Positive and energy stable schemes for Poisson-Nernst-Planck equations and related models

by

Wumaier Maimaitiyiming

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Applied Mathematics

Program of Study Committee:
Hailiang Liu, Major Professor
Jia Liu
Songting Luo
James Rossmanith
Paul Sacks

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University
Ames, Iowa
2020

Copyright © Wumaier Maimaitiyiming, 2020. All rights reserved.
DEDICATION

To my wife, Gulzar Osman.

نايالدم گولزار نوسمانغا بپخشلايمه.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>LIST OF TABLES</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>ix</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER 1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Poisson-Nernst-Planck equations</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 Multi-Dimensional PNP equations</td>
<td>2</td>
</tr>
<tr>
<td>1.1.2 Solution properties</td>
<td>3</td>
</tr>
<tr>
<td>1.1.3 Reduced PNP equations</td>
<td>3</td>
</tr>
<tr>
<td>1.1.4 Numerical challenges and existing schemes</td>
<td>6</td>
</tr>
<tr>
<td>1.2 Fokker-Planck equations with interaction potentials</td>
<td>7</td>
</tr>
<tr>
<td>1.2.1 Solution properties</td>
<td>8</td>
</tr>
<tr>
<td>1.2.2 Existing numerical schemes</td>
<td>8</td>
</tr>
<tr>
<td>1.3 Positive and energy stable schemes</td>
<td>9</td>
</tr>
<tr>
<td>1.3.1 Reformulation and spatial discretization</td>
<td>10</td>
</tr>
<tr>
<td>1.3.2 Fully discrete schemes</td>
<td>10</td>
</tr>
<tr>
<td>1.3.3 Local limiter</td>
<td>11</td>
</tr>
<tr>
<td>1.3.4 Positive and energy stable schemes for the PNP systems</td>
<td>12</td>
</tr>
<tr>
<td>1.4 Thesis organization</td>
<td>13</td>
</tr>
<tr>
<td>References</td>
<td>14</td>
</tr>
<tr>
<td>CHAPTER 2. UNCONDITIONAL POSITIVITY-PRESERVING AND ENERGY STABLE SCHEMES FOR A REDUCED POISSON-NERNST-PLANCK SYSTEM</td>
<td>22</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>22</td>
</tr>
<tr>
<td>2.1.1 Mathematical models</td>
<td>23</td>
</tr>
<tr>
<td>2.1.2 Boundary conditions and model properties</td>
<td>24</td>
</tr>
<tr>
<td>2.1.3 Related works</td>
<td>25</td>
</tr>
<tr>
<td>2.1.4 Contributions and organization of the paper</td>
<td>26</td>
</tr>
<tr>
<td>2.2 Numerical methods for a model equation</td>
<td>27</td>
</tr>
<tr>
<td>2.2.1 Scheme formulation</td>
<td>28</td>
</tr>
<tr>
<td>2.2.2 Positivity</td>
<td>30</td>
</tr>
<tr>
<td>2.3 Positive schemes for the reduced PNP-system</td>
<td>31</td>
</tr>
<tr>
<td>2.3.1 Scheme properties</td>
<td>32</td>
</tr>
<tr>
<td>2.4 Numerical tests</td>
<td>37</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.2.1 Semi-discrete scheme</td>
<td>115</td>
</tr>
<tr>
<td>5.2.2 Scheme properties</td>
<td>116</td>
</tr>
<tr>
<td>5.3 Fully discrete scheme</td>
<td>118</td>
</tr>
<tr>
<td>5.3.1 Scheme formulation and algorithm</td>
<td>119</td>
</tr>
<tr>
<td>5.3.2 Scheme properties</td>
<td>119</td>
</tr>
<tr>
<td>5.3.3 Discussion on error estimates</td>
<td>123</td>
</tr>
<tr>
<td>5.4 Numerical Method: two dimensional Case</td>
<td>124</td>
</tr>
<tr>
<td>5.4.1 Semi-discrete scheme</td>
<td>124</td>
</tr>
<tr>
<td>5.4.2 Fully discrete scheme</td>
<td>126</td>
</tr>
<tr>
<td>5.5 Second order in-time discretization</td>
<td>129</td>
</tr>
<tr>
<td>5.5.1 Second order scheme for 1D problem</td>
<td>130</td>
</tr>
<tr>
<td>5.5.2 Local limiter and algorithm</td>
<td>131</td>
</tr>
<tr>
<td>5.5.3 Second order scheme for 2D problem</td>
<td>135</td>
</tr>
<tr>
<td>5.5.4 Local limiter and algorithm</td>
<td>136</td>
</tr>
<tr>
<td>5.6 Numerical Examples</td>
<td>137</td>
</tr>
<tr>
<td>5.6.1 One-dimensional tests</td>
<td>137</td>
</tr>
<tr>
<td>5.6.2 Two-dimensional tests</td>
<td>143</td>
</tr>
<tr>
<td>5.7 Concluding remarks</td>
<td>145</td>
</tr>
<tr>
<td>5.8 Appendix: A refined time step bound for energy dissipation</td>
<td>148</td>
</tr>
<tr>
<td>References</td>
<td>150</td>
</tr>
</tbody>
</table>

**CHAPTER 6. SUMMARY AND DISCUSSION**                                  | 156  |
| 6.1 Conclusions                                                        | 156  |
| 6.2 Future Work                                                        | 157  |
| References                                                             | 157  |
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Accuracy for Example 2.4.1 with first order boundary approximations (2.3.6)</td>
<td>38</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Accuracy for Example 2.4.1 with second order boundary approximations (2.2.11)</td>
<td>39</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Times needed for reaching each steady state in Example 2.4.2 when $Q_0 = 0.2, r_f = 20$ with different channel geometry</td>
<td>41</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>Times needed for reaching each steady state on Example 2.4.3 when $\rho = 0, r_f = 20$ with different channel geometry</td>
<td>42</td>
</tr>
<tr>
<td>Table 2.5</td>
<td>Times needed for reaching steady state on Example 2.4.4 when $C = 1$ with different initial data</td>
<td>44</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Accuracy with the limiter</td>
<td>60</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Accuracy without the limiter</td>
<td>60</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Times needed for reaching steady states on Example 3.4.2</td>
<td>61</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Scheme (4.3.7) with $\tau = h$</td>
<td>92</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Scheme (4.3.7) with $\tau = h^2$</td>
<td>92</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Scheme (4.4.2) with $\tau = h$</td>
<td>92</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Accuracy of scheme (5.3.1) with $\tau = 0.1h$ and $\tau = h^2$</td>
<td>138</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Accuracy of scheme (5.5.1) with $\tau = h$</td>
<td>138</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>Efficiency of schemes (5.3.1) and (5.5.1) (CPU times in seconds)</td>
<td>143</td>
</tr>
<tr>
<td>Table 5.4</td>
<td>Accuracy of scheme (5.4.4) and (5.5.6)</td>
<td>143</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>Schematic view of K channel, membrane, and interior and exterior baths... 4</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>Computational region for channel, membrane, and baths with boundary condition types. ........................................... 4</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Diagram of 1D computational region for the channel and bath funnels ... 40</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Effects of channel geometry on steady state densities and potential with $Q_0 = 0.2$: (a)-(c) densities at $t = 0.2$, for $l_c = r_c = \frac{1}{3}$, $\frac{1}{5}$ and $\frac{1}{17}$, (d)-(f) potential profiles at $t = 0.2$ for $l_c = r_c = \frac{1}{3}, \frac{1}{5}$ and $\frac{1}{17}$. .......................... 42</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Effects of permanent charge on steady state densities and potential with $Q_0 = 0.2$: (a)-(c) are computed densities at $t = 0.2$, for $Q_0 = 0.05$, 0.1, 0.15, (d)-(f) are potential profiles at $t = 0.2$ for $Q_0 = 0.05$, 0.1, 0.15. .......................... 43</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>I-V relation: (a) voltage for $V = 0.5, 1, 3, 5$, with $l_c = r_c = \frac{1}{5}$ and $Q_0 = 0.1$ at time $t = 0.2$, (b) current voltage relation. .......................... 44</td>
</tr>
<tr>
<td>Figure 2.5</td>
<td>Effects of permanent charge and initial data on steady state densities: (a) is initial data profile (4.8), (b)-(c) are density profile at $t = 3$ for $C = 1$ and $C = 2$ respectively, (d) is random initial data profile, (e)-(f) are density profile for $C = 1$ and $C = 2$. .......................... 46</td>
</tr>
<tr>
<td>Figure 2.6</td>
<td>Energy dissipation and mass conservation: (a) density profiles at $t = 15$ for $C = 1$, (b) energy dissipation and mass conservation. .......................... 47</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Example 3.4.2: Initial data (left) and steady state solutions obtained with limiter (middle) and without limiter (right) .......................... 61</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Example 3.4.2: minimum of densities $c^{n}_1$ (left), $c^{n}_2$ (middle) and $c^{n}_3$ (right) obtained with limiter (dots) and without limiter (solid) .......................... 62</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Example 4.5.2: $\rho_1, \rho_2, \phi$ computed by scheme (4.3.7) .......................... 94</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Example 4.5.2: $\rho_1, \rho_2, \phi$ computed by scheme (4.4.2) (with limiter) .......................... 95</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>Example 4.5.3: $\rho_1, \rho_2$ computed by scheme (4.3.7) .......................... 96</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>Example 4.5.3: Mass conservation and energy dissipation .......................... 96</td>
</tr>
<tr>
<td>Figure 5.1</td>
<td>First order scheme for Example 5.6.2. .......................... 139</td>
</tr>
<tr>
<td>Figure 5.2</td>
<td>Second order scheme (with and without limiter) for Example 5.6.2. .......................... 140</td>
</tr>
<tr>
<td>Figure 5.3</td>
<td>Second order scheme energy and total mass (without limiter for $\tau = 0.01$, with limiter for $\tau = 0.1$) for Example 5.6.2. .......................... 140</td>
</tr>
<tr>
<td>Figure 5.4</td>
<td>Solution evolution and energy dissipation for Example 5.6.3 with $\alpha = 3$.</td>
</tr>
<tr>
<td>Figure 5.5</td>
<td>Solution evolution and energy dissipation for Example 5.6.3 with $\alpha = 5$.</td>
</tr>
<tr>
<td>Figure 5.6</td>
<td>Energy dissipation for Example 5.6.3 with $\tau = 0.01, 0.1, 1$.</td>
</tr>
<tr>
<td>Figure 5.7</td>
<td>Solution evolution for Example 5.6.5 (sub-critical).</td>
</tr>
<tr>
<td>Figure 5.8</td>
<td>Solution evolution for Example 5.6.5 (super-critical).</td>
</tr>
<tr>
<td>Figure 5.9</td>
<td>Values of $h | D_h W^{1/2} |^2$ for the Lipschitz kernel $W(x) = e^{-</td>
</tr>
</tbody>
</table>
I would like to take this opportunity to thank Dr. Hailiang Liu for being my advisor. Dr. Liu showed me the right attitude for doing research by his passion. He encouraged me to push myself and in turn strengthened my resilience. He has always made himself available to me, his insights and fruitful discussions often inspired me and renewed my hopes. Without his guidance, support, and encouragement I could not finish my Ph.D. program. Thank you, Dr. Liu.

My sincere thanks also go to my committee members Dr. Jia Liu, Dr. Songting Luo, Dr. James Rossmanith, and Dr. Paul Sacks for their support and oversight in this work. A special thanks to Dr. Steve Butler for his valuable suggestions and support in my teaching.

I also, once more, want to thank Dr. James Rossmanith and Dr. Paul Sacks. They have been great teachers and always solved the issues I faced throughout my graduate career.

Finally, I would like to thank my family in Xinjiang (China) and my family here in the United States for their sacrifices, support, and love.
ABSTRACT

In this thesis, we design, analyze, and numerically validate positive and energy-dissipating schemes for solving Poisson-Nernst-Planck (PNP) equations and Fokker-Planck (FP) equations with interaction potentials. These equations play an important role in modeling the dynamics of charged particles in semiconductors and biological ion channels, as well as in other applications. These model equations are nonlinear/nonlocal gradient flows in density space, and their explicit solutions are rarely available; however, solutions to such problems feature intrinsic properties such as (i) solution positivity, (ii) mass conservation, and (iii) energy dissipation. These physically relevant properties are highly desirable to be preserved at the discrete level with the least time-step restrictions.

We first construct our schemes for a reduced PNP model, then extend to multi-dimensional PNP equations and a class of FP equations with interaction potentials. The common strategies in the construction of the baseline schemes include two ingredients: (i) reformulation of each underlying model so that the resulting system is more suitable for constructing positive schemes, and (ii) integration of semi-implicit time discretization and central spatial discretization. For each model equation, we show that the semi-discrete schemes (continuous in time) preserve all three solution properties (positivity, mass conservation, and energy dissipation). The fully discrete first order schemes preserve solution positivity and mass conservation for arbitrary time steps. Moreover, there exists a discrete energy function which dissipates along time marching with an $O(1)$ bound on time steps.

We show that the second order (in both time and space) schemes preserve solution positivity for suitably small time steps; for larger time steps, we apply a local limiter to restore the solution positivity. We prove that such limiter preserves local mass and does not destroy the approximation
accuracy. In addition, the limiter provides a reliable way of restoring solution positivity for other high order conservative finite difference or finite volume schemes.

Both the first and second order schemes are linear and can be efficiently implemented without resorting to any iteration method. The second order schemes are only slight modifications of the first order schemes. Computational costs of a single time step for first and second order schemes are similar, hence our second-order in time schemes are efficient than the first-order in time schemes, given a larger time step could be utilized (to save computational cost). We conduct extensive numerical tests that support our theoretical results and illustrate the accuracy, efficiency, and capacity to preserve the solution properties of our schemes.
CHAPTER 1. INTRODUCTION

Drift-diffusion equations play an important role in modeling the movement of charged particles. In this thesis we shall focus on two classes of partial differential equations (PDE): the celebrated Poisson-Nernst-Planck (PNP) equations that arise in semiconductor physics and biology; and the Fokker-Plank (FP) equation with interaction potentials which have many applications in biology and physical chemistry.

Theoretical and numerical studies of these PDEs have a long history. For the theoretical analysis of drift-diffusion equations, see [34, 61]. For these nonlinear/non-local PDEs, in general, the exact solutions cannot be obtained. Therefore, computer simulations play an essential role in understanding the dynamics of these equations. Computational studies already started in the 1960s [15, 17], and many discretization methods have been explored for these PDE systems. The proposed algorithms include the finite difference methods [5, 6, 19, 23, 24, 31, 32, 33, 49, 58, 60], the finite volume methods [11, 42, 43, 44, 45, 46], and the finite element methods [28, 47, 48, 50, 51, 52, 55, 67], among others.

As gradient flows, the PNP equations and FP equations with interaction potentials can take very long time evolution to reach their steady states; hence, it is important to construct stable, accurate, and efficient numerical methods to solve these systems. Indeed, numerical schemes that can satisfy certain physical properties (e.g., maximum principle, mass conservation, and energy dissipation, etc.) of a model equation are proven to be more stable and produce physically relevant numerical solutions. In this thesis, we focus on designing efficient and structure-preserving schemes for PNP equations and FP equations with interaction potentials.
1.1 Poisson-Nernst-Planck equations

PNP equations is a mean-field approximation of diffusive molecules or ions, and consist of
Nernst–Planck equations that describe the drift and diffusion of ion species, and the Poisson equation
that describes the electrostatic interaction. This system of equations for multiple species has
been used in modeling biological membrane channels [16, 22, 68], electrochemical systems [2, 56],
and semiconductor devices [53, 62].

1.1.1 Multi-Dimensional PNP equations

The PNP equations have several variants, including the following form which we will consider,
\[
\begin{align*}
\partial_t \rho_i + \nabla \cdot J_i &= 0, \quad x \in \Omega \subset \mathbb{R}^d, \quad t > 0, \\
- J_i &= D_i(x) \left[ \nabla \rho_i + \frac{1}{k_B T} \rho_i (q_i \nabla \phi + \nabla \mu_i) \right], \\
- \nabla \cdot (\epsilon(x) \nabla \phi) &= 4\pi \left( f(x) + \sum_{i=1}^m q_i \rho_i \right).
\end{align*}
\]
Here \( m \) is the number of species, \( \rho_i = \rho_i(x, t) \) is the charge carrier density for the \( i \)-th species,
and \( \phi = \phi(x, t) \) the electrostatic potential. The charge carrier flux is \( J_i \), in which \( D_i(x) \) is the
diffusion coefficient, \( k_B \) is the Boltzmann constant, and \( T \) is the absolute temperature. The coupling
parameter is \( q_i = z_i e \), where \( z_i \) is the valence (with sign), \( e \) is the unit charge, and \( \mu_i(x) \) is the
chemical potential. In the Poisson equation, \( \epsilon(x) \) is the permittivity, \( f(x) \) is the permanent (fixed)
charge density of the system.

The equations are valid in a bounded domain \( \Omega \) with boundary \( \partial \Omega \) and for time \( t \geq 0 \). We
consider the initial boundary value problem for (1.1.1) subject to the initial data \( \rho_i(x, 0) = \rho_i^{in}(x) \geq 0 \)
(\( i = 1, \cdots, m \)) and Dirichlet and/or Neumann type [7] boundary conditions
\[
\begin{align*}
\rho_i(x, t) &= \rho_i^b(x, t) \geq 0, \quad \phi(x, t) = \phi^b(x, t), \quad x \in \partial \Omega_D, \\
J_i \cdot \mathbf{n} &= 0, \quad \epsilon(x) \nabla \phi \cdot \mathbf{n} = 0, \quad x \in \partial \Omega_N,
\end{align*}
\]
here \( \partial \Omega = \partial \Omega_D \cup \partial \Omega_N \), \( \mathbf{n} \) is the outward unit normal vector on the boundary \( \partial \Omega_N \). The boundary
conditions for the electrostatic potential are not unique and greatly depend on the problem under
investigation. For example, one may use non-homogeneous Neumann boundary condition \( (\nabla \phi \cdot n = \sigma, x \in \partial \Omega \) is used in [51]) or Robin boundary conditions [23, 33]. The existence and uniqueness of the solution for the nonlinear PNP boundary value problems have been studied in [35, 41, 57] for the 1D case and in [7, 34] for multi-dimensions.

### 1.1.2 Solution properties

The boundary value problem (1.1.1)-(1.1.2) possesses three important properties, one is the non-negativity

\[
\rho_i(x, t) \geq 0, \quad x \in \Omega, \ t > 0.
\]  

With zero flux \( J_i \cdot n = 0 \) and \( \epsilon(x) \nabla \phi \cdot n = 0 \) on the whole boundary, the solutions also have mass conservation and energy dissipation properties

\[
\int_{\Omega} \rho_i(x, t) dx = \int_{\Omega} \rho_i^{in}(x) dx, \quad t > 0, \quad i = 1, \cdots, m.
\]  

\[
\frac{dE}{dt} = -\int_{\Omega} \sum_{i=1}^{m} D_i(x) \rho_i |\nabla (\log \rho_i + \frac{q_i}{k_B T} \phi + \frac{1}{k_B T} \mu_i)|^2 dx \leq 0.
\]

Here the total energy \( E \) associated to (1.1.1) is given by

\[
E = \int_{\Omega} \left( \sum_{i=1}^{m} \rho_i (\log \rho_i - 1) + \frac{1}{2k_B T} (f + \sum_{i=1}^{m} q_i \rho_i) \phi + \frac{1}{k_B T} \sum_{i=1}^{m} \rho_i \mu_i \right) dx.
\]

The positivity-preserving property is of special importance since negative values in density would violate the physical meaning of the solution and may destroy the energy dissipation law (1.1.5). The PNP system (1.1.1) is the gradient flow of the free energy (1.1.6) and the energy dissipation relation (1.1.5) indicates that the solution to the boundary value problem evolves in the direction of the steepest descent of the free energy. By some energy estimate with the control of the free energy dissipation, the solution is shown to converge to the thermal equilibrium state as time becomes large if the boundary conditions are in thermal equilibrium [26].

### 1.1.3 Reduced PNP equations

The ion flow in biological ion channels can be modeled by the PNP equations. When the PNP equations (1.1.1) is used to model K channel in a biological membrane plus surrounding KCl
solution baths (see Figure 1.1), the three-dimensional channel geometry (see Figure 1.2) can be well approximated by a one-dimensional system along the axial direction $z$ [27, 63, 64]. The schematic view of the K channel structure is illustrated in Figure 1.1 [27].

![Figure 1.1 Schematic view of K channel, membrane, and interior and exterior baths.](image1)

Ion channel and regions of the bath illustrated in Figure 1.2 can be modeled by the PNP equations (1.1.1), the baths are represented by conical funnels with the opening at 45° angles on
either side of the $z$–axis. Here the boundary conditions in Figure 1.2 are defined by

$$\rho_i(r, \theta, 0, t) = \rho_{i,t} \geq 0, \quad \phi(r, \theta, 0, t) = V_i, \quad \text{(interior bath far-field BC)},$$  

(1.1.7a)

$$\rho_i(r, \theta, 1, t) = \rho_{i,r} \geq 0, \quad \phi(r, \theta, 1, t) = V_r, \quad \text{(exterior bath far-field BC)},$$  

(1.1.7b)

$$\nabla \rho_i \cdot n = 0, \quad \nabla \phi \cdot n = 0, \quad \text{(no-flux BC)}.$$  

(1.1.7c)

Let $A(z)$ be the cross sectional area at $z$ in Figure 1.2, let $\bar{\rho}_i(z, t)$ and $\bar{\phi}(z, t)$ be averages of $\rho_i$ and $\phi$ on the cross section $A(z)$, i.e.,

$$\bar{\rho}_i(z, t) = \frac{1}{A(z)} \int_\theta \int_r r \rho_i(r, \theta, z, t) dr d\theta,$$

$$\bar{\phi}(z, t) = \frac{1}{A(z)} \int_\theta \int_r r \phi(r, \theta, z, t) dr d\theta.$$

To reduce the PNP equations from three dimensions to one dimension, we take cross sectional averages of all equations in (1.1.1) on $A(z)$.

For

$$J_i = (u_i, v_i, w_i), \quad \text{with } J_i \cdot n = 0 \quad \text{no-flux BC},$$

the cross sectional average $\frac{1}{A(z)} \int_\theta \int_r r \nabla \cdot J_i dr d\theta$ reduces to

$$\frac{1}{A(z)} \int_\theta \int_r r \nabla \cdot J_i dr d\theta = \frac{1}{A(z)} \lim_{\Delta z \to 0} \int_V \nabla \cdot J_i dV \Delta z$$

$$= \frac{1}{A(z)} \lim_{\Delta z \to 0} \int_{\partial V} J_i \cdot n dS \Delta z$$

(1.1.8)

$$= \frac{1}{A(z)} \frac{\partial}{\partial z} (A(z) \bar{\rho}_i(z)).$$

Therefore

$$\frac{1}{A(z)} \int_\theta \int_r r \nabla \cdot (D_i(z) \nabla \rho_i) dr d\theta = \frac{1}{A(z)} \frac{\partial}{\partial z} (D_i(z) \partial_z (A(z) \bar{\rho}_i)).$$

(1.1.9)

The nonlinear term $\frac{D_i(z)}{k_B T} \rho_i(q_i \nabla \phi + \nabla \mu_i)$ in $J_i$ needs to be handled carefully. From (1.1.8) and the fact

$$\bar{a}b = \bar{a}b + (a - \bar{a})(b - \bar{b})$$
we have
\[ \frac{1}{A(z)} \int_{\theta} \int_{r} r \nabla \cdot \left( \frac{D_i(z) q_i}{k_B T} \rho_i \nabla \phi \right) dr d\theta = \frac{1}{A(z)} \frac{\partial}{\partial z} \left( \frac{q_i}{k_B T} D_i(z) \rho_i A(z) \phi \right) \]
\[ + \frac{1}{A(z)} \frac{\partial}{\partial z} \left( \frac{q_i}{k_B T} D_i(z) A(z) (\rho_i - \bar{\rho}_i)(\partial_z \phi - \bar{\partial}_z \phi) \right). \]  
(1.1.10)

Inserting (1.1.9) and (1.1.10) into the cross sectional average of the PNP system (1.1.1) and dropping the high-order term \[ \frac{1}{A(z)} \frac{\partial}{\partial z} \left( \frac{q_i}{k_B T} D_i(z) A(z) (\rho_i - \bar{\rho}_i)(\partial_z \phi - \bar{\partial}_z \phi) \right) \]
leads to the following 1D approximation

\[ \partial_t \bar{\rho}_i + \frac{1}{A(z)} \partial_z J_i = 0, \quad \bar{z} \in [0, 1], \quad t > 0, \]  
(1.1.11a)

\[ - J_i = D_i(z) \left[ \partial_z (A(z) \bar{\rho}_i) + \frac{1}{k_B T} \bar{\rho}_i \left( q_i \partial_z (A(z) \bar{\phi}) \right) \right], \]  
(1.1.11b)

\[ - \frac{1}{A(z)} \partial_z \left( \epsilon(z) \partial_z (A(z) \bar{\phi}) \right) = 4\pi \left( \bar{f}(z) + \sum_{i=1}^{m} q_i \bar{\rho}_i \right). \]  
(1.1.11c)

We can recover the reduced model studied in [27] from the model (1.1.11) under the assumption that the cross-sectional area \( A(z) \) changes slowly in \( z \) (as assumed in [27]). The reduced model (1.1.11) well approximates the 3D ion channel problem and has positivity, mass conservation, and energy dissipation properties. In addition, the reduced model is efficient in numerical simulations and the computed current-voltage curve for the K channel shows excellent agreement with experimental measurements [27].

1.1.4 Numerical challenges and existing schemes

The PNP system is a strongly coupled system of nonlinear equations; also, as a gradient flow, it can take a very long time evolution for the PNP systems to reach steady states. The nonlinear gradient flow nature of the PNP system makes it challenging to construct efficient and stable schemes. It is desirable to maintain (i) conservation of mass, (ii) density positivity, and (iii) free energy dissipation properties at the discrete level, preferably without or with only mild constraints on time steps. The mass conservation requires the scheme to be conservative, the non-negativity property is point-wise and also important for the energy dissipation property. In general, it is rather
challenging to obtain both unconditional positivity and discrete energy decay simultaneously. This is evidenced by several recent efforts [23, 24, 28, 31, 49, 51, 55].

The early approaches lie in simulations of PNP models in various chemical/biological applications [14, 21, 25, 29, 38, 40, 52, 65, 66, 70, 71]. Many of these existing algorithms are introduced to handle specific settings, such as discontinuous coefficients, singular charges, geometric singularities, and nonlinear couplings to accommodate various phenomena exhibited by biological ion channels. For instance, a set of finite difference algorithms were developed [70, 71] for the 3D PNP equations, in which the matched interface and boundary method (MIB) was used to handle the discontinuous property of dielectric constants on solvent and solute domain, and Dirichlet to Neumann mapping technique was applied to treat singularities of fixed charges of the protein channels. A finite element method for the PNP system with a singular charge was developed in [52].

Property-preserving numerical methods for solving the PNP system have gained more attention in recent years [23, 24, 33, 49, 51, 55]. These schemes are either explicit or fully implicit in time; the former require time step restrictions for preserving the desired solution structures while the later preserve desired properties unconditionally, but they had to be solved by some iterative solvers. More recent attempts have focused on semi-implicit schemes based on a formulation of the non-logarithmic Landau type. As a result, all schemes obtained in [19, 32, 42, 44, 45] feature unconditional positivity. In some cases, electric energy alone can be shown to decay (see [51]). Such decay has been verified for the finite difference scheme in [31] and the finite element scheme in [28], both with semi-implicit time discretization.

### 1.2 Fokker-Planck equations with interaction potentials

Consider the nonlinear nonlocal equation

\[
\partial_t \rho = \nabla \cdot (\nabla \rho + \rho \nabla (V(x) + W \ast \rho)), \quad x \in \Omega \subset \mathbb{R}^d, \quad t > 0,
\]

subject to initial data \( \rho(x, 0) = \rho_0(x) \geq 0 \) and no-flux boundary condition. Here \( \Omega \) is a bounded domain in \( \mathbb{R}^d \), \( \rho = \rho(x, t) \) is the unknown density, \( V(x) \) is a confinement potential, and \( W(x) \) is an interaction potential, which is assumed to be symmetric.
Such an equation appears in many applications. If $W$ vanishes, this model includes heat equation ($V(x) = 0$) and the classical linear Fokker–Planck equation ($V(x) \neq 0$, see e.g. [61]). With interaction potentials, the equation can model nematic phase transition of rigid rod-like polymers [20], chemotaxis [59], and aggregation in biology (see [30, 37, 69] and references therein). This equation is also related to the gradient flow for the Wasserstein metric on the space of probability measures [1].

1.2.1 Solution properties

Similar to the PNP equations, the boundary value problem (1.2.1) also has positivity, mass conservation, and free energy dissipation properties, i.e.,

$$\rho_0(x) \geq 0 \implies \rho(x, t) \geq 0, \quad t > 0,$$

(1.2.2)

$$\int_{\Omega} \rho(x, t) dx = \int_{\Omega} \rho_0(x) dx, \quad t > 0,$$

(1.2.3)

$$\frac{dE(\rho)}{dt} = -\int_{\Omega} \rho |\nabla (\log(\rho) + V(x) + W \ast \rho)|^2 dx = -I(\rho) \leq 0,$$

(1.2.4)

where the free energy associated to (1.2.1) is given by

$$E(\rho) = \int_{\Omega} \rho \log(\rho) dx + \int_{\Omega} V(x) \rho dx + \frac{1}{2} \int_{\Omega} \int_{\Omega} W(x - y) \rho(y) \rho(x) dxdy. $$

(1.2.5)

These mathematical features are crucial for the analytical study of (1.2.1), while free-energy dissipation property provides much insight into the problem and particularly important to understand the large time dynamics of solutions (see e.g., [9, 10, 54]).

1.2.2 Existing numerical schemes

There are different strategies available for obtaining structure-preserving numerical schemes for (1.2.1). One way of obtaining a structure-preserving scheme is the Chang-Cooper (CC) method, first introduced by Chang and Cooper to construct positive and steady state preserving schemes for solving linear FP equations [15]. The CC scheme for the FP equation

$$\rho_t(x, t) = \nabla \cdot \mathbf{F}(x, t),$$

with \( \mathbf{F}(x,t) = B(x,t)\rho(x,t) + C(x,t)\nabla \rho(x,t) \) is given by (in 1D)

\[
\frac{d}{dt}\rho_j(t) = \frac{F_{j+1/2}(t) - F_{j-1/2}(t)}{\Delta x},
\]

\[
F_{j+1/2} = B_{j+1/2}((1 - \delta_j)\rho_{j+1} + \delta_j\rho_j) + C_{j+1/2}\left(\frac{\rho_{j+1} - \rho_j}{\Delta x}\right),
\]

(1.2.6)

where the parameters \( \{\delta_j\} \) are chosen so that the scheme has the positivity and steady state preserving properties independent of the mesh size \( \Delta x \) [15]. Such schemes have been generalized for nonlinear FP equations in [5]. Related methods for (1.2.1) is proposed in [58].

Another class of structure preserving methods builds on a variational formulation of (1.2.1), the celebrated Jordan-Kinderlehrer-Otto (JKO) scheme (Jordan et al. [36]) has the form

\[
\rho^{n+1} = \arg\min \left\{ \frac{1}{2\tau}W^2(\rho^n, \rho) + E(\rho) \right\}.
\]

Here, at each time step, the distance of the solution update acts as a regularization to the free energy. The JKO schemes in [3, 12, 39] preserve all three properties without any time step restriction.

The particle/blob methods [4, 8, 13, 18] leverage the structural similarities between (1.2.1) and equations from fluid dynamics and naturally conserve mass and positivity, and they can also be designed to preserve the energy dissipation. Other structure preserving schemes include finite difference [58, 60], finite volume [11, 43, 46], and discontinuous Galerkin methods [47, 48, 50, 67].

### 1.3 Positive and energy stable schemes

In this section, we construct positive and free energy dissipating schemes for solving FP equations with interaction potentials (1.2.1). First, we transform the equation (1.2.1) to rewrite it in the non-logarithmic Landau form, such transformation converts the drift-diffusion operator into a self-adjoint elliptic operator. The reformulated form can be more efficiently solved and suitable for constructing positivity-preserving schemes. Then use a first order semi-implicit time discretization to construct unconditionally positive and energy stable scheme (first order in time and second order in space). Finally, we construct a fully second order scheme by using a predictor-corrector approach, where solution positivity is restored by a local scaling limiter.
1.3.1 Reformulation and spatial discretization

A key step in our scheme construction is the reformulation of the equation (1.2.1) as
\[
\partial_t \rho = \nabla \cdot \left( M \nabla \left( \frac{\rho}{M} \right) \right) =: R[\rho, M],
\]
(1.3.1)
here \( M = e^{-V(x)-W*\rho} \). In the context of Fokker-Planck equations such reformulation with \( M = e^{-V(x)} \) is termed as the non-logarithmic Landau formulation (see, e.g., [6, 46]). The advantage of the formulation (1.3.1) can be seen from both spatial and temporal discretizations [42, 43, 44, 45, 46, 49].

Due to the diffusion in the model equation, it is natural to use central spatial discretization. The symmetric spatial discretization of the one-dimensional version of (1.3.1) is
\[
h_j \frac{d}{dt} \rho_j = h_{j+1/2}^1 M_{j+1/2} \left( \frac{\rho_{j+1}}{M_{j+1}} - \frac{\rho_j}{M_j} \right) - h_{j-1/2}^1 M_{j-1/2} \left( \frac{\rho_j}{M_j} - \frac{\rho_{j-1}}{M_{j-1}} \right) =: R_h[\rho_j, M_j].
\]
(1.3.2)
Here \( \rho_j \) approximates the cell average of \( \rho(x,t) \) on \( j \)-th computational cell \([x_{j-1/2}, x_{j+1/2}]\) of size \( h_j \), we set \( h_{j+1/2} = (h_j + h_{j+1})/2 \). The evaluation of \( M \) at cell interfaces \( \{x_{j+1/2}\} \) and cell centers \( \{x_j\} \) is easily available as \( M_{j+1/2} = g(x_{j+1/2}) \) and \( M_j = g(x_j) \), here
\[
g(x) = \exp \left( -V(x) - \sum_{j=1}^{N} h_j W(x - x_j) \rho_j \right)
\]
is a globally defined function in terms of numerical solutions \( \{\rho_j\}_{j=1}^{N} \).

The semi-discrete scheme (1.3.2) is mass conservative, positive, and energy dissipating (Chapter 5, Theorem 5.2.1).

1.3.2 Fully discrete schemes

We use the semi-implicit time discretization of (1.3.2):
\[
h_j \frac{\rho_j^{n+1} - \rho_j^n}{\tau} = R_h[\rho_j^{n+1}, M_j^n],
\]
(1.3.3)
here \( \rho_j^n \) approximates \( \rho_j(t) \) at time \( t = n\tau \). The feature of such discretization is that it is a linear equation in \( \rho_j^{n+1} \), thus does not require iterative solvers. The coefficient matrix of the resulting linear system is a M-matrix and the right-hand side is a nonnegative vector, thus positivity of
the numerical solutions is ensured without any time step restriction. The scheme (1.3.3) has a fully discrete energy function which dissipates under a mild time step restriction, see Chapter 5, Theorem 5.3.1.

Let us mention some viable options for the time discretization. Solution positivity is readily available if we discretize (1.3.2) as

\[ h_j \frac{\rho_j^{n+1} - \rho_j^{n-k+1}}{k\tau} = R_h[\rho_j^{n+1}, M^*_j], \]

with a consistent choice for \( M^*_j \) and integer \( k \geq 1 \). Different options are introduced in [19, 32, 33] for obtaining their respective positive schemes for PNP equations.

The semi-implicit scheme (1.3.3) is first order accurate in time, one can design higher order in time schemes based on (1.3.1). The following linearized Crank–Nicolson method is a second order time discretization,

\[ h_j \frac{\rho_j^{n+1} - \rho_j^n}{\tau} = R_h[(\rho_j^{n+1} + \rho_j^n)/2, \frac{3}{2} M_j^n - \frac{1}{2} M_j^{n-1}], \]

which can be expressed as a predictor-corrector method:

\[ h_j \frac{\rho_j^S - \rho_j^n}{\tau/2} = R_h[\rho_j^*, \frac{3}{2} M_j^n - \frac{1}{2} M_j^{n-1}], \quad \rho_j^{n+1} = 2\rho_j^* - \rho_j^n. \quad (1.3.4) \]

Here \( \rho_j^* \) is positive for any time step. This scheme preserves solution positivity for suitably small time steps (Chapter 5, Theorem 5.5.1). For the positivity property to be preserved for large time steps, we couple the above scheme with a local limiter.

### 1.3.3 Local limiter

We begin to design a local limiter to restore positivity of \( \{c_j\}_{j=1}^N \) if \( \sum_{j=1}^N c_j > 0 \), but \( c_k < 0 \) for some \( k \). The idea is to find a neighboring index set \( S_k \) such that the local average

\[ \bar{c}_k = \frac{1}{|S_k|} \sum_{j \in S_k} c_j > 0, \]

here \( |S_k| \) denotes the minimum number of indexes for which \( c_j \neq 0 \) and \( \bar{c}_k > 0 \), then use this average as a reference to define the following scaling limiter:

\[ \tilde{c}_j = \theta c_j + (1 - \theta) \bar{c}_k, \quad j \in S_k, \quad (1.3.5) \]
The limiter \((1.3.5)\) has following properties

- \(\tilde{c}_j \geq 0\) for all \(j \in S_k\),
- \(\sum_{j \in S_k} \tilde{c}_j = \sum_{j \in S_k} c_j\), and
- \(|\tilde{c}_j - c_j| \leq |S_k|(- \min_{j \in S_k} c_j)\).

The above limiter when applied to \(\{\rho^n_j\}\) with \(c_j = h_j \rho^n_j\) gives

\[
\tilde{\rho}^n_j = \theta \rho^n_j + (1 - \theta) \frac{\tilde{c}_k}{h_j},
\]

(1.3.6)

here

\[
\theta = \min \left\{ 1, \frac{\tilde{c}_k}{\tilde{c}_k - c_{\text{min}}} \right\}, \quad c_{\text{min}} = \min_{j \in S_k} h_j \rho^n_j, \quad \tilde{c}_k = \frac{1}{|S_k|} \sum_{j \in S_k} h_j \rho^n_j.
\]

The numerical solution obtained by \((1.3.4)\) and \((1.3.6)\) have following properties:

- \(\tilde{\rho}^{n+1}_j \geq 0\) for all \(j \in S_k\),
- \(\sum_{j \in S_k} h_j \tilde{\rho}^{n+1}_j = \sum_{j \in S_k} h_j \rho^n_j\), and
- \(|\tilde{\rho}^{n+1}_j - \rho_j(t_{n+1})| \leq (1 + |S_k|\alpha) \max_{j \in S_k} |\rho^{n+1}_j - \rho_j(t_{n+1})|, \quad j \in S_k\),

here \(\alpha\) is the upper bound of the mesh ratio \(h_i/h_j\). The limiter does not destroy the second order accuracy as long as \(|S_k|\alpha\) is uniformly bounded from above. This is indeed the case for the shape-regular meshes (Chapter 5, Theorem 5.5.2). The present limiter provides a reliable way of restoring solution positivity for other high order finite difference/volume schemes as well. This approach is quite robust, and we expect that, once the theory for the accuracy propagation is further developed, it will lead to theoretical results for a more complicated schemes.

### 1.3.4 Positive and energy stable schemes for the PNP systems

The Nernst-Plank equations in the PNP systems can also be reformulated into a self-adjoint form similar to \((1.3.1)\), i.e., equations \((1.1.1a)-(1.1.1b)\) in the PNP system \((1.1.1)\) can be reformulated as

\[
\partial_t \rho_i = \nabla \cdot \left( D_i(x)e^{-\psi_i} \nabla (e^{\psi_i} \rho_i) \right) =: R^1(\rho_i, \psi_i),
\]

(1.3.7)
with

$$
\psi_i = \frac{q_i}{k_B T} \phi(x, t) + \frac{1}{k_B T} \mu_i(x).
$$

We use the semi-implicit time discretization

$$
\frac{\rho_i^{n+1} - \rho_i^n}{\tau} = \nabla \cdot \left( D_i(x) e^{-\psi_i^n} \nabla(e^{\psi_i^n} \rho_i^{n+1}) \right) =: R^1_i(\rho_i^{n+1}, \psi_i^n), \quad (1.3.8)
$$

here

$$
\psi_i^n = \frac{q_i}{k_B T} \phi^n + \frac{1}{k_B T} \mu_i,
$$

the electrostatic potential $\phi^n$ in $\psi_i^n$ is obtained by solving the Poisson equation (1.1.1c) with $\rho^n$. The semi-discrete scheme (1.3.8) is well-posed and preserves solution positivity for arbitrary time steps (see Chapter 4, Theorem 4.3.1). The second order in time scheme is

$$
\frac{\rho_i^n - \rho_i^{n-1}}{\tau/2} = R^1_i(\rho_i^n, \frac{3}{2} \psi_i^n - \frac{1}{2} \psi_i^{n-1}), \quad \rho_i^{n+1} = 2\rho_i^n - \rho_i^{n-1}. \quad (1.3.9)
$$

For the spatial discretization we use the second order central difference approximation. The fully discrete first order (first order in time and second order in space) accurate scheme preserves solution positivity for arbitrary time steps (Chapter 4, Theorem 4.3.3). For the second order (in both time and space) scheme, solution positivity for large time steps is restored by the positivity-preserving local limiter (1.3.5).

