Sussex Research

Singular solutions of the diffusion equation of population genetics

A. J. McKane, David Waxman

Publication date
01-08-2007

Licence
This work is made available under the Copyright not evaluated licence and should only be used in accordance with that licence. For more information on the specific terms, consult the repository record for this item.

Citation for this work (American Psychological Association 7th edition)
McKane, A. J., & Waxman, D. (2007). *Singular solutions of the diffusion equation of population genetics* (Version 1). University of Sussex. https://hdl.handle.net/10779/uos.23311814.v1

Published in
Journal of Theoretical Biology

Link to external publisher version
https://doi.org/10.1016/j.jtbi.2007.04.016

Copyright and reuse:
This work was downloaded from Sussex Research Open (SRO). This document is made available in line with publisher policy and may differ from the published version. Please cite the published version where possible. Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners unless otherwise stated. For more information on this work, SRO or to report an issue, you can contact the repository administrators at sro@sussex.ac.uk. Discover more of the University's research at https://sussex.figshare.com/

Accepted for publication JTB

Singular solutions of the diffusion equation of population genetics

A. J. McKane[1] and D. Waxman[2]

[1]Theoretical Physics Group, School of Physics and Astronomy, University of Manchester, Manchester M13 9PL, UK.
Email: alan.mckane@manchester.ac.uk
[2]Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Falmer, Brighton BN1 9QG, Sussex, UK.
Email: d.waxman@sussex.ac.uk

Abstract

The forward diffusion equation for gene frequency dynamics is solved subject to the condition that the total probability is conserved at all times. This can lead to solutions developing singular spikes (Dirac delta functions) at the gene frequencies 0 and 1. When such spikes appear in solutions they signal gene loss or gene fixation, with the "weight" associated with the spikes corresponding to the probability of loss or fixation. The forward diffusion equation is thus solved for all gene frequencies, namely the absorbing frequencies of 0 and 1 along with the continuous range of gene frequencies on the interval $(0, 1)$ that excludes the frequencies 0 and 1. Previously, the probabilities of the absorbing frequencies 0 and 1 were found by appeal to the backward diffusion equation, while those in the continuous range $(0, 1)$ were found from the forward diffusion equation. Our unified approach does not require two separate equations for a complete dynamical treatment of all gene frequencies within a diffusion approximation framework. For cases involving mutation, migration and selection, it is shown that a property of the deterministic part of gene frequency dynamics determines when fixation and loss can occur. It is also shown how solution of the forward equation, at long times, leads to the standard result for the fixation probability.

1 Introduction

In this work we focus on genetic drift — the process that occurs when there is random variation in the number of offspring contributed by each adult member of a finite population. At one locus in the population, the number of copies of a particular gene randomly varies from generation to generation, and undergoes a kind of random walk. The outcome is that the genetic composition of the population fluctuates over time.

Genetic drift is an evolutionary force that has the tendency to decrease the variation in a population and can influence the effectiveness of mutation and selection. One of the key mathematical approaches to dealing with genetic drift is the diffusion approximation. This was introduced into population genetics by Fisher (1922), Wright (1945), and substantially extended and developed by Kimura (1955a). Under this approximation, the proportion of individuals of a particular genetic type is treated as a continuous random variable whose distribution obeys a diffusion equation. This approach has been used to derive results that lie at the very heart of population genetics (Crow and Kimura, 1970).

Here, we aim to readdress issues that were apparently dealt with more than fifty years ago (Kimura, 1955b) and have become part of the textbook knowledge of the subject. Our aim is to provide a conceptually simple and consistent approach to solving the diffusion equation. This involves reexamining the mathematical conditions required of the solutions, as well as their nature and interpretation.

2 Basics

Consider a single genetic locus in a population of N diploid individuals. Let us focus on one allele, denoted A, at a given locus. The ratio of the total number of copies of allele A in the population, to the total number of all alleles at the locus ($2N$), is termed the gene frequency and this can only take the discrete values $0/(2N)$, $1/(2N)$, ..., $2N/(2N)$. The diffusion approximation approximates the gene frequency as a continuous variable, x, that lies in the range 0 to 1. It is a commonly held view that the *forward* diffusion approximation has doubtful validity when x lies within a distance $1/(2N)$ of the values $x = 0$ and $x = 1$ (see e.g., Chapter 10 of Gale, 1990). This is exemplified by the exact solution of the diffusion equation obtained by Kimura, for the case of a randomly mating population, where the only evolutionary force is genetic drift (Kimura, 1955b). The solution is taken to hold only on the *interior* of the possible range of x but not at the boundary values of x, i.e. not at $x = 0$ and $x = 1$. Thus while such a solution is informative about some quantities of interest, such as the level of heterozygosity in the population (Crow and Kimura, 1970), it suffers from a lack of completeness, in the sense that it does not *directly* say anything about the two gene frequencies of greatest interest. These are the frequencies corresponding to where either all copies of allele A are lost from the population (the frequency $x = 0$) or where all individuals carry two copies of allele A — corresponding to fixation (the frequency $x = 1$). More generally, solutions of the forward diffusion equation suffer from a related problem, namely, in the absence of mutations that take the population away from at least one of the frequencies $x = 0$ or $x = 1$ (or both), there is a loss of probability from the region where the forward diffusion approximation is taken to apply — i.e., all x excluding the boundary values $x = 0$ and $x = 1$. However, the same diffusion approximation, when applied in the presence of two-way mutation (i.e., mutations that go both from and to allele A) yields a distribution that applies for the full range of x, and preserves probability for all times.

The possible phenomena that can occur at the boundaries $x = 0$ and $x = 1$ have been been

previously investigated and classified (Feller, 1952), and from these, mathematical boundary conditions on solutions to the diffusion equation have been inferred (Feller, 1954; Voronka and Keller, 1975; Maruyama, 1977; Ewens, 1979; Gardiner, 2004). Here we take an alternative approach. Our fundamental guiding principle is that the probability of the gene frequency lying in the *full range* $0 \leq x \leq 1$ (i.e., all x *including* the boundary frequencies $x = 0$ and $x = 1$) should, at all times, be unity. We consistently take this viewpoint for all problems, irrespective of the pattern of mutation, selection and migration. For situations where there is no mutation, the only way for the total probability to be conserved is for probability to accumulate at the boundaries. As a consequence, the approach we adopt can lead to solutions to the forward diffusion problem that do not have the property of being smooth and well behaved. Rather, the approach can lead to solutions that possess singularities — sharp spikes (Dirac delta functions) that, when present, lie at one or other or both boundaries. The probability associated with these singularities, combined with the probability associated with the interior range of x, lead to a net probability of unity. As we show, it is completely natural to associate the probabilities associated with the spikes at the boundaries, when they exist in the solution, with the probabilities of gene loss ($x = 0$) and gene fixation ($x = 1$). Given the correctness of this association, the approach we are proposing yields a consistent and unified description of all gene frequencies, i.e., the absorbing frequencies of 0 and 1 along with the continuous range of gene frequencies on the interval $(0, 1)$ that excludes the frequencies 0 and 1. This is in contrast to all previous approaches, where the probabilities of the absorbing frequencies 0 and 1 were found by appeal to the backward diffusion equation, while those in the continuous range $(0, 1)$ were found from solving the forward diffusion equation.

