AN ENERGY STABLE AND POSITIVITY-PRESERVING SCHEME FOR THE MAXWELL-STEFAN DIFFUSION SYSTEM*

XIAOKAI HUO†, HAILIANG LIU‡, ATHANASIOS E. TZAVARAS†, AND SHUAIKUN WANG§

Abstract. We develop a new finite difference scheme for the Maxwell-Stefan diffusion system. The scheme is conservative, energy stable and positivity-preserving. These nice properties stem from a variational structure and are proved by reformulating the finite difference scheme into an equivalent optimization problem. The solution to the scheme emerges as the minimizer of the optimization problem, and as a consequence energy stability and positivity-preserving properties are obtained.

Key words. Finite difference, Maxwell-Stefan systems, Cross-diffusion, Positivity-preserving, Energy dissipation

AMS subject classifications. 35K55, 35Q79, 65M06, 35L45

1. Introduction.

Cross diffusion occurs in multicomponent systems, such as ionic liquids, wildlife populations, gas mixtures, tumor growth, etc [16, 19]. In these multicomponent systems, the diffusion happens not only in the direction from high concentration to low concentration, but also in the opposite direction due to cross diffusion. In such cases, diffusion can not be described by Fick's diffusion law and the Maxwell-Stefan diffusion model can be used instead. The Maxwell-Stefan model assumes the friction between two components is proportional to their difference in velocity and molecular fractions. It is widely used in modeling multicomponent systems.

In this work, we consider the Maxwell-Stefan diffusion system for an n-component mixture on the torus $\mathbb{T}^d$, which reads for $i = 1, ..., n$,

$$\partial_t \rho_i + \nabla \cdot (\rho_i v_i) = 0, \tag{1.1}$$

$$-\sum_{j=1}^n b_{ij}\rho_j(v_i - v_j) = \nabla \log \rho_i - \frac{1}{\sum_{j=1}^n \rho_j} \sum_{j=1}^n \rho_j \nabla \log \rho_j, \tag{1.2}$$

$$\sum_{j=1}^n \rho_j v_j = 0. \tag{1.3}$$

Here $x \in \mathbb{T}^d$, $\rho_i = \rho_i(x,t)$ and $v_i = v_i(x,t)$ are the density and velocity of the i-th component. The initial conditions are taken to be

$$\rho_i(x, 0) = \rho_{i0}(x), \ i = 1, \dots, n,$$

and we assume that

$$\rho_{i0}(x) > 0\,, \quad \text{and} \quad \sum_{j=1}^n \rho_{j0}(x) = 1 \quad \text{for } x \in \mathbb{T}^d. \tag{1.4}$$

*Submitted to the editors DATE.

Funding: Liu was partially supported by the National Science Foundation under Grant DMS1812666.

†Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia (xiaokai.huo@kaust.edu.sa, athanasios.tzavaras@kaust.edu.sa).

‡Mathematics Department, Iowa State University, Ames, IA 50011 (hliu@iastate.edu).

§School of Mathematics, Shandong University, Jinan 250100, China (skwang@email.sdu.edu.cn).

Solutions of (1.1) conserve the total density $\partial_t \sum_{i=1}^n \rho_i + \nabla \cdot \sum_{i=1}^n \rho_i v_i = 0$, and (1.3) imposes an average velocity of the mixture $v_{av} = \sum_{i=1}^n \rho_i v_i / \sum_{i=1}^n \rho_i = 0$ and that the total density $\sum_{i=1}^n \rho_i$ is conserved at each $x \in \mathbb{T}^d$. Hypothesis (1.4) then fixes the total density to

$$\sum_{j=1}^n \rho_j(x,t) = 1, \quad \text{for } x \in \mathbb{T}^d,\ t > 0. \tag{1.5}$$

Accordingly, (1.1)-(1.3) reduces to

$$\partial_t \rho_i + \nabla \cdot (\rho_i v_i) = 0, \tag{1.6}$$

$$\nabla \rho_i = -\sum_{j=1}^n b_{ij} \rho_i \rho_j (v_i - v_j), \tag{1.7}$$

$i = 1, \ldots, n$, which is the usual form of the Maxwell-Stefan diffusion system. We emphasize that assumption (1.5) is made to simplify notation. One may instead assume that the initial data satisfy $\rho_{i0}(x) > 0$ and $m(x) := \sum_{j=1}^n \rho_{j0}(x)$ is a bounded function and then all arguments are extended with the obvious modifications. The theoretical results are based on the hypothesis $\rho_{i0}(x) > 0$. Nevertheless, for initial data where some component touches zero, a scaling limiter developed in [21, 22] can be used to prepare positive initial data for the scheme, and such treatment does not destroy the scheme accuracy (this point is detailed in section 2.2).

The system (1.1)-(1.3) can be obtained as the high-friction limit of the multicomponent Euler equations [13].

$$\begin{aligned} &\partial_t \rho_i + \nabla \cdot (\rho_i v_i) = 0, \\ &\partial_t (\rho_i v_i) + \nabla \cdot (\rho v_i v_i) + \frac{\rho_i}{\varepsilon} \nabla \frac{\delta F(\rho)}{\delta \rho_i} = -\frac{1}{\varepsilon} \sum_{j=1}^n b_{ij} \rho_i \rho_j (v_i - v_j), \end{aligned} \tag{1.8}$$

when the total momentum (or the mean velocity) is zero. The energy functional $F(\rho)$ is given by

$$F(\rho) = \sum_{i=1}^n \int_{\mathbb{T}^d} \rho_i(x) \log \rho_i(x) dx. \tag{1.9}$$

It was proved in [13] that, when the total momentum is zero, the system (1.8) converges to (1.1)-(1.3) in the high-friction limit $\varepsilon \to 0$. Moreover, (1.1)-(1.3) can be regarded as a gradient flow for $F(\rho)$.

This raises the following question: Given densities $\rho^0 = (\rho_i^0)_{i=1}^n$, $\rho^1 = (\rho_i^1)_{i=1}^n$, with $\sum_i \rho_i^0 = \sum_i \rho_i^1 = 1$, consider the minimization problem

$$\min_{(\rho,v)\in K} \int_0^1 \int_{\mathbb{T}^d} \sum_{i,j=1}^n \frac{1}{4} b_{ij} \rho_i \rho_j (v_i - v_j)^2 dx dt \tag{1.10}$$

over the set

$$K = \Big\{ \rho = (\rho_1, ..., \rho_n),\, v = (v_1, ..., v_n) \ :\ \partial_t \rho_i + \nabla \cdot (\rho_i v_i) = 0, \quad i = 1, \ldots, n,$$

$$\sum_{j=1}^n \rho_j v_j = 0,\ \ \rho_i(0,x) = \rho_i^0(x),\ \rho_i(1,x) = \rho_i^1(x) \Big\}.$$

The problem (1.10) as the minimum of the frictional work is motivated by the well-known characterization of the Wasserstein distance in a one-component fluid obtained by Benamou-Brenier [1]. The study of this question will be given in a forthcoming work. The minimization (1.10) and the gradient structure of (1.1)-(1.3) detailed in [13], motivate us to use the work of friction as a building block for a numerical scheme of variational provenance – in the spirit of the well known JKO scheme [15] – in order to exploit the gradient structure of the Maxwell-Stefan system. This connection is pursued in the present work.

In this paper, we develop a new implicit-explicit finite difference scheme for the Maxwell-Stefan system (1.1)-(1.3) and prove that the scheme is energy dissipating and positivity preserving, for arbitrary time step and spatial meshes. The scheme in one dimension takes the form:

$$\frac{\rho_i^{k+1}-\rho_i^k}{\Delta t}+d_h(\hat\rho_i^k v_i^{k+1})=0, \tag{1.11}$$

$$-\sum_{j=1}^n b_{ij}\hat\rho_j^k(v_i^{k+1}-v_j^{k+1}) = D_h\log\rho_i^{k+1}-\frac{1}{\sum_{j=1}^n\hat\rho_j^k}\sum_{j=1}^n\hat\rho_j^k D_h\log\rho_j^{k+1}, \tag{1.12}$$

$$\sum_{j=1}^n \hat\rho_j^k v_j^{k+1}=0 \tag{1.13}$$

(for the d-dimensional case the reader is referred to Section 4). The subscript i refers to the i-th component and takes values $i=1,...,n$, while the superscript k refers to the k-th time step. The equations (1.11)-(1.13) are computed at spatial grid points ℓ or $\ell+\frac12$ of staggered lattices in a way precised in Section 2. The parameter Δt is the time step and h is the mesh size. The operators d_h, D_h are central difference operators, in one dimension, defined by

$$(d_h f_i)_\ell = \frac{f_{i,\ell+1/2}-f_{i,\ell-1/2}}{h}, \quad (D_h f_i)_{\ell+\frac12}=\frac{f_{i,\ell+1}-f_{i,\ell}}{h}, \tag{1.14}$$

where $\ell=\{1,\dots,N\}$, N the number of mesh intervals, and we set $(\hat f_i)_{\ell+\frac12}=\frac12(f_{i,\ell}+f_{i,\ell+1})$.

The scheme is induced by a spatial discretization of the constrained optimization problem (*cf.* (3.1))

$$\min_{\tilde K}\left\{\int_{\mathbb{T}^d}\Delta t\sum_{i,j=1}^n\frac14 b_{ij}\rho_i^k\rho_j^k|u_i-u_j|^2dx+\int_{\mathbb{T}^d}\sum_{j=1}^n\rho_j\log\rho_j\,dx\right\}, \tag{1.15}$$

where the set $\tilde K$ is defined to be

$$\tilde K=\left\{(\rho,v):\ \rho>0,\ \frac{\rho_i-\rho_i^k}{\Delta t}+\nabla\cdot(\rho_i^k u_i)=0,\ \sum_{i=1}^n\rho_i^k u_i=0\right\}.$$

The approach is motivated by the JKO-scheme [15] and the Benamou-Brenier interpretation of the Wasserstein distance [1], the latter suggesting an alternate variational scheme for nonlinear Fokker-Planck equations espoused in [20]. The novelty here is (i) that the limiting problem is a coupled parabolic system and (ii) that the mechanical friction is a complex interaction among the different components (see [2]) that is only captured in bulk by the dissipation functional (1.10). Nevertheless, this suffices in capturing the detailed interaction.

We show that there exists a discrete energy function which dissipates along time iterations, and that the numerical solutions for the densities generated by the scheme (1.11)-(1.13) preserve the positivity of the initial densities. The proof uses variational arguments and is based on the reformulation of the finite difference scheme as an equivalent optimization problem. An interesting feature is the role played by an elliptic operator $\mathcal{L}_\Phi$ defined in (2.4) and the induced dual norm (2.5). The reader familiar with the Wasserstein distance will recognize analogies with duality induced norms [23, 25, 24] appearing in the theory of nonlinear Fokker-Planck equations and induced by the metric tensor generating the Wasserstein metric.

A large literature [2, 3, 9, 10, 11, 16, 17, 18] employing diverse techniques has provided a basic theory for the Maxwell-Stefan system (1.1)-(1.3). The existence of global nonnegative weak solutions in $L^2([0,\infty); H^1(\mathbb{T}^d))$ was established in [18], while local existence of strong solutions is shown in [2, 11]. Explicit finite difference schemes were developed in [3, 9, 10]. The explicit scheme in [3] was formulated based on rewriting equation (1.6)-(1.7) with the first $n-1$ components. The scheme is easy to implement, a stability condition on the time step relative to the square of the spatial mesh size is required and no energy stability property is proved. The scheme in [9] is semi-implicit and linear, and it was shown to be mass conservative but the energy stability of the scheme is not addressed. A fully implicit Euler-Galerkin scheme is developed in [17] for the Maxwell-Stefan system coupled with a Poisson equation, which is positivity-preserving, energy stable and convergent. Recently, in [5], an implicit finite volume scheme was proposed for a cross-diffusion system similar to the Maxwell-Stefan system. A nonlinear cutoff function was used to approximate the values at cell interfaces to ensure nonnegativity of solutions. Both schemes in [17] and [5] incorporate the entropy structure to ensure the energy stable property. The scheme proposed here is positivity preserving, entropy decreasing and provides a connection between the finite difference scheme and a variational minimization problem. Both the energy stability and the positivity of solutions follow directly from the property of the variational structure. The approach is quite robust and we expect that, once the theory for the continuous problem (1.15) is further developed, it will lead to theoretical results for more complicated schemes such as finite element methods.

