HAL
open science

Alternative to the diffusion equation in population genetics.

Bahram Houchmandzadeh, Marcel Vallade

To cite this version:

Bahram Houchmandzadeh, Marcel Vallade. Alternative to the diffusion equation in population genetics.. Physical Review E : Statistical, Nonlinear, and Soft Matter Physics, 2010, 82 (5), pp.051913. 10.1103/PhysRevE.82.051913 . hal-00633718

HAL Id: hal-00633718

https://hal.science/hal-00633718v1

Submitted on 19 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An alternative to the diffusion equation in population genetics.

Bahram Houchmandzadeh, Marcel Vallade.
CNRS & Grenoble Université, Laboratoire de Spectrométrie Physique, BP87, 38402 St-Martin d'Hères Cedex, France.

Since its inception by Kimura in 1954 (M. Kimura, PNAS, **41**:144) , the diffusion equation has become a standard technique of population genetics. The diffusion equation is however only an approximation, valid in the limit of large populations and small selection. Moreover, useful quantities such as the fixation probabilities are not easily extracted from it and need the concomitant use of a forward and backward equation. We show here that the partial differential equation governing the probability generating function can be used as an alternative to the diffusion equation with none of its drawbacks: it does not involve any approximation, it has well defined initial and boundary conditions, and its solutions are finite polynomials. We apply this technique to derive analytical results for the Moran process with selection, which encompasses the Kimura diffusion equation.

I. INTRODUCTION.

The use of the diffusion equation in problems related to population genetics was first suggested by Kolmogorov to Wright [1] and was successfully applied by Kimura[2] to genetic drift. The diffusion equation is an approximation of the discrete Master Equation governing the dynamics of a stochastic system for large populations : if the size N of the population is sufficiently large to neglect terms smaller than $1/N$, then the discrete Master Equation can be written as a continuous (in allele frequency) partial differential equation. Since the resolution of partial differential equation is much more advanced than discrete equations, the diffusion equation has been proved very popular and has become a standard technique of population genetics theory [3, 4].

The Diffusion equation in population genetics is not without its drawbacks (for a thorough discussion, see [5]). First, this is an approximation of order $1/N$, and is not suitable for small populations. There are many cases where the small populations are relevant, most important among which is when the spatial scale is included in the theory. When a species is dispersed over a wide area, different alleles of a gene will be fixed in different areas, even in the absence of environment heterogeneity and geographical barriers. The isolation by distance is due to the fact that individuals compete only against those in their migration range, the number of which can be significantly smaller than the population considered as a whole [6–9].

Other problems are more technical. The original Kimura equation is a forward equation and important quantities such as the fixation probabilities of absorbing states cannot be computed directly, but one has to resort to the accompanying backward equation [3, 4], even though solutions using distribution theory have been recently proposed [10]. Moreover, the solution of Kimura's equations is given in terms of infinite series, with a low convergence rate [11], even though recent progress in algorithms has accelerated this computation [12, 13] ; in any case, it seems unnecessary to solve a finite problem involving N coupled equations by infinite series ; it would be numerically more efficient to solve directly the N original probability equations.

In the following, we show that a partial differential equation for the probability generating function (dPGF) can be obtained from the master equation ; this equation does not include any approximation and N appears only as one of its parameters ; the equation has polynomial solutions of degree N and various quantities such as the fixation probabilities of absorbing states can be easily extracted from its stationary solution. We show the usefulness of this approach by applying it to the classical model of Moran [14], which encompasses the Kimura diffusion equation in the limit of large population.

This article is organized as follows : we will first introduce the continuous time master equation for birth-death phenomena and show how various moments can be extracted from it ; we then apply it to the Moran process and show how introduction of the dPGF can circumvent the moment closure problem. The follwing section will be devoted to the asymptotic behaviour of this equation for large time, where the fixation probabilities can be found trivially ; The fourth section is devoted to the full dynamics problem in the absence of selection ; in the next section we will include selection. The concluding section is devoted to various possible generalizations.

II. MASTER EQUATION AND DPGF DERIVATION.

Consider a continuous time birth-death stochastic process in a community of fixed size N, when the probability of observing k events in an infinitesimal time interval dt is proportional to dt^k (Poissonian events). We denote note the transition rates, the probability density for the system to change its size from n to m individuals during an infinitesimal time $dt \to 0$ by [15]:

$$W(n \to n+1) = W^+(n)$$
$$W(n \to n-1) = W^-(n)$$
$$W(n \to n+k) = 0 \text{ if } |k| > 1$$

The master equation governing $P(n,t)$, the probability of observing n individuals at time t is given by

$$\frac{\partial P(n,t)}{\partial t} = W^+(n-1)P(n-1) - W^+(n)P(n) + W^-(n+1)P(n+1) - W^-(n)P(n) \quad (1)$$