3 Conservation of probability

Let $f(x, t)$ denote the probability density of the gene frequency at time t. The interpretation of $f(x, t)$ is that in a very large number of replicates of a population, that all have the same initial distribution, the fraction of such replicates where the gene frequency lies in the range a to b, at time t, is $\int_a^b f(x, t)dx$.

Generally, we can write the forward diffusion equation as

$$\frac{\partial f(x,t)}{\partial t} + \frac{\partial j(x,t)}{\partial x} = 0 \tag{1}$$

where the quantity $j(x, t)$ is the probability current density — a quantity that characterises the flow of probability density. The form that $j(x, t)$ takes for a diploid population of N randomly mating individuals is

$$j(x,t) = M(x)f(x,t) - \frac{1}{4N}\frac{\partial}{\partial x}\left[x(1-x)f(x,t)\right] \tag{2}$$

where $M(x)$ represents the deterministic part of gene frequency dynamics and is typically taken as a polynomial in x whose coefficients depend on mutation rates, migration rates and selection coefficients; Crow and Kimura (1970) use the notation $M_{\delta x}$ for this quantity.

The principle that probability is conserved means that for all times, the total probability does not change, thus

$$\frac{d}{dt}\int_0^1 f(x,t)dx = 0. \tag{3}$$

Integrating Eq. (1) over all x, and using conservation of probability, Eq. (3), yields $j(1,t)-j(0,t)=0$. Given the absence of any dynamical mechanism that connects the probability current densities at $x=0$ and $x=1$, we take the boundary conditions to be the zero current conditions:

$$\begin{aligned} j(0,t) &= 0 \\ j(1,t) &= 0. \end{aligned} \tag{4}$$

These corresponds to there being zero probability current density precisely at the boundaries, $x=0$ and $x=1$, and so no probability can flow *outside* the region $x=0$ to $x=1$ and hence be lost. Such boundary conditions were only adopted for problems with two way mutation by Crow and Kimura (1956).

We shall solve the diffusion equation, Eq. (1), subject to the conditions of Eq. (4). Once such conditions are imposed, the distribution $f(x,t)$ remains normalised for all times, in the sense that if we start at time $t=0$ with a probability distribution obeying $\int_0^1 f(x,0)dx=1$, then it automatically follows that $\int_0^1 f(x,t)dx=1$ holds for all times. There is thus no loss of probability in this approach.

4 Pure drift

We first analyse the apparently simplest case, where the only evolutionary force acting on a randomly mating diploid population is genetic drift. For this case, the probability current density is given by Eq. (2) with $M(x)=0$:

$$j(x,t) = -\frac{1}{4N}\frac{\partial}{\partial x}[x(1-x)f(x,t)]. \tag{5}$$

and the forward diffusion equation reads

$$\frac{\partial f(x,t)}{\partial t} = \frac{1}{4N}\frac{\partial^2}{\partial x^2}[x(1-x)f(x,t)] \tag{6}$$

(Crow and Kimura, 1970). We solve this equation, subject to the condition that all replicate populations initially have the gene frequency of p, so $f(x,0)$ corresponds to an initial distribution where only the single frequency p is present.

Solving Eq. (6), subject to Eq. (4) with the form for $j(x,t)$ given by Eq. (5) then leads, inescapably, to the solution containing spikes (Dirac delta functions) at the boundaries, after some time. The simplest way to see this is to look at a stationary solution of Eq. (6), i.e., a solution of the form $f(x,t)=f(x)$. For such a solution, we integrate Eq. (5) from $x=0$ to an arbitrary x. Invoking Eq. (4) leads to $x(1-x)f(x)=A$ (a constant).

For the set of well-behaved (i.e., non-singular functions), we note that if $x(1-x)f(x)=A$, then the solution for the distribution $f(x)$ is the obvious one: $f(x)=A/[x(1-x)]$. However, in the theory of probability, it is allowable for distributions to contain functions that diverge (i.e., are singular) as long as they are non-negative and integrable. The singular function that is of relevance here is the Dirac delta function $\delta(x-\alpha)$. This is a zero-width, unit area spike, that is located at $x=\alpha$ and has infinite height (and hence is singular). Such functions naturally occur. For example, if, on repeated measurement of a continuous random variable, the single value p is always obtained,

then the probability density describing this is simply $\delta(x-p)$, with all of the "mass" or "weight" of the distribution located solely at $x=p$.

Returning to the equation for the stationary solution for the distribution $f(x)$, namely $x(1-x)f(x)=A$, we proceed to solve it by dividing through by $x(1-x)$. This yields $f(x)=A/[x(1-x)]+B\delta(x)+C\delta(1-x)$ where B and C are constants that multiply Dirac delta functions located at $x=0$ and $x=1$. The Dirac delta functions, with undetermined constants multiplying them, are present since $x\delta(x)$ and $(1-x)\delta(1-x)$ are identically zero (Dirac 1958) and so must, in all generality, be included in the solution for the probability density $f(x)$. The condition that $f(x)=A/[x(1-x)]+B\delta(x)+C\delta(1-x)$ is normalisable (has a finite integral) requires $A=0$ (since $\int_0^1 1/[x(1-x)]dx=\infty$) and hence a stationary solution for $f(x)$ consists *solely* of singular solutions, namely the Dirac delta functions at $x=0$ and $x=1$. Imposing the condition of normalisation, $\int_0^1 f(x)dx=1$, on this solution yields $B+C=1$. If, furthermore, we impose the condition that the mean gene frequency, at any time, coincides with its initial value, p, since drift has no systematic direction to it (as theory can verify; see Crow and Kimura, 1970), then we arrive at $B=1-p$ and $C=p$ and the stationary solution for the distribution is $f(x)=(1-p)\delta(x)+p\delta(1-x)$. The coefficients of the delta functions are precisely the probabilities of loss or fixation of the allele A. The presence of delta functions in the solution of the forward diffusion equation is essential, in this case, if total probability is to add to unity. This example shows it is also entirely natural to associate the coefficients of $\delta(x)$ and $\delta(1-x)$ with the probability that allele A is lost or fixed.

We note that a direct numerical approach to solving the diffusion equation, Eq. (6), will inevitably run into problems, when Eq. (4) is imposed, since no standard numerical procedure can handle singularities of the delta function type, that arise in the solution.

5 Solution of the pure drift equation

Given the above arguments, the diffusion equation, Eq. (6), has solutions that

(i) consist, on the interior of the range of x, i.e., for $0<x<1$, of a function of x that is integrable

(ii) generally contains singularities (Dirac delta functions) at the boundaries of the range of x, namely $x=0$ and $x=1$

(iii) at time $t=0$, has the form $\delta(x-p)$, corresponding to an initial distribution with a single gene frequency of p being present.

The solution thus has the form (see Appendix A for mathematical details)

$$f(x,t)=\Pi_0(t)\delta(x)+\Pi_1(t)\delta(1-x)+f_K(x,t). \tag{7}$$

The quantities $\Pi_0(t)$ and $\Pi_1(t)$ are the probabilities that the gene frequency has achieved the values 0 and 1, respectively, by time t. They vanish at time $t=0$; they also depend on p and N, however we do not explicitly exhibit this dependence. The function $f_K(x,t)$ has the property $f_K(x,0)=\delta(x-p)$ and hence incorporates the condition that the only gene frequency that is initially present is p.

