ESAIM: M2AN 53 (2019) 615–634
https://doi.org/10.1051/m2an/2018058

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

NUMERICAL COMPLETE SOLUTION FOR RANDOM GENETIC DRIFT BY ENERGETIC VARIATIONAL APPROACH

Chenghua Duan[1], Chun Liu[2], Cheng Wang[3] and Xingye Yue[1,*]

Abstract. In this paper, we focus on numerical solutions for random genetic drift problem, which is governed by a degenerated convection-dominated parabolic equation. Due to the fixation phenomenon of genes, Dirac delta singularities will develop at boundary points as time evolves. Based on an energetic variational approach (EnVarA), a balance between the maximal dissipation principle (MDP) and least action principle (LAP), we obtain the trajectory equation. In turn, a numerical scheme is proposed using a convex splitting technique, with the unique solvability (on a convex set) and the energy decay property (in time) justified at a theoretical level. Numerical examples are presented for cases of pure drift and drift with semi-selection. The remarkable advantage of this method is its ability to catch the Dirac delta singularity close to machine precision over any equidistant grid.

Mathematics Subject Classification. 35K65, 92D10, 76M28, 76M30.

Received August 20, 2017. Accepted September 25, 2018.

1. Introduction

Random genetic drift is the phenomenon that the frequency of a gene variant (allele) in a population changes at the next generation due to random sampling. The process of random genetic drift plays an important role in the molecular evolution [7] and the behavior of genes in a population with a finite size [21]. From the view-point of population genetics, the most elementary step in the evolution is the change of gene frequencies. The notion and technique of random genetic drift have been widely applied to medical science [29] and other fields.

We consider a population with a finite size, which can generally cause the random genetic drift. The change in gene frequencies is treated as a stochastic process, which was first introduced by Fisher [13]. Under the assumption that generations do not overlap and each copy of gene in the new generation is chosen independently at random from all copies in the old generation, the mathematical model of genetic drift is labeled as the Wright-Fisher Model, introduced by Fisher [14] and Wright [34], and developed by Kimura [17]. This mathematical model is a formulation based on a discrete-time Markov chain. The model involves two alleles: A and a in a population with a fixed size N_e. The quantities X_t and f_t denote the proportion of A in the population and its probability distribution at generation t, respectively. Assume that the number of gene A is m at generation t,

Keywords and phrases. Random genetic drift, wright-fisher model, energetic variational approach, convex splitting scheme, Dirac delta singularity, fixation phenomenon.

[1] School of Mathematical Sciences, Soochow University, 215006 Suzhou, PR China.

[2] Department of Applied Mathematics, Illinois Institute of Technology, 60616 Chicago, IL, USA.

[3] Department of Mathematics, University of Massachusetts, Dartmouth, 02747-2300 North Dartmouth, MA, USA.

*Corresponding author: xyyue@suda.edu.cn

Article published by EDP Sciences

© EDP Sciences, SMAI 2019

then, at the next generation $t+1$, the transition probability for the number of gene A being n is given by

$$W_{n,m} \equiv P\left(X_{t+1} = \frac{n}{2N_e}\Big| X_t = \frac{m}{2N_e}\right) = \binom{2N_e}{n}\left(\frac{m}{2N_e}\right)^n \left(1-\frac{m}{2N_e}\right)^{(2N_e-n)},$$

under the circumstance that there is no factor such as mutation, migration and selection and the only evolutionary force is genetic drift. We get the distribution of probability at generation $t+1$ by the Markov chain: $f_{t+1,n} = \sum_{m=1}^{2N_e} W_{n,m} f_{t,m}$. At continuous time and space, $f_{t,n}$ is approximate to $f(x,t)$. Kimura [17,20,36] showed that for *pure drift* (the only evolutionary force is genetic drift), $f(x,t)$ obeys the diffusion equation:

$$\frac{\partial}{\partial t} f(x,t) = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}(x(1-x)f(x,t)),\ x \in (0,1), \quad t > 0, \tag{1.1}$$

where N_e is the population size. Moreover, if mutation, migration and selection effects are involved, the model becomes

$$\frac{\partial}{\partial t} f(x,t) = \frac{1}{4N_e}\frac{\partial^2}{\partial x^2}(x(1-x)f(x,t)) - \frac{\partial}{\partial x}(M(x)f(x,t)),\ x \in (0,1), \quad t > 0, \tag{1.2}$$

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.

We take the zero current boundary condition

$$\left(\frac{1}{4N_e}\partial_x(x(1-x)f(x,t)) - M(x)f(x,t)\right)|_{x=0,1} = 0, \quad t > 0,$$

with $M(x) = 0$ for pure drift and a initial state

$$f(x,0) = f_0(x) = \delta(x - x_0), \tag{1.3}$$

which means that at initial time, the proportion of gene A is $x_0 \in [0,1]$.

A complete solution, *i.e.*, the total probability is equal to unity at any time, develops sharp spikes (Dirac delta singularities) at the two boundary 0 and 1. When the sharp spikes appear, they signal gene loss or gene fixation: either all copies of gene A are finally lost, or all individuals carry A (gene a is totally lost). A complete solution is essential in Wright Fisher model, because the complete solution can include all possible outcomes whenever fixation and loss are possible, and can be extremely close correspondence with Wright-Fisher model.

For the pure drift case, it has been shown that this system keeps the conservation of the total probability and expectation, and $f(x,t) \to (1-x_0)\delta(x) + x_0\delta(x-1)$, as $t \to \infty$ which means that there is a probability of x_0 that the fixation occurs at gene A and a probability of $1-x_0$ that the fixation occurs at gene a [6,23,28].

When considering an unlinked locus with two alleles subjects to the semi-dominant selection with strength s ($\mid s \mid \ll 1$), we take $M(x) = sx(1-x)$ as in [17,20]. In this case, the probability of ultimate fixation of gene A from an initial expectation x_0 is $P_{fix}(x_0) = \frac{1-e^{-4N_e s x_0}}{1-e^{-4N_e s}}$ [19,36].

However, except for a few special cases, we could not get explicit solutions. The numerical approaches are needed to obtain the approximate solutions for the differential equation. Some attempts have been made by Kimura [17], Barakat and Wagener [1] and Wang [31], while the total probability is smaller than unity and it was also a hard work to simulate the general case including natural selection, mutation and migration. Zhao *et al.* [36] obtained a complete numerical solution by finite volume method (FVM) for a neutral locus and semi-selection. In [6], Xu *et al.* discussed three classical numerical schemes which are stable but lead to different steady state solutions. Only one of the schemes gives a true complete numerical solution and any scheme with numerical viscosity should be avoided. Therefore, a very careful analysis for the numerical scheme is necessary.

In this paper, we propose a new scheme based on energetic variational approach (EnVarA). Combining the least action principle (LAP) and maximal dissipation principle (MDP), we first obtain the trajectory equation for the Wright-Fisher model. In turn, a convex-splitting technique is applied to construct a numerical scheme that is unique solvable on a convex domain and keeps the property of energy decay in time. The numerical scheme can assure the conservation of the total probability, *i.e.*, a complete solution is obtained. Numerical examples demonstrate that we can get a complete solution and true probability of fixation. In comparison with the FVM schemes in [6, 36], the new method has a significant advantage on the approximation to the delta singularity. Over an equidistant mesh with step size h, standard finite difference methods or FVMs only present an approximation of scale $O(1/h)$ to delta singularity, while the scheme here may give an approximation of scale $O(1/\varepsilon)$ with small positive ε close to the machine precision.

The paper is organized as follows. The details of EnVarA for Wright-Fisher model are shown in Section 2. In Section 3, the numerical scheme is constructed. Then numerical examples are presented in Section 4.

2. Variational approach for the Wright Fisher model

The primary goal of this section is to derive the constitutive relation of the Wright Fisher model. We first introduce EnVarA briefly. The original work was given by Onsager [25], and then it was improved by Lord Rayleigh [27]. This method has been applied to many physical and biological problems in recent years, for instance [8, 10, 35]. In the Wright-Fisher model, $x \in [0,1]$ and $f(x,t) \geq 0$ can be viewed as the position of particles and the density of x at time t, respectively. We first introduce the different coordinate systems.

Definition 2.1. Suppose that Ω_0^X, $\Omega_t^x \subset \mathbb{R}^m$, $m \in \mathbb{N}^+$, are domains with smooth boundary and time $t > 0$, and $\mathbf{u} = (u_1, \ldots, u_m)$ is a smooth vector field in $\mathbb{R}^m$. The flow map $x(X,t) : \Omega_0^X \to \Omega_t^x$ is defined as a solution of:

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} x(X,t) = \mathbf{u}(x(X,t),t), \quad t > 0, \\ x(X,0) = X, \end{cases} \tag{2.1}$$

where $X = (X_1, \ldots, X_m) \in \Omega_0^X$ and $x = (x_1, \ldots, x_m) \in \Omega_t^x$. In turn, the coordinate system X is called the Lagrangian coordinate and the coordinate system x is called Eulerian coordinate.

EnVarA is obtained by the combination of the statistical physics and nonlinear thermodynamics. First, we define total energy

$$E^{\text{total}} := \mathcal{K} + \mathcal{H},$$

where $\mathcal{K}$ is the kinetic energy and

$$\mathcal{H} := \mathcal{U} - T\mathcal{S}$$

is the Helmholtz free energy containing the internal energy $\mathcal{U}$, temperature T and entropy S. In an isothermal system without external force, the total energy dissipation law holds:

$$\frac{\mathrm{d}}{\mathrm{d}t} E^{\text{total}} = -\Delta,$$

where $\Delta \geq 0$ is the entropy product.