The prototype of such problems is the continuous time Moran process for haploid populations [14], a process when individuals die randomly at rate μ and are immediately replaced by the duplicate of another individual . This is a broad model which generalizes the Kimura diffusion equation (where time is considered continuous, see below). The total number of individuals carrying different alleles of a given gene is fixed to N. Let us suppose that all alleles have the same fitness (= 1) except one which we call A; without loss of generality, we include the additional fitness s into the duplication probability. Denoting by n the number of individuals carrying A and by $(N-n)$ the number of all other individuals, the transition probabilities read :

$$W^-(n) = \mu n \frac{(N-n)}{N} \quad (2)$$

$$W^+(n) = \mu(N-n)\frac{n}{N}(1+s) \quad (3)$$

In the first line, μn is the probability per unit of time that one $A-$individual dies and is replaced by a non-A ; the second line corresponds to a non-A individual dying and being replaced by an A; the factor $(1+s)$ designates the different fitness of allele A in replacing another one . In the following, without loss of generality, we will measure time in units of N/μ and we therefore set $\mu/N = 1$.

Note that If N is large, equation (1) can be approximated by the Kimura diffusion equation (see mathematical details VII A)

$$\frac{\partial p(x,t)}{\partial t} = -Ns\frac{\partial\left(x(1-x)p\right)}{\partial x} + \frac{\partial^2\left(x(1-x)p\right)}{\partial x^2}$$

where $x = n/N$ and $p(x,t) = NP(n,t)$. However, as we argued above, the diffusion equation is only an approximation of order N^{-1}(The error was precisely estimated in the case of Fisher-Wright model when no selection is present[16], but to our knowledge, no precise estimation is available for $s \neq 0$). Instead of resorting to this approximation, we can directly extract exact quantities such as the probability generating function (PGF). The PGF $\phi(z,t)$ constitutes the most complete information we can have on the given stochastic process and is defined as[15, 17]

$$\phi(z,t) = \sum_n z^n P(n,t) \quad (4)$$

where z is an auxiliary continuous variable. The systems we are looking at have two absorbing states at $n = 0$ and $n = N$. Therefore the function ϕ is in fact a polynomial of degree N : If at the initial time $t = 0$, $P(n,t) = 0$ for $n < 0$ and $n > N$, the presence of the two absorbing states ensures that this will remain so.

The equation governing the PGF can be extracted from the master equation (1) [15, 18]:

$$\frac{\partial \phi}{\partial t} = \left\langle (z^{n+1} - z^n)W^+(n)\right\rangle + \left\langle (z^{n-1} - z^n)W^-(n)\right\rangle \quad (5)$$

if the transition rates $W^{\pm}(n)$ are polynomials of degree k in n, then the right hand side of equation (5) will contain partial derivatives of order k of the function ϕ with respect to z. Therefore, the discrete master equation (1) is naturally and without any approximation transformed into a partial differential equation which we call the dPGF.

Application of the above principle to the Moran process (eqs 2,3) provides the Moran dPGF (see Mathematical Details VII B) which we will investigate in this article :

$$\frac{\partial \phi}{\partial t} = \frac{1}{\sigma}(1-z)\left(\sigma - z\right)\frac{\partial}{\partial z}\left(N\phi - z\frac{\partial \phi}{\partial z}\right) \quad (6)$$

where σ is the inverse of the fitness : $\sigma = 1/(s+1)$. This is a well defined partial differential equation, first order in t and second order in z and has the same formal structure as the diffusion equation. However, the boundary conditions of this equation are unequivocally specified. If at time $t = 0$, the number of $A-$individuals is n_0, then $P(n,0) = \delta_{n,n0}$ and

$$\phi(z,0) = z^{n_0} \quad (7)$$

Moreover, from the definition of the PGF,

$$\phi(1,t) = 1 \quad (8)$$

If $s = 0$, then $W^+(n) = W^-(n)$, $\langle n(t)\rangle = n_0$ and therefore

$$\left.\frac{\partial \phi}{\partial z}\right|_{=1} = n_0 \ \text{ if } s = 0 \quad (9)$$

If $s \neq 0$, $z = \sigma$ is a fixed point of equation (6) ($\partial\phi/\partial t|_{z=\sigma} = 0$) and therefore

$$\phi(\sigma,t) = \phi(\sigma,0) = \sigma^{n_0} \quad (10)$$

The set of equations (6) with the initial condition (7) and the boundary condition (8) and (9) or (10) constitute a well defined problem ; this is not the case for the forward diffusion equation used in population genetics, where the equation is valid only for gene frequencies $x \in]0,1[$ and the terminal classes $x = 0$ and $x = 1$ have to be treated separately by some ad hoc treatment (see for example [3], p379-80).

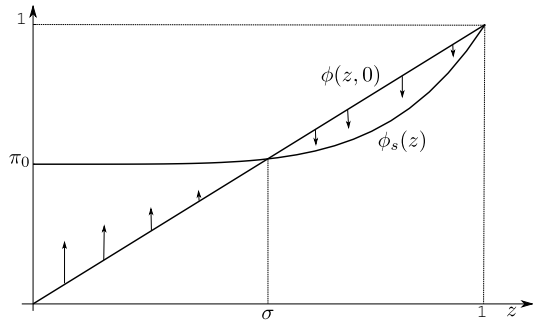

Figure 1. The evolution of the probability generating function $\phi(z,t)$ for large time. The PGF, from the initial condition $\phi(z,0) = z^{n_0}$ converges to the stationary solution $\phi_s(z) = \pi_N z^N + \pi_0$ (arrows indicate the direction of evolution of $\phi(z,t)$). In this illustration, $n_0 = 1$ where n_0 is the initial number of $A-$individual. Points $z = 1$ and $z = \sigma$ are fixed points of the evolution equation (6).

