An Asymptotic Preserving Scheme for Capturing Concentrations in Age-structured Models Arising in Adaptive Dynamics

Luís Almeida, Benoît Perthame, Xinran Ruan

AN ASYMPTOTIC PRESERVING SCHEME FOR CAPTURING CONCENTRATIONS IN AGE-STRUCTURED MODELS ARISING IN ADAPTIVE DYNAMICS*

LUIS ALMEIDA†, BENOIT PERTHAME†, AND XINRAN RUAN†

Abstract. We propose an asymptotic preserving (A-P) scheme for a population model structured by age and a phenotypical trait with or without mutation. As proved in [24], Dirac concentrations on particular phenotypical traits appear in the case without mutation, which makes the numerical resolution of the problem challenging. Inspired by its asymptotic behaviour, we apply a proper WKB representation of the solution to derive an A-P scheme, with which we can accurately capture the concentrations on a coarse, ε-independent mesh. The scheme is thoroughly analysed and important properties, including the A-P property, are rigorously proved. Furthermore, we observe nearly spectral accuracy in time in our numerical simulations. Next, we generalize the A-P scheme to the case with mutation, where a nonlinear Hamilton-Jacobi equation will be involved in the limiting model as $\varepsilon \to 0$. It can be formally shown that the generalized scheme is A-P as well, and numerical experiments indicate that we can still accurately resolve the problem on a coarse, ε-independent mesh in the phenotype space.

Key words. asymptotic preserving, Dirac concentration, age-structured population dynamics, phenotype, renewal equation, Hamilton-Jacobi equation, finite difference method

AMS subject classifications. 35F21, 35Q92, 65N06, 92D15

1. Introduction.

To describe the evolution of a phenotypical trait x in an age structured population model, a generic equation is

$$\begin{cases} \varepsilon\partial_t n_\varepsilon + \partial_a n_\varepsilon + d(a,x)n_\varepsilon = 0, \\ n_\varepsilon(t,a=0,x) = (1-m)\displaystyle\int_0^\infty b(a,x)n_\varepsilon\, da \\ \qquad\qquad + \dfrac{m}{\varepsilon}\displaystyle\int_0^\infty\int_0^\infty b(a,y)M\left(\dfrac{x-y}{\varepsilon}\right) n_\varepsilon(t,a,y)dady, \end{cases} \tag{1.1}$$

for $t \geq 0$, $a \geq 0$ and $x \geq 0$ and where $n(t,a,x)$ is the population density of individuals with age a and a phenotypical trait x at time t. The parameter $0 \leq m \leq 1$ represents the proportion of birth with mutations and the rate of change from the initial trait y to the inherited trait x is described by the probability density function $\frac{1}{\varepsilon}M(\frac{x-y}{\varepsilon})$. Here, $0 < \varepsilon \ll 1$ measures the effect of mutations on the phenotype.

We assume that M is a smooth probability density and that both the birth rate function $b(a,x)$ and the death rate function $d(a,x)$ are non-negative. Ageing, leads to assume that $b(a,x)$ is uniformly compactly supported in a and that there is a smallest value $a^* > 0$ such that

$$b(a,x) = 0, \quad \forall a \geq a^*, \quad \forall x \geq 0. \tag{1.2}$$

When the birth rate $b(a,x)$ is large enough, the population will grow exponentially. To better show the natural selection with the model, it is preferable to study

* **Funding:** BP has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 740623). This work was supported by INRIA Paris PRE 2017 MaCED

†Sorbonne Université, CNRS, Université de Paris, Inria, Laboratoire J.-L. Lions, F-75005 Paris, France. (luis.almeida@sorbonne-universite.fr, benoit.perthame@sorbonne-universite.fr, xinran.ruan@sorbonne-universite.fr).

the normalized population $\tilde{n}_\varepsilon(t,a,x)$ defined as

$$\tilde{n}_\varepsilon(t,a,x) := \frac{n_\varepsilon(t,a,x)}{\rho_\varepsilon(t)}, \qquad \rho_\varepsilon(t) = \int_0^\infty \int_0^\infty n_\varepsilon(t,a,x)\,dadx. \tag{1.3}$$

Then, one can expect that the fittest trait is selected which is expressed in terms of concentration of the population density at some moving point $\bar{x}(t)$, with a profile N, namely

$$\tilde{n}_\varepsilon(t,a,x) \to \delta(x-\bar{x}(t))N(a,\bar{x}(t)) \quad \text{as } \varepsilon \to 0^+, \tag{1.4}$$

assuming the population is monomorphic.

Because of this singular behaviour, the direct simulation of equation (1.1) requires expensive methods with an extremely fine mesh to capture the Dirac mass. Here, based on the asymptotics developed in [24], we propose a new numerical strategy which is Asymptotic Preserving, i.e., able to capture the singular limit $\varepsilon \to 0$ without refining the mesh.

Our purpose is to use the representation of solutions as a 'smooth' part multiplied by a singular part given by the solution of a Hamilton-Jacobi equation in order to compute accurately the concentration effect, as illustrated in (1.4), with a coarse grid which does not require any refining in order to be able to follow the Dirac mass. This question also appears in other types of asymptotic problems arising in kinetic equations, see [16].

This type of concentration is associated to the selection process of adaptive population dynamics. In this setting, $\bar{x}(t)$ represents the adaptive dynamics of a population which at the beginning is concentrated at a trait $\bar{x}(0)$ which might not be the fittest, and which will thus have to evolve to adapt to the environment (described by the birth rate $b(a,x)$ and death rate $d(a,x)$). The question has raised an important interest in the last decade after the initial introduction of the subject in mathematical biology (see [15, 12]). The population view has also been widely studied theoretically (see [4, 13, 11]). Several biologically-relevant examples have been treated, e.g. competitive interactions [18], the chemostat [5, 6] and the evolution of senescence [29].

The age structured equation (1.1) has also been used many times to model various phenomena in evolution biology (see for instance [32, 22, 24, 29]). In particular, it can be used to study how evolution affects the way cancer incidence depends on age (see, for instance, [14, 25]). In this setting, the trait x indicates the age beyond which the aggressiveness of the disease increases suddenly. Intuitively, the smaller the x is, the more dangerous the disease is. It was observed that, in nature, cancer mainly affects individuals beyond their reproductive age, which will correspond to age a^* in our formalism. A recent description of these issues and of the role of natural cancer prevention mechanisms in the transmission of germinally inherited cancer-causing mutant alleles is given in [2].

The paper is organized as follows. We first review the proper WKB representation of the solution in the case without mutations in Section 2. Then, we propose in Section 3 the detailed asymptotic preserving finite difference scheme based on the representation. In Section 4, we generalize the scheme to the case with mutations. In Section 5, we show numerical results to illustrate the efficiency of our method. Finally, conclusions are drawn in Section 6.

2. Case without mutation ($m = 0$). Our numerical strategy is based on the WKB representation for solutions of equation (1.1). We first present the simple

situation when $m = 0$. The dynamics of $\tilde{n}_\varepsilon(t, a, x)$ is governed by the equation

$$\begin{cases} \varepsilon\partial_t\tilde{n}_\varepsilon + \partial_a\tilde{n}_\varepsilon + d(a,x)\tilde{n}_\varepsilon = -\lambda_\varepsilon(t)\tilde{n}_\varepsilon, \\ \tilde{n}_\varepsilon(t, a = 0, x) = \displaystyle\int_0^\infty b(a,x)\tilde{n}_\varepsilon(t,a,x)\,da, \end{cases} \tag{2.1}$$

which is coupled with the dynamics of ρ_ε via

$$\lambda_\varepsilon(t) := \varepsilon\frac{\dot{\rho_\varepsilon}}{\rho_\varepsilon} = \int_0^\infty\int_0^\infty [b(a,x) - d(a,x)]\tilde{n}_\varepsilon(t,a,x)dadx. \tag{2.2}$$

It is obvious from (2.2) that

$$\inf_{a,x}\{b(a,x) - d(a,x)\} \le \lambda_\varepsilon(t) \le \sup_{a,x}\{b(a,x) - d(a,x)\}, \tag{2.3}$$

and the two bounds are independent of ε and t.

2.1. Spectral problem. It is useful and standard, as a preparation to the next steps, to introduce the following spectral problem. For each fixed x, we define the leading eigenvalue $\Lambda(x)$ and the corresponding normalized eigenfunction $N(a,x) > 0$ of the operator $\partial_a + d(a,x)$, with the same boundary condition. With assumption (1.2), thanks to an explicit representation, we may assume that we can define $(\Lambda(x), N(a,x))$ as

$$\begin{cases} \partial_a N(a,x) + d(a,x)N(a,x) = -\Lambda(x)N(a,x), \\ N(a=0,x) = \displaystyle\int_0^\infty b(a,x)N(a,x)\,da, \quad N(a,x) > 0, \quad \int_0^{+\infty} N(a,x)\,da = 1. \end{cases} \tag{2.4}$$

Also, we can define the dual eigenproblem of (2.4) as

$$\begin{cases} -\partial_a\Phi(a,x) + d(a,x)\Phi(a,x) = -\Lambda(x)\Phi(a,x) + b(a,x)\Phi(0,x), \\ \int_0^{+\infty}\Phi(a,x)N(a,x)\,da = 1, \quad \Phi(a,x) > 0. \end{cases} \tag{2.5}$$

The eigenvalue $\Lambda(x)$ only depends on values of $b(a,x)$, $d(a,x)$ for $0 < a < a^*$. Indeed, the eigenvalue Λ is defined by the relation

$$\int_0^{a^*} b(a,x)e^{-\left[\Lambda(x)a + \int_0^a d(a',x)\,da'\right]}da = 1. \tag{2.6}$$

The integrability condition for N means that, for a large, $d(a,x)$ is large enough compared to $-\Lambda(x)$. At least, we need that

$$d(a,x) > -\Lambda(x), \text{ for } a \text{ large}. \tag{2.7}$$

The above assumption is easy to be satisfied since $\Lambda(x)$ is independent of $d(a,x)$ for all $a > a^*$. It is also convenient, in order to avoid loss of mass at infinity in x, to assume that

$$\Lambda(x) < 0 \quad \text{for } x \text{ large}. \tag{2.8}$$

Furthermore, these problems have been widely studied. As shown in [27], the solutions N and Φ are bounded.

2.2. Asymptotic variable separation. The recent studies mentioned before show the asymptotic concentration of $\tilde{n}(t,a,x)$ at some $\bar{x}(t)$ as $\varepsilon \to 0$, a fact indicating the existence of a singularity according to (1.4). The approach initiated in [13], consists in writing $\tilde{n}(t,a,x)$ in the form

$$\tilde{n}_\varepsilon(t,a,x) = e^{\frac{u_\varepsilon(t,x)}{\varepsilon}} q_\varepsilon(t,a,x). \tag{2.9}$$

We assume this representation is true initially with both $q^0_\varepsilon(a,x)$ and $u^0_\varepsilon(x)$ being smooth and satisfying, for some $0 < \underline{\gamma}(x) < \overline{\gamma}(x)$,

$$0 < \underline{\gamma}(x) N(a,x) \le q^0_\varepsilon(a,x) \le \overline{\gamma}(x) N(a,x), \tag{2.10}$$

and

$$\int_0^\infty e^{\frac{u^0_\varepsilon(x)}{\varepsilon}}\, dx = 1, \qquad u^0_\varepsilon(x) \to u^0(x) \le 0 \quad \text{uniformly as } \varepsilon \to 0^+. \tag{2.11}$$

With additional technical assumptions, it implies that $\lim_{\varepsilon\to 0^+} e^{\frac{u^0_\varepsilon(x)}{\varepsilon}} = \delta(x-x_0)$ when $x_0 = \arg\max_x u^0(x)$ is unique. An example of assumption is that u^0_ε is uniformly concave in ε and x, see [21].

In the following, we establish dynamical equations for $q_\varepsilon(t,a,x)$ and $u_\varepsilon(t,x)$, respectively. We use the theory in [24] in order to get the exact equation for $u_\varepsilon(t,x)$, as $\varepsilon \to 0$, which deals with the singularity and is easy to solve. Then, it turns out that the information in a-direction, described by $q_\varepsilon(t,a,x)$, is fully regular.

Specifically, as it is standard [28, 24], we build $u_\varepsilon(t,x)$ as the solution of the equation

$$\partial_t u_\varepsilon = \Lambda(x) - \lambda_\varepsilon(t), \qquad u_\varepsilon(0,x) = u^0_\varepsilon(x) \tag{2.12}$$

where $\lambda_\varepsilon(t)$ is defined in (2.2). Then it is immediate that $q_\varepsilon(t,a,x)$ satisfies

$$\begin{cases} \varepsilon\partial_t q_\varepsilon + \partial_a q_\varepsilon + d(a,x) q_\varepsilon = -\Lambda(x) q_\varepsilon, \\ q_\varepsilon(t,a=0,x) = \displaystyle\int_0^\infty b(a,x) q_\varepsilon(t,a,x)\, da, \\ q_\varepsilon(t=0,a,x) = q^0_\varepsilon(a,x). \end{cases} \tag{2.13}$$

The dynamics of $u_\varepsilon(t,x)$ is coupled with the dynamics of $q_\varepsilon(t,a,x)$ via $\lambda_\varepsilon(t)$. In this setup, the following properties of $q_\varepsilon(t,a,x)$ and $u_\varepsilon(t,x)$ can be proved as in [24].

THEOREM 2.1. *Consider the two solutions $u_\varepsilon(t,x)$ and $q_\varepsilon(t,a,x)$ of (2.12)–(2.13), with initial constraints (2.10) and (2.11). Then, the following properties hold.*
(1) (Maximum principle of q_ε) For all $t \ge 0$, we have

$$0 < \underline{\gamma}(x) N(a,x) \le q_\varepsilon(t,a,x) \le \overline{\gamma}(x) N(a,x). \tag{2.14}$$

(2) (Conservation law) For all $t \ge 0$, we have

$$\int_0^{+\infty} q_\varepsilon(t,a,x)\Phi(a,x)\, da \equiv \int_0^{+\infty} q^0_\varepsilon(a,x)\Phi(a,x)\, da, \quad \forall x. \tag{2.15}$$

(3) (Limiting equations) We define, after extraction of a subsequence, $u(t,x) = \lim_{\varepsilon\to 0^+} u_\varepsilon(t,x)$ (uniform) and $q(t,a,x) = \lim_{\varepsilon\to 0^+} q_\varepsilon(t,a,x)$ (weak- limit). Then, the dynamics of $u(t,x)$ and $q(t,a,x)$ are governed by*

$$\partial_t u(t,x) = \Lambda(x) - \lambda(t), \tag{2.16}$$

where the function $\lambda(t)$ adapts automatically in such a way that $\max_x u(t,x) = 0$ for all $t \geq 0$, and

$$q(t,a,x) = \rho^0(x)N(a,x), \qquad \rho^0(x) := \int_0^{+\infty} q^0(a,x)\Phi(a,x)\,da. \tag{2.17}$$

As a remark, $\lim_{\varepsilon\to 0^+}\lambda_\varepsilon$ is the Lagrange multiplier associated with the constraint $\max_x u(t,x) = 0$, which is necessary for the unit mass condition $\int_0^\infty \int_0^\infty q_\varepsilon(t,a,x)da\, e^{\frac{u_\varepsilon(t,x)}{\varepsilon}}\,dx = 1$. Together with (2.17), we conclude that $\int_0^\infty q(t,a,x)da = \rho^0(x)$, and thus, for a measure $\mu(x,t)$, after extraction,

$$e^{\frac{u_\varepsilon(t,x)}{\varepsilon}} \to \mu \qquad \text{(weak limit), and } \int \rho^0(x)d\mu(x) = 1. \tag{2.18}$$