Subsequently, the least action principle (LAP) is applied: the trajectory of particles X from $x(X,0)$ at time $t = 0$ to $x(X,t^*)$ at a given time t^* in a Hamiltonian system are those which minimize the action functional defined by

$$\mathcal{A}(x(X,t)) := \int_0^{t^*} \mathcal{L}(x(X,t), x_t(X,t)) \mathrm{d}t,$$

where $\mathcal{L} := \mathcal{K} - \mathcal{H}$ is the Lagrangian functional of a conservative system and $x(X,t) \in \Omega_t^x$, $t > 0$. Moreover, in a non-Hamiltonian system here, taking variational of the action functional with respect to x, we get the conservation force

$$F_{\text{con}} = \frac{\delta \mathcal{A}}{\delta x}.$$

Next, we treat the dissipation part with maximum dissipation principle (MDP). Taking variational of Δ with respect to the velocity $\mathbf{u}$ involved in (2.1), we have the dissipative force

$$F_{\text{dis}} = \frac{\delta \frac{1}{2}\Delta}{\delta \mathbf{u}},$$

where the factor $\frac{1}{2}$ comes from a linear reponse assumption, *i.e.*, Δ is quadratic function of $\mathbf{u}$ and F_{dis} is linear in $\mathbf{u}$ [22]. According to the Newton's force balance law:

$$F_{\text{con}} = F_{\text{dis}},$$

we obtain constitutive relation. Onsager's approach [25,26] is the key point for such conclusions.

Now we revisit the Wright-Fisher model with a positive initial state in a context of EnVarA. By rescaling the time, (1.1) and (1.3) becomes:

$$\partial_t f + \partial_x (f\mathbf{u}) = 0, \tag{2.2}$$

$$f\mathbf{u} = -\partial_x \left(x(1-x)f\right), \tag{2.3}$$

$$f(x,0) = f_0(x) > 0,\ x \in [0,1], \tag{2.4}$$

$$\partial_x (x(1-x)f)\ |_{x=0,1} = 0,\ t > 0. \tag{2.5}$$

Lemma 2.2. *$f(x,t)$ is the solution of (2.2)–(2.5) if and only if f satisfies the corresponding energy dissipation law*

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 f \ln(x(1-x)f)\mathrm{d}x = -\int_0^1 \frac{f}{x(1-x)}|\boldsymbol{u}|^2 \mathrm{d}x. \tag{2.6}$$

Proof. We first prove that the energy dissipation law (2.6) holds if f is the solution of (2.2)–(2.5). Multiplying by $1 + \ln(x(1-x)f)$ and integrating on both sides of (2.2), we get

$$\int_0^1 \left(1 + \ln\left(x(1-x)f\right)\right)\partial_t f \mathrm{d}x = \int_0^1 \left(1 + \ln\left(x(1-x)f\right)\right)\partial_{xx}(x(1-x)f)\mathrm{d}x.$$

By integration by parts, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int_0^1 f \ln\left(x(1-x)f\right)\mathrm{d}x &= -\int_0^1 \frac{\partial}{\partial x}(x(1-x)f)\frac{\frac{\partial}{\partial x}(x(1-x)f)}{x(1-x)f}\mathrm{d}x \\ &= -\int_0^1 \frac{f}{x(1-x)}|\mathbf{u}|^2 \mathrm{d}x.\end{aligned}$$

Next we can derive (2.3) from the energy dissipation law (2.6) by EnVarA, while (2.2) is the conservation law which is assumed to be true.

Note that in Lagrangian coordinate, there exists an explicit formula for the solution of the conservation law (2.2),

$$f(x(X,t),t) = \frac{f_0(X)}{\frac{\partial x(X,t)}{\partial X}}, \tag{2.8}$$

where $f_0(X)$ is the initial function and $\frac{\partial x(X,t)}{\partial X}$ is *deformation gradient*, which is the Jacobian matrix of the map: $X \to x(X,t)$.

- The *total energy* of the Wright-Fisher model is given by

$$E^{\text{total}} = \mathcal{H} = \int_0^1 f \ln(x(1-x)f)\mathrm{d}x. \tag{2.9}$$

- LAP step. With (2.8), the action functional in Lagrangian coordinate becomes

$$\mathcal{A}(x) = \int_0^{t^*} (-\mathcal{H})\mathrm{d}t = -\int_0^{t^*}\int_0^1 f_0(X)\ln\left(x(1-x)\frac{f_0(X)}{\frac{\partial x(X,t)}{\partial X}}\right)\mathrm{d}X\mathrm{d}t,$$

where $t^* > 0$ is a given terminal time. Thus for any test function $y(X,t) = \widetilde{y}(x(X,t),t) \in C_0^\infty((0,1)\times(0,t^*))$ and $\varepsilon \in \mathbb{R}$, taking the variational of $\mathcal{A}(x)$ with respect to x, we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0}\mathcal{A}(x+\varepsilon y) &= -\int_0^{t^*}\int_0^1\left(f_0(X)\frac{1-2x}{x(1-x)} + \frac{\partial}{\partial X}\left(\frac{f_0(X)}{\frac{\partial x}{\partial X}}\right)\right)y\mathrm{d}X\mathrm{d}t \\ &= -\int_0^{t^*}\int_0^1\left(f\frac{1-2x}{x(1-x)} + \frac{\partial f}{\partial x}\right)\widetilde{y}\mathrm{d}x\mathrm{d}t.\end{aligned}$$

Then we obtain the conservation force

$$F_{\mathrm{con}} = \frac{\delta\mathcal{A}}{\delta x} = -\left(f\frac{1-2x}{x(1-x)} + \frac{\partial f}{\partial x}\right) = -\frac{1}{x(1-x)}\frac{\partial}{\partial x}\left(x(1-x)f\right),$$

in Eulerian coordinate, and

$$F_{\mathrm{con}} = -\left(f_0(X)\frac{1-2x}{x(1-x)} + \frac{\partial}{\partial X}\left(\frac{f_0(X)}{\frac{\partial x}{\partial X}}\right)\right),$$

in Lagrangian coordinate.

- MDP step. Let the entropy production $\Delta = \int_0^1 \frac{f}{x(1-x)}|\mathbf{u}|^2\mathrm{d}x$. Taking the variational of $\frac{1}{2}\Delta$ with respect to $\mathbf{u}$, we have the dissipation force

$$F_{\mathrm{dis}} = \frac{\delta\frac{1}{2}\Delta}{\delta\mathbf{u}} = \frac{f}{x(1-x)}\mathbf{u},$$

in Eulerian coordinate, and

$$F_{\mathrm{dis}} = \frac{\delta\frac{1}{2}\Delta}{\delta x_t} = \frac{f_0(X)}{x(1-x)}x_t,$$

in Lagrangian coordinate.

- Force balance step. We have, in Lagrangian coordinate, that

$$\frac{f_0(X)}{x(1-x)}x_t = -\frac{\partial}{\partial X}\left(\frac{f_0(X)}{\frac{\partial x}{\partial X}}\right) - f_0(X)\frac{1-2x}{x(1-x)}, \tag{2.11}$$

and in Eularian coordinate, we have

$$\frac{f(x,t)}{x(1-x)}\mathbf{u} = -\frac{1}{x(1-x)}\frac{\partial}{\partial x}\left(x(1-x)f(x,t)\right), \tag{2.12}$$

which is exactly (2.3).

□

Remark 2.3. There is an assumption that the initial state is positive in the above lemma. Otherwise, if $f_0(X) = 0$ for some $X \in (0,1)$, the argument above would be not valid any more. For example, in (2.11), the velocity x_t could be indefinite for points such that $f_0(X) = 0$. Note that in the real model, the initial state (1.3) is $f_0 = \delta(x - x_0)$, almost zero everywhere. To deal with this case, we consider two models with positive initial states $f_{0,1}, f_{0,2}$ such that $f_0 = f_{0,1} - f_{0,2}$ and correspondingly we have $f = f_1(x,t) - f_2(x,t)$.

Remark 2.4. What we really get by EnVarA is (2.11), which contains all the physics involved in this model. If we can solve (2.11) to get the trajectory $x(X,t)$, substituting it into (2.8), we obtain the solution $f(x,t)$ to (2.2)–(2.5). So in the following sections, we focus on numerical solution to (2.11).

To this purpose, we should first settle the initial and boundary condition for (2.11). From (2.5) and (2.3), we have $x_t(0,t) = x_t(1,t) = 0$, for $t > 0$. That means that a Dirichlet boundary condition should be subject to as $x(0,t) = 0, x(1,t) = 1$, for $t > 0$. So the trajectory problem is

$$\begin{cases} \frac{f_0(X)}{x(1-x)}\partial_t x = -\frac{\partial}{\partial X}\left(\frac{f_0(X)}{\frac{\partial x}{\partial X}}\right) - f_0(X)\frac{1-2x}{x(1-x)}, & X \in (0,1), \quad t > 0, \\ x(X,0) = X, & X \in [0,1], \\ x(0,t) = 0, x(1,t) = 1 & t > 0. \end{cases} \tag{2.13}$$

3. Numerical methods for trajectory equation

In this section, we consider numerical methods for (2.13).

3.1. A semi-discrete scheme in time and optimal transport

System (2.13) can be viewed as a gradient flow associated with the total energy of

$$E^{\text{total}} = \int_0^1 f_0(X)\ln\left(\frac{f_0(X)}{\frac{\partial x}{\partial X}}\right)\mathrm{d}X + \int_0^1 f_0(X)\ln\left(x(1-x)\right)\mathrm{d}X, \tag{3.1}$$

which is just the counterpart in Lagrangian coordinate of total energy (2.9) of the system (2.2)–(2.5) and can be split into convex and concave parts, that is $E^{\text{total}} = E_c - E_e$, where both E_c and E_e are convex. The canonical splitting is $E_c = \int_0^1 f_0(X)\ln\left(\frac{f_0(X)}{\frac{\partial x}{\partial X}}\right)\mathrm{d}X$ and $E_e = -\int_0^1 f_0(X)\ln\left(x(1-x)\right)\mathrm{d}X$. The convex splitting was first exploited by D. J. Eyre in [11] to craft energy stable numerical schemes for the Allen-Cahn and Cahn-Hilliard equations. The basic idea is to treat the convex part implicitly while to treat the concave part explicitly. Then a *semi-discrete scheme* for (2.13) is proposed as follows

$$\frac{f_0(X)}{x^n(1-x^n)}\frac{x^{n+1}-x^n}{\tau} = -\frac{\partial}{\partial X}\left(\frac{f_0(X)}{\frac{\partial x^{n+1}}{\partial X}}\right) - f_0(X)\frac{1-2x^n}{x^n(1-x^n)}, \tag{3.2}$$

where τ is the time step and $x^n = x(X,t^n)$ is the solution at time $t^n = n\tau$, $n \in \mathbb{N}^+$.