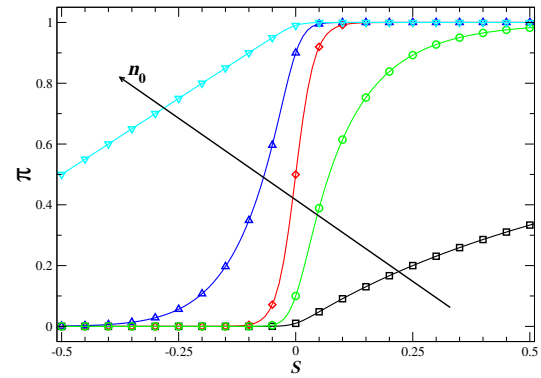

Figure 2. fixation probabilities π_N as a function of relative fitness s for different values of initial number of allele $n_0 = 1, 10, 50, 90, 99$. Symbols represent direct numerical resolution of master equation (1) for the Moran process. Continuous lines represent theoretical expressions given by equation (12). Total number of individuals $N = 100$.

III. STATIONARY SOLUTION AND THE LIMIT FOR LARGE TIMES.

Figure 1 captures the dynamics of the PGF $\phi(z,t)$ and its convergence towards the stationary solution $\phi_s(z)$. The stationary solution of equation (6) is given by

$$N\phi_s - z\phi_s' = K$$

where K is a constant. This is is an ordinary first order differential equation and its solution is:

$$\phi_s(z) = \pi_N z^N + \pi_0.$$

Using the boundary conditions (8) and (9) when $s = 0$, the two integration constants are found to be

$$\pi_N = \frac{n_0}{N} \; ; \; \pi_0 = \frac{N - n_0}{N} \tag{11}$$

When $s \neq 0$, equation (9) has to replaced by (10) and

$$\pi_N = \frac{1 - \sigma^{n_0}}{1 - \sigma^N} \; ; \; \pi_0 = \frac{\sigma^{n_0} - \sigma^N}{1 - \sigma^N} \tag{12}$$

where as mentioned, $\sigma = 1/(1+s)$. Note that as $s \to 0$, equations (12) converge to equations (11). Probabilities $P(n, t \to \infty)$ can be extracted from the stationary PGF :

$$P(n, \infty) = \pi_N \delta_{n,N} + \pi_0 \delta_{n,0}$$

specifically, π_N and π_0 are the total probability of fixation and loss of allele. With the dPGF method, the fixation probabilities are easily obtained without any approximation or hypothesis on the value of N and s. Obtaining this result is more intricate with other methods such as: (i) looking for a functional equation governing the discrete time PGF, as was originally done by Fisher [19] and reviewed by Moran ; (ii) when the backward diffusion equation [3] is used ; (iii) when two discrete Markov processes are embedded in order to transform the problem into the "gambler's ruin" one [20] as done by Moran [14]. Note that $\sigma^n = e^{n \log \sigma} \approx e^{-ns}$ for small relative fitness s, and therefore π_N given by equation (12) contains the well known result for fixation probability for haploid populations

$$u = \frac{1 - e^{-sn_0}}{1 - e^{-sN}}$$

as a particular approximation. Figure 2 shows the comparison between direct numerical resolution of the master equation (1) for the Moran process and its comparison to expressions (12) for the fixation probabilities.

IV. PURE GENETIC DRIFT.

We now turn our attention to the full solution of equation (6) when no selection is present, *i.e.* $s = 0$. The case $s \neq 0$ will be studied in the next section. The master equation (1) is a system of $N+1$ first order linear differential equations with one constraint ($\sum_n P(n,t) = 1$) and therefore its general solution is of the form

$$P(n,t) = \sum_{k=0}^{N-1} \beta_k^n e^{\lambda_k t}$$

The PGF $\phi(z,t)$ being only a combination of these probabilities weighted with functions z^n, it is natural to search for its solution as a *finite* superposition of eigenfunctions $\psi_n(z)\exp(\lambda_n t)$ where ψ_n and λ_n are solutions of the eigenvalue equation

$$\lambda\psi(z) = (1-z)^2 \frac{d}{dz}\left(N\psi(z) - z\psi'(z)\right) \tag{13}$$

The solution of the above equation can be given in terms of hypergeometric functions ${}_2F_1(y)$, where $y = 1/(z-1)$