### 1.4 Thesis organization

This thesis is organized as follows. In Chapter 2, a positive and energy stable finite-volume method is introduced for solving the reduced Poisson-Nernst-Planck system (1.1.11). By rewriting the underlying system in non-logarithmic Landau form and using a semi-implicit time discretization, we construct a simple, easy-to-implement numerical scheme which is proved to satisfy the solution positivity independent of time step size and the choice of Poisson solvers. Our scheme also preserves total mass and satisfies a discrete free energy dissipation property for zero flux boundary conditions. Numerical experiments are presented to simulate ionic channels in different settings. Chapter 3 is devoted to extending the unconditional positive scheme introduced in Chapter 2 to
a second order scheme in both space and time, for solving the reduced PNP system (1.1.11). The scheme is based on a linearized Crank-Nicolson method. The obtained numerical solutions are nonnegative for suitably small time steps; for large time steps, positivity of the numerical solutions is enforced by a novel positivity preserving limiter. In Chapter 4, we extend our results to the multi-dimensional PNP system (1.1.1) in more general settings. The semi-implicit time discretization based on a reformulation of the system gives a well-posed elliptic system, which is shown to preserve solution positivity for arbitrary time steps. The first order (in time) fully-discrete scheme is shown to preserve solution positivity and mass conservation unconditionally, and energy dissipation with only a mild $O(1)$ time step restriction. The scheme is also shown to preserve the steady-states. For the fully second order (in both time and space) scheme with large time steps, solution positivity is restored by a local scaling limiter. Moreover, we prove that the limiter does not destroy the approximation accuracy. Several three-dimensional numerical tests are conducted to verify our theoretical findings and demonstrate the accuracy, efficiency, and robustness of the proposed schemes. In Chapter 5, we further extend our results to construct positive and energy dissipating schemes for the class of FP equations with interaction potentials (1.2.1). Finally, we give concluding remarks in Chapter 6.

References


CHAPTER 2. UNCONDITIONAL POSITIVITY-PRESERVING AND ENERGY STABLE SCHEMES FOR A REDUCED POISSON-NERNST-PLANCK SYSTEM

A paper published in *Communications in Computational Physics*

Hailiang Liu and Wumaier Maimaitiyiming

Abstract

The Poisson-Nernst-Planck (PNP) system is a widely accepted model for simulation of ionic channels. In this paper, we design, analyze, and numerically validate a second order unconditional positivity-preserving scheme for solving a reduced PNP system, which can well approximate the three dimensional ion channel problem. Positivity of numerical solutions is proven to hold true independent of the size of time steps and the choice of the Poisson solver. The scheme is easy to implement without resorting to any iteration method. Several numerical examples further confirm the positivity-preserving property and demonstrate the accuracy, efficiency, and robustness of the proposed scheme, as well as the fast approach to steady states.

2.1 Introduction

Biological cells exchange chemicals and electric charge with their environments through ionic channels in the cell membrane walls. Examples include signaling in the nervous system and coordination of muscle contraction, see [5] for a comprehensive introduction. Mathematically the flow of ions can be modeled by drift-diffusion equations such as the Poisson-Nernst-Planck (PNP) system, see e.g. [4, 6, 7, 12].

In this investigation we design, analyze and numerically validate positivity-preserving algorithms to solve time-dependent drift-diffusion equations. As a first step, in this paper we focus on a reduced
model derived by Gardner et al [12] as an approximation to the full three dimensional (3D) PNP system. Let us first recall the full model and its reduction.

2.1.1 Mathematical models

The general setup in [12] is a flow of positive and negative ions in water in a channel plus surrounding baths in an electric field against a background of charged atoms on the channel protein. The distribution of charges is described by continuum particle densities \( c_i(x, t) \) for the mobile ions (such as \( K^+, Na^+, Ca^{++}, \cdots \)). The flow of ions can be modeled by the PNP system of \( m+1 \) equations

\[
\begin{align*}
\partial_t c_i &= -\nabla \cdot J_i, \quad i = 1, \cdots, m; \ x \in \Omega \subset \mathbb{R}^3, \ t > 0,
J_i &= -(D_i \nabla c_i + z_i \mu_i c_i \nabla \psi),
-\nabla \cdot (\epsilon \nabla \psi) &= \sum_{i=1}^m q_i c_i - e \rho,
\end{align*}
\] (2.1.1)

where \( J_i \) is the flux density, in which \( D_i \) is the diffusion coefficient, \( \mu_i \) the mobility coefficient which is related to the diffusion coefficient via Einstein’s relation \( \mu_i = \frac{D_i}{k_B T_0} \), where \( k_B \) is the Boltzmann constant and \( T_0 \) is the absolute temperature [5]. In the Poisson equation, \( \epsilon \) is the dielectric coefficient, \( q_i \) is the ionic charge for each ion species \( i \), \( \rho = \rho(x) \) is the permanent fixed charge density, and \( e \) is the proton charge. The coupling parameter is \( z_i = q_i/e \). In general, the physical parameters \( \epsilon, \mu_i \) and \( D_i \) are functions of \( x \). Let us mention that the case of no permanent charge does not pertain to biological channels. Even channels without permanent charge (in the form of so called acid and base side chains) have large amounts of fixed charge in their (for example) carbonyl bonds (see, e.g., [17] and references therein).

The derivation of the Nernst-Planck equation typically follows two steps, namely, using the energy variation to obtain the chemical potential and then using Fick’s laws of diffusion to attain the Nernst-Planck equation (see e.g. [11]). In the charge dynamics modeled by the traditional NP equation, mobile ions are treated as volume-less point charges. In order to incorporate more complex effects such as short-range steric effect and long range Coulomb correlation, modifications of the PNP equations were derived (see, e.g., [22] and references therein). Nonetheless, the scheme methodology proposed in this paper can well be adapted to solve such modified PNP systems.
The 3D geometry of the ion channel can be approximated by a reduced problem along the axial direction $x$, with a cross-sectional area $A(x)$ [29, 30]. Subject to a further rescaling as in [13], the corresponding PNP system (2.1.1) reduces to the following equations

$$
\frac{\partial_t c_i}{A(x)} = \frac{1}{A(x)} \partial_x (A(x) D_i (\partial_x c_i + z_i c_i \partial_x \psi)), \quad x \in \Omega = [0, 1], \quad t > 0,
$$

$$
-\frac{1}{A(x)} \partial_x (\epsilon A(x) \partial_x \psi) = \sum_{i=1}^{m} z_i c_i - \rho(x), \quad x \in \Omega, \quad t > 0.
$$

(2.1.2)

For ionic channels, an important characteristic is the so-called current-voltage relation, which can characterize permeation and selectivity properties of ionic channels (see [1] and references therein).

For (2.1.1), the electric current density (charge flux) is $J = \sum_{i=1}^{m} q_i J_i$. Such quantity for (2.1.2) reduces to

$$
J = -\sum_{i=1}^{m} z_i D_i A(x) (\partial_x c_i + z_i c_i \partial_x \psi).
$$

(2.1.3)

System (2.1.2) is a parabolic/elliptic system of partial differential equations, boundary conditions for both $c_i$ and $\psi$ can be Dirichlet or Neumann.

In order to solve the above reduced system, we consider initial data

$$
c_i(x, 0) = c_i^{in}(x) \geq 0, \quad x \in \Omega.
$$

2.1.2 Boundary conditions and model properties

We consider two types of boundary conditions. The first is the Dirichlet boundary conditions,

$$
c_i(0, t) = c_{i,l}, \quad c_i(1, t) = c_{i,r}; \quad \psi(0, t) = 0, \psi(1, t) = V, \quad t > 0,
$$

(2.1.4)

here $c_{i,l}, c_{i,r}$ are non-negative constants, and $V$ is a given constant. This is the setting adopted in [12]. One important solution property is

$$
c_i(x, t) \geq 0, \quad x \in \Omega, \quad t > 0.
$$

(2.1.5)

Another set of boundary conditions is as follows:

$$
\partial_x c_i + z_i c_i \partial_x \psi = 0, \quad x = 0, 1, \quad t > 0,
$$

$$
(-\eta \partial_x \psi + \psi)|_{x=0} = \psi_-, \quad (\eta \partial_x \psi + \psi)|_{x=1} = \psi_+, \quad t > 0,
$$

(2.1.6)
here $\psi_-, \psi_+$ are given constants, the size of parameter $\eta$ depends on the properties of the surrounding membrane [10]. In (2.1.6), the first one is the zero-flux boundary condition for the transport equation, and the second is the Robin boundary condition for the Poisson equation. Such boundary condition is adopted in [10] to model the effects of partially removing the potential from the ends of the channel. For system (2.1.2) with this boundary condition, solutions have non-negativity, mass conservation, and free energy dissipation properties, i.e., (2.1.5),

$$\int_\Omega A(x)c_i(x,t)dx = \int_\Omega A(x)c_i^{in}(x)dx, \quad t > 0, \quad i = 1, \cdots, m,$$

and

$$\frac{dE}{dt} = -\int_\Omega \sum_{i=1}^m A(x)D_i c_i |\partial_x (\log c_i + z_i \psi)|^2 dx \leq 0, \quad (2.1.8)$$

here the total energy $E$ associated to (2.1.2) is defined (see [10]) by

$$E = \int_\Omega A(x) \left( \sum_{i=1}^m c_i \log c_i + \frac{1}{2} \left( \sum_{i=1}^m z_i c_i - \rho \right) \psi \right) dx + \frac{\epsilon}{2\eta} (\psi_+ A(1) \psi(1) + \psi_- A(0) \psi(0)). \quad (2.1.9)$$

The positivity-preserving property is of special importance, since negative values in density would violate the physical meaning of the solution and may destroy the energy dissipation law (2.1.8). Numerical techniques addressing the positivity preserving property have been introduced in various application problems, see e.g. [18, 25].

In this paper, we construct second order accurate unconditional positivity-preserving schemes for solving (2.1.2) subject to two types of boundary conditions. For the zero-flux boundary condition, the schemes will be shown to satisfy mass conservation and a discrete energy dissipation law.

**2.1.3 Related works**

Numerical methods for solving the PNP system of equations have been studied extensively; see e.g., [12, 14, 16, 24, 32]. We also refer to [3] for a review on the PNP model and its generalizations for ion channel charge transport.

For the reduced PNP system (2.1.2), the finite difference scheme with TR–BDF2 time integration was first pursued in [12] to simulate an ionic channel. For the one dimensional PNP system,
the second order implicit finite difference scheme proposed in [10] can preserve total concentration of ions with the aid of a special boundary discretization, but numerical solutions may not be positive or energy dissipating. An improved scheme, further introduced in [9], can preserve a discrete form of energy dissipation law up to $O(\tau^2 + h^2)$, where $\tau$ is the time step, and $h$ is the spatial mesh size. In [2] the authors proposed an adaptive conservative finite volume method on a moving mesh that maintains solution positivity. The second order finite difference scheme in [19] is explicit and shown to preserve positivity, mass conservation, and energy dissipation, while the positivity-preserving property is ensured if $\tau = O(h^2)$. Further extension in [20] is a free energy satisfying discontinuous Galerkin scheme of any high order, where positivity-preserving property is realized by limiting techniques. The finite element scheme obtained by the method of lines approach in [26] preserves positivity of the solutions and a discrete energy dissipation law. Recently in [15] the authors presented a fully implicit finite difference method where both positivity and energy decay are preserved. In their scheme a fixed point iteration is needed for solving the resulting nonlinear system. These schemes are either explicit or fully implicit in time, the former require a time step restriction for preserving the desired properties while the later preserve desired properties unconditionally but they had to be solved by some iterative solvers.

In this paper we design schemes to preserve all three desired properties of the solutions: positivity, mass conservation, and energy dissipation. The key ingredients include a reformulation of the equation in its non-logarithmic Landau form and the use of the implicit-explicit time discretization, these together ensure the positivity-preserving property without any restriction on the size of time steps (unconditional!) and do not require iterative solvers.

2.1.4 Contributions and organization of the paper

Our scheme construction is based on the reformulation

$$A(x)\partial_t c_i(x, t) = \partial_x (A(x)D_i e^{-z_i\psi(x,t)}\partial_x(c_i(x, t)e^{z_i\psi(x,t)})$$, (2.1.10)
of the transport equation in (2.1.2). Similar formulation has been used in earlier works [19, 21].

Here we adopt a semi-implicit time discretization of (2.1.10):

\[
A(x)\frac{c_i^{n+1}(x) - c_i^n(x)}{\tau} = \partial_x \left( A(x)D_i e^{-z_i \psi(x)} \partial_x (c_i^{n+1}(x)e^{z_i \psi(x)}) \right).
\]  

(2.1.11)

The feature of such discretization is that it is a linear equation in \(c_i^{n+1}(x)\), and easy to solve numerically. For spatial discretization, we use the central finite volume approach. The coefficient matrix of the resulting linear system is a M-matrix and the right hand side is a nonnegative vector, thus positivity of the solution is ensured without any time step restriction.

The main contribution in this paper includes the model reformulation, proofs of unconditional positivity-preserving properties for two types of boundary conditions, and of mass conservation and energy dissipation properties for zero flux boundary conditions (2.1.6). In addition, the positivity-preserving property is shown to be independent of the choice of Poisson solvers. Our implicit-explicit scheme is easy to implement and efficient in computing numerical solutions over long time.

The paper is organized as follows. In section 2, we derive our numerical scheme for a model equation. Theoretical analysis of unconditional positivity is provided. In section 3, we formulate our scheme to the PNP system and prove positivity, mass conservation and energy dissipation properties of the scheme. Numerical examples are presented in section 4. Finally, concluding remarks are given in section 5.

### 2.2 Numerical methods for a model equation

In this section, we first demonstrate the key ideas through a model problem. Let \(u(x, t)\) be an unknown density, satisfying

\[
A(x)\partial_t u(x, t) = \partial_x (B(x)(\partial_x u(x, t) - u(x, t)\partial_x \phi(x, t))), \quad x \in \Omega = [0, 1], \quad t > 0,
\]

\[
u(x, 0) = u^{in}(x), \quad x \in \Omega,
\]

(2.2.1)

here \(A(x) > 0, B(x) > 0\) are given functions, and \(\phi(x, t)\) is either known or can be obtained from solving another coupled equation. For this model problem, we consider two types of boundary conditions:
(i) the Dirichlet boundary condition

\[ u(0, t) = u_l, \quad u(1, t) = u_r, \quad t > 0, \] (2.2.2)

and (ii) the zero flux boundary condition

\[ \partial_x u(x, t) - u(x, t) \partial_x \phi(x, t) = 0, \quad x = 0, 1, \quad t > 0. \] (2.2.3)

2.2.1 Scheme formulation

Let \( N \) be an integer, and the domain \( \Omega = [0, 1] \) be partitioned into computational cells \( I_j = [x_{j-1/2}, x_{j+1/2}] \) with cell center \( x_j = x_{j-1/2} + \frac{1}{2} h \), for \( j \in \{1, 2, \cdots, N\} \), \( x_{1/2} = 0 \) and \( x_{N+1/2} = 1 \). For simplicity, uniform mesh size \( h = \frac{1}{N} \) is adopted. Discretize \( t \) uniformly as \( t_n = \tau n \), here \( \tau \) is time step.

From the reformulation

\[ A(x) \partial_t u(x, t) = \partial_x (B(x) e^{\phi(x,t)} \partial_x (u(x, t) e^{-\phi(x,t)})) \] (2.2.4)

of (2.2.1), we consider a semi-implicit time discretization as follows:

\[ A(x) \frac{u^{n+1}(x) - u^n(x)}{\tau} = \partial_x \left( B(x) e^{\phi^n(x)} \partial_x (u^{n+1}(x) e^{-\phi^n(x)}) \right), \] (2.2.5)

here \( u^n(x) \approx u(x, t_n) \), \( \phi^n(x) \approx \phi(x, t_n) \). Let \( u^n_j \approx \frac{1}{h} \int_{I_j} u^n(x) dx \), and \( A_j = \frac{1}{h} \int_{I_j} A(x) dx \), then a fully-discrete scheme of (2.1.11) can be given by

\[ A_j u^{n+1}_j - u^n_j = \frac{U_{j+1/2} - U_{j-1/2}}{h}, \] (2.2.6)

the flux on interior interfaces are defined by

\[ U_{j+1/2} = B_{j+1/2} e^{\phi^{n+1}_{j+1/2}} \frac{u^{n+1}_j e^{-\phi^n_j} - u^{n+1}_j e^{-\phi^n_{j+1}}}{h}, \quad j = 1, 2, \cdots, N - 1. \] (2.2.7)

Here \( B_{j+1/2} = B(x_{j+1/2}) \); For \( \phi^n_{j+1/2} \) we either use \( \phi(x_{j+1/2}, t_n) \) if \( \phi(x, t) \) is given, or

\[ \phi^n_{j+1/2} = \frac{\phi^n_j + \phi^n_{j+1}}{2}, \]

in which \( \phi^n_j \) is a numerical approximation of \( \phi(x_{j}, t_n) \).
The boundary fluxes are given as follows:

(i) for the Dirichlet boundary condition (2.2.2)

\[ U_{1/2} = B_{1/2} e^{\phi_{1/2}} \frac{2(u_{1}^{n+1} e^{-\phi_{1}^{n}} - u_{1} e^{-\phi_{1/2}^{n}})}{h}, \]
\[ U_{N+1/2} = B_{N+1/2} e^{\phi_{N+1/2}} \frac{2(u_{N} e^{-\phi_{N+1/2}^{n}} - u_{N}^{n+1} e^{-\phi_{N}^{n}})}{h}; \]

(ii) for the zero flux boundary condition (2.2.3),

\[ U_{1/2} = U_{N+1/2} = 0. \]

In either case, the initial data are determined by

\[ u_{j}^{0} = \frac{1}{h} \int_{I_{j}} u^{m}(x) dx, \quad j = 1, \cdots, N. \]

Before turning to the analysis of solution properties, we comment on these boundary fluxes.

Remark 2.2.1. The factor 2 in the boundary flux (2.2.8) suffices to ensure the first order accuracy in the approximation of

\[ B(x) e^{\phi(x,t)} \partial_{x}(u(x,t) e^{-\phi(x,t)}) \]

at the boundary; see [8]. However, the following flux without the factor 2, i.e.

\[ U_{1/2} = B_{1/2} e^{\phi_{1/2}} \frac{4u_{1}^{n+1} e^{-\phi_{1}^{n}} - 8u_{1} e^{-\phi_{1/2}^{n}}}{h}, \]
\[ U_{N+1/2} = B_{N+1/2} e^{\phi_{N+1/2}} \frac{4u_{N} e^{-\phi_{N+1/2}^{n}} - 8u_{N}^{n+1} e^{-\phi_{N}^{n}}}{h}, \]

(2.2.10)

can produce only a zeroth order approximation at the boundary. Order loss of accuracy has been observed in our numerical tests when (2.2.10) is used.

An alternative boundary flux for (i) is a second order approximation of the form

\[ U_{1/2} = B_{1/2} e^{\phi_{1/2}} \frac{-\frac{1}{3}u_{2}^{n+1} e^{-\phi_{2}^{n}} + 3u_{1}^{n+1} e^{-\phi_{1}^{n}} - 8u_{1} e^{-\phi_{1/2}^{n}}}{h}, \]
\[ U_{N+1/2} = B_{N+1/2} e^{\phi_{N+1/2}} \frac{1u_{N-1}^{n+1} e^{-\phi_{N-1}^{n}} - 3u_{N}^{n+1} e^{-\phi_{N}^{n}} + 8u_{N} e^{-\phi_{N+1/2}^{n}}}{h}. \]

(2.2.11)

However, it is known that the first order boundary flux does not destroy the second order accuracy of the scheme, we refer to [31] for a such result regarding the Shortley-Weller method. Hence throughout the paper, we will not discuss high order boundary fluxes such as (2.2.11).
2.2.2 Positivity

It turns out that both schemes, (2.2.6)-(2.2.7)-(2.2.8) and (2.2.6)-(2.2.7)-(2.2.9), preserve positivity of numerical solutions without any time step restriction.

**Theorem 2.2.1.** Scheme (2.2.6)-(2.2.7) with either (i) (2.2.8) and \( u_t \geq 0, u_r \geq 0 \), or (ii) (2.2.9), is positivity-preserving, in the sense that if \( u^n_j \geq 0 \) for all \( j = 1, \ldots, N \), then

\[
  u^{n+1}_j \geq 0 \text{ for all } j = 1, \ldots, N.
\]

**Proof.** Set mesh ratio \( \lambda = \frac{r}{h^2} \) and introduce \( G_j = u^{n+1}_j e^{-\phi_j^n} \), so that

(i) scheme (2.2.6), (2.2.7) and (2.2.8) can be rewritten as

\[
(A_1 e^{\phi_1^n} + \lambda B_3/2 e^{\phi_3^n/2} + 2\lambda B_1/2 e^{\phi_1^n/2})G_1 - \lambda B_3/2 e^{\phi_3^n/2} G_2 = A_1 u^n_1 + 2\lambda B_1/2 u_t,
\]

\[
- \lambda B_{j-1}/2 e^{\phi_{j-1}^n} G_{j-1} + (A_j e^{\phi_j^n} + \lambda B_{j+1}/2 e^{\phi_{j+1}^n/2} + \lambda B_{j-1}/2 e^{\phi_{j-1}^n/2}) G_j - \lambda B_{j+1}/2 e^{\phi_{j+1}^n/2} G_{j+1} = A_j u^n_j,
\]

\[
- \lambda B_{N-1}/2 e^{\phi_{N-1}^n} G_{N-1} + (A_N e^{\phi_N^n} + \lambda B_{N-1}/2 e^{\phi_{N-1}^n/2} + 2\lambda B_{N+1}/2 e^{\phi_{N+1}^n/2}) G_N = a_N u^n_N + 2\lambda B_{N+1}/2 u_r.
\]

This linear system of \( \{G_j\} \) admits a unique solution since its coefficient matrix is strictly diagonally dominant. Since \( u^{n+1}_j = e^{\phi_j^n} G_j \geq e^{\phi_j^n} G_k \), where

\[
G_k = \min_{1 \leq j \leq N} \{G_j\},
\]

it suffices to prove \( G_k \geq 0 \). We discuss in cases: if \( 1 < k < N \), then from the \( k \)-th equation of (2.2.12) with \( A_k > 0 \) it follows

\[
A_k u^n_k \leq - \lambda B_{k-1}/2 e^{\phi_{k-1}^n} G_k + (A_k e^{\phi_k^n} + \lambda B_{k+1}/2 e^{\phi_{k+1}^n/2} + \lambda B_{k-1}/2 e^{\phi_{k-1}^n/2}) G_k
\]

\[- \lambda B_{k+1}/2 e^{\phi_{k+1}^n/2} G_k = A_k e^{\phi_k^n} G_k.
\]

Hence \( G_k \geq u^n_k e^{-\phi_k^n} \geq 0 \); if \( k = 1 \), from the first equation of (2.2.12) we have

\[
A_1 u^n_1 + 2\lambda B_1/2 u_t \leq (A_1 e^{\phi_1^n} + \lambda B_3/2 e^{\phi_3^n/2} + 2\lambda B_1/2 e^{\phi_1^n/2}) G_1 - \lambda B_3/2 e^{\phi_3^n/2} G_1 = (A_1 e^{\phi_1^n} + 2\lambda B_1/2 e^{\phi_1^n}) G_1.
\]

This implies \( G_1 \geq 0 \); so does the case if \( k = N \).

(ii) Likewise, scheme (2.2.6), (2.2.7) and (2.2.9) can be rewritten as

\[
(A_1 e^{\phi_1^n} + \lambda B_3/2 e^{\phi_3^n/2}) G_1 - \lambda B_3/2 e^{\phi_3^n/2} G_2 = A_1 u^n_1,
\]

\[- \lambda B_{j-1}/2 e^{\phi_{j-1}^n} G_{j-1} + (A_j e^{\phi_j^n} + \lambda B_{j+1}/2 e^{\phi_{j+1}^n/2} + \lambda B_{j-1}/2 e^{\phi_{j-1}^n/2}) G_j - \lambda B_{j+1}/2 e^{\phi_{j+1}^n/2} G_{j+1} = A_j u^n_j,
\]

\[- \lambda B_{N-1}/2 e^{\phi_{N-1}^n} G_{N-1} + (A_N e^{\phi_N^n} + \lambda B_{N-1}/2 e^{\phi_{N-1}^n/2}) G_N = A_N u^n_N.
\]
Using an entirely same argument, we can show $G_j \geq 0$, hence $u_j^{n+1} \geq 0$ for all $j$ involved.

Remark 2.2.2. The specific values or choices of $\{\phi_j^n\}$ and $\{\phi_{j+1/2}^n\}$ do not affect the unconditional positivity property of the scheme for $\{u_j^n\}$. This result thus can be applied to the case when $\phi(x,t)$ is solved by the Poisson equation, see the next section.

2.3 Positive schemes for the reduced PNP-system

The reduced PNP system (2.1.2) is reformulated as

$$
A(x) \partial_t c_i = \partial_x (A(x)D_i e^{-z_i \psi} \partial_x (c_i e^{z_i \psi})),
$$

$$
- \partial_x (\epsilon A(x) \partial_x \psi) = A(x) \left( \sum_{i=1}^m z_i c_i - \rho(x) \right).
$$

(2.3.1)

Let $c_{i,j}^n$ and $\psi_j^n$ approximate the cell average $\frac{1}{h} \int_{I_j} c_i(x,t_n)dx$ and $\frac{1}{h} \int_{I_j} \psi(x,t_n)$ respectively, then from the discretization strategy in section 2 the fully discrete scheme for system (2.3.1) follows

$$
A_j \frac{c_{i,j}^{n+1} - c_{i,j}^n}{\tau} = C_{i,j+1/2} - C_{i,j-1/2},
$$

$$
- \frac{\Psi_{j+1/2}^n - \Psi_{j-1/2}^n}{h} = A_j \left( \sum_{i=1}^m z_i c_{i,j}^n - \rho_j \right),
$$

(2.3.2)

(2.3.3)

the numerical fluxes on interior interfaces are defined by

$$
C_{i,j+1/2} = A_{j+1/2} D_i e^{-z_i \psi_{j+1/2}^n} \left( c_{i,j+1/2}^{n+1} e^{z_i \psi_{j+1/2}^n} - c_{i,j+1/2}^n e^{z_i \psi_{j+1/2}^n} \right),
$$

$$
\Psi_{j+1/2}^n = \epsilon A_{j+1/2} \frac{\psi_{j+1/2}^n - \psi_j^n}{h},
$$

(2.3.4)

(2.3.5)

here relevant terms are determined by

$$
A_j = \frac{1}{h} \int_{I_j} A(x)dx, \quad \rho_j = \frac{1}{h} \int_{I_j} \rho(x)dx
$$

$$
A_{j+1/2} = A(x_{j+1/2}), \quad \psi_{j+1/2}^n = (\psi_j^n + \psi_{j+1/2}^n)/2.
$$

For non-trivial $A(x), \rho(x)$, numerical integration of high accuracy is used to evaluate $A_j$ and $\rho_j$.

The boundary fluxes are defined as follows:
(i) for Dirichlet boundary condition (2.1.4),
\[
C_{i,1/2} = A_{1/2}D_i \frac{2(c_{n+1}^{i,1}e^{z_i\psi^n_i} - c_{i,l})}{h},
\]
\[
C_{i,N+1/2} = A_{N+1/2}D_i e^{-z_i V} \frac{2(c_{i,r}e^{z_i V} - c_{n+1}^{i,N}e^{z_i\psi^n_i})}{h},
\]
\[
\Psi_{1/2}^n = \epsilon A_{1/2} \frac{2\psi_1^n}{h},
\]
\[
\Psi_{N+1/2}^n = \epsilon A_{N+1/2} \frac{2(V - \psi_N^n)}{h},
\]
(2.3.6)

(ii) for boundary condition (2.1.6):
\[
C_{i,1/2} = 0, \quad C_{i,N+1/2} = 0,
\]
\[
\Psi_{1/2}^n = \frac{\epsilon}{\eta} A_{1/2}(\psi_1^n - \psi_-), \quad \Psi_{N+1/2}^n = \frac{\epsilon}{\eta} A_{N+1/2}(\psi_+ - \psi_N^n).
\]
(2.3.7)

### 2.3.1 Scheme properties

Scheme (2.3.2)-(2.3.5) with (2.3.6) turns out to be unconditionally positivity-preserving.

**Theorem 2.3.1.** Let \( \psi^n_j \) and \( c_{n+1}^{i,j} \) for \( i = 1, \ldots, m, j = 1, \ldots, N \) be obtained from (2.3.2)-(2.3.5) with (2.3.6). If \( c_{n}^{i,j} \geq 0 \) and \( c_{i,l} \geq 0, c_{i,r} \geq 0 \) for \( i = 1, \ldots, m, j = 1, \ldots, N \), then \( c_{n+1}^{i,j} \geq 0 \) for all \( i = 1, \ldots, m, j = 1, \ldots, N \).

**Proof.** For fixed \( i = 1, \ldots, m \), the scheme (2.3.2), (2.3.4) and (2.3.6) is of the same form as (2.2.6), (2.2.7) and (2.2.8) with \( u_j^n = c_{n}^{i,j}, B_{j+1/2} = A_{j+1/2}D_i, \phi_j^n = -z_i\psi_j^n \) and \( \phi_{j+1/2}^n = -z_i\psi_{j+1/2}^n \). From (i) in Theorem 2.2, we can conclude \( c_{n+1}^{i,j} = u_j^{n+1} \geq 0 \). \( \square \)

**Remark 2.3.1.** From the above analysis we see that positivity of \( c_{n+1}^{i,j} \) remains true even when another Poisson solver is used.

For scheme (2.3.2)-(2.3.5) with (2.3.7), it turns out that the solution \( c_{n,j}^{i} \) is mass conservative, non-negative, and energy dissipating. In order to state the energy dissipation result, we define a discrete version of the free energy (2.1.9) as

\[
E_h^n = \sum_{j=1}^{N} h A_j \left( \sum_{i=1}^{m} c_{n,i,j} \log c_{n,i,j} + \frac{1}{2} S_j^n \psi_j^n \right) + \frac{\epsilon}{2\eta} (\psi_+ A_{N+1/2} \psi_N^n + \psi_- A_{1/2} \psi_1^n),
\]
(2.3.8)
here

\[ S^n_j = \sum_{i=1}^{m} z_i c^n_{i,j} - \rho_j. \]

**Theorem 2.3.2.** Let \( \psi^n_j \) and \( c^n_{i,j} \) be obtained from (2.3.2)-(2.3.5) and (2.3.7), then we have:

1. **Conservation of mass:**
   \[
   \sum_{j=1}^{N} h A_j c^n_{i,j+1} = \sum_{j=1}^{N} h A_j c^n_{i,j} \quad \text{for } n \geq 0, i = 1, \cdots, m; \tag{2.3.9}
   \]

2. **Propagation of positivity:** if \( c^n_{i,j} \geq 0 \) for all \( j = 1, \cdots, N, \) and \( i = 1, \cdots, m, \) then
   \[ c^{n+1}_{i,j} \geq 0, \quad j = 1, \cdots, N, i = 1, \cdots, m; \]

3. **Energy dissipation:** there exists \( C^* > 0 \) depending on numerical solutions but independent on \( \tau \) and \( h, \) such that if \( \tau \leq C^* \epsilon / \eta, \) then
   \[
   E^{n+1}_h - E^n_h \leq -\frac{\tau}{2} I^n_h, \tag{2.3.10}
   \]

where

\[
I^n_h = \sum_{i=1}^{m} \sum_{j=1}^{N-1} \frac{1}{h} A_{j+1/2} D_i \left( (c^{n+1}_{i,j+1} e^{z_i \psi^n_j} + c^{n+1}_{i,j} e^{z_i \psi^n_j}) (\log c^{n+1}_{i,j+1} e^{z_i \psi^n_j} - \log c^{n+1}_{i,j} e^{z_i \psi^n_j}) \right) \geq 0.
\]

**Proof.** (1) Mass conservation follows from summing (2.3.2) over \( j = 1, \cdots, N \) and using (2.3.7).

(2) For each fixed \( i = 1, \cdots, m, \) this follows from (ii) in Theorem 2.2, by taking \( u^n_j = c^n_{i,j}, \)
\[ B_{j+1/2} = A_{j+1/2} D_i, \quad \phi^n_j = -z_i \psi^n_j \] and \( \phi^n_{j+1/2} = -z_i \psi^n_{j+1/2}. \)

(3) Using (2.3.8) we find that

\[
E^{n+1}_h - E^n_h = \sum_{j=1}^{N} h A_j \left( \sum_{i=1}^{m} (c^{n+1}_{i,j} - c^n_{i,j}) (\log c^{n+1}_{i,j} + z_i \psi^n_j) + \sum_{i=1}^{m} c^n_{i,j} \log c^{n+1}_{i,j} - \sum_{i=1}^{m} c^n_{i,j} \log c^n_{i,j} \right)
\]
\[
+ \frac{1}{2} \sum_{j=1}^{N} S^{n+1}_j \psi^{n+1}_j - \frac{1}{2} \sum_{j=1}^{N} S^n_j \psi^n_j - \sum_{i=1}^{m} z_i (c^{n+1}_{i,j} - c^n_{i,j}) \psi^n_j \n\]
\[
+ \frac{\epsilon}{2 \eta} (A_{N+1/2} \psi^n_N + A_{1/2} \psi^n_1 - A_{N+1/2} \psi^n_N + A_{1/2} \psi^n_1) =: I + II + III.
\]
We proceed to estimate term by term. For $I$, we use scheme (2.3.2)-(2.3.4) and (2.3.7) and summation by parts to obtain

\[
I = \sum_{j=1}^{N} hA_j \sum_{i=1}^{m} (c_{i,j}^{n+1} - c_{i,j}^{n})(\log c_{i,j}^{n+1} + z_i\psi_j^n)
\]

\[
= \tau \sum_{i=1}^{m} \sum_{j=1}^{N} (C_{i,j+1/2} - C_{i,j-1/2}) \log(c_{i,j}^{n+1} e^{z_i\psi_j^n})
\]

\[
= - \tau \sum_{i=1}^{m} \sum_{j=1}^{N-1} C_{i,j+1/2}(\log c_{i,j+1}^{n+1} e^{z_i\psi_j^{n+1}} - \log c_{i,j}^{n+1} e^{z_i\psi_j^n})
\]

\[
= - \tau \sum_{i=1}^{m} \sum_{j=1}^{N-1} \frac{1}{h} A_{j+1/2} D_i(c_{i,j+1}^{n+1} e^{z_i\psi_j^{n+1}} - c_{i,j}^{n+1} e^{z_i\psi_j^n})(\log c_{i,j+1}^{n+1} e^{z_i\psi_j^{n+1}} - \log c_{i,j}^{n+1} e^{z_i\psi_j^n})
\]

\[
= - \tau I^n_h \leq 0.
\]

For $II$, we use log$(X) \leq X - 1$ for $X > 0$, to obtain

\[
II = \sum_{j=1}^{N} hA_j \left( \sum_{i=1}^{m} c_{i,j}^{n} \log c_{i,j}^{n+1} - \sum_{i=1}^{m} c_{i,j}^{n} \log c_{i,j}^{n} \right)
\]

\[
= \sum_{j=1}^{N} hA_j \sum_{i=1}^{m} c_{i,j}^{n} \log \frac{c_{i,j}^{n+1}}{c_{i,j}^{n}}
\]

\[
\leq \sum_{i=1}^{m} \sum_{j=1}^{N} hA_j c_{i,j}^{n} \left( \frac{c_{i,j}^{n+1}}{c_{i,j}^{n}} - 1 \right)
\]

\[
= \sum_{i=1}^{m} \sum_{j=1}^{N} hA_j (c_{i,j}^{n+1} - c_{i,j}^{n}) = 0,
\]

in the last equality we have used conservation of mass.

Rearranging terms in $III$, we find that

\[
III = \sum_{j=1}^{N} hA_j \left( \frac{1}{2} S_j^{n+1} \psi_j^{n+1} + \frac{1}{2} S_j^n \psi_j^n - S_j^{n+1} \psi_j^n \right)
\]

\[
+ \frac{\epsilon}{2\eta} \left( A_{N+1/2} \psi_+^{n+1} + A_1 / 2 \psi_-^{n+1} \right) - \frac{\epsilon}{2\eta} \left( A_{N+1/2} \psi_+ \psi_N^n + A_1 / 2 \psi_- \psi_1^n \right)
\]

\[
= \frac{1}{2} \sum_{j=1}^{N} hA_j (S_j^{n+1} - S_j^n)(\psi_j^{n+1} - \psi_j^n) + F,
\]
here

\[
F = \frac{1}{2} \sum_{j=1}^{N} h A_j \left( S_j^n \psi_j^{n+1} - S_j^{n+1} \psi_j^n \right) + \frac{\epsilon}{2 \eta} \left( A_{N+1/2} \psi_N^{n+1} + A_{1/2} \psi_{-1}^{n+1} \right) \\
- \frac{\epsilon}{2 \eta} \left( A_{N+1/2} \psi_N^n + A_{1/2} \psi_{-1}^n \right).
\]

Tedious but elementary calculations show that \( F \equiv 0 \). Thus

\[
III = \frac{1}{2} \sum_{j=1}^{N} h A_j \left( S_j^n \psi_j^{n+1} - S_j^{n+1} \psi_j^n \right) (\psi_j^{n+1} - \psi_j^n).
\]  

(2.3.11)

Scheme (2.3.3)-(2.3.5) and (2.3.7) can be written in matrix form

\[
M \vec{\psi} = \vec{b},
\]

with

\[
M = \begin{bmatrix}
\frac{h}{\eta} A_{1/2} + A_{3/2} & -A_{3/2} \\
-A_{3/2} & A_{3/2} + A_{5/2} & -A_{5/2} \\
& \ddots & \ddots & \ddots \\
& -A_{N-3/2} & A_{N-3/2} + A_{N-1/2} & -A_{N-1/2} \\
& -A_{N-1/2} & \frac{h}{\eta} A_{N+1/2} + A_{N-1/2} & \frac{h}{\eta} A_{3/2} \\
\end{bmatrix},
\]

\[
\vec{b} = \frac{h^2}{\epsilon} \left( A_1 S_1^n + \frac{\epsilon}{h \eta} A_{1/2} \psi_-, A_2 S_2^n, \cdots, A_N S_N^n + \frac{\epsilon}{h \eta} A_{N+1/2} \psi_+ \right)^\top.
\]

Hence we have

\[
\psi_j^{n+1} - \psi_j^n = \frac{\tau h^2}{\epsilon} \sum_{k=1}^{N} (M^{-1})_{j,k} A_k D_t S_k^n, \quad \tau D_t S_j^n := S_j^{n+1} - S_j^n,
\]

thus (2.3.11) can be simplified as

\[
III = \frac{h^3 \tau^2}{2 \epsilon} \sum_{j=1}^{N} A_j D_t S_j^n \sum_{k=1}^{N} (M^{-1})_{j,k} A_k D_t S_k^n.
\]

We claim that for any \( \zeta \in \mathbb{R}^N \)

\[
\zeta \cdot M^{-1} \zeta \leq \frac{N^2 \eta}{(A_{1/2} + A_{N+1/2})} \| \zeta \|^2,
\]

(2.3.12)
with which we can bound $III$ as

$$III = \frac{h^3 \tau^2}{2\epsilon} \sum_{j=1}^{N} A_j D_t S_j^n \sum_{k=1}^{N} (M^{-1})_{j,k} A_k D_t S_k^n \leq \frac{\alpha N^2 h^3 \tau^2}{2\epsilon} \sum_{j=1}^{N} A_j^2 |D_t S_j^n|^2,$$  \hspace{1cm} (2.3.13)

here $\alpha^{-1} = A_{1/2} + A_{N+1/2}$. Note that $hN = 1$ and

$$|D_t S_j^n|^2 \leq m \sum_{i=1}^{m} z_i^2 (D_t c_i^n)^2,$$

we thus have

$$III \leq \sum_{i=1}^{m} \frac{\alpha \eta z_i^2 m \tau^2}{2\epsilon} \sum_{j=1}^{N} h A_j^2 (D_t c_i^n)^2.$$

Collecting estimates on $I, II$ and $III$ we arrive at

$$E_{n+1}^h - E_{n}^h \leq \sum_{i=1}^{m} \left( \tau \sum_{j=1}^{N} h A_j (D_t c_i^n) (\log c_i^{n+1} + z_i \psi_j^n) + \frac{\alpha \eta z_i^2 m \tau^2}{2\epsilon} \sum_{j=1}^{N} h A_j^2 (D_t c_i^n)^2 \right).$$

For (2.3.10) to hold, it remains to find a sufficient condition on time step $\tau$ so that for all $i = 1, \cdots, m$,

$$\frac{\alpha \eta z_i^2 m \tau^2}{2\epsilon} \sum_{j=1}^{N} h A_j^2 (D_t c_i^n)^2 \leq -\frac{\tau}{2} \sum_{j=1}^{N} h A_j (D_t c_i^n) (\log c_i^{n+1} + z_i \psi_j^n).$$  \hspace{1cm} (2.3.14)

This is nothing but

$$\frac{\alpha \eta z_i^2 m \tau^2}{2\epsilon} \|\xi\|^2 + \frac{\tau}{2} \vec{\xi} \cdot \vec{\mu} \leq 0,$$

here

$$\vec{\xi}_j = \sqrt{h} A_j D_t c_i^n, \quad \vec{\mu}_j = \sqrt{h} (\log c_i^{n+1} + z_i \psi_j^n).$$

Note that $I = \tau \vec{\xi} \cdot \vec{\mu} \leq 0$. One can verify using (2.3.4) and flux (2.3.2) that $\vec{\xi} \cdot \vec{\mu} = 0$ if and only if $\vec{\xi} = 0$. Therefore

$$0 < c_0 \leq \frac{-\vec{\xi} \cdot \vec{\mu}}{\|\xi\|^2} \leq \frac{\|\vec{\mu}\|}{\|\xi\|} \quad \text{for} \quad \vec{\xi} \neq 0,$$

where $c_0$ depends on the numerical solution at $t_n$ and $t_{n+1}$. We thus obtain (2.3.14) by taking

$$\tau \leq C^* \frac{\epsilon}{\eta}, \quad \text{where} \quad C^* = \min_{1 \leq i \leq m} \frac{c_0}{\alpha z_i^2 m} > 0.$$
Finally, we return to the proof of claim (2.3.12): For any $y \in \mathbb{R}^N$ with $\|y\| = 1$, we have the following

$$y \cdot M y = \frac{h}{\eta} A_{1/2} y_1^2 + \sum_{j=1}^{N-1} A_{j+1/2} (y_{j+1} - y_j)^2 + \frac{h}{\eta} A_{N+1/2} y_N^2$$

$$\geq \min_{\|y\|=1} \left\{ \frac{h}{\eta} A_{1/2} y_1^2 + \sum_{j=1}^{N-1} A_{j+1/2} (y_{j+1} - y_j)^2 + \frac{h}{\eta} A_{N+1/2} y_N^2 \right\}$$

$$= \frac{h}{N\eta} (A_{1/2} + A_{N+1/2}),$$

here the minimum is achieved at $y = (1, \cdots, 1)/\sqrt{N}$. Replacing $y$ by $y/\|y\|$ and then further set $y = M^{-1/2} \zeta$ leads to (2.3.12).