In Appendix A we determine the exact solution of Eq. (6) and show that the function $f_K(x,t)$ can be directly identified with Kimura's solution of the problem of pure drift (Kimura, 1955b) and for completeness, this function is reproduced in Eq. (A7) of Appendix A. The function $f_K(x,t)$ corresponds to the solution of the diffusion equation Eq. (6) that is normalisable and does not posses any delta function singularities at $x=0$ and $x=1$.

The functions $\Pi_0(t)$ and $\Pi_1(t)$ are shown in Appendix A to be given by

$$\Pi_0(t) = \frac{1}{4N} \int_0^t f_K(0,s)ds$$
$$\Pi_1(t) = \frac{1}{4N} \int_0^t f_K(1,s)ds. \tag{8}$$

We note that usually $(4N)^{-1} \int_0^t f_K(1,s)ds$ is identified with the fixation probability from considerations of the flow of probability density into $x = 1$ (Crow and Kimura, 1970). Here, such an identification is an automatic result of a calculational scheme where probability conservation is enforced, and in this scheme $\Pi_0(t)$ and $\Pi_1(t)$ are the coefficients of Dirac delta functions at the boundaries, in the full solution of the problem.

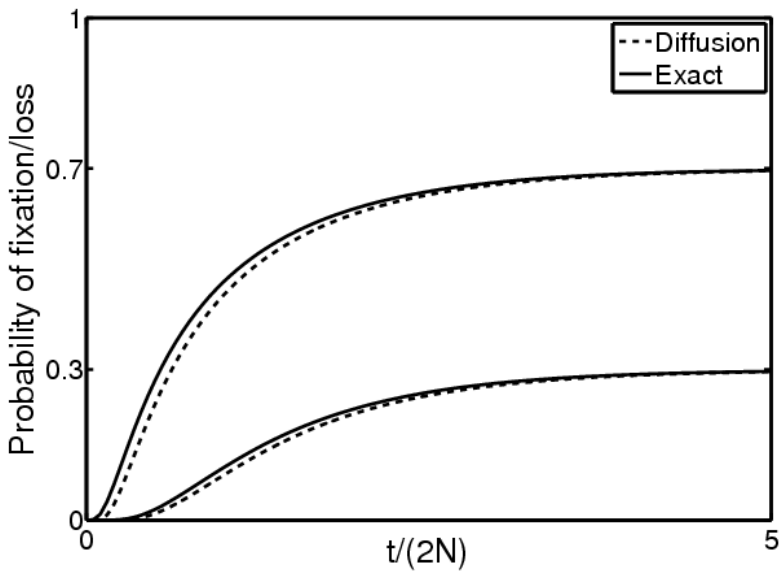

Figure 1

The functions $\Pi_0(t)$and $\Pi_1(t)$, are plotted against time. These functions are identified as the probability of loss and fixation, by time t, as follows from the diffusion analysis presented in this work. For the Figure we used an initial gene frequency of $p = 0.7$ and a population size of $N = 10$. The exact probabilities of the gene frequency taking the value 0 and 1, as follows from an exact Markov chain treatment of a Wright Fisher model (Fisher 1930; Wright 1931) are also given in the Figure. There is remarkably good agreement between the diffusion approximation for the probability of the gene frequency lying at the boundaries and the exact results for gene loss and gene fixation.

In Figure 1, the probabilities $\Pi_0(t)$ and $\Pi_1(t)$ are plotted against time. In the same Figure, plots are given of the exact probabilities of the gene frequency taking the value 0 and 1, as follows from an exact Markov chain treatment of a Wright Fisher model (Fisher, 1930; Wright, 1931). For even very small population sizes, such as the value $N = 10$, that was used in the Figure, there are very small differences between the diffusion results for the weights of the delta functions, $\Pi_0(t)$ and $\Pi_1(t)$, and the exact results for the probabilities of loss and fixation. For larger population sizes there is an even smaller discrepancy between exact results and those from diffusion analysis, with very close agreement for $N = 100$.

6 General case

For a randomly mating population, that is subject to mutation, selection and migration, the function $M(x)$ (that occurs in the equation for the probability current density, Eq. (2)) is generally non-zero and the forward diffusion equation for this case takes the form

$$\frac{\partial f(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[M(x)f(x,t)\right] + \frac{1}{4N}\frac{\partial^2}{\partial x^2}[x(1-x)f(x,t)]. \tag{9}$$

We look for a solution of this equation, subject to the boundary conditions of Eq. (4), with only the single gene frequency of p present at time $t = 0$. The form of the solution is taken to be that given in Eq. (7) with the functions $\Pi_0(t)$, $\Pi_1(t)$ and $f_K(x,t)$ to be determined. In Appendix B we show that for Eq. (7) to be a solution requires the following. (1) That $f_K(x,t)$ obeys Eq. (9). (2) That $f_K(x,t)$ corresponds to the single gene frequency p being initially present (i.e., $f_K(x,0) = \delta(x-p)$). (3) That $f_K(x,t)$ is subject to the conditions implicitly adopted by earlier workers, namely, that the function is normalisable and does not contain any delta function singularities at the boundaries. (4) For all t we have

$$\begin{aligned} M(0)\Pi_0(t) &= 0 \\ M(1)\Pi_1(t) &= 0. \end{aligned} \tag{10}$$

The two conditions in Eq. (10) yield four separate cases that govern the presence of Dirac delta functions in the solution for $f(x,t)$. With $j_K(x,t) = M(x)f_K(x,t) - (4N)^{-1}\,\partial\left[x(1-x)f_K(x,t)\right]/\partial x$ we have

(i) $M(0) \neq 0$, $M(1) \neq 0$, leading to $\Pi_0(t) = 0 = j_K(0,t)$ and $\Pi_1(t) = 0 = j_K(1,t)$.

(ii) $M(0) = 0$, $M(1) \neq 0$, leading to $\Pi_1(t) = 0 = j_K(1,t)$ and $\Pi_0(t)$ obeying $d\Pi_0(t)/dt = -j_K(0,t)$.

(iii) $M(0) \neq 0$, $M(1) = 0$, leading to $\Pi_0(t) = 0 = j_K(0,t)$ and $\Pi_1(t)$ obeying $d\Pi_1(t)/dt = j_K(1,t)$.

(iv) $M(0) = 0$, $M(1) = 0$, leading to $\Pi_0(t)$ and $\Pi_1(t)$ obeying $d\Pi_0(t)/dt = -j_K(0,t)$ and $d\Pi_1(t)/dt = j_K(1,t)$.

Case (i) corresponds to the deterministic part of gene frequency dynamics (i.e., $M(x)$) being able to move gene frequencies away from the boundary values $x = 0$ and $x = 1$, so neither loss nor fixation occurs. The outcome is that the distribution $f(x,t)$ does not develop Dirac delta functions at the boundaries. In cases (ii) and (iii) the vanishing of $M(x)$ at one boundary, as a result of vanishing deterministic dynamics there, allows gene frequencies to reach the boundary and for

Dirac delta functions to become established, over time, at that boundary. Case (iv), which includes the pure drift problem analysed above, as a special case, corresponds to a vanishing of deterministic dynamics at both boundaries. The result is that gene frequencies can reach both boundaries and both gene loss and fixation occur over time, as signalled by two delta functions that develop in $f(x,t)$ at the boundaries.