Proof. We simply recall roughly the main ideas of the proof because these are just variants of those in [24, 27, 23].
(1) The conclusion follows from the comparison principle of a transport equation.
(2) Multiplying both sides of (2.13) and (2.5) by Φ and q_ε, respectively, and subtracting, we get

$$\varepsilon\partial_t(q_\varepsilon\Phi) + \partial_a(q_\varepsilon\Phi) = -b(a,x)q_\varepsilon\Phi(0,x). \tag{2.19}$$

It follows by integrating the equation over a and noticing the boundary condition in (2.13) that

$$\frac{\partial}{\partial t}\int_0^{+\infty} q_\varepsilon(t,a,x)\Phi(a,x)\,da = 0 \tag{2.20}$$

holds true for any x and $t > 0$.
(3) Because λ_ε is bounded, u_ε converges uniformly after extraction. And q_ε is bounded noticing (2.14). Therefore we may extract a convergent subsequence as indicated. Next, we identify the limiting equation for $q(t,a,x)$, which by linearity satisfies

$$\begin{cases} \partial_a q + d(a,x)q = -\Lambda(x)q, \\ q(t,a=0,x) = \displaystyle\int_0^\infty b(a,x)q(t,a,x)\,da. \end{cases}$$

Because the dominant eigenvalue in problem (2.4) is simple, we find that $q(t,a,x) = \rho(t,x)N(a,x)$ for some function $\rho(t,x)$ which is independent of a. Then, we may pass to the limit in (2.15) and find

$$\int_0^{+\infty} q^0(a,x)\Phi(a,x)\,da \equiv \int_0^{+\infty} q(t,a,x)\Phi(a,x)\,da = \rho(t,x)\int_0^{+\infty} N(a,x)\Phi(a,x)\,da = \rho(t,x).$$

Noticing that $\int_0^{+\infty} e^{\frac{u_\varepsilon(t,x)}{\varepsilon}}\,dx$ is uniformly bounded with respect to ε since the following holds,

$$0 < \underline{\gamma}(x)\int_0^{+\infty} e^{\frac{u_\varepsilon(t,x)}{\varepsilon}}\,dx \leq 1 = \int_0^\infty\int_0^\infty q(t,a,x)e^{\frac{u_\varepsilon(t,x)}{\varepsilon}}\,dadx \leq \overline{\gamma}(x)\int_0^{+\infty} e^{\frac{u_\varepsilon(t,x)}{\varepsilon}}\,dx, \tag{2.21}$$

by (2.4), (2.14) and the normalization condition. As a result, we must have

$$\max_x\{u(t,x)\} = 0 \qquad \text{for any } t > 0, \tag{2.22}$$

which leads to the limiting equation (2.16) for $u(t,x)$. ∎

Remark 2.2. Notice that $\rho^0(x)N(a,x)e^{\frac{u_\varepsilon(t,x)}{\varepsilon}}$ are exact solutions of (2.1). Theorem 2.1 implies immediately that, these solutions attract all solutions.

Remark 2.3. The question to know if the weak limit (in measures) of $e^{\frac{u_\varepsilon(t,x)}{\varepsilon}}$ is a multiple of a Dirac mass $\delta(x-\bar{x}(t))$ is related to know if the maximum in (2.22) is unique. This holds if both u_ε^0 and Λ are strictly concave (then $u(t,\cdot)$ is) [21]. This also holds if $\Lambda(x)$ is monotonic [3].

3. Finite difference discretization: Case without mutation. Based on the dynamic equations (2.13) and (2.12) for q_ε and u_ε, respectively, we show the detailed numerical scheme discretized via finite difference method, which generates a discretized solution of (2.1), compatible with the limit $\varepsilon \to 0$.

3.1. Notations. For simplicity of notations, we introduce $[\![K_1, K_2]\!] := [K_1, K_2] \cap \mathbb{Z}$ and define δ_x^- and δ_x^+ to be the backward and forward finite difference operators approximating ∂_x, respectively. Similarly, we can define δ_a^-, δ_a^+ and δ_t^+. Let us take $\Omega = (0,\infty) \times (0,M)^2$ to be the computational domain of (t,a,x) and denote the uniformly distributed grid points as

$$t_n = n\Delta t, \quad a_j = j\Delta a, \quad x_k = k\Delta x, \tag{3.1}$$

for $n \in \mathbb{N} \cup \{0\}$, $j \in [\![0, K_a]\!]$ and $k \in [\![0, K_x]\!]$, with $\Delta a = \frac{M}{K_a}$ and $\Delta x = \frac{M}{K_x}$ being the mesh sizes in a- and x-direction, respectively. Let u_k^n and $q_{j,k}^n$ be the corresponding numerical approximations of $u_\varepsilon(t_n, x_k)$ and $q_\varepsilon(t_n, a_j, x_k)$, and denote

$$\mathbf{u}^n = (u_k^n)^T \in \mathbb{R}^{K_x+1}, \quad Q^n = (q_{j,k}^n) \in \mathbb{R}^{(K_a+1)\times(K_x+1)}. \tag{3.2}$$

Then $\tilde{n}_{j,k}^n$, which is defined as

$$\tilde{n}_{j,k}^n = q_{j,k}^n e^{\frac{u_k^n}{\varepsilon}} \tag{3.3}$$

for all $n \geq 0, j \in [\![0, K_a]\!], k \in [\![0, K_x]\!]$, is the numerical approximation of the normalized population $\tilde{n}(t_n, a_j, x_k)$. Collecting all elements $\tilde{n}_{j,k}^n$ in a matrix, we get

$$\tilde{N}^n = (\tilde{n}_{j,k}^n) \in \mathbb{R}^{(K_a+1)\times(K_x+1)}. \tag{3.4}$$

Here we introduce $I_1(\cdot)$ and $I_2(\cdot)$ to be the finite difference approximations of the 1D integral in x-direction and the 2D integral in both a- and x-directions, respectively,

$$I_1(\mathbf{u}) := \Delta x \sum_{k=0}^{K_x} w_k^{(x)} u_k, \quad I_2(Q) = \Delta a \Delta x \sum_{j=1}^{K_a} \sum_{k=0}^{K_x} w_k^{(x)} q_{j,k}, \tag{3.5}$$

with the weights

$$w_k^{(x)} = \frac{2-\delta_{k,0}}{2} = \begin{cases} \frac{1}{2}, & \text{for } k = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{3.6}$$

For each $k \in [\![0, K_x]\!]$, we consider the discretized eigenvalue problem of (2.4) as

$$\begin{cases} \delta_a^- N_{j,k} + d(a_j, x_k) N_{j,k} = -\Lambda_k N_{j,k}, & j \in [\![1, K_a]\!], \\ N_{0,k} = \Delta a \sum_{j=1}^{K_a} b(a_j, x_k) N_{j,k}, \text{ and } \Delta a \sum_{j=0}^{K_a} w_j^{(a)} N_{j,k} = 1, \end{cases} \tag{3.7}$$

where $w_j^{(a)} = \frac{2-\delta_{j,0}}{2}$, Λ_k and $N_{j,k}$ are the numerical approximations of $\Lambda(x_k)$ and $N(a_j, x_k)$, respectively. With the matrix $M_k = (m_{i,j}^{(k)}) \in \mathbb{R}^{K_a \times K_a}$ defined in (A.1) and denoting

$$\mathbf{n}_k := (N_{1,k}, N_{2,k}, \ldots, N_{K_a,k})^T \in \mathbb{R}^{K_a}, \tag{3.8}$$

the equation (3.7) can then be formulated in a more compact form as

$$M_k \mathbf{n}_k = -\Lambda_k \mathbf{n}_k, \tag{3.9}$$

and $-\Lambda_k$ is the leading eigenvalue of the matrix M_k. By the way, it would be convenient for later use to introduce the following notations

$$\boldsymbol{\lambda} = (\Lambda_k)^T \in \mathbb{R}^{K_x+1}, \quad N = (N_{j,k}) \in \mathbb{R}^{(K_a+1)\times(K_x+1)}. \tag{3.10}$$

Noticing that $m_{i,j}^{(k)} \le 0$ for all $i \neq j$, M_k, which is closely related to an M-matrix, has a positive left eigenvector $\boldsymbol{\phi}_k := (\phi_{0,k}, \phi_{1,k}, \ldots, \phi_{K_a-1,k})^T \in \mathbb{R}^{K_a}$ corresponding to the same eigenvalue $-\Lambda_k$ by the Perron-Frobenius theorem. In other words, the eigenpair $(\Lambda_k, \boldsymbol{\phi}_k)$ is the dual eigenpair of $(\Lambda_k, \mathbf{n}_k)$ and satisfies

$$\boldsymbol{\phi}_k^T M_k = -\Lambda_k \boldsymbol{\phi}_k^T, \tag{3.11}$$

or more precisely

$$\begin{cases} -\delta_a^+ \phi_{j-1,k} + d(a_j, x_k)\phi_{j-1,k} = -\Lambda_k \phi_{j-1,k} + b(a_j, x_k)\phi_{0,k}, \quad j \in [\![1, K_a - 1]\!] \\ \Delta a \sum_{j=1}^{K_a} N_{j,k}\phi_{j-1,k} = 1, \quad \phi_{K_a,k} = 0, \end{cases} \tag{3.12}$$

where the boundary condition is artificial to make the eigenvector unique. Obviously, the problem (3.12) approximates the problem (2.5) in a discrete form.

As a remark, instead of solving the eigenvalue problem directly, the leading eigenvalues can be easily computed by solving the following nonlinear equation

$$1 = \Delta a \sum_{j=1}^{K_a} b(a_j, x_k) \prod_{s=1}^{j} \frac{1}{1 + \Delta a(d(a_s, x_k) + \Lambda_k)}, \quad k \in [\![0, K_x]\!], \tag{3.13}$$

which has a unique solution in the feasible set

$$\Omega_k := \{\Lambda_k \in \mathbb{R} \,|\, 1 + \Delta a(d(a_s, x_k) + \Lambda_k) > 0\}. \tag{3.14}$$

Furthermore, the corresponding eigenvectors $\mathbf{n}_k$ and $\boldsymbol{\phi}_k$ can be computed explicitly. The derivation of the equation (3.13) and the detailed formula of $\mathbf{n}_k$ and $\boldsymbol{\phi}_k$ can be referred to Appendix B.

3.2. Numerical scheme. With the notations, the scheme works as follows from t_n to t_{n+1}.

Step 1: Update $Q^n := (q_{j,k}^n) \in \mathbb{R}^{(K_a+1)\times(K_x+1)}$ *based on the discretization of* (2.13), i.e.

$$\begin{cases} \varepsilon \frac{q_{j,k}^{n+1} - q_{j,k}^n}{\Delta t} + \delta_a^- q_{j,k}^{n+1} + d(a_j, x_k) q_{j,k}^{n+1} = -\Lambda_k q_{j,k}^{n+1}, \quad j \in [\![1, K_a]\!], \\ q_{0,k}^{n+1} = \Delta a \sum_{j=1}^{K_a} b(a_j, x_k) q_{j,k}^{n+1}. \end{cases} \tag{3.15}$$

It is worth noticing that, in (3.15), the stiff term must be treated implicitly to ensure stability.
Step 2: Update $\mathbf{u}^n := (u_0^n, u_1^n, \dots, u_{K_x}^n)^T$ *based on the discretization of* (2.12), i.e.

$$\frac{u_k^{n+1} - u_k^n}{\Delta t} = \Lambda_k - L^n, \quad k \in [\![0, K_x]\!], \tag{3.16}$$

where L^n is chosen such that

$$I_2(\tilde{N}^{n+1}) = \Delta a \Delta x \sum_{j=1}^{K_a} \sum_{k=0}^{K_x} w_k^{(x)} q_{j,k}^{n+1} e^{\frac{u_k^{n+1}}{\varepsilon}} = 1. \tag{3.17}$$

To compute effectively L^n, we set

$$\mathbf{v}^n := (v_k^n)^T \in \mathbb{R}^{K_x+1}, \qquad v_k^n = e^{\frac{u_k^n + \Delta t \Lambda_k}{\varepsilon}}, \quad k \in [\![0, K_x]\!],\ n \geq 0, \tag{3.18}$$

then

$$L^n = \frac{\varepsilon}{\Delta t} \ln \left(I_2(Q^{n+1} \circ \mathbf{v}^n) \right), \tag{3.19}$$

where the operator $\circ$ denotes the generalized Hadamard product between a matrix and a vector

$$Q^{n+1} \circ \mathbf{v}^n := (q_{j,k}^{n+1} \cdot v_k^n) = \begin{bmatrix} q_{0,0}^{n+1} v_0^n & q_{0,1}^{n+1} v_1^n & \cdots & q_{0,K_x}^{n+1} v_{K_x}^n \\ q_{1,0}^{n+1} v_0^n & q_{1,1}^{n+1} v_1^n & \cdots & q_{1,K_x}^{n+1} v_{K_x}^n \\ \vdots & \vdots & \ddots & \vdots \\ q_{K_a,0}^{n+1} v_0^n & q_{K_a,1}^{n+1} v_1^n & \cdots & q_{K_a,K_x}^{n+1} v_{K_x}^n \end{bmatrix}. \tag{3.20}$$

Then the scheme (3.16) can be rewritten as

$$u_k^{n+1} = u_k^n + \Delta t \Lambda_k - \varepsilon \ln \left(I_2(Q^{n+1} \circ \mathbf{v}^n) \right) \tag{3.21}$$

Remark 3.1. In practical computation, we need to normalize the vector $\mathbf{v}^n$ for stability issues when ε is small. Noticing that

$$\varepsilon \ln \left(\|\mathbf{v}^n\|_{l^\infty} \right) = \|\mathbf{u}^n + \Delta t \boldsymbol{\lambda}\|_{l^\infty}, \tag{3.22}$$

we can reformulate the scheme (3.16) to be

$$u_k^{n+1} = u_k^n + \Delta t \Lambda_k - \|\mathbf{u}^n + \Delta t \boldsymbol{\lambda}\|_{l^\infty} - \varepsilon \ln \left(I_2(Q^{n+1} \circ \tilde{\mathbf{v}}^n) \right), \tag{3.23}$$

where $\tilde{\mathbf{v}}^n := \frac{\mathbf{v}^n}{\|\mathbf{v}^n\|_{l^\infty}}$ is normalized in the sense $\|\tilde{\mathbf{v}}^n\|_{l^\infty} = 1$.

3.3. Theoretical properties. As an analogous to Theorem 2.1 for the continuous case, we can show the corresponding properties for the discrete case in Theorem 3.2 and Theorem 3.6.

