limit of a vanishingly small mutation rate ($\beta \to 1$). Formula (29) implies that the the limiting one-way mutation diffusion model predicts the same fixation probability at the left boundary as the mutation free diffusion model. It also gives good agreement with the exact Markov chain for 50 individuals, at least when τ is of order one or larger.

In Table 3 we investigate the accuracy of the diffusion approximation (38) close to the left and right boundaries for two mutation parameters. The lower mutation rate ($\beta = 0$) has a diffusion density v close to uniform (cf. (32)), so that the convergence rates at the two boundaries are more easily compared. The higher mutation rate ($\beta = 0.8$) has a pole at the right boundary. It is seen that the diffusion solution predicts too high probabilities close to the fixation boundary $x = 0$, whereas the convergence rate is faster at the non-fixation boundary $x = 1$. This is the case for both mutation rates, as well as for $\beta = -0.8$ (results not shown). Tyvand and Thorvaldsen [39] made a similar analysis for the mutation free Wright–Fisher model ($\beta = 1$), and found that the diffusion approximation was too large close to both fixation boundaries.

Table 4 illustrates the accuracy of the diffusion solution for the extreme initial condition of a homogeneous population of A alleles ($p = 1$), so that the distance to the steady state population is maximal. Convergence is checked towards the diffusion density (38) in the interior and at the right boundary, towards the fixation probability (34) at the left boundary, and also possible convergence

$$\Pi_{n1}(\tau) = V_{2n}^{([2n\tau])} \to \Pi_1(\tau) = 0 \tag{45}$$

at the right boundary. We see from Table 4 that the local convergence to the diffusion limit is acceptable for all values of x, with a possible exception for the non-fixation boundary $x = 1$.

In several of the tables (in particular Table 2), it can seen that the convergence rate of the fixation probability at the $x = 0$ boundary, as n increases, is slower the smaller τ is. The diffusion model generally predicts a fixation probability below that of the exact Markov chain. A similar phenomenon was observed by Ewens [13] in the context of the neutral model $\beta = 1$, and two fixation boundaries. In both cases, this is due to the fact that the Wright–Fisher model discretizes time, so that large jumps to fixation are possible in a small population, whereas the diffusion approach does not account for this. This also explains why the diffusion solution predicts too high values close to the fixation boundary, as found in Table 3.

We finally consider the case of a small but positive mutation rate μ^*. In the appendix we motivate that (45) breaks down at the right boundary when μ^* is close to 0. Instead, we suggest to perturb the diffusion solution (18) around the non-fixation boundary $x = 1$, as

$$\begin{aligned} v_n^*(x, \tau; p, \beta) = {} & \Pi_0(\tau; p, \beta)\delta(x) \\ & + v(x, \tau; p, \beta) \cdot 1_{\{0<x\le 1-1/(4n)\}} \\ & + \Pi_{n1}^*(\tau; p, \beta)\delta(x-1). \end{aligned} \tag{46}$$

This is achieved by moving a probability mass of the diffusion solution v from the interval $[1 - 1/(4n), 1)$ adjacent to 1, to a point

Table 2
Fixation probability $\Pi_{n0}(\tau)$ according to the Wright–Fisher model for different values of the mutation parameter β and the initial frequency p of the A allele. The exact Markov chain is computed for populations of $n = 50$ individuals. The lower number in each box represents the diffusion approximation $\Pi_0(\tau)$ of the fixation probability, cf. (22).

$p = 0.5$

τ	$\beta = -2$	$\beta = -1$	$\beta = 0$	$\beta = 0.5$	$\beta = 0.9$	$\beta = 1$
0.25	.0222	.0160	.0114	.0095	.0082	.0079
	.0157	.0114	.0081	.0068	.0059	.0057
0.5	.1980	.1478	.1062	.0887	.0763	.0734
	.1786	.1334	.0961	.0804	.0693	.0667
1	.5845	.4643	.3416	.2841	.2416	.2315
	.5680	.4504	.3315	.2759	.2349	.2252
2	.9058	.8034	.6289	.5172	.4241	.4014
	.9007	.7970	.6230	.5126	.4209	.3985

$\beta = -1$

τ	$p = 0.01$	$p = 0.1$	$p = 0.3$	$p = 0.7$	$p = 0.9$	$p = 0.99$
0.25	.9344	.4968	.1040	.0015	4×10^{-5}	2×10^{-6}
	.9290	.4683	.0861	.0009	2×10^{-5}	7×10^{-7}
0.5	.9686	.7202	.3481	.0509	.0114	.0041
	.9669	.7078	.3290	.0433	.0088	.0029
1	.9868	.8732	.6500	.3124	.1910	.1454
	.9864	.8688	.6393	.2977	.1775	.1330
2	.9960	.9598	.8807	.7279	.6542	.6216
	.9958	.9584	.8767	.7193	.6435	.6100

$\beta = 1$

τ	$p = 0.01$	$p = 0.1$	$p = 0.3$	$p = 0.7$	$p = 0.9$	$p = 0.99$
0.25	.9250	.4468	.0725	.0424	4×10^{-5}	10^{-8}
	.9197	.4214	.0603	.0003	2×10^{-6}	7×10^{-9}
0.5	.9589	.6480	.2431	.0150	.0011	3×10^{-5}
	.9573	.6370	.2303	.0130	.0008	3×10^{-5}
1	.9770	.7857	.4545	.0932	.0189	.0014
	.9765	.7819	.4475	.0893	.0177	.0013
2	.9860	.8637	.6062	.2181	.0653	.0062
	.9859	.8626	.6137	.2158	.0643	.0061

mass at 1, so that

$$\Pi^*_{n1}(\tau; p, \beta) = \int_{1-1/(4n)}^{1-} v(x, \tau)dx \approx \frac{C(1, \tau)}{(4n)^{1-\beta}}, \qquad (47)$$

where $C(1, \tau)$ is defined in (24). This perturbed diffusion solution (46) is mainly of interest when the population size is small and/or when $x = 1$ is a regular boundary point ($0 < \beta < 1$). For such a boundary point we notice that the right hand side of (47) will exceed the order $O(n^{-1})$ of the Markov chain probabilities at inner points.