7 Continuity of solutions

The conventional approach to solving the forward diffusion equation imposes different boundary conditions in different cases, depending on the nature of mutation (see e.g., Crow and Kimura, 1956). In the present work we have consistently imposed the same type of boundary conditions, Eq. (4), and hence do not have different cases. We have not, however, discussed how the approach presented here allows a solution to transcend what are, in the conventional approach, different cases. To consider this aspect, we have investigated a time dependent solution of the diffusion equation that is normalised for all times and which does not, for the pattern of mutation adopted, ever develop delta functions at boundary values of x. The issue is how such a solution behaves when the character of mutation is altered, such that e.g., the loss of allele A can occur with non-zero probability. To this end, consider the situation where the only evolutionary processes occurring are mutation and drift. Mutations are taken go in both directions, i.e., both from and to the allele A, with probabilities of u and v. In this case, the function $M(x)$ takes the form $M(x) = v(1-x) - ux$ (see e.g., Ewens, 1979). If, at time $t = 0$ only the gene frequency p occurs, the form of the solution which conserves probability at all times (because probability current density vanishes at both boundaries) is known (see Eqs. (8.5.8) and (8.5.9) of Crow and Kimura 1970). Given such a solution, it is possible to change the pattern of mutation, by allowing (at fixed time) first u and then v to tend to zero. In doing so, the solution develops into one that has delta function singularities at the boundaries. Exact calculations (not given here) directly show that such a solution coincides with the singular solution of the pure drift case, given above. Thus, for example, the original solution of Crow and Kimura, which has no singularity at $x = 0$, becomes a solution with a delta function, at $x = 0$, whose weight coincides precisely with the form for $\Pi_0(t)$ that was found in the the pure drift case.

We note that the analogue of the delta function that occurs, at e.g. $x = 0$, in the resulting solution, when the scaled mutation rate, $V = 4Nv$, is small but non-zero, is the function

$$\Delta(x) = Vx^{V-1}. \tag{11}$$

This is a normalised probability density over $0 \leq x \leq 1$ i.e., $\int_0^1 \Delta(x)dx = 1$. Its shape is very dependent on the value of V. If $V < 1$ then $\Delta(x)$ decreases as x is increased from $x = 0$. For $V \sim 1$ this is a relatively slow decrease, but when $V \ll 1$ the function has a rapid decrease — see Figure 2. Furthermore, for $V \ll 1$ the mean and variance of $\Delta(x)$ are both of order V. It is only as $V \to 0$ that the function $\Delta(x)$ formally approaches a Dirac delta function, $\delta(x)$ (see Eq. (1.2.15) in Barton 1989) and for small, but finite V, the function $\Delta(x)$ represents the distribution of replicate

populations where allele A is "nearly lost."

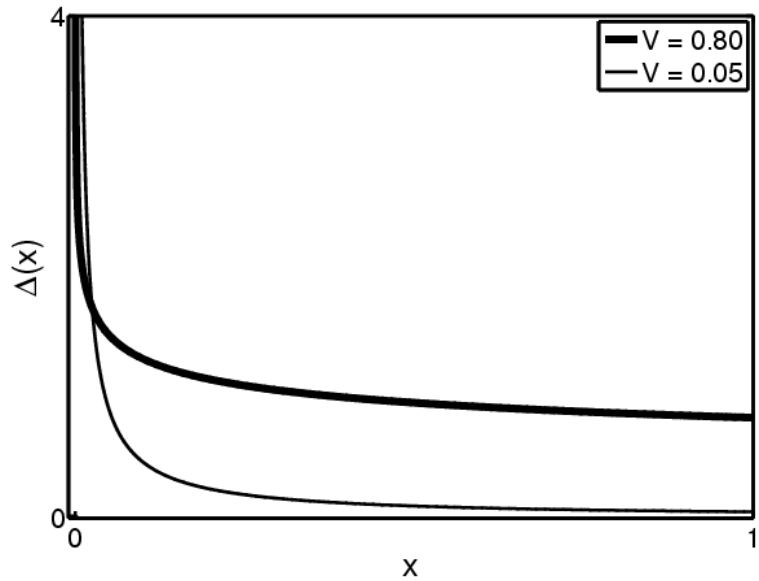

Figure 2

The function $\Delta(x)$, of Eq. (11), is the precursor of the delta function, $\delta(x)$, that becomes present in time dependent solutions of the diffusion equation when the scaled mutation rate V is taken to zero (see main text for details). Two examples of the function $\Delta(x)$ are plotted against x.

We infer from this example that there are not multiple types of boundary condition, depending on the nature of mutation, or more generally, depending on the form of the function associated with the deterministic aspect of frequency dynamics, $M(x)$. Rather, there is a single type of boundary condition, Eq. (4), and on modification of $M(x)$, time dependent solutions of the diffusion equation can be freely converted between solutions of apparently different types, i.e., converted between solutions that yield fixation and/or loss and those that do not exhibit this property. Thus the boundary condition of Eq. (4) covers all such cases.

We note that previously, the probabilities of the exact discrete terminal class frequencies ($x = 0$ and $x = 1$) have been associated with the probability, calculated from diffusion analysis, of the frequency falling into the ranges $0 < x < 1/(2N)$ and $1 - 1/(2N) < x < 1$ (Gale, 1990; pp. 281-284). We note however that in the absence of mutation, the exact solution of the diffusion equation leads to Dirac delta functions at $x = 0$ and $x = 1$. This corresponds to the range of the terminal classes being infinitesimal (the width of the delta functions) under the continuous frequency diffusion approximation. We also note that when mutation is finite, the function that becomes the delta function, at $x = 0$, is given in Eq. (11) with $V = 4Nv$. Since this function has a mean and variance of order V, it follows that when $\sqrt{V} \ll 1/(2N)$ we again find that not all of the interval $0 < x < 1/(2N)$, of the continuous x diffusion problem, contributes significantly to the probability of being in the terminal class; only a fraction $\sqrt{V}/(2N)$ contributes.

Generally, we note that under a continuous frequency diffusion approximation, there are no discrete frequency classes and we infer that detailed questions concerning particular discrete frequency classes may not be reliably answerable under such an approximation. In particular, precisely determining the range of x corresponding to a given discrete frequency class along with the associated probability, may not be unambiguously determined. Fortunately, many questions for which diffusion analysis is used are associated with averages of smooth functions of x, and these are well captured by the approximation.