THEOREM 3.2. *Taken* $\mathbf{u}^n$, Q^n *and* N *defined in* (3.2) *and* (3.10). *We assume that the initial data* Q^0 *satisfies the constraint*

$$0 < \underline{\gamma_k} N_{j,k} \leq q_{j,k}^0 \leq \overline{\gamma_k} N_{j,k} \tag{3.24}$$

for any $j \in [\![0, K_a]\!]$ and $k \in [\![0, K_x]\!]$, and update $\mathbf{u}^n$ and Q^n via (3.23) and (3.15), respectively. Then the following properties hold.
(1) (Maximum principle of Q^n) For any $n \geq 0$, we have

$$0 < \underline{\gamma_k} N_{j,k} \leq q_{j,k}^n \leq \overline{\gamma_k} N_{j,k}, \tag{3.25}$$

where $j \in [\![0, K_a]\!]$ and $k \in [\![0, K_x]\!]$.
(2) (Discrete conservation law) Define $F_k^n = \Delta a \sum_{j=1}^{K_a} q_{j,k}^n \phi_{j-1,k}$ for each $k \in [\![0, K_x]\!]$ and $n \geq 0$, where $\phi_{j,k}$ is defined in (3.12). Then we have

$$F_k^n \equiv F_k^0. \tag{3.26}$$

Proof. (1) We prove (3.25) by induction. By assumption, the conclusion (3.25) holds true at $n = 0$. As a result, we only need to show that (3.25) holds true at $t = t_{n+1}$, assuming that it holds true at $t = t_n$. The proof is based on the discrete entropy inequality generalized from [27]. For simplicity of notations, we reformulate the scheme (3.15) to be

$$\varepsilon \frac{\mathbf{q}_k^{n+1} - \mathbf{q}_k^n}{\Delta t} = \tilde{M}_k \mathbf{q}_k^{n+1}, \tag{3.27}$$

where $\tilde{M}_k = (\tilde{m}_{i,j}^{(k)}) \in \mathbb{R}^{K_a \times K_a}$ is defined to be $-M_k - \Lambda_k I$ with the matrix M_k defined in (A.1), and $\mathbf{q}_k^n := (q_{1,k}^n, q_{2,k}^n, \ldots, q_{K_a,k}^n)^T$. It is then immediate that $\tilde{M}_k \mathbf{n}_k = \mathbf{0}$ and $\boldsymbol{\phi}_k^T \tilde{M}_k = \mathbf{0}^T$ and all off-diagonal elements of $\tilde{M}_k$ are nonnegative, i.e. $\tilde{m}_{i,j} \geq 0$ if $i \neq j$. Define the discrete form of the **general relative entropy** to be

$$\sum_{j=1}^{K_a} \phi_{j-1,k} N_{j,k} H(\frac{q_{j,k}^n}{N_{j,k}}), \tag{3.28}$$

where $H(\cdot)$ is an arbitrary convex function. Then we claim that

$$\sum_{j=1}^{K_a} \phi_{j-1,k} N_{j,k} H(\frac{q_{j,k}^{n+1}}{N_{j,k}}) \leq \sum_{j=1}^{K_a} \phi_{j-1,k} N_{j,k} H(\frac{q_{j,k}^n}{N_{j,k}}). \tag{3.29}$$

In fact, a direct calculation shows that

$$\begin{aligned}
&\sum_{j=1}^{K_a} \phi_{j-1,k} N_{j,k} \left(H(\frac{q_{j,k}^{n+1}}{N_{j,k}}) - H(\frac{q_{j,k}^n}{N_{j,k}}) \right) \leq \sum_{j=1}^{K_a} \phi_{j-1,k} H'(\frac{q_{j,k}^{n+1}}{N_{j,k}}) \left(q_{j,k}^{n+1} - q_{j,k}^n \right) \\
=&\frac{\Delta t}{\varepsilon} \sum_{i,j=1}^{K_a} \phi_{j-1,k} H'(\frac{q_{j,k}^{n+1}}{N_{j,k}}) \tilde{m}_{j,i} q_{i,k}^{n+1} = \frac{\Delta t}{\varepsilon} \sum_{i,j=1}^{K_a} \phi_{j-1,k} \tilde{m}_{j,i} N_{i,k} H'(\frac{q_{j,k}^{n+1}}{N_{j,k}}) \left(\frac{q_{i,k}^{n+1}}{N_{i,k}} - \frac{q_{j,k}^{n+1}}{N_{j,k}} \right) \\
=&\frac{\Delta t}{\varepsilon} \sum_{i,j=1}^{K_a} \phi_{j-1,k} \tilde{m}_{j,i} N_{i,k} \left[H'(\frac{q_{j,k}^{n+1}}{N_{j,k}}) \left(\frac{q_{i,k}^{n+1}}{N_{i,k}} - \frac{q_{j,k}^{n+1}}{N_{j,k}} \right) + H(\frac{q_{j,k}^{n+1}}{N_{j,k}}) - H(\frac{q_{i,k}^{n+1}}{N_{i,k}}) \right] \leq 0,
\end{aligned}$$

where the equalities $\tilde{M}_k \mathbf{n}_k = \mathbf{0}$ and $\boldsymbol{\phi}_k^T \tilde{M}_k = \mathbf{0}^T$ are applied in the second and the third equation and last inequality holds true due to the convexity of $H(\cdot)$ and the fact that all diagonal terms are 0.

The upper bound of $q_{j,k}^{n+1}$ with $j \in [\![1, K_a]\!]$ can then be proven by taking

$$H(u) = (u - \overline{\gamma_k})_+^2 \geq 0. \tag{3.30}$$

By assumptions, we have (3.25) hold true at $t = t_n$, which implies that

$$\sum_{j=1}^{K_a} \phi_{j-1,k} N_{j,k} H\left(\frac{q_{j,k}^n}{N_{j,k}}\right) = 0. \tag{3.31}$$

Combining (3.31) with the entropy inequality (3.29), it is then obvious that

$$\sum_{j=1}^{K_a} \phi_{j-1,k} N_{j,k} H\left(\frac{q_{j,k}^{n+1}}{N_{j,k}}\right) \leq 0, \tag{3.32}$$

which immediately leads to the conclusion $q_{j,k}^{n+1} \leq \overline{\gamma_k} N_{j,k}$ noticing the fact that $\phi_{j-1,k} > 0$, $N_{j,k} > 0$ and $H(\cdot) \geq 0$. Similarly, we can prove the lower bound at $t = t_{n+1}$. The boundedness of the boundary terms $q_{0,k}^{n+1}$ comes from the boundary conditions in (3.7) and (3.15). As a conclusion, we have (3.25) hold true at $t = t_{n+1}$. Therefore, by induction, we proved the maximum principle (3.25) for all $n \geq 0$.

(2) For each $k \in [\![0, K_x]\!]$, multiplying both sides of (3.15) by $\phi_{j-1,k}$ and summing over $j \in [\![1, K_a]\!]$, we get

$$F_k^{n+1} - F_k^n = \frac{\Delta t}{\varepsilon} \sum_{j=1}^{K_a} \phi_{j-1,k}(-\delta_a^- q_{j,k}^{n+1} - d(a_j, x_k) q_{j,k}^{n+1} - \Lambda_k q_{j,k}^{n+1}). \tag{3.33}$$

Noticing the boundary condition in (3.15) and the fact that $\phi_{K_a,k} = 0$, we have via summation by part that

$$\begin{aligned} \sum_{j=1}^{K_a} \phi_{j-1,k} \delta_a^- q_{j,k}^{n+1} &= -\sum_{j=1}^{K_a} q_{j,k}^{n+1} \delta_a^+ \phi_{j-1,k} - \phi_{0,k} \left(\sum_{j=1}^{K_a} b(a_j, x_k) q_{j,k}^{n+1} \right) \\ &= -\sum_{j=1}^{K_a} q_{j,k}^{n+1} (\delta_a^+ \phi_{j-1,k} + b(a_j, x_k) \phi_{0,k}). \end{aligned} \tag{3.34}$$

Substituting (3.34) into (3.33), we get

$$F_k^{n+1} - F_k^n = \frac{\Delta t}{\varepsilon} \sum_{j=1}^{K_a} q_{j,k}^{n+1} (\delta_a^+ \phi_{j-1,k} + b(a_j, x_k) \phi_{0,k} - d(a_j, x_k) \phi_{j-1,k} - \Lambda_k \phi_{j-1,k}) = 0, \tag{3.35}$$

where the last step is due to (3.12). The conclusion then follows directly from (3.35).□

Remark 3.3. Analogous to Remark 2.2, when we choose the initial data to be $Q^0 = N$, we have $Q^n = N$ for all $n > 0$. All the dynamics is carried by the singular part u_k^n.

Remark 3.4. A scaling of Q^n is recommended at each step to force (3.26) holds exactly. Otherwise, noticing (3.35), the round-off error will accumulate and be no longer negligible when time is large or $0 < \varepsilon \ll 1$.

It will be shown in Proposition 3.5 that the normalization constant L^n in the scheme is uniformly bounded with whatever choice of $\varepsilon, \Delta t$ and Δx. The fact implies that our scheme will be robust not only for normal cases, but also for the limiting case $0 < \varepsilon \ll 1$, where a concentration of the normalized population $\tilde{n}(t, a, x)$ appears.

PROPOSITION 3.5. *There are two constants $\underline{L}$ and $\overline{L}$, which are independent of $\varepsilon, \Delta t$ and Δx, such that, for any $n \geq 0$,*

$$\underline{L} \leq L^n \leq \overline{L}. \tag{3.36}$$

Proof. For simplicity of notations, we introduce

$$\overline{\Lambda} = \max_{k\in[\![0,Kx]\!]}\{\Lambda_k\},\ \underline{\Lambda} = \min_{k\in[\![0,Kx]\!]}\{\Lambda_k\},\ \overline{\gamma} = \max_{k\in[\![0,Kx]\!]}\{\overline{\gamma_k}\},\ \underline{\gamma} = \min_{k\in[\![0,Kx]\!]}\{\underline{\gamma_k}\}. \tag{3.37}$$

Noticing that Λ_k is the numerical approximation of $\Lambda(x_k)$ after discretization in a-direction, $\overline{\Lambda}$ and $\underline{\Lambda}$ are independent of $\varepsilon, \Delta t$ and Δx. For each element v_k^n in $\mathbf{v}^n$ (3.18), we have

$$e^{\frac{\Delta t}{\varepsilon}\underline{\Lambda}} e^{\frac{u_k^n}{\varepsilon}} \leq v_k^n \leq e^{\frac{\Delta t}{\varepsilon}\overline{\Lambda}} e^{\frac{u_k^n}{\varepsilon}}, \tag{3.38}$$

which, combining together with (3.19), implies that

$$\underline{\Lambda} + \frac{\varepsilon}{\Delta t}\ln(I_2(Q^{n+1}\circ e^{\frac{\mathbf{u}^n}{\varepsilon}})) \leq L^n \leq \overline{\Lambda} + \frac{\varepsilon}{\Delta t}\ln(I_2(Q^{n+1}\circ e^{\frac{\mathbf{u}^n}{\varepsilon}})), \tag{3.39}$$

where $I_2(\cdot)$ is the second order numerical quadrature defined in (3.5) and the operator $\circ$ denotes the generalized Hadamard product defined in (3.20). Besides, the maximum principle of Q^n proved in Theorem 3.2 indicates that

$$\frac{\underline{\gamma}}{\overline{\gamma}} q_{j,k}^n \leq q_{j,k}^{n+1} \leq \frac{\overline{\gamma}}{\underline{\gamma}} q_{j,k}^n \tag{3.40}$$

holds true for all $n \geq 0$, $j \in [\![0, K_a]\!]$ and $k \in [\![0, K_x]\!]$. With all these preparations, now we are ready to derive an explicit upper bound and lower bound of L^n. The following two cases will be considered separately.

- Case I: $\varepsilon/\Delta t \geq C$,
- Case II: $\varepsilon/\Delta t < C$,

where the constant C can be chosen arbitrarily as long as

$$C \geq 2\frac{\overline{\gamma}}{\underline{\gamma}} \max_{j,k}\{|b(a_j,x_k) - d(a_j,x_k) - \Lambda_k|\}. \tag{3.41}$$

(I) When $\varepsilon/\Delta t \geq C$, recalling (3.15), we have that

$$\begin{aligned} I_2(Q^{n+1}\circ e^{\frac{\mathbf{u}^n}{\varepsilon}}) &= \Delta a\Delta x \sum_{j=1}^{K_a}\sum_{k=0}^{K_x} w_k^{(x)} q_{j,k}^{n+1} e^{\frac{u_k^n}{\varepsilon}} \\ &= \Delta a\Delta x \sum_{j=1}^{K_a}\sum_{k=0}^{K_x} w_k^{(x)} \left[q_{j,k}^n e^{\frac{u_k^n}{\varepsilon}} + \frac{\Delta t}{\varepsilon}(-\delta_a^- q_{j,k}^{n+1} - d(a_j,x_k) q_{j,k}^{n+1} - \Lambda_k q_{j,k}^{n+1}) e^{\frac{u_k^n}{\varepsilon}} \right] \\ &= \Delta a\Delta x \sum_{j=1}^{K_a}\sum_{k=0}^{K_x} w_k^{(x)} \left[q_{j,k}^n e^{\frac{u_k^n}{\varepsilon}} + \frac{\Delta t}{\varepsilon}(b(a_j,x_k) - d(a_j,x_k) - \Lambda_k) q_{j,k}^{n+1} e^{\frac{u_k^n}{\varepsilon}} \right], \end{aligned} \tag{3.42}$$

where the last step is due to the boundary condition in (3.15). Therefore, on one hand,

$$\begin{aligned} I_2(Q^{n+1}\circ e^{\frac{\mathbf{u}^n}{\varepsilon}}) &\geq \Delta a\Delta x \sum_{j=1}^{K_a}\sum_{k=0}^{K_x} w_k^{(x)} \left[1 - \frac{\overline{\gamma}}{\underline{\gamma}}\frac{\Delta t}{\varepsilon}|b(a_j,x_k) - d(a_j,x_k) - \Lambda_k| \right] q_{j,k}^n e^{\frac{u_k^n}{\varepsilon}} \\ &\geq \left[1 - \frac{\Delta t}{\varepsilon}\frac{C}{2} \right] I_2(\tilde{N}^n) = 1 - \frac{C\Delta t}{2\varepsilon}, \end{aligned} \tag{3.43}$$

where $\tilde{N}^n$ is defined in (3.4). Noticing the relation in (3.39) and the fact that $\ln(1-x) \geq -2x$ holds true for all $x \in (0, \frac{1}{2})$ and $\frac{C\Delta t}{2\varepsilon} \leq \frac{1}{2}$, we have that

$$L^n \geq \underline{\Lambda} + \frac{\varepsilon}{\Delta t} \ln(I_2(Q^{n+1} \circ e^{\frac{\mathbf{u}^n}{\varepsilon}})) \geq \underline{\Lambda} + \frac{\varepsilon}{\Delta t} \ln(1 - \frac{C\Delta t}{2\varepsilon}) \geq \underline{\Lambda} - C. \tag{3.44}$$

On the other hand,

$$\begin{aligned} I_2(Q^{n+1} \circ e^{\frac{\mathbf{u}^n}{\varepsilon}}) &\leq \Delta a \Delta x \sum_{j=1}^{K_a} \sum_{k=0}^{K_x} w_k^{(x)} \left[1 + \frac{\overline{\gamma}}{\underline{\gamma}} \frac{\Delta t}{\varepsilon} |b(a_j, x_k) - d(a_j, x_k) - \Lambda_k|\right] q_{j,k}^n e^{\frac{u_k^n}{\varepsilon}} \\ &\leq \left[1 + \frac{\Delta t}{\varepsilon} \frac{C}{2}\right] I_2(\tilde{N}^n) = 1 + \frac{C\Delta t}{2\varepsilon}. \end{aligned} \tag{3.45}$$

Similarly, combining the relation in (3.39) and the fact that $\ln(1+x) \leq x$ holds true for all $x \in \mathbb{R}^+$, we have that

$$L^n \leq \overline{\Lambda} + \frac{\varepsilon}{\Delta t} \ln\left(1 + \frac{C\Delta t}{2\varepsilon}\right) \leq \overline{\Lambda} + \frac{C}{2}. \tag{3.46}$$

To summarize, we have, in this case, that

$$\underline{\Lambda} - C \leq L^n \leq \overline{\Lambda} + \frac{C}{2}. \tag{3.47}$$

(II) When $\frac{\varepsilon}{\Delta t} \leq C$, recalling the relation (3.40) and the fact that $I_2(\tilde{N}^n) = I_2(Q^n \circ e^{\frac{\mathbf{u}^n}{\varepsilon}}) = 1$, we have that

$$\frac{\underline{\gamma}}{\overline{\gamma}} = \frac{\underline{\gamma}}{\overline{\gamma}} I_2(Q^n \circ e^{\frac{\mathbf{u}^n}{\varepsilon}}) \leq I_2(Q^{n+1} \circ e^{\frac{\mathbf{u}^n}{\varepsilon}}) \leq \frac{\overline{\gamma}}{\underline{\gamma}} I_2(Q^n \circ e^{\frac{\mathbf{u}^n}{\varepsilon}}) = \frac{\overline{\gamma}}{\underline{\gamma}}, \tag{3.48}$$

which immediately implies that

$$\underline{\Lambda} - C \ln\left(\frac{\overline{\gamma}}{\underline{\gamma}}\right) \leq L^n \leq \overline{\Lambda} + C \ln\left(\frac{\overline{\gamma}}{\underline{\gamma}}\right). \tag{3.49}$$

Combining the two cases, we get that there exist two constants $\overline{L}$ and $\underline{L}$ such that

$$\underline{L} \leq L^n \leq \overline{L}, \tag{3.50}$$

where the two constants can be chosen to be

$$\underline{L} = \underline{\Lambda} - C \max\left\{\ln(\frac{\overline{\gamma}}{\underline{\gamma}}), 1\right\}, \quad \overline{L} = \overline{\Lambda} + C \max\left\{\ln(\frac{\overline{\gamma}}{\underline{\gamma}}), \frac{1}{2}\right\}. \tag{3.51}$$

∎

3.4. Asymptotic Preserving property. The asymptotic preserving (A-P) schemes are extremely powerful tools as they permit the use of the same scheme to discretize a perturbation problem and its limit problem, with fixed discretization parameters [1]. Here we will show that the schemes (3.23) and (3.15) are indeed A-P schemes.

