BIT Numerical Mathematics (2019) 59:797–821
https://doi.org/10.1007/s10543-019-00749-4

BIT

Behavior of different numerical schemes for random genetic drift

Shixin Xu[1] · Minxin Chen[1] · Chun Liu[2] · Ran Zhang[3] · Xingye Yue[1]

Received: 23 May 2018 / Accepted: 25 March 2019 / Published online: 6 April 2019

Abstract
In the problem of random genetic drift, the probability density of one gene is governed by a degenerated convection-dominated diffusion equation. Dirac singularities will always be developed at boundary points as time evolves, which is known as the fixation phenomenon in genetic evolution. Three finite volume methods: FVM1-3, one central difference method: FDM1 and three finite element methods: FEM1-3 are considered. These methods lead to different equilibrium states after a long time. It is shown that only schemes FVM3 and FEM3, which are the same, preserve probability, expectation and positiveness and predict the correct probability of fixation. FVM1-2 wrongly predict the probability of fixation due to their intrinsic viscosity, even though they are unconditionally stable. Contrarily, FDM1 and FEM1-2 introduce different anti-diffusion terms, which make them unstable and fail to preserve positiveness.

Keywords Random genetic drift · Degenerate equation · Conservations of probability and expectation · Finite volume method · Finite difference method · Finite element method · Numerical viscosity and numerical anti-diffusion

Mathematics Subject Classification 35K65 · 65M06 · 92D10

1 Introduction

The number of a particular gene (allele) of one locus in a population varies randomly from generation to generation. This process is a kind of stochastic process named random genetic drift, which was first introduced by one of the founders in the field of population genetics, Wright [20]. Genetic drift plays an important role in molecular

Communicated by Elisabeth Larsson.

Partially supported by NSF of China Grants 11271281, 11301368 and 91230106 and by NSF Grants DMS-1159937, DMS-1216938 and DMS-1109107.

Extended author information available on the last page of the article

evolution [1,11]. Mathematical descriptions of genetic drift are typically built upon the Wright–Fisher model [5,21] or its diffusion limit [7,9,14]. The Wright–Fisher model describes the dynamics of a gene with two alleles, A or B in a diploid population with fixed size of N, i.e., total $2N$ alleles. The model is formulated as a discrete-time Markov chain. Let random variable X_k denote the fraction of A in k'th generation, then $1 - X_k$ is the fraction of B. At the next generation, if the effects of mutation, migration and selection are negligible, then X_{k+1} obeys

$$X_{k+1} \sim \frac{B(2N, X_k)}{2N}, \tag{1.1}$$

where $B(n, p)$ is the binomial distribution with n trails and p probability of success. The first and second conditional moments of the Wright–Fisher process satisfy [2]

$$\begin{aligned} E[X_{k+1}|X_k] &= X_k, \\ Var[X_{k+1}|X_k] &= \frac{1}{2N} X_k(1 - X_k). \end{aligned}$$

The first condition expresses the unbiased nature of neutrality. In absence of mutation, migration and selection, the *fixation* phenomenon [9] will occur, which means that only allele A is left and all copies of allele B are lost from the population, vice versa. And the allele should be fixed with probability equal to its initial frequency.

After the diffusion approximation [5,21] is introduced to model the genetic drift, Moran [14] and Kimura [7] substantially extended and developed this approach. If we describe the process by X_t, the fraction of type A gene at time t, and $f(x, t)$ denotes the probability density of $X_t = x$ at time t, Kimura [7,10] showed that $f(x, t)$ obeys the following diffusion equation,

$$\frac{\partial f(x,t)}{\partial t} + \frac{\partial j(x,t)}{\partial x} = 0, \quad x \in (0, 1), \tag{1.2}$$

where the quantity $j(x, t)$ is the current that characterizes the flow of probability density, with the form as $j(x, t) = -\frac{1}{4N} \frac{\partial}{\partial x}(x(1 - x) f(x, t))$. By rescaling of the time $t \to \frac{t}{4N}$, $j(x, t)$ has the form as

$$j(x,t) = -\frac{\partial}{\partial x}\left(x(1-x) f(x,t)\right). \tag{1.3}$$

To keep the total probability, the zero current boundary conditions [13] are imposed as

$$j(0, t) = 0, \quad j(1, t) = 0. \tag{1.4}$$

For problem (1.2)–(1.4), Tran et al. [18] proved the existence and uniqueness of a weak solution in the sense of distribution. A distribution $f(\cdot, t) \in H$ is called as a weak solution of problem (1.2)–(1.4) with a initial state $f(x, 0) = f_0(x) \in H$, if

$$\int_0^1 f(x,t)\phi(x,t)dx = \int_0^1 f_0(x)\phi(x,0)dx$$
$$+\int_0^t \int_0^1 f(\frac{\partial\phi}{\partial t} + x(1-x)\frac{\partial^2\phi}{\partial x^2})dtdx, \quad \forall\phi \in C^\infty([0,1]\times[0,\infty)), \quad (1.5)$$

where $H = \{f : [0,1] \to [0,\infty] | \int_0^1 fgdx < \infty, \ \forall g \in C^\infty[0,1]\}$ is the set of all distribution functions on $[0,1]$.

In definition (1.5), if we set $\phi = 1$, we can find that problem (1.2)–(1.4) conserve the total probability, for any time $t > 0$,

$$\int_0^1 f(x,t)\,\mathrm{d}x = \int_0^1 f(x,0)\,\mathrm{d}x = 1. \quad (1.6)$$

Furthermore, if we chose $\phi = x$ in (1.5), then we get the conservation of the mean gene frequency (expectation), i.e.,

$$\int_0^1 xf(x,t)\,\mathrm{d}x \equiv \int_0^1 xf(x,0)\,\mathrm{d}x. \quad (1.7)$$

McKane and Waxman [13] gave a closed form solution to the above problem (1.2)

$$f(x,t) = \Pi_0(t)\delta(x) + \Pi_1(t)\delta(1-x) + f_r(x,t), \quad (1.8)$$

where $\Pi_0(t)$, $\Pi_1(t)$ and $f_r(x,t)$ are smooth functions and $f_r(x,t)$ is in a form of infinite series. (1.8) means the solution has three parts: two singular part $\Pi_0(t)\delta(x)$, $\Pi_1(t)\delta(1-x)$ and a smooth part f_r. If we consider the problem (1.2)–(1.4) subject to the condition that the population has an initial A-gene frequency of $p \in (0,1)$, so $f(x,0)$ should be a Dirac delta function at $x = p$. i.e.,

$$f(x,0) = \delta(x-p). \quad (1.9)$$

In this case, by the conservation of total probability (1.6) and the conservation of expectation (1.7), Mckane and Waxman proved that

$$\lim_{t\to\infty} f(x,t) = (1-p)\delta(x) + p\delta(1-x), \quad (1.10)$$

in the sense of distribution. This is the *fixation phenomenon* corresponding to the Wright–Fisher model. However, it is not easy to compute the infinite series in (1.8). For more complex situations, such as the effects of mutation, migration and selection are involved, one can hardly derive the closed form solutions as above. Thus direct numerical solution is necessary for this problem.

A numerical scheme should be designed to find a *complete solution* [22], which preserves the conservations of total probability (1.6) and expectation (1.7). Among various numerical methods, we first choose finite volume method (FVM) [4,15] since it is easy to preserve the conservation laws numerically. Equation (1.2) looks like a

pure diffusion equation; actually it is a convection-dominated diffusion equation. If we rewrite (1.2) as

$$\frac{\partial f}{\partial t} - \frac{\partial}{\partial x}\left(x(1-x)\frac{\partial f}{\partial x}\right) + \frac{\partial}{\partial x}((2x-1)f) = 0, \tag{1.11}$$

then it is clear that (1.11) is convection-dominated near boundaries $x = 0$ and $x = 1$, where convection velocity is up to 1 while diffusion coefficient is degenerate to 0. The convection dominance will induce strongly irregular solution, for example, large jump and discontinuity.

Remark 1.1 At the boundary points, (1.11) is degenerated to a pure convection problem. Whether or not a boundary condition should be imposed is the key point to obtain a well-posed solution. For problem (1.11), the situation is that both the left and right boundaries are out-flow ones. That means we can get a unique regular solution without any boundary condition. But this solution is not a complete one: it does not maintain conservation of probability (1.6). So when the no-flux boundary condition (1.4) is imposed, we can not expect to obtain a regular solution. See the discussions on the boundary condition in the appendix of [22].

In general, a upwind finite volume method (referred to FVM1) is a better choice for the convection-dominated diffusion problem to achieve stability due to its intrinsic numerical viscosity. While it does produce a stable numerical solution, we always get the same equilibrium-state, no matter what the initial state is. This is obviously wrong for genetic drift (see (1.10)) because it can not keep the conservation of the expectation (1.7).

Then we turn to the central scheme of FVM. There are two choices: to discretize the fluxes induced by the diffusion, $-\frac{\partial}{\partial x}(x(1-x)\frac{\partial f}{\partial x})$, and convection, $\frac{\partial}{\partial x}((2x-1)f)$, respectively (referred to FVM2); or to discretize the flux $j(x,t) = -\frac{\partial}{\partial x}(x(1-x)f)$ in (1.3) as a whole (referred to FVM3). We find that both FVM2 and FVM3 are unconditionally stable, which is a surprising since we are solving a convection-dominated problem by central schemes. We also observe that the equilibrium solution of FVM2 is always the same one no matter what the initial state is, just as FVM1, but it takes a much longer time to achieve the equilibrium state. FVM3 is the simplest method but could yield a *complete solution* and predict the correct fixation probability. Dirac singularities develop at both boundary points with proper weights rather than for FVM2, same weight develop at both ends.

A careful analysis shows that FVM3 is unconditionally stable and its solution always converges to the true solution and FVM2 is equivalent to FVM3 plus a 2nd order viscosity term $-\frac{h^2}{4}f_{xx}$. That is the reason why FVM2 is also unconditionally stable and takes a much longer time to reach the equilibrium state than FVM1, which is equivalent to FVM2 plus a much larger first-order $O(h)$ viscosity term.