It can be seen from Table 5 that both $\Pi^*_{n1}(\tau)$ and $C(1, \tau)/(4n)^{1-\beta}$ are good approximation of $\Pi_1(\tau)$ when β is close to 1, although $\Pi^*_{n1}(\tau)$ is slightly more accurate for small τ. The diffusion solution, on the other hand, provides a very poor approximation for such β, but it is accurate for $\beta = 1$. In Table 4 we find that $\Pi^*_{n1}(\tau)$ is less accurate for $\beta = 0$, but it still improves upon the diffusion solution.

We interpret $\Pi^*_{n1}(\tau; p, \beta)$ as a quasi-fixation probability, and $C(1, \tau)$ as an approximate quasi-fixation probability for mutation rates so small that the denominator of (47) is close to 1. It follows from (25) that for small μ^* and large τ,

$$C(1, \tau) \approx pe^{-\mu^* \tau} = pe^{-\lambda_{min}\tau/2}$$

tends to zero at a rate determined by the smallest eigenvalue (33) of the diffusion operator. In the appendix we also provide an exact formula for $C(1, \tau)$.

It is possible to get a more probabilistic interpretation of quasi-fixation. Let $0 \le T_{n1} \le \infty$ be the stochastic and rescaled time at which the population for the first time becomes homogeneous with A alleles only. When $p < 1$, this may never happen, and in this case we put $T_{n1} = \infty$. Then, let T_{n2} be the time, after this event, that it takes for a “successful” new mutation $A \to a$ to occur, i.e. one that spreads to the whole population. The previous unsuccessful $A \to a$ mutations that happen after the A allele has taken over, will typically be short-lived, compared to the time it takes for the successful mutation to occur. This suggests that the A allele is quasi-fixed during most of the time interval $[T_{n1}, T_{n1} + T_{n2})$, and that

$$\Pi^*_{n1}(\tau; p, \beta) \approx P(T_{n1} \le \tau, T_{n1} + T_{n2} > \tau) \qquad (48)$$

is approximately the probability that up to time τ, a successful mutation has not yet occurred in an environment of A alleles only. In order to motivate (48) we need to approximate the distribution of T_{n1} and T_{n2}. We first notice that

$$P(T_{n1} \le \tau) \approx \Pi_1(\tau; p, 1) \qquad (49)$$

when $\mu^* > 0$ is small, since a Wright–Fisher model with a small one-way mutation rate behaves like a neutral model between boundaries, as a consequence of the fact that genetic drifts dominates over the systematic drift towards the left boundary. Suppose n is first chosen large, and then $\mu^* > 0$ is taken to be small. In the appendix we then show that the time T_{n2} of quasi-fixation is approximately exponentially distributed with rate parameter μ^*, and that

$$\begin{aligned}\Pi^*_{n1}(\tau; p, \beta) &\approx C(1, \tau)\\ &\approx \int_0^\tau f_{T_{n1}}(s)P(T_{n2} > \tau - s)ds\\ &\approx \int_0^\tau \Pi_1'(s; p, 1)e^{-\mu^*(\tau - s)}ds, \end{aligned} \qquad (50)$$

where $f_{T_{n1}}$ is the density function of T_{n1}. One consequence of (50) is that $C(\tau, 1)$ approaches the fixation probability $\Pi_1(\tau; p, 1)$ of the mutation free model at the right boundary, as $\mu^* \to 0$.

Table 3
Values of the approximate diffusion function $v_n(x, \tau)$ for different population sizes n and frequencies $0 < x < 1$ of the A allele that conform with (7) and are close to the left and right boundaries $x = 0$ and 1. For comparison, the diffusion density $v(x, \tau)$ is given as well. The rescaled time variable $\tau = 2$, and the initial frequency of the A allele is $p = 0.5$.

		x close to left boundary						
n	β	$\to 0$	1/320	1/160	1/80	1/40	1/20	1/10
5	0	–	–	–	–	–	–	0.3176
10		–	–	–	–	–	0.3224	0.3610
20		–	–	–	–	0.3254	0.3647	0.3797
40		–	–	–	0.3274	0.3669	0.3829	0.3892
80		–	–	0.3287	0.3683	0.3847	0.3921	0.3939
160		–	0.3294	0.3692	0.3857	0.3937	0.3967	0.3963
Diff		0.4040	0.4038	0.4036	0.4033	0.4026	0.4013	0.3987
5	0.8	–	–	–	–	–	–	0.2051
10		–	–	–	–	–	0.2049	0.2334
20		–	–	–	–	0.2055	0.2321	0.2456
40		–	–	–	0.2062	0.2320	0.2439	0.2518
80		–	–	0.2068	0.2322	0.2434	0.2499	0.2549
160		–	0.2072	0.2324	0.2433	0.2492	0.2529	0.2564
Diff		0.2541	0.2542	0.2543	0.2545	0.2549	0.2559	0.2580

		x close to right boundary						
n	β	1−1/10	1−1/20	1−1/40	1−1/80	1−1/160	1−1/320	$\to 1$
5	0	0.3190	–	–	–	–	–	–
10		0.3334	0.3306	–	–	–	–	–
20		0.3440	0.3382	0.3391	–	–	–	–
40		0.3494	0.3452	0.3421	0.3449	–	–	–
80		0.3522	0.3486	0.3465	0.3448	0.3487	–	–
160		0.3536	0.3504	0.3486	0.3476	0.3465	0.3510	–
Diff		0.3550	0.3522	0.3508	0.3501	0.3497	0.3496	0.3494
5	0.8	0.6180	–	–	–	–	–	–
10		0.5943	0.9983	–	–	–	–	–
20		0.6186	0.9469	1.6674	–	–	–	–
40		0.6240	0.9860	1.5706	2.8391	–	–	–
80		0.6271	0.9933	1.6365	2.6654	4.8857	–	–
160		0.6286	0.9976	1.6478	2.7788	4.5796	8.4551	–
Diff		0.6302	1.0020	1.6616	2.8208	4.8485	8.3869	∞

Table 4
Boundary probabilities $\Pi_{n0}(\tau)$, $\Pi_{n1}(\tau)$ and rescaled probabilities $v_n(x, \tau)$, for three equispaced choices of x, are given for $\beta = 0$ and a homogeneous founder population with only the A gamete $(p = 1)$. Results for the exact Markov chain are given for increasing population sizes n and for the diffusion limits (18) and (22) of $v_n(x, \tau)$ and $\Pi_{n0}(\tau)$. The right column refers to the quasi fixation probability in (47).