Remark 3.1. Equation (3.2) is also a Variational Particle Scheme. We explain the fact in the framework of optimal transport theory. Let $\Omega = [0,1]$. We denote by $\mathscr{P}(\Omega)$ the space of $\mathfrak{L}^1$ measure on Ω, non-negative functions with unit integral and finite second moments, where $\mathfrak{L}^1$ is the Lebesgue measure. $f^n \in \mathscr{P}(\Omega)$ is the approximation to solution of equation (2.2) and (2.3) at time $t_n = n\tau$, $n \in \mathbb{N}$. We fix a reference density f^0 and consider a time-dependent family of transport maps $x(\cdot,t^n) : [0,1] \to [0,1]$ such that $x(\cdot,t^n)\#f^0 = f^n\mathfrak{L}^1$ for all $n \in \mathbb{N}^+$, where $\#$ denotes the push-forward of measures.

Then the map from x^n to x^{n+1} is an optimal transport in the sense that x^{n+1} is the minimizer of the cost functional

$$F(x) := \int_0^1 \frac{1}{2\tau}\frac{f_0(X)}{x^n(1-x^n)}|x-x^n|^2 + f_0(X)\ln\left(\frac{f_0(X)}{\frac{\partial x}{\partial X}}\right) + f_0(X)\frac{1-2x^n}{x^n(1-x^n)}x\mathrm{d}X.$$

Some relevant descriptions on optimal transport can be found in Westdickenberg-Wilkening's work [33].

Remark 3.2. There exists another interesting approach based on a diffeomorphism mapping the uniform density to the unknown density [12], or on the pseudo-inverse of the cumulative distribution. Carrillo *et al.* [4, 5] developed some numerical schemes to solve for the diffeomorphism maps, instead of solving for Lagrangian mapping directly in (3.2). For 1D problems, the diffeomorphism is equivalent to the pseudo-inverse of the cumulative distribution. Based on this inverse, some numerical methods were constructed by Gosse and Toscani [15, 16] and Blanchet *et al.* [2]. In these methods, the diffeomorphism or the inverse play a role similar as the Lagrangian map in some sense.

3.2. The fully discrete scheme

We begin with the definition of inner-product, difference operators and summation-by-parts in one dimension. Let $h = \frac{1}{N}$, $N \in \mathbb{N}^+$ be the spatial step. Denote by $X_r = X(r) = rh$, where r takes on integer and half integer values. Let $\mathcal{E}_N$ and $\mathcal{C}_N$ be the spaces of functions whose domains are $\{X_i \mid i = 0, \ldots, N\}$ and $\{X_{i-\frac{1}{2}} \mid i = 1, \ldots, N\}$ respectively. In component form, these functions are identified *via* $l_i = l(X_i)$, $i = 0, \ldots, N$, for $l \in \mathcal{E}_N$, and $\phi_{i-\frac{1}{2}} = \phi(X_{i-\frac{1}{2}})$, $i = 1, \ldots, N$, for $\phi \in \mathcal{C}_N$.

Let l, $g \in \mathcal{E}_N$ and ϕ, $\psi \in \mathcal{C}_N$. We define the "inner-product" on space $\mathcal{E}_N$ and $\mathcal{C}_N$ respectively as

$$[l|g] = h \sum_{i=1}^{N-1} l_i g_i, \tag{3.3}$$

$$(\phi|\psi) = h \sum_{i=1}^{N} \phi_{i-\frac{1}{2}} \psi_{i-\frac{1}{2}}. \tag{3.4}$$

The difference operator $D_h : \mathcal{E}_N \to \mathcal{C}_N$ and $d_h : \mathcal{C}_N \to \mathcal{E}_N$, and the average operator $A : \mathcal{E}_N \to \mathcal{C}_N$ can be defined as respectively as

$$(D_h l)_{i-\frac{1}{2}} = (l_i - l_{i-1})/h, \qquad i = 1, \ldots, N, \tag{3.5}$$

$$(d_h \phi)_i = (\phi_{i+\frac{1}{2}} - \phi_{i-\frac{1}{2}})/h, \quad i = 1, \ldots, N-1, \tag{3.6}$$

$$(Al)_{i-\frac{1}{2}} = (l_i + l_{i-1})/2, \qquad i = 1, \ldots, N. \tag{3.7}$$

Then we have the following result of summation-by-parts.

Lemma 3.3. *Let $\phi \in \mathcal{C}_N$ and $l \in \mathcal{E}_N$. Then $(D_h l|\phi) = -[l|d_h\phi] + l_N \phi_{N-\frac{1}{2}} - l_0 \phi_{\frac{1}{2}}$.*

Let $\mathbf{Q} := \{l \in \mathcal{E}_N \mid l_{i-1} < l_i,\ 1 \le i \le N;\ l_0 = 0,\ l_N = 1\}$ and its boundary set $\partial\mathbf{Q} := \{l \in \mathcal{E}_N \mid l_{i-1} \le l_i,\ 1 \le i \le N,\ \textit{and}\ l_i = l_{i-1},\ \textit{for some}\ 1 \le i \le N; l_0 = 0,\ l_N = 1\}$. Then $\bar{\mathbf{Q}} := \mathbf{Q} \cup \partial\mathbf{Q}$ is a closed convex set.

The *fully discrete scheme* is formulated as follows: Given $x^n \in \mathbf{Q}$, find $x^{n+1} = (x_0^{n+1}, \ldots, x_N^{n+1}) \in \mathbf{Q}$ such that

$$\frac{f_0(X_i)}{x_i^n(1-x_i^n)} \frac{x_i^{n+1} - x_i^n}{\tau} = -d_h \left(\frac{A f_0(X)}{D_h x^{n+1}} \right)_i - f_0(X_i) \frac{1 - 2x_i^n}{x_i^n(1-x_i^n)}, \ 1 \le i \le N-1. \tag{3.8}$$

Equation (3.8) is still a nonlinear system. Newton's iteration method can be applied to solve it. To make sure the convergence, we choose the damped Newton's iteration method [3].

Damped Newton's iteration. Set $x^{n+1,0} = x^n$. For $k = 0, 1, 2, \ldots, x^{n+1,k+1} = x^{n+1,k} + \omega(\lambda)\delta_x$ such that

$$\begin{aligned} \frac{f_0(X_i)}{x_i^n(1-x_i^n)} \frac{\delta_{x_i}}{\tau} - d_h \left(\frac{A f_0(X)}{(D_h x^{n+1,k})^2} D_h \delta_{x_i} \right)_i &= -\frac{f_0(X_i)}{x_i^n(1-x_i^n)} \frac{x_i^{n+1,k} - x_i^n}{\tau} \\ &\quad - d_h \left(\frac{A f_0(X)}{D_h x^{n+1,k}} \right)_i - f_0(X_i) \frac{1 - 2x_i^n}{x_i^n(1-x_i^n)}, \quad 1 \le i \le N-1. \end{aligned} \tag{3.9}$$

and

$$\omega(\lambda) = \begin{cases} \frac{1}{\lambda} & \lambda > \tilde{\lambda} \\ \frac{1-\lambda}{\lambda(3-\lambda)} & \tilde{\lambda} \geq \lambda \geq \lambda^* \\ 1 & \lambda < \lambda^*, \end{cases} \tag{3.10}$$

where $\lambda(J, x^{n+1,k}) = \left(\frac{1}{a}(DJ(x^{n+1,k}))^T[D^2J(x^{n+1,k})]^{-1}DJ(x^{n+1,k})\right)^{\frac{1}{2}}$ with $J(x)$ defined in (3.21), $a = h\min_i f_0(X_i)$, $\lambda^* = 2-\sqrt{3}$, and $\tilde{\lambda} \in [\lambda^*, 1)$.

After solving (3.8), we finally get the numerical distribution $f(x^{n+1}, t^{n+1})$ from (2.8) as

$$f_i^{n+1} = \frac{f_0(X_i)}{(x_{i+1}^{n+1} - x_{i-1}^{n+1})/(2h)}, \ 1 \leq i \leq N-1, \text{ and} \tag{3.11}$$

$$f_0^{n+1} = \frac{f_0(X_0)}{(x_1^{n+1} - x_0^{n+1})/h}, \ f_N^{n+1} = \frac{f_0(X_N)}{(x_N^{n+1} - x_{N-1}^{n+1})/h}. \tag{3.12}$$

Lemma 3.4. *The density function f^{n+1} obtained from (3.11) and (3.12) keeps the conservation law of mass.*

In fact, if we define the initial mass carried by each particle $x_i^0 = X_i$ as

$$m_i^0 = hf_0(X_i), \ 1 < i < N; \ \ m_0^0 = \frac{h}{2}f_0(X_0); \ \ m_N^0 = \frac{h}{2}f_0(X_N), \tag{3.13}$$

and define the mass carried by particle x_i^n as

$$m_i^n = \frac{x_{i+1}^n - x_{i-1}^n}{2}f_i^n, \ 1 < i < N; \ \ m_0^n = \frac{x_1^n - x_0^n}{2}f_0^n; \ \ m_N^n = \frac{x_N^n - x_{N-1}^n}{2}f_N^n, \tag{3.14}$$

then we readily have from (3.11) and (3.12) that

$$m_i^n \equiv m_i^0, \ \ 0 \leq i \leq N, \ n = 1, 2, \ldots.$$

Remark 3.5. $x_i(t) = x(X_i, t)$, $0 < i < N$ are the trajectories starting from the particles X_i at time $t = 0$. From the governing equation (2.13) or (3.8), the motion of these particles is primarily determined by the second term on the right hand side since this term tends to infinity when the particle approaches to the end points $x = 0, 1$. In particular, this term tends to negative infinity around the left end $x = 0$, while the limit becomes positive infinity around the right end $x = 1$. Therefore, $x_1(t)$ and $x_{N-1}(t)$ will be closer and closer to $x_0(t) \equiv 0$ and $x_N(t) \equiv 1$, respectively.

Governed by the continuous model (2.13), the particles may touch the end points, which means that the Dirac delta singularity occurs for $f(x, t)$ from (2.8). For the discrete model (3.8), we find solution $x^{n+1} \in \mathbf{Q}$, where $x_i < x_{i+1}$ for $0 \leq i < N$. As a result, theoretically x_1 and x_{N-1} would never touch the ends. However, in the practical computations, when x_1^n and $x_0^n = 0$ are too close to distinguish from each other under the machine precision, they are bundled up and will be regarded as *one particle* which carries the mass from the original two and will be fixed at the boundary. This is the signal that the numerical Dirac delta (*i.e.*, the *fixation*) happens. In comparison with the FVMs in [6], we can now approximate the delta singularity to the scale of $1/\varepsilon$, with ε close to the machine precision, while by the standard FVMs on equidistance mesh, one can only approximate the delta singularity to the scale of $1/h$ (with the spatial mesh size h).