8 Discussion

In this work we have considered the diffusion approximation of population genetics to gene frequency dynamics. We note that doubt has persisted about validity of the solutions of the forward diffusion equation when gene frequencies are a distance $\sim 1/(2N)$ from the boundaries $x = 0$ and $x = 1$ (see e.g., Chapter 10 of Gale, 1990). An analysis of the phenomena at the boundaries $x = 0$ and $x = 1$ was performed originally by Feller (1952). However solutions of the forward diffusion equation containing singularities i.e., Dirac delta functions (Dirac, 1958; Lighthill, 1958) at the boundaries were not considered then or in the ensuing literature on the subject. In the present work we have analysed the diffusion equation under a single type of boundary conditions, Eq. (4), that follows from the requirement that probability be conserved at all times, and consequently applies, independent of whether mutation is present or absent from the equation. Consistently taking this approach can lead to singularities (Dirac delta functions) in the solution at the boundaries, that may be identified as the distributions characterising loss or fixation of allele A. The weights of the Dirac delta functions correspond to remarkably accurate approximations for the probabilities of loss and fixation (see Figure 1). Thus the diffusion approach contains essentially complete information about the full range of gene frequencies in a more consistent manner than has been previously recognised.

The present work has implicitly emphasised that the forward diffusion equation provides a complete dynamical description of all gene frequencies. In the literature there is often recourse

to the backward diffusion equation to derive some important results. Of these, one of the most important is probably the long time fixation probability, which is written $\Pi_1(\infty)$, in the notation of the present work. It is interesting and instructive to see how such a result is obtained from the solution to the forward diffusion equation. To derive $\Pi_1(\infty)$ we assume that the solution to the forward diffusion equation has the form $f(x,t) = \sum_{n=0}^{\infty} \phi_n(x)\psi_n(p)e^{-\lambda_n t}$ i.e., a spectral sum, where $\phi_n(x)$ and $\psi_n(p)$ are, respectively, eigenfunctions of the forward and backward diffusion operators that are associated with eigenvalue λ_n (see Appendix C for further details). At long times, the only part of the solution that persists is associated with vanishing eigenvalues, $\lambda_n = 0$, hence $f(x,\infty) = \sum_n' \phi_n(x)\psi_n(p)$, where the prime on the sum indicates that it only includes eigenfunctions associated with vanishing eigenvalues. The eigenfunctions associated with zero eigenvalue can be straightforwardly found (see Appendix C) with the result that the $\phi_n(x)$ are singular (contain Dirac delta functions), while the $\psi_n(p)$ are not. The $\phi_n(x)$ are necessarily singular, since a solution of the form $f(x,t) = \sum_{n=0}^{\infty} \phi_n(x)\psi_n(p)e^{-\lambda_n t}$ has to be compatible with the singular solutions of Eq. (9). The coefficient of $\delta(1-x)$ in $f(x,\infty)$ has the interpretation as the long term fixation probability, $\Pi_1(\infty)$, and we find the standard result $\Pi_1(\infty) = \int_0^p e^{-H(q)}dq / \int_0^1 e^{-H(q)}dq$ where $H(q) = 4N \int_0^q M(y)/[y(1-y)]dy$.

In summary, we have presented a unified and consistent approach to solving the forward diffusion equation. We believe this has cleared away some of the ambiguities in the literature concerning the nature of the boundary conditions that need to be imposed on solutions of the forward diffusion equation. We have demonstrated that the solutions may contain singular parts, involving Dirac delta functions, that ensure conservation of probability and which are informative about gene fixation and loss. We have given a simple classification scheme of the boundaries (in terms of the function $M(x)$) that straightforwardly determines when fixation and loss can be expected to occur and shown how standard results, that previously have been derived from the backward diffusion equation, are contained in the solution of the forward diffusion equation.

Acknowledgements

We thank Gabriel Barton, David Broomhead, Warren Ewens and Joel Peck for helpful and stimulating discussions. This work was initiated at the 2006 EPSRC Summer School on Complexity in Ambleside, UK.

Literature Cited

Abramowitz, M., Stegun, I.1965. *Handbook of Mathematical Functions.* Dover, New York.

Barton, G., 1989. *Elements of Green's Functions and Propagation, Potentials, Diffusion, and Waves.* Clarenden Press, Oxford.

Crow, J. F., Kimura, M. 1956. Some genetic problems in natural populations, in: J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 4, University of California Press, Berkeley, pp. 1–22.

Crow, J. F., and Kimura, M. 1970. *An Introduction to Population Genetics Theory.* Harper & Row, New York

Dirac, P. A. M., 1958. *Quantum Mechanics.* Oxford University Press, Oxford, Fourth edition.

Ewens, W. J., 1979. *Mathematical Population Genetics.* Springer-Verlag, Berlin.

Feller, W., 1952. The parabolic differential equations and the associated semigroup of transformations. Annals of Mathematics 55, 468-519.

Feller, W., 1954. Diffusion processes in one dimension. Transactions of the American Mathematical Society **77**: 1-31.

Fisher, R. A., 1922. On the dominance ratio. Proc. Roy. Soc. Edin, 42, 321-431.

Fisher, R. A., 1930. *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford.

Gale, J. S., 1990. *Theoretical Population Genetics.* Unwin Hyman, London.

Gardiner, C. W., 2004. *Handbook of Stochastic Methods.* Springer, Berlin, Third edition.

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

Kimura, M., 1955b. Solution of a process of random genetic drift with a continuous model. Proc. Nat. Acad. Sci. 41, 141-150.

Lighthill, M. J., 1958. *Introduction to Fourier Analysis and Generalised Functions.* Cambridge University Press, Cambridge.

Maruyama, T., 1977 *Stochastic Problems in Population Genetics.* Springer-Verlag, Berlin.

Voronka, R., Keller, J. B. 1975. Asymptotic analysis of stochastic models in population genetics. Math. Biosciences 25, 331-362.

Wright, S., 1931. Evolution in Mendelian populations, Genetics 16, 97–159.

Wright, S., 1945. The differential equation of the distribution of gene frequencies. Proc. Natl. Acad. Sci. USA 31, 382–389.

Appendix A

In this Appendix we give mathematical details of the solution of the diffusion equation for the pure drift case, Eq. (6). The basic message will be that explicit calculations lead to Dirac delta functions developing at the boundaries $x = 0$ and $x = 1$. These delta functions ensure that probability is conserved for all times.

To correctly capture the singular (delta function) parts of any solutions, we solve the diffusion equation via the technique of Fourier transformation, since the Fourier transform of a Dirac delta function is non-singular. The Fourier transformed equation turns out to be simpler to solve than the original equation, since it does not involve hypergeometric functions.

Let us begin by introducing the quantity $\tau = t/(2N)$, which measures time in units of $2N$ generations, so that the factor of $2N$ is absent from most parts of this Appendix. We shall solve for the characteristic function $\psi(k,\tau)$, which is the Fourier transform of the probability distribution function $f(x,t)$: $\psi(k,\tau) = \int_0^1 e^{ikx} f(x,t)dx$. Using Eq. (6), and the boundary condition, Eq. (4), we find, on integrating twice by parts, that $\psi(k,\tau)$ satisfies

$$\frac{\partial}{\partial\tau}\psi(k,\tau) = -\frac{k^2}{2}\left[\frac{\partial^2}{\partial k^2} - i\frac{\partial}{\partial k}\right]\psi(k,\tau). \tag{A1}$$

The boundary terms, i.e., $\lim_{x\to 0,1} x(1-x)f(x,t)$, vanish because $f(x,t)$ is normalisable for all t, hence it cannot contain sufficiently strong power law divergences at the boundaries.