Remark 2.3.2. Though $C^*$ is not explicitly given, it is about $O(1)$ as can be seen from a formal limit $\Delta t \to 0$. The sufficient condition $\tau \leq C^* \epsilon/\eta$ suggests that for smaller $\epsilon/\eta$, one should consider a smaller time step to ensure the scheme stability. This is consistent with our numerical results.

2.4 Numerical tests

In this section, we implement the fully discrete scheme (2.3.2)-(2.3.5) with different boundary conditions. Errors are measured in the following discrete $l^\infty$ norm:

$$e_f = \max_{1 \leq j \leq N} |f_j - \bar{f}_j|.$$ 

Here $\bar{f}_j$ denotes the average of $f$ on cell $I_j$. In what follows we take $f_j = c_{i,j}^n$ or $\psi_j^n$ at time $t = n\tau$.

2.4.1 Accuracy test

In this example we numerically verify the accuracy and order of schemes (2.3.2)-(2.3.5) with first order boundary flux (2.3.6) and second order boundary flux of form (2.2.11).

Example 2.4.1. Consider the initial value problem with source term
\[
\begin{aligned}
\partial_t c_1 &= \frac{1}{A(x)} \partial_x (A(x) D_1 (\partial_x c_1 + z_1 c_1 \partial_x \psi)) + f_1(x,t), \quad x \in [0, 1], \ t > 0, \\
\partial_t c_2 &= \frac{1}{A(x)} \partial_x (A(x) D_2 (\partial_x c_2 + z_2 c_2 \partial_x \psi)) + f_2(x,t), \quad x \in [0, 1], \ t > 0, \\
\frac{1}{A(x)} \partial_x (eA(x) \partial_x \psi) &= z_1 c_1 + z_2 c_2 - \rho(x) + f_3(x,t), \quad x \in [0, 1], \ t > 0,
\end{aligned}
\]

\( (2.4.1) \)

Here we take \( A(x) = (5 - 4x)^2, \ D_1 = D_2 = 1, \ z_1 = -z_2 = 1, \ \epsilon = 1 \) and \( \rho(x) = 0 \), source terms are

\[
\begin{aligned}
f_1(x,t) &= \frac{4x^4 - 9x^3 + 53x^2 - 54x + 10}{4x - 5} e^{-t} + \frac{40x^7 - 71x^6 + 30x^5}{20} e^{-2t}, \\
f_2(x,t) &= \frac{4x^5 - 13x^4 + 94x^3 - 161x^2 + 84x - 10}{5 - 4x} e^{-t} + \frac{22x^8 - 60x^7 + 53x^6 - 15x^5}{10} e^{-2t}, \\
f_3(x,t) &= -\frac{2x^4}{5} e^{-t}.
\end{aligned}
\]

The exact solution to \( (2.4.1) \) is

\[
\begin{aligned}
c_1(x,t) &= x^2(1 - x)e^{-t}, \quad c_2(x,t) = x^2(1 - x)^2e^{-t}, \quad \text{and} \ \psi(x,t) = -\frac{x^5(3 - 2x)}{60}e^{-t}.
\end{aligned}
\]

We use the time step \( \tau = h^2 \) to compute numerical solutions. The errors and orders at \( t = 1 \) are listed in Table 2.1 and Table 2.2.

Table 2.1 Accuracy for Example 2.4.1 with first order boundary approximations (2.3.6)

<table>
<thead>
<tr>
<th>N</th>
<th>( c_1 ) error</th>
<th>order</th>
<th>( c_2 ) error</th>
<th>order</th>
<th>( \psi ) error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.11184E-03</td>
<td>-</td>
<td>0.57759E-04</td>
<td>-</td>
<td>0.83275E-05</td>
<td>-</td>
</tr>
<tr>
<td>80</td>
<td>0.28354E-04</td>
<td>1.9798</td>
<td>0.14407E-04</td>
<td>2.0033</td>
<td>0.20810E-05</td>
<td>2.0006</td>
</tr>
<tr>
<td>160</td>
<td>0.71370E-05</td>
<td>1.9902</td>
<td>0.36019E-05</td>
<td>1.9999</td>
<td>0.52013E-06</td>
<td>2.0003</td>
</tr>
<tr>
<td>320</td>
<td>0.17903E-05</td>
<td>1.9951</td>
<td>0.90047E-06</td>
<td>2.0000</td>
<td>0.13002E-06</td>
<td>2.0001</td>
</tr>
</tbody>
</table>

We see from Table 2.1 and Table 2.2 that both first and second order boundary fluxes yield second order convergent solutions. The numerical errors with both fluxes are comparable.

In the remaining numerical tests we only use the first order boundary flux (2.3.6) for the Dirichlet boundary value problem.
Table 2.2  Accuracy for Example 2.4.1 with second order boundary approximations (2.2.11)

<table>
<thead>
<tr>
<th>N</th>
<th>$c_1$ error</th>
<th>order</th>
<th>$c_2$ error</th>
<th>order</th>
<th>$\psi$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.10014E-03</td>
<td>-</td>
<td>0.69633E-04</td>
<td>-</td>
<td>0.37021E-05</td>
<td>-</td>
</tr>
<tr>
<td>80</td>
<td>0.25204E-04</td>
<td>1.9903</td>
<td>0.18005E-04</td>
<td>1.9514</td>
<td>0.93954E-06</td>
<td>1.9783</td>
</tr>
<tr>
<td>160</td>
<td>0.63218E-05</td>
<td>1.9952</td>
<td>0.45767E-05</td>
<td>1.9760</td>
<td>0.23755E-06</td>
<td>1.9837</td>
</tr>
<tr>
<td>320</td>
<td>0.15830E-05</td>
<td>1.9977</td>
<td>0.11536E-05</td>
<td>1.9882</td>
<td>0.59655E-07</td>
<td>1.9935</td>
</tr>
</tbody>
</table>

2.4.2 Effects of permanent charge and channel geometry

The key structure of an ion channel includes both the channel shape and the permanent charge (see e.g. [17]). We present numerical examples to illustrate the effects from the channel geometry or the permanent charge. While we also examine dependence of the total current (2.1.3) on voltage $V$, which is known as the current-voltage (I-V) relation in [17]. Note that (2.1.3) can be reformulated as

$$J = -\sum_{i=1}^{m} z_i D_i A(x) e^{-z_i \psi} \partial_x (c_i e^{z_i \psi}).$$

Let $J_{j+1/2}^n$ be an approximation of $J(x_{j+1/2}, t_n)$, then $J_{j+1/2}^n$ can be computed by

$$J_{j+1/2}^n = -\sum_{i=1}^{m} z_i C_{i,j+1/2},$$

here $C_{i,j+1/2}$ is defined in (2.3.4) with $c_{i,j}^{n+1}$ replaced by $c_{i,j}^n$, that is,

$$C_{i,j+1/2} = A_{j+1/2} D_i e^{-z_i \psi_{j+1/2}} \left( \frac{c_{i,j+1}^{n} e^{z_i \psi_{j+1/2}} - c_{i,j}^{n} e^{z_i \psi_{j+1/2}}}{h} \right), \quad j = 1, \cdots, N-1.$$

Example 2.4.2. (Effects of channel geometry with permanent charge) We consider the system

$$A(x) \partial_t c_1 = \partial_x (A(x)(\partial_x c_1 + c_1 \partial_x \psi)), \quad x \in [0, 1], \quad t > 0,$$

$$A(x) \partial_t c_2 = \partial_x (A(x)(\partial_x c_2 - c_2 \partial_x \psi)), \quad x \in [0, 1], \quad t > 0,$$

$$-\frac{1}{A(x)} \partial_x (\varepsilon A(x) \partial_x \psi) = c_1 - c_2 - \rho(x), \quad x \in [0, 1], \quad t > 0,$$

with $\varepsilon = 5 \times 10^{-5}$, subject to initial and boundary conditions

$$c_1(x, 0) = c_2(x, 0) = 0.5 - 0.1x, \quad x \in [0, 1],$$

$$c_i(0, t) = 0.5, \quad c_i(1, t) = 0.4; \quad \psi(0, t) = 0, \psi(1, t) = 0.5, \quad t > 0.$$
This corresponds to problem (2.1.2) with $D_1 = D_2 = 1$, $z_1 = -z_2 = 1$, with $c_{i,l} = 0.5$, $c_{i,r} = 0.4$, and $V = 0.5$.

The computational domain diagram is given in Figure 2.1 [12], while the cross sectional area $A(x)$ is defined as:

$$A(x) = \begin{cases} 
2(r_f + \frac{r_c - r_f}{l_b} x), & x \in [0, l_b], \\
2r_c, & x \in (l_b, l_b + l_c), \\
2(r_c + \frac{r_f - r_c}{l_b} (x - l_b - l_c)), & x \in [l_b + l_c, 1],
\end{cases} (2.4.4)$$

here the shape parameters are allowed to vary in our numerical tests. The permanent charge $\rho(x)$ is taken as

$$\rho(x) = \begin{cases} 
0, & x \in [0, l_b], \\
2Q_0, & x \in (l_b, l_b + l_c), \\
0, & x \in [l_b + l_c, 1],
\end{cases} (2.4.5)$$

with $Q_0$ a fixed constant.
Robert Eisenberg made clear to us the great importance of the tapered representation of the baths in one dimensional versions of PNP models of channels, that became clear in his early work with Wolfgang Nonner [27, 28], followed by many other more formal treatments such as in [12].

In this numerical test, we take $h = 0.01, \tau = 5 \times 10^{-5}$. The solutions are understood to have reached steady states if $||\psi^n - \psi^{n-1}||_{\infty} \leq 10^{-6}$. Table 2.3 shows times $t_s$ needed for reaching each steady state, number of iterations, and CPU times.

Table 2.3 Times needed for reaching each steady state in Example 2.4.2 when $Q_0 = 0.2, r_f = 20$ with different channel geometry

| channel parameters | $||\psi^n - \psi^{n-1}||_{\infty}$ | time $t_s$ | iterations $n = t_s/\tau$ | CPU time (sec) |
|--------------------|-----------------------------------|------------|---------------------------|---------------|
| $r_c = l_c = \frac{3}{5}$ | 9.9822E-07                      | 0.0744     | 1488                      | 0.5244        |
| $r_c = l_c = \frac{5}{11}$ | 9.9879E-07                      | 0.0992     | 1984                      | 0.6534        |
| $r_c = l_c = \frac{1}{11}$  | 9.9920E-07                      | 0.1116     | 2232                      | 0.7035        |

From Table 2.3 we see that $t_s = 0.1116$ is the longest time needed for reaching the steady state, so we run the simulation up to $t = 0.2$.

In Figure 2.2 we take $Q_0 = 0.2, r_f = 20$, varying $l_c$ and $r_c$ inside the channel, to obtain a series of snapshots. We see that both $c_1$ and $c_2$ coincide outside the channel, but split inside the channel with the shape evolving in terms of the channel geometry. The profile of $\psi$ looks similar.

In Figure 2.3 we fix the channel shape with $r_f = 20, r_c = 1/5, l_c = 1/5$, varying $Q_0$, we observe that the difference between $c_1$ and $c_2$ inside the channel increases in terms of $Q_0$, roughly we have $c_1 - c_2 \approx 2Q_0$ inside the channel. We can also observe the effects on $\psi$.

In Figure 2.4 is the I-V relation for the PNP system with channel shape parameters $l_c = 1/5, r_c = 1/5, r_f = 20$, and $Q_0 = 0.1$. We see from the figure that the current is linear in the voltage.

Example 2.4.3. (No permanent charge in the channel $\rho = 0$) We still use problem with (2.4.2), (2.4.3), (2.4.4) and (2.4.5) to test the effects of the channel geometry, by taking $\rho = 0, r_f = 20$ and varying $l_c$ and $r_c$. In the simulation we take $h = 0.01, \tau = 5 \times 10^{-5}$.

From Table 2.4 we see that $t_s = 0.0888$ is the longest time needed for reaching the steady state, so simulation runs up to $t = 0.1$. 


In Figure 2.5 are snap shots of solutions for different channel geometry. In the case of no permanent charge, there does not seem to be any layering phenomenon on $c_1$ and $c_2$: $c_1$ and $c_2$ are rather close both inside the channel (linear) and inside the bath (constant). The profile for $\psi$ is quite similar. This is consistent with the analysis in [17], in which the authors showed that the density in the channel gets steeper as the channel gets narrower. We refer to [23] for a study of steady state solutions to (2.4.2) in the case of $\rho = 0$. They proved that for $\epsilon > 0$ small, there is a unique nonnegative steady state to problem (2.4.2) and (2.4.3).

Table 2.4 Times needed for reaching each steady state on Example 2.4.3 when $\rho = 0, r_f = 20$ with different channel geometry
Figure 2.3 Effects of permanent charge on steady state densities and potential with $l_c = r_c = \frac{1}{5}$: (a)-(c) are computed densities at $t = 0.2$, for $Q_0 = 0.05$, 0.1, 0.15, (d)-(f) are potential profiles at $t = 0.2$ for $Q_0 = 0.05$, 0.1, 0.15.

Example 2.4.4. (Variable diffusion coefficient and quadratic area function) We consider the system

\begin{align}
A(x) \partial_t c_1 &= \partial_x (A(x) D_1(x) (\partial_x c_1 + 2c_1 \partial_x \psi)), \quad x \in [-10, 10], \; t > 0 \\
A(x) \partial_t c_2 &= \partial_x (A(x) D_2(x) (\partial_x c_2 - 3c_2 \partial_x \psi)), \quad x \in [-10, 10], \; t > 0 \\
A(x) \partial_t c_3 &= \partial_x (A(x) D_3(x) (\partial_x c_3 + c_3 \partial_x \psi)), \quad x \in [-10, 10], \; t > 0 \\
- \frac{1}{A(x)} \partial_x (\epsilon A(x) \partial_x \psi) &= 2c_1 - 3c_2 + c_3 - \rho(x), \quad x \in [-10, 10], \; t > 0,
\end{align}

(2.4.6)

with $\epsilon = 0.1$, subject to boundary conditions

\begin{align}
c_i(\pm 10, t) &= 0.5, \quad \psi(\pm 10, t) = 0, \quad t > 0.
\end{align}

(2.4.7)

As in [13], we choose $A(x) = 1 + x^2$, $D_i(x) = 20(1 - 0.9e^{-x^2})$ and $\rho = Ce^{-x^4}$. This corresponds to problem (2.1.2) with $z_1 = 2$, $z_2 = -3$, $z_3 = 1$, and $c_{i,l} = c_{i,r} = 0.5$, $V = 0$. In this numerical test we take $h = 0.1$, $\tau = 10^{-3}$. 
We take two different sets of initial data, first set is given as

\[
\begin{align*}
    c_1^{in}(x) &= 0.5 - 0.5e^{-(x+4)^4}, \\
    c_2^{in}(x) &= 0.5 + 2e^{-x^4}, \\
    c_3^{in}(x) &= 0.5 + e^{-(x-4)^4}.
\end{align*}
\]  

(2.4.8)

For the second set of initial data we take uniformly distributed random initial data \( c_{i,j}^0 \in (0, 1) \).

From Table 2.5 we see that \( t_s = 2.7410 \) is the longest time needed for reaching the steady state, so simulation runs up to \( t = 3 \). We vary the parameter \( C \) to observe effects of the permanent charge.

Table 2.5  Times needed for reaching steady state on Example 2.4.4 when \( C = 1 \) with different initial data

| initial data | \( ||\psi^n - \psi^{n-1}||_\infty \) | time \( t_s \) | iterations \( n = t_s/\tau \) | CPU time (sec) |
|--------------|-------------------------------|----------------|-----------------------------|----------------|
| data (2.4.8) | 9.9999E-08                    | 2.7410         | 2741                        | 0.3874         |
| random data  | 9.9864E-08                    | 2.0440         | 2044                        | 0.3167         |

In Figure 2.6 (top three) are snap shots of solutions for initial data (2.4.8). Varying \( C \), we can see that \( \max c_1 - \min c_2 \) (or \( \max c_3 - \min c_2 \)) increases in terms of \( C \).

In Figure 2.6 (bottom three) are snap shots of solutions for random initial data, we see that the choice of initial data does not affect the steady state densities.
2.4.3 Mass conservation and free energy dissipation

In this numerical test we demonstrate mass conservation and free energy dissipation properties.

Example 2.4.5. (Zero flux + Robin boundary conditions) In this example we consider (2.4.6) with initial condition (2.4.8) and boundary condition

\[
\partial_x c_i + z_i c_i \partial_x \psi = 0, \quad x = -10, 10, \quad t > 0,
\]

\[
(-\eta \partial_x \psi + \psi)|_{x=-10} = -0.1, \quad (\eta \partial_x \psi + \psi)|_{x=10} = 0.1, \quad t > 0.
\]

We choose same \(A(x), D_i(x), \rho(x)\) and \(z_i\) as in Example 4.4 and choose \(\eta = 0.1, \epsilon = 0.1\). In this numerical test we use scheme (2.3.2)-(2.3.5) and (2.3.7), with \(h = 0.1, \tau = 10^{-3}\).

In Figure 2.7 (left) are snap shots of solutions for initial data (2.4.8), (right) is free energy for the system and total mass for each species, which confirms energy dissipation and mass conservation properties as proved in Theorem 2.3.2.

2.5 Concluding Remarks

In this paper, we have developed an unconditional positivity-preserving finite-volume method for solving initial boundary value problems for the reduced Poisson-Nernst-Planck system. Such a
reduced system has been used as a good approximation to the 3D ion channel problem. By writing the underlying system in non-logarithmic Landau form and using a semi-implicit time discretization, we constructed a simple, easy-to-implement numerical scheme which proved to satisfy positivity independent of time steps and the choice of Poisson solvers. Our scheme also preserves total mass and satisfies a free energy dissipation property for zero flux boundary conditions. Extensive numerical tests have been presented to simulate ionic channels in different settings.

Acknowledgments

The authors would like to thank Robert Eisenberg for stimulating discussions on PNP systems and their role in modeling ion channels. This research was supported by the National Science Foundation under Grant DMS1812666 and by NSF Grant RNMS (KI-Net) 1107291.
Figure 2.7  Energy dissipation and mass conservation: (a) density profiles at $t = 15$ for $C = 1$, (b) energy dissipation and mass conservation.

References


CHAPTER 3. A SECOND ORDER POSITIVE SCHEME FOR THE REDUCED POISSON-NERNST-PLANCK SYSTEM

A paper submitted to *Journal of Computational and Applied Mathematics*

Hailiang Liu and Wumaier Maimaitiyming

Abstract

We design and analyze a second order positivity preserving scheme for a reduced Poisson-Nernst-Planck (PNP) system. The scheme is based on an implicit-explicit time discretization combined with a prediction-correction technique. The obtained solution is shown nonnegative for suitably small time steps; for large time steps, solution positivity is enforced by a local positivity-preserving limiter. The limiter is shown to preserve local mass and easy to implement. Numerical tests are given to validate the proven properties of the scheme with such limiter.

3.1 Introduction

Biological cells exchange chemicals and electric charge with their environments through ionic channels in the cell membrane walls. Examples include signaling in the nervous system and coordination of muscle contraction. One of the simplest ion channel models is the Poisson-Nernst-Planck (PNP) system, which treats the ion flow as the averaged ion concentration driven by the electrostatic potential force and the ion concentration gradient (see, e.g., [1, 2, 12]). Simulating the full system is a challenging task, but in many situations the three dimensional geometry of the ion channel can be modeled by a reduced one dimensional system along the axial direction [4, 10, 11].
In this note we focus on the reduced PNP system derived in [4]:

\[
\begin{cases}
\partial_t c_i = \frac{1}{A(x)} \partial_x (A(x) D_i (\partial_x c_i + z_i c_i \partial_x \psi)), & i = 1, \ldots, m; \ x \in \Omega, \ t > 0, \\
- \frac{1}{A(x)} \partial_x (\epsilon A(x) \partial_x \psi) = \sum_{i=1}^m z_i c_i - \rho_0(x), & x \in \Omega, \ t > 0, \\
c_i(x, 0) = c^{in}_i(x) \geq 0, & x \in \Omega.
\end{cases}
\tag{3.1.1}
\]

Here \(c_i = c_i(x, t)\) is the ion density for the \(i\)-th species, \(A(x)\) is cross sectional area, \(D_i\) is the diffusion coefficient, \(z_i\) is a rescaled charge parameter. In the Poisson equation, \(\psi = \psi(x, t)\) is the electric potential, \(\epsilon\) is the dielectric coefficient, \(\rho_0 = \rho_0(x)\) is the fixed charge density. The channel interval \(\Omega\) is set as \([a, b]\). We consider the Dirichlet boundary condition for both \(c_i\) and \(\psi\) as adopted in [4]:

\[
c_i(a, t) = c_{i,l}, \quad c_i(b, t) = c_{i,r}; \quad \psi(a, t) = 0, \psi(b, t) = V, \quad t > 0,
\tag{3.1.2}
\]

where \(c_{i,l}, c_{i,r}\) are non-negative constants, and \(V\) is a given constant. One important solution property is

\[
c_i(x, t) \geq 0, \quad x \in \Omega, \ t > 0.
\tag{3.1.3}
\]

This property is naturally desired for numerical solutions, since negative values in density would violate the physical meaning of the solution.

In this paper, we construct novel positive schemes to approximate solutions to the above initial boundary value problem by finite volume approximations. Our starting point is [6], and we refer the reader to it for more references to earlier results on related positive schemes in the literature. Some more closely related works are [3, 4, 6, 7, 8]. The reformulation of the system into a self-adjoint one is crucial for our discretization in [6], and has been known also useful in numerically approximating other Fokker-Planck equations, see e.g., [7, 9]. The implicit-explicit scheme introduced in [6] is second order in space, and shown to be positivity-preserving without any restriction on time steps, but it is only first order in time. In this note we extend the scheme in [6] to a scheme which is second order in both space and time. This is achieved by a prediction-correction strategy. For large time steps, solution positivity may not be guaranteed and is restored through a local limiter. Our numerical tests show that such limiter does not destroy the second order accuracy.
The paper is organized as follows. In section 2, we present our numerical schemes for the reduced PNP system, with proven positivity properties. In section 3, we present a local limiter and show three nice properties: positivity, local mass conservation, and the factors which can affect the accuracy, followed by a positivity-preserving numerical algorithm. Numerical examples are presented in section 4. Finally, concluding remarks are given in section 5.

3.2 Numerical schemes

In this section, we construct a novel second order positive scheme for (3.1.1). To compare the results in [6] and what we do here, we need to review some conventions.

Let $N > 0$ be an integer, domain $\Omega = [a, b]$ be partitioned into cells $I_j = [x_{j-1/2}, x_{j+1/2}]$ with $h_j = |I_j|$ and cell center $x_j = x_{j-1/2} + \frac{1}{2} h_j$ for $j \in \{1, 2, \cdots, N\}$, where $x_{1/2} = a$ and $x_{N+1/2} = b$, $h_{j+1/2} = (h_j + h_{j+1})/2$. Let $\tau$ be the time step and $t_n = n\tau$, $c^n_{i,j}$ and $\psi^n_j$ approximate the cell average $\frac{1}{h_j} \int_{I_j} c_i(x, t_n) dx$ and $\frac{1}{h_j} \int_{I_j} \psi(x, t_n)$, respectively. Then the fully discrete scheme introduced in [6] for system (3.1.1), based on the following reformulation

$$
\begin{aligned}
  A(x) \partial_t c_i &= \partial_x (A(x) D_i M_i \partial_x (c_i/M_i)), \quad M_i = e^{-z_i \psi}, \\
  -\partial_x (\epsilon A(x) \partial_x \psi) &= A(x) \left( \sum_{i=1}^m z_i c_i - \rho_0(x) \right),
\end{aligned}
$$

(3.2.1)

is of the form

$$
\begin{aligned}
  - \frac{\Psi^n_{j+1/2} - \Psi^n_{j-1/2}}{h_j} &= A_j \left( \sum_{i=1}^m z_i c^n_{i,j} - \rho_{0j} \right), \\
  A_j \frac{c^n_{i,j+1} - c^n_{i,j}}{\tau} &= \frac{C_{i,j+1/2} - C_{i,j-1/2}}{h_j},
\end{aligned}
$$

(3.2.2, 3.2.3)

where

$$
\begin{aligned}
  \Psi^n_{j+1/2} &= \epsilon A_{j+1/2} \frac{\psi^n_{j+1} - \psi^n_{j}}{h_{j+1/2}}, \quad j = 1, \cdots, N - 1, \\
  C_{i,j+1/2} &= \frac{A_{j+1/2} D_{i,j+1/2} M^n_{i,j+1/2}}{h_{j+1/2}} \left( \frac{c^n_{i,j+1} - c^n_{i,j}}{M^n_{i,j+1}} - \frac{c^n_{i,j}}{M^n_{i,j}} \right), \quad j = 1, \cdots, N - 1,
\end{aligned}
$$

(3.2.4, 3.2.5)
The boundary fluxes for the Dirichlet boundary condition (3.1.2) are defined as follows:

\[ c_{ij} = \frac{1}{h_j} \int_{I_j} A(x)dx, \quad A_{j+1/2} = A(x_{j+1/2}), \quad \rho_0 = \frac{1}{h_j} \int_{I_j} \rho_0(x)dx. \]

The initial data are determined by \( c_{i,j}^0 = \frac{1}{h_j} \int_{I_j} c_i(x)dx, \) for \( i = 1, \ldots, m, \quad j = 1, \ldots, N. \) The above scheme is shown unconditionally positivity-preserving in [6], summarized in the following.

**Theorem 3.2.1.** Let \( \psi_j^n \) and \( c_{i,j}^{n+1} \) for \( i = 1, \ldots, m, \quad j = 1, \ldots, N \) be obtained from (3.2.2)-(3.2.3).

If \( c_{i,j}^n \geq 0 \) and \( c_{i,j} \geq 0, \quad c_{i,r} \geq 0 \) for \( i = 1, \ldots, m, \quad j = 1, \ldots, N, \) then \( c_{i,j}^{n+1} \geq 0 \) for all \( i = 1, \ldots, m, \quad j = 1, \ldots, N. \)

In order to extend the above scheme to a second order (in time) scheme, we replace (3.2.3) by a two step scheme

\[ A_j \frac{c_{i,j}^* - c_{i,j}^n}{\tau/2} = \frac{C_{i,j+1/2}^* - C_{i,j-1/2}^*}{h_j}, \quad c_{i,j}^{n+1} = 2c_{i,j}^* - c_{i,j}^n, \tag{3.2.7} \tag{3.2.8} \]

here

\[ C_{i,j+1/2}^* = \frac{A_{j+1/2}D_i M_{i,j+1/2}}{h_{j+1/2}} \left( \frac{c_{i,j+1}^* - c_{i,j}^*}{M_{i,j+1}^*} \right), \quad j = 1, \ldots, N - 1, \tag{3.2.9} \]

with

\[ M_{i,j+1/2}^* = e^{-z_i(\frac{3}{2}\psi_j^n - \frac{1}{2}\psi_j^{n-1})}, \quad M_{i,j}^* = e^{-z_i(\frac{3}{2}\psi_j^n - \frac{1}{2}\psi_j^{n-1})}. \]

The boundary fluxes for the Dirichlet boundary condition (3.1.2) are defined as follows:

\[ C_{i,1/2}^* = \frac{2A_{1/2}D_i}{h_{1/2}} \left( \frac{c_{i,1}^* - c_{i,j}^*}{M_{i,1}^*} \right), \tag{3.2.10} \]

\[ C_{i,N+1/2}^* = \frac{2A_{N+1/2}D_j e^{-z_i V}}{h_{N+1/2}} \left( c_{i,r} e^{z_i V} - \frac{c_{i,N}^*}{M_{i,N}^*} \right). \]
One can formally verify that this scheme is second order accurate both in space and time. In addition, we have the following.

Theorem 3.2.2. Let \( c^{n+1}_{i,j} \) be obtained from (3.2.2), (3.2.7) and (3.2.8). If \( c^n_{i,j} \geq 0 \) and \( c_{i,l} \geq 0 \), \( c_{i,r} \geq 0 \) for \( i = 1, \cdots, m \), \( j = 1, \cdots, N \), then \( c^{n+1}_{i,j} \geq 0 \) for all \( i = 1, \cdots, m \), \( j = 1, \cdots, N \), provided \( \tau \) is suitably small.

Proof. Inserting (3.2.8) into (3.2.7), and regrouping so that all terms involving \( \tau \) stay on the left hand side, and the remaining terms are put on the right hand side, we obtain when setting

\[
G_j^n = c^n_{i,j}/M_{i,j}^*, \quad g_{j+1/2}^* = A_{j+1/2} D_{i,j} M_{i,j+1/2}^* \quad (1 \leq j < N) \quad \text{and} \quad g_{1/2}^* = \frac{2A_{1/2} D_i}{h_{1/2}}, \quad g_{N+1/2}^* = \frac{2A_{N+1/2} D_i e^{-3V}}{h_{N+1/2}},
\]

the linear system

\[
\begin{align*}
(A_1 M_{i,1}^* + \frac{\tau}{2h_1} (g_{3/2}^* + g_{1/2}^*)) G_1^{n+1} - \frac{\tau}{2h_1} g_{3/2}^* G_2^{n+1} &= b_1, \\
(A_j M_{i,j}^* + \frac{\tau}{2h_j} (g_{j+1/2}^* + g_{j-1/2}^*)) G_j^{n+1} - \frac{\tau}{2h_j} g_{j+1/2}^* G_{j+1}^{n+1} - \frac{\tau}{2h_j} g_{j-1/2}^* G_{j-1}^{n+1} &= b_j, \\
(A_N M_{i,N}^* + \frac{\tau}{2h_N} (g_{N+1/2}^* + g_{N-1/2}^*)) G_N^{n+1} - \frac{\tau}{2h_N} g_{N+1/2}^* G_{N+1}^{n+1} &= b_N,
\end{align*}
\]

(3.2.11)

here \( j = 1, \cdots, N \), with the right hand side vector given by

\[
\begin{align*}
b_1 &= \left( A_1 M_{i,1}^* - \frac{\tau}{2h_1} (g_{3/2}^* + g_{1/2}^*) \right) G_1^n + \frac{\tau}{2h_1} g_{3/2}^* G_2^n + \frac{\tau}{h_1} g_{1/2}^* c_{i,l}, \\
b_j &= \left( A_j M_{i,j}^* - \frac{\tau}{2h_j} (g_{j+1/2}^* + g_{j-1/2}^*) \right) G_j^n + \frac{\tau}{2h_j} g_{j+1/2}^* G_{j+1}^n + \frac{\tau}{2h_j} g_{j-1/2}^* G_{j-1}^n, \quad j = 1, \cdots, N, \\
b_N &= \left( A_N M_{i,N}^* - \frac{\tau}{2h_N} (g_{N+1/2}^* + g_{N-1/2}^*) \right) G_N^n + \frac{\tau}{2h_N} g_{N+1/2}^* G_{N+1}^n + \frac{\tau}{h_N} g_{N+1/2}^* c_{i,r}.
\end{align*}
\]

The linear system (3.2.11) of \( \{ G_j^{n+1} \} \) admits a unique solution since its coefficient matrix is strictly diagonally dominant. Note that \( c^{n+1}_{i,j} = M_{i,j}^* G_j^{n+1} \geq M_{i,j}^* G_k^{n+1} \), here

\[
G_k^{n+1} = \min_{1 \leq j \leq N} \{ G_j^{n+1} \},
\]

it suffices to prove \( G_k^{n+1} \geq 0 \). If \( 1 \leq k < N \), replacing \( G_k^{n+1} \) by \( G_k^{n+1} \) on the left side in the \( k \)-th equation of (3.2.11) leads to

\[
A_k M_{i,k}^* G_k^{n+1} \geq b_k.
\]
Similarly, if \( k = 1 \) or \( N \), we have

\[
\left( A_1 M_{i,1}^* + \frac{\tau}{2h_1} g_{i/2}^* \right) G_1^{n+1} \geq b_1 \quad \text{or} \quad \left( A_N M_{i,N}^* + \frac{\tau}{2h_N} g_{N+1/2}^* \right) G_N^{n+1} \geq b_N.
\]

Hence \( G_k^{n+1} \geq 0 \) is ensured if each \( b_j \geq 0 \), which is the case provided

\[
\tau \leq \min_{1 \leq j \leq N} \frac{2h_j A_j M_{i,j}^*}{g_{j+1/2}^* + g_{j-1/2}^*}.
\]

The stated result thus follows.

But for large time step \( \tau \), non-negativity of \( c_{i,j}^{n+1} \) may not be guaranteed. In the next section we introduce a local limiter to deal with this problem.

### 3.3 A local limiter

In this section, we construct a positivity preserving limiter. One could take a simple cutoff where the numerical solution is negative, but such treatment would violate the mass conservation. In order to preserve the mass conservation property, a limiter should satisfy certain local conservation property.

For the zero flux boundary condition, (3.2.10) needs to be replaced by

\[
C_{i,1/2}^* = 0, \quad C_{i,N+1/2}^* = 0,
\]

with which we can show that

\[
\sum_{j=1}^{N} h_j A_j c_{i,j}^n = \sum_{j=1}^{N} h_j A_j c_{i,j,0} = \sum_{j=1}^{N} A_j \int_{I_j} c_i^n(x) dx > 0.
\]

Set \( \rho_j = h_j A_j c_{i,j}^n \), then \( \sum_{j=1}^{N} \rho_j > 0 \). If \( \rho_k < 0 \) for some \( k \), there must exist neighboring values \( \rho_j \) with index \( j \in S_k \) such that

\[
\bar{\rho} = \frac{1}{|S_k|} \sum_{j \in S_k} \rho_j > 0.
\]

This local average can be used as a reference in the following scaling limiter,

\[
\tilde{\rho}_j = \theta \rho_j + (1 - \theta) \bar{\rho}, \quad j \in S_k,
\]  

(3.3.1)
\[ \theta = \min \left\{ 1, \frac{\bar{\rho}}{\bar{\rho} - \rho_{\min}} \right\}, \quad \rho_{\min} = \min_{j \in S_k} \rho_j. \]  

Lemma 3.3.1. This limiter has the following properties:

1. \( \tilde{\rho}_j \geq 0 \) for all \( j \in S_k \),
2. \( \sum_{j \in S_k} \tilde{\rho}_j = \sum_{j \in S_k} \rho_j \), and
3. \( |\tilde{\rho}_j - \rho_j| \leq |S_k^-| (-\rho_{\min}) \), where \( S_k^- = \{ j \in S_k : \rho_j < \bar{\rho} \} \).

Proof. (1) This follows from the definition of \( \theta \) and (3.3.1).

(2) By (3.3.1) and the definition of \( \bar{\rho} \), it follows that
\[ \sum_{j \in S_k} \tilde{\rho}_j = \theta |S_k\bar{\rho}| + (1 - \theta)\bar{\rho}|S_k| = \sum_{j \in S_k} \rho_j. \]

(3) From (3.3.1) it follows that for all \( j \in S_k \),
\[ |\tilde{\rho}_j - \rho_j| = (1 - \theta)|\bar{\rho} - \rho_j| = -\rho_{\min} \frac{|\bar{\rho} - \rho_j|}{\bar{\rho} - \rho_{\min}} \leq (-\rho_{\min}) \max \left\{ 1, \frac{\rho_{\max} - \bar{\rho}}{\bar{\rho} - \rho_{\min}} \right\}, \]

here we have used the notation \( \rho_{\max} := \max_{j \in S_k} \rho_j \). We split \( \sum_{j \in S_k} (\bar{\rho} - \rho_j) = 0 \) so that
\[ \sum_{j \in S_k^+} (\rho_j - \bar{\rho}) = \sum_{j \in S_k^-} (\bar{\rho} - \rho_j), \]
in which each term involved on both sides is nonnegative. Hence,
\[ \rho_{\max} - \bar{\rho} \leq |S_k^-| (\bar{\rho} - \rho_{\min}). \]

Thus the proof is complete. \( \square \)

The above limiter when returning to \( c_j := c_{i,j}^n \), an approximation of \( c(x) := c_i(x, t_n) \geq 0 \), gives
\[ \tilde{c}_j = \theta c_j + (1 - \theta) \frac{\bar{\rho}}{h_j A_j}, \]  

which can be shown to be nonnegative and satisfy the local mass conservation
\[ \sum_{j \in S_k} h_j A_j \tilde{c}_j = \sum_{j \in S_k} h_j A_j c_j. \]
In addition, from (3) in Lemma 3.3.1 we have
\[ |\tilde{c}_j - c_j| \leq |S^-_k| \frac{h_k A_k}{h_j A_j} \max_{x \in I_j} |c(x) - c_j| \leq C|S^-_k| \max_{x \in I_j} |c(x) - c_j|, \quad j \in S_k. \]

If $S^-_k$ is uniformly bounded in terms of $N$, then this relation implies that our local limiter does not destroy the accuracy. In practice, it is indeed the case as verified by our numerical tests.

### 3.3.1 Algorithm

Our algorithm is given as follows:

1. Initialization: From the given initial condition $c^i(x)$, obtain $c^0_{i,j} = \frac{1}{h_j} \int_{I_j} c^i(x) dx$, $i = 1, \ldots, m$, $j = 1, \ldots, N$, by using proper quadrature.

2. First update to get $\{c^1_{i,j}\}$: Compute $\{\psi^0_j\}$ from (3.2.2), and then obtain $\{c^1_{i,j}\}$ by the first order scheme (3.2.3).

3. Update from $\{c^n_{i,j}\}$: Compute $\{\psi^n_j\}$ from scheme (3.2.2), and then get $\{c^{n+1}_{i,j}\}$ from (3.2.7) and (3.2.8).

4. Reconstruction: in case $c^{n+1}_{i,j} < 0$, then replace it by $\tilde{c}^{n+1}_{i,j}$ using the limiter defined in (3.3.4).

The following algorithm can be called to find an admissible set $S_k$ used in (3.3.4).

(i) Start with $S_k = \{k\}$.

(ii) If $\min\{S_k\} > 1$, then set $S_k = S_k \cup \{\min\{S_k\} - 1\}$, else go to (iii). If $\bar{\rho} > 0$, then stop, else go to (ii).

(iii) If $\max\{S_k\} < N$, then set $S_k = S_k \cup \{\max\{S_k\} + 1\}$, else go to (ii). If $\bar{\rho} > 0$, then stop, else go to (ii).

The above limiter is local and features three nice properties. Through numerical tests in next section, we demonstrate its effectiveness also for problems with the Dirichlet boundary condition (3.1.2).
3.4 Numerical tests

In this section, we present numerical examples to test both accuracy and performance of the scheme (3.2.2), (3.2.7) and (3.2.8). We apply the scaling limiter at each time step if needed.

Example 3.4.1. (Accuracy test) We consider the initial-boundary value problem with source term,

\[
\begin{aligned}
\partial_t c_1 &= \frac{1}{A(x)} \partial_x (A(x)D_1(\partial_x c_1 + z_1 c_1 \partial_x \psi) + f_1(x,t), \quad x \in [0, 1], \ t > 0, \\
\partial_t c_2 &= \frac{1}{A(x)} \partial_x (A(x)D_2(\partial_x c_2 + z_2 c_2 \partial_x \psi) + f_2(x,t), \quad x \in [0, 1], \ t > 0, \\
-\frac{1}{A(x)} \partial_x (\epsilon A(x) \partial_x \psi) &= z_1 c_1 + z_2 c_2 - \rho_0(x) + f_3(x,t), \quad x \in [0, 1], \ t > 0,
\end{aligned}
\]

(3.4.1)

\[
\begin{aligned}
c_1(0,0) &= x^2(1-x), \quad c_1(0,t) = c_1(1,t) = 0, \\
c_2(0,0) &= x^2(1-x)^2, \quad c_2(0,t) = c_2(1,t) = 0, \\
\psi(0,t) &= 0, \quad \psi(1,t) = -\frac{1}{60} e^{-t}.
\end{aligned}
\]

Here we take \( A(x) = (5 - 4x)^2, D_1 = D_2 = 1, z_1 = -z_2 = 1, \epsilon = 1 \) and \( \rho_0(x) = 0 \), source terms are

\[
\begin{aligned}
f_1(x,t) &= \frac{4x^4 - 9x^3 + 53x^2 - 54x + 10}{4x - 5} e^{-t} + \frac{40x^7 - 71x^6 + 30x^5}{20} e^{-2t}, \\
f_2(x,t) &= \frac{4x^5 - 13x^4 + 94x^3 - 161x^2 + 84x - 10}{5 - 4x} e^{-t} + \frac{22x^8 - 60x^7 + 53x^6 - 15x^5}{10} e^{-2t}, \\
f_3(x,t) &= -\frac{2x^4}{5} e^{-t}.
\end{aligned}
\]

The exact solution to (3.4.1) is

\[
\begin{aligned}
c_1(x,t) &= x^2(1-x) e^{-t}, \quad c_2(x,t) = x^2(1-x)^2 e^{-t}, \quad \text{and } \psi(x,t) = -\frac{x^5(3 - 2x)}{60} e^{-t}.
\end{aligned}
\]

We use the time step \( \tau = h \) to compute numerical solutions with and without activating the limiter. The errors and orders at \( t = 1 \) are listed in Table 3.1 (with the limiter) and Table 3.2 (without the limiter). The observed second order accuracy is for both time and space since \( \tau = h \) has been taken here.