Recently there has been a growing interest in developing energy stable and/or positivity-preserving numerical schemes for nonlinear diffusion equations [6, 7, 12, 14, 21, 22, 26]. Positivity-preserving schemes for the Poisson-Nernst-Planck systems were developed in [21, 22], where the maximum principle was used to show the non-negativity of the scheme. A series of diffusion equations satisfying a gradient flow structure was considered in [6, 7, 12, 26], where energy-stabie schemes were developed for the Cahn-Hillard equations, with positivity-preserving properties proved in [6, 7] via optimization formulations. The technique was also used in [14] to prove the positivity and energy-stability properties for a scheme associated to the quantum diffusion equation. Our approach extends these works to a setting of systems that are gradient flows by exploiting the frictional dissipation natural to the Maxwell-Stefan system.

The structure of the paper is as follows: in Section 2, we give the details of the numerical scheme and show that it conserves the total mass and is consistent. In Section 3, we first prove that the numerical scheme is equivalent to an optimization problem, in Theorem 3.1, and then show the energy stability and positivity-preserving properties in Theorem 3.6. We provide the multidimensional scheme in Section 4 and show that similar properties also hold. Finally, we give some numerical examples to verify the proved properties.

2. The scheme.

2.1. Notations. We use notations from [27]. We define the following two grids on the torus $\mathbb{T} = [0, L]$ with mesh size $h = L/N$, where N is the number of mesh intervals:

$$\mathcal{C} := \{h, 2h, \ldots, L\}, \quad \mathcal{E} := \left\{ \frac{h}{2}, \frac{3h}{2}, \ldots, (N - \tfrac{1}{2})h \right\}. \tag{2.1}$$

We define the discrete N-periodic function spaces as

$$\mathcal{C}_{\rm per} := \{f : \mathcal{C} \to \mathbb{R}\}, \quad \mathcal{E}_{\rm per} := \{f : \mathcal{E} \to \mathbb{R}\}.$$

Here we call $\mathcal{C}_{\rm per}$ the space of *cell centered functions* and $\mathcal{E}_{\rm per}$ the space of *edge centered functions*. We use f_ℓ to denote the value of function f at grid point $x_\ell = \ell h$. We also define the subspace $\mathring{\mathcal{C}}_{\rm per} := \left\{ f : f \in \mathcal{C}_{\rm per}, \sum_{\ell=1}^N f_\ell = 0 \right\}$. We can extend the above definitions to vector value functions. For example, we define $\mathcal{C}^n_{\rm per}$ by

$$\mathcal{C}^n_{\rm per} := \{f = (f_1, \ldots, f_n) : f_i \in \mathcal{C}_{\rm per}, \quad i = 1, \ldots, n\}.$$

The spaces $\mathcal{E}^n_{\rm per}$, $\mathring{\mathcal{C}}^n_{\rm per}$ are defined the same way. The discrete gradients D_h and d_h are defined in (1.14). We define the average of the function values of nearby points by

$$\hat{f}_{\ell+\frac{1}{2}} = \frac{f_\ell + f_{\ell+1}}{2}, \text{ if } f \in \mathcal{C}_{\rm per}, \quad \text{and} \quad \hat{f}_\ell = \frac{f_{\ell+\frac{1}{2}} + f_{\ell-\frac{1}{2}}}{2}, \text{ if } f \in \mathcal{E}_{\rm per}. \tag{2.2}$$

The inner products are defined by $\langle f, g\rangle := h \sum_{\ell=1}^N f_\ell g_\ell, \ \ \forall f, g \in \mathcal{C}_{\rm per}$, and $[f, g] := h \sum_{\ell=1}^N f_{\ell+\frac{1}{2}} g_{\ell+\frac{1}{2}}, \ \ \forall f, g \in \mathcal{E}_{\rm per}$. They can be also extended on $\mathcal{C}^n_{\rm per}$ and $\mathcal{E}^n_{\rm per}$ with

$$\langle f, g\rangle := h \sum_{i=1}^n \sum_{\ell=1}^N f_{i,\ell} g_{i,\ell}, \ \ \forall f, g \in \mathcal{C}^n_{\rm per}, \quad [f, g] := h \sum_{i=1}^n \sum_{\ell=1}^N f_{i,\ell+\frac{1}{2}} g_{i,\ell+\frac{1}{2}}.$$

We also take the following notation:

$$\langle f\rangle := h \sum_{\ell=1}^N f_\ell, \quad f \in \mathcal{C}_{\rm per}, \quad [f] := h \sum_{\ell=1}^N f_{\ell+\frac{1}{2}}, \quad f \in \mathcal{E}_{\rm per}.$$

Suppose $f \in \mathcal{C}_{\rm per}$ and $\phi \in \mathcal{E}_{\rm per}$, the following summation-by-parts formula holds:

$$\langle f, d_h \phi\rangle = -[D_h f, \phi]. \tag{2.3}$$

Next, we introduce a norm on $\mathring{\mathcal{C}}^{n-1}_{\rm per}$. Let Φ be an $(n-1) \times (n-1)$ symmetric, positive definite matrix, with $\Phi_{ij} \in \mathcal{E}_{\rm per}$, $i, j = 1, \ldots, n-1$. We introduce the operator $\mathcal{L}_\Phi$ on $\mathring{\mathcal{C}}^{n-1}_{\rm per}$ defined by

$$\mathcal{L}_\Phi f := -d_h(\Phi D_h f) = \left(-\sum_{j=1}^{n-1} d_h(\Phi_{ij} D_h f_j) \right), \quad \forall f \in \mathring{\mathcal{C}}^{n-1}_{\rm per}. \tag{2.4}$$

Since Φ_{ij} are nonsingular for any point on $\mathcal{E}$, $\mathcal{L}_\Phi$ is invertible on $\mathring{\mathcal{C}}^{n-1}_{\rm per}$ (by the Lax-Milgram theorem). For any $g \in \mathring{\mathcal{C}}^{n-1}_{\rm per}$, let f be determined by $g = \mathcal{L}_\Phi f$, we define the norm

$$\|g\|^2_{\mathcal{L}_\Phi^{-1}} := [D_h f, \Phi D_h f]. \tag{2.5}$$

2.2. The scheme. The scheme (1.11)-(1.13) is written in the component form as follows:

$$\frac{\rho_{i,\ell}^{k+1}-\rho_{i,\ell}^{k}}{\Delta t} = -\frac{1}{h}\left(\hat\rho^k_{i,\ell+\frac12} v^{k+1}_{i,\ell+\frac12} - \hat\rho^k_{i,\ell-\frac12} v^{k+1}_{i,\ell-\frac12}\right), \tag{2.6}$$

$$-\sum_{j=1}^n b_{ij}\hat\rho^k_{j,\ell+\frac12}(v^{k+1}_{i,\ell+\frac12}-v^{k+1}_{j,\ell+\frac12}) \tag{2.7}$$

$$= \frac{\log\rho^{k+1}_{i,\ell+1}-\log\rho^{k+1}_{i,\ell}}{h} - \frac{1}{h\sum_{j=1}^n \hat\rho^k_{j,\ell+\frac12}}\sum_{j=1}^n \hat\rho^k_{j,\ell+\frac12}(\log\rho^{k+1}_{j,\ell+1}-\log\rho^{k+1}_{j,\ell}),$$

$$\sum_{j=1}^n \hat\rho^k_{j,\ell+\frac12} v^{k+1}_{j,\ell+\frac12} = 0, \tag{2.8}$$

subject to initial data

$$\rho^0_{i,\ell} = \rho_{i0}(x_\ell), \quad i=1,\dots,n, \quad \ell=1,\dots,N. \tag{2.9}$$

if $\rho_{i0}(x_\ell) > 0$; otherwise if $\rho_{i0}(x_\ell) = 0$ for some ℓ and $\sum_{\ell=1}^N \rho_{i0}(x_\ell) > 0$, we will impose a scaling limiter so that the obtained $\rho^0_{i,\ell}$ satisfy three properties: (i) positive for all ℓ; (ii) mass is preserved in the sense that

$$\sum_{\ell=1}^N \rho^0_{i,\ell} = \sum_{\ell=1}^N \rho_{i0}(x_\ell),$$

and (iii) accuracy of the scheme is not destroyed. For instance, it suffices to have $\max_\ell|\rho^0_{i,\ell}-\rho_{i0}(x_\ell)| \le O(h^r), r>2$. To achieve this, we use the limiter in [21, 22] where the above three properties are rigorously proved. For $\sum_{\ell=1}^N \rho_{i0}(x_\ell) = 0$, we simply remove this component from the system.

Next we study the conservation properties of the scheme. First we show that, at each grid point, the total density is preserved.

Lemma 2.1. *Suppose the solutions to the scheme* (1.11)-(1.13) *are positive for* $k \ge 1$. *Then the total mass at each grid point is conserved, i.e.*

$$\sum_{i=1}^n \rho^k_{i,\ell} = \sum_{i=1}^n \rho^0_{i,\ell}, \qquad \ell=1,\dots,N \text{ and } k\ge 1. \tag{2.10}$$

Proof. From equations (2.6) and (2.8), we have for $\ell = 1,\dots,N$,

$$\sum_{i=1}^n \rho^{k+1}_{i,\ell} = \sum_{i=1}^n \rho^k_{i,\ell} - \Delta t\sum_{i=1}^n d_h(\hat\rho^k_i v^{k+1}_i)_\ell$$

$$= \sum_{i=1}^n \rho^k_{i,\ell} - \frac{\Delta t}{h}\left(\sum_{i=1}^n \hat\rho^k_{i,\ell+\frac12}v^{k+1}_{i,\ell+\frac12} - \sum_{i=1}^n \hat\rho^k_{i,\ell-\frac12}v^{k+1}_{i,\ell-\frac12}\right) = \sum_{i=1}^n \rho^k_{i,\ell}.$$

This holds for any k, hence (2.10). ∎

Next, we show that for each component, the mass is conserved, i.e. the summation over grid points is conserved. The following lemma holds.

LEMMA 2.2. *Suppose the solutions to the scheme* (1.11)-(1.13) *are positive for any* $k \geq 1$. *Then the mass for each component is conserved, i.e.,*

$$\sum_{\ell=1}^{N} \rho_{i,\ell}^{k} = \sum_{\ell=1}^{N} \rho_{i,\ell}^{0}, \qquad i = 1, \ldots, n,\ k \geq 1\,. \tag{2.11}$$

Proof. From (2.6), we get

$$\sum_{\ell=1}^{N} \rho_{i,\ell}^{k+1} = \sum_{\ell=1}^{N} \rho_{i,\ell}^{k} - \frac{\Delta t}{h} \sum_{\ell=1}^{N} \left(\hat{\rho}_{i,\ell+\frac{1}{2}}^{k} v_{i,\ell+\frac{1}{2}}^{k+1} - \hat{\rho}_{i,\ell-\frac{1}{2}}^{k} v_{i,\ell-\frac{1}{2}}^{k+1} \right) = \sum_{\ell=1}^{N} \rho_{i,\ell}^{k}.$$

Iterating in k we obtain (2.11). ∎

2.3. The scheme in $n-1$ components. We consider first the solvability of the algebraic system (1.2)-(1.3) under the hypothesis $b_{ij} > 0$. Since summing the equations (1.2) in $i = 1, \ldots, n$ equals zero, these n equations are not independent. One easily checks that for $\rho_i > 0$ the homogeneous system

$$-\sum_{j=1}^{n} b_{ij} \rho_j (v_i - v_j) = 0$$

has only the trivial solution $v_1 = \cdots = v_n$. Hence the null space has dimension one. The solution of (1.2)-(1.3) is given by the following lemma.