Criteria for particles meet the boundary. Though we can choose the machine precision as a criterion to judge whether two particles touch each other, it is not practical. For example, in (3.12), when $x_1^{n+1} - x_0^{n+1}$ is close to machine precision, we will lose all the accuracy of f_0^{n+1}. So we will choose a criterion with $\varepsilon_0 = 10^{-10}$ in double precision system as:

$$\text{Criteria: } \begin{cases} \text{If } x_i^{n+1} \in \mathbf{B}_l = [0, \varepsilon_0], & \text{it will be fixed at left boundary for ever,} \\ \text{If } x_i^{n+1} \in \mathbf{B}_r = [1-\varepsilon_0, 1], & \text{it will be fixed at right boundary for ever.} \end{cases} \tag{3.15}$$

Equivalently, we have a rearrangement on the position of the particles as

$$x_i^{n+1} = \begin{cases} 0, & \text{if } x_i^{n+1} \in \mathbf{B}_l, \\ 1, & \text{if } x_i^{n+1} \in \mathbf{B}_r, \\ x_i^{n+1}, & \text{otherwise.} \end{cases} \tag{3.16}$$

At the next time step, we only need to determine the position of particles from $(\varepsilon_0, 1-\varepsilon_0)$.

With the above rearrangement, the formulas (3.12) for the density function at the boundary points don't work any more. To define the revised formulas, we need to count the total number of particles accumulated at the boundary points. Let

$$\begin{cases} \text{starting point} & i_s^{n+1} = \max\{\ i\ |x_i^{n+1} \in \mathbf{B}_l,\ 0 \le i < N\}, \\ \text{ending point} & i_e^{n+1} = \min\{\ i\ |x_i^{n+1} \in \mathbf{B}_r,\ 0 < i \le N\}. \end{cases} \tag{3.17}$$

If $i_s^{n+1} > 0$ or $i_e^{n+1} < N$, there must be some particles which touched the boundary points at time t^{n+1}. Then the revised formula for the density function $f^{n+1} = (f_{i_s^{n+1}}^{n+1}, f_{i_s^{n+1}+1}^{n+1}, \ldots, f_{i_e^{n+1}}^{n+1})$ become

$$f_i^{n+1} = \frac{f_0(X_i)}{(x_{i+1}^{n+1} - x_{i-1}^{n+1})/(2h)} = \frac{m_i^0}{(x_{i+1}^{n+1} - x_{i-1}^{n+1})/2},\ i_s^{n+1} < i < i_e^{n+1}, \tag{3.18}$$

$$f_i^{n+1} = \frac{2}{\varepsilon_0}\sum_{k=0}^{i-1} m_k^0 + \frac{m_i^0}{(x_{i+1} - x_i)/2},\ \text{for } i = i_s^{n+1}, \text{and} \tag{3.19}$$

$$f_i^{n+1} = \frac{2}{\varepsilon_0}\sum_{k=i+1}^{N} m_k^0 + \frac{m_i^0}{(x_i - x_{i-1})/2},\ \text{for } i = i_e^{n+1}. \tag{3.20}$$

Remark 3.6. The treatment in (3.18)–(3.20) keeps the conservation law of total mass naturally and means that only the last fixed particle can feel the free nearest particle inside and the effect of all former fixed particles is confined to the ε_0 neighbor of boundary points.

Combining all the discussions above together, we can now present the final algorithm as follows.

Algorithm 3.7.

- Initialization.
  For $0 \le i \le N$, we get the initial particle position $x_i^0 = X_i$, the initial density distribution $f_i^0 = f_0(X_i)$, and the initial mass m_i^0 by (3.13).
  Set starting point $i_s = 0$ and ending point $i_e = N$.
- Time Stepping.
  For $n = 0, 1, 2, \ldots$, find the density distribution at next time step $f^{n+1} = (f_{i_s}^{n+1}, f_{i_s+1}^{n+1}, \ldots, f_{i_e}^{n+1})$ by the following procedures.
  (1) Obtain the position of particles x_i^{n+1}, $i_s \le i \le i_e$, *via* solving the fully discrete system (3.8) by Newton's iteration (3.9), with $x_{i_s}^{n+1} = 0,\ x_{i_e}^{n+1} = 1$.
  (2) Check whether a particle meets the boundary by the criteria (3.15), re-arrange the position by (3.16) and update the starting point i_s and the ending point i_e by (3.17) if necessary.
  (3) Obtain the density distribution f^{n+1} by (3.18)–(3.20).

3.3. Unique solvability and energy decay of fully discrete scheme

In this subsection, we provide some analyses on the unique solvability and energy decay of the fully discrete scheme (3.8), and the convergence of the Newton method (3.9) with (3.10).

Theorem 3.8. *The numerical scheme* (3.8) *is unique solvable in* $\boldsymbol{Q}$.

Proof. We first consider the following optimization problem:

$$\min_{y\in\bar{\mathbf{Q}}} J(y) := \frac{1}{2\tau}\left[\frac{f_0(X)}{x^n(1-x^n)}(y-x^n)\Big|(y-x^n)\right] + \left(Af_0(X)\Big|\ln\left(\frac{Af_0(X)}{D_h y}\right)\right) + \left[f_0(X)\frac{1-2x^n}{x^n(1-x^n)}\Big|y\right], \tag{3.21}$$

where $f_0(X) \in \mathcal{E}_N$ is the initial distribution and $x^n \in \mathbf{Q}$ is the known position of particles at time t^n. It is easy to verify that $J(y)$ is a convex function on the closed convex set $\bar{\mathbf{Q}}$. Hence there exists a unique minimizer $x \in \bar{\mathbf{Q}}$. We must have the minimizer $x \in \mathbf{Q}$ since if $y \in \partial\mathbf{Q}$, then there exists some $i > 0$ such that $(D_h y)_{i-1/2} = (y_i - y_{i-1})/h = 0$, and $J(y) = +\infty$.

We first claim that $x \in \mathbf{Q}$ is the minimizer of $J(y)$ if and only if it is a solution of scheme (3.8). Hence the fully discrete scheme (3.8) has a unique solution.

In fact, if $x \in \mathbf{Q}$ is the minimizer of $J(y)$, then for $\forall y \in \bar{\mathbf{Q}}$, there exists a sufficiently small $\varepsilon_0 > 0$, such that for any $\varepsilon \in (-\varepsilon_0, \varepsilon_0)$, $x + \varepsilon(y - x) \in \mathbf{Q}$ since $\mathbf{Q}$ is a open set. Then $j(\varepsilon) = J(x + \varepsilon(y - x))$ achieves its minimal at $\varepsilon = 0$. So we have $j'(0) = 0$ and using summation by parts, we obtain

$$\frac{1}{\tau}\left[\frac{f_0(X)}{x^n(1-x^n)}(x-x^n)\Big|y-x\right] + \left[d_h\left(\frac{Af_0(X)}{D_h x}\right)\Big|y-x\right] + \left[f_0(X)\frac{1-2x^n}{x^n(1-x^n)}\Big|y-x\right] = 0,$$

for any $y \in \bar{\mathbf{Q}}$. This implies that $x \in \mathbf{Q}$ satisfies (3.8).

Conversely, let $x \in \mathbf{Q}$ be the solution to scheme (3.8). We need to prove that x is the minimizer of $J(y)$ on $\bar{\mathbf{Q}}$.

For any $y \in \partial\mathbf{Q}$, $J(y) = +\infty$. We always have $J(y) \geq J(x)$. Then for any $y \in \mathbf{Q}$, taking the inner product of (3.8) with $y - x$ and using summation by parts, we get

$$\frac{1}{\tau}\left[\frac{f_0(X)}{x^n(1-x^n)}(x-x^n)\Big|y-x\right] - \left(\frac{Af_0(X)}{D_h x}\Big|D_h(y-x)\right) + \left[f_0(X)\frac{1-2x^n}{x^n(1-x^n)}\Big|y-x\right] = 0. \tag{3.22}$$

After direct calculation, we see that, for any $y \in \mathbf{Q}$

$$\begin{aligned} J(y) &= J(x + (y-x)) \\ &= J(x) + \frac{1}{2\tau}\left[\frac{f_0(X)}{x^n(1-x^n)}(y-x)\Big|(y-x)\right] + \frac{1}{\tau}\left[\frac{f_0(X)}{x^n(1-x^n)}(x-x^n)\Big|y-x\right] \\ &\quad + \left(Af_0(X)\Big|\ln\left(\frac{D_h x}{D_h y}\right)\right) + \left[f_0(X)\frac{1-2x^n}{x^n(1-x^n)}\Big|y-x\right] \\ &\geq J(x), \end{aligned} \tag{3.23}$$

where the last inequality is obtained from (3.22) and the fact $\ln(p) \geq 1 - \frac{1}{p}$, for $p > 0$, which leads to

$$\left(Af_0(X)\Big|\ln\left(\frac{D_h x}{D_h y}\right)\right) \geq \left(Af_0(X)\Big|1 - \frac{D_h y}{D_h x}\right) = -\left(\frac{Af_0(X)}{D_h x}\Big|D_h(y-x)\right).$$

The proof is finished. □

We define the discrete total energy $E_N : \mathbf{Q} \to \mathbb{R}$ of (2.9) as

$$E_N(x) := \left(Af_0(X)\Big|\ln\left(\frac{Af_0(X)}{D_h x}\right)\right) + \left[f_0(X)\Big|\ln(x(1-x))\right] \equiv E_{N,c}(x) - E_{N,e}(x),$$

where $E_{N,c}(x)$ and $E_{N,e}(x)$ are both convex and their first order variations are

$$\delta_x E_{N,c}(x) = d_h\left(\frac{Af_0(X)}{D_h x}\right),\ \delta_x E_{N,e}(x) = -f_0(X)\frac{1-2x}{x(1-x)}. \tag{3.24}$$

Theorem 3.9. *Suppose $x^n = (x_0^n, \dots, x_N^n) \in \boldsymbol{Q}$ be the solution to scheme (3.8) at time t_n. Then the discrete energy dissipation law holds, i.e.,*

$$\frac{E_N(x^{n+1}) - E_N(x^n)}{\Delta t} \le -\left[\frac{f_0(X)}{x^n(1-x^n)}\frac{x^{n+1}-x^n}{\Delta t}\middle|\frac{x^{n+1}-x^n}{\Delta t}\right],\quad n = 0, 1, \dots.$$

This is the discrete counterpart of the dissipation law in Lemma 2.2.