THEOREM 3.6. *(Asymptotic preserving) When the discretization parameters Δa, Δx and Δt are fixed and $0 < \underline{\gamma_k} N_{j,k} \leq Q_{j,k}^0 \leq \overline{\gamma_k} N_{j,k}$ holds true for any $j \in [\![0, K_a]\!]$*

and $k \in [\![0, K_x]\!]$, then we have
(1) the scheme (3.23) tends to the following scheme as ε goes to 0

$$\frac{u_k^{n+1} - u_k^n}{\Delta t} = \Lambda_k - L^n, \tag{3.52}$$

where L^n is chosen such that $\max_k u_k^{n+1} = 0$.
(2) the scheme (3.15) tends to the following scheme as ε goes to 0 for any $k \in [\![0, K_x]\!]$,

$$\begin{cases} \delta_a^- q_{j,k}^{n+1} + d(a_j, x_k) q_{j,k}^{n+1} = -\Lambda_k q_{j,k}^{n+1}, \quad j \in [\![1, K_a]\!], \\ q_{0,k}^{n+1} = \Delta a \sum_{j=0}^{K_a} b(a_j, x_k) q_{j,k}^{n+1}, \end{cases} \tag{3.53}$$

which is identical to (3.7) and implies that

$$q_{j,k}^n = \rho_k^0 N_{j,k}, \tag{3.54}$$

where $\rho_k^0 = \Delta a \sum_{j=1}^{K_a} q_{j,k}^0 \phi_{j-1,k}$.

Proof. (1) Recalling the scheme (3.23), it is sufficient to show that

$$\lim_{\varepsilon \to 0} \varepsilon \ln \left(I_2(Q^{n+1} \circ \tilde{\mathbf{v}}^n) \right) = 0. \tag{3.55}$$

On one hand, noticing the fact that $\|\tilde{\mathbf{v}}^n\|_{l^\infty} \leq 1$ and the maximum principle of Q^n in Theorem 3.2, we have that

$$I_2(Q^{n+1} \circ \tilde{\mathbf{v}}^n) \leq \max_k \{\overline{\gamma_k}\} I_2(N). \tag{3.56}$$

On the other hand, since both Q^{n+1} and $\tilde{\mathbf{v}}^n$ are nonnegative, we have

$$I_2(Q^{n+1} \circ \tilde{\mathbf{v}}^n) \geq 0. \tag{3.57}$$

Since both the upper bound and the lower bound of $I_2(Q^{n+1} \circ \tilde{\mathbf{v}}^n)$ are independent of ε, the limit (3.55) is then obvious. The limiting scheme (3.52) follows directly by combining (3.23) and (3.55).

(2) The maximum principle of Q^n in Theorem 3.2 shows that Q^{n+1} is uniformly bounded for all $n \geq 0$. Therefore, it is obvious that $\lim_{\varepsilon \to 0} \varepsilon \delta_t^+ q_{j,k}^n = 0$ for any $j \in [\![0, K_a]\!]$, $k \in [\![0, K_x]\!]$ and $n \geq 0$. The limiting scheme (3.53) then follows directly. Since the limiting scheme (3.53) is identical to (3.7), we have, for each $n \geq 0$, that

$$q_{j,k}^n = \rho_k^n N_{j,k} \tag{3.58}$$

for some ρ_k^n. The boundary condition in (3.12) implies that

$$\rho_k^n = \Delta a \sum_{j=1}^{K_a} q_{j,k}^n \phi_{j-1,k}. \tag{3.59}$$

Then, by Theorem 3.2, we have that $\rho_k^n \equiv \rho_k^0$. ∎

4. Case with mutation ($m > 0$). A more interesting and realistic model is to include the mutation effect, i.e. $m \neq 0$. In this case, the equation for the normalized solution $\tilde{n}(t, a, x)$ becomes

$$
\begin{cases}
\varepsilon \partial_t \tilde{n}_\varepsilon + \partial_a \tilde{n}_\varepsilon + d(a,x)\tilde{n}_\varepsilon = -\lambda_\varepsilon(t)\tilde{n}_\varepsilon, \\
\tilde{n}_\varepsilon(t, a=0, x) = (1-m)\displaystyle\int_0^\infty b(a,x)\tilde{n}_\varepsilon(t,a,x)\,da \\
\qquad\qquad + \dfrac{m}{\varepsilon}\displaystyle\int_0^\infty\int_0^\infty b(a,y) M\left(\dfrac{x-y}{\varepsilon}\right)\tilde{n}_\varepsilon(t,a,y)\,dady,
\end{cases} \tag{4.1}
$$

where $\lambda_\varepsilon(t) := \varepsilon\frac{\dot{\rho_\varepsilon}}{\rho_\varepsilon} = \int [b(a,x) - d(a,x)]\tilde{n}_\varepsilon(t,a,x)dadx$. However the theory is not fully understood in these cases and the general proof of convergence of the correctors, beyond the caustics, is a long standing open question. Formally, however the same method can be applied and uniform bounds can be obtained which are enough to define an A-P scheme.

4.1. Asymptotic variable separation. Inspired by the case without mutation, we take the same ansatz (2.9) for $\tilde{n}(t, a, x)$, i.e.

$$
\tilde{n}_\varepsilon(t,a,x) = e^{\frac{u_\varepsilon(t,x)}{\varepsilon}} q_\varepsilon(t,a,x), \tag{4.2}
$$

which leads to the following equation by substituting (4.2) into (4.1)

$$
\begin{cases}
q_\varepsilon \partial_t u_\varepsilon + \varepsilon \partial_t q_\varepsilon + \partial_a q_\varepsilon + d(a,x) q_\varepsilon + \lambda_\varepsilon(t) q_\varepsilon = 0, \\
q_\varepsilon(t, a=0, x) = (1-m)\displaystyle\int_0^\infty b(a,x) q_\varepsilon(t,a,x)\,da \\
\qquad\qquad + \dfrac{m}{\varepsilon}\displaystyle\int_0^\infty\int_0^\infty b(a,y) M\left(\dfrac{x-y}{\varepsilon}\right) q_\varepsilon(t,a,y) e^{\frac{u_\varepsilon(t,y)-u_\varepsilon(t,x)}{\varepsilon}}\,dady.
\end{cases} \tag{4.3}
$$

For convenience of the numerical computation, we introduce $z = (x-y)/\varepsilon$ and rewrite the boundary condition in (4.3) as

$$
\begin{aligned}
q_\varepsilon(t, a=0, x) =& (1-m)\int_0^\infty b(a,x) q_\varepsilon(t,a,x)\,da \\
&+ m\int_{-\infty}^\infty \left(\int_0^\infty b(a, x-\varepsilon z) q_\varepsilon(t,a,x-\varepsilon z)\,da\right) M(z)\, e^{\frac{u_\varepsilon(t,x-\varepsilon z)-u_\varepsilon(t,x)}{\varepsilon}}\,dz,
\end{aligned} \tag{4.4}
$$

where we force $q_\varepsilon(t,a,x) = 0$ when $x < 0$. Following the case without mutations, we will establish the dynamical equations for $u_\varepsilon(t,x)$ and $q_\varepsilon(t,a,x)$, respectively, based on (4.3) and (4.4).

Assuming that $\partial_x u_\varepsilon(t,x)$ is independent of ε, which will be shown later to be possible, and introducing $\bar{\eta}[\partial_x u_\varepsilon]$ as

$$
\bar{\eta}[\partial_x u_\varepsilon] := \int_{-\infty}^\infty M(z)\, e^{-z\partial_x u_\varepsilon(t,x)}\,dz \tag{4.5}
$$

for any given x and t, we can write (4.4) of first order as $\varepsilon \to 0$

$$
q_\varepsilon(t, a=0, x) = (1 - m + m\bar{\eta}[\partial_x u_\varepsilon])\int_0^\infty b(a,x) q_\varepsilon(t,a,x)\,da. \tag{4.6}
$$

Following Section 2.1, it would be useful to consider the following spectral problem in variable a, where x is a parameter and the parameter η will take the value $\bar{\eta}[\partial_x u_\varepsilon]$ defined in (4.5),

$$\begin{cases} \partial_a N(a,x,\eta) + d(a,x)N(a,x,\eta) = -\Lambda(x,\eta)N(a,x,\eta), \\ N(a=0,x,\eta) = (1-m+m\eta)\int_0^\infty b(a,x)N(a,x,\eta)\,da, \\ N(a,x,\eta) > 0 \quad \text{and} \quad \int_0^\infty N(a,x,\eta)\,da = 1. \end{cases} \tag{4.7}$$

In words, $\Lambda(x,\bar{\eta}[\partial_x u_\varepsilon])$ is the leading eigenvalue of the operator $\partial_a + d(a,x)$ with a parameter dependent boundary condition, which approximates (4.4), and the corresponding normalized eigenfunction is $N(a,x,\bar{\eta}[\partial_x u_\varepsilon])$. Theoretical analysis shows that $\partial_\eta \Lambda > 0$ [24].

Now we choose the Hamilton-Jacobi equation for $u_\varepsilon(t,x)$ as

$$\partial_t u_\varepsilon(t,x) = \Lambda(x,\bar{\eta}[\partial_x u_\varepsilon]) - \lambda_\varepsilon(t), \tag{4.8}$$

where $\lambda_\varepsilon(t)$ is independent of x and only changes $u_\varepsilon(t,x)$ by a time dependent value in order to ensure the mass 1 conservation law. As shown in [8], the viscosity solution for the Hamilton-Jacobi equation exists and is unique. Besides, it is easy to verify that $\partial_x u_\varepsilon(t,x)$ is independent of ε noticing the fact that the only term related to ε is $\lambda_\varepsilon(t)$, which is independent of x.

Combining (4.3) and (4.8), the function $q_\varepsilon(t,a,x)$ satisfies

$$\begin{cases} \varepsilon\partial_t q_\varepsilon + \partial_a q_\varepsilon + d(a,x)q_\varepsilon = -\Lambda(x,\bar{\eta}[\partial_x u_\varepsilon])q_\varepsilon, \\ q_\varepsilon(t,a=0,x) = (1-m)\displaystyle\int_0^\infty b(a,x)q_\varepsilon(t,a,x)\,da \\ \qquad\qquad + m\displaystyle\int_{-\infty}^{\infty}\int_0^\infty b(a,x-\varepsilon z)M(z)\,q_\varepsilon(t,a,x-\varepsilon z)e^{\frac{u_\varepsilon(t,x-\varepsilon z)-u_\varepsilon(t,x)}{\varepsilon}}\,dadz. \end{cases} \tag{4.9}$$

The limiting equations of $u_\varepsilon(t,x)$ and $q_\varepsilon(t,a,x)$ as $\varepsilon \to 0^+$ can be derived. Denote $u(t,x)$ and $q(t,a,x)$ to be the limit function of $u_\varepsilon(t,x)$ and $q_\varepsilon(t,a,x)$, respectively. Similar to the case without mutation, we must have

$$\sup_x\{u(t,x)\} = 0. \tag{4.10}$$

By taking the limit $\varepsilon \to 0^+$ in (4.9), we can easily get the equation for $q(t,a,x)$ to be

$$\begin{cases} \partial_a q + d(a,x)q = -\Lambda(x,\bar{\eta}[\partial_x u])q, \\ q(t,a=0,x) = (1-m+m\bar{\eta}[\partial_x u]))\int_0^\infty b(a,x)q_\varepsilon(t,a,x)\,da. \end{cases} \tag{4.11}$$

4.2. Finite difference discretization. Based on the dynamical equation (4.8) for $u_\varepsilon(t,x)$ and the one (4.9) for $q_\varepsilon(t,a,x)$, we are ready to detail the numerical scheme with finite difference discretization in space.

4.2.1. Notations. For simplicity, we choose the same notations as in Section 3 and assume the mutation function $M(z)$ to be compactly supported on the interval $(-Z,Z)$. Discretize the interval with uniformly distributed grid points as

$$z_l = l\Delta z, \quad \text{for } l \in [\![-K_z, K_z]\!], \tag{4.12}$$

where $\Delta z = \frac{Z}{K_z}$, and denote M_l to be the numerical approximation of $M(z_l)$.

The parameter $\bar{\eta}[\partial_x u_\varepsilon]$ defined in (4.5) plays an important role in the case with mutation. For each $k \in [\![0, K_x]\!]$ and $n > 0$, we compute η_k^n, the numerical approximation of $\bar{\eta}[\partial_x u_\varepsilon](t_n, x_k)$, via the finite difference method. The first-order upwind scheme [9, 10, 26, 30] is applied for computing the first order derivatives. To be more specific, we set

$$\eta_k^n = g(\delta_x^- u_k^n, \delta_x^+ u_k^n), \tag{4.13}$$

where the function $g(\alpha, \beta)$ is defined as

$$g(\alpha, \beta) = \Delta z \sum_{l=0}^{K_z} w_l^{(z)} \left(M_{(-l)} e^{-z_{(-l)}\beta} + M_l e^{-z_l \alpha} \right), \tag{4.14}$$

with weights

$$w_l^{(z)} = 1 - \frac{\delta_{l,0} + \delta_{l,K_z}}{2} = \begin{cases} \frac{1}{2}, & \text{if } l = 0 \text{ or } K_z, \\ 1, & \text{otherwise.} \end{cases} \tag{4.15}$$

To avoid the difficulties with the boundary conditions, we simply assume $\delta_x^- u_0^n = \delta_x^+ u_{K_x}^n = 0$ for any $n \geq 1$. This does not have much effect because it corresponds to very negative values of $\mathbf{u}^n$, whose effect on the normalized population is exponentially small and is thus negligible. Obviously, the scheme (4.13) is independent of ε. More accurate finite difference approximations for the first order derivatives, such as WENO [19, 20, 31], as well as approximations via other methods, including the finite volume method and DG method [7, 17, 33], could be applied as well. However, the corresponding schemes would obviously be much more complicated noticing that all u_k^n must be updated implicitly.