By the method of vanishing viscosity, i.e., a small viscosity term is first added in, then the limit behavior of the solutions is considered when the added viscosity tends to zero, we see that the limit of the equilibrium state solution is uniquely determined and has nothing to do with the initial states. This means that the long time behavior of

the original problem will be changed by any added infinitesimal viscosity. This is the reason why the central scheme FVM2 does not work for the genetic drift problem. The upwind scheme FVM1 introduces a more complicated viscosity. Further investigation shows that the viscosity acts as an artificial two-way mutation near two boundary points, which also leads to a unique artificial equilibrium state.

However, the fact that a central scheme is unconditionally stable for a convection-dominated problem beyond the common understanding on the constrain of Peclet's number, which is necessary for a central scheme to achieve stability. So we check a central finite difference method (referred to FDM1). For this scheme, we see that the constrain on Peclet's number does matter. Due to the degeneration, the Peclet's number can never be less than 1. Comparing with FVM3, FDM1 introduces an anti-diffusion term, which makes the scheme unstable and can not keep the positiveness of the solution. We also discuss three finite element methods (referred to FEM1-3) with or without numerical quadrature. We find that if the mass matrix is got by quadrature and stiffness matrix is got by exact integration (referred to FEM3), we exactly duplicate FVM3; if the mass matrix is got by exact integration (FEM1) or the stiffness matrix is got by quadrature (FEM2), anti-diffusion terms must be introduced, which make both schemes unstable and can not keep the positiveness of the solution.

Our study shows that for this kind of problems, numerical methods must be carefully chosen and any method with intrinsic numerical viscosity or contrarily with numerical anti-diffusion should be avoided.

This paper is organized as follows. We present three different FVMs for genetic drift problem in Sect. 2. Numerical results and analysis for FVMs are presented in Sect. 3. In Sect. 4, we discuss one central FDM and three FEMs. The final section gives some concluding remarks.

2 Numerical methods

In order to maintain the total probability, we start with a finite volume method (FVM) [4]. A uniform grid, with grid spacing $h = 1/M$ and grid points $x_i = ih$, $i = 0, \ldots, M$, is used to discretize the space domain $[0, 1]$. Likewise, the time domain is uniformly discretized with step size τ and the grid points are $t_n = n\tau$, $n = 0, 1, \ldots$. Let j_i^n and f_i^n be the numerical approximations of $j(x_i, t_n)$, $f(x_i, t_n)$, respectively. For inner mesh point x_i $(0 < i < M)$, the control volume is

$$\mathscr{D}_i = \left\{ x |\, x_{i-\frac{1}{2}} \le x \le x_{i+\frac{1}{2}} \right\},$$

where $x_{i+\frac{1}{2}} = (i + \frac{1}{2})h$. Given f^{n-1}, by FVM, we have, for $n = 1, 2, \ldots,$

$$\frac{f_i^n - f_i^{n-1}}{\tau} + \frac{j_{i+\frac{1}{2}}^n - j_{i-\frac{1}{2}}^n}{h} = 0. \tag{2.1}$$

For the boundary points $x_0 = 0$ and $x_M = 1$, the control volumes are

$$\mathscr{D}_0 = \left\{x|\ 0 \le x \le x_{\frac{1}{2}}\right\},\ \text{and}\ \mathscr{D}_M = \left\{x|\ x_{M-\frac{1}{2}} \le x \le 1\right\}.$$

By the boundary conditions $j(0,t) = j(1,t) = 0$, we have

$$\frac{f_0^n - f_0^{n-1}}{\tau} + \frac{j_{\frac{1}{2}}^n - 0}{h/2} = 0 \text{ and } \frac{f_M^n - f_M^{n-1}}{\tau} + \frac{0 - j_{M-\frac{1}{2}}^n}{h/2} = 0. \tag{2.2}$$

To get a fully discrete scheme, we still need to approximate the flux $j_{i+\frac{1}{2}}^n$, for $i = 0, \ldots, M-1$. It is treated differently by the following three schemes:

– FVM1: approximate $j(x,t) = -x(1-x)\frac{\partial f}{\partial x} + (2x-1)f$ at point $x_{i+\frac{1}{2}}$ by upwind scheme

$$j_{i+\frac{1}{2}}^n = \begin{cases} -x_{i+\frac{1}{2}}(1-x_{i+\frac{1}{2}})\frac{f_{i+1}^n - f_i^n}{h} + (2x_{i+\frac{1}{2}} - 1)f_{i+1}^n, & 2x_{i+\frac{1}{2}} - 1 < 0, \\ -x_{i+\frac{1}{2}}(1-x_{i+\frac{1}{2}})\frac{f_{i+1}^n - f_i^n}{h} + (2x_{i+\frac{1}{2}} - 1)f_i^n, & 2x_{i+\frac{1}{2}} - 1 > 0. \end{cases} \tag{2.3}$$

– FVM2: approximate $j(x,t) = -(x(1-x))\frac{\partial f}{\partial x} + (2x-1)f$ at point $x_{i+\frac{1}{2}}$ by central scheme

$$j_{i+\frac{1}{2}}^n = -x_{i+\frac{1}{2}}(1-x_{i+\frac{1}{2}})\frac{f_{i+1}^n - f_i^n}{h} + (2x_{i+\frac{1}{2}} - 1)\frac{f_{i+1}^n + f_i^n}{2}. \tag{2.4}$$

– FVM3: approximate $j(x,t) = -\frac{\partial}{\partial x}(x(1-x)f(x,t))$ at point $x_{i+\frac{1}{2}}$ by central scheme

$$j_{i+\frac{1}{2}}^n = -\frac{x_{i+1}(1-x_{i+1})f_{i+1}^n - x_i(1-x_i)f_i^n}{h}. \tag{2.5}$$

FVM3 was recently used in [22] for a numerical investigation on the random genetic problems, where some applications can be found on more complicated topics such as time-dependent probability of fixation for a neutral locus or in the presence of selection effect within a population of constant size; probability of fixation in the presence of selection and demographic change. In this paper, we confine ourselves to the simplest case to see the behaviors of different schemes.

3 Numerical results and analysis

3.1 Numerical results

The numerical results of different schemes for the genetic drift problem (1.2)–(1.4) and (1.9) are presented in this section. We first approximate the initial Dirac delta function (1.9) by a normal distribution function:

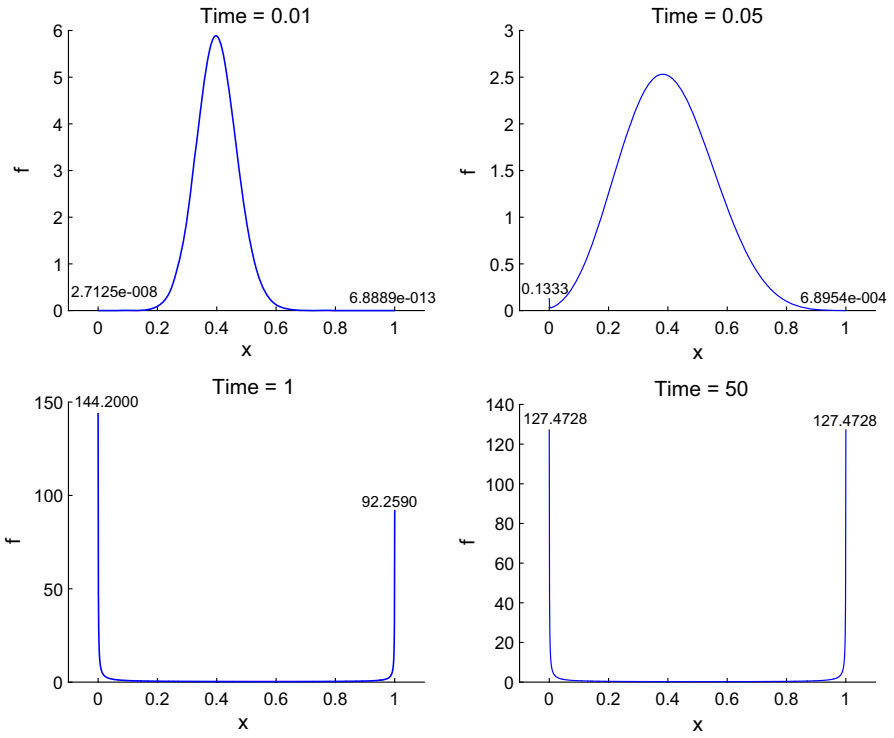

Fig. 1 Numerical results of FVM1 with initial state $f(x,0) \sim \mathcal{N}(0.4, 0.01^2)$ at different time $t = 0.01, 0.05, 1, 50$. The step sizes are $h = 1/1000$, $\tau = 1/1000$. At the equilibrium, heights of two pikes are equal due to the $O(h)$ numerical viscosity

$$f(x,0) \sim \mathcal{N}(p, \sigma^2) \tag{3.1}$$

with $\sigma << 1$. Later, we will show the dependence on the mean value p of the results.

Spike developing In Figs. 1, 2, 3, 4, 5 and 6, the numerical probability density at different time with various initial states are presented for three methods FVM1-3. These figures show that, for all three schemes, two spikes are formed at the boundary ($x = 0$ and $x = 1$) as time evolves. But the heights of the spikes at the equilibrium state are different. For FVM3, the heights of the two spikes at the boundary of equilibrium-state are dependent on the mean of initial probability density. For FVM1 and FVM2, the heights are equal and have nothing to do with the initial condition. This means FVM1 and FVM2 can not keep the conservation of expectation, and can not give a *complete solution*.