We see from Table 3.1 and Table 3.2 that the limiter does not destroy the second order accuracy.
Table 3.1 Accuracy with the limiter

<table>
<thead>
<tr>
<th>N</th>
<th>$L^1$ error of $c_1$ order</th>
<th>$L^1$ error of $c_2$ order</th>
<th>$L^1$ error of $\psi$ order</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>4.1646E-05 -</td>
<td>2.7583E-05 -</td>
<td>4.0670E-06 -</td>
</tr>
<tr>
<td>80</td>
<td>1.0286E-05 2.0175</td>
<td>6.8252E-06 2.0148</td>
<td>9.9109E-07 2.0369</td>
</tr>
<tr>
<td>160</td>
<td>2.5895E-06 1.9899</td>
<td>4.3208E-07 1.9907</td>
<td>6.2121E-08 1.9956</td>
</tr>
<tr>
<td>320</td>
<td>6.5156E-07 1.9907</td>
<td>4.3208E-07 1.9907</td>
<td>6.2121E-08 1.9956</td>
</tr>
</tbody>
</table>

Table 3.2 Accuracy without the limiter

<table>
<thead>
<tr>
<th>N</th>
<th>$L^1$ error of $c_1$ order</th>
<th>$L^1$ error of $c_2$ order</th>
<th>$L^1$ error of $\psi$ order</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>4.1630E-05 -</td>
<td>2.7602E-05 -</td>
<td>4.0681E-06 -</td>
</tr>
<tr>
<td>80</td>
<td>1.0288E-05 2.0166</td>
<td>6.8315E-06 2.0145</td>
<td>9.9115E-07 2.0372</td>
</tr>
<tr>
<td>160</td>
<td>2.5897E-06 1.9901</td>
<td>1.7179E-06 1.9916</td>
<td>2.4773E-07 2.0003</td>
</tr>
<tr>
<td>320</td>
<td>6.5155E-07 1.9908</td>
<td>4.3208E-07 1.9913</td>
<td>6.2121E-08 1.9956</td>
</tr>
</tbody>
</table>

Example 3.4.2. (Positivity) We consider the system

\[
\begin{align*}
\partial_t c_1 &= \frac{1}{A(x)} \partial_x (A(x) D_1(x)(\partial_x c_1 + c_1 \partial_x \psi)), \quad x \in [-10, 10], \quad t > 0, \\
\partial_t c_2 &= \frac{1}{A(x)} \partial_x (A(x) D_2(x)(\partial_x c_2 - 2c_2 \partial_x \psi)), \quad x \in [-10, 10], \quad t > 0, \\
\partial_t c_3 &= \frac{1}{A(x)} \partial_x (A(x) D_3(x)(\partial_x c_3 + c_3 \partial_x \psi)), \quad x \in [-10, 10], \quad t > 0, \\
-\frac{1}{A(x)} \partial_x (\epsilon A(x) \partial_x \psi) &= c_1 - 2c_2 + c_3 - \rho_0(x), \quad x \in [-10, 10], \quad t > 0,
\end{align*}
\]

(3.4.2)

subject to boundary conditions

\[
c_i(\pm 10, t) = 0, \quad \psi(\pm 10, t) = 0, \quad t > 0,
\]

(3.4.3)

and the initial data

\[
\begin{align*}
c_1^{in}(x) &= 0.5 - 0.25e^{-(x+4)^4}, \\
c_2^{in}(x) &= 0.5 + 0.5e^{-x^4}, \\
c_3^{in}(x) &= 0.5 - 0.5e^{-(x-3)^4}.
\end{align*}
\]

(3.4.4)

As in [5], we choose $\epsilon = 0.05$, $A(x) = 1 + 0.5x^2$, $D_1(x) = 20(1 - 0.9e^{-x^4})$ and $\rho_0(x) = 20e^{-x^4}$. This corresponds to problem (3.1.1) with $z_1 = z_3 = 1$, $z_2 = -2$, with $c_{i,l} = c_{i,r} = 0.5$ and $V = 0$. In this numerical test we take $h = 0.05$, $\tau = 10^{-4}$. The solutions are understood to have reached steady states if $||\psi^n - \psi^{n-1}||_{\infty} \leq 10^{-7}$. 

Table 3.3 shows times $t_s$ needed for reaching each steady state, number of time steps, and CPU times. The second order scheme both with and without limiter appears faster than the first order scheme, and the positivity preserving limiter looks quite efficient.

Table 3.3  Times needed for reaching steady states on Example 3.4.2

| Scheme                        | $||\psi^n - \psi^{n-1}||_\infty$ | time $t_s$ | $n = t_s/\tau$ | CPU time (sec) |
|-------------------------------|-----------------------------------|------------|----------------|---------------|
| first order scheme            | 9.9998E-08                        | 3.6513     | 36513          | 11.7332       |
| second order scheme without limiter | 9.9990E-08                        | 2.0134     | 20134          | 6.4808        |
| second order scheme with limiter | 9.9967E-08                        | 2.0135     | 20135          | 6.6138        |

We compute up to $t = 3$ to get steady state solutions. From Figure 3.1 we see that both second order schemes with limiter (middle) and without limiter (right) produce the same steady state solutions. As shown in Figure 3.2, the scheme without limiter produces negative solutions (solid lines) at some time steps while the scheme with the limiter always produces nonnegative solutions (dots).

Figure 3.1  Example 3.4.2: Initial data (left) and steady state solutions obtained with limiter (middle) and without limiter (right)

3.5 Concluding remarks

In this paper, we have extended the unconditional positive scheme introduced in [6] to a novel scheme which is second order in both space and time, for solving the reduced PNP system (3.1.1). The scheme is based on an implicit-explicit time discretization combined with a prediction-correction technique. The obtained solution is shown nonnegative for suitably small time steps;
Figure 3.2  Example 3.4.2: minimum of densities $c^n_1$ (left), $c^n_2$ (middle) and $c^n_3$ (right) obtained with limiter (dots) and without limiter (solid)

for large time steps, positivity of the numerical solution is enforced by a positivity preserving limiter. This local limiter has been shown to preserve both local mass and solution non-negativity. In addition, we observe from our numerical tests that the limiter does not destroy the scheme accuracy.

Acknowledgments

This research was supported by the National Science Foundation under Grant DMS1812666.

References


CHAPTER 4. EFFICIENT, POSITIVE, AND ENERGY STABLE SCHEMES FOR MULTI-D POISSON-NERNST-PLANCK SYSTEMS

A paper submitted to Journal of Scientific Computing
Hailiang Liu and Wumaier Maimaitiyiming

Abstract

In this paper, we design, analyze, and numerically validate positive and energy-dissipating schemes for solving the time-dependent multi-dimensional system of Poisson-Nernst-Planck (PNP) equations, which has found much use in the modeling of biological membrane channels and semiconductor devices. The semi-implicit time discretization based on a reformulation of the system gives a well-posed elliptic system, which is shown to preserve solution positivity for arbitrary time steps. The first order (in time) fully-discrete scheme is shown to preserve solution positivity and mass conservation unconditionally, and energy dissipation with only a mild $O(1)$ time step restriction. The scheme is also shown to preserve the steady-states. For the fully second order (in both time and space) scheme with large time steps, solution positivity is restored by a local scaling limiter, which is shown to maintain the spatial accuracy. These schemes are easy to implement. Several three-dimensional numerical examples verify our theoretical findings and demonstrate the accuracy, efficiency, and robustness of the proposed schemes, as well as the fast approach to steady states.

4.1 Introduction

In this paper, we are concerned with efficient and structure-preserving numerical approximations to a multi-dimensional time-dependent system of Poisson-Nernst-Planck (PNP) equations. Such system has been widely used to describe charge transport in diverse applications such as biological
membrane channels [4, 7, 43], electrochemical systems [1, 33], and semiconductor devices [30, 38]. In the semiconductor modeling, it is often called the Poisson-drift-diffusion system.

PNP equations consist of Nernst–Planck (NP) equations that describe the drift and diffusion of ion species, and the Poisson equation that describes the electrostatic interaction. Such mean field approximation of diffusive ions admits several variants, including the following form which we will consider,

\[
\partial_t \rho_i + \nabla \cdot J_i = 0, \quad x \in \Omega \subset \mathbb{R}^d, \quad t > 0,
\]

\[
- J_i = D_i(x) \left[ \nabla \rho_i + \frac{1}{k_B T} \rho_i q_i \nabla \phi + \nabla \mu_i \right],
\]

\[
- \nabla \cdot (\epsilon(x) \nabla \phi) = 4\pi \left( f(x) + \sum_{i=1}^m q_i \rho_i \right),
\]

subject to initial data \( \rho_i(x, 0) = \rho_i^0(x) \geq 0 \) \((i = 1, \cdots, m)\) and appropriate boundary conditions to be specified in section 4.2.1. Here \( m \) is the number of species, \( \rho_i = \rho_i(x, t) \) is the charge carrier density for the \( i \)-th species, and \( \phi = \phi(x, t) \) is the electrostatic potential. The charge carrier flux is \( J_i \), with which \( D_i(x) \) is the diffusion coefficient, \( k_B \) is the Boltzmann constant, and \( T \) is the absolute temperature. The coupling parameter is \( q_i = z_i e \), where \( z_i \) is the valence (with sign), \( e \) is the unit charge. In the Poisson equation, \( \epsilon(x) \) is the permittivity, \( f(x) \) is the permanent (fixed) charge density of the system. The equations are valid in a bounded domain \( \Omega \) with boundary \( \partial \Omega \) and for time \( t \geq 0 \). For more accurate modeling of collective interactions of charged particles, the chemical potential \( \mu_i \) is often included and can be modeled by other means (see section 4.2.3 for more details).

Due to the wide variety of devices modeled by the PNP equations, computer simulation for this system of differential equations is of great interest. However, the PNP system is a strongly coupled system of nonlinear equations, also, the PNP system as a gradient flow can take very long time evolution to reach steady states. Hence, designing efficient and stable methods with comprehensive numerical analysis for the PNP system is highly desirable. This is what we plan to do in this work.
4.1.1 Related work

In the literature, there are different numerical solvers available for solving both steady and time-dependent PNP problems; see, e.g., [11, 13, 17, 29, 31, 40, 41, 47]. Many existing algorithms were introduced to handle specific issues in complex applications, in which one may encounter different numerical obstacles, such as discontinuous coefficients, singular charges, geometric singularities, and nonlinear couplings to accommodate various phenomena exhibited by biological ion channels. We refer the interested reader to [44] for some variational multiscale models on charge transport and related algorithms.

Solutions to the PNP equations are known to satisfy some important physical properties. It is desirable to maintain these properties at the discrete level, preferably without or with only mild constraints on time step relative to spatial meshes. Under natural boundary conditions, three main properties for the PNP equations are known as (i) conservation of mass, (ii) density positivity, and (iii) free energy dissipation law. The first property requires the scheme to be conservative. The second property is point-wise and also important for the third property. In general, it is rather challenging to obtain both unconditional positivity and discrete energy decay simultaneously. This is evidenced by several recent efforts [8, 9, 10, 14, 24, 25, 32], in which these properties have been partially addressed at the discrete level for PNP equations. With explicit time discretization, the finite difference scheme in [24] preserves solution positivity under a CFL condition $\Delta t = O(\Delta x^2)$ and the energy decay was shown for the semi-discrete scheme (time is continuous). An arbitrary high order DG scheme in [25] was shown to dissipate the free energy, with solution positivity restored with the aid of a scaling limiter. With implicit time discretization, the second order finite difference scheme in [9] preserves positivity under a CFL condition $\Delta t = O(\Delta x^2)$ and a constraint on spatial meshes. An energy-preserving version was further given in [8] with a proven second order energy decay rate. The finite element method in [32] employs the fully implicit backward Euler scheme to obtain solution positivity and the discrete energy decay. In some cases, electric energy alone can be shown to decay (see [25]). Such decay has been verified for the finite difference scheme in [14] and the finite element scheme in [10], both with semi-implicit time discretization.
More recent attempts have focused on semi-implicit schemes based on a formulation of the nonlogarithmic Landau type. As a result, all schemes obtained in \([6, 15, 16, 21, 23]\) have been shown to feature unconditional positivity (see further discussion in section 4.1.2).

Our goal here is to construct and analyze structure-preserving numerical schemes for PNP equations in a more general setting: multi-dimension, multi-species, also subject to other chemical forces.

### 4.1.2 Our contributions

A key step is to reformulate \((4.1.1a)-(4.1.1b)\) as

\[
\partial_t \rho_i = \nabla \cdot (D_i(x)e^{-\psi_i} \nabla (\rho_i e^{\psi_i})),
\]

with

\[
\psi_i(x, t) = \frac{q_i}{k_B T} \phi(x, t) + \frac{1}{k_B T} \mu_i.
\]

Such reformulation, called the Slotboom transformation in the semiconductor literature, converts a drift-diffusion operator into a self-adjoint elliptic operator. It can be more efficiently solved, and in particular more suitable for keeping the positivity-preserving property. In the context of Fokker-Planck equations it is termed as the nonlogarithmic Landau formulation (see, e.g., \([2, 26]\)).

Using such reformulation in \([26]\) Liu and Yu constructed an implicit scheme for a singular Fokker-Planck equation and proved that all three solution properties hold for arbitrary time steps, for which implicit time-discretization is essential. Inspired by \([24, 26]\), we adopted a semi-implicit discretization of \((4.1.2)\) in \([23]\) to construct a first order in time and second order in space scheme for a reduced PNP system, and proved all three solution properties for the resulting scheme with only a mild \(O(1)\) time step restriction. We further introduced a second order (in time) extension in \([21]\) again for the reduced PNP system, and a fully second order scheme in \([22]\) for a class of nonlinear nonlocal Fokker-Planck type equations. All schemes in \([21, 22, 23]\) feature unconditional positivity and a conditional discrete energy dissipation law simultaneously.
This paper improves upon the existing results in [21, 22, 23] in the study of (4.1.1). We first present a semi-implicit time discretization of form

$$\frac{\rho_i^{n+1} - \rho_i^n}{\tau} = \nabla \cdot (D_i(x)e^{-\psi^n_i} \nabla (e^{\psi^n_i} \rho_i^{n+1})) =: R[\rho_i^{n+1}, \psi^n_i], \quad (4.1.3)$$

which is shown to be well-posed and positivity-preserving for time steps of arbitrary size and independent of the Poisson solver. We further construct the following second order scheme

$$\frac{\rho_i^* - \rho_i^n}{\tau/2} = R[\rho_i^*, \frac{3}{2} \psi_i^n - \frac{1}{2} \psi_i^{n-1}], \quad \rho_i^{n+1} = 2\rho_i^* - \rho_i^n, \quad (4.1.4)$$

for which solution positivity for large time steps is restored by a positivity-preserving local limiter. For the spatial discretization we use the 2nd order central difference approximation.

Before stating the main results, let us mention some viable options in the use of reformulation (4.1.2), i.e.,

$$\partial_t \rho_i = R[\rho_i, \psi_i],$$

which is linear in \( \rho_i \) if \( \psi_i \) is a priori given. With the second order central difference in spatial discretization, there are several ways to define \( \psi_i \) on cell interfaces (see section 4.3.3). For the time discretization, solution positivity is readily available if we take

$$\frac{\rho_i^{n+1} - \rho_i^{n-k+1}}{k\tau} = R[\rho_i^{n+1}, \psi_i^*], \quad (4.1.5)$$

with a consistent choice for \( \psi_i^* \) and integer \( k \geq 1 \). Different options are introduced in [6, 15, 16] for obtaining their respective positive schemes.

It is natural and simple to take \( k = 1 \) and \( \psi^* = \psi^n \) in (4.1.5), that is (4.1.3) (again with further central difference in space). But it is subtle to establish a discrete energy dissipation law. A fully discrete scheme using (4.1.3) was studied in [6], where no energy dissipation law was established. Nonetheless, a discrete energy dissipation law can be verified with other options. Indeed, (4.1.5) with \( k = 2 \) and \( \psi^*_i = \psi_i^n \) was considered in [15], where the authors proved unconditional energy decay for a modified energy. In [16], (4.1.5) with \( k = 1 \) and \( \psi^*_i = (\psi_i^{n+1} + \psi_i^n)/2 \) was considered, and all three properties are shown to hold simultaneously even for general boundary conditions for the Poisson equation. Obviously these options can bring further computational overheads.
In this work, we formulate simple finite volume schemes for (4.1.1) by integrating the central
difference method for spatial discretization with the semi-implicit time discretization of the refor-
mulation (4.1.2). We have strived to advance these numerical schemes by presenting a series of
theoretical results. We summarize the main contributions as follows:

- We show that the first order time discretization results a well-posed elliptic system (4.1.3)
at each time step, and features solution positivity independent of the time steps (Theorem
4.3.1). Upper bound of numerical solutions for some cases is established as well (Theorem
4.3.2).

- For the first order (in time) fully-discrete scheme, beyond the unconditional solution positivity
(Theorem 4.3.3), we further establish a discrete energy dissipation law for time steps of size
$O(1/M)$, where $M$ is the upper-bound of the numerical solutions (Theorem 4.3.4). This result
sharpens the previous estimates in [23] for the reduced PNP system. We also prove that the
scheme preserves steady-states, and numerical solutions converge to a steady state as $n \to \infty$
(Theorem 4.3.5).

- We design a fully second order (both in time and space) scheme, and solution positivity is
shown for small time steps (Theorem 4.4.1). While solution positivity for large time steps is
ensured by using a local limiter. We prove that such limiter does not destroy the 2nd order
spatial accuracy (Theorem 4.4.2).

- Three-dimensional numerical tests are conducted to evaluate the scheme performance and ver-
ify our theoretical findings. The computational cost of the second order scheme is comparable
to that of the first order semi-implicit schemes (see section 4.5).

4.1.3 Organization.

We organize this paper as follows: In section 2, we present primary problem settings and
solution properties, as well as model variations. In section 3, we formulate a unified finite volume
method for the PNP system subject to mixed boundary conditions and establish solution positivity,
energy dissipation, mass conservation, and steady-state preserving properties for the case of natural boundary conditions. Extension to a second order scheme is given in section 4. In section 5, we numerically verify good performance of the schemes. Finally in section 6 some concluding remarks are given.

Throughout this paper, we denote \( \rho \) as vector \((\rho_1, \cdots, \rho_m)\), \( \partial \Omega \) as the boundary of domain \( \Omega \) includes both the Dirichlet boundary \( \partial \Omega_D \) and the Neumann boundary \( \partial \Omega_N \). \(|K|\) denotes the volume of domain \( K \). We use \( g_\alpha \) to denote \( g_\alpha = 1/|K_\alpha| \int_{K_\alpha} g(x) dx \), for an integral average of function \( g(x) \) over a cell \( K_\alpha \).

### 4.2 Models and related work

#### 4.2.1 Boundary conditions

Boundary conditions are a critical component of the PNP model and determine important qualitative behavior of the solution. Here we consider the simplest form of boundary conditions of Dirichlet and/or Neumann type \([3]\).

Let \( \Omega \) be a bounded domain with Lipschitz boundary \( \partial \Omega \). The external electrostatic potential \( \phi \) is influenced by applied potential, which can be modeled by prescribing a Dirichlet boundary condition

\[
\phi(x, t) = \phi^b(x, t), \quad x \in \partial \Omega_D. \tag{4.2.1}
\]

For the remaining part of the boundary \( \partial \Omega_N = \partial \Omega \setminus \partial \Omega_D \), a no-flux boundary condition is applied:

\[
\epsilon(x) \nabla \phi \cdot \mathbf{n} = 0, \quad x \in \partial \Omega_N. \tag{4.2.2}
\]

This boundary condition models surface charges, where \( \mathbf{n} \) is the outward unit normal vector on the boundary \( \partial \Omega_N \). Same types of boundary conditions are imposed for \( \rho_i \) as

\[
\rho_i(x, t) = \rho_i^b(x, t) \geq 0, \quad x \in \partial \Omega_D, \tag{4.2.3}
\]

\[
J_i \cdot \mathbf{n} = 0, \quad x \in \partial \Omega_N. \tag{4.2.4}
\]
In this work we present our schemes by restricting to a rectangular computational domain \( \Omega = (0, L_1) \times \cdots \times (0, L_d) \), with \( \partial \Omega_D = \{ x \in \bar{\Omega} \mid x_1 = 0, x_1 = L_1 \} \).

We remark that the boundary conditions for the electrostatic potential are not unique and greatly depend on the problem under investigation. For example, one may use a non-homogeneous Neumann boundary condition \((\nabla \phi \cdot n = \sigma \) is used in [25]) or Robin boundary conditions [9, 16]. The existence and uniqueness of the solution for the nonlinear PNP boundary value problems have been studied in [19, 27, 34] for the 1D case and in [3, 18] for multi-dimensions.

### 4.2.2 Positivity and energy dissipation law

One important solution property is

\[
\rho_i(x, t) \geq 0, \quad x \in \Omega, \; t > 0. \tag{4.2.5}
\]

Integration of each density equation gives

\[
\frac{d}{dt} \int_{\Omega} \rho_i(x, t) \, dx = \int_{\partial \Omega} J_i \cdot n \, ds,
\]

which with zero flux \( J_i \cdot n = 0 \) on the whole boundary leads to the mass conservation:

\[
\int_{\Omega} \rho_i(x, t) \, dx = \int_{\Omega} \rho_i^n(x) \, dx, \quad t > 0, \quad i = 1, \cdots, m. \tag{4.2.6}
\]

We consider the free energy functional \( E \) associated to (4.1.1) with \( \mu_i = \mu_i(x) \):

\[
E = \int_{\Omega} \left( \sum_{i=1}^{m} \rho_i (\log \rho_i - 1) + \frac{1}{2k_BT} (f + \sum_{i=1}^{m} q_i \rho_i) \phi + \frac{1}{k_BT} \sum_{i=1}^{m} \rho_i \mu_i \right) \, dx. \tag{4.2.7}
\]

In virtue of the Poisson equation (4.1.1c), the free energy may be written as

\[
E = \int_{\Omega} \left( \sum_{i=1}^{m} \rho_i (\log \rho_i - 1) + \frac{\epsilon}{8\pi k_BT} |\nabla \phi|^2 + \frac{1}{k_BT} \sum_{i=1}^{m} \rho_i \mu_i \right) \, dx.
\]

Note that the unscaled free energy \( F = k_B T E \) is also often used, see [28]. A formal calculation gives

\[
\frac{dE}{dt} = - \int_{\Omega} \sum_{i=1}^{m} D_i(x) \rho_i |\nabla \psi_i^*|^2 \, dx + \int_{\partial \Omega} \sum_{i=1}^{m} \psi_i^2 J_i \cdot n \, ds
\]

\[
+ \frac{1}{8\pi k_BT} \int_{\partial \Omega} \epsilon(x) [\phi(\partial_n \phi)_t - \partial_n \phi \phi_t] \, ds,
\]
where

$$\psi_i^*: = \log \rho_i + \frac{q_i}{k_B T} \phi + \frac{1}{k_B T} \mu_i.$$ 

Clearly, with $\partial \Omega^D = \emptyset$, we have the following energy dissipation law:

$$\frac{dE}{dt} = - \int_\Omega \sum_{i=1}^m D_i(x) \rho_i |\nabla \psi_i^*|^2 dx \leq 0.$$ (4.2.8)

Otherwise, the Dirichlet boundary condition needs to be carefully handled (see, e.g., [28]).

For time dependent chemical potentials $\mu_i(x,t)$, the total free energy and its dissipation law needs to be modified depending on how the chemical potential is determined.

### 4.2.3 Chemical potential

In application, the chemical potential $\mu_i$ often includes the ideal chemical potential $\mu_i^{id}(x,t)$ and the excess chemical potential $\mu_i^{ex}(x,t)$ of the charged particles:

$$\mu_i(x,t) = \mu_i^{id}(x,t) + \mu_i^{ex}(x,t),$$

with

$$\mu_i^{id}(x,t) = - \log[\gamma_i \rho_i(x,t)/\rho_i^{bulk}],$$

where the activity coefficient $\gamma_i$ described by the extended Debye-Hückel theory depends on $\rho$ in nonlinear manner. Meanwhile,

$$\mu_i^{ex}(x,t) = \frac{\delta F^{ex}(\rho(x,t))}{\delta \rho_i(x,t)}$$

is the $L^2$ variational derivative of the excess chemical functional $F^{ex}$, which may include hard-sphere components, short-range interactions, Coulomb interactions and electrostatic correlations, where the expression of each component can be found in [31, 42].

We remark that the steric interactions between ions of different species are important in the modeling of ion channels [17, 20]. Such effects can be described by choosing

$$F^{ex} = \frac{1}{2} \int_\Omega \omega_{ij} \rho_i \rho_j,$$
where \( \omega_{ij} \) are the second-order virial coefficients for hard spheres, depending on the size of \( i \)-th and \( j \)-th ion species [49]. With this addition alone, the flux becomes

\[
-J_i = D_i(x) \left( \nabla \rho_i + \frac{1}{k_B T} q_i \rho_i \nabla \phi + \rho_i \sum_{j=1}^{m} \omega_{ij} \nabla \rho_j \right).
\]

The PNP system with this modified flux has been studied numerically first in [39] without cross steric interactions, and then in [6] with cross interactions.

Our schemes will be constructed so that numerical solutions are updated in an explicit-implicit manner while \( \mu_i \) needs only to be evaluated off-line. For simplicity, we shall present our schemes assuming \( \mu_i \) is given while keeping in mind that it can be applied to complex chemical potentials without difficulty.

### 4.2.4 Steady states

By the free energy dissipation law (4.2.8), the solution to (4.1.1) with zero-flux boundary conditions is expected to converge to the steady-states as time becomes large. In such case the steady states formally satisfy (4.1.1) with \( \partial_t \rho_i = 0 \); i.e.,

\[
\nabla \cdot (D_i(x) \rho_i \nabla \psi_i^*) = \nabla \cdot J_i = 0, \quad J_i \cdot n = 0, \quad x \in \partial \Omega.
\]

This yields \( \int_{\Omega} J_i \cdot \nabla \psi_i^* dx = 0 \), which ensures that \( \psi_i^* \) must be a constant. This gives the well-known Boltzmann distribution

\[
\rho_i = c_i e^{-\frac{1}{k_B T} (q_i \phi + \mu_i)},
\]

(4.2.9)

here \( c_i \) is any constant. Such constant can be uniquely determined by the initial data in the PNP system (4.1.1) if such steady-state is approached by the solution at large times. Indeed, mass conservation simply gives

\[
c_i = \frac{\int_{\Omega} \rho_i^{in} dx}{\int_{\Omega} e^{-\frac{1}{k_B T} (q_i \phi + \mu_i)} dx}.
\]

(4.2.10)

This allows us to obtain a closed Poisson-Boltzmann equation (PBE) of form

\[
-\nabla \cdot (\epsilon(x) \nabla \phi) = 4\pi \left( f(x) + \sum_{i=1}^{m} q_i c_i e^{-\frac{1}{k_B T} (q_i \phi + \mu_i)} \right), \quad \partial_n \phi|_{\partial \Omega} = 0.
\]

(4.2.11)
We should point out that the numerical method presented in this paper may be used as an iterative algorithm to numerically compute the nonlocal PBE (4.2.11); hence it serves as a simpler alternative to the iterative DG methods recently developed in [45, 46].

In practical applications, one may describe ions of less interest using the Boltzmann distribution and still solve the NP equations for the target ions so to reduce the computational cost, see [48] for further details on related models. Our numerical method thus provides an alternative path to simulate such models.

4.3 Numerical method

In this section we will construct positive and energy stable schemes.

4.3.1 Reformulation

By setting
\[ \psi_i(x, t) = \frac{1}{k_B T} (q_i \phi(x, t) + \mu_i), \]
we reformulate the density equation (4.1.1a)-(4.1.1b) as:
\[ \partial_t \rho_i = \nabla \cdot (D_i(x) e^{-\psi_i} \nabla (e^{\psi_i} \rho_i)). \] (4.3.1)

In spite of the aforementioned advantages of such reformulation, possible large variation of the transformed diffusion coefficients could result in large condition number of the stiffness matrix [29]. This issue has been recently investigated in [6, 36].

4.3.2 Time discretization

Let \( \tau > 0 \) be the time step, and \( t_n = \tau n, n = 0, 1 \cdots \), be the corresponding temporal grids. We initialize by taking \( \rho^0(x) = \rho^{in}(x) \), and obtaining \( \phi^0 \) by solving the Poisson equation (4.1.1c) using \( \rho^0(x) \).
Let $\rho^n$ and $\phi^n$ be numerical approximations of $\rho(x, t_n)$ and $\phi(x, t_n)$, respectively, we first obtain $\rho^{n+1}$ by solving the following elliptic system:

\[
\frac{\rho_i^{n+1} - \rho_i^n}{\tau} = \nabla \cdot (D_i(x)e^{-\psi_i^n} \nabla (e^{\psi_i^n} \rho_i^{n+1})) =: R[\rho_i^{n+1}, \psi_i^n],
\]

(4.3.2a)

\[
\rho_i^{n+1} = \rho_i^b(x, t_{n+1}), \quad x \in \partial \Omega_D,
\]

(4.3.2b)

\[
\nabla (e^{\psi_i^n} \rho_i^{n+1}) \cdot n = 0, \quad x \in \partial \Omega_N,
\]

(4.3.2c)

here

\[
\psi_i^n = \frac{1}{k_B T}(q_i \phi^n + \mu_i).
\]

Using this obtained $\rho^{n+1}$, we update to obtain $\phi^{n+1}$ from solving

\[
-\nabla \cdot (\epsilon(x) \nabla \phi^{n+1}) = 4\pi \left( f(x) + \sum_{i=1}^m q_i \rho_i^{n+1} \right),
\]

(4.3.3a)

\[
\phi^{n+1}(x) = \phi^b(x, t_{n+1}), \quad x \in \partial \Omega_D,
\]

(4.3.3b)

\[
\nabla \phi^{n+1} \cdot n = 0, \quad x \in \partial \Omega_N.
\]

(4.3.3c)

This scheme is well-defined for any $\tau > 0$ with $\rho^n \geq 0$ for all $n \in \mathbb{N}$. More precisely, we have

**Theorem 4.3.1.** Assume $D_i(x) \geq D_0 > 0$ and $\epsilon(x) \geq \epsilon_0 > 0$, and $\mu_i(x) \in C(\bar{\Omega})$. Then for given $(\rho^n, \phi^n) \in C(\bar{\Omega}) \cap C^2(\Omega)$, there exists a unique solution $(\rho^{n+1}, \phi^{n+1}) \in C(\bar{\Omega}) \cap C^2(\Omega)$. If $\rho^n \geq 0$ and $\rho^b(x, t) \geq 0$, $x \in \partial \Omega_D$, then $\rho^{n+1} \geq 0$ for $x \in \Omega$.

The proof is deferred to the appendix A.

In some cases density for the PNP problem is known to be uniformly bounded for all time. We shall show this bound property also for the semi-discrete scheme (4.3.2).

**Theorem 4.3.2.** Let $0 \leq \rho_i^{m}(x) \leq B_i$, $0 \leq \rho_i^b(x, t) \leq B_i^b$, $D_i(x)/\epsilon(x) = \sigma_i$ be constants, $\Omega$ be $C^1$ convex domain, all $q_i$ have the same sign, and $\mu_i$ is smooth with $(\nabla \mu_i) \cdot n \geq 0$ on $\partial \Omega_N$. If

\[
\tau < \frac{1}{Q_{i, \max}},
\]

then $\rho^n$ obtained by scheme (4.3.2) is uniformly bounded, i.e.,

\[
\rho_i^n(x) \leq \max \left\{ B_i^b, \quad B_i, \quad \frac{Q_{i, \max}}{\gamma_i} \right\},
\]

(4.3.4)
here $Q_{i,\text{max}} = \max_{x \in \Omega} Q_i(x)$ with

$$Q_i(x) = \frac{1}{k_B T} \left[ \nabla \cdot \left( D_i(x) \nabla \mu_i \right) - 4\pi q_i \sigma_i f(x) \right], \quad \gamma_i = \frac{4\pi q_i^2 \sigma_i}{k_B T}.$$

**Remark 4.3.1.** In the case of $q_i$ with different sign, density $\rho_i$ in (4.1.1) may not be bounded.

**Proof.** We rewrite the semi-discrete scheme

$$\frac{\rho_i^{n+1} - \rho_i^n}{\tau} = \nabla \cdot \left( D_i(x) e^{-\psi_i^n} \nabla \left( \rho_i^{n+1} e^{\psi_i^n} \right) \right)$$

into

$$\frac{\rho_i^{n+1} - \rho_i^n}{\tau} = D_i(x) \Delta \rho_i^{n+1} + b_i \cdot \nabla \rho_i^{n+1} + c_i \rho_i^{n+1},$$

with

$$b_i = (\nabla D_i(x) + D_i(x) \nabla \psi_i^n), \quad c_i = \nabla \cdot (D_i(x) \nabla \psi_i^n).$$

In virtue of $\psi_i^n = \frac{q_i}{k_B T} \phi^n + \frac{1}{k_B T} \mu_i$ and $D_i(x)/\epsilon(x) = \sigma_i$, the coefficient $c_i$ can be estimated as

$$c_i = \frac{1}{k_B T} \left[ \nabla \cdot \left( q_i D_i(x) \nabla \phi^n + \nabla \cdot \left( D_i(x) \nabla \mu_i \right) \right) \right]$$

$$= \frac{1}{k_B T} \left[ q_i \sigma_i \nabla \cdot (\epsilon(x) \nabla \phi^n) + \nabla \cdot (D_i(x) \nabla \mu_i) \right]$$

(using (4.3.3a))

$$= \frac{1}{k_B T} \left[ -4\pi q_i \sigma_i \left( f(x) + \sum_{j=1}^m q_j \rho_j^n \right) + \nabla \cdot (D_i(x) \nabla \mu_i) \right]$$

(using $q_i q_j > 0$ and $\rho_j^2 \geq 0$)

$$\leq \frac{1}{k_B T} \left[ \nabla \cdot (D_i(x) \nabla \mu_i) - 4\pi q_i \sigma_i f(x) \right] - 4\pi q_i^2 \sigma_i \rho_i^n \right]$$

$$= Q_i(x) - \gamma_i \rho_i^n.$$

Hence

$$\frac{\rho_i^{n+1} - \rho_i^n}{\tau} \leq D_i(x) \Delta \rho_i^{n+1} + b_i \cdot \nabla \rho_i^{n+1} + \rho_i^{n+1} \left( Q_{i,\text{max}} - \gamma_i \rho_i^n \right). \quad (4.3.5)$$

We proceed to distinct three cases, by letting $x^* = \arg\max_{x \in \Omega} \rho_i^{n+1}(x)$:

(i) If $x^* \in \partial \Omega_D$ we have

$$\rho_i^{n+1}(x^*) = \rho_i^b(x^*, t_{n+1}) \leq B_i^b.$$
(ii) If \( x^* \in \Omega \), then (4.3.5) can be reduced to
\[
\frac{\rho_i^{n+1}(x^*) - \rho_i^n(x^*)}{\tau} \leq \rho_i^{n+1}(x^*) (Q_{i,max} - \gamma_i \rho_i^n(x^*)).
\]
This using notation \( \rho_{i,max} = \max_{x \in \bar{\Omega}} \rho_i^n \) yields
\[
\rho_i^{n+1}(x) \leq \rho_i^{n+1}(x^*) \leq \frac{\rho_{i,max}^n}{1 - \tau Q_{i,max} + \tau \gamma_i \rho_i^n} =: P(\rho_{i,max}),
\]
(4.3.6)
here we used the fact that \( P(\cdot) : \mathbb{R}^+ \to \mathbb{R}^+ \) is non-decreasing.

(iii) If \( x^* \in \partial \Omega_N \), we must have \( \rho_i^{n+1}(x^*) \leq P(\rho_{i,max}) \). Otherwise assume \( \rho_i^{n+1}(x^*) > P(\rho_{i,max}) \).
Set
\[
U(x) = \rho_i^{n+1}(x) - \rho_i^{n+1}(x^*),
\]
and introduce the differential operator
\[
L_\xi := \tau D_i(x) \Delta \xi + \tau b_i \cdot \nabla \xi - (1 - \tau Q_{i,max} + \tau \gamma_i \rho_i^n) \xi.
\]
From (4.3.5) we have
\[
L \rho_i^{n+1} \geq -\rho_i^n,
\]
and using (4.3.6) we obtain
\[
LU(x) = L \rho_i^{n+1}(x) - L \rho_i^{n+1}(x^*) \geq - \rho_i^n + (1 - \tau Q_{i,max} + \tau \gamma_i \rho_i^n) \rho_i^{n+1}(x^*) \geq - \rho_i^n + (1 - \tau Q_{i,max} + \tau \gamma_i \rho_i^n) P(\rho_{i,max}) \geq 0.
\]
Note that \( U(x) \leq 0 \) on \( \partial \Omega \) and \( U(x^*) = 0 \). Apply the maximum-principle [35, Theorem 8] we have
\[
(\nabla U(x^*)) \cdot n = (\nabla \rho_i^{n+1}(x^*)) \cdot n > 0.
\]
On the other hand, from the no-flux boundary condition (4.3.2c) and using (4.3.3c), we have
\[
0 = \left( \nabla \left( \rho_i^{n+1} e^{\psi_i^n} \right) \right) \cdot n = \left( e^{\psi_i^n} \nabla \rho_i^{n+1} + \frac{1}{k_B T} e^{\psi_i^n} \rho_i^{n+1} (q_i \nabla \phi^n + \nabla \mu_i) \right) \cdot n = e^{\psi_i^n} \left( \nabla \rho_i^{n+1} \cdot n + \frac{1}{k_B T} \nabla \mu_i \cdot n \right) > e^{\psi_i^n} \frac{1}{k_B T} \nabla \mu_i \cdot n \quad x \in \partial \Omega_N.
\]
This is a contradiction to the assumption \((\nabla \mu_i) \cdot n \geq 0\). Hence for \(x \in \Omega \cup \partial \Omega_N \cup \partial \Omega_D = \bar{\Omega}\), we have

\[
\rho_{i,\max}^{n+1} \leq \max \left\{ B_i, \ Pn_i \right\}.
\]

Again by the monotonicity of \(P(\cdot)\), we obtain

\[
\rho_{i,\max}^{n+1} \leq \max \left\{ B_i, \ \max \left\{ \rho_{i,\max}^n, Q_{i,\max}^\gamma \right\} \right\}.
\]

The stated result (4.3.4) thus follows by induction.

A discrete energy dissipation law can be established by precisely quantifying a sufficient bound on the time step. In order to save space, we present a detailed analysis of the energy dissipation property only for the fully discrete scheme in the next section.

### 4.3.3 Spatial discretization

For given positive integers \(N_j\) \((j = 1, \cdots, d)\), let \(h_j = L_j/N_j\) be the mesh size in \(j\)-th direction, \(\alpha \in \mathbb{Z}^d\) be the index vector with \(\alpha(j) \in \{1, \cdots, N_j\}\), and \(e_j \in \mathbb{Z}^d\) be a vector with \(j\)-th entry equal to one and all other entries equal to zero. We partition the domain \(\Omega\) into computational cells

\[
K_\alpha = \left[ (\alpha(1) - 1/2)h_1, (\alpha(1) + 1/2)h_1 \right] \times \cdots \times \left[ (\alpha(d) - 1/2)h_d, (\alpha(d) + 1/2)h_d \right]
\]

with cell size \(|K_\alpha| = \prod_{j=1}^d h_j\) such that \(\bigcup_{\alpha \in A} K_\alpha = \Omega\), here \(A\) denotes the set of all indices \(\alpha\).

#### 4.3.3.1 Density update

A finite volume approximation of (4.3.2a) over each cell \(K_\alpha\) with \(\alpha \in A\) gives

\[
\frac{\rho_{i,\alpha}^{n+1} - \rho_{i,\alpha}^n}{\tau} = \sum_{j=1}^d \frac{C_{i,\alpha + e_j/2} - C_{i,\alpha - e_j/2}}{h_j} =: R_{\alpha}[\rho_{i}^{n+1}, \psi^n_i], \tag{4.3.7a}
\]

here \(\rho_{i,\alpha}^0 := \rho_{i,\alpha}^n\).

Numerical fluxes on interfaces are defined by:

(i) on the interior interfaces,

\[
C_{i,\alpha + e_j/2} = \frac{D_i(x_{\alpha + e_j/2})e^{-\psi^n_{i,\alpha + e_j/2}}}{h_j} \left( \rho_{i,\alpha + e_j}^n e^{\psi^n_{i,\alpha + e_j}} - \rho_{i,\alpha}^{n+1} e^{\psi^n_{i,\alpha}} \right), \quad \text{for } 1 < \alpha(j) < N_j; \tag{4.3.7b}
\]
(ii) on the boundary \( \partial \Omega_D \),
\[
C_{i,\alpha-e_1/2} = \frac{2D_i(x_{\alpha-e_1/2})e^{-\psi_i^\text{h}(x_{\alpha-e_1/2}, t_n)}}{h_1} \left( \frac{\rho_{i,\alpha}^{n+1} - \rho_{i,\alpha}^n(x_{\alpha-e_1/2}, t_n+1)}{e^{-\psi_i^\text{h}(x_{\alpha-e_1/2}, t_n+1)}} \right), \quad \alpha(1) = 1,
\]
\[
C_{i,\alpha+e_1/2} = \frac{2D_i(x_{\alpha+e_1/2})e^{-\psi_i^\text{h}(x_{\alpha+e_1/2}, t_n)}}{h_1} \left( \frac{\rho_{i,\alpha}^{n+1} - \rho_{i,\alpha}^n(x_{\alpha+e_1/2}, t_n+1)}{e^{-\psi_i^\text{h}(x_{\alpha+e_1/2}, t_n+1)}} \right), \quad \alpha(1) = N_1; \tag{4.3.7c}
\]

(iii) on the boundary \( \partial \Omega_N \),
\[
C_{i,\alpha-e_j/2} = 0, \quad \text{for } \alpha(j) = 1,
\]
\[
C_{i,\alpha+e_j/2} = 0, \quad \text{for } \alpha(j) = N_j. \tag{4.3.7d}
\]

In (4.3.7b), \( e^{-\psi_i^n_{\alpha+\epsilon_j/2}} \) needs to be evaluated using numerical solutions \( \phi_i^n \). There are three choices, all are second order approximations:

(i) the harmonic mean
\[
e^{-\psi_i^n_{\alpha+\epsilon_j/2}} = \frac{2e^{-\psi_i^n_{\alpha+\epsilon_j}} - e^{-\psi_i^n_{\alpha}}}{e^{-\psi_i^n_{\alpha+\epsilon_j}} + e^{-\psi_i^n_{\alpha}}}, \tag{4.3.8}
\]

(ii) the geometric mean
\[
e^{-\psi_i^n_{\alpha+\epsilon_j/2}} = \sqrt{e^{-\psi_i^n_{\alpha+\epsilon_j}} - e^{-\psi_i^n_{\alpha}}}, \tag{4.3.9}
\]

(iii) the algebraic mean
\[
e^{-\psi_i^n_{\alpha+\epsilon_j/2}} = \frac{e^{-\psi_i^n_{\alpha+\epsilon_j}} + e^{-\psi_i^n_{\alpha}}}{2}. \tag{4.3.10}
\]

It is reported in [36] that the harmonic mean results in a linear system with better condition number than that of the geometric mean. We use the harmonic mean in our numerical tests.