LEMMA 2.3. *Let* $\rho_i(x,t) > 0$, $x \in \mathbb{T}^d, t > 0$, $i = 1, \ldots, n$, *and suppose that* $b_{ij} > 0$ *and* $b_{ij} = b_{ji}$, *for* $i \neq j$ *and* $i, j = 1, ..., n$. *Then the algebraic system* (1.2), (1.3) *has a unique solution that is explicitly expressed by*

$$\rho_i v_i = -\sum_{j=1}^{n-1} D_{ij} \nabla(\log \rho_j - \log \rho_n),\ \ i = 1, \ldots, n-1, \quad \textit{and} \quad \rho_n v_n = -\sum_{i=1}^{n-1} \rho_i v_i,$$

where

$$D_{ij} = D_{ij}(\rho) = \sum_{s,m=1}^{n-1} Q_{is}^{-T} B_{sm}^{-1} Q_{mj}^{-1}, \quad i, j = 1, \ldots, n-1, \tag{2.12}$$

and

$$B_{ij} = B_{ij}(\rho) = \delta_{ij} \sum_{m=1}^{n} b_{im} \rho_i \rho_m - b_{ij} \rho_i \rho_j, \tag{2.13}$$

$$Q_{ij} = Q_{ij}(\rho) = \frac{1}{\rho_i} \delta_{ij} + \frac{1}{\rho_n} \tag{2.14}$$

$$(Q^{-1})_{ij} = Q_{ij}^{-1}(\rho) = \delta_{ij} \rho_i - \frac{\rho_i \rho_j}{\sum_{j=1}^{n} \rho_j}. \tag{2.15}$$

For $\rho > 0$, B is diagonally dominant and thus invertible. We note that $Q^T = Q$ and that by a direct computation $QQ^{-1} = Q^{-1}Q = \mathbb{I}$, where Q^{-1} is determined by (2.15); hence, Q is also invertible. The proof can be found in [13] or [28]. A similar formula is established for the numerical scheme (1.11)-(1.13):

LEMMA 2.4. *Assume $b_{ij} > 0$ and $b_{ij} = b_{ji}$ for $i \neq j$ and $i, j = 1, \ldots, n$. Suppose $\rho^k_{i,\ell} > 0$ for $i = 1, \ldots, n$, $\ell = 1, \ldots, N$. The solutions of (1.12)-(1.13) are calculated by the explicit formula*

$$\hat\rho^k_i v^{k+1}_i = -\sum_{j=1}^{n-1} \hat D^k_{ij} D_h(\log \rho^{k+1}_j - \log \rho^{k+1}_n),\ i = 1, \ldots, n-1, \tag{2.16}$$

and $\hat\rho^k_n v^{k+1}_n = -\sum_{i=1}^{n-1} \hat\rho^k_i v^{k+1}_i$. Here

$$\hat D^k_{ij} = \sum_{s,m=1}^{n-1} (\hat Q^k)^{-T}_{is} (\hat B^k)^{-1}_{sm} (\hat Q^k)^{-1}_{mj}\,, \tag{2.17}$$

and $\hat Q^k_{ij} = Q_{ij}(\hat\rho^k)$, $\hat B^k_{ij} = B_{ij}(\hat\rho^k)$, $(\hat Q^k)^{-1}_{ij} = Q^{-1}_{ij}(\hat\rho^k)$ are the corresponding matrices (2.13)-(2.15) with ρ_i replaced by $\hat\rho^k_i$.

Notice that formulas (2.16) hold at each grid point $\ell+1/2 = 3/2, \ldots, N/2+1$(or $1/2$); to simplify the notation, we do not write the subscript $\ell + 1/2$.

Proof. Multiplying (1.12) by $\hat\rho^k_i$ gives

$$\hat\rho^k_i D_h \log \rho^{k+1}_i - \frac{\hat\rho^k_i}{\sum_{s=1}^n \hat\rho^k_s} \sum_{j=1}^n \hat\rho^k_j D_h \log \rho^{k+1}_j = -\sum_{j=1}^n b_{ij} \hat\rho^k_i \hat\rho^k_j (v^{k+1}_i - v^{k+1}_j)\,,$$

which is rewritten as

$$\sum_{j=1}^n \left(\delta_{ij} \hat\rho^k_i - \frac{\hat\rho^k_i \hat\rho^k_j}{\sum_{s=1}^n \hat\rho^k_s} \right) D_h \log \rho^{k+1}_j = -\sum_{j=1}^n \left(\delta_{ij} \sum_{m=1}^n b_{im} \hat\rho^k_i \hat\rho^k_m - b_{ij} \hat\rho^k_i \hat\rho^k_j \right) v^{k+1}_j. \tag{2.18}$$

Setting $\hat B^k_{ij} = B_{ij}(\hat\rho^k) = \delta_{ij} \sum_{m=1}^n b_{im} \hat\rho^k_i \hat\rho^k_m - b_{ij} \hat\rho^k_i \hat\rho^k_j$, the right side of (2.18) is expressed as

$$-\sum_{j=1}^n \hat B^k_{ij} v^{k+1}_j = -\sum_{j=1}^{n-1} \hat B^k_{ij} v^{k+1}_j - \hat B^k_{in} v^{k+1}_n = -\sum_{j=1}^{n-1} \hat B^k_{ij} (v^{k+1}_j - v^{k+1}_n). \tag{2.19}$$

Using (1.13) we get

$$\begin{aligned} -\sum_{j=1}^{n-1} \hat B^k_{ij} (v^{k+1}_j - v^{k+1}_n) &= -\sum_{j=1}^{n-1} \hat B^k_{ij} (v^{k+1}_j + \frac{1}{\hat\rho^k_n} \sum_{m=1}^{n-1} \hat\rho^k_m v^{k+1}_m) \\ &= -\sum_{j=1}^{n-1} \hat B^k_{ij} \sum_{m=1}^{n-1} (\frac{1}{\hat\rho^k_m} \delta_{jm} + \frac{1}{\hat\rho^k_n}) \hat\rho^k_m v^{k+1}_m = -\sum_{j,m=1}^{n-1} \hat B^k_{ij} \hat Q^k_{jm} \hat\rho^k_m v^{k+1}_m\,, \end{aligned} \tag{2.20}$$

where $\hat Q^k_{jm} = Q_{jm}(\hat\rho^k) = \frac{1}{\hat\rho^k_m} \delta_{jm} + \frac{1}{\hat\rho^k_n}$.By direct calculation it is shown that $\hat Q^k_{jm}$ is invertible with inverse $(\hat Q^k)^{-1}_{ij} = \left(\delta_{ij} \hat\rho^k_i - \frac{\hat\rho^k_i \hat\rho^k_j}{\sum_{s=1}^n \hat\rho^k_s} \right)$. The left side of (2.18) is rewritten for $i \neq n$ as

$$\sum_{j=1}^{n}\left(\delta_{ij}\hat\rho_i^k - \frac{\hat\rho_i^k\hat\rho_j^k}{\sum_{s=1}^n \hat\rho_s^k}\right) D_h \log\rho_j^{k+1}$$
$$= \sum_{j=1}^{n-1}(\hat Q^k)^{-1}_{ij} D_h\log\rho_j^{k+1} - \frac{\hat\rho_i^k(\sum_{j=1}^n\hat\rho_j^k - \sum_{j=1}^{n-1}\hat\rho_j^k)}{\sum_{s=1}^n\hat\rho_s^k} D_h\log\rho_n^{k+1}$$
$$= \sum_{j=1}^{n-1}(\hat Q^k)^{-1}_{ij} D_h(\log\rho_j^{k+1} - \log\rho_n^{k+1}).$$

This leads to expressing (2.18) as

$$\sum_{j=1}^{n-1}(\hat Q^k)^{-1}_{ij} D_h(\log\rho_j^{k+1} - \log\rho_n^{k+1}) = -\sum_{j,m=1}^{n-1}\hat B^k_{ij}\hat Q^k_{jm}\hat\rho^k_m v_m^{k+1}\,.$$

Since $\hat B^k$ and $\hat Q^k = (\hat Q^k)^T$ are invertible, we conclude that (2.16) holds. ∎

We adopt the notation

$$\tilde f = (f_1,\dots,f_{n-1}) \text{ for } f = (f_1,\dots,f_n). \tag{2.21}$$

With Lemma 2.4, the scheme (1.11)-(1.13) can be written as

$$\frac{\tilde\rho^{k+1}-\tilde\rho^k}{\Delta t} = -d_h\left(\hat D^k D_h\left(\frac{1}{h}\frac{\partial F_h}{\partial\tilde\rho}(\tilde\rho^{k+1})\right)\right),$$

where

$$F_h = F_h(\tilde\rho) := \left\langle \sum_{i=1}^{n-1}\rho_i\log\rho_i\right\rangle + \left\langle\left(1-\sum_{i=1}^{n-1}\rho_i\right)\log\left(1-\sum_{i=1}^{n-1}\rho_i\right)\right\rangle. \tag{2.22}$$

2.4. Consistency. Let (P,V) be the exact smooth solution of the equations (1.1)-(1.2) in the space $P,V\in C^3_{t,x}([0,T]\times\mathbb{T})$. The values at grid points are $P^k_{i,\ell} := P_i(x_\ell,k\Delta t), V^k_{i,\ell} := V_i(x_\ell,k\Delta t)$. The local truncation errors are defined by

$$\tau_i^1 = \frac{P_i^{k+1}-P_i^k}{\Delta t} + d_h(\hat P_i^k V_i^{k+1}),$$
$$\tau_i^2 = D_h\log P_i^{k+1} - \frac{1}{\sum_{j=1}^n\hat P_j^k}\sum_{i=1}^n\hat P_i^k D_h\log P_i^{k+1} + \sum_{j=1}^n b_{ij}\hat P_j^k(V_i^{k+1}-V_j^{k+1}),$$
$$\tau_i^3 = \sum_{i=1}^n\hat P_i^k V_i^{k+1}.$$

We have the following lemma.

LEMMA 2.5. *Suppose the solutions (P,V) to the system (1.1)-(1.3) are smooth in time and space, with $P,V\in C^3_{t,x}$ and $P_i(x,t)>0$ for $x\in\mathbb{T}$ and $t>0$ and for any $i=1,\dots,n$. Suppose (P,V) satisfies the condition (1.4). Then the local truncation errors satisfy*

$$|\tau^1_{i,\ell}|,\ |\tau^2_{i,\ell+\frac12}|,\ |\tau^3_{i,\ell+\frac12}| \le C(\Delta t + h^2).$$

Here $C>0$ is a positive constant depending on (P,V).

The elementary proof of this lemma is provided in the Supplementary material.

3. Optimization formulation.

3.1. Formulation via an optimization problem. In this section, we give an optimization formulation of the scheme (1.11)-(1.13). We recall that the system (1.1)-(1.3) can be written as the gradient flow of the energy functional (1.9), see [13]. Consider the minimization problem

$$\rho^{k+1} = \underset{\rho\geq 0, w}{\arg\min} \left\{ \frac{1}{\Delta t} \int_{\mathbb{T}^d} \sum_{i,j=1}^{n} \frac{1}{4} b_{ij} \rho_i^k \rho_j^k (w_i - w_j)^2 dx + F(\rho) \right\},$$

with $F(\rho)$ defined in (1.9), subject to the constraints

$$\rho_i - \rho_i^k + \nabla \cdot (\rho_i^k w_i) = 0, \quad i = 1, \dots, n, \quad \text{and} \quad \sum_{i=1}^{n} \rho_i^k w_i = 0.$$

The idea is to calculate minimizers of the free energy penalized by the work consumed by friction. The variational scheme is related to the Jordan-Kinderlehrer-Otto scheme [15], an analogy due to the connection between frictional dissipation and the Wasserstein distance offered by the Benamou-Brenier interpretation [1] of the Monge-Kantorovich mass transfer problem. There is however one important difference, as the frictional dissipation is more elaborate in the multi-component mixture situation.