Expectation evolving The evolution of the discrete expectation was presented in Fig. 7 for all the schemes. It is clear that FVM3 preserves expectation while FVM1 and FVM2 do not. This means only FVM3 may yield a *complete solution*. And it is

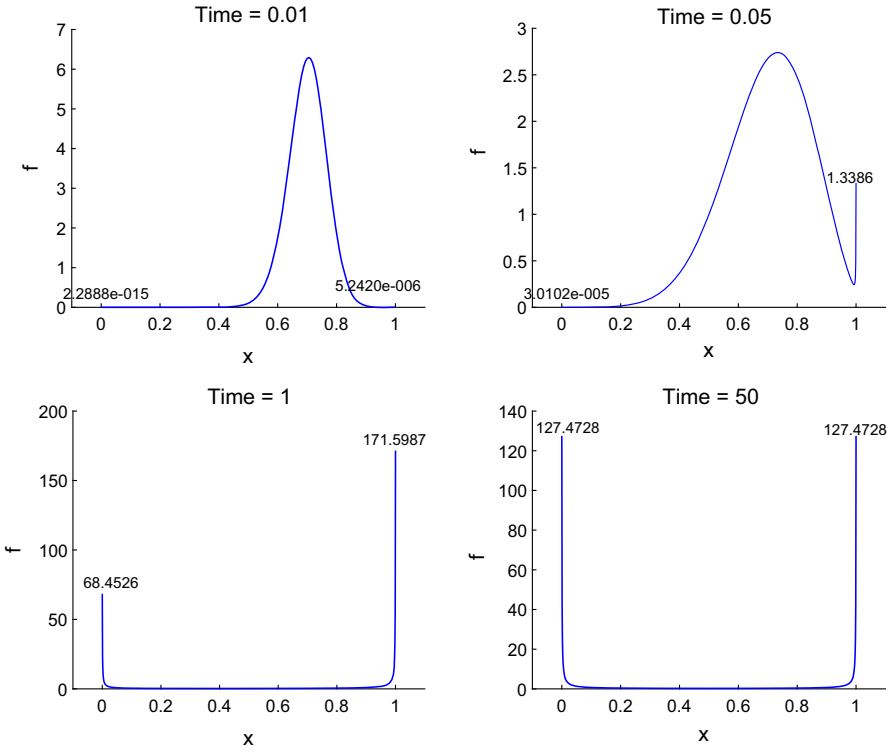

Fig. 2 Numerical results of FVM1 with initial state $f(x,0) \sim \mathcal{N}(0.7, 0.01^2)$ at different time $t = 0.01, 0.05, 1, 50$. The step sizes are $h = 1/1000$, $\tau = 1/1000$. At the equilibrium, heights of two pikes are equal due to the $O(h)$ numerical viscosity

observed that it takes a much longer time for FVM2 to achieve the equilibrium state than for FVM1.

Delta singularity checking Table 1 presents the probability density and focused probability at the boundaries (f_0^n, f_M^n and $\frac{h}{2} f_0^n$, $\frac{h}{2} f_M^n$) for FVM3, with different space grid sizes ($h = \frac{1}{100}, \frac{1}{1000}, \frac{1}{10,000}$) and different initial states ($f(x,0) \sim \mathcal{N}(0.4, 0.01^2)$ and $\mathcal{N}(0.7, 0.01^2)$) at $t_n = 6$ ($\tau = 0.0001$). It is shown that the probabilities concentrated on the boundary at the equilibrium state are independent of the space grid size h. This verifies that Dirac delta singularities do develop at the boundary points. This is just the fixation phenomena. It is also shown that the fixation probabilities are dependent on the expectation of the initial condition, and conservations of probability and expectation are always kept.

Table 2 shows the equilibrium state solution got by FVM2 at time $t_n = 20{,}000$ with time step $\tau = 1/100$ and initial state $f(x,0) = \mathcal{N}(0.7, 0.01^2)$. Combining with Figs. 3 and 4, the equilibrium state tends to $\frac{1}{2}\delta(x) + \frac{1}{2}\delta(1-x)$ as h tends to zero, no matter what the initial state is. This scheme predicts a wrong fixation probability.

Table 3 shows the equilibrium state solution got by FVM1 at time $t_n = 50$ with time step $\tau = 1/100$ and initial state $f(x,0) = \mathcal{N}(0.7, 0.01^2)$. Singularity but not

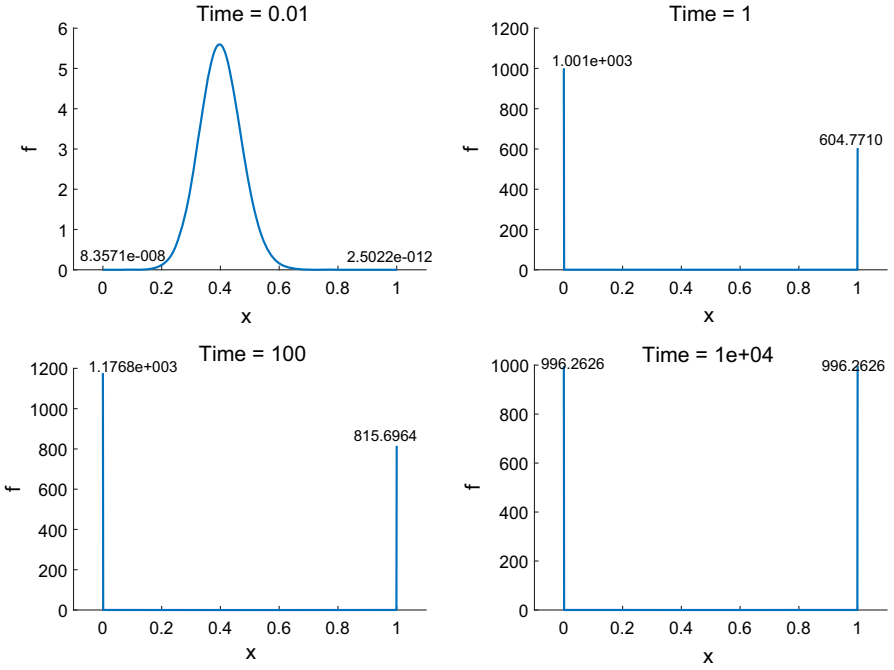

Fig. 3 Numerical results of FVM2 with initial state $f(x,0) \sim \mathcal{N}(0.4, 0.01^2)$ at different time $t = 0.01, 1, 100, 10{,}000$. The step sizes are $h = 1/1000$, $\tau = 1/1000$. At the equilibrium, heights of two pikes are equal due to the $O(h^2)$ numerical viscosity

Dirac delta is developed, i.e., no fixation happens. This scheme fails to predict the fixation phenomena.

Stability checking. In Table 4, for fixed grid spacing $h = 1/1000$ and two different initial states $f_0 \sim \mathcal{N}(0.4, 0.01^2)$ and $f_0 \sim \mathcal{N}(0.7, 0.01^2)$, the time step τ is changed from $\frac{1}{10}$ to $\frac{1}{10{,}000}$. The results shows that FVM 3 can get the same equilibrium state and keep conservation of total probability and expectation for different mesh ratio $\gamma = \frac{\tau}{h^2}$. This means that FVM3 for the genetic drift problem (1.2), (1.3), (1.4) and (1.9) is stable and independent of the mesh ratio γ, i.e., unconditional stable.

Dynamics checking In Fig. 8, we compare the results of Monte Carlo simulation for Wright–Fisher model (1.1) with numerical solution of FVM3. The initial state is $f(x,0) \sim \mathcal{N}(0.3, 0.01^2)$. The Monte Carlo simulations are done with $N = 100$ alleles, 5000 time steps and 5000 samples. The left panel shows the dynamics of mass at points $x = 0$ and $x = 1$; the right panel shows the dynamics of inner region $(0, 1)$ mass. The results validate that FVM3 could predict the correct long-time probability behavior (Fig. 7).

3.2 Numerical analysis

In this subsection, we will present the numerical analysis of above 3 FVM schemes. FVM3 can predict the correct long-time probability because it preserves probability,

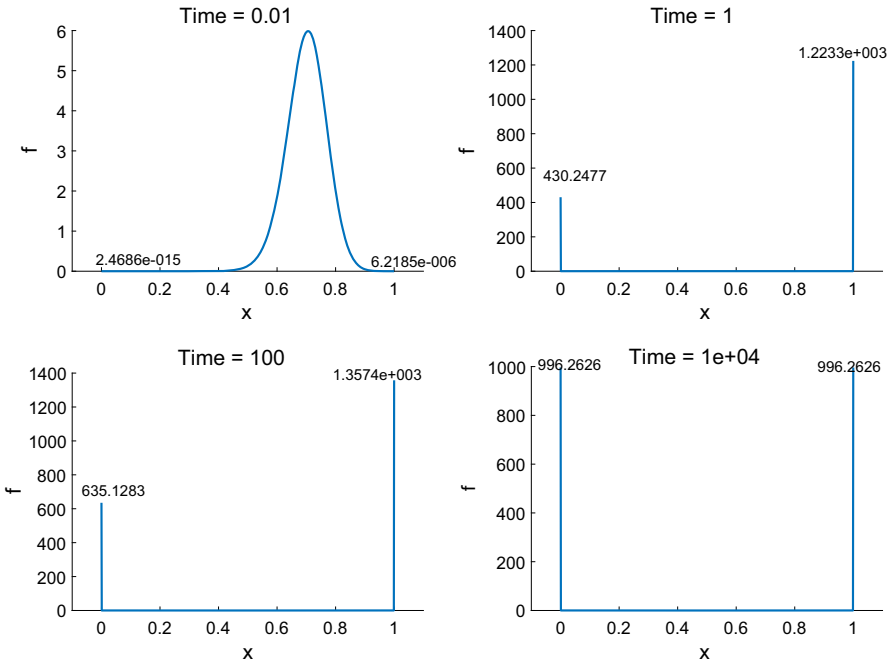

Fig. 4 Numerical results of FVM2 with initial state $f(x,0) \sim \mathcal{N}(0.7, 0.01^2)$ at different time $t = 0.01, 1, 100, 10{,}000$. The step sizes are $h = 1/1000$, $\tau = 1/1000$. At the equilibrium, heights of two pikes are equal due to the $O(h^2)$ numerical viscosity

expectation and positiveness. At same time, we also prove that any extra infinitesimal viscosity leads to a unique artificial equilibrium state, no matter the initial states. So FVM1-2 predict wrong probabilities of fixation due to their intrinsic viscosity, even though they are stable.