$\beta = 0,\ p = 1$							
n	τ	$\Pi_{n0}(\tau)$	$v_n(0.25, \tau)$	$v_n(0.5, \tau)$	$v_n(0.75, \tau)$	$\Pi_{n1}(\tau)$	$\Pi^*_{n1}(\tau)$
4	1	0.0814	0.5956	0.8799	1.1678	0.1492	0.1098
32		0.0424	0.5814	0.9065	1.3013	0.0241	0.0142
128		0.0379	0.5795	0.9092	1.3169	0.0063	0.0036
Diff		0.0361	0.5788	0.9101	1.3221	0.0000	0.0000
4	2	0.3802	0.5928	0.6477	0.6697	0.0715	0.0502
32		0.3136	0.6383	0.6923	0.7425	0.0109	0.0063
128		0.3044	0.6431	0.6971	0.7510	0.0028	0.0016
Diff		0.3006	0.6446	0.6988	0.7538	0.0000	0.0000
4	4	0.7727	0.2332	0.2407	0.2354	0.0240	0.0170
32		0.7370	0.2644	0.2663	0.2663	0.0037	0.0021
128		0.7320	0.2678	0.2691	0.2698	0.0009	0.0005
Diff		0.7300	0.2690	0.2700	0.2710	0.0000	0.0000

Recall from (34) that the distribution of the fixation time T_n of the a-allele is approximated by the diffusion solution, as

$$P(T_n \le \tau) \approx \Pi_0(\tau; p, \beta).$$

This distribution will be very skewed for small mutation rates when $p < 1$. This can be seen by rewriting it as a mixture of two components;

$$\begin{aligned} P(T_n \le \tau) &= P(T_n \le \tau, T_{n1} = \infty) + P(T_n \le \tau, T_{n1} < \infty) \\ &= P(T_n \le \tau, T_{n1} = \infty) + P(T_n \le \tau, T_{n1} \le \tau) \\ &\approx P(T_n \le \tau, T_{n1} = \infty) + P(T_{n1} + T_{n2} \le \tau, T_{n1} \le \tau) \\ &\approx \Pi_0(\tau; p, 1) + [\Pi_1(\tau; p, 1) - \Pi^*_{n1}(\tau; p, \beta)]. \end{aligned} \quad (51)$$

The first term of (51) approximates the probability of immediate fixation of the a-allele up to time τ, whereas the second term approximates the probability that quasi-fixation of the A-allele occurs at first, followed by a successful mutation $A \to a$ that spreads to the whole population. In the third step of (51) we ignored the time $T_n - (T_{n1} + T_{n2})$ it takes for the successful mutation to spread, one it has occurred, and in the fourth step we used (48) and (49).

Table 5
Probabilities of a homogeneous population of A alleles are shown for different population sizes n, time points τ, mutation parameters β and initial frequencies p of A alleles. Four types of probabilities are displayed; exact values $\Pi_{n1}(\tau)$ from the Markov chain (upper rows marked $n = 4, 128$), the quasi fixation probability $\Pi^*_{n1}(\tau)$ (middle rows marked $n = 4, 128$), the approximate quasi fixation probability $C(1,\tau)/(4n)^{1-\beta}$ (lower rows marked $n = 4, 128$) and the diffusion solution $\Pi_1(\tau)$ (rows marked Diff). The quasi fixation probabilities for the mutation free models refer limits $\beta \to 1-$.

		$p = 0.5$			$p = 1$		
n	τ	$\beta = 0.95$	$\beta = 0.99$	$\beta = 1$	$\beta = 0.95$	$\beta = 0.99$	$\beta = 1$
4	1	0.2519	0.2707	0.2756	0.9104	0.9814	1.0000
		0.2366	0.2544	0.2590	0.9142	0.9825	1.0000
		0.2078	0.2216	0.2252	0.9179	0.9833	1.0000
128		0.1773	0.2168	0.2280	0.7736	0.9499	1.0000
		0.1755	0.2151	0.2262	0.7717	0.9498	1.0000
		0.1747	0.2141	0.2252	0.7718	0.9498	1.0000
Diff		0.0000	0.0000	0.2252	0.0000	0.0000	1.0000
4	3	0.4037	0.4586	0.4735	0.8506	0.9681	1.0000
		0.3984	0.4528	0.4673	0.8507	0.9684	1.0000
		0.3949	0.4483	0.4627	0.8510	0.9685	1.0000
128		0.3333	0.4337	0.4632	0.7175	0.9357	1.0000
		0.3322	0.4332	0.4628	0.7156	0.9355	1.0000
		0.3321	0.4331	0.4627	0.7156	0.9355	1.0000
Diff		0.0000	0.0000	0.4627	0.0000	0.0000	1.0000
4	10	0.3561	0.4672	0.5000	0.7123	0.9344	1.0000
		0.3559	0.4672	0.5000	0.7119	0.9345	1.0000
		0.3559	0.4672	0.5000	0.7119	0.9345	1.0000
128		0.3001	0.4515	0.5000	0.6002	0.9028	1.0000
		0.2993	0.4513	0.5000	0.5986	0.9026	1.0000
		0.2993	0.4513	0.5000	0.5986	0.9026	1.0000
Diff		0.0000	0.0000	0.5000	0.0000	0.0000	1.0000
4	100	0.0374	0.2979	0.5000	0.0748	0.5957	1.0000
		0.0375	0.2979	0.5000	0.0750	0.5958	1.0000
		0.0375	0.2979	0.5000	0.0750	0.5958	1.0000
128		0.0316	0.2878	0.5000	0.0633	0.5757	1.0000
		0.0315	0.2878	0.5000	0.0631	0.5755	1.0000
		0.0315	0.2878	0.5000	0.0631	0.5755	1.0000
Diff		0.0000	0.0000	0.5000	0.0000	0.0000	1.0000
4	500	0.0000	0.0403	0.5000	0.0000	0.0806	1.0000
		0.0000	0.0403	0.5000	0.0000	0.0806	1.0000
		0.0000	0.0403	0.5000	0.0000	0.0806	1.0000
128		0.0000	0.0390	0.5000	0.0000	0.0779	1.0000
		0.0000	0.0389	0.5000	0.0000	0.0779	1.0000
		0.0000	0.0389	0.5000	0.0000	0.0779	1.0000
Diff		0.0000	0.0000	0.5000	0.0000	0.0000	1.0000

6. Summary and conclusions

We have analyzed the continuum representation of the biallelic and haploid Wright–Fisher model with a constant one-way mutation rate, and no selection. In order to investigate the accuracy of this diffusion solution, we compared it to the exact Wright–Fisher Markov chain for a wide range of parameter values. In general, the agreement is good at fixed sample allele frequency points $0 < x < 1$, with a rapid local convergence of the exact stochastic process towards the diffusion density, in the limit of very large populations. We also found that the local convergence of the one-way mutation model is rapid close to the non-fixation boundary, but slower close to the fixation boundary. The local convergence result was justified by comparing the Jordan decomposition of the Markov chain's transition matrix with the infinite series representation of the diffusion solution.