Proof. Thanks to the convexity of $E_{N,c}^n$ and $E_{N,e}^n$, we have

$$\begin{aligned} E_{N,c}(x^n) - E_{N,c}(x^{n+1}) &\ge \left[\delta_x E_{N,c}(x^{n+1})\middle| x^n - x^{n+1}\right], \\ E_{N,e}(x^{n+1}) - E_{N,e}(x^n) &\ge \left[\delta_x E_{N,e}(x^n)\middle| x^{n+1} - x^n\right]. \end{aligned}$$

Then from (3.24) and (3.8),

$$\begin{aligned} E_N(x^{n+1}) - E_N(x^n) &= (E_{N,c}(x^{n+1}) - E_{N,e}(x^{n+1})) - (E_{N,c}(x^n) - E_{N,e}(x^n)) \\ &\le \left[\delta_x E_{N,c}(x^{n+1}) - \delta_x E_{N,e}(x^n)\middle| x^{n+1} - x^n\right] \\ &= -\left[\frac{f_0(X)}{x^n(1-x^n)}\frac{x^{n+1}-x^n}{\Delta t}\middle| x^{n+1} - x^n\right] \le 0. \end{aligned}$$

Then the proof is completed. □

Hence the numerical scheme (3.8) for $x \in \mathbf{Q}$ is uniquely solvable. And regardless of time step, the energy decays in time: $E_N(x^{n+1}) \le E_N(x^n)$.

Before we analyse the convergence of damped Newton's iteration (3.9), the definition of *self-concordant* should be involved.

Definition 3.10 ([24])**.** Let $\mathcal{G}$ be a finite-dimensional real vector space, $\mathcal{Q}$ be an open nonempty convex subset of $\mathcal{G}$, $\Lambda : \mathcal{Q} \to \mathbb{R}$ be a function, $a > 0$. Λ is called self-concordant on $\mathcal{Q}$ with the parameter value a, if $\Lambda \in C^3$ is a convex function on $\mathcal{Q}$, and, for all $x \in \mathcal{Q}$ and all $u \in \mathcal{G}$, the following inequality holds:

$$|D^3\Lambda(x)[u,u,u]| \le 2a^{-1/2}(D^2\Lambda(x)[u,u])^{3/2}$$

($D^k\Lambda(x)[u_1, \dots, u_k]$ henceforth denotes the kth directional derivative of Λ taken at x along the collection of directions $u_1, \dots, u_k \in \mathcal{G}$).

Theorem 3.11. *Suppose $f_0(X) \in \mathcal{E}_N$ is the initial distribution with a positive lower bound for $X \in Q$, then $J(y)$, defined in (3.21), is a self-concordant function and the damped Newton's iteration (3.9) and (3.10) is convergent in Q.*

Proof. Let $C_0 := \min_{X \in Q} f_0(X) > 0$ and $J(y) := J_1(y) + J_2(y) + J_3(y)$ with

$$\begin{aligned} J_1(y) &:= \frac{1}{2\tau}\left[\frac{f_0(X)}{x^n(1-x^n)}(y - x^n)\middle|(y - x^n)\right], \\ J_2(y) &:= \left(Af_0(X)\middle|\ln\left(\frac{Af_0(X)}{D_h y}\right)\right), \\ J_3(y) &:= \left[f_0(X)\frac{1-2x^n}{x^n(1-x^n)}\middle| y\right]. \end{aligned}$$

Since linear and quadratic functions have zero third derivative, $J_1(y)$ and $J_3(y)$ are self-concordant for all $y \in Q$. We just need to prove $J_2(y)$ is a self-concordant function in Q.

Based on the Definition (3.10), a function $J_2 : Q \to \mathbb{R}$ is self-concordant if it is self concordant along every line in its domain, *i.e.*, $\tilde{J}_2(\xi) = J_2(y + \xi u)$ is a self-concordant function of $\xi \in \mathbb{R}$ for all $y \in Q$ and for all $u \in Q$ [3].

Combining with the definition of "inner-product" (3.4), we have

$$\tilde{J}_2(\xi) = J_2(y + \xi u) = h \sum_{i=1}^{N} b_i \ln \left(\frac{h b_i}{y_i + \xi u_i - y_{i-1} - \xi u_{i-1}} \right), \tag{3.26}$$

and

$$\tilde{J}_2''(\xi) = h \sum_{i=1}^{N} \frac{b_i (u_i - u_{i-1})^2}{(y_i + \xi u_i - y_{i-1} - \xi u_{i-1})^2}, \tag{3.27}$$

and

$$\tilde{J}_3'''(\xi) = -2h \sum_{i=1}^{N} \frac{b_i (u_i - u_{i-1})^3}{(y_i + \xi u_i - y_{i-1} - \xi u_{i-1})^3}, \tag{3.28}$$

where $b_i = (A f_0(X))_{i-\frac{1}{2}}$, $i = 1, \ldots, N$. By Cauchy inequality, we have for any $w_i \in \mathbb{R}$,

$$\left| \sum_{i=1}^{N} w_i^3 \right| \leq \left(\sum_{i=1}^{N} w_i^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{N} w_i^4 \right)^{\frac{1}{2}} \leq \left(\sum_{i=1}^{N} w_i^2 \right)^{\frac{3}{2}}.$$

Then, we have

$$\begin{aligned} \left| \sum_{i=1}^{N} \frac{h b_i (u_i - u_{i-1})^3}{(y_i + \xi u_i - y_{i-1} - \xi u_{i-1})^3} \right| &\leq \left(\sum_{i=1}^{N} \frac{(h b_i)^{\frac{2}{3}} (u_i - u_{i-1})^2}{(y_i + \xi u_i - y_{i-1} - \xi u_{i-1})^2} \right)^{\frac{3}{2}} \\ &\leq \frac{1}{\sqrt{a}} \left(\sum_{i=1}^{N} \frac{h b_i (u_i - u_{i-1})^2}{(y_i + \xi u_i - y_{i-1} - \xi u_{i-1})^2} \right)^{\frac{3}{2}} \end{aligned} \tag{3.29}$$

where $a = hC_0$. So $J(y)$ is self-concordant for $y \in Q$.

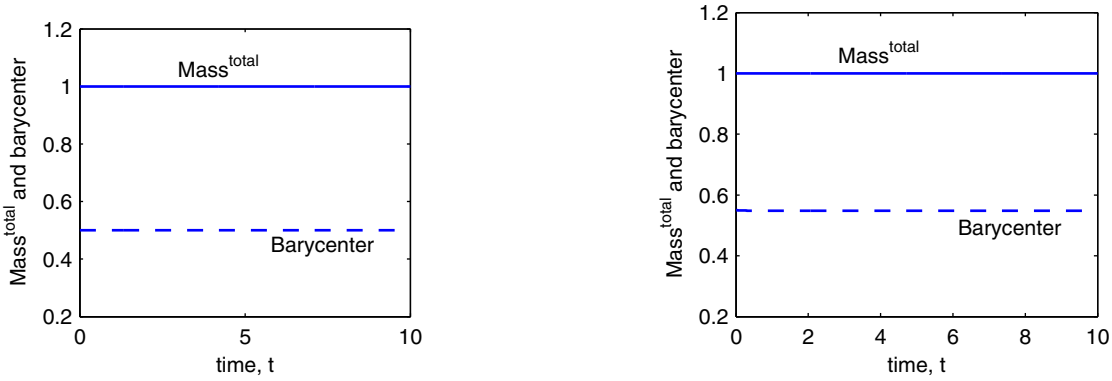
Based on Theorem 2.2.3 in [24], the self-concordant property is sufficient for the convergence of the damped Newton's iteration (3.9) and (3.10). □

4. Numerical results

4.1. Numerical results for positive initial functions

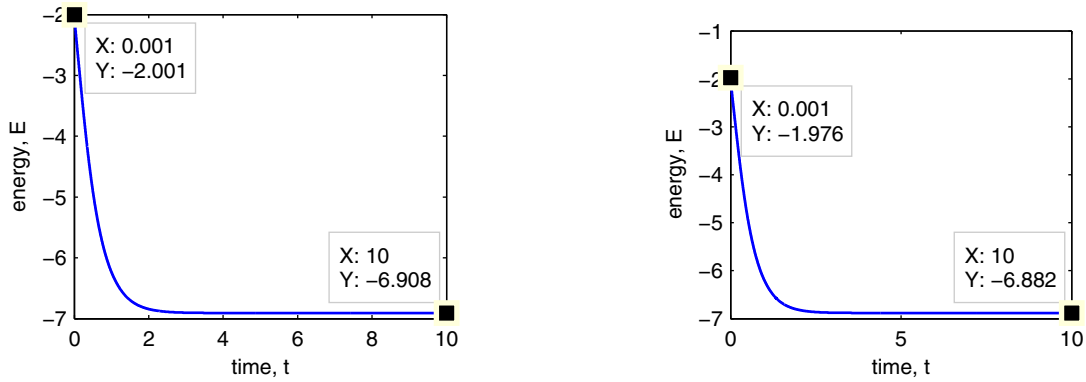
In this subsection, we present some numerical results for equation (2.2)–(2.5) with positive initial functions by Algorithm 3.7. We take $f_0^1(x) = 1$, $f_0^2(x) = \frac{1}{5}(2 + 6x + \frac{\pi}{2}\sin(2\pi x))$ as examples and choose the space mesh size $h = 1/1000$, time step size $\tau = 1/1000$ under a criterion $\varepsilon_0 = 10^{-10}$. Also note that, although the total mass of the system is equal to unity, it is not the total probability since the initial function is not in the probability measure. At the same time, the first moment (the mean) stands for barycenter instead of expectation.