3.2.1 Stability, expectation conservation and long-time behavior of FVM3

In this part, we will prove the stability and long-time convergence for FVM3. (2.1), (2.2) and (2.5) can be split into three independent parts. For inner points, $0 < i < M$,

$$\frac{f_i^n - f_i^{n-1}}{\tau} - \frac{D_{i+1} f_{i+1}^n - 2D_i f_i^n + D_{i-1} f_{i-1}^n}{h^2} = 0, \tag{3.2}$$

with $D_i = x_i(1 - x_i)$; For the boundary points,

$$\frac{f_0^n - f_0^{n-1}}{\tau} - \frac{2D_1 f_1^n}{h^2} = 0, \quad \frac{f_M^n - f_M^{n-1}}{\tau} - \frac{2D_{M-1} f_{M-1}^n}{h^2} = 0. \tag{3.3}$$

Due to $D_0 = D_M = 0$, the unknowns at inner points $f_1^n, \ldots, f_{M-1}^n$ form a closed linear system which can be solved first. Then the boundary points f_0^n, f_M^n can be updated by the inner points f_1^n, f_{M-1}^n respectively.

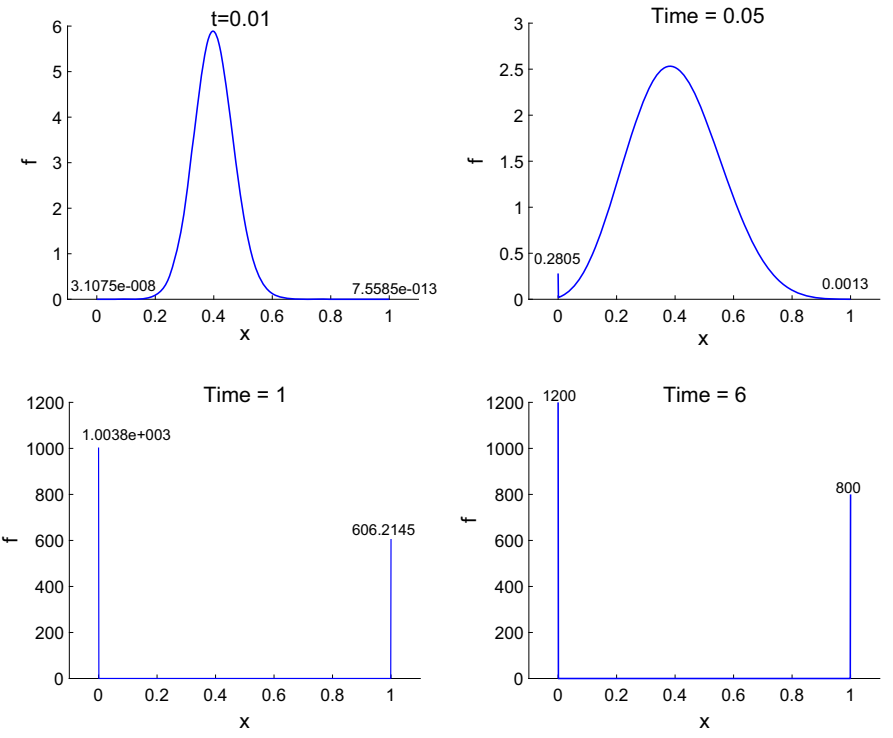

Fig. 5 Numerical results of FVM3 with initial state $f(x,0) \sim \mathcal{N}(0.4, 0.01^2)$ at different time $t = 0.01, 0.05, 1, 6$. The step sizes are $h = 1/1000$, $\tau = 1/1000$. At the equilibrium, the ratio of two pikes' heights is 3:2 and the expectation is 0.4 which is equal to the initial one

Theorem 3.1 *The central scheme FVM3 is unconditionally stable. Its solution keeps non-negative, if the initial state is non-negative.*

Proof Denote by $\mathbb{K}$ the matrix of the inner system (3.2). We have that $\mathbb{K}$ is tri-diagonal and with entries

$$k_{ii} = 1 + 2\gamma D_i,\ k_{i,i\pm1} = -\gamma D_{i\pm1},$$

where the mesh ratio $\gamma = \frac{\tau}{h^2}$. The system (3.2) yields a linear system

$$\mathbb{K} * (f_1^n, \ldots, f_{M-1}^n)^T = (f_1^{n-1}, \ldots, f_{M-1}^{n-1})^T.$$

Noting that $2D_i - D_{i+1} - D_{i-1} = 2h^2 > 0, \forall i = 1, \ldots, M-1$, we get $\mathbb{K}$ is a M-matrix, i.e., a diagonal dominant matrix with positiveness diagonal entries and non-positive off-diagonal entries. This guarantees the unconditional stability and positive preservation of FVM3. □

In the following, we prove all FVMs preserves discrete total probability and FVM3 also preserves the expectation. The discrete total probability and expectation at step n are defined as follows.

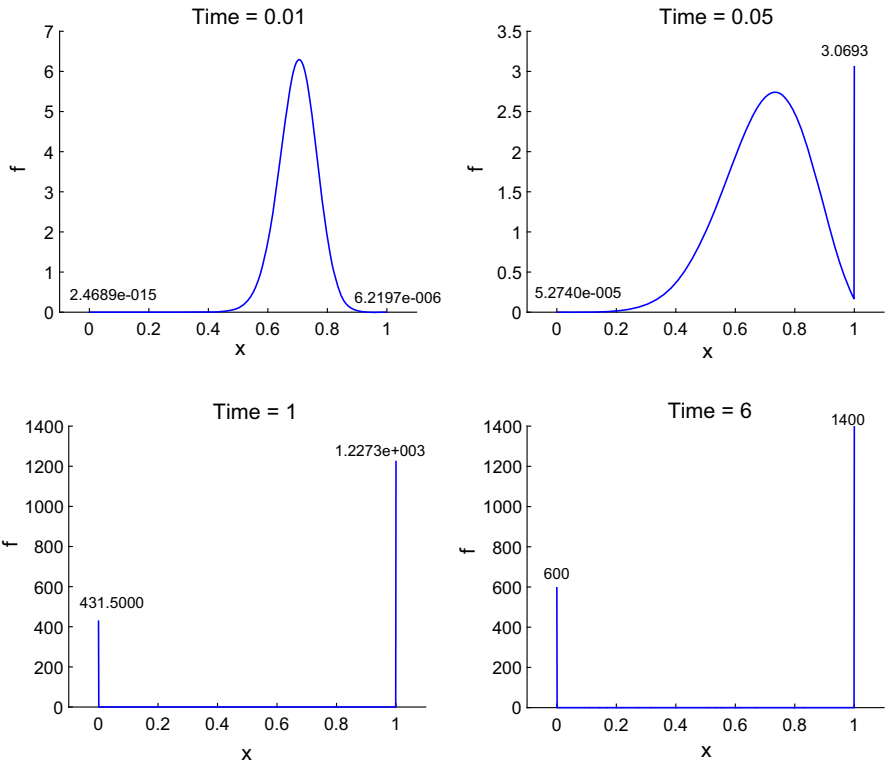

Fig. 6 Numerical results of FVM3 with initial state $f(x,0) \sim \mathcal{N}(0.7, 0.01^2)$ at different time $t = 0.01, 0.05, 1, 6$. The step sizes are $h = 1/1000$, $\tau = 1/1000$. At the equilibrium, the ratio of two pikes' heights is 3:7 and the expectation is 0.7 which is equal to the initial one

$$\begin{cases} P_n = \frac{h}{2} f_0^n + \frac{h}{2} f_M^n + \sum_{i=1}^{M-1} f_i^n h, \\ E_n = \frac{h}{2} x_0 f_0^n + \frac{h}{2} x_M f_M^n + \sum_{i=1}^{M-1} x_i f_i^n h, \end{cases} \tag{3.4}$$

Theorem 3.2 *The finite volume methods (2.1), (2.2) keep the discrete total probability, $P_n = P_{n-1}$, for $n = 1, 2, \ldots$. Furthermore, FVM3 yields a complete solution, which preserves the discrete expectation, $E_n = E_{n-1}$, for $n = 1, 2, \ldots$.*

Proof By the definition of P_n,

$$P_n - P_{n-1} = \frac{h}{2}(f_0^n - f_0^{n-1}) + \frac{h}{2}(f_M^n - f_M^{n-1}) + \sum_{i=1}^{M-1}(f_i^n - f_i^{n-1})h.$$

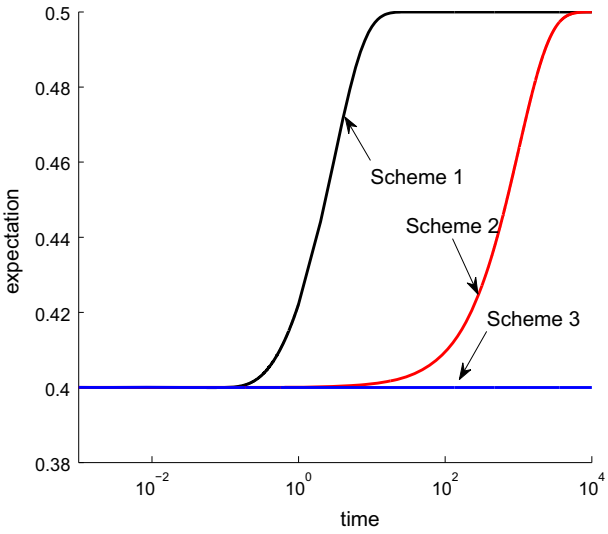

Fig. 7 Expectation produced by different schemes with initial state $f(x,0) \sim \mathcal{N}(0.4, 0.01^2)$ under the logarithm time scale. The discrete expectation is defined in (3.4). Here Schemes 1,2 and 3 represent FVM1, FVM2 and FVM3. Only FVM3 preserves the expectation. Both FVM1 and FVM2 achieve the same expectation which is not dependent on the initial value. And it can also be found that it takes a much longer time for FVM2 to achieve the equilibrium state than for FVM1