### 4.3.3.2 Solving Poisson’s equation

In order to complete the scheme, we need to evaluate \( \psi_i^n_{\alpha} \) by
\[
\psi_i^n_{\alpha} = \frac{1}{k_B T}(q_i \phi_i^n + \mu_i),
\]
and \( \phi_i^n \) is determined from \( \rho_i^n \) by using the following discretization of the equation (4.3.3a):
\[
-\sum_{j=1}^{d} \frac{\Phi_i^{n+\epsilon_j/2} - \Phi_i^{n-\epsilon_j/2}}{h_j} = 4\pi \left( f_{\alpha} + \sum_{i=1}^{m} q_i \rho_i^{n} \right), \tag{4.3.11a}
\]
where numerical fluxes on cell interfaces are defined by:

(i) on the interior interfaces,

\[ \Phi^{n}_{a+e_{j}/2} = \epsilon (x_{a+e_{j}/2}) \frac{\phi^{n}_{a+e_{j}/2} - \phi^{n}_{a}}{h_{j}}, \quad \text{for } 1 < \alpha(j) < N_{j}, \]  

(ii) on the boundary \( \partial \Omega_{D} \),

\[ \Phi^{n}_{a-e_{1}/2} = \epsilon (x_{a-e_{1}/2}) \frac{2(\phi^{n}_{a} - \phi^{b}(x_{a-e_{1}/2},t_{n}))}{h_{1}}, \quad \text{for } \alpha(1) = 1, \]

\[ \Phi^{n}_{a+e_{1}/2} = \epsilon (x_{a+e_{1}/2}) \frac{2(\phi^{b}(x_{a+e_{1}/2},t_{n}) - \phi^{n}_{a})}{h_{1}}, \quad \text{for } \alpha(1) = N_{1}, \]

(iii) on the boundary \( \partial \Omega_{N} \),

\[ \Phi^{n}_{a-e_{j}/2} = 0, \quad \text{for } \alpha(j) = 1, \]

\[ \Phi^{n}_{a+e_{j}/2} = 0, \quad \text{for } \alpha(j) = N_{j}. \]

Note that in the case of \( \partial \Omega_{D} = \emptyset \), the solution to (4.3.11) is unique only up to an additive constant, in such case we take \( \phi^{n}_{(1,\ldots,1)} = 0 \) to obtain a unique solution \( \phi^{n}_{a} \).

### 4.3.3.3 Positivity

The following theorem states that the scheme (4.3.7) preserves positivity of numerical solutions without any time step restriction.

**Theorem 4.3.3.** Let \( \rho^{n+1}_{\alpha} \) be obtained from (4.3.7). If \( \rho^{n}_{\alpha} \geq 0 \) for all \( \alpha \in A \), and \( \rho^{b}(x,t_{n}) \geq 0 \), \( x \in \partial \Omega_{D} \), then

\[ \rho^{n+1}_{\alpha} \geq 0 \quad \text{for all } \alpha \in A. \]

**Proof.** This proof mimics that in [26] for the Fokker-Planck equation. Set \( \lambda_{j} = \frac{\tau}{h_{j}} \), \( \bar{D}_{i,\alpha + e_{j}/2} = D_{i}(x_{a+e_{j}/2})e^{-\psi^{n}_{i,\alpha + e_{j}/2}}, \psi^{n}_{i,\alpha} = \psi^{n}_{i,\alpha}, \) and

\[ G_{i,\alpha} = \rho^{n+1}_{i,\alpha} g^{n}_{i,\alpha}, \quad \alpha \in A. \]

Let \( \beta \) be such that

\[ G_{i,\beta} = \min_{\alpha \in A} G_{i,\alpha}, \]
it suffices to prove $G_{i,\beta} \geq 0$. We discuss in cases:

(i) $K_\beta$ is an interior cell. On the cell $K_\beta$ we have

$$g^n_{i,\beta} G_{i,\beta} = \sum_{j=1}^{d} \lambda_j [\bar{D}_{i,\beta+e_j/2}(G_{i,\beta+e_j} - G_{i,\beta}) - \bar{D}_{i,\beta-e_j/2}(G_{i,\beta} - G_{i,\beta-e_j})] + \rho^n_{i,\beta}$$

$$\geq \rho^n_{i,\beta},$$

here we used the fact $G_{i,\beta} \leq G_{i,\beta+e_j}$ and $\bar{D}_{i,\beta+e_j/2} > 0$. Since $g^n_{i,\beta} > 0$, so $G_{i,\beta} \geq 0$.

(ii) $K_\beta$ is a boundary cell ($K_\beta \cap \partial \Omega_D \neq \emptyset$). We only deal with the case $\beta(1) = 1$, remaining cases are similar. In such case,

$$g^n_{i,\beta} G_{i,\beta} = \sum_{j=1}^{d} \lambda_j [\bar{D}_{i,\beta+e_j/2}(G_{i,\beta+e_j} - G_{i,\beta}) - \bar{D}_{i,\beta-e_j/2}(G_{i,\beta} - G_{i,\beta-e_j})]$$

$$+ \lambda_1 \bar{D}_{i,\beta+\epsilon_1/2}(G_{i,\beta+\epsilon_1} - G_{i,\beta})$$

$$- 2\lambda_1 D_i(x_{\beta-\epsilon_1/2})g^b_i(x_{\beta-\epsilon_1/2}, t_n) \left(G_{i,\beta} - \frac{\rho_i^b(x_{\beta-\epsilon_1/2}, t_{n+1})}{g^b_i(x_{\beta-\epsilon_1/2}, t_n)}\right) + \rho^n_{i,\beta}.$$ 

Due to $G_{i,\beta} \leq G_{i,\beta+\epsilon_j}$ and $\bar{D}_{i,\beta+\epsilon_j/2} \geq 0$, we have

$$\left(g^n_{i,\beta} + 2\lambda_1 D_i(x_{\beta-\epsilon_1/2})g^b_i(x_{\beta-\epsilon_1/2}, t_n)\right) G_{i,\beta} \geq 2\lambda_1 D_i(x_{\beta-\epsilon_1/2}) \rho_i^b(x_{\beta-\epsilon_1/2}, t_{n+1}) + \rho^n_{i,\beta} \geq 0,$$

which with $g^n_{i,\beta} + 2\lambda_1 D_i(x_{\beta-\epsilon_1/2})g^b_i(x_{\beta-\epsilon_1/2}, t_n) > 0$ ensures $G_{i,\beta} \geq 0$.

(iii) $K_\beta$ is a boundary cell ($K_\beta \cap \partial \Omega_N \neq \emptyset$). Again we only deal with the case $\beta(l) = 1$. In such case,

$$g^n_{i,\beta} G_{i,\beta} = \sum_{j=1,j \neq l}^{d} \lambda_j [\bar{D}_{i,\beta+e_j/2}(G_{i,\beta+e_j} - G_{i,\beta}) - \bar{D}_{i,\beta-e_j/2}(G_{i,\beta} - G_{i,\beta-e_j})]$$

$$+ \lambda_1 \bar{D}_{i,\beta+1/2\epsilon_1}(G_{i,\beta+\epsilon_1} - G_{i,\beta}) + \rho^n_{i,\beta}$$

$$\geq \rho^n_{i,\beta} \geq 0.$$ 

This also gives $G_{i,\beta} \geq 0$. The proof is thus complete. \qed
4.3.3.4 Energy dissipation

If \( \partial \Omega_D = \emptyset \), then solutions \( \rho_n^{n+1} \) obtained by (4.3.7) are conservative and energy dissipating in addition to the non-negativity. Let a discrete version of the free energy (4.2.7) be defined as

\[
E_h^n = \sum_{\alpha \in A} |K_\alpha| \left[ \sum_{i=1}^m \rho_i^n (\log \rho_i^n - 1) + \frac{1}{2k_BT} \left( f_\alpha + \sum_{i=1}^m q_i \rho_i^n \phi_\alpha^n + \frac{1}{k_BT} \sum_{i=1}^m \rho_i^n \mu_{i,\alpha} \right) \right], \tag{4.3.12}
\]

we have the following result.

Theorem 4.3.4. Let \( \rho^n_\alpha \) be obtained from (4.3.7) by using either (4.3.8), (4.3.9), or (4.3.10) for \( e^{-\psi_{i,\alpha}^n + \epsilon_j/2} \). Let \( \phi^n_\alpha \) be obtained from (4.3.11). If \( \partial \Omega_D = \emptyset \), then we have:

(i) Mass conservation: \( \sum_{\alpha \in A} |K_\alpha| \rho_{i,\alpha}^{n+1} = \sum_{\alpha \in A} |K_\alpha| \rho_{i,\alpha}^n \) for \( n \geq 0, i = 1, \ldots, m \);

(ii) Energy dissipation: There exists \( \tau^* > 0 \) such that if \( \tau \in (0, \tau^*) \), then

\[
E_h^{n+1} - E_h^n \leq -\frac{\tau^n}{2} I^n, \tag{4.3.13}
\]

here

\[
I^n = \sum_{i=1}^m \sum_{j=1}^d \sum_{\alpha(j) \neq N_j} |K_\alpha| \frac{C_i^{\alpha + \epsilon_j/2}}{h_j} \left( \log(\rho_{i,\alpha + \epsilon_j}^{n+1} e^{\psi_{i,\alpha}^n + \epsilon_j}) - \log(\rho_{i,\alpha}^n e^{\psi_{i,\alpha}^n}) \right) \geq 0.
\]

If we let

\[
\epsilon_{\min} = \min_{x \in \Omega} \epsilon(x), \quad \epsilon_{\max} = \max_{x \in \Omega} \epsilon(x), \quad D_{\max} = \max_{i,x \in \Omega} D_i(x),
\]

then \( \tau^* \) can be quantified by

\[
\tau^* = \frac{k_BT \epsilon_{\min}^2}{4\pi \epsilon_{\max} D_{\max} \max_{i,\alpha,n} \rho_i^n \sum_{i=1}^m q_i^2 e^{-\max_{i,j,\alpha,n} |\psi_{i,\alpha}^{n+\epsilon_j} - \psi_{i,\alpha}^n|}}.
\]

Remark 4.3.2. We remark that \( \tau^* \) is of size \( O(1) \), though it appears to be dependent on numerical solutions. For \( h_j \) small, the exponential term is only of size \( e^{O(h)} \), therefore bounded. As \( n \) increases, the solution \( \{\rho^n_\alpha\} \) is expected to converge to the steady-state and therefore bounded from above, hence we simply use the notation \( \max_{i,\alpha,n} \rho_i^n \). The boundedness of \( \rho^n \) in \( n \) for some cases has been established in Theorem 4.3.2 for the corresponding semi-discrete scheme.

The proof is deferred to Appendix B.
4.3.3.5 Preservation of steady-states

With no-flux boundary conditions, scheme (4.3.7) can be shown to be steady-state preserving. Based on the discussion in section 4.2.4, we say a discrete function \( \rho_\alpha \) is at steady-state if

\[
\rho_{i,\alpha} = c_i e^{-\frac{1}{k_BT}(q_i \phi_\alpha + \mu_{i,\alpha})}, \quad i = 1, \cdots, m, \quad \alpha \in A,
\]

(4.3.14)

here \( \phi_\alpha \) satisfies (4.3.11) with \( \rho_{i,\alpha} \) replaced by the above relation, which is a nonlinear algebraic equation for \( \phi_\alpha \) uniquely determined for each \((c_1, \cdots, c_m)\). We have the following theorem.

Theorem 4.3.5. Let the assumptions in Theorem 4.3.4 be met, then

(i) If \( \rho_{i,\alpha}^0 \) is already at steady-state, then \( \rho_{i,\alpha}^n = \rho_{i,\alpha}^0 \) for \( n \geq 1 \).

(ii) If \( E_{h_i}^{n+1} = E_{h_i}^n \), then \( \rho_{i,\alpha}^n \) must be at steady-state.

(iii) If \( \rho_{i,\alpha}^n, \phi_\alpha^n \) converge as \( n \to \infty \), then their limits are determined by

\[
\rho_{i,\alpha}^\infty = c_i e^{-\frac{1}{k_BT}(q_i \phi_\alpha^\infty + \mu_{i,\alpha})}, \quad \phi_\alpha^\infty = \frac{\sum_{\alpha \in A} |K_\alpha| \rho_{i,\alpha}^0}{\sum_{\alpha \in A} |K_\alpha| e^{-\frac{1}{k_BT}(q_i \phi_\alpha^0 + \mu_{i,\alpha})}},
\]

here \( \phi_\alpha^\infty \) is obtained by solving (4.3.11) by using \( \rho_{i,\alpha}^\infty \).

Proof. (i) We only need to prove \( \rho_{i,\alpha}^1 = \rho_{i,\alpha}^0 \), for all \( i = 1, \cdots, m, \alpha \in A \). Summing (4.3.7) with \( n = 0 \) against \( |K_\alpha| \rho_{i,\alpha}^1 / \rho_{i,\alpha}^0 \), using summation by parts, we obtain

\[
\sum_{\alpha \in A} |K_\alpha|(\rho_{i,\alpha}^1 - \rho_{i,\alpha}^0) \rho_{i,\alpha}^1 \rho_{i,\alpha}^0 = \tau \sum_{j=1}^d \sum_{\alpha \in A} |K_\alpha| \frac{1}{h_j} (C_{i,\alpha+j/2} - C_{i,\alpha-j/2}) \left( \frac{\rho_{i,\alpha}^1}{\rho_{i,\alpha}^0} \right).
\]

Substituting \( \rho_{i,\alpha}^0 = c_i e^{-\psi_{i,\alpha}} \) into \( C_{i,\alpha+j/2} \), the right hand side of (4.3.15) becomes

\[
RHS = -\tau c_i \sum_{j=1}^d \sum_{\alpha(j) \neq N_j} |K_\alpha| \frac{D_{i,\alpha+j/2} e^{-\phi_{i,\alpha}^0}}{h_j^2} \left( \frac{\rho_{i,\alpha}^1}{\rho_{i,\alpha}^0} \right) \leq 0.
\]

Adding \( \sum_{\alpha \in A} |K_\alpha|(\rho_{i,\alpha}^1 - \rho_{i,\alpha}^0) = 0 \) to the left hand side of (4.3.15) leads to

\[
LHS = \sum_{\alpha \in A} |K_\alpha| \left[ \left( \frac{\rho_{i,\alpha}^1 - \rho_{i,\alpha}^0}{\rho_{i,\alpha}^0} \right) \right] \geq 0.
\]
Hence $LHS = RHS \equiv 0$, we must have
\[
\rho_{i,\alpha}^1 = \rho_{i,\alpha}^0, \quad i = 1, \ldots, m, \quad \alpha \in \mathcal{A}.
\]

(ii) The inequality (4.3.13) when combined with $E_{h}^{n+1} = E_{h}^{n}$ leads to $I^{n} = 0$. From the proof of Theorem 4.3.4 in Appendix B it follows
\[
\rho_{i,\alpha}^{n+1} = \rho_{i,\alpha}^{n}.
\]

(iii) Since $E_{h}^{n}$ is non-increasing in $n$, and we can verify that $E_{h}^{n}$ is bounded from below, hence
\[
\lim_{n \to \infty} E_{h}^{n} = \inf \{E_{h}^{n}\}.
\]
Taking the limit in (4.3.13), we have $\lim_{n \to \infty} I^{n} = 0$, which implies
\[
\rho_{i,\alpha}^{\infty} = c_{i}^{\infty} e^{-\psi_{i,\alpha}^{\infty}}.
\]
Conservation of mass gives
\[
c_{i}^{\infty} = \frac{\sum_{\alpha \in \mathcal{A}} |K_{\alpha}| \rho_{i,\alpha}^{0}}{\sum_{\alpha \in \mathcal{A}} |K_{\alpha}| e^{-\psi_{i,\alpha}^{\infty}}}, \quad i = 1, \ldots, m, \quad \alpha \in \mathcal{A},
\]
here $\phi_{\alpha}^{\infty}$ in $\psi_{i,\alpha}^{\infty} = \frac{1}{k_{B} T} (q_{i} \phi_{\alpha}^{\infty} + \mu_{i,\alpha})$ is obtained by solving (4.3.11) using $\rho_{i,\alpha}^{\infty}$.

4.4 Second order in time discretization

The semi-discrete scheme (4.3.2a) is first order accurate, one can design higher order in time schemes based on (4.3.1).

The following is a second order time discretization,
\[
\frac{\rho_{i}^{n+1} - \rho_{i}^{n}}{\tau} = R[\rho_{i}^{n+1} + \rho_{i}^{n}] / 2, \quad \frac{3}{2} \psi_{i}^{n} - \frac{1}{2} \psi_{i}^{n-1}.
\]
This can be expressed as a prediction-correction method,
\[
\frac{\rho_{i}^{*} - \rho_{i}^{n}}{\tau / 2} = R[\rho_{i}^{*} + \frac{3}{2} \psi_{i}^{n} - \frac{1}{2} \psi_{i}^{n-1}], \quad \rho_{i}^{n+1} = 2 \rho_{i}^{*} - \rho_{i}^{n}.
\]
As argued for the first order scheme, this scheme is well-defined.
4.4.1 Second order fully-discrete scheme

With central spatial difference, our fully discrete second order (in both space and time) scheme reads

\[ \frac{\rho_i^{n+1} - \rho_i^n}{\tau/2} = R_\alpha[\rho_i^*, \frac{3}{2}\psi_i^n - \frac{1}{2}\psi_i^{n-1}], \tag{4.4.2a} \]
\[ \rho_i^{n+1} = 2\rho_i^* - \rho_i^n. \tag{4.4.2b} \]

Positivity of \( \rho_i^{n+1} \) can be ensured if time steps are sufficient small.

**Theorem 4.4.1.** Let \( \rho_i^{n+1} \) be obtained from (4.4.2). If \( \rho^n_\alpha \geq 0 \) for all \( \alpha \in \mathcal{A} \), and \( \rho^b(x,t) \geq 0 \) for \( x \in \partial\Omega_D \), then

\[ \rho_i^{n+1} \geq 0, \quad \alpha \in \mathcal{A} \]

provided \( \tau \) is sufficiently small.

**Proof.** Inserting (4.4.2b) into (4.4.2a) leads to the following compact form of the scheme (4.4.2):

\[ \rho_i^{n+1} - \frac{\tau}{2} R_\alpha[\rho_i^{n+1}, \frac{3}{2}\psi_i^n - \frac{1}{2}\psi_i^{n-1}] = \rho_i^n + \frac{\tau}{2} R_\alpha[\rho_i^*, \frac{3}{2}\psi_i^n - \frac{1}{2}\psi_i^{n-1}], \tag{4.4.3} \]

here we have used the linearity of \( R_\alpha[\cdot,\cdot] \) on the first entry.

Set

\[ g_i^* = e^{\frac{3}{2}\psi_i^n - \frac{1}{2}\psi_i^{n-1}}, \quad \bar{D}_{i,\alpha+e_j/2}^* = D_{i,\alpha+e_j/2} e^{-\frac{3}{2}\psi_i^{n+1} + \frac{1}{2}\psi_i^{n-1}} \geq 0, \quad G_{i,\alpha}^n = \rho_i^n g_i^*, \]

then the scheme (4.4.3) can be rewritten as

\[ g_i^* G_{i,\alpha}^{n+1} - \sum_{j=1}^d \tau h_j^2 [\bar{D}_{i,\alpha+e_j/2} G_{i,\alpha+e_j}^{n+1} - G_{i,\alpha+e_j}^n - \bar{D}_{i,\alpha-e_j/2} (G_{i,\alpha}^{n+1} - G_{i,\alpha-e_j}^n)] = g_i^* G_{i,\alpha}^n + \sum_{j=1}^d \tau h_j^2 [\bar{D}_{i,\alpha+e_j/2} G_{i,\alpha+e_j}^n - G_{i,\alpha+e_j}^n - \bar{D}_{i,\alpha-e_j/2} (G_{i,\alpha}^n - G_{i,\alpha-e_j}^n)]. \tag{4.4.4} \]

Let \( \beta \) be such that

\[ G_{i,\beta}^{n+1} = \min_{\alpha \in \mathcal{A}} G_{i,\alpha}^{n+1}, \]

it suffices to prove \( G_{i,\beta}^{n+1} \geq 0 \). We prove the result when \( K_\beta \) is an interior cell, the result for boundary cells can be proved similarly.
Since $G_{i,\beta}^{n+1} \leq G_{i,\beta \pm j}^{n+1}$ and $G_{i,\beta \pm j}^n \geq 0$, thus equation (4.4.4) on cell $K_\beta$ reduces to the inequality:

$$g_{i,\beta}^* G_{i,\beta}^{n+1} \geq \left( g_{i,\beta}^* - \tau \sum_{j=1}^d \frac{1}{h_j^2} \left( D_{i,\beta + e_j/2}^* + D_{i,\beta - e_j/2}^* \right) \right) G_{i,\beta}^n,$$

we see that $G_{i,\beta}^{n+1} \geq 0$ is insured if

$$\tau \leq \min_{\alpha} \left\{ \frac{g_{i,\alpha}^*}{\sum_{j=1}^d \frac{1}{h_j^2} \left( D_{i,\alpha + e_j/2}^* + D_{i,\alpha - e_j/2}^* \right)} \right\}.$$

The stated result thus follows.

We should point out that numerical density $\{\rho_\alpha^n\}$ obtained by the second order scheme (4.4.2) may not be non-negative for large time step $\tau$, though $\{\rho_\alpha^*\}$ stays positive. We shall restore solution positivity by using a local limiter, which was first introduced in [21] for one-dimensional case.

4.4.2 Positivity-preserving limiter

We present a local limiter to restore positivity of $\rho$ if

$$\sum_{\alpha \in A} |K_\alpha| \rho_\alpha > 0,$$

but $\rho_\beta < 0$ for some $\beta \in A$. The idea is to find a neighboring index set $S_\beta$ such that the local average

$$\bar{\rho}_\beta = \frac{1}{|S_\beta|} \sum_{\gamma \in S_\beta} |K_\gamma| \rho_\gamma > 0,$$

here $|S_\beta|$ denotes the minimum number of indices for which $\rho_\gamma \neq 0$ and $\bar{\rho}_\beta > 0$, then use this local average as a reference to define the following scaling limiter

$$\tilde{\rho}_\alpha = \theta \rho_\alpha + \left( 1 - \theta \right) \frac{\bar{\rho}_\beta}{|K_\alpha|}, \quad \alpha \in S_\beta,$$

with

$$\theta = \min \left\{ 1, \frac{\bar{\rho}_\beta}{\tilde{\rho}_\beta - \rho_{\min}} \right\}, \quad \rho_{\min} = \min_{\gamma \in S_\beta} |K_\gamma| \rho_\gamma.$$
Recall the result stated in Lemma 5.1 in [22], such limiter restores solution positivity and respects the local mass conservation. In addition, for any sequence $g_\alpha$ with $g_\alpha \geq 0$, we have

$$|\tilde{\rho}_\alpha - g_\alpha| \leq (1 + |S_\beta|\Lambda) \max_{\gamma \in S_\beta} |\rho_\gamma - g_\gamma|, \quad \alpha \in S_\beta,$$

(4.4.6)

where $\Lambda$ is the upper bound of mesh ratio $|K_\gamma|/|K_\alpha|$. Let $\rho_\alpha$ be the approximation of $\rho(x) \geq 0$, we let $g_\alpha = \rho(x_\alpha)$ or the average of $\rho$ on $K_\alpha$, so we can assert that the accuracy is not destroyed by the limiter as long as $|S_\beta|\Lambda$ is uniformly bounded. Boundedness of $|S_\beta|$ for shape-regular meshes was rigorously proved in [22] for the one-dimensional case. We restate such result in the present setting in the following.

**Theorem 4.4.2.** Let $\{\rho_\alpha\}$ be an approximation of $\rho(x) \geq 0$ over shape regular meshes, and $\rho \in C^k(\Omega)$ ($k \geq 2$). If $\rho_\beta < 0$ (or only finite number of neighboring values are negative), then there exists $C^* > 0$ finite such that

$$|\tilde{\rho}_\alpha - \rho(x_\alpha)| \leq C^* \max_{\alpha \in S_\beta} |\rho_\alpha - \rho(x_\alpha)|, \quad \forall \alpha \in S_\beta,$$

where $C^*$ may depend on the local meshes associated with $S_\beta$.

**Remark 4.4.1.** This limiter is of independent interest, and can be applied to other high order finite volume or conservative finite difference schemes. For a particular scheme of choice, it would be interesting to further study accuracy propagation when applying such limiter.

**Proof.** For simplicity, we prove only for the case of uniform meshes (e.g. uniform in each dimension). Let $h = \min_{1 \leq j \leq d} h_j \leq 1$ and $h_j \leq \Lambda h$ for some $\Lambda > 0$. From (4.4.6) we see that it suffices to show there exists $A^* > 0$ finite such that $|S_\beta| \leq A^*$, with which we will have $C^* = 1 + A^*\Lambda$. Under the smoothness assumption of $\rho$ we may assume $|\rho_\alpha - \rho(x_\alpha)| \leq C h^k$. Under the assumption $\rho_\beta < 0$, $\rho$ must touch zero near $x_\beta$. We discuss the case where $\rho(x^*) = 0$ and $\nabla \rho(x^*) = \vec{0}$ with $\rho(x) > 0$ for $x(j) \geq x^*(j), \ j = 1, \cdots, d$, locally with $x^* \in K_\beta$. To be concrete, we consider $\beta = (1, \cdots, 1)$ and $\int_{K_\beta} \rho(x)dx > 0$. From the limiter construction we have $S_\beta$ such that

$$\sum_{\alpha \in S_\beta} |K_\alpha| \rho_\alpha > 0.$$  

(4.4.7)
The rest of the proof is devoted to bounding $|S_\beta|$. The assumed error bound gives

$$\rho_\alpha \geq \rho(x_\alpha) - Ch^k. \quad (4.4.8)$$

From $\rho \in C^k(\Omega)(k \geq 2)$, we have

$$\rho(x_\alpha) \geq \bar{\rho}_\alpha - \lambda \Lambda^2 h^2, \quad (4.4.9)$$

with $\lambda = \frac{d}{24} \max_{j=1,\ldots,d} |\partial_{x_j} x_\beta|$ and the cell average $\bar{\rho}_\alpha = \frac{1}{|K_\alpha|} \int_{K_\alpha} \rho(x) dx$. From (4.4.8) and (4.4.9), we see that the left hand side of (4.4.7) is bounded from below by

$$\sum_{\alpha \in S_\beta} |K_\alpha| \rho_\alpha \geq \sum_{\alpha \in S_\beta} |K_\alpha| \left( \bar{\rho}_\alpha - (C + \lambda \Lambda^2) h^2 \right)$$

$$= \int_{\bigcup_{\alpha \in S_\beta} K_\alpha} \rho(x) dx - (C + \lambda \Lambda^2) h^2 \sum_{\alpha \in S_\beta} |K_\alpha|. \quad (4.4.10)$$

Without loss of generality we assume $\bigcup_{\alpha \in S_\beta} K_\alpha$ is a rectangle in $\mathbb{R}^d$; otherwise we could add more cells to complete the rectangle. Let

$$|\bigcup_{\alpha \in S_\beta} K_\alpha| = \Pi_{j=1}^d \eta_j, \quad h_j \leq \eta_j \leq L_j,$$

and $\bar{\eta} = (\eta_1, \ldots, \eta_d)$, $\bar{h} = (h_1, \ldots, h_d)$. Rewriting integral in (4.4.10) we have

$$\sum_{\alpha \in S_\beta} |K_\alpha| \rho_\alpha \geq \left[ g(\eta) - (C + \lambda \Lambda^2) h^2 \right] \sum_{\alpha \in S_\beta} |K_\alpha|,$$

here

$$g(\eta) := \int_0^1 \cdots \int_0^1 \rho \left( \text{diag}(\bar{\theta}) \bar{\eta} + x_\beta - \frac{1}{2} \bar{h} \right) d\theta_1 \cdots d\theta_d.$$

From the fact $h^d \leq \frac{n_{\beta} \cdots n_d}{|S_\beta|}$, we can see that the term in the bracket is bounded from below by

$$g(\eta) - (C + \lambda \Lambda^2) \left( \frac{\eta_1 \cdots \eta_d}{|S_\beta|} \right)^{2/d},$$

which is positive if

$$|S_\beta| > (C + \lambda \Lambda^2)^{d/2} g(\eta)^{-d/2} \eta_1 \cdots \eta_d.$$

This can be insured if we take

$$|S_\beta| = |A| + 1,$$
here

\[ A = (C + \lambda A^2)^{d/2} \max_{\eta_j \in [h_j, L_j], j = 1, \ldots , d} g(\eta)^{-d/2} \eta_1 \cdots \eta_d. \]

This is bounded and may depend on the local mesh of \( K_\beta \). \( \square \)

Note that our numerical solutions feature the following property: if \( \rho_{i,\alpha}^n = 0 \), then

\[ \rho_{i,\alpha}^{n+1} = 2\rho_{i,\alpha}^* - \rho_{i,\alpha}^n \geq 0 \]

due to the fact \( \rho_{i,\alpha}^* \geq 0 \). This means that if \( \rho_{i,\alpha}(x) = 0 \) on an interval, then \( \rho_{i,\alpha}^1 \) cannot be negative in most of nearby cells. Thus negative values appear only where the exact solution turns from zero to a positive value, and the number of these values are finitely many. Our result in Theorem 4.4.2 is thus applicable.

### 4.4.3 Algorithm

The following algorithm is only for the second order scheme with limiter.

1. Initialization: From the initial data \( \rho_{i,\alpha}^0(x) \), obtain

\[ \rho_{i,\alpha}^0 = \frac{1}{|K_\alpha|} \int_{K_\alpha} \rho_{i,\alpha}^0(x) dx, \quad i = 1, \ldots , m, \quad \alpha \in \mathcal{A}, \]

by using central point quadrature.

2. Update to get \( \{\rho_{i,\alpha}^1\} \): Compute \( \{\phi_{\alpha}^0\} \) from (4.3.11), then obtain \( \{\rho_{i,\alpha}^1\} \) by the first order scheme (4.3.7).

3. Update from \( \{\rho_{i,\alpha}^n\} \): For \( n \geq 1 \), compute \( \{\phi_{\alpha}^n\} \) from scheme (4.3.11) then get \( \{\rho_{i,\alpha}^{n+1}\} \) from (4.4.2).

4. Reconstruction: if necessary, locally replace \( \rho_{i,\alpha}^{n+1} \) by \( \tilde{\rho}_{i,\alpha}^{n+1} \) using the limiter defined in (4.4.5).

The following algorithm can be called to find an admissible set \( S_\alpha \) used in (4.4.5).

(i) Start with \( S_\beta = \{\beta\} \), \( p = 1 \).
(ii) For \( l_j = \max\{1, \alpha(j) - p\} : \min\{\alpha(j) + p, N_j\} \) with \( j = 1, \cdots, d \).

If \( \alpha := (l_1, \cdots, l_d) \notin S_\beta \) and \( \rho^{n+1}_{t,\alpha} \neq 0 \), then set \( S_\beta = S_\beta \cup \{\alpha\} \).

If \( \bar{\rho}_\beta > 0 \), then stop, else go to (iii).

(iii) Set \( p = p + 1 \) and go to (ii).

Remark 4.4.2. The second-order scheme (4.4.2) is only a slight modification of the first-order scheme (4.3.7), hence computationally more efficient. This is critical in 3D simulations. In addition, the energy dissipation was also observed numerically for the second order scheme, see Figure 4.4 (right).

Remark 4.4.3. The coefficient matrices of the linear systems obtained by (4.3.7), (4.3.11), and (4.4.2a) are sparse, diagonally dominant, and symmetric, hence more efficient linear system solvers, such as the ILU preconditioner + FGMRES (see e.g., [37]), ILU preconditioner + Bicgstab (see e.g., [5]), can be used.

4.5 Numerical tests

In this section, we implement the fully discrete schemes (4.3.7) and (4.4.2) to demonstrate their orders of convergence and capacity to preserve solution properties. In both schemes the numerical solution \( \phi^n_\alpha \) is computed by the scheme (4.3.11). Errors in the accuracy tests are measured in the following discrete \( l^1 \) norm:

\[
\text{error} = \sum_{\alpha \in A} |K_\alpha||\tilde{g}_\alpha - g_\alpha|.
\]

Here \( g_\alpha \) denotes the numerical solution, say \( g_\alpha = \rho^n_{t,\alpha} \) or \( \phi^n_\alpha \) at time \( t = n\tau \), and \( \tilde{g}_\alpha \) indicates the cell average of the corresponding exact solutions.

In our numerical tests, the sparse linear systems obtained by (4.3.7), (4.3.11), and (4.4.2a) are solved by ILU preconditioned FGMRES [37] algorithm using compressed row format of the coefficient matrices. In the three-dimensional case, the coefficient matrices of the linear systems are 7-diagonal matrices. It is worth to mention that the compressed row format allows us to store a \( l \times l \) 7-diagonal matrix by using at most 15\( l \) storage locations with \( l = N_x \times N_y \times N_z \). With
In our three examples below we consider the computational domain
\[ \Omega = (0, 1) \times (0, 1) \times (0, 1). \]

**Example 4.5.1.** (Accuracy test) In this test we numerically verify the accuracy and order of schemes (4.3.7) and (4.4.2) by using manufactured solutions. Consider
\[
\begin{align*}
\rho_1(x, t) &= 4(x^2(1 - x)^2 + y(1 - y))e^{-t}, \\
\rho_2(x, t) &= (y(1 - y) + z^2(1 - z)^2)e^{-t}, \\
\phi(x, t) &= (x^2(1 - x)^2 + y(1 - y) + z^2(1 - z)^2)e^{-t}
\end{align*}
\]
and
\[ \partial \Omega_D = \{ x \in \bar{\Omega} : y = 0, 1 \}, \quad \partial \Omega_N = \partial \bar{\Omega} \setminus \partial \Omega_D, \]
then they are exact solutions to the following problem
\[
\begin{align*}
\partial_t \rho_1 &= \nabla \cdot (\nabla \rho_1 + \rho_1 \nabla \phi) + f_1(x, t), \quad x \in \Omega, \ t > 0, \\
\partial_t \rho_2 &= \nabla \cdot (\nabla \rho_2 - \rho_2 \nabla \phi) + f_2(x, t), \quad x \in \Omega, \ t > 0, \\
-\Delta \psi &= \rho_1 - \rho_2 + f_3(x, t), \quad x \in \Omega, \ t > 0,
\end{align*}
\]here source terms \( f_1(x, t), f_2(x, t) \) and \( f_3(x, t) \), and the initial and boundary conditions are determined by the exact solutions.

We first test the accuracy of the semi-implicit scheme (4.3.7) by using various spatial step size \( h \), errors and orders at \( t = 1 \) are listed in Table 4.1 (with \( \tau = h \)) and in Table 4.2 (with \( \tau = h^2 \)), respectively. We observe the first order accuracy in time and the second order accuracy in space. We then test the accuracy of the scheme (4.4.2) with time step size \( \tau = h \). From Table 4.3, we see the second order accuracy in both time and space.
Table 4.1 Scheme (4.3.7) with $\tau = h$

<table>
<thead>
<tr>
<th>$N_x \times N_y \times N_z$</th>
<th>$\rho_1$ error</th>
<th>order</th>
<th>$\rho_2$ error</th>
<th>order</th>
<th>$\phi$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$8 \times 8 \times 8$</td>
<td>4.7508E-02</td>
<td>-</td>
<td>1.3904E-02</td>
<td>-</td>
<td>5.7213E-03</td>
<td>-</td>
</tr>
<tr>
<td>$16 \times 16 \times 16$</td>
<td>2.1283E-02</td>
<td>1.1585</td>
<td>5.8701E-03</td>
<td>1.2440</td>
<td>2.0987E-03</td>
<td>1.4468</td>
</tr>
<tr>
<td>$32 \times 32 \times 32$</td>
<td>1.0060E-02</td>
<td>1.0811</td>
<td>2.6956E-03</td>
<td>1.1228</td>
<td>8.6460E-04</td>
<td>1.2794</td>
</tr>
<tr>
<td>$64 \times 64 \times 64$</td>
<td>4.8890E-03</td>
<td>1.0410</td>
<td>1.2915E-03</td>
<td>1.0616</td>
<td>3.8667E-04</td>
<td>1.1609</td>
</tr>
</tbody>
</table>

Table 4.2 Scheme (4.3.7) with $\tau = h^2$

<table>
<thead>
<tr>
<th>$N_x \times N_y \times N_z$</th>
<th>$\rho_1$ error</th>
<th>order</th>
<th>$\rho_2$ error</th>
<th>order</th>
<th>$\phi$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$8 \times 8 \times 8$</td>
<td>1.1252E-02</td>
<td>-</td>
<td>4.0301E-03</td>
<td>-</td>
<td>3.1194E-03</td>
<td>-</td>
</tr>
<tr>
<td>$16 \times 16 \times 16$</td>
<td>2.7824E-03</td>
<td>2.0158</td>
<td>9.8548E-04</td>
<td>2.0319</td>
<td>7.7117E-04</td>
<td>2.0161</td>
</tr>
<tr>
<td>$32 \times 32 \times 32$</td>
<td>6.9369E-04</td>
<td>2.0040</td>
<td>2.4502E-04</td>
<td>2.0079</td>
<td>1.9225E-04</td>
<td>2.0041</td>
</tr>
<tr>
<td>$64 \times 64 \times 64$</td>
<td>1.7330E-04</td>
<td>2.0010</td>
<td>6.1170E-05</td>
<td>2.0020</td>
<td>4.8028E-05</td>
<td>2.0010</td>
</tr>
</tbody>
</table>

**Example 4.5.2.** (Solution positivity) We consider the two-species PNP system with initial data of form

\[ \begin{align*}
\partial_t \rho_1 &= \nabla \cdot (\nabla \rho_1 + \rho_1 \nabla \phi), \quad x \in \Omega, \ t > 0, \\
\partial_t \rho_2 &= \nabla \cdot (\nabla \rho_2 - \rho_2 \nabla \phi), \quad x \in \Omega, \ t > 0, \\
-\Delta \psi &= \rho_1 - \rho_2 + 10 \chi_{[0.2,0.4] \times [0.2,0.4] \times [0.2,0.4]}, \quad x \in \Omega, \ t > 0, \\
\rho_1^{in}(x) &= \chi_{[0.0,0.25] \times [0.0,0.25] \times [0.0,0.25]}, \\
\rho_2^{in}(x) &= 2 \chi_{[0.0,0.25] \times [0.0,0.25] \times [0.0,0.25]}.
\end{align*} \]

This corresponds to (4.1.1) with $D_1 = D_2 = 1$, $q_1 = -q_2 = 1$, $k_B T = 1$, $\epsilon(x) = 4\pi$, $\mu_i = 0$, and $f(x) = 10 \chi_{[0.2,0.4] \times [0.2,0.4] \times [0.2,0.4]}.$

Table 4.3 Scheme (4.4.2) with $\tau = h$

<table>
<thead>
<tr>
<th>$N_x \times N_y \times N_z$</th>
<th>$\rho_1$ error</th>
<th>order</th>
<th>$\rho_2$ error</th>
<th>order</th>
<th>$\phi$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>$8 \times 8 \times 8$</td>
<td>5.5476E-03</td>
<td>-</td>
<td>2.3247E-03</td>
<td>-</td>
<td>2.7378E-03</td>
<td>-</td>
</tr>
<tr>
<td>$16 \times 16 \times 16$</td>
<td>1.5073E-03</td>
<td>1.8799</td>
<td>6.0465E-04</td>
<td>1.9429</td>
<td>6.7758E-04</td>
<td>2.0146</td>
</tr>
<tr>
<td>$32 \times 32 \times 32$</td>
<td>4.9635E-04</td>
<td>1.9271</td>
<td>1.5851E-04</td>
<td>1.9315</td>
<td>1.6895E-04</td>
<td>2.0038</td>
</tr>
<tr>
<td>$64 \times 64 \times 64$</td>
<td>1.0182E-04</td>
<td>1.9608</td>
<td>4.0875E-05</td>
<td>1.9553</td>
<td>4.2206E-05</td>
<td>2.0011</td>
</tr>
</tbody>
</table>
With $\partial \Omega_D = \{ x \in \bar{\Omega} : y = 0, 1 \}$, and $\partial \Omega_N = \partial \bar{\Omega} \setminus \partial \Omega_D$, we solve the problem subject to mixed boundary conditions

$$
\begin{cases}
(\nabla \phi) \cdot n = 0, & (\nabla \rho_1 + \rho_1 \nabla \phi) \cdot n = 0, & (\nabla \rho_2 - \rho_2 \nabla \phi) \cdot n = 0, & x \in \partial \Omega_N, \\
\phi^b(x, t) = (x^2(1 - x)^2 + z^2(1 - z)^2)e^{-t}, & x \in \partial \Omega_D, \\
\rho_1^b(x, t) = 4x^2(1 - x)^2e^{-t}, & x \in \partial \Omega_D, \\
\rho_2^b(x, t) = z^2(1 - z)^2e^{-t}, & x \in \partial \Omega_D.
\end{cases}
$$

We use $30 \times 30 \times 30$ cells with $\tau = 0.5h$ to compute numerical solutions up to $t = 2$. Given in Figure 4.1 are the time evolution of numerical solutions (top three rows) and the minimum of $\rho_1, \rho_2$ (bottom row) obtained by the scheme (4.3.7), showing non-negative approximations for both $\rho_1$ and $\rho_2$. Results obtained by the scheme (4.4.2) are given in Figure 4.2. Note that the positivity preserving limiter keeps being invoked when we use the scheme (4.4.2). The CPU time (average of 10 simulations) needed for running the schemes (4.3.7) and (4.4.2) are 207.27 seconds and 203.15 seconds, respectively, from which we see that the second-order scheme is as efficient as the first order scheme.