Figure 1 shows that the total mass is unity all the time and the mean value keeps the conservation for both the positive initial functions. Figure 2 shows the total energy of the two systems decay as time evolves. The solutions of the two initial functions at time $t = 0.002$, $t = 0.01$ and the steady state $t = 10$ are shown in Figures 3 and 4, respectively: singularities develop at two boundaries and the heights are dependent on the mean of initial state. Figure 5 shows the motion of particles which is influenced by the initial state. After certain



(a) $f_0^1(x) = 1$ (b) $f_0^2(x) = \frac{1}{5}(2 + 6x + \frac{\pi}{2}\sin(2\pi x))$

FIGURE 1. Total Mass ($\text{Mass}^{\text{total}}$) and Barycenter of positive initial functions over time with $h = 1/1000$, $\tau = 1/1000$.

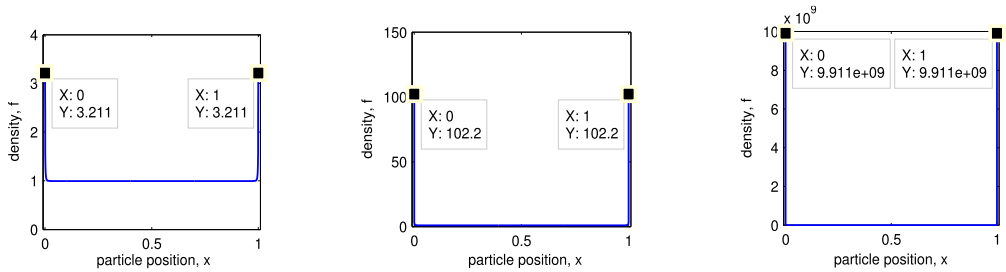


(a) $f_0^1(x)$ (b) $f_0^2(x)$

FIGURE 2. Energy of positive initial functions over time with $h = 1/1000$, $\tau = 1/1000$.

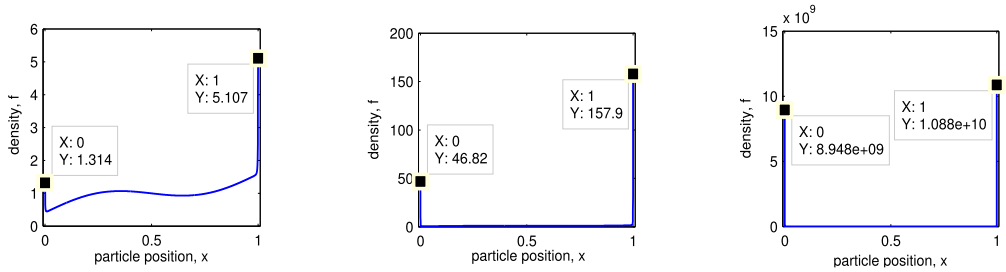
time, almost all particles stay at the two boundaries, which causes *fixation phenomenon*. This result means that we obtain the numerical complete solution, with the numerical scheme (3.8) satisfying energy decay over time. Moreover, we can approximate the delta singularity to the scale of $1e + 10$.

Table 1 presents the total mass (M^{total}), barycenter (Barycenter), the density and the mass at the two boundary points (f_l, f_r, M_l, M_r) of the two initial functions with different grid size ($h = 1/100$, $\tau = 1/100$; $h = 1/1000$, $\tau = 1/1000$; $h = 1/10\,000$, $\tau = 1/10\,000$) at time $t = 10$. It shows that the total mass keeps unity regardless of the grid size, and the barycenter approximates to its own initial mean at the level of the grid size. It also shows that delta singularities at boundaries can be simulated at the level of $1e+10$ regardless of the grid size and the values are influenced by the initial expectation. Moreover, the sum of M_l and M_r is approximate to unity, which verifies the development of Dirac delta functions.



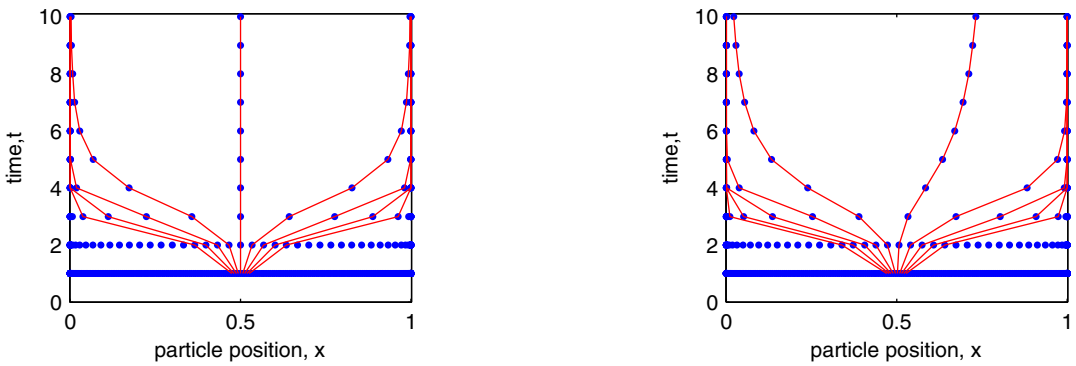
(a) $t = 0.002$ for $f_0^1(x)$ (b) $t = 0.01$ for $f_0^1(x)$ (c) $t = 10$ for $f_0^1(x)$

FIGURE 3. Density over time for $f_0^1(x)$ with $h = 1/1000$, $\tau = 1/1000$.



(a) $t = 0.002$ for $f_0^2(x)$ (b) $t = 0.01$ for $f_0^2(x)$ (c) $t = 10$ for $f_0^2(x)$

FIGURE 4. Density over time for $f_0^2(x)$ with $h = 1/1000$, $\tau = 1/1000$.



(a) x for $f_0^1(x)$ (b) x for $f_0^2(x)$

FIGURE 5. Particle position, x, over time for $f_0^1(x)$ and $f_0^2(x)$ with $h = 1/1000$, $\tau = 1/1000$.

4.2. Numerical results for pure drift

In this section, we focus on $f_0(x) = \delta(x - x_0)$ $(0 < x_0 < 1)$ and use normal distribution $N(x_0, \sigma^2)$ $(\sigma = 0.01)$ to approximate $\delta(x - x_0)$. Based on Remark 2.3, we split the problem (2.2)–(2.5) into two positive initial value

TABLE 1. Results for positive initial functions f_0^1, f_0^2 at time $t = 10$ with different grid sizes.

h	τ	M^{total}	Barycenter	f_l	f_r	M_l	M_r
$f_0^1 = 1$							
1/100	1/100	1.0000	0.5000	8.2235e+09	8.2235e+09	0.4150	0.4150
1/1000	1/1000	1.0000	0.5000	9.9105e+09	9.9105e+09	0.4965	0.4965
1/10 000	1/10 000	1.0000	0.5000	9.9930e+09	9.9930e+09	0.4998	0.4998
$f_0^2 = \frac{1}{5}(2 + 6x + \frac{\pi}{2}\sin(2\pi x))$							
1/100	1/100	1.0000	0.5316	7.4881e+09	8.9220e+09	0.3834	0.4489
1/1000	1/1000	1.0000	0.5483	8.9477e+09	1.0879e+10	0.4475	0.5445
1/10 000	1/10 000	1.0000	0.5498	8.9952e+09	1.0989e+10	0.4499	0.5496

Notes. M^{total} denote by Total Mass. M_l and M_r are the mass at left and right boundaries, respectively.

problems:

$$\begin{cases} g_t = \partial_{xx}(x(1-x)g), & x \in (0,1), & t > 0, \\ g(x,0) = 10, & x \in [0,1], & \\ \partial_x(x(1-x)g)\mid_{x=0} = 0, & \partial_x(x(1-x)g)\mid_{x=1} = 0, & t > 0, \end{cases} \tag{4.1}$$

$$\begin{cases} w_t = \partial_{xx}(x(1-x)w), & x \in (0,1), & t > 0, \\ w(x,0) = 10 + N(x_0, \sigma^2), & x \in [0,1], & \\ \partial_x(x(1-x)w)\mid_{x=0} = 0, & \partial_x(x(1-x)w)\mid_{x=1} = 0, & t > 0. \end{cases} \tag{4.2}$$

Then we have the solution $f = w - g$. Because of this fact, we first obtain the numerical solutions $G(x^n, t^n)$ and $W(y^n, t^n)$ of two problems (4.1) and (4.2) by Algorithm 3.7, respectively, where x^n and y^n are the particle positions at time t^n. We cannot take the difference between $G(x^n, t^n)$ and $W(y^n, t^n)$ directly since x^n and y^n may be different. We need to get the value of G at y^n by the mass-conserved interpolation.

The details of the mass-conserved interpolation are shown as follows:

Algorithm 4.1 (Mass-conserved interpolation)**.**

- Input: the particle positions $x = (x_0, x_1, \ldots, x_N)$ and $y = (y_0, y_1, \ldots, y_N)$; Starting point i_s and ending point i_e of free particles in x; Starting point j_s and ending point j_e of free particles in y; Mass $m_x = (m_{x_0}, m_{x_1}, \ldots, m_{x_N})$ for each particle of x.
  Output: $m_y = (m_{y_0}, m_{y_1}, \ldots, m_{y_N})$, the re-assigned mass carried by particles y; $G(y) = (G_{j_s}, \ldots, G_{j_e})$, the value of G at y.
- Re-assign the mass from particles x to y.
  (1) Define the mean mass density function $\bar{m}(s), s \in [0,1]$. Let $\Delta x_i = x_{i+1} - x_i$.

$$\begin{aligned} \bar{m}(s) &= \tfrac{m_{x_i}}{(x_{i+1}-x_{i-1})/2}, & \text{for } s \in \left(x_i - \tfrac{\Delta x_{i-1}}{2},\ x_i + \tfrac{\Delta x_i}{2}\right), & \quad i_s < i < i_e; \\ \bar{m}(s) &= \tfrac{m_{x_i}}{(x_{i+1}-x_i)/2}, & \text{for } s \in \left(x_i,\ x_i + \tfrac{\Delta x_i}{2}\right), & \quad i = i_s; \\ \bar{m}(s) &= \tfrac{m_{x_i}}{(x_i-x_{i-1})/2}, & \text{for } s \in \left(x_i - \tfrac{\Delta x_{i-1}}{2},\ x_i\right), & \quad i = i_e. \end{aligned}$$

  Note that $x_{i_s} = 0$ and $x_{i_e} = 1$.
  (2) Collect mass for particles at $y = (y_0, y_1, \ldots, y_N)$. Let $\Delta y_j = y_{j+1} - y_j$.
  For free particles,

$$\begin{aligned} m_{y_j} &= \int_{y_j - \frac{\Delta y_{j-1}}{2}}^{y_j + \frac{\Delta y_j}{2}} \bar{m}(s)\mathrm{d}s, & j_s < j < j_e; \\ m_{y_j} &= \int_{y_j}^{y_j + \frac{\Delta y_j}{2}} \bar{m}(s)\mathrm{d}s, & j = j_s; \\ m_{y_j} &= \int_{y_j - \frac{\Delta y_{j-1}}{2}}^{y_j} \bar{m}(s)\mathrm{d}s, & j = j_e. \end{aligned}$$

TABLE 2. The comparision of numerical results with FVM in grid size $h = 1/10\,000$, $\tau = 1/10\,000$ for $x_0 = 0.4$ at $t = 10$.