Table 1 Numerical results of FVM3 at the boundaries at equilibrium state ($t_n = 6$, time step $\tau = 1/10{,}000$) with different initial states and space grid sizes

Space step h	$f(x,0) = \mathcal{N}(0.4, 0.01^2)$			
	f_0^n	f_M^n	$\frac{h}{2} f_0^n$	$\frac{h}{2} f_M^n$
1/100	1.19999123e2	7.99991237e1	0.59999562	0.39999562
1/1000	1.19999115e3	7.99991154e2	0.59999558	0.39999558
1/10,000	1.19999115e4	7.99991146e3	0.59999557	0.39999557
Space step h	$f(x,0) = \mathcal{N}(0.7, 0.01^2)$			
	f_0^n	f_M^n	$\frac{h}{2} f_0^n$	$\frac{h}{2} f_M^n$
1/100	5.99992332e1	1.39999236e2	0.29999617	0.69999617
1/1000	5.99992260e2	1.39999226e3	0.29999613	0.69999613
1/10,000	5.99992253e3	1.39999225e4	0.29999613	0.69999613

The equilibrium state is $(1-p)\delta(x) + p\delta(1-x)$ if the initial state is with an expectation of p

Using (2.1) and (2.2), we have

$$P_n - P_{n-1} = -\tau \sum_{i=1}^{M-1} (j^n_{i+\frac{1}{2}} - j^n_{i-\frac{1}{2}}) - \tau j^n_{\frac{1}{2}} + \tau j^n_{M-\frac{1}{2}} = 0.$$

Table 2 Numerical results of FVM2 at the boundaries at equilibrium state ($t_n = 20{,}000$, time step $\tau = 1/100$) with initial state $f(x,0) = \mathcal{N}(0.7, 0.01^2)$

Step h	f_0^n	f_M^n	$\frac{h}{2} f_0^n$	$\frac{h}{2} f_M^n$
1/200	197.10	197.10	0.4928	0.4928
1/400	396.74	396.74	0.4959	0.4959
1/800	796.39	796.39	0.4977	0.4977
1/1600	1596.0	1596.0	0.4988	0.4988
1/3200	3193.2	3198.2	0.4989	0.4997

The equilibrium state tends to $\frac{1}{2}\delta(x) + \frac{1}{2}\delta(1-x)$ as h tends to zero, no matter what the initial state is

Table 3 Numerical results of upwind scheme FVM1 at the boundaries at equilibrium state ($t_n = 50$, time step $\tau = 1/100$) with $f(x,0) = \mathcal{N}(0.7, 0.01^2)$

Step h	f_0^n	f_M^n	$\frac{h}{2} f_0^n$	$\frac{h}{2} f_M^n$
1/200	32.348	32.348	0.08087	0.08087
1/400	57.931	57.931	0.07241	0.07241
1/800	105.03	105.03	0.06564	0.06564
1/1600	192.25	192.25	0.06008	0.06008
1/3200	354.64	354.64	0.05541	0.05541

Singularity but not Dirac delta is developed, i.e., no fixation happens

Table 4 Numerical results of FVM3 at the boundaries at equilibrium state ($t_n = 6$) with space grid size $h = 1/1000$

Time step τ	$f(x,0) = \mathcal{N}(0.4, 0.01^2)$			
	f_0^n	f_M^n	$\frac{h}{2} f_0^n$	$\frac{h}{2} f_M^n$
1/10	1.19997448e3	7.99974480e2	0.59998724	0.39998724
1/100	1.19999005e3	7.99990054e2	0.59999503	0.39999502
1/1000	1.19999106e3	7.99991058e2	0.59999553	0.39999553
1/10,000	1.19999115e3	7.99991154e2	0.59999558	0.39999558
Time step τ	$f(x,0) = \mathcal{N}(0.7, 0.01^2)$			
	f_0^n	f_M^n	$\frac{h}{2} f_0^n$	$\frac{h}{2} f_M^n$
1/10	5.99977672e2	1.39997767e3	0.29998884	0.69998884
1/100	5.99991298e2	1.39999130e3	0.29999565	0.69999565
1/1000	5.99992177e2	1.39999218e3	0.29999609	0.69999608
1/10,000	5.99992260e2	1.39999226e3	0.29999613	0.69999613

The scheme works well for any ratio of grid

On the other hand,

$$E_n - E_{n-1} = \frac{h}{2} x_0 (f_0^n - f_0^{n-1}) + \frac{h}{2} x_M (f_M^n - f_M^{n-1}) + \sum_{i=1}^{M-1} x_i (f_i^n - f_i^{n-1}) h.$$

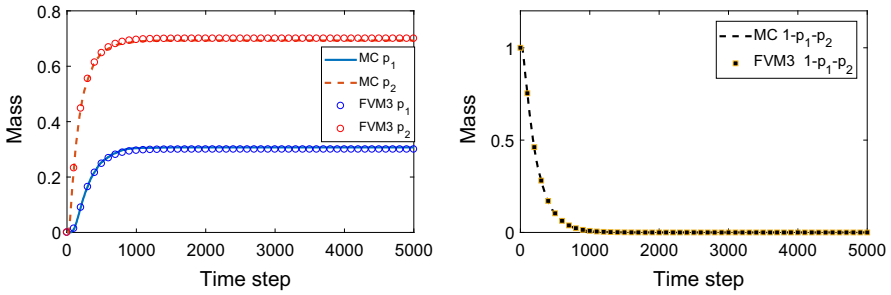

Fig. 8 Consistency between Mont Carlo simulations of Wright–Fisher model and FVM3 numerical solutions with initial state $f(x,0) \sim \mathcal{N}(0.3, 0.01^2)$. Left: the dynamics of mass at points $x = 0$ and $x = 1$; Right: the dynamics of inner region $(0, 1)$ mass. The Mont Carlo simulations are done with number of alleles $N = 100$, number of sample 5000 and time step 5000. FVM3 is set to be $h = 1/200,\ \tau = 1/400$

Using (2.1), (2.2) and (2.5), we have, by summation by parts, that

$$\begin{aligned} E_n - E_{n-1} &= -\tau \sum_{i=1}^{M-1} x_i (j^n_{i+\frac{1}{2}} - j^n_{i-\frac{1}{2}}) + \tau x_M j^n_{M-\frac{1}{2}} \\ &= h\tau \sum_{i=1}^{M-1} j^n_{i+\frac{1}{2}} = \tau \sum_{i=1}^{M-1} (D_{i+1} f^n_{i+1} - D_i f^n_i) = 0. \end{aligned}$$

The proof is completed. □

We now turn to the long-time behavior of FVM3. First we have the following decay estimate on the inner part of the solution.

Lemma 3.3 *Let f^n be the solution of the central scheme FVM3. Then for any step sizes $h, \tau > 0$,*

$$\sum_{i=1}^{M-1} |f^n_i|^2 \le \left(\frac{1}{1+4\tau}\right)^n \sum_{i=1}^{M-1} |f^0_i|^2,\ t_n = n\tau.$$

Proof Denote by $\Delta_h v_i = \dfrac{v_{i-1} - 2v_i + v_{i+1}}{h^2}$ the discrete Laplacian. Multiplying by f^n_i on both sides of (3.2), summing from $i = 1$ to $M - 1$, we have that

$$\frac{1}{2\tau} \sum_{i=1}^{M-1} |f^n_i|^2 - \sum_{i=1}^{M-1} \Delta_h (D_i f^n_i) f^n_i \le \frac{1}{2\tau} \sum_{i=1}^{M-1} |f^{n-1}_i|^2. \tag{3.5}$$

We claim that

$$-\sum_{i=1}^{M-1} \Delta_h (D_i v_i) v_i \ge 2 \sum_{i=1}^{M-1} |v_i|^2,\ \forall\ v \in \mathbb{R}^{M-1}. \tag{3.6}$$

Actually in matrix form, the above formula is

$$v^T \tilde{\mathbb{M}} v \geq 2 v^T v, \ \forall \ v \in \mathbb{R}^{M-1},$$

where $\tilde{\mathbb{M}}$ is a tri-diagonal matrix with entries

$$\tilde{m}_{i,i} = 2D_i/h^2, \quad \tilde{m}_{i,i\pm1} = -D_{i\pm1}/h^2.$$

Noting that $-\Delta_h D_i = -\Delta_h x_i(1-x_i) = \tilde{m}_{i,i-1} + \tilde{m}_{i,i} + \tilde{m}_{i,i+1} \equiv 2$, we have $\tilde{\mathbb{M}} - 2I$ is still diagonal dominant and semi-positive definite. So inequality (3.6) is true.

Combining (3.5) and (3.6), we obtain

$$\sum_{i=1}^{M-1} |f_i^n|^2 \leq \frac{1}{1+4\tau} \sum_{i=1}^{M-1} |f_i^{n-1}|^2 \leq \left(\frac{1}{1+4\tau}\right)^n \sum_{i=1}^{M-1} |f_i^0|^2.$$

So the proof is completed. □

Theorem 3.4 *As to the long-time behavior of the solution f^n of the central scheme FVM3, we have, for any fixed step sizes h and τ, as $n \to \infty$,*

$$\begin{cases} f_i^n \to 0, \text{ for } i = 1, \ldots, M-1, \\ \frac{h}{2} f_0^n + \frac{h}{2} f_M^n \to P_0, \\ \frac{h}{2} x_0 f_0^n + \frac{h}{2} x_M f_M^n \to E_0. \end{cases}$$

If the initial state is approximated by (3.1), we further have

$$\begin{cases} \frac{h}{2} f_0^n \to (P_0 - E_0) \approx 1 - p, \\ \frac{h}{2} f_M^n \to E_0 \approx p. \end{cases}$$

The means that FVM3 predicts the right fixation probability.