[18] or Hahn's polynomials [21] ; for the purpose of this article and having in mind the case $s \neq 0$, it is more fruitful to solve it directly using the polynomial nature of the solutions. We already know the stationary solution $\lambda_0 = 0$, $\psi_0(z) = \pi_0 + \pi_N z^N$. For $\lambda \neq 0$, as $z = 1$ is a double zero of $\psi(z)$, we look for solutions as polynomials of $(1-z)$, *i.e.*

$$\psi(z) = \sum_{k=0}^{N-1} a_k (1-z)^{k+1} \tag{14}$$

which gives rise to the following *two* term recurrence relations between the coefficients a_k :

$$a_0 = 0 \tag{15}$$

$$[\lambda + k(k+1)]\, a_k = k(k-N) a_{k-1} \quad k = 1, ...N-1 \tag{16}$$

As $a_0 = 0$, non trivial solutions are found only if $\lambda = -n(n+1)$ for some integer n ; we use this integer to order the eigenvalues λ_n and eigenfunctions $\psi_n(z)$:

$$\lambda_n = -n(n+1) \quad n = 1, 2, ...N-1 \tag{17}$$

$$\psi_n(z) = \sum_{k=n}^{N-1} a_k^n (1-z)^{k+1} \tag{18}$$

$$a_n^n = 1$$

$$a_k^n = \frac{k(N-k)}{n(n+1) - k(k+1)} a_{k-1}^n \quad k = n+1, ..., N-1 \tag{19}$$

The coefficients a_k^n can be put into explicit form in terms of binomial coefficients (see Mathematical details VII C). The PGF is given in terms of the above eigenfunctions as

$$\phi(z,t) = \pi_0 + \pi_N z^N + \sum_{n=1}^{N-1} C_n \psi_n(z) e^{\lambda_n t} \tag{20}$$

where the coefficients C_n are determined from the initial condition $\phi(z,0) = z^{n_0}$. The a_k^n matrix is triangular and therefore this determination is straightforward (see Mathematical details VII C).

Expanding $\psi_n(z)$ using the binomial development of $(1-z)^{k+1}$ and identifying the result with the PGF definition (4) we obtain the probabilities $P(n,t)$ as :

$$P(n,t) = \pi_0 \delta_{n,0} + \pi_N \delta_{n,N} + (-1)^n \sum_{k=n}^{N} \alpha_{k-1}(t) \binom{k}{n} \tag{21}$$

where

$$\alpha_k(t) = \sum_{n=1}^{N-1} C_n a_k^n e^{\lambda_n t} \quad k = 1, ..., N-1$$

and $\alpha_{-1}(t) = \alpha_0(t) = 0$.

Note that the above expressions are exact solutions. However, as the eigenvalues $\lambda_n = -n(n+1)$ increase rapidly, these expressions can be approximated by taking into account only the first few eigenfunctions, depending on the degree of accuracy required. Figures (3,4) show the accuracy of the above formula by comparing them to the numerical resolution of the Master equation (1).

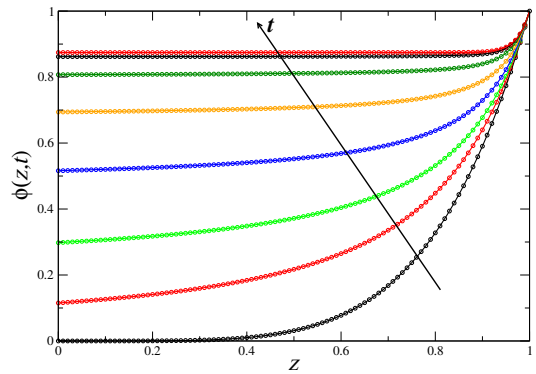

Figure 3. The PGF $\phi(z,t)$ as a function of z for various times t_i for genetic drift $s = 0$. The PGF is computed directly by numerical resolution of the Master equation (1) (continuous line) and is compared to its theoretical expression given by eq.(20) (circle). $N = 40$, $n_0 = 5$, times $t_i = 0, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2$ (in units of N/μ)

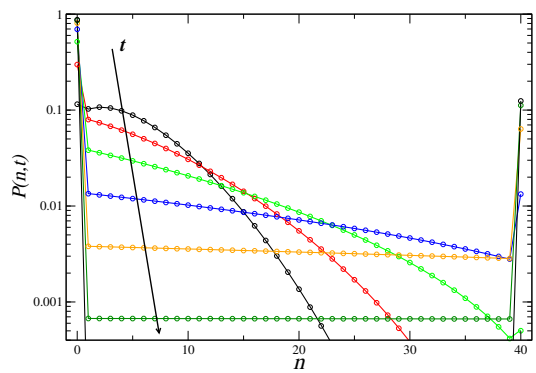

Figure 4. Probabilities $P(n,t)$ as a function of n at various times for genetic drift $s = 0$. The probabilities are computed by direct numerical resolution of the Master equation (1) (continuous lines) and are compared to their theoretical expression given by eq.(20) (circles). $N = 40$, $n_0 = 5$, times $t_i = 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2$ (in units of N/μ)