The characteristic function satisfies the usual conditions $\psi(0,\tau) = 1$ and $|\psi(k,\tau)| \leq 1$ for all k and τ, and if the initial condition is that we begin with the single frequency, $x = p$, then $f(x,0) = \delta(x-p)$ and so $\psi(k,0) = e^{ikp}$.

Assuming a separable form, $\psi(k,\tau) = \phi(k)e^{-\lambda\tau}$, for the solution of Eq. (A1) yields $\left[d^2/dk^2 - id/dk - 2\lambda/k^2\right]\phi(k) = 0$. Solutions of this equation may be written down in terms of the Bessel functions J and Y (Abramowitz and Stegun, 1965): $\phi(k) = Ae^{ik/2}\sqrt{k/2}J_{\sqrt{1+8\lambda}/2}(k/2)$ $+Be^{ik/2}\sqrt{k/2}Y_{\sqrt{1+8\lambda}/2}(k/2)$ where A and B are constants. Given that $\phi(k)$ is, up to factors, the k dependent part of a characteristic function where all moments exist (since x only ranges over a finite interval), it must contain only integer powers of k, hence $\sqrt{1+8\lambda}/2 = n + 1/2$ with $n = 0, 1, 2, ...$ i.e., λ only takes the discrete values $\lambda_n = n(n+1)/2$, $n = 0, 1, 2, \ldots$. In the analogous calculation for $f(x,t)$, this condition comes about because of boundary conditions on a hypergeometric function.

The full solution of Eq. (A1) is the linear combination $\psi(k,\tau) = \sum_{n=0}^{\infty} e^{ik/2}e^{-\lambda_n\tau}\sqrt{k/2}\left[A_n J_{n+1/2}(k/2) + B_n Y_{n+1/2}(k/2)\right]$. Since $\psi(k,\tau)$ is a characteristic function, it is bounded ($|\psi(k,\tau)| \leq 1$), which requires $B_n = 0$ for $n \geq 1$. It also satisfies $\psi(0,\tau) = 1$, which requires $B_0 = -\sqrt{\pi/2}$. Therefore

$$\psi(k,\tau) = e^{ik/2}\cos\left(\frac{k}{2}\right) + \sum_{n=0}^{\infty} e^{ik/2}\sqrt{k/2}A_n J_{n+1/2}(k/2)e^{-\lambda_n\tau}. \tag{A2}$$

We determine A_0 by differentiating (A2) with respect to k and then set both k and τ equal to zero: $ip = \partial\psi(k,\tau)/\partial k|_{k=0,\tau=0} = i/2$ $+A_0\, d\left[\sqrt{k/2}J_{1/2}(k/2)\right]/dk\Big|_{k=0} = i/2 + A_0/\sqrt{2\pi}$. Thus $A_0 = \sqrt{2\pi}i(2p-1)/2$. The remaining unknown A_n could also be obtained by using the initial condition $\psi(k,0) = e^{ikp}$, since the solution in Eq. (A2) can be expressed in terms of spherical Bessel functions $j_n(k) = \sqrt{\pi/(2k)}J_{n+1/2}(k)$, which form an orthogonal set. However it is simpler to first transform back to x dependent functions and then determine the A_n.

We note that the probability density, $f(x,t)$, was defined on the interval $0 \leq x \leq 1$, and given boundary conditions and initial data appropriate to this interval. However solving the diffusion equation via Fourier transformation, for the function $\psi(k,\tau)$ and then taking the inverse Fourier transform of $\psi(k,\tau)$ has the effect of *artificially extending* the range of x to $-\infty < x < \infty$. The boundary conditions and initial data ensure no probability density ever starts outside $0 \leq x \leq 1$, nor can ever get outside this range. An automatic consequence is that the solution for $f(x,t)$ is zero outside the interval $0 \leq x \leq 1$, as the calculations below show. This is indicated, in the solution, by the presence of the Heaviside step function, $\theta(x)$, which has the value of unity for $x > 0$, and vanishes otherwise.

Proceeding, the inverse Fourier transformation of Eq. (A2) yields

$$f(x,t) = \int_{-\infty}^{\infty} e^{-ikx}\psi(k,\tau)\frac{dk}{2\pi}$$

$$= \frac{1}{2}\left[\delta(1-x)+\delta(x)\right] + i\frac{d}{dx}\sum_{n=0}^{\infty} A_n e^{-\lambda_n \tau}\int_{-\infty}^{\infty} e^{-ir(2x-1)} r^{-1/2} J_{n+1/2}(r)\frac{dr}{2\pi}. \tag{A3}$$

The integral appearing above may be evaluated in terms of Legendre polynomials $P_n(x)$ (Abramowitz and Stegun, 1965) $\int_{-\infty}^{\infty} e^{-iwr} r^{-1/2} J_{n+1/2}(r)dr = (-i)^n\sqrt{2\pi}P_n(w)\theta\left(1-w^2\right)$. Noting that the derivative of the Heaviside step function, $\theta(x)$, is the Dirac delta function, $\delta(x)$, we find, on carrying out the differentiation in Eq. (A3), that

$$f(x,t) = \frac{1}{2}\left[\delta(1-x)+\delta(x)\right] + \sum_{n=0}^{\infty}\frac{i^{n+1}A_n}{\sqrt{2\pi}}e^{-\lambda_n\tau}\left[\delta\left(x\right)-(-1)^n\delta\left(1-x\right)\right]$$

$$+\sum_{n=1}^{\infty}\frac{i^{n+1}A_n}{\sqrt{2\pi}}e^{-\lambda_n\tau}\theta\left(x\right)\theta\left(1-x\right)\frac{d}{dx}P_n(1-2x), \tag{A4}$$

with $n \geq 1$ in the last sum since $P_0(x) = 1$. The expression for $f(x,t)$ may be simplified by introducing the Gegenbauer polynomial (Abramowitz and Stegun, 1965) $C_{n-1}^{(3/2)}(y) = (d/dy)P_n(y)$, for $n > 0$ and setting $a_n = -2(i)^{n+1}A_n/\sqrt{2\pi}$. This implies $a_0 = (2p-1)$ and gives the result

$$f(x,t) = \left[p\delta(1-x)+(1-p)\delta(x)\right] - \frac{1}{2}\sum_{n=0}^{\infty} a_{n+1}e^{-\lambda_{n+1}\tau}\left[\delta\left(x\right)+(-1)^n\delta\left(1-x\right)\right]$$

$$+\sum_{n=0}^{\infty} a_{n+1}e^{-\lambda_{n+1}\tau}C_n^{(3/2)}(1-2x)\theta\left(x\right)\theta\left(1-x\right). \tag{A5}$$