	FVM		Varitional Particle Scheme (3.8)	
Time	f_l	f_r	f_l	f_r
$t = 1.0000$	1.0039e+04	6.0629e+03	9.2680e+09	5.9800e+09
$t = 2.0000$	1.1736e+04	7.7362e+03	1.1680e+10	7.7400e+09
$t = 3.0000$	1.1964e+04	7.9643e+03	1.1930e+10	7.7983e+09
$t = 4.0000$	1.1995e+04	7.9952e+03	1.1980e+10	8.0200e+09
$t = 5.0000$	1.1999e+04	7.9993e+03	1.1980e+10	8.0238e+09

For particles accumulated at left end,

$$m_{y_0} = \ldots = m_{y_{j_s-1}} = \sum_{i=1}^{i_s-1} m_{x_i}/(i_s - 1).$$

For particles accumulated at right end,

$$m_{y_{j_e+1}} = \ldots = m_{y_N} = \sum_{i=i_e+1}^{N} m_{x_i}/(N - i_e).$$

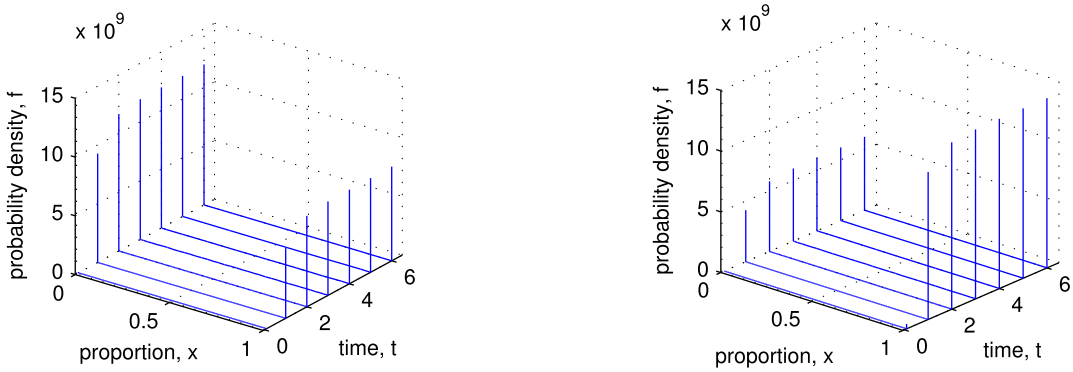
- Recover $G(y) = (G_{j_s}, \ldots, G_{j_e})$ from $m_{y_0}, \ldots, m_{y_N}$ by the same rules as in (3.18)–(3.20).

Remark 4.2. When the initial state has a compact support, the particle may have an infinite speed. To avoid this difficulty, here we make use of the linearity to split the original problem into two problems with positive initial data. There exist some other approach to handle this issue. For a method based on the diffeomorphism [5], the authors suggested to regularize the initial state. Similar idea may be used to our problem: just add a small positive term to the initial state. However, the velocity is almost singular and not easy to handle numerically.

On the other hand, when we consider some nonlinear problems, such as porous media equation [9], where the interface of the initial support has a finite propagation speed and even badly, the waiting time phenomenon may occur, the regularization of the initial state may destroy the waiting time phenomenon. Our method can treat this kind of problem naturally: just solve for the trajectories start from the initial support [9].

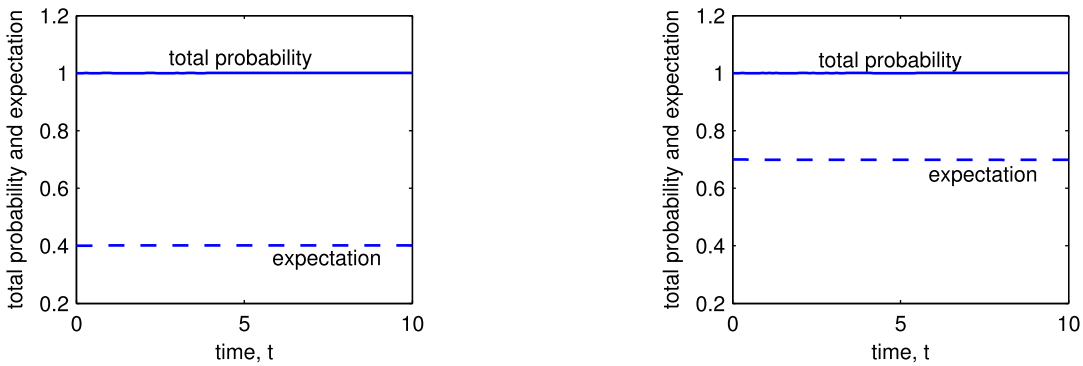
Then we simulate pure drift (2.2)–(2.5) for $x_0 = 0.4$ and $x_0 = 0.7$ with $\varepsilon_0 = 10^{-10}$ and the step size $h = 1/10\,000$, $\tau = 1/10\,000$. Figure 6 shows the evolution of distribution of probability: the density almost vanishes in $(0, 1)$, and singularities develop at the boundary points. Moreover, the values of singularities depend on their initial states. As shown in Figure 7, their total probabilities are equal to unity and expectations keep the conservation based on their own initial expectations. This means that the numerical solution is a complete solution. Figure 8 also shows the behavior of probabilities at two boundaries as time evolves: the value increases to a state where the sum of both is close to unity. That causes the development of Dirac delta singularities.

Table 2 presents the comparison of the density at two boundary points (f_l, f_r) with scheme (3) in [6], which is a FVM scheme with central difference method. For $x_0 = 0.4$ and a fixed grid size $h = 1/10\,000$, $\tau = 1/10\,000$ with $\varepsilon_0 = 10^{-10}$, it shows that f_l, f_r obtained by scheme (3) is at the level of $1e + 04$, while that scale becomes $1e+10$ by scheme (3.8) in the present paper. This fact indicates that, the numerical solution obtained by scheme (3.8) is an approximation of scale $O(1/\varepsilon_0)$ to the delta singularity, with a small positive $\varepsilon_0 > 0$ close to the machine precision.



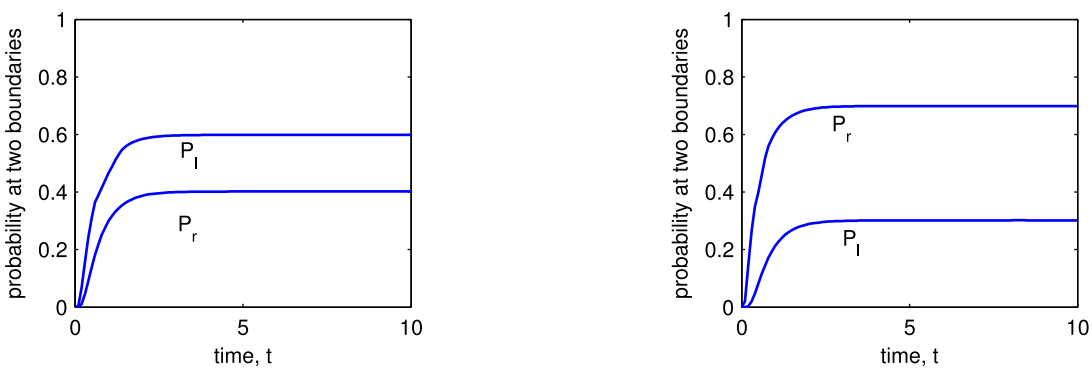
(a) $x_0 = 0.4$ (b) $x_0 = 0.7$

FIGURE 6. Distribution of probability for pure drift over time with $h = 1/10\,000$, $\tau = 1/10\,000$.



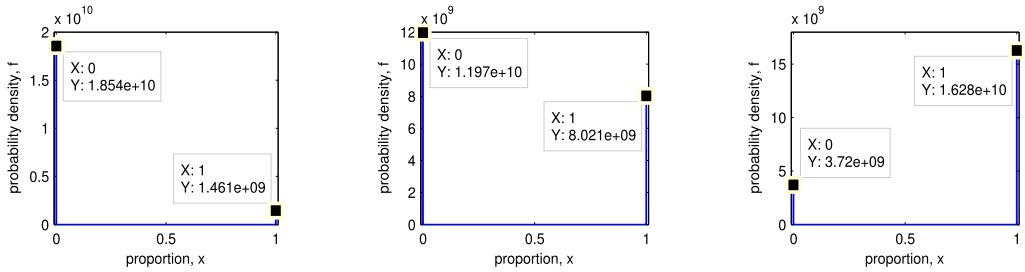
(a) TP and Exp for $x_0 = 0.4$ (b) TP and Exp for $x_0 = 0.7$

FIGURE 7. Total probability (TP) and expectation (Exp) for the pure drift as time evolves with $h = 1/10\,000$, $\tau = 1/10\,000$.



(a) P_l, P_r for $x_0 = 0.4$ (b) P_l, P_r for $x_0 = 0.7$

FIGURE 8. Probability at two boundaries as time evolves with $h = 1/10\,000$, $\tau = 1/10\,000$ (P_l and P_r denote the fixation probability at left boundary and right boundary, respectively).