The eigenpair $(\Lambda_k(\eta_k^n), \mathbf{n}_k(\eta_k^n))$, which now depends on the parameter η_k^n, satisfies the following discretized eigenvalue problem of (4.7)

$$\begin{cases} \delta_a^- N_{j,k}(\eta_k^n) + d(a_j, x_k) N_{j,k}(\eta_k^n) = -\Lambda_k(\eta_k^n) N_{j,k}(\eta_k^n), \quad j \in [\![1, K_a]\!], \\ N_{0,k}(\eta_k^n) = (1 - m + m\eta_k^n) \Delta a \sum_{j=1}^{K_a} b(a_j, x_k) N_{j,k}(\eta_k^n), \\ \Delta a \sum_{j=0}^{K_a} w_j^{(a)} N_{j,k}(\eta_k^n) = 1, \end{cases} \tag{4.16}$$

where $\mathbf{n}_k(\eta_k^n) = (N_{1,k}(\eta_k^n), N_{2,k}(\eta_k^n), \ldots, N_{K_a,k}(\eta_k^n))^T$ and $k \in [\![0, K_x]\!]$. The equation (4.16) can be written in a more compact form with $M_k(\eta_k^n) \in \mathbb{R}^{K_a \times K_a}$ defined in (A.1) as

$$M_k(\eta_k^n) \mathbf{n}_k(\eta_k^n) = -\Lambda_k(\eta_k^n) \mathbf{n}_k(\eta_k^n). \tag{4.17}$$

Then $\Lambda_k(\eta_k^n)$ is defined to be the leading eigenvalue of the matrix $M_k(\eta_k^n)$. Again the Perron-Frobenius theorem indicates the existence of the positive eigenvector $\mathbf{n}_k(\eta_k^n)$ for each $k \in [\![0, K_x]\!]$ and any $\eta_k^n > 0$.

4.2.2. Numerical Scheme. With all these preparations, now we are ready to show the detailed scheme to update $\mathbf{u}^n$ and Q^n.

Step 1a: Theoretical computation of $\tilde{u}_k^n$, an approximation of u_k^{n+1}. The computation of $\tilde{u}_k^n$ is based on the finite difference discretization of

$$\partial_t \tilde{u}(t, x) = \Lambda(x, \bar{\eta}[\partial_x \tilde{u}]). \tag{4.18}$$

The backward-Euler method is applied in time to ensure the stability of the scheme. Denote $\tilde{\mathbf{u}}^n := (\tilde{u}_0^n, \tilde{u}_1^n, \ldots, \tilde{u}_{K_x}^n)^T$ and define

$$\tilde{\eta}_k^n = g(\delta_x^- \tilde{u}_k^n, \delta_x^+ \tilde{u}_k^n), \tag{4.19}$$

where the function $g(\cdot,\cdot)$ is defined in (4.14). Then, the scheme works as follows

$$\frac{\tilde{u}_k^n - u_k^n}{\Delta t} = \Lambda_k(\tilde{\eta}_k^n) = \Lambda_k(g(\delta_x^- \tilde{u}_k^n, \delta_x^+ \tilde{u}_k^n)), \quad k \in [\![0, K_x]\!],\ n \geq 0, \tag{4.20}$$

where $\Lambda_k(\tilde{\eta}_k^n)$ is the leading eigenvalue of the matrix $M_k(\tilde{\eta}_k^n)$.

Intuitively, the solution of the equation (4.18) differs from the solution of (4.8) by a function of time, which is independent of x. As a result, u_k^{n+1} can be obtained from $\tilde{u}_k^n$ by adding a constant which will be determined later for the mass 1 normalization.

The following lemma shows that the scheme (4.20) is an unconditionally monotone scheme.

LEMMA 4.1. *The scheme* (4.20) *is a monotone scheme for any* $\Delta t > 0$.

Proof. On one hand, $\mathbf{u}^n$ can be viewed as a function of $\tilde{\mathbf{u}}^n$ via (4.20). Differentiating both sides of (4.20), we get

$$\frac{\partial u_k^n}{\partial \tilde{u}_k^n} = 1 - \Delta t \Lambda_k'(\tilde{\eta}_k^n)\frac{\partial \tilde{\eta}_k^n}{\partial \tilde{u}_k^n}, \quad \frac{\partial u_k^n}{\partial \tilde{u}_{k-1}^n} = -\Delta t \Lambda_k'(\tilde{\eta}_k^n)\frac{\partial \tilde{\eta}_k^n}{\partial \tilde{u}_{k-1}^n}, \quad \frac{\partial u_k^n}{\partial \tilde{u}_{k+1}^n} = -\Delta t \Lambda_k'(\tilde{\eta}_k^n)\frac{\partial \tilde{\eta}_k^n}{\partial \tilde{u}_{k+1}^n}. \tag{4.21}$$

Noticing that $\Lambda_k'(\eta) > 0$, $\frac{\partial \tilde{\eta}_k^n}{\partial \tilde{u}_{k\pm1}^n} > 0$ and $\frac{\partial \tilde{\eta}_k^n}{\partial \tilde{u}_k^n} = -\frac{\partial \tilde{\eta}_k^n}{\partial \tilde{u}_{k-1}^n} - \frac{\partial \tilde{\eta}_k^n}{\partial \tilde{u}_{k+1}^n} < 0$, the tridiagonal Jacobian matrix $\mathbf{J}_{\mathbf{u}^n}$ is an M-matrix for whatever $\Delta t > 0$, which implies that $\mathbf{J}_{\mathbf{u}^n}^{-1}$ exists and all its elements are non-negative.

On the other hand, for each $\mathbf{u}^n$, there exists a unique solution $\tilde{\mathbf{u}}^n$, where the existence and uniqueness of the solution will be proven in Lemma C.1 later. Therefore, $\tilde{\mathbf{u}}^n$ can be viewed as a function of $\mathbf{u}^n$ as well, whose Jacobian matrix,

$$\mathbf{J}_{\tilde{\mathbf{u}}^n}(\mathbf{u}^n) = [\mathbf{J}_{\mathbf{u}^n}(\tilde{\mathbf{u}}^n)]^{-1}, \tag{4.22}$$

is a matrix with all elements to be non-negative. In other words, we have $\frac{\partial \tilde{u}_{\tilde{k}}^n}{\partial u_k^n} \geq 0$ for any $k, \tilde{k} \in [\![0, K_x]\!]$, which means that the scheme (4.20) is a monotone scheme. ∎

Step 1b: Practical computation of $\tilde{u}_k^n$. Iterative techniques will be used to solve the implicitly formulated equation (4.20). However, it is difficult to solve (4.20) directly since there is no explicit formula of the function $\Lambda_k(\eta)$. As an alternative, we update $\Lambda_k(\eta_k^{n+1})$ first and then compute $\tilde{u}_k^n$ via (4.20). Following a similar procedure as in Appendix B, we get the following linear system for $\Lambda_k(\eta_k^{n+1})$,

$$1 = (1 - m + m\tilde{\eta}_k^n)\Delta a \sum_{j=1}^{K_a} b(a_j, x_k) \prod_{s=1}^{j} \frac{1}{1 + \Delta a(d(a_s, x_k) + \Lambda_k(\tilde{\eta}_k^n))}, \quad k \in [\![0, K_x]\!], \tag{4.23}$$

where

$$\tilde{\eta}_k^n = g(\delta_x^- u_k^n + \Delta t \delta_x^- \Lambda_k(\tilde{\eta}_k^n), \delta_x^+ u_k^n + \Delta t \delta_x^+ \Lambda_k(\tilde{\eta}_k^n)), \tag{4.24}$$

by combining (4.13) and (4.20). Then iterative methods, such as the Newton method, can be applied for the system (4.23)-(4.24) to solve $\Lambda_k(\eta_k^{n+1})$.

For completeness, here we present the detailed Newton's method for solving the system (4.23)-(4.24). For simplicity of notations, we introduce

$$\boldsymbol{\lambda}^{(l)} = (\Lambda_0^{(l)}, \Lambda_1^{(l)}, \dots, \Lambda_{K_x}^{(l)})^T \in \Omega, \tag{4.25}$$

where $\Omega := \{\boldsymbol{\lambda} \in \mathbb{R}^n \,|\, 1 + \Delta a(d(a_s, x_k) + \Lambda_k) > 0,\ k \in [\![0, K_x]\!]\}$ is the feasible set of $\boldsymbol{\lambda}$, and introduce the functions $Y_k^n(\boldsymbol{\lambda})$, where $k \in [\![0, K_x]\!]$, to be the right-hand side of (4.23). Then the system (4.23)-(4.24) can be rewritten as

$$\mathbf{Y}^n(\boldsymbol{\lambda}) = \mathbf{1}_{K_x+1}, \tag{4.26}$$

where $\mathbf{1}_{K_x+1} = (1, 1, \dots, 1)^T \in \mathbb{R}^{K_x+1}$ and $\mathbf{Y}^n(\boldsymbol{\lambda}) = (Y_0^n(\boldsymbol{\lambda}), Y_1^n(\boldsymbol{\lambda}), \dots, Y_{K_x}^n(\boldsymbol{\lambda}))^T \in \mathbb{R}^{K_x+1}$. The Jacobian matrix $\mathbf{J}_{\mathbf{Y}^n}$, which is defined as

$$\mathbf{J}_{\mathbf{Y}^n}(\boldsymbol{\lambda}) = \begin{bmatrix} \frac{\partial Y_0^n}{\partial \Lambda_0} & \frac{\partial Y_0^n}{\partial \Lambda_1} & \frac{\partial Y_0^n}{\partial \Lambda_2} & \cdots & \frac{\partial Y_0^n}{\partial \Lambda_{K_x}} \\ \frac{\partial Y_1^n}{\partial \Lambda_0} & \frac{\partial Y_1^n}{\partial \Lambda_1} & \frac{\partial Y_1^n}{\partial \Lambda_2} & \cdots & \frac{\partial Y_1^n}{\partial \Lambda_{K_x}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial Y_{K_x}^n}{\partial \Lambda_0} & \frac{\partial Y_{K_x}^n}{\partial \Lambda_1} & \frac{\partial Y_{K_x}^n}{\partial \Lambda_2} & \cdots & \frac{\partial Y_{K_x}^n}{\partial \Lambda_{K_x}} \end{bmatrix} \in \mathbb{R}^{(K_x+1)\times(K_x+1)}, \tag{4.27}$$

is tridiagonal, where all non-zero elements can be explicitly computed, and strictly diagonally dominate with negative diagonal elements, which implies that its inverse $[\mathbf{J}_{\mathbf{Y}^n}(\boldsymbol{\lambda})]^{-1}$ exists everywhere with all elements non-positive. Then we update an intermediate solution $\boldsymbol{\lambda}^{(l)}$ via

$$\boldsymbol{\lambda}^{(l+1)} = \boldsymbol{\lambda}^{(l)} + [\mathbf{J}_{\mathbf{Y}^n}(\boldsymbol{\lambda}^{(l)})]^{-1}(\mathbf{1}_{K_x+1} - \mathbf{Y}^n(\boldsymbol{\lambda}^{(l)})), \quad l \geq 0, \tag{4.28}$$

and the detailed Newton's method works as follows. It will be shown in Appendix C that the sequence $\{\boldsymbol{\lambda}^{(l)}\}$ will converge to the desired solution.

Algorithm 4.1 Newton's iteration for solving the system (4.23)-(4.24) (or equivalently (4.26))

1: For each $n \geq 1$ and $k \in [\![0, K_x]\!]$, compute $\Lambda_k^{(0)}$ such that $(1-m)S_k(\Lambda_k^{(0)}) = 1$
2: $l \leftarrow 0$, $\boldsymbol{\lambda}^{(0)} \leftarrow (\Lambda_0^{(0)}, \Lambda_1^{(0)}, \dots, \Lambda_{K_x}^{(0)})^T$
3: **while** $\|\mathbf{1}_{K_x+1} - \mathbf{Y}^n(\boldsymbol{\lambda}^{(l)})\|_{l^\infty} > tolerance$ **do**
4: $\quad \boldsymbol{\lambda}^{(l+1)} \leftarrow \boldsymbol{\lambda}^{(l)} + [\mathbf{J}_{\mathbf{Y}^n}(\boldsymbol{\lambda}^{(l)})]^{-1}(\mathbf{1}_{K_x+1} - \mathbf{Y}^n(\boldsymbol{\lambda}^{(l)}))$
5: $\quad l \leftarrow l + 1$
6: **end while**

Remark 4.2. There is no need to compute $[\mathbf{J}_{\mathbf{Y}^n}(\boldsymbol{\lambda})]^{-1}$ in Algorithm 4.1. Instead, we can evaluate the part $[\mathbf{J}_{\mathbf{Y}^n}(\boldsymbol{\lambda}^{(l)})]^{-1}(\mathbf{1}_{K_x+1} - \mathbf{Y}^n(\boldsymbol{\lambda}^{(l)}))$ as a whole via the Tridiagonal matrix algorithm (TDMA), which is much more computationally efficient.

Step 2: Update $Q^n := (q_{j,k}^n) \in \mathbb{R}^{K_a \times K_x}$ based on the finite difference discretization of (4.9). Again, the stiff terms must be treated implicitly to ensure the stability of the

scheme.

$$
\begin{cases}
\varepsilon\frac{q_{j,k}^{n+1}-q_{j,k}^{n}}{\Delta t}+\delta_a^- q_{j,k}^{n+1}+d(a_j,x_k)q_{j,k}^{n+1}=-\Lambda_k(\tilde{\eta}_k^n)q_{j,k}^{n+1},\\
q_{0,k}^{n+1}=(1-m)\Delta a\sum_{j=1}^{K_a}b(a_j,x_k)q_{j,k}^{n+1}+\\
\qquad m\Delta z\sum_{l=-K_z}^{K_z}\tilde{w}_l^{(z)}\left[\Delta a\sum_{j=1}^{K_a}b(a_j,x_k-\varepsilon z_l)q_{j,k-\tilde{\varepsilon}l}^{n+1}\right]M_l e^{\frac{\tilde{u}_{k-\tilde{\varepsilon}l}^n-\tilde{u}_k^n}{\varepsilon}},
\end{cases}
\tag{4.29}
$$

where $\tilde{\varepsilon}=\varepsilon\Delta z/\Delta x$ and $\tilde{u}_{k-\tilde{\varepsilon}l}^n-\tilde{u}_k^n$ and $q_{j,k-\tilde{\varepsilon}l}^{n+1}$ are the numerical approximations of $u_\varepsilon(t_{n+1},x_k-\varepsilon z_l)-u_\varepsilon(t_{n+1},x_k)$ and $q_\varepsilon(t_{n+1},a_j,x_k-\varepsilon z_l)$, respectively, and the new weights $\tilde{w}_l^{(z)}$ for $l\in[\![-K_z,K_z]\!]$ are defined as

$$
\tilde{w}_l^{(z)}=\begin{cases}w_{|l|}^{(z)}, & \text{if } l\neq 0,\\ 2w_0^{(z)}, & \text{if } l=0.\end{cases}
\tag{4.30}
$$

Here we assume $q_\varepsilon(t,\cdot,\cdot)$ is constantly 0 outside the computational domain $(0,M)^2$.