Historically, problems of evolution were formulated in the framework of Fisher Wright (FW) model. Moran and FW are equivalent at the large population limit, where both are well approximated by the same diffusion equation. The exact solution $P(n,t|n_0,0)$ dervied above (eq. 21) allows for a direct comparison between them. FW is a discrete time, non-overlapping generations, $N-$step model where the probability of having n individuals at generation $T+1$ given that there are n_0 at generation T is

$$P_{FW}(n|n_0) = \left(\frac{n_0}{N}\right)^n \left(1 - \frac{n_0}{N}\right)^{N-n} \binom{N}{n} \tag{22}$$

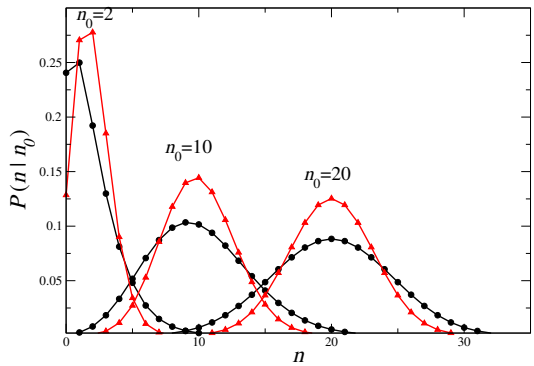

Figure 5. Comparison between the discrete time, $N-$step Moran distribution $P_M(n|n_0) = P(n, t = 1/N|n_0)$ (black, circles) and Fisher-Wright (red, triangle) distribution $P_{FW}(n|n_0)$ for three different inital values n_0.

The Moran process is a one-step model over infinitesimal time; it transforms into a $N-$step one if we consider it over the finite time of one generation, $P_M(n|n_0) = P(n, t = 1/N|n_0, 0)$. Figure 5 shows the comparison between these two processes where it can be observed that the FW process has a norrower distribution than the Moran one. Moran[14] had pointed to this difference by computing the probability of the number of descendant of one individual in both processes.

V. INCLUDING SELECTION.

When selection is present and $s \neq 0$, the spectral decomposition is achieved by solving the eigenvalue equation

$$\lambda\sigma\psi = (1-z)(\sigma - z)\frac{d}{dz}(N\psi - z\psi') \tag{23}$$

where $\sigma = 1/(1+s)$ as defined before. This equation is called Heun's equation [22]. Heun's polynomials and their eigenvalues have been less studied than for example the oblong spheroïd ; there is, to our knowledge, no explicit formula or fast algorithm for their computation. However, we are interested in the small s limit ($s \ll 1$) and therefore we can compute the solution of (23) by the perturbation technique in powers of s. The first order perturbation solution, satisfactory for $Ns \lesssim 1$, is directly obtained from the pure genetic drift solution by a simple scaling : note that if we set $y = 1 - z/\sqrt{\sigma}$, equation (23), transforms into

$$-\sqrt{\sigma}\lambda\psi = \left(y^2 - \epsilon(y-1)\right)\frac{d}{dy}(N\psi + (1-y)\psi') \tag{24}$$

where

$$\epsilon = \sqrt{\sigma} + 1/\sqrt{\sigma} - 2 = \frac{s^2}{4} + \mathcal{O}(s^3) \tag{25}$$

The ϵ term in the transformed equation (24) is $\sim s^2$ and is therefore neglected in the first order (in s) calculation.

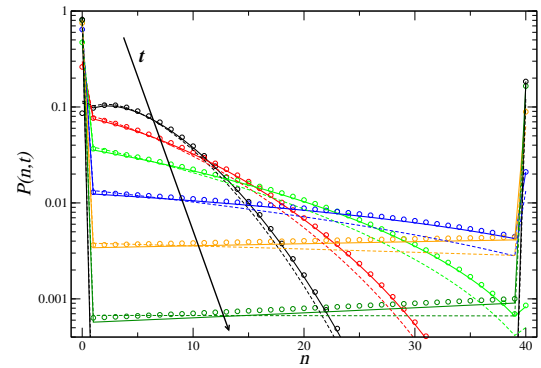

Figure 6. The probabilities $P(n, t)$ as a function of n at various times t_i, for $s = 2.5\times10^{-2}$ ($Ns = 1$). (i) continuous line : direct numerical resolution of the Master equation (1) ; (ii) circles : theoretical expression (26) corresponding to first order perturbations ; (iii) dotted lines : solutions for $s = 0$ (from Fig. 4). $N = 40$, $n_0 = 5$, times $t_i = 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2$ (in units of N/μ).

Neglecting $\mathcal{O}(\epsilon)$ terms, equation (24) acquires the same structure as the equation (13) for the pure genetic drift which we have already solved. The PGF $\phi(z, t)$ therefore reads

$$\phi(z,t) = \pi_0 + \pi_N z^N + \sum_{n=1}^{N-1} C_n^{(1)}\psi_n^{(1)}(z)e^{-n(n+1)t/\sqrt{\sigma}} + \mathcal{O}(s^2) \tag{26}$$

where

$$\psi_n^{(1)} = \sum_{k=n}^{N-1} a_k^n (1 - z/\sqrt{\sigma})^{k+1}$$

The coefficients a_k^n are the same as in (19) ; the amplitudes $C_n^{(1)}$ are obtained as before by using the initial condition $\phi(z, 0) = z^{n_0}$. Figure 6 shows the accuracy of the first order solution for $Ns = 1$.