**Example 4.5.3.** (Mass conservation and energy dissipation) In this numerical example we test both mass conservation and energy dissipation properties of our schemes.

We consider system (4.5.3) with zero flux boundary conditions:

$$(\nabla \phi) \cdot n = 0, & (\nabla \rho_1 + \rho_1 \nabla \phi) \cdot n = 0, & (\nabla \rho_2 - \rho_2 \nabla \phi) \cdot n = 0, & x \in \partial \Omega.$$
Figure 4.1  Example 4.5.2: $\rho_1, \rho_2, \phi$ computed by scheme (4.3.7)
Figure 4.2  Example 4.5.2: $\rho_1, \rho_2, \phi$ computed by scheme (4.4.2) (with limiter)
4.6 Concluding remarks

In this paper, we have developed unconditional structure-preserving schemes for PNP equations in more general settings. These schemes are shown to preserve several important physical laws at the fully discrete level including: mass conservation, solution positivity, and free energy dissipation. The non-logarithmic Landau reformulation of the model is important, enabling us to construct a simple, easy-to-implement fully discrete scheme (first order in time, second order in space), which proved to satisfy all three desired properties of the continuous model with only $O(1)$ time step restriction. We further designed a second order (in both time and space) scheme, which has the same
computational complexity as the first order scheme. For such second order scheme, we employed a local scaling limiter to restore solution positivity where necessary. Moreover, we rigorously proved that the limiter does not destroy the desired accuracy. Three-dimensional numerical tests are conducted to evaluate the scheme performance and verify our theoretical findings. Our schemes presented with $\mu_i$ given can be applied to complex chemical potentials without difficulty.

Acknowledgments

This research was partially supported by the National Science Foundation under Grant DMS1812666.

4.7 Appendix A: Proof of Theorem 4.3.1

Proof. The elliptic problem (4.3.2) can be rewritten in $w = \rho_i^{n+1} e^{\psi_i^n}$ as

\begin{align*}
  e^{-\psi_i^n} w - \tau \nabla \cdot (D_i(x) e^{-\psi_i^n} \nabla w) &= \rho_i^n, & (4.7.1a) \\
  w &= \rho_i^b(x, t_{n+1}) e^{\psi_i^b(x, t_n)}, \quad x \in \partial \Omega_D, & (4.7.1b) \\
  (\nabla w) \cdot n &= 0, \quad x \in \partial \Omega_N. & (4.7.1c)
\end{align*}

Let $\gamma_0$ be the trace operator on $\partial \Omega_D$. The above problem admits a variational formulation of form

\[ B[u, v] = Lv, \quad u, v \in H, \tag{4.7.2} \]

where for a Dirichlet lift $G \in H^2(\Omega)$ with trace $\gamma_0(G) = \rho_i^b(x, t_{n+1}) e^{\psi_i^b(x, t_n)}$, we find

\[ w = u + G. \]

Here

\begin{align*}
  H &= \{ v \in H^1(\Omega) : \gamma_0(v) = 0 \text{ on } \partial \Omega_D \}, \tag{4.7.3a} \\
  B[u, v] &= \int_{\Omega} (\tau D_i(x) e^{-\psi_i^n} \nabla u \cdot \nabla v + e^{-\psi_i^n} uv) dx, \tag{4.7.3b} \\
  Lv &= \int_{\Omega} (\rho_i^n - e^{-\psi_i^n} G) v - \tau D_i(x) e^{-\psi_i^n} \nabla G \cdot \nabla v dx. \tag{4.7.3c}
\end{align*}
Under the assumptions, the celebrated Lax-Milgram theorem ([12] Theorem 5.8) ensures that the variational problem (4.7.2) admits a unique solution \( u \in H \). We thus obtain

\[
\rho_i^{n+1} = e^{-\psi_i^n} (u + G).
\]

Regularity for \( \rho_i^{n+1} \) follows from the classical elliptic regularity for \( u \).

Similarly, the variational problem for (4.3.3) can also be written as (4.7.2) with

\[
B[u, v] = \int_{\Omega} \epsilon(x) \nabla u \cdot \nabla v dx,
\]

\[
Lv = \int_{\Omega} 4\pi \left( f(x) + \sum_{i=1}^{m} q_i \rho_i^{n+1} \right) v - \epsilon(x) \nabla G \cdot \nabla v dx,
\]

where the Dirichlet lift \( G \in H^2(\Omega) \) with \( \gamma_0(G) = \phi^b(x, t_{n+1}) \) on \( x \in \partial\Omega_D \). Here one can use the Poincaré-Friedrichs’ inequality of form \( \|u\|_{L^2} \leq C_F \|\nabla u\|_{L^2} \), which holds if \( u = 0 \) on a set of \( \partial\Omega \) with non-vanishing measure, to regain coercivity of \( B \) on \( H \). The variational problem is thus well-posed, and we obtain

\[
\phi^{n+1} = u + G.
\]

Regularity for \( \phi^{n+1} \) follows from the classical elliptic regularity for \( u \) and regularity for \( \rho^{n+1} \).

If \( \partial\Omega_D = \emptyset \), then \( B[u, 1] = 0 \) requires the compatibility condition for the source

\[
\int_{\Omega} \left( f(x) + \sum_{i=1}^{m} q_i \rho_i^{n+1} \right) dx = 0.
\]

Due to conservation of mass, this can be ensured by

\[
\int_{\Omega} \left( f(x) + \sum_{i=1}^{m} q_i \rho_i^{in} \right) dx = 0.
\]

With such compatibility condition the solution of this variational formulation exists but is not unique. In such case one can replace \( H \) by

\[
H_* = \left\{ v \in H^1, \quad \int_{\Omega} v dx = 0 \right\},
\]

then by the Poincaré-Wirtinger inequality, \( B \) is actually \( H_* \)-coercive. The new variational problem hence admits a unique solution and is well-posed.
Finally we prove positivity of $\rho^{n+1}$ if $\rho^n \geq 0$. Since $w = \rho_i^{n+1}e^{\psi^n_i} \in C(\bar{\Omega}) \cap C^2(\Omega)$, we let $x^* = \arg\min_{x \in \bar{\Omega}} w(x)$, and distinct three cases:

(i) If $x^* \in \partial \Omega_D$, then

$$w(x) \geq w(x^*) = \rho_i^b(x^*, t_{n+1})e^{\psi_i^b(x^*, t_n)} \geq 0, \quad x \in \bar{\Omega}.$$ 

(ii) If $x^* \in \Omega$, then we can show that

$$w(x) \geq w(x^*) \geq \rho_i^n(x^*)e^{\psi_i^n(x^*)} \geq 0, \quad x \in \bar{\Omega}.$$ 

In fact, from (4.7.1a) it follows

$$\rho_i^n(x) = e^{-\psi_i^n(x)}w(x) - \tau \nabla \cdot (D_i(x)e^{-\psi_i^n(x)}\nabla w(x)) \quad \text{or} \quad \rho_i^n(x) = e^{-\psi_i^n(x)}w(x) - \tau \nabla (D_i(x)e^{-\psi_i^n(x)}\nabla w(x)) - \tau D_i(x)e^{-\psi_i^n(x)}\Delta w(x).$$

This when evaluated at $x^*$, using $\nabla w(x^*) = 0$ and $\Delta w(x^*) \geq 0$, gives

$$\rho_i^n(x^*) \leq e^{-\psi_i^n(x^*)}w(x^*).$$

(iii) For $x^* \in \partial \Omega_N$. If $w(x^*) \geq 0$, the proof is complete. We proceed with the case that

$$w(x^*) < 0, \quad x^* \in \partial \Omega_N.$$ 

This is possible by the Hopf strong minimum principle.

Define the differential operator

$$L\xi := \tau D_i(x)e^{-\psi_i^n(x)}\Delta \xi + \tau \nabla (D_i(x)e^{-\psi_i^n(x)}) \cdot \nabla \xi - e^{-\psi_i^n(x)}\xi.$$ 

We then have $Lw = -\rho_i^n(x) \leq 0$, and $w(x) \geq w(x^*)$ for all $x \in \Omega$. These together with $w(x^*) < 0$ allow us to apply Theorem 8 in [35] to conclude $(\nabla w(x^*)) \cdot \mathbf{n} < 0$. This is a contradiction.

Collecting all three cases, we have $w(x) \geq 0$ for all $x \in \bar{\Omega}$. 

4.8 Appendix B: Proof of Theorem 4.3.4.

Proof. (i) For fixed $i$ we sum (4.3.7) over all cells to get

$$\sum_{\alpha \in A} |K_\alpha| (\rho_i^{n+1} - \rho_i^n) = \tau \sum_{j=1}^d \sum_{\alpha \in A} |K_\alpha| \left( C_{i,\alpha + \epsilon_j/2} - C_{i,\alpha - \epsilon_j/2} \right) = 0,$$
where we used summation by parts and $C_{i,\alpha+e_j/2} = 0$ for $x_{\alpha+e_j/2} \in \partial \Omega$.

(ii) Set

$$S^n_\alpha = f_\alpha + \sum_{i=1}^m q_i \rho^n_i \alpha$$

and

$$\psi^*_{i,\alpha} = \log \rho^{n+1}_i \alpha + \frac{1}{k_B T} q_i \phi^n_i + \frac{1}{k_B T} \mu_{i,\alpha}.$$

Using (4.3.12) we find that

$$E^{n+1}_h - E^n_h = \sum_{\alpha \in A} \sum_{i=1}^m |K_\alpha| \left( (\rho^{n+1}_i \alpha - \rho^n_i \alpha) \psi^*_{i,\alpha} + \rho^n_i \alpha \log \frac{\rho^{n+1}_i \alpha}{\rho^n_i \alpha} \right)$$

$$+ \frac{1}{k_B T} \sum_{\alpha \in A} |K_\alpha| \left( \frac{1}{2} S^{n+1}_\alpha \phi^n_\alpha + \frac{1}{2} S^n_\alpha \phi^n_\alpha - S^{n+1}_\alpha \phi^n_\alpha \right).$$

Using $\log X \leq X - 1$ for $X > 0$ and the mass conservation, we have

$$\sum_{\alpha \in A} |K_\alpha| \rho^{n+1}_i \alpha \log \frac{\rho^{n+1}_i \alpha}{\rho^n_i \alpha} \leq \sum_{\alpha \in A} |K_\alpha| (\rho^{n+1}_i \alpha - \rho^n_i \alpha) = 0.$$

Also one can verify that

$$\sum_{\alpha \in A} |K_\alpha| S^{n+1}_\alpha \phi^n_\alpha = \sum_{\alpha \in A} |K_\alpha| S^n_\alpha \phi^{n+1}_\alpha,$$

with which we obtain

$$\sum_{\alpha \in A} |K_\alpha| \left( \frac{1}{2} S^{n+1}_\alpha \phi^n_\alpha + \frac{1}{2} S^n_\alpha \phi^n_\alpha - S^{n+1}_\alpha \phi^n_\alpha \right) = \frac{1}{2} \sum_{\alpha \in A} |K_\alpha| (S^{n+1}_\alpha - S^n_\alpha)(\phi^{n+1}_\alpha - \phi^n_\alpha).$$

Insertion of these into (4.8.1) gives

$$E^{n+1}_h - E^n_h \leq -\tau I^n + \tau^2 II^n,$$

where

$$I^n = -\sum_{\alpha \in A} \sum_{i=1}^m |K_\alpha| \left( \frac{\rho^{n+1}_i \alpha - \rho^n_i \alpha}{\tau} \right) \psi^*_{i,\alpha},$$

$$II^n = \frac{1}{2k_B T} \sum_{\alpha \in A} |K_\alpha| \left( \frac{S^{n+1}_\alpha - S^n_\alpha}{\tau} \right) \left( \frac{\phi^{n+1}_\alpha - \phi^n_\alpha}{\tau} \right).$$
By using (4.3.7) and summation by parts, we have

\[
I^n = -\sum_{i=1}^{m} \sum_{\alpha \in A} \sum_{j=1}^{d} |K_{\alpha}| \left( \frac{C_{i,\alpha+e_j/2} - C_{i,\alpha-e_j/2}}{h_j} \right) \psi_{i,\alpha}^n
\]

\[
= \sum_{i=1}^{m} \sum_{j=1}^{d} \sum_{\alpha(j) \neq N_j} |K_{\alpha}| \left( \frac{\psi_{i,\alpha+e_j}^n - \psi_{i,\alpha}^n}{h_j} \right) C_{i,\alpha+e_j/2}.
\]

Note that

\[
C_{i,\alpha+e_j/2} = \frac{D_i(x_{i,\alpha+e_j/2}) e^{-\psi_{i,\alpha+e_j/2}^n}}{h_j} \left( e^{\psi_{i,\alpha+e_j}^n} - e^{\psi_{i,\alpha}^n} \right),
\]

hence \(I^n \geq 0\).

We pause to discuss the special case with \(I^n = 0\). In such case we must have \(\psi_{i,\alpha+e_j}^n = \psi_{i,\alpha}^n\) for each \(i, j\) and \(\alpha \in A\), which implies \(C_{i,\alpha+e_j/2} = 0\) for each \(i, j\) and \(\alpha \in A\). Thus, we have \(\rho_{i,\alpha}^{n+1} = \rho_{i,\alpha}^n\), hence

\[
S_{\alpha} = f_{\alpha} + \sum_{i=1}^{m} q_i \rho_{i,\alpha}^{n+1} = f_{\alpha} + \sum_{i=1}^{m} q_i \rho_{i,\alpha}^n = S_{\alpha}^{n+1}, \quad \forall \alpha \in A,
\]

therefore \(H^n = 0\) and \(E_h^{n+1} - E_h^n \leq 0\), this is (4.3.13) with \(I^n = 0\).

From now on we only consider the case \(I^n > 0\). We proceed to estimate \(H^n\),

\[
H^n = \frac{1}{2k_BT} \sum_{\alpha \in A} |K_{\alpha}| \left( \frac{S_{\alpha}^{n+1} - S_{\alpha}^n}{\tau} \right) \left( \frac{\phi_{\alpha}^{n+1} - \phi_{\alpha}^n}{\tau} \right)
\]

\[
= \frac{1}{8\pi k_BT} \sum_{\alpha \in A} \sum_{j=1}^{d} |K_{\alpha}| \left( \frac{\phi_{\alpha+e_j/2}^{n+1} - \phi_{\alpha-e_j/2}^{n+1} - \phi_{\alpha+e_j/2}^n + \phi_{\alpha-e_j/2}^n}{\tau^2 h_j} \right) (\phi_{\alpha}^{n+1} - \phi_{\alpha}^n)
\]

\[
= \frac{1}{8\pi k_BT} \sum_{j=1}^{d} \sum_{\alpha(j) \neq N_j} |K_{\alpha}| \left( \frac{\phi_{\alpha+e_j/2}^{n+1} - \phi_{\alpha-e_j/2}^{n+1} - \phi_{\alpha+e_j/2}^n + \phi_{\alpha-e_j/2}^n}{\tau^2 h_j} \right) (\phi_{\alpha}^{n+1} - \phi_{\alpha}^n)
\]

\[
= \frac{1}{8\pi k_BT} \sum_{j=1}^{d} \sum_{\alpha(j) \neq N_j} |K_{\alpha}| \left( \frac{\phi_{\alpha+e_j/2}^{n+1} - \phi_{\alpha-e_j/2}^{n+1} - \phi_{\alpha+e_j}^n + \phi_{\alpha-e_j}^n}{\tau^2 h_j} \right) (\phi_{\alpha}^{n+1} - \phi_{\alpha}^n).
\]

Here the second equality is obtained by using the equation (4.3.11), the last equality is obtained by using the definition (4.3.11b) of \(\Phi_{\alpha+e_j/2}^n\).

From (4.8.3) and (4.8.4), we see that the energy dissipation inequality (4.3.13) is satisfied if

\[
\tau \leq \tau^* \leq \frac{I^n}{2H^n}.
\]
In the remaining of the proof we will quantify $\tau^*$ from estimating the lower bound of $\frac{I_n^c}{2\pi \tau}$. Subtracting (4.3.11) at time level $t = t_{n+1}$ and $t = t_n$, one has

$$-\frac{d}{j=1} \frac{\Phi_{\alpha+e_j/2}^{n+1} - \Phi_{\alpha-e_j/2}^{n+1} - \Phi_{\alpha+e_j/2}^n + \Phi_{\alpha-e_j/2}^n}{h_j} = 4\pi \sum_{i=1}^{m} q_i (\rho_{i,\alpha}^{n+1} - \rho_{i,\alpha}^n), \quad (4.8.6)$$

multiplying by $|K_\alpha|(\phi_{\alpha}^{n+1} - \phi_{\alpha}^n)$ and summing over $\alpha \in A$ leads to

$$-\sum_{i=1}^{m} \sum_{\alpha \in A} K_\alpha \frac{\Phi_{\alpha+e_j/2}^{n+1} - \Phi_{\alpha-e_j/2}^{n+1} - \Phi_{\alpha+e_j/2}^n + \Phi_{\alpha-e_j/2}^n}{h_j} (\phi_{\alpha}^{n+1} - \phi_{\alpha}^n)$$

$$= 4\pi \sum_{i=1}^{m} \sum_{\alpha \in A} q_i |K_\alpha| (\rho_{i,\alpha}^{n+1} - \rho_{i,\alpha}^n)(\phi_{\alpha}^{n+1} - \phi_{\alpha}^n). \quad (4.8.7)$$

Similar to (4.8.4), the left hand side of (4.8.7) reduces to

$$LHS = \sum_{j=1}^{d} \sum_{\alpha \in A, \alpha \neq N_j} |K_\alpha| \frac{\epsilon_{\alpha+e_j/2}}{h_j^2} (\phi_{\alpha+e_j}^{n+1} - \phi_{\alpha+e_j}^n - \phi_{\alpha+e_j}^{n+1} + \phi_{\alpha}^{n+1})^2. \quad (4.8.8)$$

We estimate the right hand side of (4.8.7) by using the equation (4.3.7):

$$RHS = 4\pi \sum_{i=1}^{m} \sum_{\alpha \in A} q_i |K_\alpha| (\rho_{i,\alpha}^{n+1} - \rho_{i,\alpha}^n)(\phi_{\alpha}^{n+1} - \phi_{\alpha}^n)$$

$$= 4\pi \sum_{i=1}^{m} \sum_{\alpha \in A} \sum_{j=1}^{d} q_i |K_\alpha| \frac{1}{h_j} (C_{i,\alpha+e_j/2} - C_{i,\alpha-e_j/2}) (\phi_{\alpha}^{n+1} - \phi_{\alpha}^n) \quad (4.8.9)$$

$$= -4\pi \sum_{i=1}^{m} \sum_{j=1}^{d} \sum_{\alpha \in A, \alpha \neq N_j} q_i |K_\alpha| \frac{1}{h_j} C_{i,\alpha+e_j/2} (\phi_{\alpha+e_j}^{n+1} - \phi_{\alpha+e_j}^n - \phi_{\alpha}^{n+1} + \phi_{\alpha}^n).$$

Note that

$$LHS \geq \epsilon_{min} \sum_{j=1}^{d} \sum_{\alpha \neq N_j} |K_\alpha| \left( \frac{\phi_{\alpha+e_j}^{n+1} - \phi_{\alpha+e_j}^n - \phi_{\alpha}^{n+1} + \phi_{\alpha}^n}{h_j} \right)^2.$$

Using the Cauchy-Schwarz inequality we see that

$$RHS \leq 4\pi \sum_{i=1}^{m} |q_i| \left[ \sum_{j=1}^{d} \sum_{\alpha \neq N_j} |K_\alpha| C_{i,\alpha+e_j/2} \left( \phi_{\alpha+e_j}^{n+1} - \phi_{\alpha+e_j}^n - \phi_{\alpha}^{n+1} + \phi_{\alpha}^n \right) \right]^{1/2} \quad (4.8.10)$$

$$\leq 4\pi \sum_{i=1}^{m} |q_i| \left[ \sum_{j=1}^{d} \sum_{\alpha \neq N_j} |K_\alpha| C_{i,\alpha+e_j/2}^2 \right]^{1/2} \times \left[ \sum_{j=1}^{d} \sum_{\alpha \neq N_j} |K_\alpha| \left( \frac{\phi_{\alpha+e_j}^{n+1} - \phi_{\alpha+e_j}^n - \phi_{\alpha}^{n+1} + \phi_{\alpha}^n}{h_j} \right) \right]^{1/2}.$$
We thus obtain
\[
\sum_{j=1}^{d} \sum_{\alpha(j) \neq N_j} |K_\alpha| \left( \frac{\phi^{n+1}_{\alpha+\epsilon_j} - \phi^{n+1}_{\alpha} + \phi^n_{\alpha}}{h_j} \right)^2 \leq 16\pi^2 \frac{\sum_{i=1}^{m} q_i}{\epsilon_{min}^2} \left( \sum_{j=1}^{d} \sum_{\alpha(j) \neq N_j} |K_\alpha| C^2_{i,\alpha+\epsilon_j/2} \right)^{1/2)^2} \tag{4.8.11}
\]

\[
\leq 16\pi^2 \frac{\sum_{i=1}^{m} q_i^2}{\epsilon_{min}^2} \left( \sum_{i=1}^{m} q_i^2 \right)^{1/2} \sum_{i=1}^{d} \sum_{j=1}^{m} \sum_{\alpha(j) \neq N_j} |K_\alpha| C^2_{i,\alpha+\epsilon_j/2}.
\]

Upon insertion into (4.8.4)
\[
II^n \leq C \sum_{i=1}^{m} \sum_{j=1}^{d} \sum_{\alpha(j) \neq N_j} |K_\alpha| C^2_{i,\alpha+\epsilon_j/2}, \tag{4.8.12}
\]

here \( C = \frac{2e_{max} \pi \sum_{i=1}^{m} q_i^2}{\epsilon_{min}^2 kB^4} \). We use (4.8.3) and (4.8.12) to obtain:

\[
\frac{I^n}{2II^n} \geq \frac{\sum_{i=1}^{m} \sum_{j=1}^{d} \sum_{\alpha(j) \neq N_j} |K_\alpha| C_{i,\alpha+\epsilon_j/2} (\psi^n_{i,\alpha+\epsilon_j} - \psi^n_{i,\alpha})}{2C \sum_{i=1}^{m} \sum_{j=1}^{d} \sum_{\alpha(j) \neq N_j} |K_\alpha| C^2_{i,\alpha+\epsilon_j/2}} \geq \frac{1}{2C} \min_{i,j,\alpha} \left\{ \psi^n_{i,\alpha+\epsilon_j} - \psi^n_{i,\alpha} \right\} \left( \frac{h_j}{C_{i,\alpha+\epsilon_j/2}} \right)
\]

\[
= \frac{1}{2C} \min_{i,j,\alpha} \left\{ \frac{\psi^n_{i,\alpha+\epsilon_j} - \psi^n_{i,\alpha}}{D_{i,\alpha+\epsilon_j/2} e^{-\psi^n_{i,\alpha} + \epsilon_j/2} (e^{\psi^n_{i,\alpha} - \psi^n_{i,\alpha}}) } \right\} \text{ by the mean-value theorem}
\]

\[
= \frac{1}{2C} \min_{i,j,\alpha} \left\{ \frac{1}{D_{i,\alpha+\epsilon_j/2} e^{-\psi^n_{i,\alpha} + \epsilon_j/2} (e^{\theta \psi^n_{i,\alpha} + (1-\theta)\psi^n_{i,\alpha}}) } \right\},
\]

here \( \theta \in (0, 1) \). By using the harmonic mean for \( e^{-\psi^n_{i,\alpha} + \epsilon_j/2} \), we have
\[
\frac{1}{e^{-\psi^n_{i,\alpha} + \epsilon_j/2} (e^{\theta \psi^n_{i,\alpha} + (1-\theta)\psi^n_{i,\alpha}})} = \frac{e^{(\theta-1)\psi^n_{i,\alpha} - \theta \psi^n_{i,\alpha} + (1-\theta)\psi^n_{i,\alpha}}}{\rho^{n+1}_{i,\alpha+\epsilon_j}} \rho^{n+1}_{i,\alpha} \geq \frac{1}{2C} \min_{i,j,\alpha} \left\{ \psi^n_{i,\alpha+\epsilon_j} - \psi^n_{i,\alpha} \right\}
\]

\[
\geq \frac{2e}{2Me} \max_{i,j,\alpha} \left\{ \psi^n_{i,\alpha+\epsilon_j} - \psi^n_{i,\alpha} \right\} = \frac{\rho^{n+1}_{i,\alpha+\epsilon_j}}{M},
\]
here $M = \max_{i,\alpha,n} \rho^m_{i,\alpha}$, thus

$$
\frac{I^n}{2I^n} \geq \frac{1}{2CD_{\text{max}}M} e^{-\max_{i,j,\alpha} |\psi_{i,\alpha}^m - \psi_{i,\alpha}^n|}.
$$

(4.8.14)

For geometric mean or algebraic mean when used for the evaluation of $e^{-\psi_{i,\alpha}^m + \psi_{i,\alpha}^n}$ we can verify either the same or bigger bound than the right hand side of in (4.8.14).

References


[9] A. Flavell, M. Machen, B. Eisenberg, J. Kabre, C. Liu, and X. Li. A conservative finite
difference scheme for poisson–nernst–planck equations. *Journal of Computational Electronics*,


2015.


[15] D. He, K. Pan, and X. Yue. A positivity preserving and free energy dissipative difference scheme

[16] J. Hu and X. Huang. A fully discrete positivity-preserving and energy-dissipative finite difference


for hard-sphere potential: I–v relations and critical potentials. part i: analysis. *Journal of


CHAPTER 5. POSITIVE AND FREE ENERGY SATISFYING SCHEMES
FOR DIFFUSION WITH INTERACTION POTENTIALS

A paper accepted for publication in Journal of Computational Physics
Hailiang Liu and Wumaier Maimaitiyiming

Abstract

In this paper, we design and analyze second order positive and free energy satisfying schemes for solving diffusion equations with interaction potentials. The semi-discrete scheme is shown to conserve mass, preserve solution positivity, and satisfy a discrete free energy dissipation law for nonuniform meshes. These properties for the fully-discrete scheme (first order in time) remain preserved without a strict restriction on time steps. For the fully second order (in both time and space) scheme, a local scaling limiter is introduced to restore solution positivity when necessary. It is proved that such limiter does not destroy the second order accuracy. In addition, these schemes are easy to implement, and efficient in simulations. Both one and two dimensional numerical examples are presented to demonstrate the performance of these schemes.

5.1 Introduction

This paper is concerned with efficient numerical approximations to the following problem,

\[
\begin{aligned}
\partial_t \rho &= \nabla \cdot (\nabla \rho + \rho \nabla (V(x) + W * \rho)), & \quad x \in \Omega \subset \mathbb{R}^d, \quad t > 0, \\
\rho(x, 0) &= \rho_0(x), & \quad x \in \Omega \subset \mathbb{R}^d,
\end{aligned}
\]  

(5.1.1)

subject to zero flux boundary conditions. Here \( \Omega \) is a bounded domain in \( \mathbb{R}^d \), \( \rho = \rho(x, t) \) is the unknown density, \( V(x) \) is a confinement potential, and \( W(x) \) is an interaction potential, which is assumed to be symmetric.
Such problems appear in many applications. If $W$ vanishes, this model includes heat equation $(V(x) = 0)$ and the Fokker–Planck equation $(V(x) \neq 0$, see e.g. [47]). With interaction potentials, the equation can model nematic phase transition of rigid rod-like polymers [18], chemotaxis [46], and aggregation in biology (see [22, 28, 51] and references therein). For chemotaxis, a wide literature exists in relation to the Patlak-Keller-Segel system [29, 45], and for rod-like polymers, the Doi-Onsager equation [15, 18, 37, 41] is a well studied model.

Main properties of the solution to (5.1.1) are non-negativity, mass conservation and free energy dissipation, i.e.,

$$\rho_0(x) \geq 0 \implies \rho(x,t) \geq 0, \quad t > 0, \quad (5.1.2)$$

$$\int_\Omega \rho(x,t)dx = \int_\Omega \rho_0(x)dx, \quad t > 0, \quad (5.1.3)$$

$$\frac{dE(\rho)}{dt} = -\int_\Omega \rho|\nabla(\log(\rho) + V(x) + W*\rho)|^2dx = -I(\rho) \leq 0, \quad (5.1.4)$$

here the free energy associated to (5.1.1) is given by

$$E(\rho) = \int_\Omega \rho \log(\rho)dx + \int_\Omega V(x)\rho dx + \frac{1}{2} \int_\Omega \int_\Omega W(x - y)\rho(y)\rho(x)dydx. \quad (5.1.5)$$

This energy functional is a sum of internal energy, potential energy, and the interaction energy. The functional $I$ is referred to as the energy dissipation. The nice mathematical features (5.1.2)-(5.1.4) are crucial for the analytical study of (5.1.1), while free-energy dissipation inequality (5.1.4) is particularly important to understand the large time dynamics of solutions of (5.1.1)( see e.g., [6, 7, 39]). There have been many studies about the connection between the free energy, the Fokker-Planck equation, and optimal transportation in a continuous state space (see e.g., [4, 21, 27, 42, 52]). These properties are also desired to be preserved by the numerical methods, and they are particularly important in the accuracy of long time numerical simulation.

One way of obtaining a structure-preserving numerical scheme is the minimizing movement approximation (see [1] and the references therein), also named Jordan-Kinderlehrer-Otto (JKO)
scheme (Jordan et al. [27]), which is given by
\[
\rho^{n+1} = \text{argmin} \left\{ \frac{1}{2\tau} W^2(\rho^n, \rho) + E(\rho) \right\}
\]
Here, at each time step, the distance of the solution update acts as a regularization to the free energy. Yet such problems involving the Wasserstein distance \( W(\rho^n, \rho) \) are computationally demanding, see, e.g., [2, 9, 17, 38] for some recent advances.

The second way of obtaining a structure-preserving numerical scheme is by a direct discretization of (5.1.1) so that these solution properties are preserved at the discrete level. This way has gained increasing attention in recent years, some closely related works include [8, 32, 33, 34, 35, 36, 50]. In [32], second order implicit numerical schemes designed for linear (yet singular) Fokker-Planck equations satisfy all three solution properties without any time step restriction. In [35], the authors extended the idea in [32] to a system of Poisson-Nernst-Planck equations using the explicit time discretization. For a more general class of nonlinear nonlocal equations,
\[
\partial_t \rho = \nabla \cdot \left( \rho \nabla (H'(\rho) + V(x) + W * \rho) \right),
\]  
where \( H \) is a smooth convex function, a second order finite-volume method was constructed in [8], where positivity is enforced by using piecewise linear polynomials interpolating interface values. Structure preserving schemes based on the Chang-Cooper scheme [10] have been constructed in [44] to numerically solve nonlinear Fokker-Planck equations. Note that in [8, 35, 44] different time step restrictions are imposed in order to preserve the desired solution properties.

The construction of higher order schemes using the discontinuous Galerkin (DG) framework has recently been carried out for Fokker-Planck-type equations. We refer to [34] for energy satisfying DG schemes of arbitrary high order, and to [33] for a DG scheme of third order to satisfy the discrete maximum principle for linear Fokker-Planck equations. In [36], the authors designed free energy satisfying DG schemes of any high order for Poisson-Nernst-Planck equations, but positive cell averages are shown to propagate in time only for special cases. While in [50], a high order nodal DG method for (5.1.6) was constructed using \( k + 1 \) Gauss–Lobatto quadrature points for degree \( k \) polynomials in order to preserve both the energy dissipation and the solution positivity;
somehow degeneracy of accuracy in some cases was reported. Despite some well-known advantages of the DG method, structural properties of the above fully discrete DG schemes are verified under some CFL conditions. It would be interesting to explore some explicit-implicit strategies for DG schemes.

In this paper we extend the idea in [32] to construct explicit-implicit schemes which are proven to preserve three main properties of (5.1.1) without a strict restriction on time steps. This therefore has improved upon the work [35]. Our main results include the scheme formulation, proofs of mass conservation, solution non-negativity, and the discrete free-energy dissipation law for both semi-discrete and fully discrete methods. In particular, the fully-discrete scheme (first order in time) is shown to satisfy three desired properties without strict restriction on time steps, in both one and two dimensional cases with nonuniform meshes. For the fully second order (in both time and space) scheme, we design a local scaling limiter to restore solution positivity, the limiter is build upon the one introduced in [31] and shown to preserve the second order accuracy.

More precisely, our scheme construction is based on a reformulation

$$
\partial_t \rho = \nabla \cdot \left( M \nabla \left( \frac{\rho}{M} \right) \right),
$$

(5.1.7)

with $M = e^{-V(x) - W*\rho}$, motivated by the fact that the equilibrium solutions of (5.1.1) may be expressed as $\rho = C e^{-V(x) - W*\rho}$. For linear Fokker-Planck equations, such reformulation with $M = e^{-V(x)}$ (so called non-logarithmic Landau form) has been used in [32], as well as in earlier works (see e.g., [5]). We note that for the general nonlinear nonlocal model (5.1.6), our scheme construction remains valid if we take $M = \rho e^{-H'(\rho) - V(x) - W*\rho}$ in the reformulation (5.1.7).

The advantage of formulation (5.1.7) can be seen from both spatial and temporal discretization. The symmetric spatial discretization of the one-dimensional version of (5.1.7) yields the semi-discrete scheme

$$
h_j^d \frac{d}{dt} \rho_j = h_{j+1/2}^{d-1} M_{j+1/2} \left( \frac{\rho_{j+1}}{M_{j+1}} - \frac{\rho_j}{M_j} \right) - h_{j-1/2}^{d-1} M_{j-1/2} \left( \frac{\rho_j}{M_j} - \frac{\rho_{j-1}}{M_{j-1}} \right),
$$

(5.1.8)
in which the evaluation of $M$ at cell interfaces $\{x_{j+1/2}\}$ and cell centers $\{x_j\}$ is easily available as defined in (5.2.4). Here $\rho_j$ approximates the cell average of $\rho(x,t)$ on $j$-th computational cell $[x_{j-1/2}, x_{j+1/2}]$ of size $h_j$, and $h_{j+1/2} = (h_j + h_{j+1})/2$.

For time discretization of (5.1.8), we adopt an implicit-explicit approach to obtain

$$h_j \frac{\rho_j^{n+1} - \rho_j^n}{\tau} = h_{j+1/2} M_{j+1}^{n+1/2} \left( \frac{\rho_j^{n+1}}{M_j^{n+1}} - \frac{\rho_j^{n+1}}{M_j^{n}} \right) - h_{j-1/2} M_j^{n-1/2} \left( \frac{\rho_j^{n+1}}{M_j^{n}} - \frac{\rho_j^{n+1}}{M_j^{n-1}} \right), \quad (5.1.9)$$

here $\rho_j^n$ approximates $\rho_j(t)$ at time $t = n\tau$, see (5.3.1). This scheme is easy to implement, and is shown to preserve all three desired properties without a strict time step restriction. However, the scheme (5.1.9) is only first order in time. We further propose a fully second order scheme:

$$h_j \frac{\rho_j^* - \rho_j^n}{\tau/2} = h_{j+1/2} M_{j+1}^* M_{j+1/2} \left( \frac{\rho_j^*}{M_j^*} - \frac{\rho_j^n}{M_j^n} \right) - h_{j-1/2} M_j^{*} M_{j-1/2} \left( \frac{\rho_j^*}{M_j^*} - \frac{\rho_j^{n-1}}{M_j^{n-1}} \right), \quad (5.1.10)$$

$$\rho_j^{n+1} = 2\rho_j^* - \rho_j^n,$$

based on the predictor-corrector methodology, where $M_j^*$ and $M_{j+1/2}^*$ are given in (5.5.1). This scheme is second order in both time and space, and it preserves solution positivity for small time steps. For large time steps, we use a local scaling limiter to restore the solution positivity.

Although we derive the schemes for the model equation (5.1.1), the methods can be applied to a larger class of PDE problems of drift-diffusion type; see [30].

Finally, we point out that the energy stability has always played an essential role in the accuracy of long time simulations of a gradient flow. The related works could also be found for other physical models such as the phase field equations [11, 49, 53], the thin film growth equations [12, 54], and the Cahn-Hillard models [14, 19, 20, 23, 24, 48, 55, 56]. In the case of the Cahn-Hilliard equation with a singular potential such as the Flory-Huggins potential, which is defined only when the solution lies strictly within an interval, we refer to [14, 19] for theoretical justification of the positivity-preserving property of some finite difference schemes. Different from the present work, the key ingredient used in [14, 19] is the singular nature of the logarithmic term around the boundary values which prevents the numerical solution from reaching these singular values.

The rest of the paper is organized as follows. In section 2, we present a semi-discrete scheme for one dimensional problems. Theoretical analysis of three properties is provided. In section 3,
we present fully discrete implicit-explicit schemes for one dimensional case and prove the desired properties. Section 4 is devoted to numerical schemes for two dimensional problems. In section 5, we extend the scheme to a fully second order (in both time and space) scheme, a mass conserving local limiter is also introduced to restore solution positivity. Numerical examples for one and two dimensional problems are presented in section 6. Finally, concluding remarks are given in section 7.

5.2 Numerical Method: one dimensional case

We begin with

\[
\begin{align*}
\partial_t \rho &= \partial_x (\partial_x \rho + \rho \partial_x (V(x) + W * \rho)), \quad x \in \Omega, \quad t > 0, \\
\rho(x, 0) &= \rho_0(x), \quad x \in \Omega, \\
\partial_x \rho + \rho \partial_x (V(x) + W * \rho) &= 0, \quad x \in \partial \Omega, \quad t > 0.
\end{align*}
\]  

(5.2.1)

and reformulate (5.2.1) as

\[
\begin{align*}
\partial_t \rho &= \partial_x (M \partial_x (\rho/M)), \quad x \in \Omega, \quad t > 0, \\
\rho(x, 0) &= \rho_0(x), \quad x \in \Omega, \\
M \partial_x (\rho/M) &= 0, \quad x \in \partial \Omega, \quad t > 0.
\end{align*}
\]  

(5.2.2)

here \( M = e^{-V(x)-W*\rho} \). We propose a finite volume scheme for (5.2.2) over the interval \( \Omega = [a, b] \).

For a given positive integer \( N \), we partition domain \( \Omega \) into computational cells \( I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \) with mesh size \( h_j = |I_j| \) and cell center at \( x_j = x_{j-\frac{1}{2}} + \frac{1}{2} h_j, \quad j \in \{1, 2, \cdots, N\} \), we set \( h_{j+1/2} = (h_j + h_{j+1})/2 \).

5.2.1 Semi-discrete scheme

We integrate on each computational cell \( I_j \) to obtain

\[
\frac{d}{dt} \int_{I_j} \rho(x, t) dx = M \partial_x (\rho/M)|_{x_{j+1/2}} - M \partial_x (\rho/M)|_{x_{j-1/2}}.
\]
Let $\rho(t) = \{\rho_1, \cdots, \rho_N\}$ be the numerical solution approximating all cell averages and $C_{j+1/2}$ be an approximation to $M\partial_x(\rho/M)|_{x_{j+1/2}}$, then one has the following semi-discrete scheme,

$$\frac{d}{dt}\rho_j(t) = \frac{C_{j+1/2} - C_{j-1/2}}{h_j}, \quad j = 1, 2, \cdots, N,$$  \hspace{1cm} (5.2.3)

we define

$$C_{j+1/2} = \frac{M_{j+1/2}}{h_{j+1/2}} \left( \frac{\rho_{j+1}}{M_{j+1}} - \frac{\rho_j}{M_j} \right) \quad \text{for} \quad j = 1, 2, \cdots, N - 1,$$

$$C_{1/2} = 0, \quad C_{N+1/2} = 0.$$

Here $M_{j+1/2} = Q_1(x_{j+1/2}, \rho)$ and $M_j = Q_1(x_j, \rho)$ with

$$Q_1(x, v) = e^{-V(x) - \sum_{i=1}^{N} h_i W(x_i - x)v_i}, \quad \text{for} \quad x \in \mathbb{R}, \ v \in \mathbb{R}^N.$$  \hspace{1cm} (5.2.4)

Note that the zero flux boundary conditions have been weakly enforced.

### 5.2.2 Scheme properties

We investigate three desired properties for this semi-discrete scheme. For the energy dissipation property, we define a semi-discrete version of the free energy (5.1.5) as

$$E_h(t) = \sum_{j=1}^{N} h_j \left( \rho_j \log(\rho_j) + V_j \rho_j + \frac{1}{2} g_j \rho_j \right),$$  \hspace{1cm} (5.2.5)

here $g_j = \sum_{i=1}^{N} h_i W(x_i - x_j)\rho_i$ is a second order approximation of the convolution $(W * \rho)(x_j)$.

The following theorem states that the semi-discrete scheme (5.2.3) is conservative, positive, and energy dissipating.

**Theorem 5.2.1.** The semi-discrete scheme (5.2.3) satisfies the following properties:

1. Conservation of mass: for any $t > 0$ we have

$$\sum_{j=1}^{N} h_j \rho_j(t) = \sum_{j=1}^{N} h_j \rho_j(0).$$  \hspace{1cm} (5.2.6)

2. Positivity preserving: if $\rho_j(0) \geq 0$ for all $j \in \{1, \cdots, N\}$, then $\rho_j(t) \geq 0$ for any $t > 0$.

3. Energy dissipation: $\frac{dE_h(t)}{dt} \leq -I_h$, where

$$I_h = \sum_{j=1}^{N-1} C_{j+1/2} \left( \log \left( \frac{\rho_{j+1}}{M_{j+1}} \right) - \log \left( \frac{\rho_j}{M_j} \right) \right) \geq 0.$$  \hspace{1cm} (5.2.7)
Proof. (1) Summing all equations in (5.2.3), we have
\[ \frac{d}{dt} \sum_{j=1}^{N} h_j \rho_j(t) = \sum_{j=1}^{N} \frac{d}{dt} h_j \rho_j(t) = 0, \]
Therefore (5.2.6) holds true for any \( t > 0 \).