We also observed convergence of the fixation probability towards the diffusion solution at the left fixation boundary, over a wide range of parameter values, although the convergence rate is slower when the rescaled time parameter τ is small. On the other hand, the accuracy of the diffusion approximation is very poor at the non-fixed right boundary when the mutation rate is small, since the mutant allele may temporarily be fixed for a very long time. We developed an accurate modification of the diffusion solution that incorporates such quasi-fixation. This phenomenon is due to a discontinuity of the diffusion solution at zero mutation rates ($\beta = 1$). The solution switches from fixation at only one boundary as $\beta \to 1^-$, to fixation at both boundaries. As long as mutations of the one-way model remain, the asymptotic limit (23) is unique, with a steady state of only non-mutant alleles, although the time it takes to reach this limit increases as the mutation probability goes to zero.

The main application of our results is the infinite alleles model, where a fixed mutant allele A is of interest, and a represents all the other alleles. We argue that the quasi-fixation phenomenon will very often occur, unless the population n is very large. Indeed, it follows from (17), (46) and (47) that quasi-fixation can be neglected only if

$$4n\log(4n) \geq \frac{\log[C(1,\tau)/\varepsilon]}{L\mu_{site}}, \tag{52}$$

for a DNA sequence of L base pairs or sites, where $0 < \varepsilon < 1$ is a tolerance parameter quantifying the maximal quasi-fixation probability that can be neglected, and μ_{site} is the mutation probability per generation and site, so that the mutation probability for the whole sequence is $\mu = L\mu_{site}$. For instance, putting $C(1,\tau) = 0.5$, $\varepsilon = 0.05$ and $\mu_{site} = 10^{-8}$, the order of the mutation rate of humans ([3,25]), the right hand side of (52) simplifies to $10^8 \log(10)/L = 2.30 \cdot 10^5$ for a string of $L = 1000$ base pairs. This value is of the same order of magnitude as if the effective population size n_e of humans is plugged into the left hand side of (52). Although estimates of n_e differ between studies and types of effective sizes, it is believed to be of the order 10^4 ([37]), which gives $4.24 \cdot 10^5$ in (52). Hence it is clear that quasi-fixation of a 1 kb DNA string cannot be neglected for populations smaller than n_e.

Several extensions of our work are possible. First, Moran [32] introduced a haploid model with overlapping generations where only one gamete at a time is replaced. Karlin and McGregor [24] showed that the Moran and Wright–Fisher models share a common diffusion limit, with appropriate rescaling of time. Tyvand and Thorvaldsen [39] compared both models when no mutations are present, and showed numerically that the diffusion solution approximates the Moran model much more accurately than the Wright–Fisher model, close to both fixation boundaries. This is due to the nearest-neighbor interactions of the Moran model, whereas the many neighbor interactions of the Wright–Fisher model slows down its convergence rate close to boundaries. We conjecture that the same is true in the case of one-way mutations, so that the Moran model converges more rapidly to the diffusion solution at the fixation boundary $x = 0$ as well as in its close vicinity.

Second, quasi-fixation will also occur for a (Wright–Fisher or Moran) model with two-way mutations. Let ν^* be the normalized mutation rate $a \to A$, defined in the same way as μ^* in (15). It is well known that the steady state solution of the diffusion solution has a beta distribution

$$v(x,\infty) = \frac{\Gamma(2\mu^* + 2\nu^*)}{\Gamma(2\mu^*)\Gamma(2\nu^*)} x^{2\nu^*-1}(1-x)^{2\mu^*-1}, \quad 0 < x < 1, \tag{53}$$

where Γ is the gamma function, see for instance Sections 8.5 and 9.3 of [8]. This solution is independent of initial conditions and possesses a discontinuity as μ^* and ν^* tend to 0, since the limiting mutation free model has a two-point stationary distribution (28) that depends on the initial frequencies of the two alleles. For small but positive mutation rates and/or small populations, quasi-fixation occurs at both boundaries of the two-way model, with non-negligible probabilities. In analogy with (46), these probabilities can be approximated by integrating the diffusion density $v(\cdot, \tau)$ over intervals $(0, 1/(4n))$ and $(1 - 1/(4n), 1)$ that surround the two boundary points. These quasi-fixation probabilities will depend on initial conditions, but as $\tau \to \infty$ they converge to limiting nonzero values, independently of the founder population composition. The limiting values are obtained by integrating (53) over the same two intervals.

Third, population genetic diffusion models with $K+1$ alleles are defined on a K-simplex. They have been well studied for neutral models with mutations, see for instance [19,21,34] and [2]. A one-way model is obtained by assuming that alleles $A_1, \ldots, A_K$ are mutant, whereas all remaining alleles are collected into a final state a. For an infinite alleles model, only K types of mutations $A_i \to a$ are possible for $i = 1, \ldots, K$. In this case it would be of interest to study quasi-fixation of all mutant alleles jointly. The quasi-fixation boundary is then the $K-1$ simplex defined by all allele frequency configurations for which a is absent.

Fourth, it is possible to address quasi-fixation for a larger class of models. This is achieved by replacing the diffusion solution v by a refined approximation of the Markov chain ([40,43]). See also [18], where coalescence theory is used to improve the diffusion solution approximation of time to fixation. These refined approximations typically have a finite allele frequency space, a subset of which could represent the quasi-fixation boundary.

Acknowledgements

Ola Hössjer was supported by the Swedish Research Council, contracts no. 621-2008-4946 and 621-2013-4633, and the Gustafsson Foundation for Research in Natural Sciences and Medicine. The authors wish to thank three anonymous reviewers for valuable comments that considerably improved the presentation of the paper.