The computation can be extended to second order perturbations in s (see Mathematical Details VII C). Note however that for large value of Ns, the term $z\partial_z\phi$ in equation (6), is comparable to $N\phi$ only in the vicinity of $z = 1$. Therefore, for $z \in [0, \sigma]$, we can neglect this term and use the approximate equation

$$\sigma\frac{\partial\phi}{\partial t} = N(1-z)(\sigma - z)\frac{\partial\phi}{\partial z}$$

which is a first order differential equation and can be solved exactly :

$$\phi(z,t) = \left(\frac{(\sigma - z)e^{-Nst} - \sigma(1-z)}{(\sigma - z)e^{-Nst} - (1-z)}\right)^{n_0}$$

This is indeed a good approximation of the PGF for $Ns \gtrsim 2$ in the interval $[0, \sigma]$. As ϕ is not $z-$polynomial anymore retrieving the probabilities $P(n, t)$ from this function by successive derivation is numerically fragile ; the formula

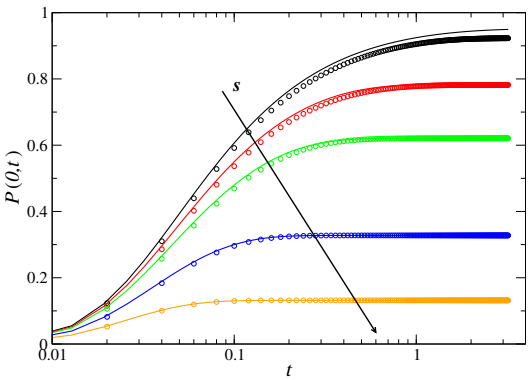

Figure 7. Loss of allele probability $P(0,t)$ as a function of time, measured in (N/μ) units, for various values of the additional fitness s. Continuous curves correspond to eq. (27) , circles to direct numerical resolution of the Master equation 1). $N = 100$; $n_0 = 5$; $s = 0.01, 0.05, 0.1, 0.25, 0.5$.

however is very accurate for small n, and specially for the loss of allele probability as a function of time which takes the simple form

$$P(0,t) = \left(\frac{1 - e^{-Nst}}{1 + s - e^{-Nst}} \right)^{n_0} \tag{27}$$

Figure (7) show the accuracy of this approximation for $Ns \in [1, 50]$ interval.

VI. CONCLUDING REMARKS.

We have shown the usefullness of dPGF technique in population genetics through the example of the Moran process. We have shown that a partial differential equation can be obtained for the probability generating function that is not an approximation and which has finite polynomial solutions. The solutions can be computed exactly for pure genetic drift and with the perturbation techniques when $s \neq 0$, and we have shown the agreement with the numerical solution of the original Master Equation.

The usefulness of the dPGF technique is very broad and can be used to capture many features of population dynamics. For example mutations can be included by considering two alleles A and a, with mutation rate from one to others being ν_1 and $\nu 2$. Denoting by n the number of $A-$alleles, the transition rates read :

$$W^+(n) = (N - n)\left(n(1 + s)(1 - \nu_1) + (N - n - 1)\nu_2\right)$$
$$W^-(n) = n\left((N - n)(1 - \nu_2) - (n - 1)\nu_1\right)$$

and give rise to a dPGF equation which has a similar structure to equation (6) and can be studied by the same methods. The diploid populations can be studied by including transition rates $W(n \to n \pm 2)$. More importantly, we could include the spatial extension of the ecosystem by dividing the ecosystem into patches and modifying the transition rates to include migrations from adjacent patches, which again will include linear terms in the transition rates [9, 23]. This would be an important step to show the possibility of sympatry in speciation. Other problems which could be modeled by this technique are the selection of social behaviour and the control of the cheaters. The dPGF technique has the potential to investigate by simple means a large number of problems of population genetics.

VII. MATHEMATICAL DETAILS.

A. Diffusion equation.

To transform the discrete master equation (1) into a continuous diffusion equation for large N, set $x = n/N$, $dx = 1/N$, $p(x,t)dx = P(n,t)$, $w^{\pm}(x) = W(n)$. Developing equation (1) into powers of dx, one finds

$$\frac{\partial p(x,t)}{\partial t} = -\frac{1}{N}\frac{\partial\left(a(x)p(x,t)\right)}{\partial x} + \frac{1}{2N^2}\frac{\partial^2\left(b(x)p(x,t)\right)}{\partial x^2} + O(dx^3)$$

where $a(x) = w^+(x) - w^-(x)$ and $b(x) = w^+(x) + w^-(x)$. In the particular case of the Moran Process, $a(x) = N^2 sx(1-x)$, $b(x) = 2N^2 x(1-x)(1+s/2)$; neglecting higher order terms the above equation reads

$$\frac{\partial p(x,t)}{\partial t} = -Ns\frac{\partial\left(x(1-x)p(x,t)\right)}{\partial x} + (1+\frac{s}{2})\frac{\partial^2\left(x(1-x)p(x,t)\right)}{\partial x^2}$$