Proof From the decay property of Lemma 3.3, we have as $n \to \infty$, $f_i^n \to 0$, for $i = 1, \ldots, M-1$. The other convergence results come from the conservation properties of Theorem 3.2. The proof is completed. □

3.2.2 Effect of additional viscosity

As shown in Figs. 1, 2, 3 and 4, in the results of the first two schemes, the values of equilibrium-state solutions at boundaries $x = 0$ and $x = 1$ are of the same height with different initial states. This is not consistent with the equilibrium state of the singular solutions given in (1.10), and also does not satisfy conservation of expectation.

To see what's wrong here, taking the difference between these schemes, we have the following results.

Theorem 3.5 *Denote by $D(x) = x(1-x)$ and by $b(x) = D'(x) = 1-2x$. The central Scheme FVM2, (2.1), (2.2) and (2.4) can be regraded as FVM3 plus a viscosity term $\Lambda_i = -\frac{h^2}{4} * \frac{f_{i+1}^n - 2f_i^n + f_{i-1}^n}{h^2}$, i.e.,*

$$FVM2: \quad \frac{f_i^n - f_i^{n-1}}{\tau} - \frac{D_{i+1} f_{i+1}^n - 2D_i f_i^n + D_{i-1} f_{i-1}^n}{h^2} + \Lambda_i = 0.$$

The upwind Scheme FVM1, (2.1)–(2.3) can be regarded as FVM2 plus a viscosity term

$$\tilde{\Lambda}_i = -\frac{h}{2} * \frac{|b_{i+\frac{1}{2}}| f_{i+1}^n - \left(|b_{i+\frac{1}{2}}| + |b_{i-\frac{1}{2}}|\right) |f_i^n + |b_{i-\frac{1}{2}}| f_{i-1}^n}{h^2}, \quad i.e.,$$

$$FVM1: \quad \frac{f_i^n - f_i^{n-1}}{\tau} - \frac{D_{i+1} f_{i+1}^n - 2D_i f_i^n + D_{i-1} f_{i-1}^n}{h^2} + \Lambda_i + \tilde{\Lambda}_i = 0.$$

Proof of Theorem 3.5. Λ_i is the difference between FVM2 and FVM3. We have after direct computation that

$$\begin{aligned}
\Lambda_i &= \frac{1}{h^2}\left(\left(D_{i+1} - D_{i+\frac{1}{2}} - \frac{h}{2} b_{i+\frac{1}{2}}\right) f_{i+1}^n + \left(D_{i-1} - D_{i-\frac{1}{2}} + \frac{h}{2} b_{i-\frac{1}{2}}\right) f_{i-1}^n\right) \\
&\quad + \frac{1}{h^2}\left(\left(D_{i+\frac{1}{2}} + D_{i-\frac{1}{2}} - 2D_i - \frac{h}{2}\left(b_{i+\frac{1}{2}} - b_{i-\frac{1}{2}}\right)\right) f_i^n\right) \\
&= \frac{1}{h^2}\left(\frac{h^2}{8} D''(x) f_{i+1}^n + \frac{h^2}{8} D''(x) f_{i-1}^n - \frac{h^2}{4} D''(x) f_i^n\right) \\
&= -\frac{h^2}{4} * \frac{f_{i+1}^n - 2f_i^n + f_{i-1}^n}{h^2},
\end{aligned}$$

where we have used the facts that $D''(x) = -2$ and $\frac{d^n}{dx^n} D = 0,\ n > 2$.

Note that $\tilde{\Lambda}_i$ is the difference between FVM1 and FVM2, which, as well-known, is a first-order ($O(h)$) viscosity term introduced by the upwind technique [12]. □

So FVM1 and FVM2 are unconditionally stable since they introduce extra viscosities to the stable scheme FVM3. FVM2 is equivalent to discretize the problem $\partial_t f - \partial_{xx}\left(\left(x(1-x) + \frac{h^2}{4}\right) f\right) = 0$ by FVM3. To see the effect of this constant viscosity, we consider a viscosity vanishing procedure to the equilibrium state problem. First, a infinitesimal diffusion is added, then the limit behavior of the equilibrium-state solution is investigated.

For $\varepsilon > 0$, find non-zero $f_\varepsilon(x)$ such that

$$\begin{cases} -\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(\left(x(1-x) + \varepsilon\right) f_\varepsilon\right) = 0, x \in (0,1), \\ \frac{\mathrm{d}}{\mathrm{d}x}\left(\left(x(1-x) + \varepsilon\right) f_\varepsilon\right)\Big|_{x=0,1} = 0, \end{cases} \tag{3.7}$$

with a constraint $\int_0^1 f_\varepsilon(x)\mathrm{d}x = 1$.

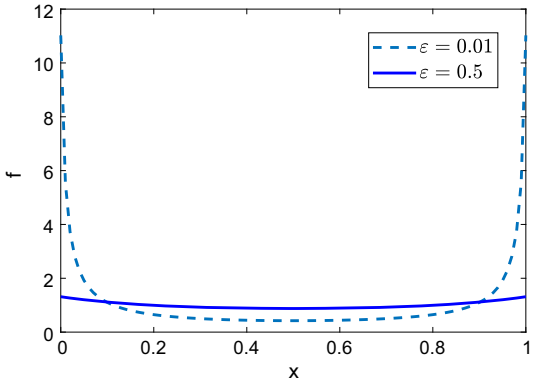

Fig. 9 Two equilibrium-state solutions f_ε of equation (3.7) with artificial viscosity $\varepsilon = 0.5$, and 0.01

Theorem 3.6 *If f_ε is a non-zero solution of* (3.7)*, then* $\lim_{\varepsilon\to 0} f_\varepsilon(x) = \frac{1}{2}\delta(x) + \frac{1}{2}\delta(x-1)$ *in the sense of distribution.*

Proof Integrating the problem (3.7), we get from the boundary condition that

$$f_\varepsilon(x) = \frac{b_\varepsilon}{x(1-x)+\varepsilon}. \tag{3.8}$$

The constraint on the total probability yields that

$$b_\varepsilon = \frac{\sqrt{1/4+\varepsilon}}{\ln(\sqrt{1/4+\varepsilon}+1/2) - \ln(\sqrt{1/4+\varepsilon}-1/2)}.$$

See Fig. 9 for some profiles of f_ε.

For any $\varphi \in C^\infty[0,1]$, we want to prove that $\langle f_\varepsilon, \varphi\rangle \to \frac{1}{2}\varphi(0) + \frac{1}{2}\varphi(1)$ as $\varepsilon \to 0$. By symmetry, $\int_0^{\frac{1}{2}} f_\varepsilon dx = \int_{\frac{1}{2}}^1 f_\varepsilon dx = \frac{1}{2}$, so we have

$$\begin{aligned}\langle f_\varepsilon, \varphi\rangle = &\int_0^{1/2} f_\varepsilon(x)(\varphi(x)-\varphi(0))\mathrm{d}x + \int_{1/2}^1 f_\varepsilon(x)(\varphi(x)-\varphi(1))\mathrm{d}x \\ &+ \frac{1}{2}\varphi(0) + \frac{1}{2}\varphi(1).\end{aligned}$$

Now we need to prove $\lim_{\varepsilon\to 0}\int_0^{\frac{1}{2}} f_\varepsilon\,(\varphi(x)-\varphi(0))\mathrm{d}x = \lim_{\varepsilon\to 0}\int_{\frac{1}{2}}^1 f_\varepsilon\,(\varphi(x)-\varphi(1))\mathrm{d}x = 0$. Actually, we have, by denoting by $M_1 = \max_{x\in[0,\frac{1}{2}]} |\varphi'(x)|$,

$$\begin{aligned}\left|\int_0^{\frac{1}{2}} f_\varepsilon(\varphi(x)-\varphi(0))\mathrm{d}x\right| &= \left|b_\varepsilon\int_0^{\frac{1}{2}} \frac{1}{x(1-x)+\varepsilon}\,(\varphi(x)-\varphi(0))\,\mathrm{d}x\right| \\ &\le b_\varepsilon M_1\int_0^{\frac{1}{2}} \frac{x}{x(1-x)+\varepsilon}\,\mathrm{d}x \le b_\varepsilon M_1\int_0^{\frac{1}{2}} \frac{x}{x(1-x)}\mathrm{d}x \\ &= b_\varepsilon M_1 * \ln 2 \to 0, \quad \text{as } \varepsilon \to 0,\end{aligned}$$

Table 5 Numerical results of upwind scheme FVM1 at the left boundary at equilibrium state ($t = 50$) with $f(x, 0) = \mathcal{N}(0.7, 0.01^2)$ and time step $\tau = 1/100$. $Df_i \equiv \frac{f_{i+1}-f_i}{h}$

Step$h$	1/200	1/400	1/800	1/1600	1/3200
Df_0/Df_1	5.0103	5.0051	5.0025	5.0013	5.0006

since $b_\varepsilon \to 0$.

Similarly, we also have $\lim_{\varepsilon\to 0} \int_{\frac{1}{2}}^{1} f_\varepsilon\ (\varphi(x) - \varphi(1))\mathrm{d}x = 0$.

Thus we have that $\lim_{\varepsilon\to 0} \langle f_\varepsilon, \varphi\rangle = \frac{1}{2}\varphi(0) + \frac{1}{2}\varphi(1), \forall \varphi(x) \in \mathrm{C}^\infty[0, 1]$. The proof is completed. □

The above theorem verifies the observation in Figs. 3 and 4 and Table 2, i.e., the equilibrium state got by FVM2 tends to $\frac{1}{2}\delta(x) + \frac{1}{2}\delta(1 - x)$ as h tends to zero, *no matter what the initial state is*. This means that conservation of expectation is violated by any extra viscosity.