A simple way to evaluate $\tilde{u}_{k-\tilde{\varepsilon}l}^n-\tilde{u}_k^n$ and $q_{j,k-\tilde{\varepsilon}l}^{n+1}$ is via linear interpolation. When $0<\varepsilon<\frac{\Delta x}{K_z\Delta z}$, we have $k-\tilde{\varepsilon}l\in[k-1,k+1]$ for all $l\in[-K_z,K_z]$. Therefore,

$$
q_{j,k-\tilde{\varepsilon}l}^{n+1}=\begin{cases}q_{j,k}^{n+1}-\varepsilon z_l\delta_x^+ q_{j,k}^{n+1},\\ q_{j,k}^{n+1}-\varepsilon z_l\delta_x^- q_{j,k}^{n+1},\end{cases}
\qquad
\tilde{u}_{k-\tilde{\varepsilon}l}^n-\tilde{u}_k^n=\begin{cases}-\varepsilon z_l\delta_x^+\tilde{u}_k^n, & \text{for } -K_z\le l\le 0,\\ -\varepsilon z_l\delta_x^-\tilde{u}_k^n, & \text{for } 0<l\le K_z.\end{cases}
\tag{4.31}
$$

Then, the boundary condition in (4.29) can be reformulated as

$$
q_{0,k}^{n+1}=(1-m)\Delta a\sum_{j=1}^{K_a}b(a_j,x_k)q_{j,k}^{n+1}+m\Delta a\sum_{j=1}^{K_a}(C_{j,k}^{n,(0)}q_{j,k}^{n+1}+C_{j,k}^{n,(-1)}q_{j,k-1}^{n+1}+C_{j,k}^{n,(1)}q_{j,k+1}^{n+1}),
\tag{4.32}
$$

where the coefficients can be explicitly computed as

$$
\begin{aligned}
C_{j,k}^{n,(0)}=\Delta z\sum_{l=0}^{K_z}w_l^{(z)}(1-\frac{\varepsilon z_l}{\Delta x})\Big[&b(a_j,x_k-\varepsilon z_l)M_l e^{-z_l\delta_x^-\tilde{u}_k^n}\\
&+b(a_j,x_k+\varepsilon z_l)M_{(-l)}e^{z_l\delta_x^+\tilde{u}_k^n}\Big]>0,
\end{aligned}
\tag{4.33}
$$

and

$$
C_{j,k}^{n,(-1)}=\Delta z\sum_{l=0}^{K_z}w_l^{(z)}\frac{\varepsilon z_l}{\Delta x}b(a_j,x_k-\varepsilon z_l)M_l e^{-z_l\delta_x^-\tilde{u}_k^n}>0,
\tag{4.34}
$$

$$
C_{j,k}^{n,(1)}=\Delta z\sum_{l=0}^{K_z}w_l^{(z)}\frac{\varepsilon z_l}{\Delta x}b(a_j,x_k+\varepsilon z_l)M_{(-l)}e^{z_l\delta_x^+\tilde{u}_k^n}>0.
\tag{4.35}
$$

Similarly, when $\varepsilon\ge\frac{\Delta x}{K_z\Delta z}$, the boundary condition in (4.29) can be proved to be of form

$$
q_{0,k}^{n+1}=(1-m)\Delta a\sum_{j=1}^{K_a}b(a_j,x_k)q_{j,k}^{n+1}+m\Delta a\sum_{j=1}^{K_a}\sum_{l=-k}^{K_x-k}C_{j,k}^{n,(l)}q_{j,k+l}^{n+1},
\tag{4.36}
$$

where $C_{j,k}^{n,(l)} \geq 0$ for all $j \in [\![0, K_a]\!]$, $k \in [\![0, K_x]\!]$, $l \in [\![-k, K_x - k]\!]$ and $n \geq 0$. The nonnegativity of the coefficients $C_{j,k}^{n,(l)}$ comes from the fact that $q_{j,k-\bar{\varepsilon}l}^{n+1}$ is obtained by the linear interpolation of Q^{n+1}, which implies that $\frac{\partial q_{j,k-\bar{\varepsilon}l}^{n+1}}{\partial q_{\tilde{j},\tilde{k}}^{n+1}} \geq 0$ holds true for all $j, \tilde{j} \in [\![0, K_a]\!]$ and $k, \tilde{k} \in [\![0, K_x]\!]$. The explicit formula for the coefficients is omitted here for brevity.

Remark 4.3. Unlike the case without mutation, now we no longer have the same maximum principle of q_ε on both the discrete and continuous level because of the coupling in x-direction that appeared in the boundary conditions.

Step 3: Compute $\mathbf{u}^{n+1}$ *via normalization.* For each $k \in [\![0, K_x]\!]$, we define

$$u_k^{n+1} = \tilde{u}_k^n - \Delta t\, L^n, \tag{4.37}$$

where L^n is the normalization constant making sure that $I_2(\tilde{N}^{n+1}) = 1$ with $\tilde{N}^{n+1}$ defined in (3.4). The constant L^n can be computed explicitly as in (3.19), i.e.

$$L^n = \frac{\varepsilon}{\Delta t} \ln\left(I_2(Q^{n+1} \circ \mathbf{v}^n)\right), \tag{4.38}$$

where $\mathbf{v}^n := (v_k^n)^T \in \mathbb{R}^{K_x+1}$ with $v_k^n = e^{\frac{\tilde{u}_k^n}{\varepsilon}}$, and the operator $\circ$ denotes the generalized Hadamard product defined in (3.20). Similar to the case without mutation, we can then reformulate the scheme (4.37) to be a more robust one

$$u_k^{n+1} = \tilde{u}_k^n - \|\tilde{\mathbf{u}}^n\|_{l^\infty} - \varepsilon \ln\left(I_2(Q^{n+1} \circ \tilde{\mathbf{v}}^n)\right). \tag{4.39}$$

where $\tilde{\mathbf{v}}^n := \frac{\mathbf{v}^n}{\|\mathbf{v}^n\|_{l^\infty}}$ is the normalized vector in the sense $\|\tilde{\mathbf{v}}^n\|_{l^\infty} = 1$.

4.3. Asymptotic preserving (A-P) property. Due to the lack of the maximum principle of Q^{n+1}, we can only formally verify the A-P property of the schemes proposed in Section 4.2. Assuming Q^{n+1} is bounded, the limiting scheme for $\mathbf{u}^n$ can be derived similarly as in Theorem 3.6. As for the limiting scheme for Q^n, it is enough to consider the case $0 < \varepsilon < \frac{\Delta x}{z_{K_z}}$, where the updating scheme (4.29) becomes

$$\begin{cases} \varepsilon \frac{q_{j,k}^{n+1} - q_{j,k}^n}{\Delta t} + \delta_a^- q_{j,k}^{n+1} + d(a_j, x_k) q_{j,k}^{n+1} = -\Lambda_k(\tilde{\eta}_k^n) q_{j,k}^{n+1}, \\ q_{0,k}^{n+1} = (1-m)\Delta a \sum_{j=1}^{K_a} b(a_j, x_k) q_{j,k}^{n+1} + \\ \qquad m\Delta a \sum_{j=1}^{K_a} (C_{j,k}^{n,(0)} q_{j,k}^{n+1} + C_{j,k}^{n,(-1)} q_{j,k-1}^{n+1} + C_{j,k}^{n,(1)} q_{j,k+1}^{n+1}), \end{cases} \tag{4.40}$$

where the positive coefficients $C_{j,k}^{n,(-1)}$, $C_{j,k}^{n,(0)}$ and $C_{j,k}^{n,(1)}$ are defined in (4.33)-(4.35). Obviously, by taking $\varepsilon \to 0$, we have

$$\lim_{\varepsilon \to 0} C_{j,k}^{n,(1)} = \lim_{\varepsilon \to 0} C_{j,k}^{n,(-1)} = 0, \quad \lim_{\varepsilon \to 0} C_{j,k}^{n,(0)} = b(a_j, x_k)\eta_k^{n+1}, \tag{4.41}$$

where $\eta_k^{n+1} = \tilde{\eta}_k^n$ is defined in (4.19). Therefore, by formally taking the limit $\varepsilon \to 0$ in the scheme (4.40), we get the limiting scheme

$$\begin{cases} \delta_a^- q_{j,k}^{n+1} + d(a_j, x_k) q_{j,k}^{n+1} = -\Lambda_k(\eta_k^{n+1}) q_{j,k}^{n+1}, \\ q_{0,k}^{n+1} = (1-m)\Delta a \sum_{j=1}^{K_a} b(a_j, x_k) q_{j,k}^{n+1} + m\Delta a \sum_{j=1}^{K_a} b(a_j, x_k)\eta_k^{n+1} q_{j,k}^{n+1}, \\ \qquad = (1 - m + m\eta_k^{n+1})\Delta a \sum_{j=1}^{K_a} b(a_j, x_k) q_{j,k}^{n+1}. \end{cases} \tag{4.42}$$

It is easy to check that the limiting scheme (4.42) is indeed the direct finite difference discretization of the limiting equation for q_ε (4.11) at $t = t_{n+1}$.

5. Numerical experiments. We illustrate the performance of the scheme and its A-P property with extensive numerical experiments. In particular, we focus on the case without mutation, where the concentration of the normalized population is more clear both theoretically and numerically. We show the efficiency and accuracy of the proposed scheme by comparing it with the standard explicit scheme, which solves the equation (2.1) directly. The case with mutation will be studied as well at the end.

5.1. Case without mutation. For the numerical test, we choose

$$b(a,x) = \max\left\{\frac{1-(a-2)^2}{1+(x)_+}, 0\right\}, \quad d(a,x) = 0.5a + 10(a-(x)_+)_+^2, \tag{5.1}$$

where $(x)_+ := \max\{x, 0\}$. Here we choose the initial data

$$\tilde{n}(0,a,x) = e^{-0.8a - \frac{(x-0.5)^2}{2}}, \tag{5.2}$$

and the computation domain of (a,x) to be $\Omega = (0,5)\times(0,5)$. Obviously, the birth rate function is uniformly compact supported in the interval $(1,3)$ in a-direction and the death rate function satisfies the assumptions as well. As shown in Figure 5.1, the concentration of the normalized population in x-direction will be more and more obvious as time progresses.

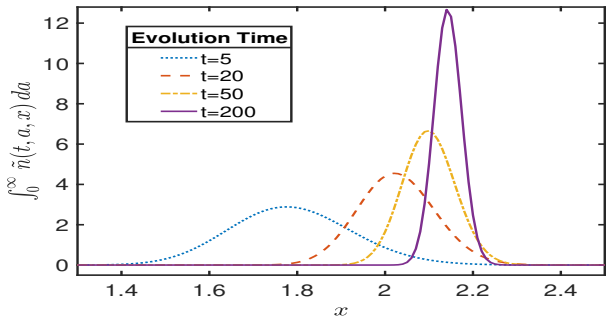

FIG. 5.1. *With $\varepsilon = 0.1$ fixed, we observe sharper concentration as time progresses.*

To begin with, we confirm that our method (3.15)-(3.23) does solve the original equation (2.1) correctly by comparing it with a standard implicit discretization of the equation (2.1), i.e.

$$\begin{cases} \varepsilon\frac{\tilde{n}_{j,k}^{n+1}-\tilde{n}_{j,k}^{n}}{\Delta t} + \delta_a^-\tilde{n}_{j,k}^{n+1} + d(a_j,x_k)\tilde{n}_{j,k}^{n+1} = -\lambda^n\tilde{n}_{j,k}^{n+1}, \quad j\in[\![1,K_a]\!], \\ \tilde{n}_{0,k}^{n+1} = \Delta a\sum_{j=1}^{K_a} b(a_j,x_k)\tilde{n}_{j,k}^{n+1}, \end{cases} \tag{5.3}$$

where $\tilde{n}_{j,k}^n$ is the numerical approximation of $\tilde{n}(t_n,a_j,x_k)$, $j\in[\![0,K_a]\!]$, $k\in[\![0,K_x]\!]$, $n\geq 0$, and λ^n is a constant making sure that

$$\Delta a\Delta x\sum_{j=1}^{K_a}\sum_{k=0}^{K_x} w_k^{(x)}\tilde{n}_{j,k}^n = 1. \tag{5.4}$$

For numerical experiments, we fix $\varepsilon = 0.01$ and do the computation till $t = 1$. Though the implicit scheme (5.3) is stable with large time steps, a small time step is still necessary for accuracy reasons. Here we choose $\Delta t = 10^{-4}$ for the scheme (5.3) under a mesh $\Delta a = \Delta x = 0.05$, which is fine enough for the direct scheme to catch the concentration of the normalized population in x-direcction during the time. For our scheme (3.15)-(3.23), we apply $\Delta t = 10^{-1}$ and choose a coarse mesh in x-direcction with $\Delta x = 0.5$ and the same mesh in a-direction, i.e. $\Delta a = 0.05$. Then the solution can be accurately reconstructed on the fine mesh via an accurate interpolation, such as the spline interpolation. Figure 5.2 shows the normalized population computed via the two methods at time $t = 0.1$ and $t = 1$. As shown in the figure, the numerical solutions computed in two different ways are almost identical, which indicates not only the validity of our scheme (3.15)-(3.23), but also the efficiency since a much larger time step and an ε-independent mesh can be applied.

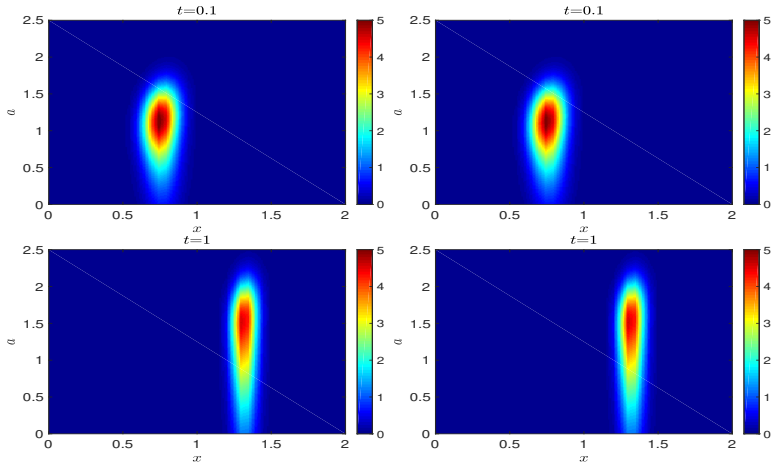

FIG. 5.2. *Comparison of the normalized population density $\tilde{n}(t, a, x)$ computed via a direct implicit scheme (left column) or via the A-P scheme (right column). For the the implicit scheme, we choose $\Delta t = 10^{-4}$ and $\Delta x = 0.05$, while for the A-P scheme, we choose $\Delta t = 10^{-1}$ and $\Delta x = 0.5$.*

Figure 5.3 shows the accuracy of Q^n, $\mathbf{u}$ and $\tilde{N}^n$ in a- and x-direction. Here we don't do any interpolation in x-direction. A roughly first-order accuracy in a-direction can be observed, which is consistent with our expectations since the upwind method, which is only first-order accurate, is applied for the derivatives in (3.15). The accuracy in x-direction is second order, which is consistent with our expectation as well since the system is completely decoupled in x-direction and the second order accurate quadrature rule is applied to the numerical integral in (3.17) for normalization. Besides, Q^n seems to converge as $\varepsilon \to 0^+$, which is consistent with Theorem 3.6.

One of the great advantage of our scheme (3.15)-(3.23) is that we can easily capture the concentration of the normalized population in x-direction with a coarse mesh. Since both the functions $q_\varepsilon(t, a, x)$ and $u_\varepsilon(t, x)$ are regular, an interpolation of Q^n and $\mathbf{u}^n$ computed on a coarse mesh can help accurately reconstruct the normalized population $\tilde{N}^n$ on a fine mesh. Figure 5.4 shows the accuracy with different interpolation methods. Obviously, the spline interpolation seems to be a good choice. Figure 5.5 compares the interpolated normalized population with the 'exact' one. Here we apply the spline interpolation in two ways. One way is to apply the spline interpolation for both $\mathbf{u}^n$ and Q^n as mentioned before, and the other way is to apply the spline inter-

polation for $\tilde{N}^n$ directly. As shown in the figure, the direct interpolation of $\tilde{N}^n$ would fail when the mesh is coarse while the interpolation performed on $\mathbf{u}^n$ and Q^n always works perfectly to catch the concentration. It implies that the WKB representation (3.3) plays an essential role in our A-P scheme, which enables us to use a coarse, probably ε-independent mesh in the x-direction to accurately capture the solution, and thus makes our A-P scheme efficient.