The Kimura equation is an approximation of the above diffusion equation for small s, when the term $s/2$ can be neglected compared to unity. To go beyond the limit of small s, we renormalize the time $t' = t(1 + s/2)$ and the fitness $s' = s/(1 + s/2)$ to find

$$\frac{\partial p(x,t)}{\partial t'} = -Ns'\frac{\partial\left(x(1-x)p(x,t)\right)}{\partial x} + \frac{\partial^2\left(x(1-x)p(x,t)\right)}{\partial x^2}$$

which is again similar to the classical Kimura equation, valid for arbitrary s. We have to keep in mind however that the coefficient s' can be markedly different from the fitness s, when the latter is not small compared to unity.

B. Moran dPGF derivation.

Consider the master equation (1) with the Moran transition rates (2,3). We define the PGF as

$$\phi(z,t) = \sum_n z^n P(n,t) \tag{28}$$

where z is an auxiliary continuous variable. To derive the PGF equation, multiply both sides of equation (1) by z^n and sum over the index n. The left hand side of the equation is

$$\sum_n z^n \frac{\partial P(n,t)}{\partial t} = \frac{\partial}{\partial t}\sum_n z^n P(n,t) = \frac{\partial \phi(z,t)}{\partial t}$$

For the right hand side, consider for example the term

$$\sum_n z^n W^+(n-1)P(n-1) = \sum_n z^{n+1} W^+(n)P(n) \quad (29)$$

Recall that because of the existence of the two absorbing state $n = 0$ and $n = N$ and the initial condition

$$P(n, 0) = 0 \text{ if } n < 0 \text{ or } n > N$$

the sum can be extended to $n \in \mathbb{Z}$ and therefore the change of the summation variable from n to $n+1$ in (29) has no effect on the boundaries of the summation. Performing this change of variable on all terms, the equation for the PGF reads :

$$\frac{\partial \phi}{\partial t} = \left\langle (z^{n+1} - z^n) W^+(n) \right\rangle + \left\langle (z^{n-1} - z^n) W^-(n) \right\rangle \quad (30)$$

For the Moran process, transition rates are of the form

$$W^{\pm}(n) = k^{\pm} n(N-n)$$

where $k^- = 1$ and $k^+ = (1+s)$ are constant (Recall that time is measured in units of N/μ). Consider the general term

$$\left\langle z^{n+\alpha} kn(N-n) \right\rangle = kz^{\alpha} \left(N \left\langle nz^n \right\rangle - \left\langle n^2 z^n \right\rangle \right) \quad (31)$$

From the definition (28), it is easily shown that

$$\left\langle nz^n \right\rangle = z \frac{\partial \phi}{\partial z}$$

or in general terms, $\left\langle n^k z^n \right\rangle = (z\partial_z)^k \phi(z,t)$. Replacing these terms in (31) reads

$$\left\langle z^{n+\alpha} kn(N-n) \right\rangle = kz^{\alpha} z \frac{\partial}{\partial z} \left[N\phi - z \frac{\partial \phi}{\partial z} \right]$$

Replacing the above terms in equation (30) we obtain

$$\begin{aligned} \frac{\partial \phi}{\partial t} &= \left(k^+ z^2 - k^+ z + k^- - k^- z \right) \frac{\partial}{\partial z} \left[N\phi - z \frac{\partial \phi}{\partial z} \right] \\ &= (1-z)(k^- - k^+ z) \frac{\partial}{\partial z} \left[N\phi - z \frac{\partial \phi}{\partial z} \right] \\ &= (1-z)\left(1 - (1+s)z\right) \frac{\partial}{\partial z} \left[N\phi - z \frac{\partial \phi}{\partial z} \right] \end{aligned}$$

which is the displayed equation (6).

C. Explicit expression for coefficients.

The recurrence relation for the coefficients of the eigenfunctions of equation (13) is

$$a_k^n = -\frac{k(N-k)}{k(k+1) - n(n+1)} a_{k-1}^n \quad k = n+1, ..., N-1$$

where $a_n^n = 1$. We can therefore compute the product directly

$$a_k^n = \frac{(1-N+n)_{k-n}}{(2n+2)_{k-n}} \binom{k}{n}$$

where it is assumed that Binomial$(k, n) = 0$ if $k < n$ and $(\alpha)_\beta$ is the Pochhammer symbol $\Gamma(\alpha+\beta)/\Gamma(\beta)$.