Appendix

The hypergeometric function. The infinite series representation of the hypergeometric function can be found for instance in formula 15.1.1 of [1]. It is defined as

$${}_2F_1(a, b; c; x) = \sum_{m=0}^{\infty} \frac{(a)_m (b)_m}{(c)_m m!} x^m, \tag{54}$$

where

$$(x)_m = \frac{\Gamma(x+m)}{\Gamma(x)} \tag{55}$$

is the rising factorial. Hence $(x)_0 = 1$ and $(x)_m = x(x+1) \cdot \ldots \cdot (x+m-1)$ for $m > 0$.

Hypergeometric functions and Jacobi polynomials. We can use formulas 15.3.3 and 15.4.6 of [1] to express the hypergeometric functions in (18), as

$$\begin{aligned}\phi_m(x) &:= {}_2F_1(\beta - m, m+2; 2; x) \\ &= (1-x)^{-\beta} {}_2F_1(m - \beta + 2, -m; 2; x) \\ &= \frac{1}{m+1}(1-x)^{-\beta} P_m^{(1,-\beta)}(1-2x),\end{aligned} \tag{56}$$

where $P_m^{(1,-\beta)}(z)$ is a Jacobi polynomial of order m, and in particular

$$P_0^{(1,-\beta)}(z) = 1. \tag{57}$$

The Jacobi polynomials of integer order are orthogonal over the interval $-1 < z < 1$ with respect to the weight function $(1-z)(1+z)^{-\beta}$, and constitute an orthogonal and complete set of functions over $|z| < 1$.

Equivalence of (18) and (22) with previous diffusion solutions. Formulas 8.5.17 and 8.5.19 of [8] contain the diffusion limit of the Wright–Fisher model with one-way mutations. Adapted to our notation, these two equations read

$$\begin{aligned}v(x,\tau) &= \sum_{m=0}^{\infty} \frac{(2-\beta+2m)\Gamma(2-\beta+m)\Gamma(1-\beta+m)}{m!(m+1)!\Gamma(1-\beta)^2} \\ &\quad \cdot p \cdot {}_2F_1(-m, 2-\beta+m; 1-\beta; 1-p) \\ &\quad \cdot (1-x)^{-\beta} {}_2F_1(-m, 2-\beta+m; 1-\beta; 1-x) e^{-\lambda_m \tau/2} \\ &= \sum_{m=0}^{\infty} (2m+2-\beta)(m+1-\beta)(m+1) \\ &\quad \cdot p \frac{(1-\beta)_m}{(m+1)!} {}_2F_1(-m, 2-\beta+m; 1-\beta; 1-p) \\ &\quad \cdot (1-x)^{-\beta} \frac{(1-\beta)_m}{(m+1)!} {}_2F_1(-m, 2-\beta+m; 1-\beta; 1-x) e^{-\lambda_m \tau/2}\end{aligned} \tag{58}$$

for the interior $(0 < x < 1)$ and

$$\begin{aligned}\Pi_0(\tau) &= 1 - p \sum_{m=0}^{\infty} (-1)^m \frac{\Gamma(m+1-\beta)(2m+2-\beta)}{\Gamma(1-\beta)(m+1)!} \\ &\quad \cdot {}_2F_1(-m, 2+m-\beta; 1-\beta; 1-p) e^{-\lambda_m \tau/2} \\ &= 1 - p \sum_{m=0}^{\infty} (-1)^m \frac{(1-\beta)_m (2m+2-\beta)}{(m+1)!} \\ &\quad \cdot {}_2F_1(-m, 2+m-\beta; 1-\beta; 1-p) e^{-\lambda_m \tau/2}\end{aligned} \tag{59}$$

at the fixation boundary ($x = 0$). In the second steps of formulas (58) and (59) we used property (55) of the gamma function $\Gamma(x)$.

It follows from equation 15.3.6 of [1] that

$$
\begin{aligned}
&{}_2F_1(-m, 2-\beta+m; 1-\beta; 1-x) \\
&\quad = \frac{(-1)^m (m+1)!}{(1-\beta)_m}\, {}_2F_1(-m, 2-\beta+m; 2; x).
\end{aligned} \tag{60}
$$

Inserting (60) into (58) and (59), we find that

$$
\begin{aligned}
v(x,\tau) = &\sum_{m=0}^{\infty} (2m+2-\beta)(m+1-\beta)(m+1) \\
&\cdot p \cdot {}_2F_1(-m, 2-\beta+m; 2; p) \\
&\cdot (1-x)^{-\beta}\, {}_2F_1(-m, 2-\beta+m; 2; x) e^{-\lambda_m \tau/2}
\end{aligned} \tag{61}
$$

and

$$
\begin{aligned}
\Pi_0(\tau) = 1 - p &\sum_{m=0}^{\infty} (2m+2-\beta) \\
&\cdot {}_2F_1(-m, 2+m-\beta; 2; p) e^{-\lambda_m \tau/2}.
\end{aligned} \tag{62}
$$

We finally apply the second step of (56) to each term of (61) and (62), in order to find that (61) agrees with (18) in the interior ($0 < x < 1$), and that (62) coincides with (22).

Motivation of (24) and (25). Inserting (56) into (18), we find that

$$
\begin{aligned}
v(x,\tau) = &\ \Pi_0(\tau)\delta(x) \\
&+ (1-x)^{-\beta} \sum_{m=0}^{\infty} \frac{A_m}{m+1} P_m^{(1,-\beta)}(1-2x) e^{-\lambda_m \tau/2},
\end{aligned}
$$

where

$$
A_m = p(1-p)^{\beta}(2m+2-\beta)(m+1-\beta)(m+1)\phi_m(p). \tag{63}
$$

This agrees with (24), if

$$
\begin{aligned}
C(x,\tau) &= \frac{1}{1-\beta} \sum_{m=0}^{\infty} \frac{A_m}{m+1} P_m^{(1,-\beta)}(1-2x) e^{-\lambda_m \tau/2} \\
&= \frac{p(1-p)^{\beta}}{1-\beta} \sum_{m=0}^{\infty} (2m+2-\beta)(m+1-\beta)\phi_m(p) \\
&\qquad \cdot P_m^{(1,-\beta)}(1-2x) e^{-\lambda_m \tau/2} \\
&= \frac{p}{1-\beta} \sum_{m=0}^{\infty} \frac{(2m+2-\beta)(m+1-\beta)}{m+1} \\
&\qquad \cdot P_m^{(1,-\beta)}(1-2p) P_m^{(1,-\beta)}(1-2x) e^{-\lambda_m \tau/2}.
\end{aligned} \tag{64}
$$

In the second and third steps we made use of (63) and (56). The $m = 0$ term of the expansion for $C(x, \tau)$ equals the leading term of (25), because of $\lambda_0 = 1 - \beta$ and (57).