The minimizers of the above constraint problem can be calculated by considering the min-max augmented Lagrangian

$$\begin{aligned}\min_{\rho,w} \max_{\alpha,\beta} L(\rho, w, \alpha, \beta) =& \frac{1}{\Delta t} \int_{\mathbb{T}^d} \sum_{i,j=1}^{n} \frac{1}{4} b_{ij} \rho_i^k \rho_j^k (w_i - w_j)^2 dx + \int_{\mathbb{T}^d} \sum_{j=1}^{n} \rho_j \log \rho_j \, dx \\ &+ \int_{\mathbb{T}^d} \alpha \sum_{i=1}^{n} \rho_i^k w_i dx + \int_{\mathbb{T}^d} \sum_{i=1}^{n} \big(\beta_i (\rho_i - \rho_i^k) - \nabla \beta_i \cdot (\rho_i^k w_i) \big) \, dx,\end{aligned}$$

Computing the variational derivatives gives:

$$\begin{aligned}
\frac{\delta L}{\delta \rho_i} = 0 \qquad &\text{implies} \qquad \log \rho_i + 1 + \beta_i = 0, \\
\frac{\delta L}{\delta w_i} = 0 \qquad &\text{implies} \qquad \frac{1}{\Delta t} \sum_{j=1}^{n} b_{ij} \rho_i^k \rho_j^k (w_i - w_j) + \alpha \rho_i^k - \rho_i^k \nabla \beta_i = 0, \\
\frac{\delta L}{\delta \alpha} = 0 \qquad &\text{implies} \qquad \sum_{i=1}^{n} \rho_i^k w_i = 0, \\
\frac{\delta L}{\delta \beta_i} = 0 \qquad &\text{implies} \qquad \rho_i - \rho_i^k + \nabla \cdot (\rho_i^k w_i) = 0.
\end{aligned}$$

Let $(\rho_i^{k+1}, w_i^{k+1})$ be the minimizer of the variational problem. Summing the second of the above equations over the index i and using the first implies

$$\alpha \sum_{i=1}^{n} \rho_i^k + \sum_{i=1}^{n} \rho_i^k \nabla \log \rho_i^{k+1} = 0\,.$$

Taking $v_i = w_i/\Delta t$, we get

$$\frac{\rho_i^{k+1} - \rho_i^k}{\Delta t} + \nabla \cdot (\rho_i^k v_i^{k+1}) = 0,$$

$$-\sum_{j=1}^n b_{ij}\rho_i^k \rho_j^k (v_i^{k+1} - v_j^{k+1}) = \rho_i^k \nabla \log \rho_i^{k+1} - \frac{\rho_i^k}{\sum_{j=1}^n \rho_j^k} \sum_{i=1}^n \rho_j^k \nabla \log \rho_j^{k+1},$$

$$\sum_{i=1}^n \rho_i^k v_i^{k+1} = 0.$$

The latter corresponds to an implicit-explicit discretization in time of the system (1.1)-(1.3).

Next we will give details of the optimization formulation for the fully discretized scheme (1.11)-(1.13).

We prove the following theorem.

THEOREM 3.1. *Assume $b_{ij} > 0$ and $b_{ij} = b_{ji}$ for $i \neq j$ and $i, j = 1, \ldots, n$. Given $\rho^k \in \mathcal{C}_{\mathrm{per}}$ with $\rho^k > 0$. There exists $\delta_0 > 0$ such that $\rho^{k+1} > 0$ is a solution of the numerical scheme* (1.11)-(1.13) *if and only if it is a minimizer of the optimization problem:*

$$\rho^{k+1} = \underset{(\rho,w)\in K_\delta}{\arg\min} \left\{ J = \frac{1}{4\Delta t} \left[\sum_{i,j=1}^n b_{ij} \hat{\rho}_i^k \hat{\rho}_j^k (w_i - w_j)^2 \right] + F_h(\rho) \right\}, \tag{3.1}$$

where $F_h(\rho) = \langle \sum_{i=1}^n \rho_i \log \rho_i \rangle$, and

$$K_\delta = \Big\{ (\rho, w) : \ \rho \in \mathcal{C}_{\mathrm{per}}^n, \ w \in \mathcal{E}_{\mathrm{per}}^n; \ \rho_{i,\ell} \geq \delta, \quad \rho_{i,\ell} - \rho_{i,\ell}^k + d_h(\hat{\rho}_i^k w_i)_\ell = 0,$$

$$\sum_{i=1}^n \hat{\rho}_{i,\ell+\frac{1}{2}}^k w_{i,\ell+\frac{1}{2}} = 0 \ \text{and} \ \sum_{i=1}^n \rho_{i,\ell} = 1, \ \forall i = 1, \ldots, n, \ \forall \ell = 1, \ldots, N \Big\},$$

for any $0 < \delta \leq \delta_0$.

We first prove a lemma that will be used later in the proof.

LEMMA 3.2. *Suppose Φ is an $(n-1)\times(n-1)$ symmetric positive definite matrix, with $\Phi_{ij} \in \mathcal{E}_{\mathrm{per}}$ for $i, j = 1, \ldots, n-1$. Suppose $\phi \in \mathring{\mathcal{C}}_{\mathrm{per}}^{n-1}$ is bounded in L^∞ satisfying $\|\phi\|_{L^\infty} \leq M$, where $\|\cdot\|_{L^\infty}$ is defined by*

$$\|\phi\|_{L^\infty} := \max_{\substack{i=1,\ldots,n-1\\ \ell=1,\ldots,N}} |\phi_{i,\ell}|.$$

Then the following estimate holds

$$\|\mathcal{L}_\Phi^{-1}\phi\|_{L^\infty} \leq \frac{CM}{\lambda_{\min}} h^{-\frac{1}{2}} (n-1)^{\frac{1}{2}},$$

where $C > 0$ is a constant independent of h, λ_{min} the minimum of all eigenvalues of Φ:

$$\lambda_{\min} = \min_{\ell=1,\ldots,N} \left\{ \lambda_\ell : \lambda_\ell \ \textit{is the eigenvalue of} \ (\Phi_{ij,\ell+\frac{1}{2}})_{(n-1)\times(n-1)} \right\}.$$

Proof. Since $\|\phi\|_{L^\infty} \le M$,

$$\|\phi\|_{L^2}^2 := h \sum_{\substack{i=1,\dots,n-1\\ \ell=1,\dots,N}} |\phi_{i,\ell}|^2 = h \sum_{\substack{i=1,\dots,n-1\\ \ell=1,\dots,N}} |M|^2 \le (n-1)hN|M|^2 = (n-1)L|M|^2.$$

Set $g = \phi \in \mathring{\mathcal{C}}_{\rm per}^{n-1}$, and $f = \mathcal{L}_\Phi^{-1} g$ in (2.5), we get

$$\|\phi\|^2_{\mathcal{L}_\Phi^{-1}} = [D_h f, \Phi D_h f].$$

Since Φ is positive definite so its minimum eigenvalue $\lambda_{\min} > 0$, we get

$$\lambda_{\min} \|D_h f\|_{L^2}^2 \le [D_h f, \Phi D_h f] = -\langle f, d_h(\Phi D_h f)\rangle = \langle f, \phi\rangle \le \|f\|_{L^2}\|\phi\|_{L^2}.$$

The use of the discrete Poincaré inequality gives $\|f\|_{L^2} \le C_P \|D_h f\|_{L^2}$. Therefore, we get

$$\|D_h f\|_{L^2} \le \frac{C_P}{\lambda_{\min}} \|\phi\|_{L^2}.$$

We can use the inequality $\|f\|_{L^\infty} \le C_P h^{-1/2} \|D_h f\|_{L^2}$, which follows from $\|f\|_{L^\infty}^2 = \max_{i=1,\dots,n} f_i^2 \le \sum_{i=1}^n f_i^2 \le h^{-1}\|f\|_{L^2}^2$ and the discrete Poincáre inequality. Applying this inverse inequality leads to

$$\|f\|_{L^\infty} \le C_P h^{-\frac12} \|D_h f\|_{L^2} \le \frac{C_P^2}{\lambda_{\min}} h^{-\frac12} L^{\frac12} M (n-1)^{\frac12} \le \frac{CM}{\lambda_{\min}} h^{-\frac12}(n-1)^{\frac12}. \qquad \square$$

Proof of Theorem 3.1. The proof is divided into three steps. In the first two steps, we prove that the optimization problem (3.1) has a unique interior minimizer and, in the last step, we prove that this minimizer is equivalent to the solution of the numerical scheme (1.11)-(1.13).

Step 1. Existence of the optimization problem. First we show existence for the optimization problem (3.1) for any $\delta > 0$. Notice that the objective function J in (3.1) is convex in w but it is not strictly convex. However, we can rewrite the optimization problem by using the first $n-1$ components of w and get an equivalent convex optimization problem. We introduce

$$W = (W_1, \dots, W_n), \quad W_i = \hat\rho_i^k w_i, \quad i = 1, \dots, n,$$

and so $\sum_{i=1}^n W_i = 0$. We adopt the notation (2.21) and define $\tilde W = (W_1, \dots, W_{n-1})$. We have the following lemma.

LEMMA 3.3. *The following formula holds:*

$$I(\tilde W) := \frac12 \sum_{i=1}^n b_{ij} \hat\rho_i^k \hat\rho_j^k (w_i - w_j)^2 = \tilde W^T (\hat Q^k)^T \hat B^k \hat Q^k \tilde W = \tilde W^T (\hat D^k)^{-1} \tilde W. \tag{3.2}$$

For $\hat\rho^k > 0$, the function $I : \mathbb{R}^{n-1} \to \mathbb{R}^+$ is strictly convex.

Proof. By the assumption that b_{ij} is symmetric, the following formula holds

$$\frac12 \sum_{i,j=1}^n b_{ij} \hat\rho_i^k \hat\rho_j^k (w_i - w_j)^2 = \sum_{i=1}^n w_i \sum_{j=1}^n b_{ij} \hat\rho_i^k \hat\rho_j^k (w_i - w_j).$$

Recalling (2.19), (2.20), we also have

$$\sum_{j=1}^{n} b_{ij}\hat{\rho}_i^k\hat{\rho}_j^k(w_i - w_j) = \sum_{j,m=1}^{n-1} \hat{B}_{ij}^k\hat{Q}_{jm}^k\hat{\rho}_m w_m$$

Therefore,

$$\begin{aligned}
&\frac{1}{2}\sum_{i,j=1}^{n} b_{ij}\hat{\rho}_i^k\hat{\rho}_j^k(w_i - w_j)^2 \\
&\quad= \sum_{i=1}^{n} w_i \sum_{j,m=1}^{n-1} \hat{B}_{ij}^k\hat{Q}_{jm}^k\hat{\rho}_m w_m \\
&\quad= \sum_{i=1}^{n-1} w_i \sum_{j,m=1}^{n-1} \hat{B}_{ij}^k\hat{Q}_{jm}^k\hat{\rho}_m w_m - \sum_{s=1}^{n-1}\frac{\hat{\rho}_s^k w_s}{\hat{\rho}_n^k}\sum_{j,m=1}^{n-1}\left(-\sum_{i=1}^{n-1}\hat{B}_{ij}^k\hat{Q}_{jm}^k\hat{\rho}_m^k w_m\right) \\
&\quad= \sum_{s,i,j,m=1}^{n-1} \hat{\rho}_s^k w_s\left(\frac{\delta_{is}}{\hat{\rho}_s^k} + \frac{1}{\hat{\rho}_n^k}\right)\hat{B}_{ij}^k\hat{Q}_{jm}^k\hat{\rho}_m^k w_m \\
&\quad= \sum_{s,i,j,m=1}^{n-1} \hat{\rho}_s^k w_s\hat{Q}_{is}^k\hat{B}_{ij}^k\hat{Q}_{jm}^k\hat{\rho}_m^k w_m = \tilde{W}^T(\hat{Q}^k)^T\hat{B}^k\hat{Q}^k\tilde{W}.
\end{aligned}$$

Notice that $\hat{B}^k$ is a symmetric strictly diagonally dominant matrix with positive diagonal entries since $\rho^k > 0$ and thus is positive definite. Because of this and since $\hat{Q}^k$ is non-singular, we have

$$(\hat{Q}^k)^T\hat{B}^k\hat{Q}^k \text{ is positive definite.}$$

Therefore, (3.2) is a convex function of $\tilde{W}$. ∎

We also need a lemma on the convexity of the discretized energy function $F_h(\tilde{\rho})$, defined by (2.22) that incorporates the constraint $\sum_{i=1}^n \rho_i = 1$.