The constants a_{n+1} may be determined by using the initial condition $f(x,0) = \delta(x-p)$ together with the orthogonality of the Gegenbauer polynomials: $\int_0^1 x(1-x)C_m^{(3/2)}(1-2x)C_n^{(3/2)}(1-2x)dx = (n+1)(n+2)\delta_{nm}/[4(2n+3)]$. Multiplying $f(x,0)$, as given by Eq. (A5), by $x(1-x)C_m^{(3/2}(1-2x)$ eliminates the contributions from the delta functions at $x = 0$ and $x = 1$, and on integrating

between $x = 0$ and $x = 1$ yields $p(1-p)C_m^{(3/2)}(1-2p) = (m+1)(m+2)a_{m+1}/\left[4(2m+3)\right]$, for $m \geq 0$. Substituting this back into Eq. (A5) gives

$$f(x,t) = \delta(x)(1-p)\left[1 - \sum_{n=0}^{\infty} \frac{2p(2n+3)}{(n+1)(n+2)} C_n^{(3/2)}(1-2p)e^{-\lambda_{n+1}\tau}\right]$$

$$+ \delta(1-x)p\left[1 - \sum_{n=0}^{\infty} \frac{2(1-p)(2n+3)}{(n+1)(n+2)} (-1)^n C_n^{(3/2)}(1-2p)e^{-\lambda_{n+1}\tau}\right]$$

$$+ \theta(x)\theta(1-x)p(1-p)\sum_{n=0}^{\infty} \frac{4(2n+3)}{(n+1)(n+2)}$$

$$\times C_n^{(3/2)}(1-2p)C_n^{(3/2)}(1-2x)e^{-\lambda_{n+1}\tau}. \tag{A6}$$

The last term in Eq. (A6) coincides with the result Kimura obtained by solving the diffusion equation (6) directly (Kimura, 1955b). To see this we use the relation between the hypergeometric function and the Gegenbauer polynomials (Abramowitz and Stegun 1965): $F(-n, n+3; 2; x) = [2/(n+1)(n+2)]C_n^{(3/2)}(1-2x)$. Then, omitting the Heaviside functions, $\theta(x)\theta(1-x)$, which are irrelevant for x confined to the range 0 to 1, this third term reads

$$f_K(x,t) = p(1-p)\sum_{n=0}^{\infty}(2n+3)(n+1)(n+2)$$

$$\times F(-n, n+3; 2; p)F(-n, n+3, 2; x)e^{-\lambda_{n+1}\tau} \tag{A7}$$

which is equivalent to the result found by Kimura (1955b).

To prove the results of Eq. (8) in the main text, first consider

$$\int_0^t f_K(0,s)ds = 2Np(1-p)\sum_{n=0}^{\infty} \frac{4(2n+3)}{(n+1)(n+2)}$$

$$\times \frac{C_n^{(3/2)}(1-2p)C_n^{(3/2)}(1)}{\lambda_{n+1}}\left[1 - e^{-\lambda_{n+1}t/(2N)}\right]. \tag{A8}$$

We note that $C_n^{(3/2)}(1)/\lambda_{n+1} = 1$ since $C_n^{(3/2)}(1) = (n+1)(n+2)/2$. The t independent sum may be carried out by using the generating function for Gegenbauer polynomials, which is given by (Abramowitz and Stegun, 1965)

$$\sum_{n=0}^{\infty} C_n^{(3/2)}(y)z^n = \frac{1}{(1-2yz+z^2)^{3/2}}. \tag{A9}$$

From this we can deduce that

$$\sum_{n=0}^{\infty}\left(\frac{1}{n+1}+\frac{1}{n+2}\right)C_n^{(3/2)}(y)=\int_0^1\frac{1+z}{(1-2yz+z^2)^{3/2}}dz=\frac{1}{1-y}. \tag{A10}$$

Therefore Eq. (A8) becomes

$$\int_0^t f_K(0,s)ds=4N(1-p)-2Np(1-p)\sum_{n=0}^{\infty}\frac{4(2n+3)}{(n+1)(n+2)}$$

$$\times C_n^{(3/2)}(1-2p)e^{-\lambda_{n+1}t/(2N)} \tag{A11}$$

which is $4N$ times the coefficient of $\delta(x)$ in Eq. (A6), as required. The analogous result at the $x=1$ boundary can be proved in a similar fashion. The only difference is that the term $C_n^{(3/2)}(1)$ in Eq. (A8) is replaced by $C_n^{(3/2)}(-1)=(-1)^nC_n^{(3/2)}(1)$. The extra factor of $(-1)^n$ is equivalent to replacing y by $-y$ in Eqs. (A9) and (A10). This allows us to show that $\int_0^t f_K(1,s)ds$ is $4N$ times the coefficient of $\delta(1-x)$ in Eq. (A6).

Appendix B

In this Appendix we consider solutions of the general diffusion equation, Eq. (9), that incorporates mutation, selection and migration. We note that the solution, $f(x,t)$, is defined on the interval $0 \leq x \leq 1$. However, noting that the method adopted for solving the diffusion equation in Appendix A (Fourier transformation, followed some steps later, by inverse Fourier transformation) has the effect of *artificially extending* the range of x to $-\infty < x < \infty$, we adopt this extended range of x here. Given that no probability density ever starts outside the interval $0 \leq x \leq 1$, nor can ever get outside this range, we look for a solution of the form

$$f(x,t) = \Pi_0(t)\delta(x) + \Pi_1(t)\delta(1-x) + D(x)f_K(x,t). \tag{B1}$$

Here the function $f_K(x,t)$ is normalisable over $0 < x < 1$ and does not contain singularities at $x = 0$ and $x = 1$. The function $D(x) = \theta(x)\theta(1-x)$ has the value unity for $0 < x < 1$ and is zero outside this range. The presence of the function $D(x)$ in Eq. (B1) ensures the solution vanishes outside the interval $0 \leq x \leq 1$. Note that a property of $D(x)$ is that its derivative is $\delta(x) - \delta(1-x)$.

In the main text, we omit $D(x)$ from the solutions, since for $0 < x < 1$ the function $D(x)$ has the value of unity.

We proceed by deriving equations that determine the functions $\Pi_0(t)$, $\Pi_1(t)$ and $f_K(x,t)$ that appear in Eq. (B1), sometimes using a prime, $'$, or an overdot, $\bullet$, on a function to denote differentiation with respect to x or t.

The diffusion equation takes the form given in Eq. (9). Substituting the solution of the form Eq. (B1) into Eq. (9) leads to a left hand side of $\dot{f}_K(x,t) + \dot{\Pi}_0(t)\delta(x) + \dot{\Pi}_1(t)\delta(1-x)$.

The right hand side of the diffusion equation obtains a contribution from the pure drift term of $(4N)^{-1}\frac{\partial}{\partial x}\left[\left(\frac{\partial}{\partial x}\left[x(1-x)f_K(x,t)\right]\right)D(x)\right.$
$\left.+x(1-x)f_K(x,t)\delta(x) - \delta(1-x)\right]$. The second term in this expression is identically zero given $\lim_{x\to 0,1} x(1-x)f_K(x,t) = 0$, since $f_K(x,t)$ cannot contain sufficiently strong power law divergences at the boundaries that would prevent it being normalisable.

Carrying out the second differentiation yields $(4N)^{-1}\left(\frac{\partial^2}{\partial x^2}\left[x(1-x)f_K(x,t)\right]\right)D(x) +$ $(4N)^{-1}\left(\frac{\partial}{\partial x}\left[x(1-x)f_K(x,t)\right]\right)\left[\delta(x) - \delta(1-x)\right]$. The right hand side of the diffusion equation also obtains a contribution from the term in the diffusion equation involving M of $-\left(\frac{\partial}{\partial x}\left[M(x)f_K(x,t)\right]\right)D(x) -$ $M(x)f_K(x,t)\left[\delta(x) - \delta(1-x)\right] - \left[M(0)\Pi_0\delta'(x) + M(1)\Pi_1\delta'(1-x)\right]$.