Derivation of (29). Since the first and third equations of (29) follow easily from (18) and (26)–(28), we concentrate on the middle one. In order to verify this second equation of (29), we need some properties of hypergeometric functions. The first one

$$
{}_2F_1(1, 2; 2; p) = \frac{1}{1-p}, \tag{65}
$$

follows directly from (54). The second formula

$$
{}_2F_1(1-m, m+2; 2; 1-p) = -(-1)^m\, {}_2F_1(1-m, m+2; 2; p) \tag{66}
$$

can be deduced from 15.3.6 of [1]. In conjunction with (22), we find that

$$
\begin{aligned}
\Pi_0(\tau; p, 1-) &= 1 - p(1-p) \sum_{m=0}^{\infty} (2m+1) e^{-m(m+1)\tau/2} \\
&\qquad \cdot {}_2F_1(1-m, m+2; 2; p) \\
&\overset{(65)}{=} 1 - p - \sum_{m=1}^{\infty} (2m+1) e^{-m(m+1)\tau/2} \\
&\qquad \cdot {}_2F_1(1-m, m+2; 2; p) \\
&\overset{(66)}{=} 1 - p + \sum_{m=1}^{\infty} (2m+1)(-1)^m e^{-m(m+1)\tau/2} \\
&\qquad \cdot {}_2F_1(1-m, m+2; 2; 1-p) \\
&= \Pi_0(\tau; p, 1),
\end{aligned}
$$

as was to be proved.

Proof of (35). We have that

$$
\begin{aligned}
E(T_n) &= \frac{1}{2n} \sum_{t=0}^{\infty} P\Big(T_n > \frac{t}{2n}\Big) \\
&= \frac{1}{2n} \sum_{t=0}^{\infty} \Big[1 - \Pi_{n0}(\frac{t}{2n})\Big] \\
&\to \int_0^{\infty} [1 - \Pi_0(\tau)] d\tau \\
&= E(T),
\end{aligned}
$$

where in the third step we used (34) to deduce that the integrand converges pointwise. The fact that the integral converges as well follows from dominated convergence. Indeed, let $\tilde{V}^{(0)}$ be the row vector of length $2n$ that excludes the first component of $V^{(0)}$ in (5), and $\tilde{M}$ the square matrix of order $2n$ that excludes the first column and the first row of the transition matrix M in (2). The Jordan decomposition (40) of M implies that

$$
\begin{aligned}
1 - \Pi_{n0}\Big(\frac{t}{2n}\Big) &= \tilde{V}^{(0)} \tilde{M}^t (1, \dots, 1)' \\
&\le K d_1^t \\
&= K d_1^{2n\tau} \\
&= K \exp(-\lambda_{n0}\tau/2) \\
&\le 2K \exp(-\lambda_0 \tau/2),
\end{aligned} \tag{67}
$$

for some constant K not depending on n, where in the last step of (67) we used (43). Clearly, the right hand side of (67) is an integrable function of τ.

Proof of (42). In order to prove (42) we insert the Jordan decomposition (40) into (5) and (9), and use the relation between x and i in (7). In this way, we can rewrite the renormalized absolute probability vector as

$$
\begin{aligned}
v_n(x,\tau) &= 2n\Big(V^{(0)} M^{2n\tau}\Big)_i \\
&= 2n\Big(V^{(0)} Q^{-1} D^{2n\tau} Q\Big)_{2nx} \\
&= 2n \sum_{k=0}^{2n} (V^{(0)} Q^{-1})_k \cdot d_k^{2n\tau} l_{k,2nx} \\
&\overset{x>0}{=} \sum_{k=1}^{2n} r_{k,2np} \cdot 2n l_{k,2nx} \cdot d_k^{2n\tau} \\
&= \sum_{m=0}^{2n-1} h_{nm}(x; p) \exp(-\lambda_{nm}\tau/2),
\end{aligned}
$$

where in the second last step we used (6), (11), and that $l_{1,2nx} = 0$ when $x > 0$. In the last step we changed summation index from k to $m = k - 1$.

Proof of (43). It follows from (19) and (41) that

$$\begin{aligned} d_{m+1} &= 1 - \frac{(1-\beta)(m+1)}{4n} - \frac{m(m+1)}{4n} + \frac{c_m}{(2n)^2} + o(n^{-2}) \\ &= 1 - \frac{\lambda_m}{4n} + \frac{c_m}{(2n)^2} + o(n^{-2}) \qquad (68) \end{aligned}$$

as $n \to \infty$, for any fixed $m \geq 0$ and some constant $c_m = c_m(\beta)$. Formula (68) and the definition of λ_{nm} below (42), imply

$$\begin{aligned} \exp(-\lambda_{nm}\tau/2) &= d_{m+1}^{2n\tau} \\ &= \left[1 - \frac{\lambda_m}{4n} + \frac{c_m}{(2n)^2} + o(n^{-2})\right]^{2n\tau} \\ &= \exp(-\lambda_m\tau/2)\left[1 + \frac{b_m}{2n} + o(n^{-1})\right], \end{aligned}$$

with $b_m = c_m\tau - \lambda_m^2\tau/8$.

Motivation of (47). In view of (24) and the continuity of $x \to C(x, \tau)$ on [0, 1], we find that the probability mass of the diffusion density v that is removed from $[1 - 1/(4n), 1)$ is

$$\begin{aligned} \Pi_{n1}^*(\tau) &= \int_{1-1/(4n)}^{1-} v(x, \tau)dx \\ &= C(1, \tau)\int_{1-1/(4n)}^{1} (1-\beta)(1-x)^{-\beta}dx \cdot (1 + o(1)) \\ &= C(1, \tau) \cdot \frac{1}{(4n)^{1-\beta}} \cdot (1 + o(1)), \end{aligned}$$

as $n \to \infty$.