LEMMA 3.4. *The energy function $F_h = F_h(\tilde{\rho})$ is a convex function of $\tilde{\rho}$.*

Proof. Considering the function

$$f = \sum_{i=1}^{n-1}\rho_i\log\rho_i + \rho_n\log\rho_n, \quad \rho_n = 1 - \sum_{i=1}^{n-1}\rho_i,$$

we have

$$\frac{\partial f}{\partial\rho_i} = \log\rho_i + 1 - (\log\rho_n + 1) = \log\rho_i - \log\rho_n, \quad \frac{\partial^2 f}{\partial\rho_i\partial\rho_j} = \frac{1}{\rho_i}\delta_{ij} + \frac{1}{\rho_n}.$$

Since for any $z \in \mathbb{R}^{n-1}$ and $z \neq 0$,

$$\sum_{i,j=1}^{n-1}\frac{\partial^2 f}{\partial\rho_i\partial\rho_j}z_i z_j = \sum_{i,j=1}^{n-1}\left(\frac{1}{\rho_i}\delta_{ij} + \frac{1}{\rho_n}\right)z_i z_j = \sum_{i=1}^{n-1}\frac{1}{\rho_i}z_i^2 + \frac{1}{\rho_n}\left(\sum_{i=1}^{n-1}z_i\right)^2 > 0,$$

the function f is a convex function of $\tilde{\rho}$. Therefore, $F_h(\tilde{\rho})$ is convex in $\tilde{\rho}$. ∎

Using Lemmas 3.3 and 3.4, we deduce that the optimization problem (3.1) is equivalent to

$$\min_{(\tilde\rho,\tilde W)\in\tilde K_\delta}\left\{J=\frac{1}{2\Delta t}\left[\tilde W^T(\hat Q^k)^T\hat B^k\hat Q^k\tilde W\right]+F_h(\tilde\rho)\right\}, \tag{3.3}$$

where

$$\tilde K_\delta=\{(\tilde\rho,\tilde W):\ \tilde\rho\in\mathcal{C}^{n-1}_{\rm per},\tilde W\in\mathcal{E}^{n-1}_{\rm per};\rho_{i,\ell}\ge\delta,\quad\sum_{i=1}^{n-1}\rho_{i,\ell}\le 1-\delta \text{ and}$$
$$\rho_{i,\ell}-\rho^k_{i,\ell}+d_h(W_i)_\ell=0,\ \forall i=1,\dots,n-1,\ell=1,\dots,N\}.$$

Due to the above lemmas, the objective function J is a convex function of $\tilde W$ and $\tilde\rho$ (note that $(\hat Q^k)^T\hat B^k\hat Q^k$ is a fixed matrix determined from the previous step). The domain $\tilde K_\delta$ is affine in $\tilde W$ and it is convex and bounded in $\tilde\rho$. The optimization problem (3.3) has a unique minimizer according to standard optimization theory [4]. Since the problems (3.1) and (3.3) are equivalent, there also exists a unique solution to the optimization problem (3.1).

Step2. The minimizer does not touch the boundary. Next, we show that there exists a constant $\delta_0>0$ such that the solution of the optimization problem (3.1) could not touch the boundary of K_δ for $\delta\le\delta_0$. Recall that on the set $\tilde K_\delta$,

$$\rho_i-\rho_i^k+d_h(W_i)=0.$$

Hence, if we set

$$\tilde W=\hat D^kD_h\tilde f,\quad \tilde g=\tilde\rho-\tilde\rho^k\in\mathring{\mathcal{C}}^{n-1}_{\rm per},$$

where $\tilde f\in\mathring{\mathcal{C}}^{n-1}_{\rm per}$ is uniquely defined by the first equation above, then according to the definition (2.5),

$$\left[\tilde W^T(\hat Q^k)^T\hat B^k\hat Q^k\tilde W\right]=[(D_h\tilde f)^T\hat D^kD_h\tilde f]=\|\tilde\rho-\tilde\rho^k\|^2_{\mathcal{L}^{-1}_{\hat D^k}}. \tag{3.4}$$

Therefore, the optimization problem (3.3) is equivalent to

$$\min_{\tilde\rho\in\mathring{\tilde K}_\delta}\left\{J=\frac{1}{2\Delta t}\|\tilde\rho-\tilde\rho^k\|^2_{\mathcal{L}^{-1}_{\hat D^k}}+F_h(\tilde\rho)\right\}, \tag{3.5}$$

over the set

$$\mathring{\tilde K}_\delta=\left\{\tilde\rho:\tilde\rho-\tilde\rho^k\in\mathring{\mathcal{C}}^{n-1}_{\rm per};\rho_{i,\ell}\ge\delta,\ \sum_{i=1}^{n-1}\rho_{i,\ell}\le1-\delta,\ \forall i=1,\dots,n-1,\ell=1,\dots,N\right\}.$$

Recall the notation $\tilde\rho=(\rho_1,...,\rho_{n-1})$ stands for the vector of the first $n-1$ densities which are computed at the grid points $l=1,...,N$. The density ρ_n appears in the formulation (3.5) only indirectly through the constraint (1.5). Also, $\tilde\rho-\tilde\rho^k\in\mathring{\mathcal{C}}^{n-1}_{\rm per}$ means $\sum_{\ell=1}^N(\rho_{i,\ell}-\rho^k_{i,\ell})=0$ for any $i=1,...,n-1$.

Let $\tilde\rho^\star\in\mathring{\tilde K}_\delta$ be a minimizer of the optimization problem (3.5). We will show that $\tilde\rho^\star$ does not lie on the boundary of $\mathring{\tilde K}_\delta$. If it lies on the boundary:

(i) either $\rho^\star_{i,\ell}=\delta$ for some $i=1,\dots,n-1$ at some grid point ℓ,

(ii) or $\sum_{i=1}^{n-1} \rho^\star_{i,\ell} = 1-\delta$ at some grid point ℓ.

First consider the case (i). Suppose that $\tilde{\rho}^\star$ touches the boundary at the grid point ℓ_0 for the i_0-th component, that is

$$\rho^\star_{i_0,\ell_0} = \delta. \tag{3.6}$$

We calculate the directional derivative of the objective function J at $\tilde{\rho}^\star$ along the direction $\{\nu : \nu \in \mathbb{R}^{(n-1)\times N}\}$ with $\tilde{\rho}^\star + s\nu \in \mathring{\tilde{K}}_\delta$ as

$$\begin{aligned}
(3.7)\quad & \frac{d}{ds} J(\tilde{\rho}^\star + s\nu)\Big|_{s=0} \\
&= \frac{d}{ds}\Big|_{s=0}\left(\frac{1}{2\Delta t}\|\tilde{\rho}^\star + s\nu - \tilde{\rho}^k\|^2_{\mathcal{L}^{-1}_{\hat{D}^k}} + F_h(\tilde{\rho}^\star + s\nu)\right) \\
&= \frac{1}{\Delta t}\left\langle \mathcal{L}^{-1}_{\hat{D}^k}(\tilde{\rho}^\star - \tilde{\rho}^k), \nu\right\rangle + \sum_{i=1}^{n-1}\left\langle \log\rho^\star_i + 1 - \log\left(1 - \sum_{j=1}^{n-1}\rho^\star_j\right) - 1, \nu_i\right\rangle \\
&= \frac{1}{\Delta t}\left\langle \mathcal{L}^{-1}_{\hat{D}^k}(\tilde{\rho}^\star - \tilde{\rho}^k), \nu\right\rangle + \sum_{i=1}^{n-1}\left\langle \left(\log\rho^\star_i - \log\left(1 - \sum_{j=1}^{n-1}\rho^\star_j\right)\right), \nu_i\right\rangle.
\end{aligned}$$

Here we use a contradiction argument, for which it suffices to find a direction ν such that the above directional derivative is negative. The first term on the right hand side of the above equation is bounded by Lemma 3.2, but the second term may become sufficiently negative as $\rho^\star_i = \delta$ or $1 - \sum_{i=1}^{n-1}\rho^\star_j = \delta$, at some point with a proper choice of ν. Based on this we argue in two cases respectively.

We divide the first case further into the following two cases:

(a)

$$\sum_{i=1}^{n-1}\rho^\star_{i,\ell_0} \geq \frac{1}{2},$$

(b)

$$\sum_{i=1}^{n-1}\rho^\star_{i,\ell_0} < \frac{1}{2}.$$

Case (i) **and** (a). Suppose $\{\rho^\star_{i,\ell_0}\}_{i=1}^{n-1}$ achieves its maximum at the i_1-th component while $\{\rho^\star_{i_0,\ell}\}_{\ell=1}^{N}$ achieves its maximum at ℓ_1. Define ν by

$$\nu_{i,\ell} = \begin{cases} 1, & \text{for } i = i_0,\ \ell = \ell_0, \\ -1, & \text{for } i = i_1,\ \ell = \ell_0, \\ -1, & \text{for } i = i_0,\ \ell = \ell_1, \\ 1, & \text{for } i = i_1,\ \ell = \ell_1, \\ 0, & \text{otherwise.} \end{cases}$$

Taking a variation in this direction, (3.7) becomes

$$\begin{aligned}
(3.8)\quad & \frac{1}{h}\frac{d}{ds} J(\tilde{\rho}^\star + s\nu)\Big|_{s=0} \\
&= \frac{1}{\Delta t}(\mathcal{L}^{-1}_{\hat{D}^k}(\tilde{\rho}^\star - \tilde{\rho}^k))_{i_0,\ell_0} - \frac{1}{\Delta t}(\mathcal{L}^{-1}_{\hat{D}^k}(\tilde{\rho}^\star - \tilde{\rho}^k))_{i_1,\ell_0} - \frac{1}{\Delta t}(\mathcal{L}^{-1}_{\hat{D}^k}(\tilde{\rho}^\star - \tilde{\rho}^k))_{i_0,\ell_1} \\
&\quad + \frac{1}{\Delta t}((\mathcal{L}^{-1}_{\hat{D}^k}(\tilde{\rho}^\star - \tilde{\rho}^k))_{i_1,\ell_1} + \log\rho^\star_{i_0,\ell_0} - \log\rho^\star_{i_1,\ell_0} - \log\rho^\star_{i_0,\ell_1} + \log\rho^\star_{i_1,\ell_1}.
\end{aligned}$$

Note the variation $\nu_{i,l}$ along which we calculate (3.7) is selected so that the contributions of the terms $\log\left(1-\sum_{j=1}^{n-1}\rho_j^*\right)$ cancel out.

Since $\{\rho^\star_{i,\ell_0}\}_{i=1}^{n-1}$ achieves its maximum for the i_1-th component, in the case (a) $\sum_{i=1}^{n-1}\rho^\star_{i,\ell_0}\geq\frac{1}{2}$ implies

$$\rho^\star_{i_1,\ell_0}\geq\frac{1}{2(n-1)}. \tag{3.9}$$

Since $\{\rho^\star_{i_0,\ell}\}_{\ell=1}^N$ achieves its maximum at the grid point ℓ_1 and $\tilde{\rho}^\star-\tilde{\rho}^k\in\mathring{\mathcal{C}}^{n-1}_{\text{per}}$,

$$\rho^\star_{i_0,\ell_1}\geq\frac{1}{N}\sum_{\ell=1}^N\rho^\star_{i_0,\ell}=\frac{1}{N}\sum_{\ell=1}^N\rho^k_{i_0,\ell}\geq\frac{m}{hN}, \tag{3.10}$$

where m is set to be $m:=\min_{i\in\{1,\ldots,n-1\}}\left\{h\sum_{\ell=1}^N\rho^k_{i,\ell}\right\}$. Moreover, for $\tilde{\rho}^\star\in\mathring{\tilde{K}}_\delta$ the constraint $\sum_{i=1}^{n-1}\rho^\star_{i,\ell_1}\leq 1-\delta$ implies

$$\rho^\star_{i_1,\ell_1}<1. \tag{3.11}$$

Next, we show that for δ satisfying

$$\delta\leq\min\left\{\frac{m}{2hN},\frac{1}{4(n-1)}\right\}, \tag{3.12}$$

if $s>0$ is selected sufficiently small and ν as above we have $\tilde{\rho}^\star+s\nu\in\mathring{\tilde{K}}_\delta$. Indeed,

$$\rho^\star_{i_0,\ell_0}+s=\delta+s\geq\delta,\quad \rho^\star_{i_1,\ell_1}+s\geq\delta+s,$$
$$\rho^\star_{i_0,\ell_1}-s\geq\frac{m}{hN}-s\geq\delta,\quad \rho^\star_{i_1,\ell_0}-s\geq\frac{1}{2(n-1)}-s\geq\delta,$$
$$\sum_{i=1}^{n-1}(\rho^\star_{i,\ell_0}+s\nu_{i,\ell_0})=\sum_{i=1}^{n-1}\rho^\star_{i,\ell_0}\leq 1-\delta,\quad \sum_{i=1}^{n-1}(\rho^\star_{i,\ell_1}+s\nu_{i,\ell_1})=\sum_{i=1}^{n-1}\rho^\star_{i,\ell_1}\leq 1-\delta,$$

imply that if δ satisfying (3.12) and for $s>0$ small we have $\tilde{\rho}^\star+s\nu\in\mathring{\tilde{K}}_\delta$.