The upwind scheme FVM1 also introduces a viscosity term $\tilde{\Lambda}_i$ (see Theorem 3.5). Its behavior in Figs. 1 and 2 and Table 3 can not be explained by Theorem 3.6 since this viscosity is much more complicated than the constant viscosity introduced by FVM2. Let us rewrite $\tilde{\Lambda}_i$ as

$$\begin{aligned}\tilde{\Lambda}_i &= -\frac{1}{2}|b_{i+\frac{1}{2}}|\frac{f_{i+1}^n - f_i^n}{h} + \frac{1}{2}|b_{i-\frac{1}{2}}|\frac{f_i^n - f_{i-1}^n}{h}\\ &\approx -\frac{1}{2}\frac{f_{i+1}^n - f_i^n}{h} + \frac{1}{2}\frac{f_i^n - f_{i-1}^n}{h}, \text{ near boundary points.}\end{aligned}$$

Denote by $Df_i = \frac{f_{i+1}-f_i}{h}$. From Table 5, we have that at the left boundary, $Df_0 \approx 5Df_1$, i.e.,

$$\tilde{\Lambda}_1 \approx 2\frac{f_1^n - f_0^n}{h}.$$

Symmetrically, we have that, at the right boundary,

$$\tilde{\Lambda}_{M-1} \approx -2\frac{f_M^n - f_{M-1}^n}{h}.$$

It means that the extra viscosity term $\tilde{\Lambda}_i$ behaviors as a one-way mutation term $\partial_x(2(1-x)f)$ near the left boundary and as another one-way mutation term $-\partial_x(2xf)$ near the right boundary. This artificial two-way mutation leads to the unique equilibrium state as in Figs. 1 and 2, no matter what the initial state is (see [6]).

Finally, FVM1 introduces a viscosity of first-order ($O(h)$) while FVM2 introduces a viscosity of second-order ($O(h^2)$). This is the reason why FVM2 takes a much longer time than FVM1 to achieve the equilibrium state (see Fig. 7).

Remark 3.7 The equilibrium state problem (3.7) is actually to find an eigenfunction for the zero eigenvalue. For the original diffusion operator, i.e., $\varepsilon = 0$, there exists infinite eigenfunctions with the form as $(1-p)\delta(x) + p\delta(1-x)$, $\forall p \in [0,1]$, corresponding to the zero eigenvalue. But for any $\varepsilon > 0$, (3.7) has a unique eigenfunction (3.8). So numerical schemes with any extra viscosity will lead to a unique equilibrium state no matter what the initial state is.

4 Discussion of finite difference methods and finite element methods

We have shown that a central scheme FVM2 is unconditionally stable for a convection-dominated problem. It seems beyond the common understanding on the constraint of Peclet's number [15], which is necessary for a central scheme to achieve stability. In the following, we will check a central finite difference method (FDM). We will see that the constrain on Peclet's number does matter. We also discuss some finite element methods (FEM) for this problem.

4.1 Finite difference method

If the diffusion term and convection term in (1.11) is discretized by central difference directly, we get

$$\text{FDM1:}\quad \frac{f_i^n - f_i^{n-1}}{\tau} - \frac{D_{i+1/2} f_{i+1}^n - (D_{i+1/2} + D_{i-1/2}) f_i^n + D_{i-1/2} f_{i-1}^n}{h^2} + \frac{b_{i+1} f_{i+1}^n - b_{i-1} f_{i-1}^n}{2h} = 0, \quad 1 < i < M, \tag{4.1}$$

Noting that

$$\frac{b_{i+1} f_{i+1}^n - b_{i-1} f_{i-1}^n}{2h} = \frac{b_{i+1/2} \frac{f_{i+1}^n + f_i^n}{2} - b_{i-1/2} \frac{f_i^n + f_{i-1}^n}{2}}{h} + \frac{h^2}{2} \frac{f_{i-1}^n - 2 f_i^n + f_{i+1}^n}{h^2},$$

so we have from Theorem 3.5 that

Proposition 4.1 *The following results are always true for FDM1.*

- *FDM1 is just FVM2 plus an anti-diffusion* $\frac{h^2}{2} \frac{f_{i-1}^n - 2 f_i^n + f_{i+1}^n}{h^2}$.
- *FDM1 is just FVM3 plus an anti-diffusion* $\frac{h^2}{4} \frac{f_{i-1}^n - 2 f_i^n + f_{i+1}^n}{h^2}$.
- *The Peclet's number* $P_e \equiv \max_i \{\frac{h}{2} \frac{|b_{i+1}|}{D_{i+1/2}}, \frac{h}{2} \frac{|b_{i-1}|}{D_{i-1/2}}\} = \frac{1}{1-h/2} > 1$ *for any space step* h.

For non-degenerated convection-diffusion problems, we can choose the space step h sufficiently small to make sure the Peclet's number $P_e \leq 1$, which is sufficient to get a stable scheme. Due to the degeneration of the diffusion coefficient $D(x)$, the Peclet's number here can never be less than 1. The numerical results in Fig. 10 also show that FDM1 is not stable and can not preserve the positivity of the solution.

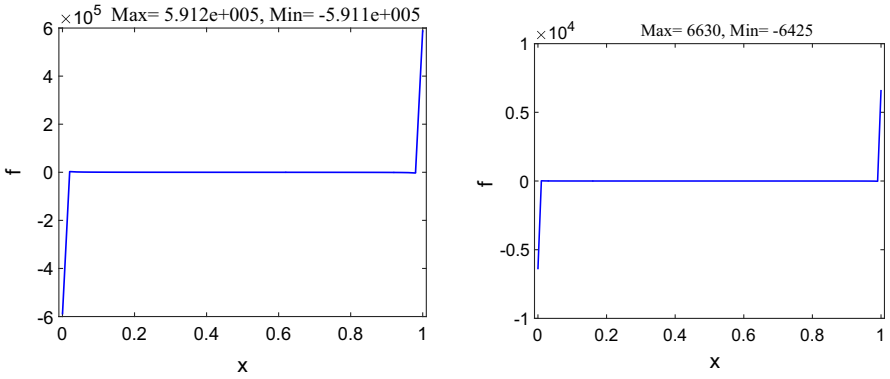

Fig. 10 Numerical solution at $t = 500$ by FDM1 with initial state $f(x,0) \sim \mathcal{N}(0.7, 0.01^2)$. $\tau = 0.01$. Left: $h = 0.02$; Right: $h = 0.01$

Remark 4.2 The scheme in (4.1) is not completed. Only inner point equations are involved. FDM is not easy to treat the no-flow boundary conditions in (1.3)–(1.4). Improper treatment may even break the conservation of mass. In the numerical tests of Fig. 10, to be consistent with the inner points, Proposition 4.1 indicate to discretize the equation $f_t - ((D(x) - h^2/4)f)_{xx} = 0$ by FVM3 at both boundary points.

4.2 Finite element methods

In this subsection, we consider three linear FEM schemes with or without numerical quadrature (see [16] for standard language on FEM). Taking the same uniform partition as before, we have the grid points as $x_i = ih, t_n = n\tau,\ i = 0, \ldots, M,\ n = 0, 1, \ldots$. We now have the elements as $e_i = [x_i, x_{i+1}],\ i = 0, \ldots, M-1$. The linear FE space is defined as $S_h = \{v_h \in C[0,1] : v_h|_{e_i}$ is linear for $i = 0, \ldots, M-1\} = \text{span}\{\varphi_0, \ldots, \varphi_M\}$, where φ_i is the nodal basis of piecewise linear interpolation at node x_i. The L^2-inner product by exact integration and discrete L^2-inner product by quadrature are defined respectively as:

$$(u_h, v_h) \equiv \int_0^1 u_h(x)v_h(x)dx; \quad (u_h, v_h)_h \equiv (u_0v_0 + u_Mv_M)\frac{h}{2} + \sum_{i=1}^{M-1} u_iv_ih.$$

Three FEM schemes are defined as follows: Given $f_h^0 = \sum_{i=0}^M f_i^0\varphi_i(x)$, find $f_h^n \in S_h$, for $n = 1, 2, \ldots$, such that

$$\text{FEM1}: \left(\frac{f_h^n - f_h^{n-1}}{\tau}, v_h\right) + \left(\partial_x(D(x)f_h^n), \partial_x v_h\right) = 0,\ \forall\, v_h \in S_h, \tag{4.2}$$

$$\text{FEM2}: \left(\frac{f_h^n - f_h^{n-1}}{\tau}, v_h\right)_h + \left(\partial_x(D(x)f_h^n), \partial_x v_h\right)_h = 0,\ \forall\, v_h \in S_h, \tag{4.3}$$

$$\text{FEM3}: \left(\frac{f_h^n - f_h^{n-1}}{\tau}, v_h\right)_h + \left(\partial_x(D(x) f_h^n), \partial_x v_h\right) = 0, \ \forall\, v_h \in S_h. \quad (4.4)$$

FEM1 is the standard P_1 finite element method. In FEM2 and FEM3, quadrature with trapezoidal rule is used. In matrix form, they are equivalent to

$$\begin{aligned}
\text{FEM1}: \ & \mathbb{M}\frac{f^n - f^{n-1}}{\tau} + \mathbb{K} f^n = 0, \\
\text{FEM2}: \ & \mathbb{M}_h \frac{f^n - f^{n-1}}{\tau} + \mathbb{K}_h f^n = 0, \\
\text{FEM3}: \ & \mathbb{M}_h \frac{f^n - f^{n-1}}{\tau} + \mathbb{K} f^n = 0,
\end{aligned}$$

where $f^n = (f_0^n, \ldots, f_M^n)^T$ and the mass matrix $\mathbb{M}$, the stiffness matrices $\mathbb{K}$ and $\mathbb{K}_h$ are tri-diagonal, the lumped mass matrix $\mathbb{M}_h$ is diagonal, their entries are respectively $m_{i,j} = (\varphi_j, \varphi_i)$, $k_{i,j} = (\partial_x(D(x)\varphi_j), \partial_x\varphi_i), k_{i,j}^h = (\partial_x(D(x)\varphi_j), \partial_x\varphi_i)_h$, and $m_{i,i}^h = (\varphi_i, \varphi_i)_h$, $i, j = 0, \ldots, M$.