Motivation of (50). We will first derive an explicit expression for $C(1, \tau)$, and a useful approximation of this quantity for small mutation rates. In view of (64), we need the following values

$$\begin{aligned} P_m^{(1,-\beta)}(-1) &= (-1)^m \binom{m-\beta}{m} \\ &= (-1)^m \frac{\Gamma(m+1-\beta)}{\Gamma(m+1)\Gamma(1-\beta)} \\ &\approx (-1)^m \frac{1-\beta}{m} \qquad (69) \end{aligned}$$

of the Jacobi polynomials at the left boundary -1 for all $m \geq 1$, where the last approximative step requires that $2\mu^* = 1 - \beta$ is small. Insertion of (56), (57) and (69) into (64), we get

$$\begin{aligned} C(1, \tau) &= \frac{p(1-p)^\beta}{1-\beta} \sum_{m=0}^{\infty} (2m + 2 - \beta)(m + 1 - \beta)P_m^{(1,-\beta)}(-1) \\ &\qquad \cdot {}_2F_1(\beta - m, m + 2; 2; p)e^{-\lambda_m\tau/2} \\ &\approx p(2-\beta)e^{-\mu^*\tau} \\ &\quad + p(1-p)^\beta \sum_{m=1}^{\infty} \frac{(2m + 2 - \beta)(m + 1 - \beta)}{m}(-1)^m \\ &\qquad \cdot {}_2F_1(\beta - m, m + 2; 2; p)e^{-\lambda_m\tau/2}. \qquad (70) \end{aligned}$$

In order to prove the first step of (50), we initially keep n fixed. It then follows from (27), (29), (30) and (70) that

$$\lim_{\beta \to 1-} \Pi_{n1}^*(\tau) = \int_{1-1/(4n)}^{1-} v(x, \tau; p, 1)dx + \Pi_1(\tau; p, 1)$$

and

$$\lim_{\beta \to 1-} \frac{C(1, \tau)}{(4n)^{1-\beta}} = \lim_{\beta \to 1-} C(1, \tau) = \Pi_1(\tau; p, 1).$$

By choosing n large enough, the above two limits can be made arbitrarily close.

In order to motivate the third step of (50), we need to show that T_{n2} is asymptotically exponentially distributed with rate μ^* for small mutation rates. We will first look at $2nT_{n2}$, which is the number of generations t after the population gets homogeneous for the A allele, that a successful mutation occurs. We argue that $2nT_{n2}$ has a geometric distribution with a success probability close to

$$\left[1 - (1-\mu)^{2n}\right] \cdot \frac{1}{2n} \approx \mu,$$

where the first factor is the probability that at least one mutation $A \to a$ occurs per generation in a homogeneous population of A alleles, and the second factor approximates the probability that one a allele subsequently spreads to the whole population. Since μ^* is small, this follows from (28), with $p = 1 - 1/(2n)$. The claimed exponential distribution follows, since

$$\begin{aligned} P(T_{n2} > \tau) &= P(2nT_{n2} > 2n\tau) \approx (1-\mu)^{2n\tau} \\ &= \left(1 - \frac{\mu^*}{2n}\right)^{2n\tau} \to e^{-\mu^*\tau} \end{aligned}$$

as $n \to \infty$, for a fixed positive value of the rescaled time parameter $\tau = t/(2n)$.

We finish by showing that the right hand side of (50) approximates $C(1, \tau)$ well for small mutation rates. To this end, we first differentiate (27) in order to find the first factor

$$\begin{aligned} \Pi_1'(\tau; p, 1) &= p(1-p)\sum_{m=1}^{\infty} \frac{(2m+1)m(m+1)}{2}(-1)^{m-1} \\ &\qquad \cdot {}_2F_1(1 - m, m + 2; 2; p)e^{-m(m+1)\tau/2} \end{aligned}$$

of the integrand of the right hand side of (50). Let I denote the value of this integral. Inserting the above series expansion into the integrand, we obtain

$$\begin{aligned} I &=: \int_0^\tau \Pi_1'(s; p, 1)e^{-\mu^*(\tau - s)}ds \\ &= p(1-p)e^{-\mu^*\tau}\sum_{m=1}^{\infty} \frac{(2m+1)m(m+1)}{\lambda_m^*}(-1)^{m-1} \\ &\qquad \cdot {}_2F_1(1 - m, m + 2; 2; p)(1 - e^{-\lambda_m^*\tau/2}), \qquad (71) \end{aligned}$$

after some straightforward but tedious calculations, where the order of summation and integration is exchanged, and

$$\lambda_m^* = m(m+1) - 2\mu^* \overset{\mu^* \text{ small}}{\approx} m(m+1).$$

Inserting this approximation of λ_m^* into (71), we get

$$\begin{aligned} I &\approx p(1-p)e^{-\mu^*\tau}\sum_{m=1}^{\infty}(2m+1)(-1)^m \\ &\qquad \cdot {}_2F_1(1 - m, m + 2; 2; p)(1 - e^{-\lambda_m^*\tau/2}) \\ &= pe^{-\mu^*\tau} \\ &\quad + p(1-p)\sum_{m=1}^{\infty}(2m+1)(-1)^m \\ &\qquad \cdot {}_2F_1(1 - m, m + 2; 2; p)e^{-m(m+1)\tau/2}, \end{aligned}$$

where in the last step we used (27) in order to recognize the coefficient of $e^{-\mu^*\tau}$ as $[p - \Pi_1(0; p, 1)] = p$. But in view of (70), the last approximation of I is close to $C(1, \tau)$ when β is close to 1, as was to be proved.

References

[1] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions, tenth printing, United States Department of Commerce, National Bureau of Standards, Applied Mathematics Series, 55, Washington D.C., 1972.

[2] G.J. Baxter, R.A. Blythe, A.J. McKane, Exact solution of the multi-allelic diffusion model, Math. Biosc. 209 (2007) 124–170.

[3] C.D. Campbell, E.E. Eichler, Properties and rates of germline mutations in humans, Trends Genet. 29 (10) (2013) 575–584.

[4] C. Cannings, The latent roots of certain Markov chains arising in genetics: A new approach, I. Haploid models, Adv. Appl. Prob. 6 (1974) 260–290.