Since $\tilde{\rho}^\star-\tilde{\rho}^k\in\mathring{\mathcal{C}}^{n-1}_{\text{per}}$ and $\|\tilde{\rho}^\star\|_{L^\infty},\|\tilde{\rho}^k\|_{L^\infty}\leq 1$, we can apply Lemma 3.2 to (3.8) with $\phi=\tilde{\rho}^\star-\tilde{\rho}^k$ and $\Phi=\hat{D}^k$ and use (3.6) and (3.9)-(3.11) to get

$$\frac{1}{h}\frac{d}{ds}J(\tilde{\rho}^\star+s\nu)\bigg|_{s=0}\leq\frac{8C}{\lambda^k_{\min}\Delta t}h^{-\frac{1}{2}}(n-1)^{\frac{1}{2}}+\log\delta-\log\frac{1}{2(n-1)}-\log\frac{m}{hN}+\log 1.$$

Here $\lambda^k_{\min}$ is the minimum eigenvalue of $\hat{D}^k$. Taking

$$\delta_0\leq\min\left\{\frac{m}{4(n-1)hN}e^{-\frac{8C}{\lambda^k_{\min}\Delta t}h^{-\frac{1}{2}}(n-1)^{\frac{1}{2}}},\frac{m}{2hN},\frac{1}{4(n-1)}\right\}, \tag{3.13}$$

we have for $\delta\leq\delta_0$, $\tilde{\rho}^\star+s\nu\in\mathring{\tilde{K}}_\delta$ and

$$\frac{1}{h}\frac{d}{ds}J(\tilde{\rho}^\star+s\nu)\bigg|_{s=0}\leq-\log 2<0. \tag{3.14}$$

This contradicts the assumption that $\tilde{\rho}^\star$ is a minimizer, and so the situation (a) cannot occur.

Case (i) **and** (b). Again $\rho_{i_0,\ell_0} = \delta$ and suppose now that $\{\rho^\star_{i_0,\ell}\}_{\ell=1}^N$ achieves its maximum at the ℓ_1-th grid point. We take

$$\nu_{i,\ell} = \begin{cases} 1, & \text{for } i = i_0,\ \ell = \ell_0, \\ -1, & \text{for } i = i_0,\ \ell = \ell_1, \\ 0, & \text{otherwise,} \end{cases}$$

and note that (3.10) still holds in the present setting. Using (3.6), (b), (3.10), and the inequality $1 - \sum_{i=1}^{n-1} \rho^\star_{i,\ell_1} \le 1 - (n-1)\delta \le 1$, we obtain

$$\begin{aligned} &\frac{1}{h}\frac{d}{ds} J(\tilde{\rho}^\star + s\nu)\bigg|_{s=0} \\ &\quad = \frac{1}{\Delta t}(\mathcal{L}^{-1}_{\hat{D}^k}(\tilde{\rho}^\star - \tilde{\rho}^k))_{i_0,\ell_0} + \log \rho^\star_{i_0,\ell_0} - \log\left(1 - \sum_{i=1}^{n-1} \rho^\star_{i,\ell_0}\right) \\ &\qquad - \frac{1}{\Delta t}(\mathcal{L}^{-1}_{\hat{D}^k}(\tilde{\rho}^\star - \tilde{\rho}^k))_{i_0,\ell_1} - \log \rho^\star_{i_0,\ell_1} + \log\left(1 - \sum_{i=1}^{n-1} \rho^\star_{i,\ell_1}\right) \\ &\quad \le \frac{4C}{\lambda^k_{\min}\Delta t} h^{-\frac{1}{2}}(n-1)^{\frac{1}{2}} + \log\delta - \log\frac{1}{2} - \log\frac{m}{hN} + \log 1 \\ &\quad \le \frac{4C}{\lambda^k_{\min}\Delta t} h^{-\frac{1}{2}}(n-1)^{\frac{1}{2}} + \log\delta - \log\frac{m}{2hN}. \end{aligned}$$

Taking

$$\delta_0 \le \min\left\{ \frac{m}{4hN} e^{-\frac{4C}{\lambda^k_{\min}\Delta t} h^{-\frac{1}{2}}(n-1)^{\frac{1}{2}}}, \frac{m}{2hN} \right\} \tag{3.15}$$

leads to $\tilde{\rho}^\star + s\nu \in \mathring{\tilde{K}}_\delta$ and

$$\frac{1}{h}\frac{d}{ds} J(\tilde{\rho}^\star + s\nu)\bigg|_{s=0} = -\log 2 < 0,$$

which contradicts the hypothesis that $\tilde{\rho}^\star$ is a minimizer; so the situation (b) cannot occur.

Case (ii). Assume there exists a grid index ℓ_0 such that

$$\sum_{i=1}^{n-1} \rho^\star_{i,\ell_0} = 1 - \delta, \tag{3.16}$$

and suppose the maximum value of $\{\rho^\star_{i,\ell_0}\}_{i=1}^{n-1}$ occurs at the index i_0. Then (3.16) implies that for $\delta \le 1/2$ equation (3.9) holds, that is

$$\rho^\star_{i_0,\ell_0} \ge \frac{1-\delta}{n-1} \ge \frac{1}{2(n-1)}. \tag{3.17}$$

Setting $\rho^k_{\min} := \min_{\substack{i=1,\dots,n,\\ \ell=1,\dots,N}} \rho^k_{i,\ell} > 0$, we have $\sum_{i=1}^{n-1} \rho^k_{i,\ell} = 1 - \rho^k_{n,\ell} \le 1 - \rho^k_{\min}$. Since $\tilde{\rho}^\star - \tilde{\rho}^k \in \mathring{C}^{n-1}_{\text{per}}$, we have

$$\sum_{\ell=1}^{N}\sum_{i=1}^{n-1} \rho^\star_{i,\ell} = \sum_{\ell=1}^{N}\sum_{i=1}^{n-1} \rho^k_{i,\ell} \le N(1 - \rho^k_{\min}).$$

Suppose $\left\{\sum_{i=1}^{n-1} \rho^\star_{i,\ell}\right\}_{\ell=1}^{N}$ achieves its minimum at the grid point ℓ_1. Then using (3.16) it follows for $\delta \le \frac{1}{2}\rho^k_{\min}$,

$$
\begin{aligned}
\sum_{i=1}^{n-1} \rho^\star_{i,\ell_1} &\le \tfrac{1}{N-1} \sum_{\substack{\ell=1,,,,,N\\ \ell\neq\ell_0}} \sum_{i=1}^{n-1} \rho^\star_{i,\ell} \\
&= \tfrac{1}{N-1}\left(\sum_{\ell=1}^{N}\sum_{i=1}^{n-1}\rho^\star_{i,\ell} - \sum_{i=1}^{n-1}\rho^\star_{i,\ell_0}\right) \\
&\le \tfrac{1}{N-1}\left(N(1-\rho^k_{\min}) - (1-\delta)\right) \\
&\le 1 - \frac{N\rho^k_{\min} - \delta}{N-1} \\
&\le 1 - \frac{2N-1}{2(N-1)}\rho^k_{\min}\,. \qquad (3.18)
\end{aligned}
$$

Taking now

$$
\nu_{i,\ell} = \begin{cases} -1, & \text{for } i = i_0,\ \ell = \ell_0, \\ 1, & \text{for } i = i_0,\ \ell = \ell_1, \\ 0, & \text{otherwise}, \end{cases}
$$

into (3.7) and using (3.16), (3.17), (3.18), Lemma 3.2, and the inequality $\rho^\star_{i_0,\ell_1} \le 1-\delta \le 1$ we obtain

$$
\begin{aligned}
&\frac{1}{h}\frac{d}{ds}J(\tilde\rho^\star + s\nu)\Big|_{s=0} \\
&\quad = -\frac{1}{\Delta t}(\mathcal{L}^{-1}_{\hat D^k}(\tilde\rho^\star - \tilde\rho^k))_{i_0,\ell_0} - \log\rho^\star_{i_0,\ell_0} + \log\left(1 - \sum_{i=1}^{n-1}\rho^\star_{i,\ell_0}\right) \\
&\qquad + \frac{1}{\Delta t}(\mathcal{L}^{-1}_{\hat D^k}(\tilde\rho^\star - \tilde\rho^k))_{i_0,\ell_1} + \log\rho^\star_{i_0,\ell_1} - \log\left(1 - \sum_{i=1}^{n-1}\rho^\star_{i,\ell_1}\right) \\
&\quad \le \frac{4C}{\lambda^k_{\min}\Delta t}h^{-\frac12}(n-1)^{\frac12} - \log\frac{1}{2(n-1)} + \log\delta + \log 1 - \log\frac{2N-1}{2(N-1)}\rho^k_{\min}.
\end{aligned}
$$

Taking

$$
\delta_0 \le \min\left\{\frac{(2N-1)\rho^k_{\min}}{8(N-1)(n-1)}e^{-\frac{4C}{\lambda^k_{\min}\Delta t}h^{-\frac12}(n-1)^{\frac12}}, \frac{1}{2}\rho^k_{\min}, \frac{1}{4(n-1)}\right\}, \qquad (3.19)
$$

we see that for $\delta < \delta_0$ the above inequality becomes negative. In addition,

$$
\begin{aligned}
&\rho^\star_{i_0,\ell_0} - s \ge \frac{1}{2(n-1)} - s \ge \delta, \quad \rho^\star_{i_0,\ell_1} + s \ge \delta + s \ge \delta, \\
&\sum_{i=1}^{n}\rho^\star_{i,\ell_0} - s = 1 - \delta - s \le 1 - \delta, \quad \sum_{i=1}^{n-1}\rho^\star_{i,\ell_1} + s \le 1 - \frac{2N-1}{N-1}\delta + s \le 1 - \delta,
\end{aligned}
$$

imply that for $\delta < \delta_0$ the variation $\tilde\rho^\star + s\nu \in \mathring{\tilde K}_\delta$ for sufficiently small $s > 0$. This contradicts the assumption that $\tilde\rho^\star$ is a minimizer and thus case (ii) cannot occur.

In summary, setting δ_0 to be the minimum among (3.13), (3.15) and (3.19) we conclude that (i) and (ii) cannot occur. Consequently, for $\delta \le \delta_0$, the minimizer to the optimization problem (3.5), or equivalently (3.1), does not occur at the boundary.

Step 3. The equivalence with the numerical scheme. Any interior minimizer $\tilde{\rho}^*$ of (3.5) must satisfies

$$\left\langle \frac{\partial J}{\partial \tilde{\rho}}(\tilde{\rho}^\star), \nu \right\rangle = 0, \tag{3.20}$$

for any $\nu \in \mathring{\mathcal{C}}_{\text{per}}^{n-1}$ which is its tangent space, i.e., (3.7) equals zero. Due to the arbitary choice of ν, we get

$$\frac{1}{\Delta t}\mathcal{L}_{\hat{D}^k}^{-1}(\tilde{\rho}^\star - \tilde{\rho}^k)_i + \log \rho_i^\star - \log\left(1 - \sum_{j=1}^n \rho_j^\star\right) = C_i,$$

with $C_i, i = 1, \ldots n-1$ being constants, from which it follows that for $i = 1, \ldots, n-1$,

$$\frac{\rho_i^\star - \rho_i^k}{\Delta t} = -\,\mathcal{L}_{\hat{D}^k}\left(\log \tilde{\rho}^\star - \log\left(1 - \sum_{j=1}^n \tilde{\rho}_j^\star\right)\right)_i = \sum_{j=1}^{n-1} d_h(\hat{D}_{ij}^k D_h(\log \rho_j^\star - \log \rho_n^\star)).$$

By Lemma 2.4, $\tilde{\rho}^\star$ satisfies the numerical scheme (1.11)-(1.13).