As the eigenfunctions are given as polynomials of $(1-z)$, let us set $y = 1 - z$. The coefficients $a_k^n \neq 0$ only for $k \geq n$, and the matrix a_k^n is a $(N-1) \times (N-1)$ triangular matrix where its diagonal elements are unity. To determine the coefficients C_n in the equation (20) we use the initial condition $\phi(z, 0) = z^{n_0}$:

$$\begin{aligned} \sum_{n=1}^{N-1} \sum_{k=1}^{N-1} C_n a_k^n y^{k+1} &= (1-y)^{n_0} - \pi_0 - \pi_N (1-y)^N \\ &= \sum_{k=1}^{N-1} b_k y^{k+1} \end{aligned}$$

b_k is the result of the binomial development of the above expression and reads

$$b_k = (-1)^k \left(\pi_N \binom{N}{k+1} - \binom{n_0}{k+1} \right)$$

The C_n are then extracted from the linear triangular system

$$\sum_{n=1}^{N-1} C_n a_k^n = b_k \quad k = 1, ..., N-1 \quad (32)$$

and can also be given explicitly,

$$\begin{aligned} C_n = (-1)^{n+1} n_0 \frac{(1-N)_n}{(n+1)_n} \times \\ {}_3F_2(1-n_0, -n, n+1; 2, 1-N; 1) \end{aligned}$$

It is more efficient to solve directly the linear triangular system (32).

When $s \neq 0$, the first order (in s) amplitudes $C_n^{(1)}$ are obtained by the same procedure, except that now the coefficients b_k are defined as

$$b_k = (-1)^k \left(\pi_N \binom{N}{k+1} \sigma^{N/2} - \binom{n_0}{k+1} \sigma^{n_0/2} \right)$$

The same procedure can be extended to perturbations of order $\sim s^2$ and it extends the range of validity to $Ns \lesssim 10$. The computation is more cumbersome and we give here only the results on the eigenvalues :

$$\lambda_n = -n(n+1) \left(1 + \epsilon \frac{N^2 - 1 + n(n+1)}{2(2n-1)(2n+3)} \right) \sigma^{-1/2}$$

For $Ns = 10$, the relative deviations from exact values are at most 4% for the first eigenvalues and become negligible for large n's.

a. Acknowledgements. We thank O. Rivoire for fruitful discussion of the manuscript. This work was partly funded by Agence Nationale de la Recherche Française (ANR) grant "Evo-Div".

[1] S. Wright. The differential equation of the distribution of gene frequencies. *PNAS*, 31:382–389, 1945.

[2] M. Kimura. Solution of a process of random genetic drift with a continuous model. *Proc. Nat. Ac. Sci. (USA)*, 41:144, 1955.

[3] J. F. Crow and M. Kimura. *An Introduction to Population Genetics Theory*. The Blackburn Press, Cal, 2009.

[4] W. J. Ewens. *Mathematical Population Genetics*. Springer-Verlag, 2004.

[5] J. Wakely. The limits of theoretical population genetics. *Genetics*, 169:1–7, 2005.

[6] G. Malécot. *Les mathématiques de l'hérédité*. Masson, Paris, 1948.

[7] M. Kimura and G. H. Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49:561–576, 1964.

[8] J. Felsenstein. A pain in the torus : some difficulties with models of isolation by distance. *American Naturalist*, 109:359–368, 1975.

[9] B. Houchmandzadeh and M. Vallade. Clustering in neutral ecology. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(6 Pt 1):061912, Dec 2003.

[10] A. J. McKane and D. Waxman. Singular solutions of the diffusion equation of population genetics. *J Theor Biol*, 247(4):849–858, Aug 2007.

[11] F. Chalub and M. O. Souza. From discrete to continuous evolution models: A unifying approach to drift-diffusion and replicator dynamics. *Theor Popul Biol*, 76(4):268–277, Dec 2009.

[12] L. Li, M. Leong, T. Yeo, P. Kooi, and K. Tan. Computations of spheroidal harmonics with complex arguments: A review with an algorithm. *Phys. Rev. E*, 58(5):6792–6806, Nov 1998.

[13] Y. Wang and B. Rannala. A novel solution for the time-dependent probability of gene fixation or loss under natural selection. *Genetics*, 168(2):1081–1084, Oct 2004.

[14] P.A.P. Moran. *The Statistical processes of of evolutionary theory*. Oxford University Press, 1962.

[15] C. Gardiner. *Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences*. Springer, 2004.

[16] S. N. Ethier and M. F. Norman. Error estimate for the diffusion approximation of the wright–fisher model. *Proc Natl Acad Sci U S A*, 74(11):5096–5098, Nov 1977.

[17] P.S. Laplace. *Théorie analytiques des probabilités*. 1812.

[18] A. McKane, D. Alonso, and R. V. Solé. Mean-field stochastic theory for species-rich assembled communities. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, 62:8466–8484, 2000.

[19] R.A. Fisher. *The genetical theory of natural selection, a complete variorum edition*. Oxford University Press, 1999.

[20] W. Feller. *An introduction into probability theory and its applications*. John Wiley & Sons, 1962.

[21] S. Karlin and J. McGregor. On a genetics model of moran. *Math. Proc. Camb. Phil. Soc.*, 58:299 –311, 1962.

[22] A. Ronveaux, editor. *Heun's Differential Equations*. Oxford University Press, 1995.

[23] M. Vallade and B. Houchmandzadeh. Analytical solution of a neutral model of biodiversity. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68(6 Pt 1):061902, Dec 2003.