The result of substituting Eq. (7) into Eq. (9) can be written

$$\begin{aligned}
&\dot{f}_K(x,t)D(x) + \dot{\Pi}_0(t)\delta(x) + \dot{\Pi}_1(t)\delta(1-x) \\
&= -j_K'(x,t)D(x) - \left[j_K(0,t)\delta(x) - j_K(1,t)\delta(1-x)\right] \\
&- \left[M(0)\Pi_0(t)\delta'(x) - M(1)\Pi_1(t)\delta'(1-x)\right]
\end{aligned} \tag{B2}$$

where $j_K(x,t)$ is the probability current density of Eq. (2) with $f_K(x,t)$ used in place of $f(x,t)$.

A comparison of the terms in Eq. (B2) indicates that generally $f_K(x,t)$ obeys $\dot{f}_K(x,t) = -j_K'(x,t)$ which is of identical form to the general diffusion equation Eq. (9). Furthermore, to avoid unbalanced derivatives of delta functions, it is necessary that $M(0)\Pi_0(t) = 0$ and $M(1)\Pi_1(t) = 0$. These are conditions that determine whether delta functions can be present in the solution. When

$M(0) \neq 0$ we require $\Pi_0(t) = 0$ and $j_K(0,t) = 0$ but when $M(0) = 0$ we have $\dot{\Pi}_0(t) = -j_K(0,t)$. Similarly, when $M(1) \neq 0$ we require $\Pi_1(t) = 0$ and $j_K(1,t) = 0$, but when $M(1) = 0$ we have $\dot{\Pi}_1(t) = j_K(1,t)$.

Appendix C

In this Appendix, we derive an expression for the fixation probability at long times, from solution of the forward diffusion equation, Eq. (9). The long time fixation probability is usually derived only from the backward diffusion equation.

The analysis presented in this Appendix is restricted to the case $M(0) = 0 = M(1)$, so that both gene fixation and gene loss can occur.

We begin by assuming, without proof, a solution to Eq. (9) in the form

$$f(x,t) = \sum_{n=0}^{\infty} \phi_n(x)\psi_n(p)e^{-\lambda_n t} \tag{C1}$$

i.e., a spectral sum where the functions $\phi_n(x)$ and $\psi_n(p)$ obey

$$\frac{d}{dx}\left[M(x)\phi_n(x)\right] - \frac{1}{4N}\frac{d^2}{dx^2}[x(1-x)\phi_n(x)] = \lambda_n\phi_n(x) \tag{C2}$$

$$-M(p)\frac{d}{dp}\psi_n(p) - \frac{p(1-p)}{4N}\frac{d^2}{dp^2}\psi_n(p) = \lambda_n\psi_n(p) \tag{C3}$$

and so are eigenfunctions of forward and backward diffusion operators and are both associated with eigenvalue λ_n. We make the further assumption that the smallest value of the λ_n is zero.

For Eq. (C1) to be a solution of the general diffusion equation, Eq. (9) the $\phi_n(x)$ must inherit the properties of $f(x,t)$ of having vanishing probability current at $x = 0$ and $x = 1$, i.e., $\lim_{x\to 0,1}\left(M(x)\phi_n(x) - (4N)^{-1}\, d\left[x(1-x)\phi_n(x)\right]/dx\right) = 0$ and also having the normalisability property $\lim_{x\to 0,1} x(1-x)\phi_n(x) = 0$. The required condition on the $\psi_n(p)$ is simply that they remain bounded.

For large times, we arrive at $f(x) \equiv f(x,\infty) = \sum_n' \phi_n(x)\psi_n(p)$ where the prime on the sum indicates that it only includes eigenfunctions associated with vanishing eigenvalues. Since this long time solution consists solely of eigenfunctions of the forward equation associated with vanishing eigenvalues, we have $d\left[M(x)f(x)\right]/dx - (4N)^{-1}\, d^2[x(1-x)f(x)]/dx^2 = 0$. Integrating this equation from $x = 0$ to an arbitrary x and noting that the probability current density vanishes at $x = 0$ yields $M(x)f(x) - (4N)^{-1}\, d[x(1-x)f(x)]/dx = 0$. To solve this equation, we introduce the function $g(x) = x(1-x)f(x)$, which obeys $dg(x)/dx = 4NM(x)g(x)/[x(1-x)]$. This equation has the solution $g(x) = A\exp(H(x))$ where A is independent of x and

$$H(x) = 4N\int_0^x \frac{M(y)}{y(1-y)}dy. \tag{C4}$$

It follows that $x(1-x)f(x) = A\exp(H(x))$ and as discussed in Section 4, the solution for $f(x)$ consists of the regular part $A\exp(H(x))/[x(1-x)]$ and a singular part involving Dirac delta functions, i.e., $f(x) = A\exp(H(x))/[x(1-x)] + B\delta(x) + C\delta(1-x)$ where B and C are independent of x. We note that because $H(0)$ and $H(1)$ are finite, normalisation of the solution requires $A = 0$ and $B = 1 - C$ hence $f(x)$ has the form

$$f(x) = (1-C)\delta(x) + C\delta(1-x). \tag{C5}$$

The coefficient C in this equation is generally a function of p: $C = C(p)$ and as a function of p it must be associated with vanishing eigenvalues of Eq. (C3). Thus it must obey $-M(p)dC(p)/dp -$

$(4N)^{-1}\, p(1-p)d^2C(p)/dp^2 = 0$ with the bounded solution $C(p) = DG(p) + E$ where D and E are constants and

$$G(p) = \int_0^p e^{-H(q)} dq. \tag{C6}$$

Thus $f(x) = (1 - DG(p) - E)\delta(x) + (DG(p) + E)\,\delta(1-x)$.

Lastly, we note that when $p = 0$ we must have $f(x) = \delta(x)$ so $E = 0$, similarly, when $p = 1$ we must have $f(x) = \delta(1-x)$ so $D = 1/G(1)$ hence the overall solution is

$$f(x) = \left(1 - \frac{G(p)}{G(1)}\right)\delta(x) + \frac{G(p)}{G(1)}\delta(1-x). \tag{C7}$$

As established in this work, the long time fixation probability is the coefficient of $\delta(1-x)$ in the solution of the forward diffusion equation, i.e., $G(p)/G(1)$, which is the standard result (Crow and Kimura, 1970).

As it stands, Eq. (C7) does not appear to be of the form $\sum_n' \phi_n(x)\psi_n(p)$, however it turns out that there are two eigenfunctions of Eqs. (C2) and (C3) that are associated with vanishing eigenvalues. Thus the right hand side of Eq. (C7) can be written $\phi_0(x)\psi_0(p) + \phi_1(x)\psi_1(p)$ and a possible choice of the eigenfunctions is $\phi_0(x) = 2^{-1}\left[\delta(x) + \delta(1-x)\right]$, $\psi_0(p) = 1$, $\phi_1(x) = 2^{-1}\left[\delta(x) - \delta(1-x)\right]$ and $\psi_1(p) = 1 - 2G(p)/G(1)$.