Conversely, assume $\rho^{k+1} > 0$ is a solution of the numerical scheme (1.11)-(1.13), we can reverse the above calculation with $C_i = 0$ to show that (3.20) holds, which together with the fact that the convex optimization problem (3.5) has a unique interior minimizer, implies that ρ^{k+1} is also the minimizer of (3.5), or equivalently of (3.1). ∎

Remark 3.5. The assumption (1.5) is not necessary in the above proof. Suppose $\sum_{j=1}^n \rho_{j0}(x) = m(x) > 0$, the condition is discretized as $\sum_{j=1}^n \rho_{j,\ell}^0 = m_\ell$, $\ell = 1, \ldots, N$. The corresponding condition in the set $\tilde{K}_\delta$ is replaced by $\sum_{i=1}^n \rho_{i,\ell} \le m_\ell - \delta$. The right hand side of (3.7) is again bounded by Lemma 3.2 and the second term becomes sufficiently negative when $\rho_{i,\ell}^\star = \delta$ or $m_\ell - \sum_{i=1}^{n-1} \rho_{i,\ell}^\star = \delta$. The proof is divided into similar cases. For example, for the case $\rho_{i_0,\ell_0}^\star = \delta$ and $\sum_{i=1}^{n-1} \rho_{i,\ell_0}^\star \ge m_{\ell_0}/2$, the terms $\rho_{i_1,\ell_0}^\star \ge m_{\ell_0}/(2(n-1))$ and $\rho_{i_1,\ell_1}^\star \le m_{\ell_1}$ and (3.8) is negative when δ is small.

3.2. Properties of the scheme . The positivity-preserving and energy stability properties of the scheme follow directly from Theorem 3.1.

THEOREM 3.6. *Assume ρ^0 defined in (2.9) is positive, the solution of the numerical scheme* (1.11)-(1.12) *then satisfies*

1. *(Positivity-preserving) $\rho^k > 0$ for any $k \ge 1$,*
2. *(Unconditionally energy stability) the inequality*

$$F_h(\rho^k) + \|\tilde{\rho}^k - \tilde{\rho}^{k-1}\|_{\mathcal{L}_{\hat{D}^k}^{-1}}^2 \le F_h(\rho^{k-1}) \tag{3.21}$$

holds for any $k \ge 1$.

Proof. 1. Starting from ρ_0, we apply Theorem 3.1 recursively to obtain

$$\rho^k \in K_{\delta_k}$$

for some constant δ_k that is chosen for each step by the minimum among (3.13), (3.15) and (3.19). This yields for every k,

$$\rho^k \in \bigcap_{k=1}^{\infty} K_{\delta_k} \subset K_0\backslash\{0\},$$

so that $\rho^k > 0$.

2. Since the solution of the numerical scheme (1.11)-(1.13) is the minimizer of (3.5), we have

$$J(\rho^{k+1}) \leq J(\rho^k),$$

which is (3.21). ∎

4. Multidimensional case. The scheme can be generalized to the multidimensional case and similar proprties can be established. Before we present the multidimensional scheme, we introduce some notations following [27]. Consider two multidimensional grids define by

$$\mathcal{C}^d := \underbrace{\mathcal{C} \times \cdots \times \mathcal{C}}_{d}, \quad \mathcal{E}_{x_s} := \underbrace{\mathcal{C} \times \cdots \times \mathcal{E} \times \cdots \times \mathcal{C}}_{d}, \ s = 1, \ldots, d,$$

and the functions on them

$$\mathcal{C}^d_{\text{per}} := \{f : \mathcal{C}^d \to \mathbb{R}\}, \quad \mathcal{E}^d_{x_s,\text{per}} := \{f : \mathcal{E}^d_{x_s} \to \mathbb{R}\}, \quad \mathcal{E}^d_{\text{per}} := \left\{ f : \bigcup_{s=1}^{d} \mathcal{E}^d_{x_s} \to \mathbb{R} \right\},$$

as well as the vector functions, $(\mathcal{C}^d_{\text{per}})^n := \{f = (f_1, \ldots, f_n) : f_i \in \mathcal{C}^d_{\text{per}}, i = 1, \ldots, n\}$, $(\mathcal{E}^d_{\text{per}})^n := \{f = (f_1, \ldots, f_n) : f_i \in \mathcal{E}^d_{\text{per}}, i = 1, \ldots, n\}$. We also define the space

$$(\mathring{\mathcal{C}}^d_{\text{per}})^n := \left\{ f \in (\mathcal{C}^d_{\text{per}})^n : \sum_{\ell \in \{1,\ldots,N\}^d} f_{i,\ell} = 0, i = 1, \ldots, n \right\}.$$

We use $f_{\ell_1,\ldots,\ell_d}$ to denote the value of a function f at the grid point $(x_1 = \ell_1 h, \ldots, x_d = \ell_d h)$. We introduce the finite difference operators $D_h : \mathcal{C}^d_{\text{per}} \mapsto \mathcal{E}^d_{\text{per}}$ and $d_h : \mathcal{E}^d_{\text{per}} \mapsto \mathcal{C}^d_{\text{per}}$ as

$$D_h f_{\ell_1,\ldots,\ell_s+\frac{1}{2},\ldots,\ell^d} = \frac{f_{\ell^1,\ldots,\ell^s+1,\ldots,\ell^d} - f_{\ell^1,\ldots,\ell^s,\ldots,\ell^d}}{h},$$

and

$$d_h f_{\ell_1,\ldots,\ell_d} := \sum_{s=1}^{d} \frac{f_{\ell^1,\ldots,\ell^s+\frac{1}{2},\ldots,\ell^d} - f_{\ell^1,\ldots,\ell^s-\frac{1}{2},\ldots,\ell^d}}{h}.$$

We also define for $f \in \mathcal{C}^d_{\text{per}}$, $\hat{f}_{\ell^1,\ldots,\ell^s+\frac{1}{2},\ldots,\ell^d} = \frac{f_{\ell^1,\ldots,\ell^s+1,\ldots,\ell^d} + f_{\ell^1,\ldots,\ell^s,\ldots,\ell^d}}{2}$, $s = 1, \ldots, d$, so that $\hat{f} \in \mathcal{E}^d_{\text{per}}$. We define the inner products

$$\langle f, g \rangle := h^d \sum_{i=1}^{n} \sum_{\ell \in \{1,\ldots,N\}^d} f_{i,\ell} g_{i,\ell}, \quad \forall f, g \in (\mathcal{C}^d_{\text{per}})^n,$$

$$[f, g] := h^d \sum_{i=1}^{n} \sum_{\ell_1,\ldots,\ell_n=1}^{N} f_{i,\ell_1,\ldots,\ell_s+\frac{1}{2},\ldots,\ell_d} g_{i,\ell_1,\ldots,\ell_s+\frac{1}{2},\ldots,\ell_d}, \quad \forall f, g \in (\mathcal{E}^d_{\text{per}})^n.$$

The following summation-by-parts formula holds for any $f \in (\mathcal{C}^d_{\text{per}})^n$ and $\phi \in (\mathcal{E}^d_{\text{per}})^n$,

$$\langle f, d_h \phi \rangle = -[D_h f, \phi].$$

Next we define a norm on $(\mathring{\mathcal{C}}^d_{\text{per}})^{n-1}$. Suppose Φ is a $(n-1) \times (n-1)$ symmetric positive definite matrix, with $\Phi_{ij} \in \mathcal{E}^d_{\text{per}}$. We introduce the following operator

$$\mathcal{L}_\Phi f = -d_h(\Phi D_h f) = -\sum_{j=1}^n d_h(\Phi_{ij} D_h f_j),$$

where the multiplication $\Phi_{ij} D_h f_j$ is taken elementwise on the grid points. For any $g \in (\mathring{\mathcal{C}}^d_{\text{per}})^{n-1}$, let f be determined by $g = \mathcal{L}_\Phi f$, we define the following norm

$$\|g\|^2_{\mathcal{L}_\Phi^{-1}} := [D_h f, \Phi D_h f]. \tag{4.1}$$

With the above notations, the numerical scheme for the system (1.1)-(1.2) is

$$\frac{\rho_i^{k+1} - \rho_i^k}{\Delta t} + d_h(\hat{\rho}_i^k v_i^{k+1}) = 0, \tag{4.2}$$

$$D_h \log \rho_i^{k+1} - \frac{1}{\sum_{i=1}^n \hat{\rho}_i^k} \sum_{j=1}^n \hat{\rho}_j^k D_h \log \rho_i^{k+1} = -\sum_{j=1}^n b_{ij} \hat{\rho}_j^k (v_i^{k+1} - v_j^{k+1}), \tag{4.3}$$

$$\sum_{i=1}^n \hat{\rho}_i^k v_i^{k+1} = 0, \tag{4.4}$$

subject to initial data

$$\rho^0_{i,\ell} = \rho_{i0}(x_\ell), \quad i = 1, \ldots, n, \quad \ell = \{1, \ldots, N\}^d. \tag{4.5}$$

All properties proved for the one dimensional case carry over the d-dimensional case. The following theorem holds.

THEOREM 4.1. *Suppose $\rho^0 > 0$. The solution of the numerical scheme (4.2)-(4.4) satisfies*

1. *(Conservation of mass.) For $k \geq 1$,*

$$\sum_{i=1}^n \rho^k_{i,\ell} = \sum_{i=1}^n \rho^0_{i,\ell}, \quad \textit{for all } \ell \in \{1, \ldots, d\}^N,$$

and

$$\sum_{\ell \in \{1,\ldots,d\}^N} \rho^k_{i,\ell} = \sum_{\ell \in \{1,\ldots,d\}^N} \rho^0_{i,\ell}, \quad \textit{for all } i = 1, \ldots, n.$$

2. *(Positivity-preserving.) For $k \geq 1$, $\rho^k > 0$.*
3. *(Unconditional energy stability.) For $k \geq 1$, the following inequality holds:*

$$F_h(\rho^k) + \|\tilde{\rho}^k - \tilde{\rho}^{k-1}\|^2_{\mathcal{L}^{-1}_{\hat{D}^k}} \leq F_h(\rho^{k-1}),$$

where $F_h(\rho) := \langle \sum_{i=1}^n \rho_i \log \rho_i \rangle$.

The proof of the above theorem is based on the following.

THEOREM 4.2. *Assume $b_{ij} > 0$ and $b_{ij} = b_{ji}$ for $i \neq j$ and $i, j = 1, \dots, n$. Assume $\rho^k \in (\mathcal{C}^d_{\mathrm{per}})^n$ be positive. Then there exists a constant $\delta_0 > 0$, such that $\rho^{k+1} > 0$ is a solution of the numerical scheme (4.2)-(4.4) if and only if it is a minimizer of the optimization problem:*

$$\rho^{k+1} = \underset{(\rho,w)\in K_\delta}{\arg\min} \left\{ J = \frac{1}{4\Delta t}\left[\sum_{i,j=1}^{n} b_{ij}\hat{\rho}_i^k\hat{\rho}_j^k(w_i - w_j)^2\right] + F_h(\rho)\right\}, \tag{4.6}$$

where

$$\begin{aligned} K_\delta = \Big\{(\rho, w):\ & \rho \in (\mathcal{C}^d_{\mathrm{per}})^n,\ w \in (\mathcal{E}^d_{\mathrm{per}})^n;\ \rho_{i,\ell} \geq \delta,\quad \rho_{i,\ell} - \rho^k_{i,\ell} + d_h(\hat{\rho}^k_i w_i)_\ell = 0, \\ & \sum_{i=1}^n \hat{\rho}^k_{i,\ell_1,\dots,\ell_s+\frac{1}{2},\dots,\ell_d} w_{i,\ell_1,\dots,\ell_s+\frac{1}{2},\dots,\ell_d} = 0 \text{ and } \sum_{i=1}^n \rho_{i,\ell} = 1, \\ & \forall i = 1,\dots,n,\ \forall \ell = (\ell_1,\dots,\ell_d) \in \{1,\dots,N\}^d, s = 1,\dots,d\Big\}, \end{aligned}$$

for any $0 < \delta \leq \delta_0$.

The proof of these multi-dimensional results is similar and provided in the Supplementary material.

5. Numerical Examples. We numerically validate our theoretical findings using numerical examples in both one and two dimensions.

5.1. One dimension. We perform the simulation on the Duncan and Toor experiment [3, 8]. We extend the domain from $[0, 1]$ to $[0, 2]$ by reflection to make the solution symmetric on $\mathbb{T} = [0, 2]$ and the initial conditions are taken to be

$$\begin{aligned} \rho_{10}(x) &= \begin{cases} 0.8, & \text{for } 0 \leq x < 0.25 \\ 1.6(0.75 - x), & \text{for } 0.25 \leq x < 0.75, \\ 0, & \text{for } 0.75 \leq x \leq 1.25, \\ 1.6(x - 1.25), & \text{for } 1.25 < x \leq 1.75, \\ 0.8, & \text{for } 1.75 < x \leq 2. \end{cases} \\ \rho_{20}(x) &= 0.2, \\ \rho_{30}(x) &= 1 - \rho_{10}(x) - \rho_{20}(x). \end{aligned}$$

The parameters $(b_{ij})_{n\times n}$ are $b_{12} = b_{13} = 1/0.833, b_{23} = 1/0.168$.