(2) Let \( \vec{F}(\vec{\rho}) \) be the vector field defined by the right hand side of (5.2.3), then
\[ \frac{d}{dt} \vec{\rho} = \vec{F}(\vec{\rho}). \tag{5.2.8} \]
Note that the hyperplane \( \Sigma = \{ \vec{\rho} : \sum_{j=1}^{N} h_j \rho_j = \sum_{j=1}^{N} h_j \rho_j(0) \} \) is an invariant region of (5.2.8).
We define a closed set \( \Sigma_1 \) on this hyperplane by
\[ \Sigma_1 = \left\{ \vec{\rho} : \rho_j \geq 0, j = 1, 2, \ldots, N, \text{and} \sum_{j=1}^{N} h_j \rho_j = \sum_{j=1}^{N} h_j \rho_j(0) \right\}. \]
It suffices to show that \( \Sigma_1 \) is invariant under system (5.2.8). This is the case if the vector field \( \vec{F}(\vec{\rho}) \)
strictly points to interior of \( \Sigma_1 \) on its boundary \( \partial \Sigma_1 \): i.e.,
\[ \vec{F}(\vec{\rho}) \cdot \vec{v} < 0, \]
where \( \vec{v} \) is outward normal vector on any part of \( \partial \Sigma_1 \).
A direct calculation using (5.2.3) gives
\[ \begin{align*}
\vec{F}(\vec{\rho}) \cdot \vec{v} &= \sum_{j=1}^{N-1} \frac{v_j}{h_j} C_{j+1/2} - \sum_{j=2}^{N} \frac{v_j}{h_j} C_{j-1/2} \\
&= - \sum_{j=1}^{N-1} \left( \frac{v_{j+1}}{h_{j+1}} - \frac{v_j}{h_j} \right) C_{j+1/2}.
\end{align*} \tag{5.2.9} \]
For each \( \vec{\mu} \in \partial \Sigma_1 \), we define the set \( S = \{ j : 1 \leq j \leq N \text{ and } \mu_j = 0 \} \), then the outward normal
vector at \( \vec{\mu} \) has the form
\[ \vec{v} = (v_1, v_2, \ldots, v_N)^T \text{ with } v_i = \begin{cases} -\alpha_i, & i \in S, \\ 0, & i \notin S, \end{cases} \]
and \( \alpha_i > 0 \) if \( i \in S \).
Note that if \( j, j + 1 \in S \), then \( \rho_j = \rho_{j+1} = 0 \) implies \( C_{j+1/2} = 0 \); if \( j, j + 1 \notin S \), then \( v_{j+1} = v_j = 0 \). Therefore nonzero terms in (5.2.9) are those with \( j \in S, j + 1 \notin S \) or \( j \notin S, j + 1 \in S \). Hence

\[
\vec{F}(\vec{\rho}) \cdot \vec{v} = - \sum_{j \in S, j+1 \notin S} \frac{\alpha_j}{h_j} \frac{M_{j+1/2}}{M_{j+1}} \rho_{j+1} - \sum_{j \notin S, j+1 \in S} \frac{\alpha_{j+1}}{h_{j+1}} \frac{M_{j+1/2}}{M_{j+1}} \rho_j < 0.
\]

Therefore \( \Sigma_1 \) is an invariant region of (5.2.3), this completes the proof of (2).

(3) From the fact that \( W(x) = W(-x) \), it follows

\[
\frac{d}{dt} \sum_{j=1}^{N} \frac{h_j}{2} g_j \rho_j = \sum_{j=1}^{N} h_j g_j \frac{d\rho_j}{dt}.
\]

Differentiating the discrete free energy (5.2.5) with respect to time and using (5.2.10) we obtain

\[
\frac{dE_h(t)}{dt} = \sum_{j=1}^{N} \left[ \log(\rho_j) + 1 + V_j + g_j \right] h_j \frac{d\rho_j}{dt} = \sum_{j=1}^{N} \left( \log \left( \frac{\rho_j}{M_j} \right) + 1 \right) \left( C_{j+1/2} - C_{j-1/2} \right)
\]

\[
= - \sum_{j=1}^{N-1} C_{j+1/2} \left( \log \left( \frac{\rho_{j+1}}{M_{j+1}} \right) - \log \left( \frac{\rho_j}{M_j} \right) \right) = -I_h \leq 0.
\]

Note that

\[
I_h = \sum_{j=1}^{N-1} C_{j+1/2} \left( \log \left( \frac{\rho_{j+1}}{M_{j+1}} \right) - \log \left( \frac{\rho_j}{M_j} \right) \right) = \sum_{j=1}^{N-1} \frac{1}{h_{j+1/2}} \frac{M_{j+1/2}}{M_{j+1}} \left( \frac{\rho_{j+1}}{M_{j+1}} - \frac{\rho_j}{M_j} \right) \left( \log \left( \frac{\rho_{j+1}}{M_{j+1}} \right) - \log \left( \frac{\rho_j}{M_j} \right) \right)
\]

and \((x - y)(\log x - \log y) \geq 0\) for \( x, y \in \mathbb{R}^+ \), so we have \( I_h \geq 0 \).

\( \square \)

### 5.3 Fully discrete scheme

For time discretization of (5.2.3), we use an implicit-explicit time discretization in order to construct an easy to implement yet stable numerical scheme without time step restriction.
5.3.1 Scheme formulation and algorithm

Let $\tau$ be time step and $\rho^n_j$ be the numerical solution at $t_n = n\tau$ to approximate $\rho_j(t_n)$. From given $\rho^n_j$, $j = 1, 2, \cdots, N$, we update to get $\rho^{n+1}_j$ by

$$\frac{\rho^{n+1}_j - \rho^n_j}{\tau} = \frac{C^{n,*}_{j+1/2} - C^{n,*}_{j-1/2}}{h_j}, \quad j = 1, 2, \cdots, N. \quad (5.3.1)$$

with

$$C^{n,*}_{j+1/2} = \frac{M^n_{j+1/2}}{h_{j+1/2}} \left( \frac{\rho^{n+1}_{j+1}}{M^n_{j+1}} - \frac{\rho^{n+1}_j}{M^n_j} \right) \text{ for } j = 1, 2, \cdots, N-1,$$  

$$C^{n,*}_{1/2} = C^{n,*}_{N+1/2} = 0,$$

where $M^n_{j+1/2} = Q_1(x_{j+1/2}, \rho^n)$ and $M^n_j = Q_1(x_j, \rho^n)$. The initial data is chosen by

$$\rho^n_0 = \frac{1}{h_j} \int_{I_j} \rho_0(x) dx, \quad j = 1, 2, \cdots, N. \quad (5.3.3)$$

5.3.2 Scheme properties

Define a fully discrete version $E^n_h$ of the free energy (5.1.5) as

$$E^n_h = \sum_{j=1}^N h_j \left( \rho^n_j \log(\rho^n_j) + V_j \rho^n_j + \frac{1}{2} g^n_j \rho^n_j \right), \quad (5.3.4)$$

where $g^n_j = \sum_{i=1}^N h_i W(x_i - x_j) \rho^n_i$.

The following theorem states that the three desired properties are preserved by the scheme (5.3.1) without strict time step restriction.

Theorem 5.3.1. The fully discrete scheme (5.3.1) has the following properties:

1. Conservation of mass:

$$\sum_{j=1}^N h_j \rho^n_j = \int_\Omega \rho_0(x) dx \quad \text{for } n \geq 1. \quad (5.3.5)$$

2. Positivity preserving: If $\rho^n_j \geq 0$ for all $j = 1, \cdots, N$, then

$$\rho^{n+1}_j \geq 0, \quad j = 1, \cdots, N.$$
(3) Energy dissipation: there exists $\tau^* > 0$ such that if $\tau \in (0, \tau^*)$, then

$$E_h^{n+1} - E_h^n \leq -\frac{\tau}{2} I_h^n,$$

where

$$I_h^n = \sum_{j=1}^{N-1} C_{j+1/2} \left( \log \left( \frac{\rho_j^{n+1}}{M_{j+1}} \right) - \log \left( \frac{\rho_j^{n+1}}{M_j} \right) \right) \geq 0.$$

\textbf{Proof.} Set $G_j^{n,*} = \rho_j^{n+1}/M_j^n$ and $\lambda_{j+1/2} = \tau/h_{j+1/2}$, so the fully discrete scheme (5.3.1) can be rewritten into the following linear system:

$$
\begin{align*}
  h_1 \rho_1^n &= (h_1 M_1^n + \lambda_{1+1/2} M_{1+1/2}^n) G_1^{n,*} - \lambda_{1+1/2} M_{1+1/2}^n G_2^{n,*}, \\
  h_j \rho_j^n &= -\lambda_{j-1/2} M_{j-1/2}^n G_{j-1}^{n,*} + (h_j M_j^n + \lambda_{j-1/2} M_{j-1/2}^n + \lambda_{j-1/2} M_{j+1/2}^n) G_j^{n,*} \\
  &\quad - \lambda_{j+1/2} M_{j+1/2}^n G_{j+1}^{n,*} \quad j = 2, 3, \ldots, N - 1, \\
  h_N \rho_N^n &= -\lambda_{N-1/2} M_{N-1/2}^n G_{N-1}^{n,*} + (h_N M_N^n + \lambda_{N-1/2} M_{N-1/2}^n) G_N^{n,*}.
\end{align*}
$$

(5.3.7)

Note that the coefficient matrix of linear system (5.3.7) is strictly diagonally dominant, therefore (5.3.7) has a unique solution for whatever $\tau$ a priori chosen so does (5.3.1) because $\rho_j^{n+1} = G_j^{n,*} M_j^n$.

(1) (5.3.5) follows from adding all equations in system (5.3.7) and using (5.3.3).

(2) Since $\rho_j^{n+1} = M_j^n G_j^{n,*}$ and $M_j^n > 0$, it suffices to prove that

$$G_j^{n,*} = \min_{1 \leq j \leq N} \{G_j^{n,*}\} \geq 0.$$

Assume $1 < i < N$, from $i$-th equation of (5.3.7) we have

$$
\begin{align*}
  h_i \rho_i^n &= -\lambda_{i-1/2} M_{i-1/2}^n G_{i-1}^{n,*} + (h_i M_i^n + \lambda_{i-1/2} M_{i-1/2}^n + \lambda_{i+1/2} M_{i+1/2}^n) G_i^{n,*} - \lambda_{i+1/2} M_{i+1/2}^n G_{i+1}^{n,*} \\
  &\leq -\lambda_{i-1/2} M_{i-1/2}^n G_{i-1}^{n,*} + (h_i M_i^n + \lambda_{i-1/2} M_{i-1/2}^n + \lambda_{i+1/2} M_{i+1/2}^n) G_i^{n,*} - \lambda_{i+1/2} M_{i+1/2}^n G_{i+1}^{n,*} \\
  &= h_i M_i^n G_i^{n,*}.
\end{align*}
$$

Thus $G_i^{n,*} \geq \frac{\rho_i^n}{M_i^n} \geq 0$. A similar argument applies if $i = 1$ or $i = N$. 


(3) A direct calculation using (5.3.4) gives

\[ E_{n+1}^h - E_n^h = \sum_{j=1}^{N} h_j \left( \rho_j^{n+1} \log(\rho_j^{n+1}) - \rho_j^n \log(\rho_j^n) \right) + V_j \rho_j^{n+1} - V_j \rho_j^n + \frac{1}{2} g_j^{n+1} \rho_j^{n+1} - \frac{1}{2} g_j^n \rho_j^n \]

\[ = \sum_{j=1}^{N} h_j \left( (\rho_j^{n+1} - \rho_j^n) \log(\rho_j^{n+1}) + (\rho_j^{n+1} - \rho_j^n) V_j + (\rho_j^{n+1} - \rho_j^n) g_j^n \right) \]

\[ + \frac{1}{2} g_j^n \rho_j^n - g_j^n \rho_j^{n+1} + \frac{1}{2} g_j^{n+1} \rho_j^n + \rho_j^n \log(\frac{\rho_j^{n+1}}{\rho_j^n}) \]

\[ \leq \sum_{j=1}^{N} h_j \left( (\rho_j^{n+1} - \rho_j^n) \log(G_{j,n}^*) + \frac{1}{2} g_j^n \rho_j^n - g_j^n \rho_j^{n+1} + \frac{1}{2} g_j^{n+1} \rho_j^n \right), \]

here we have used \( \rho_j^n \log(\frac{\rho_j^{n+1}}{\rho_j^n}) \leq \rho_j^n (\frac{\rho_j^{n+1}}{\rho_j^n} - 1) \) and mass conservation \( \sum_{j=1}^{N} h_j (\rho_j^{n+1} - \rho_j^n) = 0 \). We proceed with

\[ \tau \sum_{j=1}^{N} \left( \frac{h_j \rho_j^{n+1} - h_j \rho_j^n}{\tau} \right) \log(G_{j,n}^*) = \tau \sum_{j=1}^{N} \left( \log(G_{j,n}^*)(h_j^{n+1} M_j^{n+1/2}(G_{j,n}^* - G_{j,n}^*) - h_j^{n+1} M_j^{n+1/2}(G_{j,n}^* - G_{j,n}^*)) \right) \]

\[ = -\tau \sum_{j=1}^{N} h_j^{n+1} M_j^{n+1/2}(G_{j,n}^* - G_{j,n}^*) \log(G_{j,n}^* - G_{j,n}^*) \]

\[ = -\tau I_h^n \leq 0. \]

(5.3.8)

Here the sign of \( I_h^n \) is implied by the monotonicity of the logarithmic function.

It remains to find a sufficient condition on time step \( \tau \) so that

\[ \sum_{j=1}^{N} h_j \left( \frac{1}{2} g_j^n \rho_j^n - g_j^n \rho_j^{n+1} + \frac{1}{2} g_j^{n+1} \rho_j^{n+1} \right) \leq -\frac{\tau}{2} \sum_{j=1}^{N} \left( \frac{h_j \rho_j^{n+1} - h_j \rho_j^n}{\tau} \right) \log(G_{j,n}^*). \]

(5.3.9)
From $\sum_{j=1}^N h_j g_j^n \rho_j^{n+1} = \sum_{j=1}^N h_j g_j^{n+1} \rho_j^n$ it follows that

$$\sum_{j=1}^N h_j (\frac{1}{2} g_j^n \rho_j^n - g_j^n \rho_j^{n+1} + \frac{1}{2} g_j^{n+1} \rho_j^n) = \frac{1}{2} \sum_{j=1}^N h_j (g_j^{n+1} - g_j^n)(\rho_j^{n+1} - \rho_j^n)$$

$$= \frac{1}{2} \sum_{j=1}^N h_j \sum_{i=1}^N h_i W(x_i - x_j)(\rho_i^{n+1} - \rho_i^n)(\rho_j^{n+1} - \rho_j^n)$$

$$\leq \frac{||W||_\infty}{2} \sum_{j=1}^N h_j \sum_{i=1}^N h_i |\rho_i^{n+1} - \rho_i^n||\rho_j^{n+1} - \rho_j^n|$$

$$\leq \frac{||W||_\infty}{2} (b-a) \tau^2 \sum_{j=1}^N h_j \left(\frac{\rho_j^{n+1} - \rho_j^n}{\tau}\right)^2,$$

where we have used the Cauchy-Schwarz inequality and $b - a = \sum_{j=1}^N h_j$. Let $\vec{\xi}, \vec{\eta} \in \mathbb{R}^N$ be vectors defined as $\vec{\xi}_j = \sqrt{h_j (\rho_j^{n+1} - \rho_j^n)/\tau}$, $\vec{\eta}_j = \sqrt{h_j \log G_j^{n,*}}$, then (5.3.9) is satisfied if

$$\frac{||W||_\infty (b-a) \tau^2}{2} |\vec{\xi}|^2 + \frac{\tau^2}{2} \vec{\xi} \cdot \vec{\eta} \leq 0.$$

We claim that

$$\vec{\xi} \cdot \vec{\eta} = 0 \quad \text{if and only if} \quad \vec{\xi} = 0. \tag{5.3.10}$$

Therefore

$$0 < c_0 \leq -\frac{\vec{\xi} \cdot \vec{\eta}}{|\vec{\xi}|^2} \leq \frac{||\eta||}{|\vec{\xi}|} \quad \text{for} \quad \vec{\xi} \neq 0,$$

where $c_0$ may depend on numerical solutions at $t_n$ and $t_{n+1}$. We thus obtain (5.3.9) by taking

$$\tau \leq \tau^* = \frac{c_0}{||W||_\infty (b-a)}.$$

Finally, we verify claim (5.3.10). If $\vec{\xi} \cdot \vec{\eta} = 0$, then from (5.3.8) we have

$$0 = \vec{\xi} \cdot \vec{\eta} = -\sum_{j=1}^{N-1} h_{j+1/2}^{-1} M_{j+1/2}^n (\log G_{j+1}^{n,*} - \log G_j^{n,*})(G_{j+1}^{n,*} - G_j^{n,*}) \leq 0,$$

therefore we must have $G_j^{n,*} = constant$ for all $j \in \{1, 2, \cdots, N\}$. This when inserted into scheme (5.3.1) leads to

$$\rho_j^{n+1} = \rho_j^n \quad \text{for all} \quad j \in \{1, 2, \cdots, N\},$$

thus $\vec{\xi} = 0$. \qed
Remark 5.3.1. Though $\tau^*$ in the above proof is not explicitly given, it is expected to be $O(1)$ since $-\frac{\tilde{\xi} \cdot \tilde{\eta}}{\| \tilde{\xi} \|^2}$ tends to a quantity of size $O(1)$ as meshes are refined. More precisely, we have

$$-\frac{\tilde{\xi} \cdot \tilde{\eta}}{\| \tilde{\xi} \|^2} \to \| \partial_t E(\rho) \|,$$

which is valid before reaching the steady state. This remark applies to Theorem 5.4.2 as well. Numerically energy dissipation was observed for large time steps relative to the spatial mesh sizes, see Example 5.6.3. Furthermore, a more precise bound for $\tau^*$ can be obtained with additional structural conditions on $W$; see the appendix.

Remark 5.3.2. One could take the Euler forward time discretization to obtain an explicit scheme: From $\rho^n_j, j = 1, 2, \cdots, N$, update to get $\rho^{n+1}_j$ by

$$\frac{\rho^{n+1}_j - \rho^n_j}{\tau} = \frac{C^n_{j+1/2} - C^n_{j-1/2}}{h_j}, \ j = 1, 2, \cdots, N.$$

where

$$C^n_{j+1/2} = \frac{M^n_{j+1/2}}{h_{j+1/2}} \left( \frac{\rho^n_{j+1}}{M^n_{j+1}} - \frac{\rho^n_j}{M^n_j} \right) \text{ for } j = 1, 2, \cdots, N - 1,$$

$$C^n_{1/2} = C^n_{N+1/2} = 0,$$

with $M^n_{j+1/2} = Q_1(x_{j+1/2}, \rho^n)$ and $M^n_j = Q_1(x_j, \rho^n)$. One can show that the positivity preserving property is still met yet under a CFL condition like $\tau \leq \gamma h^2$.

5.3.3 Discussion on error estimates

It is desirable to obtain global-in-time error estimates by using the established energy dissipation law (5.3.6). But this appears rather difficult for the nonlinear term in the scheme fits more for the positivity-preserving property than the energy dissipation property. This said, we can obtain the local-in-time error estimate. The analysis includes both the truncation error estimate and the energy estimate for the error equation, yet estimates of the nonlinear terms are much more involved. We therefore only state the main result for (5.3.1), leaving detailed analysis to a separate publication.
Theorem 5.3.2. Assume that both \( W \) and \( V \) are Lipschitz continuous. Given smooth initial data \( \rho_0(x) \), suppose the unique, smooth solution for (5.1.1) is given by \( \rho(x, t) \) on \( \Omega \times [0, T] \) for some \( T \) finite, and the numerical solution for (5.3.1) is given by \( \rho^n_j \) with \( \rho^n_j = \frac{1}{h_j} \int_{I_j} \rho_0(x) \, dx \). Then, provided \( \tau \) and \( h = \max_j h_j \) are sufficiently small, for all positive integers \( n \) such that \( n \tau \leq T \), we have

\[
\sum_{j=1}^{N} |\rho(x_j, t^n) - \rho^n_j|^2 h_j \leq C(\tau + h^2)^2.
\]

where \( C > 0 \) is independent of \( h \) and \( \tau \).

To see the complex nature of estimates in handing nonlinear terms, we refer to [23, 24] for the local-in-time error estimates of finite difference schemes to the nonlocal Cahn-Hilliard equation.

5.4 Numerical Method: two dimensional Case

In this section, we extend our method to multi-dimensional problems. For simplicity, we only present schemes for the two dimensional initial value problem,

\[
\begin{aligned}
\partial_t \rho &= \nabla \cdot (\nabla \rho + \rho \nabla (V(x, y) + W \ast \rho)), & (x, y) &\in \Omega \subset \mathbb{R}^2, \quad t > 0, \\
\rho(x, y, 0) &= \rho_0(x, y), & (x, y) &\in \Omega,
\end{aligned}
\]

(5.4.1)
on a rectangular domain \( \Omega = [a, b] \times [c, d] \) subject to zero flux boundary conditions.

For given positive integers \( N_x, N_y \), we partition \( \Omega \) by a Cartesian mesh with computational cells

\[
I_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}],
\]

where \( i \in \{1, 2, \cdots, N_x\}, j \in \{1, 2, \cdots, N_y\} \). The mesh size is \( |I_{i,j}| = h_i^x h_j^y \) with the cell center at

\[
(x_i, y_j) = (x_{i-1/2} + \frac{1}{2} h_i^x, y_{j-1/2} + \frac{1}{2} h_j^y),
\]

we set \( h_{i+1/2}^x = (h_i^x + h_{i+1}^x)/2, h_{j+1/2}^y = (h_j^y + h_{j+1}^y)/2 \).

5.4.1 Semi-discrete scheme

Let \( \rho(t) = \{\rho_{i,j}\} \) be the numerical solution, then dimension by dimension spatial discretization of

\[
\partial_t \rho = \nabla \cdot \left( M \nabla \left( \frac{\rho}{M} \right) \right), \quad \text{with} \quad M = e^{-V(x,y)-W\ast\rho},
\]
yields the following semi-discrete scheme
\[ \frac{d}{dt} \rho_{i,j} = \frac{C_{i+1/2,j} - C_{i-1/2,j}}{h_i^x} + \frac{C_{i,j+1/2} - C_{i,j-1/2}}{h_j^y}, \quad (5.4.2) \]

where
\[ C_{i+1/2,j} = \frac{M_{i+1/2,j}}{h_i^x} \left( \rho_{i+1,j} - \frac{\rho_{i,j}}{M_{i,j}} \right), \quad i = 1, \ldots, N_x - 1, \quad j = 1, \ldots, N_y, \]
\[ C_{i,j+1/2} = \frac{M_{i,j+1/2}}{h_j^y} \left( \rho_{i,j+1} - \frac{\rho_{i,j}}{M_{i,j}} \right), \quad i = 1, \ldots, N_x, \quad j = 1, \ldots, N_y - 1, \]
\[ C_{1/2,j} = C_{N_x+1/2,j} = C_{i,1/2} = C_{i,N_y+1/2} = 0, \quad i = 1, \ldots, N_x, \quad j = 1, \ldots, N_y, \]

with \( M_{i+1/2,j} = Q_2(x_{i+1/2}, y_j, \rho), \) \( M_{i,j+1/2} = Q_2(x_i, y_{j+1/2}, \rho), \) and \( M_{i,j} = Q_2(x_i, y_j, \rho). \) Where
\[ Q_2(x, y, v) = e^{-V(x, y) - \sum_{k=1}^{N_y} h_k^x h_k^y w(x_k - x, y_k - y) v_k}, \quad \text{for} \ x, y \in \mathbb{R}, v \in \mathbb{R}^{N_x \times N_y}. \quad (5.4.3) \]

Let
\[ E_h(t) = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y \left( \rho_{i,j} \log(\rho_{i,j}) + V_{i,j} \rho_{i,j} + \frac{1}{2} g_{i,j} \rho_{i,j} \right) + 1, \]
be an approximation of the energy functional (5.1.5), with
\[ g_{i,j} = \sum_{k=1}^{N_y} \sum_{l=1}^{N_y} h_k^x h_l^y w(x_k - x_i, y_l - y_j) \rho_{k,l}. \]

The following theorem states that the semi-discrete scheme (5.4.2) is conservative, positive, and energy dissipating.

**Theorem 5.4.1.** The semi-discrete scheme (5.4.2) satisfies the following properties:

1. **Conservation of mass:** for any \( t > 0, \)
\[ \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y \rho_{i,j}(t) = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y \rho_{i,j}(0). \]

2. **Positivity preserving:** if \( \rho_{i,j}(0) \geq 0 \) for all \( i \in \{1, \ldots, N_x\}, j \in \{1, \ldots, N_y\}, \) then \( \rho_{i,j}(t) \geq 0 \) for any \( t > 0. \)

3. **Energy dissipation:** \( \frac{d E_h(t)}{dt} \leq -I_h, \)
\[ I_h = \sum_{j=1}^{N_y} \sum_{i=1}^{N_y-1} h_j^y C_{i+1/2,j} \left( \log \left( \frac{\rho_{i+1,j}}{M_{i+1,j}} \right) - \log \left( \frac{\rho_{i,j}}{M_{i,j}} \right) \right) \]
\[ + \sum_{i=1}^{N_x} \sum_{j=1}^{N_x-1} h_i^x C_{i,j+1/2} \left( \log \left( \frac{\rho_{i,j+1}}{M_{i,j+1}} \right) - \log \left( \frac{\rho_{i,j}}{M_{i,j}} \right) \right) \geq 0. \]

**Proof.** The proof is similar to that of Theorem 5.2.1, details are therefore omitted.
5.4.2 Fully discrete scheme

Let $\rho_{i,j}^n$ approximate $\rho_{i,j}(t_n)$, then (5.4.2) gives the following fully discrete scheme,

$$
\frac{\rho_{i,j}^{n+1} - \rho_{i,j}^n}{\tau} = \frac{C_{i+1/2,j}^{n,*} - C_{i-1/2,j}^{n,*}}{h_i^x} + \frac{C_{i,j+1/2}^{n,*} - C_{i,j-1/2}^{n,*}}{h_j^y},
$$

(5.4.4)

where

$$
C_{i+1/2,j}^{n,*} = \frac{M_{i+1/2,j}^{n+1}}{h_{i+1/2}^x} \left( \frac{\rho_{i+1,j}^{n+1}}{M_{i,j}^{n+1}} - \frac{\rho_{i,j}^{n+1}}{M_{i,j}^{n+1}} \right), \quad i = 1, \ldots, N_x - 1, j = 1, \ldots, N_y,
$$

$$
C_{i,j+1/2}^{n,*} = \frac{M_{i,j+1/2}^{n+1}}{h_{j+1/2}^y} \left( \frac{\rho_{i,j+1}^{n+1}}{M_{i,j}^{n+1}} - \frac{\rho_{i,j}^{n+1}}{M_{i,j}^{n+1}} \right), \quad i = 1, \ldots, N_x, j = 1, \ldots, N_y - 1,
$$

$$
C_{1/2,j}^{n,*} = C_{N_x+1/2,j}^{n,*} = C_{i,1/2}^{n,*} = C_{i,N_y+1/2}^{n,*} = 0, \quad i = 1, \ldots, N_x, j = 1, \ldots, N_y,
$$

with $M_{i+1/2,j}^n = Q_2(x_{i+1/2}, y_j, \rho^n)$, $M_{i,j+1/2}^n = Q_2(x_i, y_{j+1/2}, \rho^n)$, and $M_{i,j}^n = Q_2(x_i, y_j, \rho^n)$.

The initial data is chosen as

$$
\rho_{i,j}^0 = \frac{1}{|I_{i,j}|} \int_{I_{i,j}} \rho_0(x, y) dx dy.
$$

(5.4.5)

In 2D case, a discrete version of energy (5.1.5) may be defined as

$$
E_h^n = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y \left( \rho_{i,j}^n \log(\rho_{i,j}^n) + V_{i,j} \rho_{i,j}^n + \frac{1}{2} g_{i,j}^n \rho_{i,j}^n \right),
$$

(5.4.6)

where

$$
g_{i,j}^n = \sum_{k=1}^{N_x} \sum_{l=1}^{N_y} h_k^x h_l^y W(x_k - x_i, y_l - y_j) \rho_{k,l}^n.
$$

**Theorem 5.4.2.** The fully discrete scheme (5.4.4) has the following properties:

1. Conservation of mass:

$$
\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y \rho_{i,j}^n = \int_{\Omega} \rho_0(x, y) dx dy, \quad \text{for all } n \geq 1.
$$

(5.4.7)

2. Positivity preserving: if $\rho_{i,j}^n \geq 0$ for all $i \in \{1, \ldots, N_x\}$ and $j \in \{1, \ldots, N_y\}$, then

$$
\rho_{i,j}^{n+1} \geq 0.
$$

(3) energy dissipation: there exists $\tau^* > 0$ such that if $\tau \in (0, \tau^*)$, then

$$
E_h^{n+1} - E_h^n \leq -\frac{\tau}{2} I_h^n,
$$

(5.4.8)
therefore the corresponding equation is

\[ I^n_h = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h^y_j C^n_{i,j+1/2,j} (\log \frac{\rho^n_{i+1,j}}{M^n_{i+1,j}} - \log \frac{\rho^n_{i,j}}{M^n_{i,j}}) + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y-1} h^x_i C^n_{i,j+1/2} (\log \frac{\rho^n_{i,j+1}}{M^n_{i,j+1}} - \log \frac{\rho^n_{i,j}}{M^n_{i,j}}) \geq 0. \]

**Proof.** For simplicity of analysis we rewrite the scheme (5.4.4) as

\[
\begin{align*}
\sum_{i,j} h^x_i h^y_j \rho^n_{i,j} &= (h^x_i h^y_j M^n_{i,j} + \tau \bar{M}^n_{i+1/2,j} + \tau \bar{M}^n_{i-1/2,j} + \tau \bar{M}^n_{i,j+1/2} + \tau \bar{M}^n_{i,j-1/2} \rho^n_{i,j}) \quad \text{(5.4.9)} \\
&- \tau \bar{M}^n_{i+1/2,j} G^n_{i,j+1} - \tau \bar{M}^n_{i-1/2,j} G^n_{i,j-1} - \tau \bar{M}^n_{i,j+1/2} G^n_{i,j} - \tau \bar{M}^n_{i,j-1/2} G^n_{i,j},
\end{align*}
\]

with the following notations

\[
\bar{M}^n_{i+1/2,j} = \frac{h^y_j}{h^x_i} M^n_{i+1/2,j}, \quad \bar{M}^n_{i,j+1/2} = \frac{h^x_i}{h^y_j} M^n_{i,j+1/2}, \quad G^n_{i,j} = \frac{\rho^n_{i,j}}{M^n_{i,j}}.
\]

Note that the coefficient matrix of the linear system (5.4.9) (when consider \( G^n_{i,j} \) as unknowns) is strictly diagonally dominant, therefore (5.4.9) always has a unique solution.

1. Adding all equations in (5.4.4) and using (5.4.5) lead to (5.4.7).
2. Since \( \rho^n_{i,j} = M^n_{i,j} G^n_{i,j} \) and \( M^n_{i,j} > 0 \), it suffices to prove that \( G^n_{k,l} = \min_{(i,j)} G^n_{i,j} \geq 0 \), the corresponding equation is

\[
\begin{align*}
\sum_{i,j} h^x_i h^y_j \rho^n_{k,l} &= (h^x_i h^y_j M^n_{k,l} + \tau \bar{M}^n_{k+1/2,l} + \tau \bar{M}^n_{k-1/2,l} + \tau \bar{M}^n_{k,l+1/2} + \tau \bar{M}^n_{k,l-1/2} \rho^n_{k,l}) \\
&- \tau \bar{M}^n_{k+1/2,l} G^n_{k+1,l} - \tau \bar{M}^n_{k-1/2,l} G^n_{k-1,l} - \tau \bar{M}^n_{k,l+1/2} G^n_{k,l+1} - \tau \bar{M}^n_{k,l-1/2} G^n_{k,l-1} \\
&\leq h^x_i h^y_j M^n_{k,l} G^n_{k,l}.
\end{align*}
\]

therefore \( G^n_{k,l} \geq 0 \).

3. A direct calculation using (5.4.6) gives

\[
E^{n+1}_h - E^n_h = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h^x_i h^y_j (\rho^{n+1}_{i,j} \log(\rho^{n+1}_{i,j}) - \rho^n_{i,j} \log(\rho^n_{i,j}) + \delta^{n+1}_{i,j} \log(\frac{\rho^{n+1}_{i,j}}{\rho^n_{i,j}})) \\
+ V_j \rho^{n+1}_{i,j} + \frac{1}{2} g^n_{i,j} \rho^{n+1}_{i,j} - V_i \rho^n_{i,j} - \frac{1}{2} g^n_{i,j} \rho^n_{i,j} \quad \text{(5.4.10)}
\]

\[
\leq \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h^x_i h^y_j (\log(G^n_{i,j})(\rho^{n+1}_{i,j} - \rho^n_{i,j}) + \frac{1}{2} g^n_{i,j} \rho^n_{i,j} - g^n_{i,j} \rho^{n+1}_{i,j} + \frac{1}{2} g^n_{i,j} \rho^{n+1}_{i,j}),
\]
where we have used $\log(x) \leq x - 1$ and mass conservation property. By the symmetrical property of $W(x,y)$ we have

$$\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y g_{i,j}^n \rho_{i,j}^{n+1} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y g_{i,j}^{n+1} \rho_{i,j}^n,$$

so that

$$\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y (\frac{1}{2} g_{i,j}^n \rho_{i,j}^n - g_{i,j}^{n+1} \rho_{i,j}^n + \frac{1}{2} g_{i,j}^{n+1} \rho_{i,j}^{n+1})$$

$$= \frac{1}{2} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y (g_{i,j}^{n+1} - g_{i,j}^n) (\rho_{i,j}^{n+1} - \rho_{i,j}^n)$$

$$= \frac{1}{2} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y (\sum_{k=1}^{N_x} \sum_{l=1}^{N_y} h_k^x h_l^y W(x_k - x_j, y_j - y_l) (\rho_{k,l}^{n+1} - \rho_{k,l}^n)) (\rho_{i,j}^{n+1} - \rho_{i,j}^n)$$

$$\leq \frac{||W||_{\infty}}{2} \left( \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y \rho_{i,j}^{n+1} - \rho_{i,j}^n \right)^2$$

where $|\Omega| = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y$. Substitution of the above inequality into (5.4.10) yields

$$E_h^{n+1} - E_h^n \leq \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y \log(G_{i,j}^n) (\rho_{i,j}^{n+1} - \rho_{i,j}^n) + \frac{||W||_{\infty} |\Omega|}{2} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y (\rho_{i,j}^{n+1} - \rho_{i,j}^n)^2$$

$$= F_1^n + F_2^n.$$

We proceed using summation by parts and boundary conditions so that

$$F_1^n = \tau \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \log(G_{i,j}^n) (\tilde{M}_{i,j+1/2}^n (G_{i,j+1}^n - G_{i,j}^n) - \tilde{M}_{i,j-1/2}^n (G_{i,j}^n - G_{i,j-1}^n))$$

$$+ \tau \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \log(G_{i,j}^n) (\tilde{M}_{i+1,j}^n (G_{i,j+1}^n - G_{i,j}^n) - \tilde{M}_{i,j-1}^n (G_{i,j}^n - G_{i,j-1}^n))$$

$$= - \tau \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \tilde{M}_{i,j+1/2}^n (\log(G_{i,j+1}^n) - \log(G_{i,j}^n)) (G_{i,j+1}^n - G_{i,j}^n)$$

$$- \tau \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \tilde{M}_{i,j+1/2}^n (\log(G_{i,j+1}^n) - \log(G_{i,j}^n)) (G_{i,j+1}^n - G_{i,j}^n)$$

$$= - \tau I_1^n.$$
It remains to figure out a condition on $\tau$ so that $F_2^n + \frac{1}{2}F_1^n \leq 0$. Let $\vec{\xi}, \vec{\eta} \in \mathbb{R}^{N_x N_y}$ be vectors defined as:

$$
\vec{\xi} = \left( \sqrt{h_x} \frac{\rho_{n+1}^{1,1} - \rho_{1,1}^n}{\tau}, \cdots, \sqrt{h_x} \frac{\rho_{N_x,1}^{n+1} - \rho_{N_x,1}^n}{\tau}, \cdots, \sqrt{h_x} \frac{\rho_{N_x,N_y}^{n+1} - \rho_{N_x,N_y}^n}{\tau} \right)^T
$$

$$
\vec{\eta} = \left( \sqrt{h_y} \log(G_{1,1}^{n,*}), \cdots, \sqrt{h_y} \log(G_{N_x,1}^{n,*}, \cdots, \sqrt{h_y} \log(G_{N_x,N_y}^{n,*}) \right)^T,
$$

then $F_2^n + \frac{1}{2}F_1^n \leq 0$ if

$$
\tau^2 \|W\|_{\infty} |\vec{\xi}|^2 + \tau \vec{\xi} \cdot \vec{\eta} \leq 0.
$$

In similar manner as in 1D case, we can show that $\vec{\xi} \cdot \vec{\eta} = 0$ if and only if $\vec{\xi} = 0$. Therefore

$$
0 < c_0 \leq \frac{\vec{\xi} \cdot \vec{\eta}}{|\vec{\xi}|^2} \leq \frac{|\vec{\eta}|}{|\vec{\xi}|} \quad \text{for } \xi \neq 0,
$$

where $c_0$ may depend on numerical solutions at $t_n$ and $t_{n+1}$. We thus obtain the desired result (5.4.8) by taking $\tau \leq \tau^* = \frac{c_0}{\|W\|_{\infty} |\vec{\eta}|}$.

**Remark 5.4.1.** The schemes presented so far may be applied to the general class of nonlinear nonlocal equations (5.1.6), based on the reformulation

$$
\partial_t \rho = \nabla \cdot (M \nabla \frac{\rho}{M}),
$$

where $M = \rho e^{-H'(\rho) - V(x) - W* \rho}$ for $\rho$ away from zero. The numerical solution may be oscillatory at low density, for which one could use either upwind numerical fluxes or non-oscillatory limiters as a remedy [8]. Note that for the aggregation equation (in the absence of diffusion), particle methods have been developed in [16, 43]; Particle methods naturally conserve mass and positivity, yet a large number of particles is often required to resolve finer properties of solutions.

### 5.5 Second order in-time discretization

The numerical schemes presented so far are only first order in time. In this section we extend these schemes with a second order in time discretization.
5.5.1 Second order scheme for 1D problem

We replace (5.3.1) by a two step scheme

\[ \frac{\rho_j^n - \rho_j^0}{\tau/2} = \frac{C_{j+1/2}^* - C_{j-1/2}^*}{h_j}, \quad j = 1, 2, \ldots, N, \]  
(5.5.1a)

\[ \rho_j^{n+1} = 2\rho_j^* - \rho_j^n, \quad j = 1, 2, \ldots, N, \]  
(5.5.1b)

where

\[ C_{j+1/2}^* = \frac{M_{j+1/2}^*}{h_{j+1/2}} \left( \frac{\rho_{j+1}^*}{M_{j+1}^*} - \frac{\rho_j^*}{M_j^*} \right), \quad \text{for} \quad j = 1, 2, \ldots, N - 1, \]

\[ C_{1/2}^* = 0, \quad C_{N+1/2}^* = 0, \]

with \( M_{j+1/2}^* = Q_1(x_{j+1/2}, \frac{3}{2}\rho^n - \frac{1}{2}\rho^{n-1}) \) and \( M_j^* = Q_1(x_j, \frac{3}{2}\rho^n - \frac{1}{2}\rho^{n-1}) \). The scheme (5.5.1) has following properties.

**Theorem 5.5.1.** Let \( \rho^{n+1} \) be obtained from (5.5.1), then

1. Conservation of mass:

\[ \sum_{j=1}^{N} h_j \rho_j^n = \int_{\Omega} \rho_0(x) dx, \quad \text{for} \quad n \geq 1. \]

2. Positivity preserving: if \( \rho_j^n \geq 0 \) for all \( j = 1, \ldots, N \), then

\[ \rho_j^{n+1} \geq 0, \quad j = 1, \ldots, N, \]

provided \( \tau \) is sufficiently small.

**Proof.** (1) From the scheme construction, the conservation property remains hold.

(2) Setting

\[ G_j^n = \frac{\rho_j^n}{M_j^*}, \quad g_{j+1/2}^* = \frac{M_{j+1/2}^*}{h_{j+1/2}}, \]

and a careful regrouping leads to the following linear system

\[
\begin{align*}
(M_1^* + \frac{\tau}{2h_1}g_{3/2})G_{1}^{n+1} - \frac{\tau}{2h_1}g_{3/2}G_{2}^{n+1} &= b_1, \\
(M_j^* + \frac{\tau}{2h_j}(g_{j+1/2}^* + g_{j-1/2}^*))G_{j}^{n+1} - \frac{\tau}{2h_j}g_{j+1/2}G_{j+1}^n - \frac{\tau}{2h_j}g_{j-1/2}G_{j-1}^{n+1} &= b_j, \\
(M_N^* + \frac{\tau}{2h_N}g_{N-1/2})G_{N}^{n+1} - \frac{\tau}{2h_N}g_{N-1/2}G_{N-1}^{n+1} &= b_N,
\end{align*}
\]
(5.5.2)
where \( j = 1, \cdots, N - 1 \), with the right hand side vector given by
\[
b_1 = \left( M_1^* - \frac{\tau}{2h_1} g_{3/2}^* \right) G_1^m + \frac{\tau}{2h_1} g_{3/2}^* G_2^m,
\]
\[
b_j = \left( M_j^* - \frac{\tau}{2h_j} (g_{j+1/2}^* + g_{j-1/2}^*) \right) G_j^m + \frac{\tau}{2h_j} g_{j+1/2}^* G_{j+1}^m + \frac{\tau}{2h_j} g_{j-1/2}^* G_{j-1}^m, \quad j = 1, \cdots, N - 1,
\]
\[
b_N = \left( M_N^* - \frac{\tau}{2h_N} g_{N-1/2}^* \right) G_N^m + \frac{\tau}{2h_N} g_{N-1/2}^* G_{N-1}^m.
\]

The linear system (5.5.2) admits a unique solution \( \{ G_j^{n+1} \} \) since its coefficient matrix is strictly diagonally dominant. Following the proof of (2) in Theorem 5.3.1, we see that \( G_j^{n+1} \geq 0 \) is ensured if each \( b_j \geq 0 \), which is the case provided
\[
\tau \leq \min \left\{ \frac{2h_1 M_1^*}{g_{3/2}^*}, \min_{1 < j < N} \frac{2h_j M_j^*}{g_{j+1/2}^* + g_{j-1/2}^*}, \frac{2h_N M_N^*}{g_{N-1/2}^*} \right\}.
\]

The stated result thus follows. \( \square \)

**Remark 5.5.1.** We expect the energy dissipation to still hold for smaller time steps, as can be seen in Figure 5.3(a) in our numerical tests. Moreover, the energy dissipation was also observed for relatively larger time steps, see Figure 5.3(b).