[5] P. Collet, S. Martinez, J. San Martin, Quasi Stationary Distributions, Markov Chains, Diffusions and Dynamical Systems, Springer, Berlin, 2013.

[6] D.R. Cox, H.D. Miller, The Theory of Stochastic Processes, Methuen & Co Ltd, London, 1965.

[7] J. Crow, M. Kimura, Some genetic problems in natural populations, in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 4, University of California Press, 1956, pp. 1–22.

[8] J. Crow, M. Kimura, An Introduction to Population Genetics Theory, The Blackburn Press, Caldwell, New Jersey, 1970.

[9] R. Durrett, Probability Models for DNA Sequence Evolution, second edition, Springer, New York, 2008.

[10] S.N. Ethier, R.C. Griffiths, The transition function of a Fleming–Viot process, Ann. Probab. 21 (1993) 1571–1590.

[11] S.N. Ethier, T.G. Kurz, Markov Processes, Characterization and Convergence, John Wiley & Sons, Hoboken, New Jersey, 1986.

[12] S.N. Ethier, M.F. Norman, Error estimate for the diffusion approximation of the Wright–Fisher model, Proc. Natl. Acad. Sci. USA 74 (1977) 5096–5098.

[13] W.J. Ewens, Numerical results and diffusion approximation in a genetic process, Biometrika 50 (1963) 241–249.

[14] W.J. Ewens, The adequacy of the diffusion approximation to certain distributions in genetics, Biometrics 21 (2) (1965) 386–394.

[15] W.J. Ewens, Mathematical Population Genetics 1: Theoretical Introduction, second edition, Springer, New York, 2004.

[16] W. Feller, Diffusion processes in genetics, in: J. Neyman (Ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1951, pp. 227–246.

[17] R.A. Fisher, On the dominance ratio, Proc. Roy. Soc. Edinburgh 42 (1922) 321–431.

[18] G. Greenbaum, Revisiting the time until fixation of a neutral mutant in a finite population – a coalescent theory approach, J. Theor. Biol. 380 (2015) 98–102.

[19] R. Griffiths, A transition density expansion for a multi-allele diffusion model, Adv. Appl. Prob. 11 (1979) 310–325.

[20] R. Griffiths, Line of descent in the diffusion approximation of neutral Wright–Fisher models, Theor. Pop. Biol. 17 (1980) 37–50.

[21] R. Griffiths, Allele frequencies in multidimensional Wright–Fisher models with a general symmetric mutation structure, Theor. Pop. Biol. 17 (1980) 51–70.

[22] R. Griffiths, D. Spanó, Diffusion processes and coalescence trees, in: N.H. Bingman, C.M. Goldie (Eds.), Probability and Mathematical Genetics, Papers in Honour of Sir John Kingman, LMS Lecture Note Series 378, Cambridge University Press, 2010, pp. 358–375.

[23] S. Karlin, A First Course in Stochastic Processes, Academic Press, New York, 1966.

[24] S. Karlin, J. McGregor, On a genetics model of Moran, Proc. Camb. Phil. Soc. 58 (1962) 299–311.

[25] M. Kellis, B. Wold, M.P. Snyder, B.E. Bertstein, A. Kundaje, G.K. Marinov, Defining functional DNA elements in the human genome, Proc. Natl. Acad. Sci. 111 (17) (2014) 6131–6138.

[26] M. Kimura, Stochastic processes and distribution of gene frequencies under natural selection, Cold Spring Harbour Symp. Quant. Biol. 20 (1955) 33–53.

[27] M. Kimura, Solution of a process of random genetic drift with a continuous model, Proc. Natl. Acad. Sci. USA 41 (1955) 141–150.

[28] M. Kimura, Diffusion models in population genetics, J. Appl. Prob. 1 (1964) 177–232.

[29] M. Kimura, Theoretical foundations of population genetics at the molecular level, Theor. Pop. Biol. 2 (1971) 174–208.

[30] M. Kimura, Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical and pseudo-sampling methods, Proc. Natl. Acad. Sci. USA 77 (1) (1980) 522–526.

[31] A.J. McKane, D. Waxman, Singular solutions of the diffusion equation of population genetics, J. Theor. Biol. 247 (2007) 849–858.

[32] P.A.P. Moran, Random processes in genetics, Proc. Camb. Phil. Soc. 54 (1958) 60–71.

[33] F. Papangelou, Tracing the path of a Wright–Fisher process with one-way mutation in the case of large deviation, in: I. Karatzas, B.S. Rajput, M.S. Taqqu (Eds.), Stochastic Processes and Related Topics, In Memory of Stambanis Cambanis, Springer Science+Business Media New York, 1978, pp. 315–330.

[34] N. Shimakura, Equations différentielles provenant de la génétique des populations, Tohoku Math. J. 29 (1977) 287–318.

[35] N. Shimakura, Formulas for diffusion approximations of some gene frequency models, J. Math. Kyoto Univ. 21-1 (1981) 19–45.

[36] S. Tavaré, Line-of-descent and genealogical processes, and their application in population genetics models, Theor. Popul. Biol. 26 (1984) 119–164.

[37] A. Tenesa, P. Navarro, B.J. Hayes, D.L. Duffy, G.M. Clarke, M.E. Goddard, P.M. Visscher, Recent human effective population size estimated from linkage disequilibrium, Genome Res. 17 (4) (2007) 520–526.

[38] P.A. Tyvand, An exact algebraic theory of genetic drift in finite diploid populations with random mating, J. Theor. Biol. 163 (1993) 315–331.

[39] P.A. Tyvand, S. Thorvaldsen, Exact Markov chains versus diffusion theory for haploid random mating, Math. Biosci. 225 (2010) 18–23.

[40] D. Waxman, Comparison and content of the Wright-Fisher model of random genetic drift, the diffusion approximation, and an intermediate genetic model, J. Theor. Biol. 269 (2011) 79–87.

[41] S. Wright, Evolution in Mendelian populations, Genetics 16 (1931) 97–159.

[42] S. Wright, The differential equation of the distribution of gene frequencies, Proc. Natl. Acad. Sci. 31 (1945) 382–389.

[43] L. Zhao, X. Yue, D. Waxman, Complete numerical solution of the diffusion equation of random genetic drift, Genetics 194 (2013) 973–985.