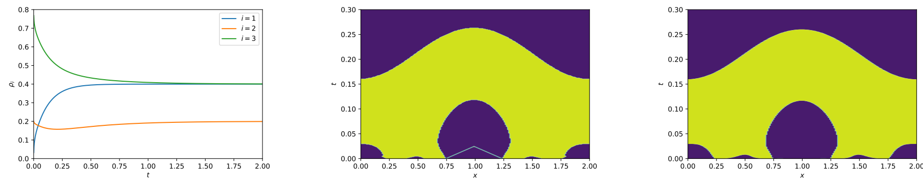
Since $\rho_{10}(x) = 0$ on the subinterval $[0, 75, 1.25]$ and $\rho_{30}(x) = 0$ on $[0, 0.25) \cup (1.75, 2]$, we reinitialize the data following the procedure described in section 2.2, which we outline as below:

For $f_i \geq 0$, but $f_\ell = 0$ for some ℓ, we find a neighboring index set S_ℓ such that the local average

$$\bar{f}_\ell = \frac{1}{|S_\ell|}\sum_{j\in S_\ell} f_j > \eta,$$

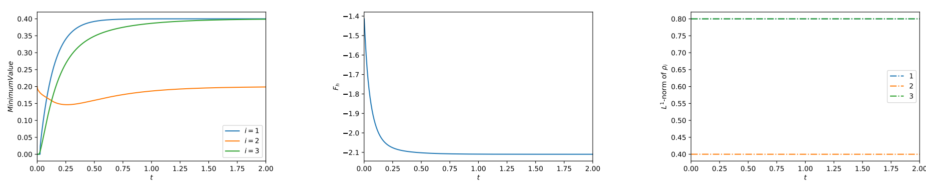
with η being a small number less than $O(h^r), r > 2$. Here $|S_\ell|$ is the number of indices for which $f_j > 0$. We use $\bar{f}_\ell$ as a reference to define the scaling limiter

$$\tilde{f}_j = \theta f_j + (1 - \theta)\bar{f}_\ell, \text{ for all } j \in S_\ell,$$



(a) (b) (c)

FIG. 5.1. *Simulations Results at* $x = 0.72$ *(a) and the uphill diffusion region* $\rho_2 v_2 D_h \rho_2 \leq 0$*(calculated with* $h = 0.001$ *in (b) and* $h = 0.0001$ *in (c))*



(a) (b) (c)

FIG. 5.2. *Minimum value (a), discrete energy (b) and mass (c)*

where $\theta = (\bar{f}_\ell - \eta)/\bar{f}_\ell$. Such a limiter is positive and does not destroy the numerical accuracy [21, 22].

When $\rho_{i,\ell} = 0$ is modified by the above method, we also need to make sure the total density $\sum_{i=1}^{n} \rho_{i,\ell} = 1$ is still preserved. Specifically, for all $j \in S_\ell$, we set $\tilde{\rho}_{s,j} = \rho_{s,j} - (\tilde{\rho}_{i,j} - \rho_{i,j})$ for some index s satisfying $\rho_{s,j} > \eta$. Here we take $\eta = 10^{-15}$ with mesh size $h \geq 10^{-6}$.

We first take the mesh size to be $h = 0.01$ and time step $\Delta t = 0.001$ and compute till $t = 2$. The solution at $x = 0.72$ and the uphill diffusion zone defined by $\rho_2 v_2 D_h \rho_2 \leq 0$ are plotted in Figure 5.1. The solution approximately reaches equilibrium at $t = 2$ and the uphill diffusion zone is almost the same compared to the result in [3, 10]. Notice that here for any time step and mesh size, the scheme is stable. However, for the scheme in [3, 9], Δt and Δx must be carefully set to make the scheme stable. For example, $\Delta t \leq b_{23} h^2/2$ was needed in [3] to make the explicit scheme stable.

To verify the properties of the scheme, we plot the minimum value of ρ over time, the discrete energy function $F_h(\rho)$ and the total mass in Figure 5.2. It can be seen from the figures that the numerical solutions are positive, energy dissipative and conservative.

In order to compute the convergence order, we take $\Delta t = 0.00001$ and the mesh sizes to be 32, 64, 128, 256, 512 and 1024. This small time step is taken so that the numerical error is dominated by the spatial discretization. We compare solutions at $t = 0.01$. The last solution with 1024 meshes is taken as the reference solution. The errors are plotted in Figure 5.3. The figure shows that the scheme is approximately of second order in space.

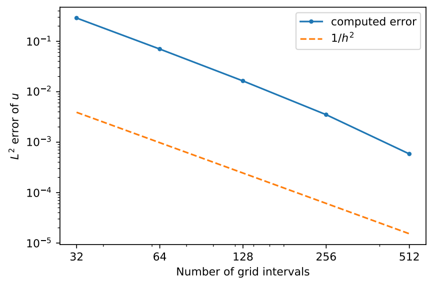
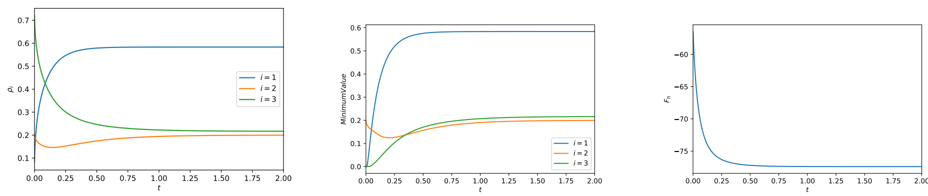


FIG. 5.3. *Numerical errors*



(a) (b) (c)

FIG. 5.4. *The numerical solution at $x = y = 0.7$ (a) and the minimum value (b), discrete energy (c)*

5.2. Two dimensions. We take the same (b_{ij}) as in the one dimensional example and the initial data on $\mathbb{T}^2 = [0,2] \times [0,2]$ to be

$$\rho_{10}(x,y) = \begin{cases} 0.8, & \text{for } x \le 0.25 \text{ or } x \ge 1.75 \text{ and } x \le 0.25 \text{ or } y \ge 1.75, \\ 0, & \text{for } 0.75 \le x \le 1.25 \text{ and } 0.75 \le y \le 1.25, \\ 1.6(0.75 - x), & \text{for } 0.25 \le x < 0.75 \text{ and } x \le y < 2 - x, \\ 1.6(x - 1.25), & \text{for } 1.25 < x < 1.75 \text{ and } 2 - x < y \le x, \\ 1.6(0.75 - y), & \text{for } 0.25 \le y < 0.75 \text{ and } y < x \le 2 - y, \\ 1.6(y - 1.25), & \text{for } 1.25 < y < 1.75 \text{ and } 2 - y \le x < y, \end{cases}$$

$$\rho_{20}(x,y) = 0.2,$$

$$\rho_{30}(x,y) = 1 - \rho_{10}(x,y) - \rho_{20}(x,y).$$

The mesh size is taken to be $h = 0.05$ and time step $\Delta t = 0.001$. We calculate for 500 time steps. The energy and minimum values are shown in Figure 5.4. We can see that the minimum values are all positive and the energy is decaying.

Acknowledgments. The fourth author would like to thank the support from KAUST, where he worked at when this work was done.

REFERENCES

[1] J.-D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the monge-kantorovich mass transfer problem*, Numer. Math., 84 (2000), pp. 375–393.

[2] D. Bothe, *On the maxwell-stefan approach to multicomponent diffusion*, in Parabolic problems, Springer, 2011, pp. 81–93.

[3] L. Boudin, B. Grec, and F. Salvarani, *A mathematical and numerical analysis of the maxwell-stefan diffusion equations*, Discrete Contin. Dyn. Syst. Ser. B, 17 (2012), pp. 1427–1440.

[4] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.
[5] C. Cancès and B. Gaudeul, *A convergent entropy diminishing finite volume scheme for a cross-diffusion system*, SIAM J. Numer. Anal., 58 (2020), pp. 2684–2710.
[6] W. Chen, C. Wang, X. Wang, and S. M. Wise, *Positivity-preserving, energy stable numerical schemes for the cahn-hilliard equation with logarithmic potential*, J. Comput. Phys. X, 3 (2019), p. 100031.
[7] L. Dong, C. Wang, H. Zhang, and Z. Zhang, *A positivity-preserving, energy stable and convergent numerical scheme for the cahn–hilliard equation with a flory–huggins–degennes energy*, Commun. Math. Sci., 17 (2019), pp. 921–939.
[8] J. B. Duncan and H. Toor, *An experimental study of three component gas diffusion*, AIChE Journal, 8 (1962), pp. 38–41.
[9] K. Ehrhardt, K. Klusáček, and P. Schneider, *Finite-difference scheme for solving dynamic multicomponent diffusion problems*, Comput. Chem. Eng., 12 (1988), pp. 1151–1155.
[10] J. Geiser, *Numerical methods of the Maxwell-Stefan diffusion equations and applications in plasma and particle transport*, arXiv preprint arXiv:1501.05792, (2015).
[11] V. Giovangigli and M. Massot, *The local cauchy problem for multicomponent reactive flows in full vibrational non-equilibrium*, Math. Methods Appl. Sci., 21 (1998), pp. 1415–1439.
[12] Y. Gong, J. Zhao, and Q. Wang, *Arbitrarily high-order unconditionally energy stable schemes for thermodynamically consistent gradient flow models*, SIAM J. Sci. Comput., 42 (2020), pp. B135–B156.
[13] X. Huo, A. Jüngel, and A. E. Tzavaras, *High-friction limits of euler flows for multicomponent systems*, Nonlinearity, 32 (2019), p. 2875.
[14] X. Huo and H. Liu, *A positivity-preserving and energy stable scheme for a quantum diffusion equation*, Numer. Methods Partial Differential Equations, (2021), p. in publication.
[15] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the fokker–planck equation*, SIAM J. Math. Anal., 29 (1998), pp. 1–17.
[16] A. Jüngel, *Entropy methods for diffusive partial differential equations*, Springer, 2016.
[17] A. Jüngel and O. Leingang, *Convergence of an implicit euler galerkin scheme for poisson–maxwell–stefan systems*, Adv. Comput. Math., 45 (2019), pp. 1469–1498.
[18] A. Jungel and I. V. Stelzer, *Existence analysis of maxwell–stefan systems for multicomponent mixtures*, SIAM J. Math. Anal., 45 (2013), pp. 2421–2440.
[19] R. Krishna and J. Wesselingh, *The maxwell-stefan approach to mass transfer*, Chem. Eng. Sci., 52 (1997), pp. 861–911.
[20] W. Li, J. Lu, and L. Wang, *Fisher information regularization schemes for wasserstein gradient flows*, J. Comput. Phys., 416 (2020), p. 109449.
[21] H. Liu and W. Maimaitiyiming, *Positive and free energy satisfying schemes for diffusion with interaction potentials*, J. Comput. Phys., 419 (2020), p. 109483.
[22] H. Liu and W. Maimaitiyiming, *Efficient, positive, and energy stable schemes for multi-d poisson–nernst–planck systems*, J. Sci. Comput., 87 (2021), pp. 1–36.
[23] F. Otto, *The geometry of dissipative evolution equations: the porous medium equation*, Comm. Partial Differential Equations, 26 (2001), pp. 101–174.
[24] F. Otto and A. E. Tzavaras, *Continuity of velocity gradients in suspensions of rod–like molecules*, Comm. Math. Phys., 277 (2008), pp. 729–758.
[25] F. Otto and M. Westdickenberg, *Eulerian calculus for the contraction in the wasserstein distance*, SIAM J. Math. Anal., 37 (2006), pp. 1227–1255.
[26] J. Shen, J. Xu, and J. Yang, *A new class of efficient and robust energy stable schemes for gradient flows*, SIAM Rev., 61 (2019), pp. 474–506.
[27] S. M. Wise, C. Wang, and J. S. Lowengrub, *An energy-stable and convergent finite-difference scheme for the phase field crystal equation*, SIAM J. Numer. Anal., 47 (2009), pp. 2269–2288.
[28] Z. Yang, W.-A. Yong, and Y. Zhu, *A rigorous derivation of multicomponent diffusion laws*, arXiv preprint arXiv:1502.03516, (2015).