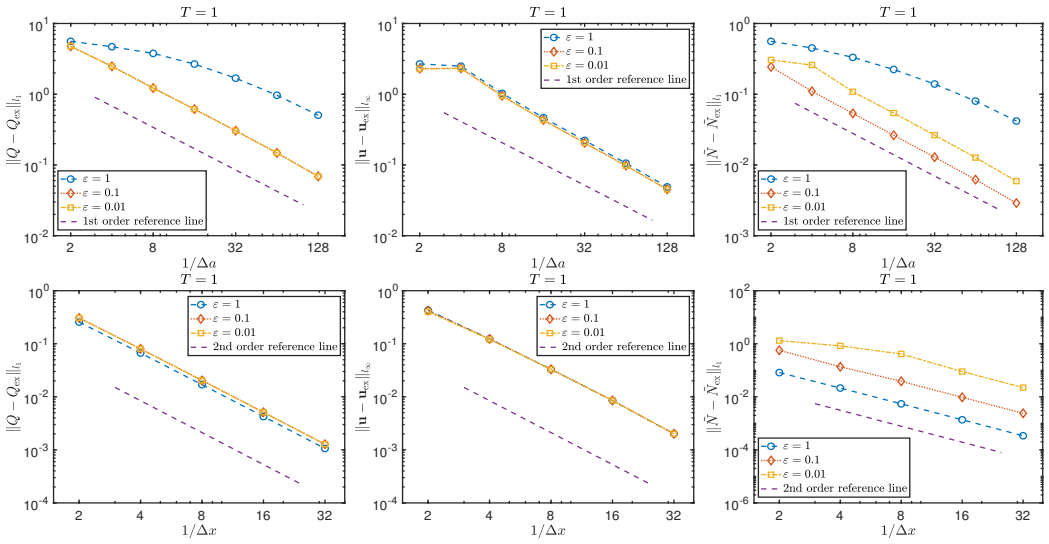

FIG. 5.3. *Accuracy of Q (left), $\mathbf{u}$ (middle) and $\tilde{N}$ (right) at time $T = 1$ with fixed $\Delta t = 10^{-3}$. The figures on the first row show the accuracy in a-direction, where we choose $\Delta x = 0.01$ and the 'exact' one computed with $\Delta a = 0.001$. The figures on the second row show the accuracy in x-direction, where we choose $\Delta a = 0.01$ and the 'exact' one computed with $\Delta x = 0.01$.*

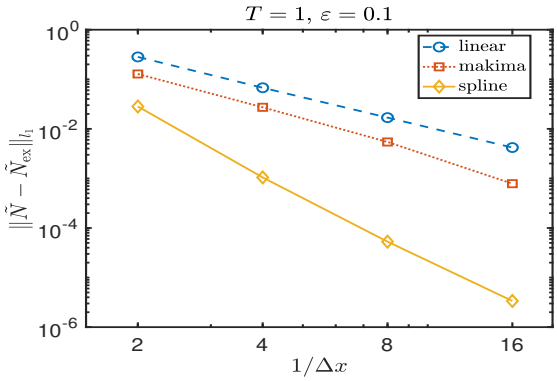

FIG. 5.4. *Numerical accuracy of the normalized population $\tilde{N}$, which are computed via the WKB representation (3.3) with Q and $\mathbf{u}$ computed on a coarse mesh in x-direction and then interpolated onto a fine mesh via different methods. Here we choose $T = 1$, $\varepsilon = 0.1$ and $\Delta t = \Delta a = 0.01$.*

Another great advantage of our method is that our scheme is unconditionally stable and the numerical solutions at some fixed time T would converge extremely fast as $\Delta t \to 0^+$ if $T \gg \varepsilon$, which is somewhat surprising since we applied the backward Euler method, which is only first-order accurate, in our scheme (3.15). Figure 5.6

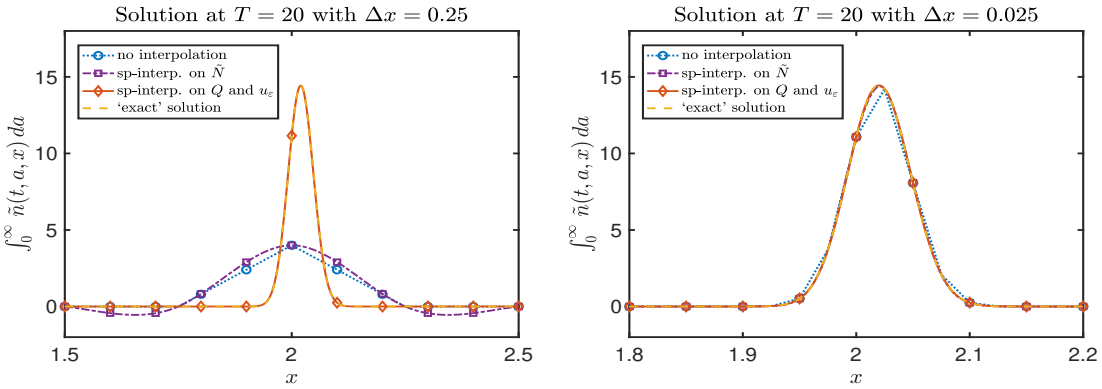

FIG. 5.5. *Comparison of the reconstructed solutions* $\int_0^\infty \tilde{n}(t,a,x)\,da$ *at* $T = 20$ *with the 'exact' one. Here we choose* $\varepsilon = 0.01$, $\Delta t = 0.5$, $\Delta a = 0.02$ *and compute the 'exact' solution on a fine mesh with* $\Delta x = 0.001$. *The reconstructed solutions are firstly computed on a coarse mesh and then interpolated onto a fine mesh with the spline interpolation method ('sp-interp.'). The left figure and the right figure show the solutions computed on coarse meshes with* $\Delta x = 0.25$ *and* $\Delta x = 0.025$, *respectively. Here we do the interpolation in two ways. In one way, we interpolate* $\tilde{N}$ *directly. In the other way, we do interpolation to* Q *and* $\mathbf{u}$ *and then reconstruct* $\tilde{N}$ *via* (3.3).

shows the comparison of the numerical solutions computed via our scheme (3.15)-(3.23) with a fixed mesh size $\Delta a = \Delta x = 0.01$ and different choices of the time step Δt. The 'exact' solution is assumed to be the one computed with $\Delta t = 0.1$. On one hand, it is easy to observe from the figure the spectral accuracy in time. On the other hand, we find our scheme is robust and efficient since the solution is somewhat accurate even with only one time step, i.e. the time step is chosen to be equal to the final time. Intuitively, the high accuracy of Q^n is due to its asymptotic behaviour shown in Theorem 3.6. As a remark, although $\tilde{N}^n$ will finally be less accurate with a smaller ε, which is reasonable since the numerical error is amplified by the exponential part $e^{\frac{u_\varepsilon(t,x)}{\varepsilon}}$, the error is still small and we can observe the spectral accuracy in time as well. To sum up, our scheme is stable, efficient and accurate, and thus works perfectly in the case without mutation.

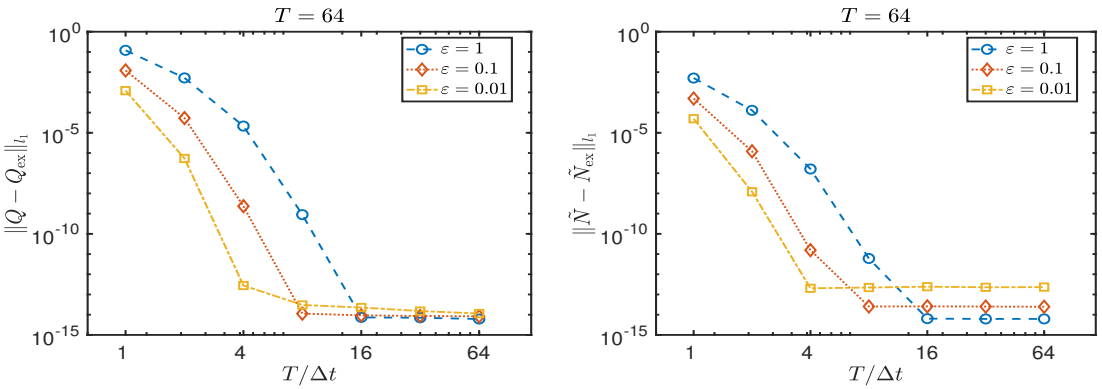

FIG. 5.6. *Temporal convergence of* Q *(left) and* $\tilde{N}$ *(right) under the* l_1*-norm at time* $T = 64$ *with a fixed mesh and different choices of time steps. The 'exact' solutions* Q_{ex} *and* $\tilde{N}_{\text{ex}}$ *are chosen to be the one computed with the same mesh and an extremely small time step.*

5.2. Case with mutation. Now we apply our method to the case where the mutation effect is included. The mutation between different traits prevents the normalized population to converge to a multiple of the Dirac function in the x-direction. In our model, two parameters, i.e. m and ε, are applied to describe the mutation effect. Intuitively, the parameter m measures the frequency of the mutation and the parameter ε measures the effect of mutations on the phenotype of new borns. Figure 5.7 shows the effect of the two parameters on the normalized population. Obviously, the concentration of the normalized population will be less obvious with a stronger mutation effect, i.e. when m and ε are large.

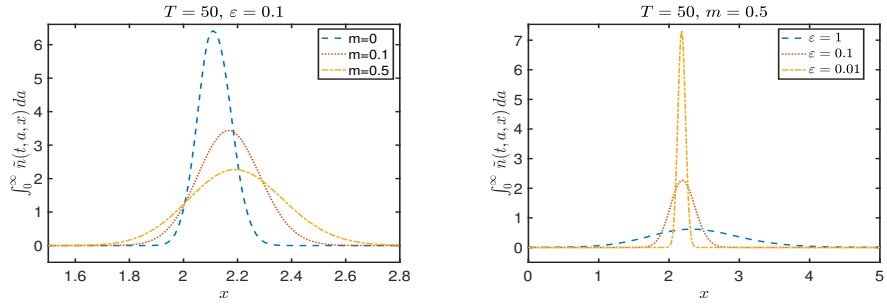

FIG. 5.7. *Illustration of the mutation effect parameterized via ε and m.*

Figure 5.8 shows the accuracy of $\tilde{N}^n$ computed via our method in a- and x-directions. As shown in the figure, we have a roughly first order accuracy in a-direction but a nearly second order accuracy in x-direction, which is similar to the case without mutation but somewhat unexpected since the upwind scheme, which is only first order accurate, is applied to evaluate $\tilde{\eta}_k^n$ (4.19).

As in the case without mutation, we can rebuild accurately the solution on a fine mesh in x-direction via a high-accurate interpolation method as long as we evaluate $\delta_x^{\pm}\Lambda_k(\tilde{\eta}_k^n)$ in (4.24) in a more accurate way and replace the linear interpolations of $q_{j,k-\tilde{\varepsilon}l}^{n+1}$ and $\tilde{u}_{k-\tilde{\varepsilon}l}^{n+1}$ (4.31) by the corresponding high-accurate interpolations. Unlike the linear interpolation case, where the system (4.29) is linear and the matrix can be formulated explicitly, the system now becomes nonlinear. One simple way to solve the system is via an iterative semi-implicit solver, where the mutation part in (4.29) is treated explicitly while all the other parts are treated implicitly as before. Figure 5.9 shows the accuracy of the normalized population $\tilde{N}$, which are interpolated in different ways in x-direction. As shown in the figure, we observe that we can use few points in the x-direction to accurately rebuild the solution on a fine mesh and the spline interpolation obviously benefits from the high accuracy.

Figure 5.10 shows the temporal accuracy of the normalized population $\tilde{N}$ at $T = 64$. Unfortunately, we no longer have the spectral convergence in time as shown in Figure 5.6, where no mutation effect is considered. It is due to the fact that the eigenpair $(\Lambda_k(\eta_k), \mathbf{n}_k(\eta_k))$, where $k \in [\![0, N_x]\!]$, now depends on time, which indicates that we can no longer evaluate the integral

$$\int_{t_n}^{t_{n+1}} \Lambda_k(\bar{\eta}[\partial_x u_\varepsilon|_{x=x_k}])\,ds \tag{5.5}$$

exactly to accurately approximate u_ε. Though the first order accuracy is expected since the backward Euler method is applied in the schemes (4.20) and (4.29), Figure 5.10 shows a nearly second order accuracy in time. Besides, Figure 5.10 shows

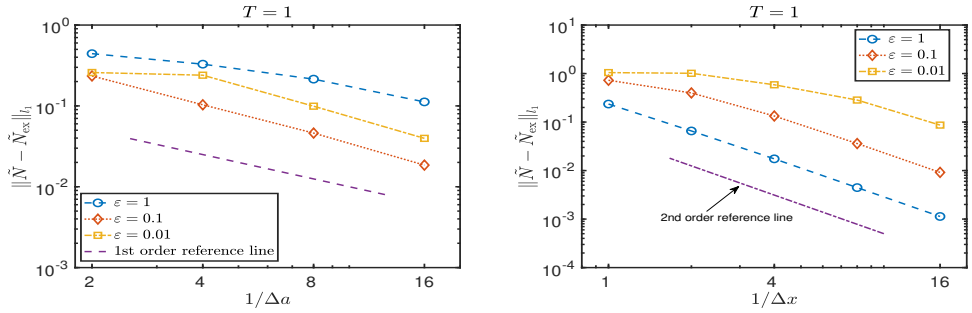

FIG. 5.8. *Accuracy of $\tilde{N}$ at time $T = 1$ with fixed $\Delta t = 0.02$ and $m = 0.5$. The figure on the left shows the accuracy in a-direction, where we fix $\Delta x = 0.02$ and compute the 'exact' one with $\Delta a = 0.02$. The figure on the right shows the accuracy in x-direction, where we fix $\Delta a = 0.02$ and compute the 'exact' one with $\Delta x = 0.02$.*

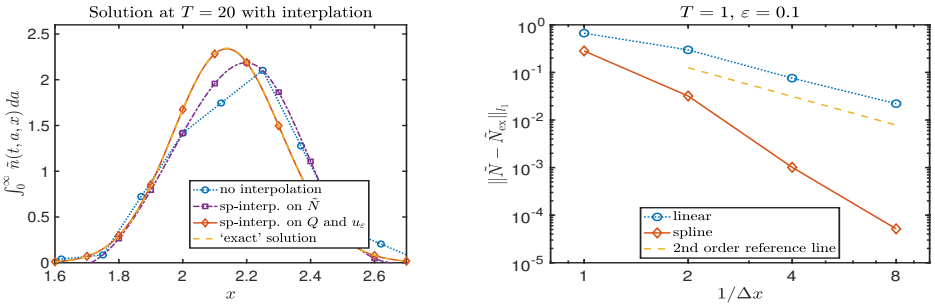

FIG. 5.9. *The left figure compares $\int_0^\infty \tilde{n}(t, a, x)\, da$ at $T = 20$. Here we fix $\varepsilon = 0.1$ and $m = 0.5$. The interpolated solutions are computed on a coarse mesh with $\Delta x = 0.25$ while the 'exact' solution is computed on a fine mesh with $\Delta x = 0.01$. Again, we do the spline interpolation (sp-interp.) in two ways — on $\tilde{N}$ directly or on Q and u_ε seperately. For numerical efficiency, here we choose $\Delta t = 0.5$ and $\Delta a = 0.02$. The right figure shows the order of accuracy at $T = 1$ of two different interpolation methods in x-direction with $\Delta t = \Delta a = 0.02$.*

that the normalized population will be less accurate with a smaller ε, which is reasonable since the inaccuracy of the approximation of the exponent $u_\varepsilon(t, x)/\varepsilon$ dominates the error.