For large time step \( \tau \), non-negativity of \( \rho^{n+1} \) obtained by the second order scheme (5.5.1) may not be guaranteed, we introduce a local limiter to resolve the solution positivity.

### 5.5.2 Local limiter and algorithm

We begin to design a local limiter to restore positivity of \( \{ c_j \}_{j=1}^N \) if \( \sum_{j=1}^N c_j > 0 \), but \( c_k < 0 \) for some \( k \). The idea is to find a neighboring index set \( S_k \) such that the local average
\[
\bar{c}_k = \frac{1}{|S_k|} \sum_{j \in S_k} c_j > 0,
\]
where \( |S_k| \) denotes the minimum number of indexes for which \( c_j \neq 0 \) and \( \bar{c}_k > 0 \), then use this as a reference to define the following scaling limiter,
\[
\tilde{c}_j = \theta c_j + (1 - \theta) \bar{c}_k, \quad j \in S_k, \quad (5.5.3)
\]
where
\[
\theta = \min \left\{ 1, \frac{\bar{c}_k}{c_k - c_{\min}} \right\}, \quad c_{\min} = \min_{j \in S_k} c_j.
\]
Lemma 5.5.1. This limiter has the following properties:

(1) $\tilde{c}_j \geq 0$ for all $j \in S_k$,
(2) $\sum_{j \in S_k} \tilde{c}_j = \sum_{j \in S_k} c_j$, and
(3) $|\tilde{c}_j - c_j| \leq |S_k|(-\min_{j \in S_k} c_j)$.

Proof. (1) This follows from the definition of $\theta$ and (5.5.3).

(2) By (5.5.3) and the definition of $\bar{c}_k$, it follows that
\[
\sum_{j \in S_k} \tilde{c}_j = \theta |S_k| \bar{c}_k + (1 - \theta) \bar{c}_k |S_k| = \sum_{j \in S_k} c_j.
\]

(3) From (5.5.3) it follows that for all $j \in S_k$,
\[
|\tilde{c}_j - c_j| = (1 - \theta) |\tilde{c}_k - c_j| = -c_{\min} \frac{|\tilde{c}_k - c_j|}{(\bar{c}_k - c_{\min})}
\leq (-c_{\min}) \max \left\{1, \frac{c_{\max} - \bar{c}_k}{\bar{c}_k - c_{\min}}\right\},
\]
where $c_{\max} := \max_{j \in S_k} c_j$ and $c_{\min} := \min_{j \in S_k} c_j$. Note that $\sum_{j \in S_k} (\tilde{c}_k - c_j) = 0$ implies
\[
\sum_{j \in S_k^+} (c_j - \bar{c}_k) = \sum_{j \in S_k^-} (\bar{c}_k - c_j),
\]
in which each term involved on both sides is nonnegative. Hence, $c_{\max} - \bar{c}_k \leq |S_k|(\bar{c}_k - c_{\min})$.
Obviously, $|S_k| \geq 1$. Hence the claimed bound follows. \qed

Remark 5.5.2. In general, $|S_k|$ may not be bounded. For instance, we let
\[
c_j = \frac{1}{2^j} \text{ for } j = 1, \cdots, N - 1, \text{ and } c_N = -\frac{1}{2},
\]
then $\sum_{j=1}^{N} c_j = \frac{1}{2} - \frac{1}{2^{N+1}} > 0$, but $\sum_{j=2}^{N} c_j = -\frac{1}{2^{N+1}} < 0$. This implies that $|S_N| = N$ since $S_N = \{1, \cdots, N\}$.

The above limiter when applied to $\{\rho_j\}$ with $c_j = h_j \rho_j$ gives
\[
\tilde{\rho}_j = \theta \rho_j + (1 - \theta) \frac{\bar{c}_k}{h_j}, \tag{5.5.4}
\]
where
\[
\theta = \min \left\{1, \frac{\bar{c}_k}{\bar{c}_k - c_{\min}}\right\}, \quad c_{\min} = \min_{j \in S_k} h_j \rho_j, \quad \bar{c}_k = \frac{1}{|S_k|} \sum_{j \in S_k} h_j \rho_j.
Such limiter still respects the local mass conservation. In addition, for any sequence $g_j$ with $g_j \geq 0$, we have
\[
|\tilde{\rho}_j - g_j| \leq (1 + |S_k|\alpha) \max_{j \in S_k} |\rho_j - g_j|, \quad j \in S_k,
\]
where $\alpha$ is the upper bound of mesh ratio $h_i/h_j$. Let $\rho_j$ be the approximation of $\rho(x) \geq 0$, we let $g_j = \rho(x_j)$ or the average of $\rho$ on $I_j$, so we can assert that the accuracy is not destroyed by the limiter as long as $|S_k|\alpha$ is uniformly bounded. In practice, it is indeed the case as verified by our numerical tests when using shape-regular meshes.

Indeed, the boundedness of $|S_k|$ can be proved rigorously for shape-regular meshes.

**Theorem 5.5.2.** Let $\rho(x) \geq 0$, be in $C^2(\Omega)$, and $\{\rho_j\}$ be an approximation of $\rho(x)$ such that
\[
|\rho_j - \rho(x_j)| \leq Ch^2,
\]
where $h = \min_{1 \leq j \leq N} h_j$ and $h_j \leq \alpha h$ for some $\alpha > 0$. If $\rho_k < 0$ (or only finite number of neighboring values are negative), then there exists $K^* > 0$ finite such that
\[
|S_k| \leq K^*.
\]
where $K^*$ may depend on the local meshes associated with $S_k$.

**Proof.** Under the assumption $\rho_k < 0$, $\rho$ must touch zero near $x_k$. We discuss the case where $\rho(x^*) = 0$ and $\rho'(x^*) = 0$ with $\rho(x) > 0$ for $x > x^*$ locally with $x^* \in I_k$. The case where $\rho(x) > 0$ for $x < x^*$ can be handled as well. Without loss of generality, we consider $k = 1$ with $x^* \in I_1$, and $\int_{I_1} \rho(x)dx > 0$. It suffices to find $K$ such that
\[
\sum_{j=1}^{K} h_j \rho_j > 0. \tag{5.5.5}
\]
Using the error bound we have
\[
\rho_j \geq \rho(x_j) - Ch^2.
\]
Also from $\rho \in C^2$ we can deduce that
\[
\rho(x_j) \geq \tilde{\rho}_j - \lambda h_j^2,
\]
with $\lambda = \frac{1}{24} \max_{x \in \Omega} |\rho''|$ and the cell average $\bar{\rho}_j = \frac{1}{h_j} \int_{I_j} \rho(x) dx$. Combining these we see that the left hand side of (5.5.5) is bounded from below by

$$\sum_{j=1}^{K} h_j \rho_j \geq \sum_{j=1}^{K} h_j (\bar{\rho}_j - Ch^2 - \lambda h_j^2)$$

$$\geq \int_{x_{1/2}}^{2K+1/2} \rho(x) dx - (\lambda + C) \sum_{j=1}^{K} h_j^3$$

$$\geq \int_{x_{1/2}}^{2K+1/2} \rho(x) dx - (\lambda + C) h^2 \alpha^2 \sum_{j=1}^{K} h_j$$

$$= \left[ \int_{1}^{1} \rho \left( \theta \eta + x_{1/2} \right) d\theta - (\lambda + C) h^2 \alpha^2 \right] \eta,$$

where $\eta := \sum_{j=1}^{K} h_j$, and we have used $h_j \leq h \alpha$. Using the fact $Kh \leq \eta$, the term in the bracket is bounded below by

$$\int_{0}^{1} \rho \left( \theta \eta + x_{1/2} \right) d\theta - (\lambda + C) \eta^2 \alpha^2 / K^2,$$

which is positive if

$$K > \frac{\alpha \sqrt{\lambda + C} \eta}{\sqrt{\int_{0}^{1} \rho \left( \theta \eta + x_{1/2} \right) d\theta}}.$$ 

This can be ensured if we take

$$K = \lceil A \rceil + 1,$$

where for $\Omega = [a,b]$,

$$A = \max_{z \in [h_1, b-a]} \frac{\alpha \sqrt{\lambda + C} z}{\sqrt{\int_{0}^{1} \rho \left( \theta z + x_{1/2} \right) d\theta}}$$

which is bounded and depends on $h_1$. For general cases a different bound can be identified and it may depend on local meshes.

Note that our numerical solutions feature the following property: if $\rho_j^n = 0$, then $\rho_{j}^{n+1} = 2\rho_j^n - \rho_j^n \geq 0$ due to the fact that $\rho_j^n \geq 0$ for all $j = 1, \cdots, N$. This means that if $\rho_0(x) = 0$ on an interval, then $\rho_1^n$ cannot be negative in most of nearby cells. Thus negative values appear only where the exact solution turns from zero to a positive value, and the number of these values are finitely many. Our result in Theorem 5.5.2 is thus applicable.
Algorithm. We have the following algorithm:

1. Initialization: From initial data $\rho_0(x)$, obtain $\rho_j^0 = \frac{1}{h_j} \int_{I_j} \rho_0(x) dx$, $j = 1, \cdots, N$, by using a second order quadrature.

2. Update to get $\{\rho_j^1\}$ by the first order scheme (5.3.1).

3. Marching from $\{\rho_j^n\}$ to $\{\rho_j^{n+1}\}$ for $n = 1, 2, \cdots$, based on (5.5.1).

4. Reconstruction: if necessary, locally replace $\rho_j^{n+1}$ by $\tilde{\rho}_j^{n+1}$ using the limiter defined in (5.5.4).

The following algorithm can be called to find an admissible set $S_k$ used in (5.5.4).

(i) Start with $S_k = \{k\}$, $m = 1$.

(ii) If $k - m \geq 1$ and $c_{k-m} \neq 0$, then set $S_k = S_k \cup \{k - m\}$. If $\bar{c}_k > 0$, then stop, else go to (iii).

(iii) If $k + m \leq N$ and $c_{k+m} \neq 0$, then set $S_k = S_k \cup \{k + m\}$. If $\bar{c}_k > 0$, then stop, else set $m = m + 1$ and go to (ii).

5.5.3 Second order scheme for 2D problem

A similar two step time-discretization technique can be applied to higher dimensional problems. In the 2D case, that with scheme (5.4.2) gives the following fully discrete scheme,

$$\frac{\rho_{i,j}^* - \rho_{i,j}^n}{\tau/2} = \frac{C_{i+1/2,j}^* - C_{i-1/2,j}^*}{h_i^x} + \frac{C_{i,j+1/2}^* - C_{i,j-1/2}^*}{h_j^y},$$

(5.5.6a)

$$\rho_{i,j}^{n+1} = 2\rho_{i,j}^* - \rho_{i,j}^n,$$

(5.5.6b)

where

$$C_{i+1/2,j}^* = \frac{M_{i+1/2,j}^*}{h_i^{x+1/2}} \left( \frac{\rho_{i+1,j}^* - \rho_{i,j}^*}{\frac{M_{i+1,j}^*}{M_{i,j}^*}} \right), \quad i = 1, \cdots, N_x - 1, \quad j = 1, \cdots, N_y,$$

$$C_{i,j+1/2}^* = \frac{M_{i,j+1/2}^*}{h_j^{y+1/2}} \left( \frac{\rho_{i,j+1}^* - \rho_{i,j}^*}{\frac{M_{i,j+1}^*}{M_{i,j}^*}} \right), \quad i = 1, \cdots, N_x, \quad j = 1, \cdots, N_y - 1,$$

$$C_{1/2,j}^* = C_{N_x+1/2,j}^* = C_{i,1/2}^* = C_{i,N_y+1/2}^* = 0, \quad i = 1, \cdots, N_x, \quad j = 1, \cdots, N_y,$$
with \( M_{i+1/2,j}^* = Q_2(x_{i+1/2}, y_j, \frac{3}{2} \rho_{n} - \frac{1}{2} \rho_{n-1}) \), \( M_{i,j+1/2}^* = Q_2(x_i, y_{j+1/2}, \frac{3}{2} \rho_{n} - \frac{1}{2} \rho_{n-1}) \), and \( M_{i,j}^* = Q_2(x_i, y_j, \frac{3}{2} \rho_{n} - \frac{1}{2} \rho_{n-1}) \). In an entirely similar fashion (details are therefore omitted), we can prove the following.

**Theorem 5.5.3.** The fully discrete scheme (5.5.6) has the following properties:

1. **Conservation of mass:**
   \[
   \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y \rho_{i,j}^n = \int_{\Omega} \rho_0(x, y) dxdy, \quad \text{for } n \geq 1.
   \]

2. **Positivity preserving:** if \( \rho_{i,j}^n \geq 0 \) for all \( i \in \{1, \cdots, N_x\} \) and \( j \in \{1, \cdots, N_y\} \), then
   \[\rho_{i,j}^{n+1} \geq 0,\]
   provided \( \tau \) is sufficiently small.

### 5.5.4 Local limiter and algorithm

If the time step \( \tau \) is not small, positivity of \( \rho_{i,j}^n \) is not guaranteed for \( n \geq 2 \). We use the following limiter to resolve this issue:

\[
\tilde{\rho}_{i,j} = \theta \rho_{i,j} + (1 - \theta) \frac{\overline{c}_{k,l}}{h_i^x h_j^y},
\]
with

\[
\theta = \min \left\{ 1, \frac{\overline{c}_{k,l}}{\overline{c}_{k,l} - c_{\min}} \right\}, \quad c_{\min} = \min_{(i,j) \in S_{k,l}} h_i^x h_j^y \rho_{i,j}, \quad \overline{c}_{k,l} = \frac{1}{|S_{k,l}|} \sum_{(i,j) \in S_{k,l}} h_i^x h_j^y \rho_{i,j},
\]

where \( S_{k,l} \) denotes the minimum number of indexes for which \( \rho_{i,j} \neq 0 \) and \( \overline{c}_{k,l} > 0 \).

The limiter (5.5.7) can be shown to be nonnegative and satisfy the local mass conservation. In addition, for any \( g_{i,j} \geq 0 \) we have

\[
|\tilde{\rho}_{i,j} - g_{i,j}| \leq (1 + |S_{k,l}|) \alpha \max_{(i,j) \in S_{k,l}} |\rho_{i,j} - g_{i,j}|, \quad (i,j) \in S_{k,l},
\]

where \( \alpha \) is the upper bound of 2D mesh ratios. Hence the second order accuracy remains for shape-regular meshes since \( |S_{k,l}| \) can be shown bounded as in the one-dimensional case.

**Algorithm** Our algorithm for 2D problem is given as follows:
1. Initialization: From initial data \( \rho_0(x,y) \), obtain \( \rho^0_{i,j} = \frac{1}{I_i,j} \int_{I_i,j} \rho_0(x,y) dx dy, \) \( i = 1, \cdots, N_x, j = 1, \cdots, N_y \), by using a second order quadrature.

2. Update to get \{ \rho^1_{i,j} \} by the first order scheme (5.4.4).

3. March from \{ \rho^n_{i,j} \} to \{ \rho^{n+1}_{i,j} \} based on the scheme (5.5.6).

4. Reconstruction: if necessary, locally replace \( \rho^{n+1}_{i,j} \) by \( \tilde{\rho}^{n+1}_{i,j} \) using the limiter defined in (5.5.7).

The following algorithm can be called to find an admissible set \( S_{k,l} \) used in (5.5.7).

(i) Start with \( S_{k,l} = \{(k,l)\}, m = 1 \).

(ii) For \( d_y = \max\{1, l - m\} : \min\{l + m, N_y\} \) and \( d_x = \max\{1, k - m\} : \min\{k + m, N_x\} \),

If \( (d_x, d_y) \notin S \) and \( c_{k-m} \neq 0 \), then set \( S_{k,l} = S_{k,l} \cup \{(d_x, d_y)\} \).

If \( \bar{c}_{k,l} > 0 \), then stop, else go to (iii).

(iii) Set \( m = m + 1 \) and go to (ii).

5.6 Numerical Examples

In this section, we implement the fully discrete schemes (5.3.1) and (5.4.4) and second order extensions (5.5.1) and (5.5.6). Errors in 1-D case are measured in the following discrete norms:

\[ e_{l1} = h \sum_{i=1}^{N} |\rho^n_i - \tilde{\rho}^n_i|, \]
\[ e_{l\infty} = \max_{1 \leq i \leq N} |\rho^n_i - \tilde{\rho}^n_i|. \]

Here \( \tilde{\rho}^n_i \) is cell average of the exact solution on \( I_i \) at time \( t = n\tau \).

5.6.1 One-dimensional tests

Example 5.6.1. (Accuracy test) In this example we test the accuracy of scheme (5.3.1) and scheme (5.5.1) Consider the initial value problem with source term

\[
\begin{aligned}
\partial_t \rho &= \partial_x (\partial_x \rho + \rho \partial_x (V(x) + W \ast \rho)) + F(x,t), \quad t > 0, \; x \in [-\pi, \pi], \\
\rho(x,0) &= 2 + \cos(x), \quad x \in [-\pi, \pi],
\end{aligned}
\]
subject to zero flux boundary conditions. Here we take $V(x) = \cos(x), W(x) = \cos(x)$, and

$$F(x, t) = \pi e^{-2t}(2\cos^2(x) + 2\cos(x) - 1) + e^{-t}(2\cos^2(x) + 2\cos(x) - 3).$$

One can check that the exact solution to (5.6.1) is

$$\rho(x, t) = e^{-t}(2 + \cos(x)).$$

We compute to $t = 1$, first use time step $\tau = 0.1h$ and $\tau = h^2$ to check accuracy of scheme (5.3.1), then use $\tau = h$ to check accuracy of scheme (5.5.1), results are reported in Table 5.1 and Table 5.2 respectively. We see that the scheme (5.3.1) is first order accurate in time and second order accurate in space, while the scheme (5.5.1) is second order accurate both in time and space.

Note that the exact solution is $\rho(x, t) = e^{-t}(2 + \cos(x))$, which is far above 0 for $t \in [0, 1]$. Hence the positivity-preserving limiter is not activated in this test.

<table>
<thead>
<tr>
<th>N</th>
<th>$l^1$ error</th>
<th>order</th>
<th>$l^\infty$ error</th>
<th>order</th>
<th>$l^1$ error</th>
<th>order</th>
<th>$l^\infty$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.70474E-01</td>
<td>-</td>
<td>0.26268E-01</td>
<td>-</td>
<td>0.10451E-00</td>
<td>-</td>
<td>0.46075E-01</td>
<td>-</td>
</tr>
<tr>
<td>80</td>
<td>0.32212E-01</td>
<td>1.1295</td>
<td>0.15021E-01</td>
<td>0.8063</td>
<td>0.25847E-01</td>
<td>2.0156</td>
<td>0.11397E-01</td>
<td>2.0153</td>
</tr>
<tr>
<td>160</td>
<td>0.15796E-01</td>
<td>1.0280</td>
<td>0.79593E-02</td>
<td>0.9163</td>
<td>0.64441E-02</td>
<td>2.0039</td>
<td>0.28433E-02</td>
<td>2.0030</td>
</tr>
<tr>
<td>320</td>
<td>0.78955E-02</td>
<td>1.0005</td>
<td>0.40881E-02</td>
<td>0.9612</td>
<td>0.16098E-02</td>
<td>2.0011</td>
<td>0.71027E-03</td>
<td>2.0011</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>$l^1$ error</th>
<th>order</th>
<th>$l^\infty$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.14049E-00</td>
<td>-</td>
<td>0.43022E-01</td>
<td>-</td>
</tr>
<tr>
<td>80</td>
<td>0.35941E-01</td>
<td>1.9668</td>
<td>0.10729E-01</td>
<td>2.0036</td>
</tr>
<tr>
<td>160</td>
<td>0.90784E-02</td>
<td>1.9851</td>
<td>0.26805E-02</td>
<td>2.0009</td>
</tr>
<tr>
<td>320</td>
<td>0.22814E-02</td>
<td>1.9925</td>
<td>0.67108E-03</td>
<td>1.9980</td>
</tr>
</tbody>
</table>

**Example 5.6.2.** In this example, we study dynamics of linear Fokker-Plank equations by considering the following problem

$$\partial_t \rho = \partial_x(\partial_x \rho + x \rho), \quad t > 0, \quad x \in [-5, 5],$$  \hspace{1cm} (5.6.2)
with initial condition

\[
\rho(x, 0) = \begin{cases} 
\frac{1}{7} \int_{\Omega} e^{-\frac{x^2}{2}} \, dx, & x \in [-3.5, 3.5], \\
n0, & \text{otherwise,}
\end{cases}
\]  

(5.6.3)

and zero flux boundary conditions \((\partial_x \rho + x \rho)|_{x=\pm5} = 0\).

This is (5.2.1) with \(V(x) = \frac{x^2}{2}\) and \(W(x) = 0\). The steady state to (5.6.2) is \(\rho_{eq}(x) = e^{-\frac{x^2}{2}}\).

We use the time step \(\tau = 0.1\) to compute solutions up to \(t = 4\), with \(N = 200\). In Figure 5.1(a) are snap shots of \(\rho\) obtained by the first order scheme (5.3.1) at \(t = 0, 0.2, 0.5, 1, 4\), and the steady state. Figure 5.1(b) shows the mass conservation and energy decay. We observe from this figure that the solution of problem (5.6.2) becomes indistinguishable from the steady state after \(t = 2\). Compared in Figure 5.2 are numerical solutions obtained by the second order scheme (5.5.1) with and without the local limiter. We see that the limiter produces positive solutions and reduces solution oscillations.

Both mass conservation and energy dissipation for the second order scheme are given in Figure 5.3. In Figure 5.3(a), we take \(\tau = 0.01\), for which no limiter is needed. In Figure 5.3(b), we take \(\tau = 0.1\), the limiter keeps being invoked at each step.
Figure 5.2 Second order scheme (with and without limiter) for Example 5.6.2.

Figure 5.3 Second order scheme energy and total mass (without limiter for $\tau = 0.01$, with limiter for $\tau = 0.1$) for Example 5.6.2.
Example 5.6.3. (Doi-Onsager equation with the Maier-Saupe potential) In this example, we consider the Doi-Onsager equation with Maier-Saupe potential

\[
\begin{align*}
\partial_t \rho &= \partial_x (\partial_x \rho + \alpha \rho \partial_x (W * \rho)), \\
\rho(x, 0) &= \frac{x+1}{2\pi(\pi+1)}, \\
W(x) &= \sin^2(x), \\
W(x) &= \sin^2(x) t > 0, x \in [0, 2\pi]
\end{align*}
\]

subject to zero flux boundary conditions. Here $\alpha$ is the intensity parameter. Stationary solutions of (5.6.4) have been an interesting subject of study, since when $\alpha$ increases, phase transition from isotropic state to nematic state will appear. A detailed characterization of solutions can be found in [37]: for $0 < \alpha \leq \alpha^* = 4$, the only stationary solution is the isotropic state $\rho_{eq}(x) = \frac{1}{2\pi}$. When $\alpha > \alpha^*$ besides the constant solution $\rho_{eq}(x) = \frac{1}{2\pi}$, there are other solutions given by

\[
\rho_{eq}(x) = \frac{e^{-\eta^* \cos(2(x-x_0))}}{\int_0^{2\pi} e^{-\eta^* \cos(2x)} dx},
\]

where $x_0$ is arbitrary, $\eta^* > \frac{\alpha}{2} \sqrt{1 - 4/\alpha}$ is uniquely determined by

\[
\frac{\int_0^{2\pi} \cos(2x) e^{-\eta^* \cos(2x)} dx}{\int_0^{2\pi} e^{-\eta^* \cos(2x)} dx} + \frac{2\eta}{\alpha} = 0.
\]

We use scheme (5.3.1) and choose the time step $\tau = 0.1$ to compute up to $T = 30$ with $N = 80$. In Figure 5.4(a) are snap shots of solutions to (5.6.4) for $\alpha = 3 < \alpha^*$ at $t = 0, 0.5, 5, 15, 25, 30$. Figure 5.4(b) shows mass conservation and energy decay, from which we can observe that the problem (5.6.4) is already at steady state $\rho_{eq}(x) = \frac{1}{2\pi}$ after $t = 20$. In Figure 5.5(a) are snap shots of solutions to (5.6.4) for $\alpha = 5 > \alpha^*$ at $t = 0, 0.5, 1, 5, 25, 35$. Figure 5.5(b) shows mass conservation and energy decay, which tells that problem (5.6.4) is at already steady state after $t = 30$. In Figure 5.6 are the free energy plots for different time steps, we observe energy dissipation even for large time steps. Our method gives satisfying results for the problem, consistent with the numerical results obtained in [13] by an explicit scheme with Euler forward time discretization.

In Table 5.3, we compare the efficiency of schemes (5.3.1) and (5.5.1) using Example 5.6.2 and Example 5.6.3. We choose $\tau = 0.01$ for the first order scheme and $\tau = 0.1$ for the second order scheme so that they have same accuracy. We see from Table 5.3 that the second order scheme is more efficient than the first order scheme in all three cases. We also see that, in Example 5.6.2, the
Figure 5.4  Solution evolution and energy dissipation for Example 5.6.3 with $\alpha = 3$.

Figure 5.5  Solution evolution and energy dissipation for Example 5.6.3 with $\alpha = 5$.

Figure 5.6  Energy dissipation for Example 5.6.3 with $\tau = 0.01, 0.1, 1$. 
limiter in the second order scheme takes about 13% of the total computational time, but no limiter is used in Example 5.6.3 because the exact solutions are away from zero.

Table 5.3  Efficiency of schemes (5.3.1) and (5.5.1) (CPU times in seconds).

<table>
<thead>
<tr>
<th></th>
<th>Scheme (5.3.1)</th>
<th>Scheme (5.5.1)</th>
<th>Limiter in (5.5.1)</th>
<th>Cost for limiter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 5.6.2, $T = 4$</td>
<td>0.31307E-00</td>
<td>0.48877E-01</td>
<td>0.63181E-02</td>
<td>13%</td>
</tr>
<tr>
<td>Example 5.6.3 $\alpha = 3, T = 25$</td>
<td>0.15611E+02</td>
<td>0.16166E+01</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Example 5.6.3, $\alpha = 5, T = 25$</td>
<td>0.14853E+02</td>
<td>0.16315E+01</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

5.6.2 Two-dimensional tests

Example 5.6.4. (Accuracy test) We consider the initial value problem with source term,

$$\begin{aligned}
\partial_t \rho &= \nabla \cdot (\nabla \rho + \rho \nabla V(x,y)) + F(x,y,t), \quad t > 0, \quad (x, y) \in [-\frac{\pi}{2}, \frac{\pi}{2}] \times [-\frac{\pi}{2}, \frac{\pi}{2}], \\
\rho(x,y,0) &= 2 + \sin(x)\sin(y), \quad (x, y) \in [-\frac{\pi}{2}, \frac{\pi}{2}] \times [-\frac{\pi}{2}, \frac{\pi}{2}],
\end{aligned}$$

subject to zero flux boundary conditions, here $V(x,y) = \sin(x)\sin(y)$, and

$$F(x,y,t) = e^{-t}(2\sin^2(x)\sin^2(y) + 5\sin(x)\sin(y) - \cos^2(x)\sin^2(y) - \sin^2(x)\cos^2(y) - 2).$$

This problem has the exact solution

$$\rho(x,t) = e^{-t}(2 + \sin(x)\sin(y)).$$

We choose $\tau = 0.1h^2$ in scheme (5.4.4) and $\tau = 0.1h$ in scheme (5.5.6). Errors and orders at $t = 1$ are listed in Table 5.4, in this test uniform meshes with $h = h^x = h^y = \pi/N$ have been used.

Table 5.4  Accuracy of scheme (5.4.4) and (5.5.6).

<table>
<thead>
<tr>
<th>$N \times N$</th>
<th>scheme (5.4.4) with $\tau = 0.1h^2$</th>
<th>scheme (5.5.6) with $\tau = 0.1h$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$l^1$ error</td>
<td>order</td>
</tr>
<tr>
<td>10 $\times$ 10</td>
<td>0.927816E-1</td>
<td>-</td>
</tr>
<tr>
<td>20 $\times$ 20</td>
<td>0.232384E-1</td>
<td>1.997</td>
</tr>
<tr>
<td>40 $\times$ 40</td>
<td>0.581196E-2</td>
<td>1.999</td>
</tr>
<tr>
<td>80 $\times$ 80</td>
<td>0.145297E-2</td>
<td>2.000</td>
</tr>
</tbody>
</table>
Finally we mention that there is a class of equations in which the interaction is modeled through a potential governed by the Poisson equation. The celebrated model is the Patlak-Keller-Segel system of the chemotaxis [25, 26]. The original model is a coupled parabolic system, and the one related to our model equation (5.1.1) is the parabolic-elliptic version of the form (see e.g., [40])

\[
\begin{cases}
\partial_t \rho &= \Delta \rho - \nabla \cdot (\chi \rho \nabla c), \quad t > 0, \; x \in \mathbb{R}^2, \\
-\Delta c &= \rho, \\
\rho(x, 0) &= \rho_0(x), \quad x \in \mathbb{R}^2.
\end{cases}
\] (5.6.6)

Here, \(\rho(x,t)\) is the cell density, \(c(x,t)\) is the chemical attractant concentration, the parameter \(\chi > 0\) is the sensitivity of bacteria to the chemical attractant. It has been shown in [3] that the solution behavior of problem (5.6.6) is quite different when crossing a critical mass. If the initial mass \(M = \int_{\mathbb{R}^2} \rho_0(x,y) \, dx \, dy\) is smaller than a critical value \(M_c = 8\pi/\chi\), then the solution exists globally. When \(M > M_c\), the solution will blow up in finite time, which is referred to as chemotactic collapse.

**Example 5.6.5.** (Patlak–Keller–Segal system). In this example, we test the method’s capacity in capturing solution concentrations for the Patlak–Keller–Segal system (5.6.6). Using the Green function for the Poisson equation, this system can be reformulated as (5.1.1) with \(V = 0\) and

\[
W(x,y) = \frac{\chi}{2\pi} \log(\sqrt{x^2 + y^2}).
\] (5.6.7)

In our simulation, we restrict to a bounded domain \(\Omega\) subject to zero flux boundary conditions, using formulation (5.4.1) with \(V(x,y) = 0\) and \(W\) defined in (5.6.7). We fix \(\chi = 1\) and consider both the sub-critical case with

\[
\rho_0(x,y) = \begin{cases}
2(\pi - 0.2), & (x,y) \in [-1,1] \times [-1,1], \\
0, & (x,y) \in \Omega \setminus [-1,1] \times [-1,1],
\end{cases}
\]
on \(\Omega = [-5,5] \times [-5,5]\), and super-critical case with

\[
\rho_0(x,y) = \begin{cases}
2(\pi + 0.2), & (x,y) \in [-1,1] \times [-1,1], \\
0, & (x,y) \in \Omega \setminus [-1,1] \times [-1,1],
\end{cases}
\]
on $\Omega = [-1.5, 1.5] \times [-1.5, 1.5]$, for which we know that the solution blows-up at finite time.

We take time step $\tau = 0.01$, and set $N_x = N_y = 51$ so that a single cell is located at the center of the computational domain, where one can view a clear picture of the blow-up phenomena in super-critical case. In Figure 5.7 are snap shots of numerical solutions in the sub-critical case at $t = 0, 2, 8, 12, 16$, from which we observe that the numerical solution dissipates in time, the last picture in Figure 5.7 shows mass conservation and energy dissipation. In Figure 5.8 are snap shots of numerical solutions in super-critical case at $t = 0, 0.5, 1, 1.5, 2$, we observe that numerical solutions tend to concentrate at the origin.

Let us remark that in [50] the same concentration phenomena was observed, using a DG method for this problem with periodic boundary conditions. Different boundary conditions do not affect the concentration profile since the solution is compactly supported in our setting. In the super-critical case, the peak in our result is slightly lower than that captured in [50], this is expected because the solution is concentrated at a single point, and cell averaging near the origin can decrease the height of the peak.

5.7 Concluding remarks

In this paper, we have developed positive and free energy satisfying schemes for diffusion equations with interaction potentials; since such equations are governed by a free energy dissipation law and are featured with non-negative solutions. Based on the non-logarithmic Landau reformulation of the model, we constructed a simple, easy-to-implement fully discrete numerical scheme (first order in time) which proved to satisfy all three desired properties of the continuous model: mass conservation, free energy dissipation and non-negativity, without a strict time step restriction. For a fully second order (in both time and space) scheme, we designed a local scaling limiter to restore solution positivity when necessary. Moreover, we proved rigorously that the limiter does not destroy the second order approximation accuracy. Numerical examples have demonstrated the superior performance of these schemes, in particular, the three solution properties numerically confirmed are consistent with our theoretical findings.
Figure 5.7  Solution evolution for Example 5.6.5 (sub-critical).
Figure 5.8  Solution evolution for Example 5.6.5 (super-critical).
Acknowledgments

This research was supported by the National Science Foundation under Grant DMS1312636.

5.8 Appendix: A refined time step bound for energy dissipation

Here we present an alternative proof of (5.3.9), i.e.,

$$\sum_{j=1}^{N} h_j \left( \frac{1}{2} g_j^n \rho_j^n - g_j^{n+1} \rho_j^{n+1} + \frac{1}{2} g_j^{n+1} \rho_j^{n+1} \right) \leq -\frac{\tau}{2} \sum_{j=1}^{N} \left( \frac{h_j \rho_j^{n+1} - h_j \rho_j^n}{\tau} \right) \psi_j^*, \quad \psi_j^* := \log(G_j^*), \quad (5.8.1)$$

in order to have a more precise bound on $\tau^*$. To this end we make the following assumptions:

- The matrix $W = (W_{i,j})$ with $W_{i,j} = W(x_i - x_j)$ is positive definite; both $W$ and $V$ are Lipschitz continuous.

- Meshes are shape-regular so that $\alpha^{-1} \leq h_i/h_j \leq \alpha$ for some $\alpha \geq 1$.

In addition, we assume that

$$\max_j h_j \|D_h W^{1/2}\|^2$$

is uniformly bounded with respect to $h_j$. Here $D_h$ denotes a finite difference operator $(D_h \phi)_j := \frac{\phi_{i+1} - \phi_i}{h_j}$. Our numerical tests suggest that such bound may always be true if $W$ is Lipschitz continuous, see Figure 5.9 for a typical example.

We now proceed to bound $\tau^*$. First using $h_j \rho_j^n = (W^{-1} g^n)_j$, the left hand side in (5.8.1) can be rewritten as

$$II_h^n = \frac{1}{2} \sum_{j=1}^{N} [W^{-1}(g_j^{n+1} - g^n)]_j (g_j^{n+1} - g^n) = \frac{1}{2} \|W^{-\frac{1}{2}} \phi\|^2, \quad (5.8.2)$$

where $\phi := g_j^{n+1} - g^n$. On the other hand, using the scheme (5.3.1) and summation by parts, we have

$$II_h^n = \frac{\tau}{2} \sum_{j=1}^{N} (C_j^* - C_{j+1/2}^*) \phi_j = -\frac{\tau}{2} \sum_{j=1}^{N-1} C_{j+1/2}^* (\phi_{j+1} - \phi_j)$$

$$\leq \frac{\tau}{2} \left( \sum_{j=1}^{N-1} h_j (C_j^*)^2 \right)^{\frac{1}{2}} \left( \sum_{j=1}^{N-1} \frac{(\phi_{j+1} - \phi_j)^2}{h_j} \right)^{\frac{1}{2}}, \quad (5.8.3)$$
where we used the Cauchy-Schwarz inequality. Under the assumption on $W$, we have
\[
\sum_{j=1}^{N-1} \frac{(\phi_{j+1} - \phi_j)^2}{h_j} \leq C||W^{-\frac{1}{2}}\phi||^2.
\] (5.8.4)

(5.8.4) when inserted into (5.8.3) and using (5.8.2) allows us to obtain
\[
II_h^2 \leq \frac{C\tau^2}{2} \sum_{j=1}^{N-1} h_j (C_j^{*+1/2})^2.
\] (5.8.5)

Finally (5.3.6), or (5.8.1) is satisfied if
\[
II_h^2 \leq \frac{\tau^2}{2} \sum_{j=1}^{N-1} h_j (C_j^{*+1/2})^2 \leq \frac{\tau}{2} \sum_{j=1}^{N-1} C_j^{*+1/2} (\psi_{j+1}^* - \psi_j^*).
\]

Thus it suffices to bound from below the following
\[
\frac{\sum_{j=1}^{N-1} C_j^{*+1/2} (\psi_{j+1}^* - \psi_j^*)}{C \sum_{j=1}^{N-1} h_j (C_j^{*+1/2})^2} \geq \frac{1}{C} \min_j \left\{ \frac{\psi_{j+1}^* - \psi_j^*}{h_j C_j^{*+1/2}} \right\}
\]
\[
= \frac{1}{C} \min_j \left\{ \frac{h_j}{h_{j+1/2}} \frac{\psi_{j+1}^* - \psi_j^*}{M_{j+1/2}^n (e^{\psi_{j+1}^*} - e^{\psi_j^*})} \right\}
\] (5.8.6)

using $\alpha = \max\{h_i/h_j\}$ and the mean-value theorem
\[
\geq \frac{1}{\alpha C} \min_j \left\{ \frac{1}{M_{j+1/2}^n e^{(\theta \psi_{j+1}^* + (1-\theta)\psi_j^*)}} \right\},
\]
where \( \theta \in (0, 1) \). By using \( M_{j+1/2}^n = e^{-V_j+1/2-g_j^n+1/2} \), we have

\[
\frac{1}{M_{j+1/2}^n e^{(\theta)\psi_{j+1}^n+(1-\theta)\psi_j^n}} = e^{[(\theta-1)(V_j+g_j^n)-\theta(V_j+1+g_j^n+1)]} e^{V_j+1/2+g_j^n+1/2} = e^{[V_j+1/2-(1-\theta)V_j-\theta V_j+1]} e^{[g_j^n+1/2-(1-\theta)g_j^n-\theta g_j^n+1]}
\]

\[
\geq \frac{1}{\max_{k,n} \rho_k^n} e^{-2\alpha L (1+\int \rho_0(x) dx) h}.
\]

Hence we may take

\[
\tau^* = \frac{1}{\alpha C \max_{k,n} \rho_k^n} e^{-2\alpha L (1+\int \rho_0(x) dx) h}.
\]

Remark 5.8.1. This is only a sufficient bound to ensure the energy stability, yet it indicates a need to carefully tune the time step when numerical density becomes large.

References


6.1 Conclusions

In this thesis, we constructed positive and energy dissipating schemes for solving PNP equations and FP equations with interaction potentials, motivated by the solution properties (positivity, mass conservation, and energy dissipation) of such equations. The non-logarithmic Landau reformulation of the model equations is important, which is suitable for constructing positive schemes.

The semi-discrete schemes, obtained by integrating the non-logarithmic Landau form under the standard finite volume framework, are shown to conserve mass, preserve solution positivity, and satisfy energy dissipation law. Although the model problems are nonlinear, we were able to construct fully discrete linear schemes by implicit-explicit time discretization. An important feature of our fully discrete first order schemes is solution positivity for arbitrary time steps. Moreover, we proved that our first order schemes satisfy discrete energy dissipation law with mild $O(1)$ time step restrictions.

We constructed fully second order (in both time and space) schemes by using prediction-correction methodology, where the first order schemes were used as prediction steps. Our second order schemes produce positive solutions for suitably small time steps, for larger time steps, we designed a local scaling limiter to restore solution positivity where necessary. We proved that the limiter does not destroy the approximation accuracy. Moreover, the limiter can be used for other high-order conservative finite difference or finite volume schemes.

Numerical examples have demonstrated the superior performance of these schemes, in particular, the three solution properties numerically confirmed are consistent with our theoretical findings. Our schemes were able to capture the PNP system’s layering phenomena and two-dimensional Patlak-Keller-Segel system’s blow-up phenomena as well.
6.2 Future Work

The FP equations with interaction potentials studied in Chapter 5 is the special case \( H(\rho) = \rho \log \rho \) of the general class of nonlinear nonlocal equation

\[
\partial_t \rho = \nabla \cdot \left( \rho \nabla \left( H' (\rho) + V(x) + W * \rho \right) \right). \tag{6.2.1}
\]

The schemes presented in this thesis may be applied to the general class (6.2.1) based on the reformulation

\[
\partial_t \rho = \nabla \cdot \left( M \nabla \frac{\rho}{M} \right),
\]

where \( M = \rho e^{-H' (\rho) - V(x) - W * \rho} \) for \( \rho \) away from zero. However, the numerical solution may be oscillatory at low density, it would be interesting to extend the present schemes to (6.2.1) even when near zero density.

The Jordan-Kinderlehrer-Otto (JKO) scheme introduced in [3] linear FP equations is a remarkable tool for the study of measure-valued solutions to gradient flows in density space. Such a JKO scheme naturally satisfies mass conservation, energy dissipation, and solution positivity. However, the Wasserstein distance involved in such schemes is known to be computationally subtle. Recent progress on the dynamic reformulation of the Wasserstein distance led to works [1, 2, 4], in which the JKO-type scheme has been explored to solve (6.2.1) with \( H(\rho) = \rho \log \rho \). A JKO scheme has been proposed for drift-diffusion systems in [5], we are working in numerically investigating such schemes for PNP systems.

References