After direct calculation, we have

Proposition 4.3 *The following results are valid for FEMs.*

- *FEM3 is exactly FVM3* (3.2) *and* (3.3)*, which works well.*
- *FEM2 is just FVM3 plus an anti-diffusion* $\frac{h^2}{2}\frac{f_{i-1}^n - 2f_i^n + f_{i+1}^n}{h^2}$ *at the left hand side, which is induced by the numerical quadrature applying to the diffusion term.*
- *FEM1 is FVM3 plus an anti-diffusion* $\frac{h^2}{6\tau}\frac{f_{i-1}^n - 2f_i^n + f_{i+1}^n}{h^2}$ *at the left hand side and plus an term* $\frac{h^2}{6\tau}\frac{f_{i-1}^{n-1} - 2f_i^{n-1} + f_{i+1}^{n-1}}{h^2}$ *at right hand side.*

The existence of the anti-diffusion in FEM1 and FEM2 leads to non-stability and fails to preserve the positiveness of the solution. The numerical results are depicted in Fig. 11. In the left figure, the numerical solution by FEM1 becomes to negative very

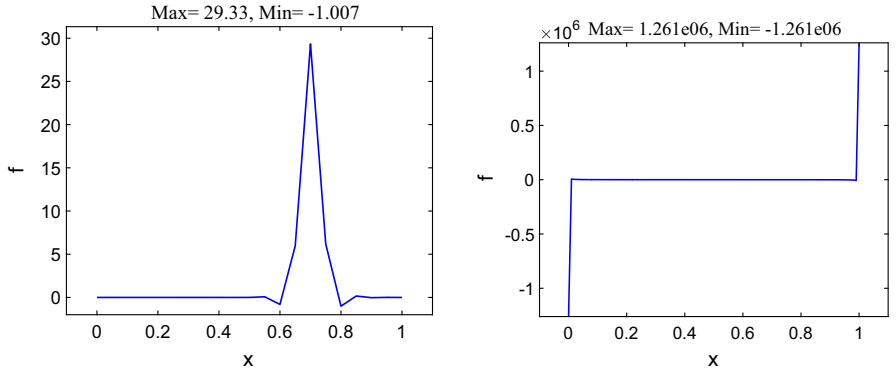

Fig. 11 Numerical solution by FEM with initial state $f(x, 0) \sim \mathcal{N}(0.7, 0.01^2)$. Left: FEM1 at $t = 0.001$ with $\tau = 0.0001$, $h = 0.05$; Right: FEM2 at $t = 500$ with $\tau = 0.01$, $h = 0.01$

early. In the right one, the numerical solution by FEM2 becomes very negative after a long time.

Remark 4.4 Taking $v_h = 1$ and then $v_h = x$ in FEM1 (4.2), we have that FEM1 preserve the total of mass (or total of probability) and the mass center (or expectation). Unfortunately, it is not locally conservative and can not preserve the positiveness of solution. $\mathbb{M} + \tau\mathbb{K}$ is not a M-matrix for sufficiently small $\tau > 0$.

In FEM3 (4.4), mass matrix is determined by numerical quadrature, while the stiffness matrix is determined by exact integration. It is very tricky. Otherwise, anti-diffusion will always be introduced if we calculate the mass matrix by exact integration as FEM1, or calculate the stiffness matrix by quadrature as FEM2.

5 Conclusions

We have considered three FVM schemes, one central FDM scheme and three FEM schemes for the genetic drift problem. FVM1 and FVM2 discretize the flux in convection and diffusion form using upwind and central methods, respectively. FVM3 discretizes the flux as a whole using the central method. FDM1 is a central difference scheme. FEM1,2, and 3 are P_1 finite element methods with or without quadrature.

Numerical experiments and numerical analysis show that

- Only schemes, preserving all the probability, expectation and the positivity, yield the correct fixation probabilities, such as FVM3 and FEM3.
- Any numerical viscosity will lead to an artificial equilibrium state. Comparing with FVM3, FVM1 introduces a first-order of $O(h)$ viscosity term since it is a upwind scheme, while FVM2, as a central scheme, introduces a second-order of $O(h^2)$ viscosity term. That is the reason why they are also unconditionally stable and it takes quite different time for them to reach the wrong equilibrium state.
- All the three FVM schemes are unconditionally stable and preserve the total probability.
- FDM1, as a central difference scheme, is unconditionally unstable since it introduces an anti-diffusion term of $O(h^2)$ order and the numerical Peclet's number is always larger than 1.
- FEM1 and FEM2 are unstable, since they introduce anti-diffusion terms. In FEM1, anti-diffusion comes from the exact integration of the mass matrix, while in FEM2, anti-diffusion comes from the quadrature error to the stiffness matrix.

All the complexity comes from diffusion degeneration and convection dominance. For this kind of problem, the numerical method should be carefully chosen. Any method with intrinsic numerical viscosity or anti-diffusion must be avoided, though they might be ignorable for non-degenerated problems.

Recently, we construct another numerical method for the 1-d genetic drift problem with pure diffusion or semi-selection by a framework of energetic variational approach [3]. If a population or species of organisms typically includes multiple alleles at each locus among various individuals, which is called multiple alleles, the problem is high dimensional [8,17,19]. It is a great challenge to find a *complete solution* since the

singularity will always be developed on the boundary surface rather than only two points for 1-D case. The numerical methods for the multiple alleles include the fixation phenomena will be addressed in the future work.

Acknowledgements The authors benefitted a great deal from discussions with Prof. David Waxman, Prof. Xinfu Chen and Prof. Xiaobing Feng. The authors thank the anonymous referees for their most valuable comments which improve the paper.

References

1. Crow, J.F., Kimura, M.: An Introduction to Population Genetics Theory. Harper & Row, New York (1970)
2. Der, R., Epstein, C.L., Plotkin, J.B.: Generalized population models and the nature of genetic drift. Theor. Popul. Biol. **80**(2), 80–99 (2011)
3. Duan, C., Liu, C., Wang, C., Yue, X.: Numerical complete solution for random genetic drift by energetic variational approach. arXiv:1803.09436. Mathematical modelling and numerical analysis accepted and published online (2018) https://doi.org/10.1051/m2an/2018058
4. Eymard, R., Gallouët, T., Herbin, R.: The finite volume method. In: Ciarlet, P., Lions, J.L. (eds.) Handbook for Numerical Analysis, pp. 715–1022. North Holland, Amsterdam (2000)
5. Fisher, R.A.: On the dominance ratio. Proc. R. Soc. Edinb. **42**, 321–431 (1922)
6. Hössjer, O., Tyvand, P.A., Miloh, T.: Exact Markov chain and approximate diffusion solution for hapliod genetic drift with one-way mutation. Math. Biosci. **272**, 100–112 (2016)
7. Kimura, M.: Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harb. Symp. Quant. Biol. **20**, 33–53 (1955)
8. Kimura, M.: Random genetic drift in multi-allelic locus. Evolution **9**(4), 419–435 (1955)
9. Kimura, M.: On the probability of fixation of mutant genes in a population. Genetics **47**(6), 713–719 (1962)
10. Kimura, M.: Diffusion models in population genetics. J. Appl. Probab. **1**, 177–232 (1964)
11. Kimura, M.: The Neutral Theory of Molecular Evolution: A Review of Recent Evidence. Cambridge University Press, Cambridge (1983)
12. LeVeque, R.: Finite-Volume Methods for Hyperbolic Problems. Cambridge University Press, Cambridge (2002)
13. McKane, A.J., Waxman, D.: Sigular solution of the diffusion equation of population genetics. J. Theor. Biol. **247**, 849–858 (2007)
14. Moran, P.A.P.: Random processes in genetics. Proc. Camb. Philos. Soc. **54**, 60–72 (1958)
15. Roos, H.G., Stynes, M., Tobiska, L.: Numerical Methods for Singularly Perturbed Differential Equations. Springer, New York (1996)
16. Thomee, V.: Galerkin Finite Element Methods for Parabolic Problems. Springer Series in Computational Mathematics, vol. 25. Springer, New York (2006)
17. Tran, T.D., Hofrichter, J., Jost, J.: A general solution of the Wright-Fisher model of random genetic drift. Differ. Equ. Dyn. Syst. (2012). https://doi.org/10.1007/s12591-016-0289-7
18. Tran, T.D., Hofrichter, J., Jost, J.: An introduction to the mathematical structure of the Wright–Fisher model of population genetics. Theory Biosci. **132**, 73–82 (2013)
19. Waxman, D.: Fixation at a locus with multiple alleles: structure and solution of the Wright Fisher model. J. Theor. Biol. **257**, 245–251 (2009)
20. Wright, S.: The evolution of dominace. Am. Nat. **63**(689), 556–561 (1929)
21. Wright, S.: The differential equation of the distribution of gene frequencies. Proc. Natl. Acad. Sci. U.S.A. **31**, 382–389 (1945)
22. Zhao, L., Yue, X., Waxman, D.: Complete numerical solution of the diffusion equation of random genetic drift. Genetics **194**(4), 973–985 (2013)

Affiliations

Shixin Xu[1] · Minxin Chen[1] · Chun Liu[2] · Ran Zhang[3] · Xingye Yue[1]

✉ Xingye Yue
xyyue@suda.edu.cn

Shixin Xu
xsxztr@hotmail.com

Minxin Chen
chenminxin@suda.edu.cn

Chun Liu
cliu124@iit.edu

Ran Zhang
zhang.ran@mail.shufe.edu.cn

1 School of Mathematical Sciences, Soochow University, Suzhou 215006, China

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

3 School of Mathematics, Shanghai University of Finance and Economics, Shanghai 200433, China