(a) $s = -0.0001$ (b) $s = 0.0000$ (c) $s = 0.0001$

FIGURE 9. Distribution of probability influenced by s for $x_0 = 0.4$ at $t = 10$ with $h = 1/10\,000$ and $\tau = 1/10\,000$.

4.3. Numerical results for semi-selection case

In this part, we consider the semi-selection case with $M(x) = sx(1-x)$ (s is the strength of semi-dominant selection) in a population with the fixed size $N_e = 10\,000$. By rescaling the time, we have the following initial-boundary value problem:

$$\begin{cases} \partial_t f(x,t) = \frac{\partial^2}{\partial x^2}(x(1-x)f(x,t)) - \frac{\partial}{\partial x}(4N_e M(x) f(x,t)), & x \in (0,1), \quad t > 0, \\ f(x,0) = f_0(x), & x \in [0,1], \\ \{\partial_x(x(1-x)f(x,t)) - 4N_e M(x) f(x,t)\} \mid_{x=0,1} = 0, & t > 0, \end{cases} \tag{4.3}$$

and the corresponding energy dissipation law is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_0^1 f \ln(x(1-x)f) - 4N_e sxf \mathrm{d}x\right) = -\int_0^1 \frac{f}{x(1-x)}|\mathbf{u}|^2 \mathrm{d}x,$$

where $\mathbf{u} := -\frac{\partial_x(x(1-x)f)}{f} + 4N_e sx(1-x)$. Based on Energetic Variational Approach, Problem (4.3) is transformed into

$$\begin{cases} \frac{f_0(X)}{x(1-x)} x_t = 4sN_e f_0(X) - \left(\frac{\partial}{\partial X}\left(\frac{f_0(X)}{\frac{\partial x}{\partial X}}\right) + f_0(X)\frac{1-2x}{x(1-x)}\right), & X \in (0,1), \quad t > 0, \\ x(X,0) = X, & X \in [0,1], \\ x(0,t) = 0, x(1,t) = 1 & t > 0, \end{cases} \tag{4.4}$$

in the Lagrangian coordinate. Furthermore, the distribution of probability $\{f(x_i^n, t^n)\}_{i=0}^N$ $(n > 0)$ can be also calculated by (3.18)–(3.20).

Figure 9 shows the probability distribution of the steady state at $t = 10$ with $s = -0.0001$, $s = 0.0000$, and $s = 0.0001$ for initial state $f_0 = \delta(x - x_0)$ and $x_0 = 0.4$. It shows that semi-selection with $s = -0.0001$ prefers alleles a, while it is more willing to favor alleles A if $s = 0.0001$. Figure 10 implies that the total probabilities always keep unity whatever the value of s is, while the expectation does not keep conservative any more. Figure 11 shows how the expectations are associated with the values of s when $x_0 = 0.4$ at time $t = 10$. It also shows that the expectation is a good approximation of the true probability of ultimate fixation P_{fix} given by

$$P_{fix}(x_0) = \frac{1 - \mathrm{e}^{-4x_0 sN_e}}{1 - \mathrm{e}^{-4sN_e}}. \tag{4.5}$$

This means that we also get a complete solution for the semi-selection case.

5. CONCLUSION

In this paper, we simulate the Wright-Fisher model for pure drift and semi-selection. We first obtain the trajectory equation of the model based on EnVarA and then get the numerical scheme by the convex splitting

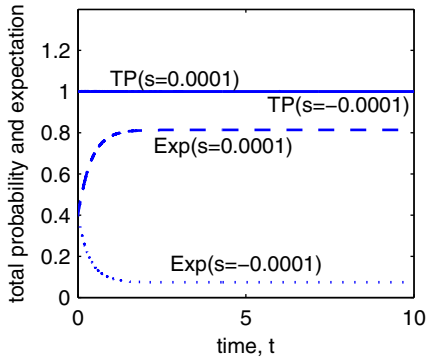


FIGURE 10. Total probability (TP) and expectation (Exp) over time for $x_0 = 0.4$ under $s = 0.0001$ and $s = -0.0001$ with $h = 1/10\,000$ and $\tau = 1/10\,000$.

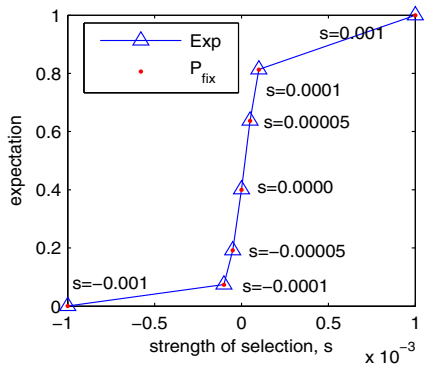


FIGURE 11. Expectation at time $t = 10$ under different s for $x_0 = 0.4$ at $t = 10$ with $h = 1/10\,000$ and $\tau = 1/10\,000$; P_{fix} is given by (4.5).

technique. The scheme is uniquely solvable and satisfies energy decay on a convex set where the position of particles is strictly increasing. Then we obtain the numerical complete solutions and true probability of fixation. Moreover, at any equidistant grid, Dirac delta singularities can be measured of scale close to machine precision.

Acknowledgements. It is grateful to Prof. Xinfu Chen for helpful discussions. This work is supported in part by NSF of China under the grants 11271281. Chun Liu is supported by NSF-DMS 1759535 and NSF-DMS 1759536. And Cheng Wang is supported by NSF grants DMS-1418689.

REFERENCES

[1] R. Barakat and D. Wagener, Solutions of the forward diallelic diffusion equation in population genetics. *Math. Biosci.* **41** (1978) 65–79.

[2] A. Blanchet, V. Calvez and J.A. Carrillo, Convergence of the mass-transport steepest descent scheme for the subcritical Patlak-Keller-Segel model. *SIAM J. Numer. Anal.* **46** (2008) 691–721.

[3] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge Univ. Press, Cambridge (2004).

[4] J.A. Carrillo and J.S. Moll, Numerical simulation of diffusive and aggregation phenomena in nonlinear continuity equations by evolving diffeomorphisms. *SIAM J. Sci. Comput.* **31** (2009) 4305–4329.

[5] J.A. Carrillo, H. Ranetbauer and M.-T. Wolfram, Numerical simulation of nonlinear continuity equations by evolving diffeomorphisms. *J. Comput. Phys.* **327** (2016) 186–202.

[6] M. Chen, C. Liu, S. Xu, X. Yue and R. Zhang, Behavior of different numerical schemes for population genetic drift problems. Preprint arXiv:1410.5527 (2018).

[7] J.F. Crow and M. Kimura, An introduction to population genetics theory. *Population (French Edition)* **26** (1971) 977–978.

[8] Q. Du, C. Liu, R. Ryham and X. Wang, Energetic variational approaches in modeling vesicle and fluid interactions. *Phys. D* **238** (2009) 923–930.

[9] C. Duan, C. Liu, C. Wang and X. Yue, Numerical methods for porous medium equation by an energetic variational approach. Preprint arXiv:1806.10775 (2018).

[10] B. Eisenberg, Y.K. Hyon and C. Liu, Energy variational analysis of ions in water and channels: Field theory for primitive models of complex ionic fluids. *J. Chem. Phys.* **133** (2010) 104104.

[11] D.J. Eyre, Unconditionally gradient stable time marching the Cahn-Hilliard equation. *MRS Proceedings*. Cambridge Univ. Press **529** (1998) 39.

[12] L. Evans, O. Savin and W. Gangbo, Diffeomorphisms and nonlinear heat fows. *SIAM J. Math. Anal.* **37** (2005) 737–751.

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

[14] R.A. Fisher, The Genetical Theory of Natural Selection. Clarendon Press, Oxford (1930).

[15] L. Gosse and G. Toscani, Lagrangian numerical approximations to one-dimensional convolution-diffusion equations. *SIAM J. Sci. Comput.* **28** (2006) 1203–1227.

[16] L. Gosse and G. Toscani, Identification of asymptotic decay to self-similarity for one-dimensional filtration equations. *SIAM J. Numer. Anal.* **43** (2006) 2590–2606.

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

[18] M. Kimura, Random genetic drift in multi-allelic locus. *Evolution* **9** (1955) 419-435.

[19] M. Kimura, On the probability of fixation of mutant genes in a population. *Genetics* **47** (1962) 713.

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

[21] M. Kimura, The Neutral Theory of Molecular Evolution. Cambridge Univ. Press, Cambridge (1983).

[22] R. Kubo, Thermodynamics: An Advanced Course with Problems and Solutions. North-Holland Pub. Co., Amsterdam (1976).

[23] A.J. McKane and D. Waxman, Sigular solution of the diffusion equation of population genetics. *J. Theor. Biol.* **247** (2007) 849–858.

[24] Y. Nesterov and A. Nemirovskii, Interior-point Polynomial Algorithms in Convex Programming. SIAM, Pjiladelphia, PA **13** (1994).

[25] L. Onsager, Reciprocal relations in irreversible processes. II. *Phys. Rev.* **38** (1931) 2265–2279.

[26] L. Onsager, Reciprocal relations in irreversible processes. I. *Phys. Rev.* **37** (1931) 405.

[27] J.W. Strutt, Some general theorems relating to vibrations. *Proc. Lond. Math. Soc.* **4** (1873) 357–368.

[28] T.D. Tran, J. Hofrichter and J. Jost, An introduction to the mathematical structure of the Wright Fisher model of population genetics. *Theory Biosci.* **132** (2013) 73-82.

[29] A. Traulsen, T. Lenaerts, J.M. Pacheco and D. Dingli, On the dynamics of neutral mutations in a mathematical model for a homogeneous stem cell population. *J. R. Soc. Interface* **10** (2013) 20120810.

[30] J.L. Vzquez, The Porous Medium Equation: Mathematical Theory. Oxford Univ. Press, Oxford (2007).

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

[32] D. Waxman, Fixation at a locus with multiple alleles: structure and solution of the Wright-Fisher model. *J. Theor. Biol.* **257** (2009) 245–251.

[33] M. Westdickenberg and J. Wilkening, Variational particle schemes for the porous medium equation and for the system of isentropic Euler equations. *ESAIM: M2AN* **44** (2010) 133–166.

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

[35] X.F. Yang, J.J. Feng, C. Liu and J. Shen, Numerical simulations of jet pinching-off and drop formation using an energetic variational phase-field method. *J. Comput. Phys.* **218** (2006) 417–428.

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