6. Conclusion. We have presented an asymptotic preserving (A-P) scheme for capturing concentrations arising in the adaptive dynamics of the age-structured population. A proper WKB representation of the solution, which could perfectly describe the asymptotic behaviour of the solution in the case without mutation, was adopted to derive our scheme. Important properties of the scheme, including the A-P property, have been rigorously proved for the case. Extensive numerical experiments have been presented to show the robustness and efficiency of our scheme. In particular, we found that the concentrations on some particular phenotypical traits can be accurately captured with a rather coarse mesh and a nearly spectral accuracy in time can be observed. Then we generalized our scheme to the case with mutation. Though complicated, we showed in details the efficient and stable way to update the solution in each step. The A-P property was formally shown for the case. Numerical experiments showed that, though we would lose the spectral accuracy in time, we can still rebuild accurately the solution on a fine mesh with much fewer points in the

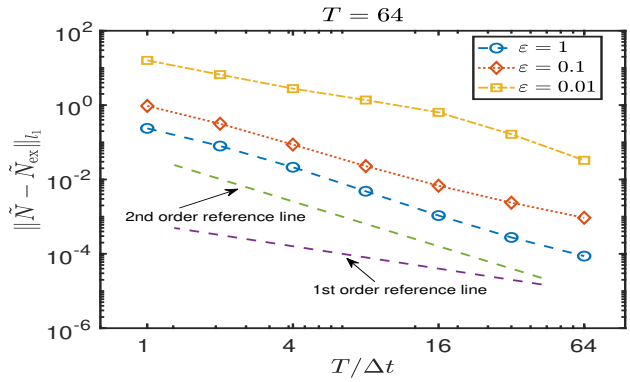

FIG. 5.10. *Temporal accuracy of $\tilde{N}$ at time $T = 64$ with $m = 0.5$, $\Delta a = 0.05$, $\Delta x = 0.001$ and different choices of ε. The 'exact' solutions $Q_{\rm ex}$ and $u_{\rm ex}$ are chosen to be the one computed with the same mesh and a small time step $\Delta t = 0.1$.*

phenotype space.

Appendix A. Reformulation of (3.7) and (4.16) with matrices. Define the parameter-dependent matrix $M_k(\eta)$ to be

$$M_k(\eta) = \begin{bmatrix} \tilde{d}_{1,k} - \tilde{b}_{1,k} & -\tilde{b}_{2,k} & -\tilde{b}_{3,k} & \dots & -\tilde{b}_{K_a-1,k} & -\tilde{b}_{K_a,k} \\ -\frac{1}{\Delta a} & \tilde{d}_{2,k} & 0 & \dots & 0 & 0 \\ 0 & -\frac{1}{\Delta a} & \tilde{d}_{3,k} & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \tilde{d}_{K_a-1,k} & 0 \\ 0 & 0 & 0 & \dots & -\frac{1}{\Delta a} & \tilde{d}_{K_a,k} \end{bmatrix}, \tag{A.1}$$

where $\tilde{b}_{j,k} = (1-m+m\eta)b(a_j, x_k)$ and $\tilde{d}_{j,k} = \frac{1}{\Delta a} + d(a_j, x_k)$. Then the equation (3.7) can be reformulated as

$$M_k \mathbf{n}_k = -\Lambda_k \mathbf{n}_k, \tag{A.2}$$

where $M_k := M_k(1)$ and $(-\Lambda_k, \mathbf{n}_k)$ is the leading eigenpair of the matrix M_k. Similarly, the equation (4.16) can be reformulated in the same form by replacing M_k in (A.2) by $M_k(\eta_k^n)$.

Appendix B. Fast computation of the eigenpair $(\Lambda_k, \mathbf{n}_k)$. For each $k \in [\![1, K_x]\!]$, we get the following relation by solving the first equation in (3.7) directly

$$N_{j,k} = \prod_{s=1}^{j} \frac{1}{1 + \Delta a(d(a_s, x_k) + \Lambda_k)} N_{0,k}. \tag{B.1}$$

As a result, we only need to solve Λ_k and $N_{0,k}$.

Substituting (B.1) into the boundary condition in (3.7) and then canceling out $N_{0,k}$ on both sides of the equation, we get

$$1 = \Delta a \sum_{j=1}^{K_a} b(a_j, x_k) \prod_{s=1}^{j} \frac{1}{1 + \Delta a(d(a_s, x_k) + \Lambda_k)}, \quad k \in [\![0, K_x]\!]. \tag{B.2}$$

Noticing that the right-hand side of the equation is monotone decreasing from ∞ to 0 over the feasible set Ω_k, the equation (B.2) has a unique solution inside Ω_k (3.14). Combining the normalization condition of $\mathbf{n}_k$ in (3.7) and the relation (B.1), we can further get

$$N_{0,k} = \frac{1}{\Delta a \left[w_0^{(a)} + \sum_{j=1}^{K_a} w_j^{(a)} \prod_{s=1}^{j} \frac{1}{1+\Delta a(d(a_s,x_k)+\Lambda_k)} \right]}. \tag{B.3}$$

Similarly, if we denote $\phi_{j,k} = r_{j,k}\phi_{0,k}$, then the first equation in (3.12) indicates that the coefficients $\{r_{j,k}\}$ can be computed recursively in the backward direction as

$$r_{K_a,k} = 0, \quad r_{j-1,k} = \frac{r_{j,k} + b(a_j, x_k)\Delta a}{1 + (d(a_j, x_k) + \Lambda_k)\Delta a}. \tag{B.4}$$

Then $r_{0,k} = 1$ is exactly the equation (B.2). The normalization condition in (3.12) implies that

$$\phi_{0,k} = \frac{1}{\Delta a \sum_{j=1}^{K_a} N_{j,k} r_{j-1,k}}. \tag{B.5}$$

As a remark, the recursion (B.4) in the forward direction will fail due to numerical instability.

Appendix C. Properties of the sequence $\{\boldsymbol{\lambda}^{(l)}\}$ in Algorithm 4.1. A direct observation of the sequence $\{\boldsymbol{\lambda}^{(l)}\}$ is that, for any $l \geq 0$, we have

$$\mathbf{Y}^n(\boldsymbol{\lambda}^{(l)}) \geq \mathbf{1}_{K_x+1}. \tag{C.1}$$

In fact, when $l = 0$, the conclusion holds true due to the non-negativity of $f_k^n(\boldsymbol{\lambda})$. When $l \geq 1$, noticing the relation (4.28) and the fact that $Y_k^n(\boldsymbol{\lambda})$ is convex, we have

$$\mathbf{Y}^n(\boldsymbol{\lambda}^{(l)}) \geq \mathbf{Y}^n(\boldsymbol{\lambda}^{(l-1)}) + [\mathbf{J}_{\mathbf{Y}^n}(\boldsymbol{\lambda}^{(l-1)})](\boldsymbol{\lambda}^{(l)} - \boldsymbol{\lambda}^{(l-1)}) = \mathbf{1}_{K_x+1}. \tag{C.2}$$

With this observation, we can further show the convergence of the sequence $\{\boldsymbol{\lambda}^{(l)}\}$. The result is summarized as the following lemma.

LEMMA C.1. *The sequence $\{\boldsymbol{\lambda}^{(l)}\}$ computed via Algorithm 4.1 converges. To be more specific,*

$$\lim_{l\to\infty} \boldsymbol{\lambda}^{(l)} = \boldsymbol{\lambda}^* \tag{C.3}$$

for some vector $\boldsymbol{\lambda}^*$, which is the unique solution of the nonlinear system (4.26) in the feasible set Ω.

Proof. On one hand, the sequence $\{\boldsymbol{\lambda}^{(l)}\}$ is non-decreasing in the sense

$$\Lambda_k^{(l+1)} \geq \Lambda_k^{(l)}, \quad k \in [\![0, K_x]\!], \tag{C.4}$$

since $[\mathbf{J}_{\mathbf{Y}^n}(\boldsymbol{\lambda}^{(l)})]^{-1}$ is elementwisely non-positive and $\mathbf{Y}^n(\boldsymbol{\lambda}^{(l)}) \geq \mathbf{1}_{K_x+1}$ (C.1).

On the other hand, we claim that, for each $n \geq 1$, there exists some bounded domain Ω^n such that $\{\boldsymbol{\lambda}^{(l)}\} \subset \Omega^n$. For any given vector $\boldsymbol{\lambda}$, denote $\bar{k}$ to be the index such that

$$\Lambda_{\bar{k}} = \|\boldsymbol{\lambda}\|_{l_\infty}. \tag{C.5}$$

As a result, $\delta_x^- \Lambda_{\bar{k}} \geq 0$ and $\delta_x^+ \Lambda_{\bar{k}} \leq 0$, which implies that

$$\tilde{\eta}_{\bar{k}}^n \leq g(\delta_x^- u_{\bar{k}}^n, \delta_x^+ u_{\bar{k}}^n) \leq g(\min_k\{\delta_x^- u_k^n\}, \max_k\{\delta_x^+ u_k^n\}), \tag{C.6}$$

where the last bound is independent of k. Therefore,

$$Y_{\bar{k}}^n(\boldsymbol{\lambda}) \leq [1 - m + mg(\min_k\{\delta_x^- u_k^n\}, \max_k\{\delta_x^+ u_k^n\})] S_{\bar{k}}(\Lambda_{\bar{k}}). \tag{C.7}$$

Obviously, the right-hand side of (C.7) goes to 0 as $\Lambda_{\bar{k}} \to \infty$, which indicates that there exists some number R^n such that $Y_{\bar{k}}^n(\boldsymbol{\lambda}) < 1$ if $\Lambda_{\bar{k}} > R^n$ (or equivalently $\|\boldsymbol{\lambda}\|_{l^\infty} > R^n$). Now we take $\Omega^n = \{\boldsymbol{\lambda} \,|\, \|\boldsymbol{\lambda}\|_{l^\infty} \leq R^n\} \cap \Omega$. Then the fact (C.1) implies that $\{\boldsymbol{\lambda}^{(l)}\} \subset \Omega^n$.

As a result, there exists a limit $\boldsymbol{\lambda}^*$ of the sequence $\{\boldsymbol{\lambda}^{(l)}\}$ in Ω. It is easy to see that $\boldsymbol{\lambda}^*$ is the solution of the system (4.26) by taking the limit $l \to \infty$ in (4.28). Assume that there exists another solution $\tilde{\boldsymbol{\lambda}}^*$, then we must have

$$\mathbf{Y}^n(\tilde{\boldsymbol{\lambda}}^*) = \mathbf{Y}^n(\boldsymbol{\lambda}^*) + \mathbf{J}_{\mathbf{Y}^n}(\tilde{\boldsymbol{\lambda}})(\tilde{\boldsymbol{\lambda}}^* - \boldsymbol{\lambda}^*) \tag{C.8}$$

for some $\tilde{\boldsymbol{\lambda}} \in \Omega$. Since $\mathbf{Y}^n(\tilde{\boldsymbol{\lambda}}^*) = \mathbf{Y}^n(\boldsymbol{\lambda}^*) = \mathbf{1}_{K_x+1}$ and the matrix $[\mathbf{J}_{\mathbf{Y}^n}(\boldsymbol{\lambda})]^{-1}$ exists everywhere, we must have $\tilde{\boldsymbol{\lambda}}^* = \boldsymbol{\lambda}^*$, which immediately implies the uniqueness of the solution. ∎

REFERENCES

[1] R. Abgrall and C.-W. Shu, eds., *Handbook of numerical methods for hyperbolic problems*, vol. 18 of Handbook of Numerical Analysis, North-Holland, Amsterdam, 2017.

[2] A. Arnal, T. Tissot, B. Ujvari, L. Nunney, E. Solary, L. Laplane, F. Bonhomme, M. Vittecoq, A. Tasiemski, F. Renaud, P. Pujol, B. Roche, and F. Thomas, *The guardians of inherited oncogenic vulnerabilities.*, Evolution, 70 (2016), pp. 1–6.

[3] G. Barles and B. Perthame, *Dirac concentrations in Lotka-Volterra parabolic PDEs*, Indiana Univ. Math J., 57 (2008), pp. 3275–3301.

[4] A. Calsina and S. Cuadrado, *Asymptotic stability of equilibria of selection-mutation equations*, J. Math. Biol., 54 (2007), pp. 489–511.

[5] N. Champagnat and P.-E. Jabin, *The evolutionary limit for models of populations interacting competitively via several resources*, J. Differ. Equations, 251 (2011), pp. 176–195.

[6] N. Champagnat, P.-E. Jabin, and G. Raoul, *Convergence to equilibrium in competitive Lotka-Volterra equations and chemostat systems*, C. R. Math., 348 (2010), pp. 1267–1272.

[7] Y. Cheng and C.-W. Shu, *A discontinuous Galerkin finite element method for directly solving the Hamilton-Jacobi equations*, J. Comput. Phys., 223 (2007), pp. 398–415.

[8] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, T. Am. Math. Soc., 277 (1983), pp. 1–42.

[9] ———, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comput., 43 (1984), pp. 1–19.

[10] M. G. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comput., 34 (1980), pp. 1–21.

[11] L. Desvillettes, P.-E. Jabin, S. Mischler, and G. Raoul, *On selection dynamics for continuous structured populations*, Commun. Math. Sci., 6 (2008), pp. 729–747.

[12] O. Diekmann, *A beginner's guide to adaptive dynamics*, Banach Cent., 63 (2004), pp. 47–86.

[13] O. Diekmann, P.-E. Jabin, S. Mischler, and B. Perthame, *The dynamics of adaptation: an illuminating example and a Hamilton-Jacobi approach*, Theor. Popul. Biol., 67 (2005), pp. 257–271.

[14] S. A. Frank, *Age-specific acceleration of cancer.*, Curr Biol, 14 (2004), pp. 242–246.

[15] S. A. H. Geritz, E. Kisdi, G. Meszena, and J. A. J. Metz, *Evolutionarily singular strategies and the adaptive growth and branching of the evolutionary tree*, Evol. Ecol., 12 (1998), pp. 35–57.

[16] H. Hivert, *A first-order asymptotic preserving scheme for front propagation in a one-dimensional kinetic reaction-transport equation*, J. Comput. Phys., 367 (2018), pp. 253–278.

[17] C. Hu and C.-W. Shu, *A discontinuous Galerkin finite element method for Hamilton-Jacobi equations*, SIAM J. Sci. Comput., 21 (1999), pp. 666–690.
[18] P.-E. Jabin and G. Raoul, *On selection dynamics for competitive interactions*, J. Math. Biol., 63 (2011), pp. 493–517.
[19] G.-S. Jiang and D. Peng, *Weighted ENO schemes for Hamilton-Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2126–2143.
[20] G.-S. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.
[21] A. Lorz, S. Mirrahimi, and B. Perthame, *Dirac mass dynamics in multidimensional nonlocal parabolic equations*, Commun. Part. Diff. Eq., 36 (2011), pp. 1071–1098.
[22] S. Méléard and V. C. Tran, *Slow and fast scales for superprocess limits of age-structured populations*, Stoch. Proc. Appl., 122 (2012), pp. 250–276.
[23] P. Michel, S. Mischler, and B. Perthame, *General relative entropy inequality: an illustration on growth models*, J. Math. Pure. Appl., 84 (2005), pp. 1235–1260.
[24] S. Nordmann, B. Perthame, and C. Taing, *Dynamics of concentration in a population model structured by age and a phenotypical trait*, Acta Appl. Math., 155 (2018), pp. 197–225.
[25] L. Nunney, *The real war on cancer: the evolutionary dynamics of cancer suppression.*, Evol Appl, 6 (2013), pp. 11–19.
[26] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton-Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.
[27] B. Perthame, *Transport equations in biology*, Frontiers in Mathematics, Birkhäuser Verlag, Basel, 2007.
[28] B. Perthame and P. E. Souganidis, *Rare mutations limit of a steady state dispersal evolution model*, Math. Model. Nat. Pheno., 11 (2016), pp. 154–166.
[29] T. Roget, *On the long-time behaviour of an age and trait structured population dynamics*. working paper or preprint, Nov. 2017.
[30] C.-W. Shu, *High order numerical methods for time dependent Hamilton-Jacobi equations*, in Mathematics and computation in imaging science and information processing, vol. 11 of Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, World Sci. Publ., Hackensack, NJ, 2007, pp. 47–91.
[31] ———, *High order weighted essentially nonoscillatory schemes for convection dominated problems*, SIAM Rev., 51 (2009), pp. 82–126.
[32] V. C. Tran, *Large population limit and time behaviour of a stochastic particle model describing an age-structured population*, ESAIM-Probab. Stat., 12 (2008), pp. 345–386.
[33] J. Yan and S. Osher, *A local discontinuous Galerkin method for directly solving Hamilton-Jacobi equations*, J. Comput. Phys., 230 (2011), pp. 232–244.