On entropy satisfying and maximum-principle-satisfying high-order methods for Fokker-Planck equations

Hui Yu
Iowa State University
On entropy satisfying and maximum-principle-satisfying high-order methods for 
Fokker–Planck equations

by

Hui Yu

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
James Evans
Scott Hansen
Zhi Jian Wang
Jue Yan

Iowa State University
Ames, Iowa
2013
Copyright © Hui Yu, 2013. All rights reserved.
DEDICATION

I would like to dedicate this thesis to my parents, Jianlin Yu and Yunxian Tu. I would also like to thank my friends and family for their loving guidance and support during the writing of this work.
# TABLE OF CONTENTS

**LIST OF TABLES** ................................................................. vii

**LIST OF FIGURES** .............................................................. x

**ACKNOWLEDGEMENTS** ......................................................... xii

**ABSTRACT** ............................................................................ xiii

**CHAPTER 1. GENERAL INTRODUCTION** ................................. 1

1.1 General background .......................................................... 1

1.2 Thesis organization .......................................................... 5

1.3 Entropy satisfying conservative methods for finitely extensible nonlinear elastic dumbbell model .............................................. 6

1.4 Entropy satisfying discontinuous Galerkin methods for Fokker–Planck equations .............................................................. 7

1.5 Maximum-principle-satisfying methods for Fokker–Planck equations ................................................................. 8

1.6 Entropy/energy stable methods for evolutionary dispersal models ................................................................. 11

**CHAPTER 2. AN ENTROPY SATISFYING CONSERVATIVE METHOD FOR THE FOKKER–PLANCK EQUATION OF FINITELY EXTENSIBLE NONLINEAR ELASTIC DUMBBELL MODEL** ................................. 14

2.1 Introduction ........................................................................ 14

2.1.1 Related work ............................................................. 18

2.1.2 Contents .................................................................... 20

2.2 One-dimensional Fokker–Planck Equation ............................ 20

2.2.1 Semidiscrete scheme ................................................ 20

2.2.2 Fully discrete scheme ................................................ 25

2.3 Extension to the multidimensional FENE model .................. 28
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1</td>
<td>Reformulation</td>
<td>28</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Discretization in $m \in B$</td>
<td>30</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Time discretization</td>
<td>38</td>
</tr>
<tr>
<td>2.4</td>
<td>Numerical implementation and results</td>
<td>42</td>
</tr>
<tr>
<td>2.4.1</td>
<td>One-dimensional tests</td>
<td>43</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Two-dimensional tests</td>
<td>46</td>
</tr>
<tr>
<td>2.5</td>
<td>Conclusion</td>
<td>53</td>
</tr>
</tbody>
</table>

**CHAPTER 3. THE ENTROPY SATISFYING DISCONTINUOUS GALERKIN METHOD FOR FOKKER–PLANCK EQUATIONS, WITH APPLICATIONS TO THE FINITELY EXTENSIBLE NONLINEAR ELASTIC DUMBBELL MODEL**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>56</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Related work</td>
<td>59</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Contents</td>
<td>62</td>
</tr>
<tr>
<td>3.2</td>
<td>Reformulation</td>
<td>62</td>
</tr>
<tr>
<td>3.3</td>
<td>One-dimensional Fokker–Planck Equation</td>
<td>63</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Semidiscrete DG</td>
<td>63</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Numerical flux and coercivity of $A$</td>
<td>64</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Entropy stability</td>
<td>67</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Fully discrete DG</td>
<td>70</td>
</tr>
<tr>
<td>3.3.5</td>
<td>A positive approximation</td>
<td>72</td>
</tr>
<tr>
<td>3.4</td>
<td>Extension to the multidimensional FENE Model</td>
<td>73</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Discretization</td>
<td>74</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Numerical flux and coercivity of $A$</td>
<td>76</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Entropy stability</td>
<td>80</td>
</tr>
<tr>
<td>3.5</td>
<td>Numerical tests</td>
<td>81</td>
</tr>
<tr>
<td>3.5.1</td>
<td>One-dimensional tests</td>
<td>81</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Two-dimensional tests</td>
<td>83</td>
</tr>
<tr>
<td>3.6</td>
<td>Concluding remarks</td>
<td>86</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Error and order of accuracy for Example 1 on a uniform mesh of (N) cells: (b = 16, \Delta t = 0.1, ) final time (t = 1.8).</td>
</tr>
<tr>
<td>2.2</td>
<td>Error and order of accuracy for Example 2 on a uniform mesh of (N) cells: (b = 50, \Delta t = 0.1, ) final time (t = 1.8).</td>
</tr>
<tr>
<td>2.3</td>
<td>Numerical convergence to the equilibrium solution measured by distances for Example 3: (b = 16, \varepsilon = 0.01\sqrt{b}, ) and (N = 160).</td>
</tr>
<tr>
<td>2.4</td>
<td>Relative entropy in Example 4: (b = 16, N = 640) and (\Delta t = 0.1).</td>
</tr>
<tr>
<td>2.5</td>
<td>Error and order of accuracy for Example 5: (b = 40, \kappa = 0, ) final time (t = 4) and the reference solution is given by (P = Q = 320).</td>
</tr>
<tr>
<td>2.6</td>
<td>Error and order of accuracy for Example 5: (k_{11} = 0.5, k_{12} = k_{21} = 0.15, \Delta t = 0.05, ) final time (t = 4).</td>
</tr>
<tr>
<td>2.7</td>
<td>Relative entropy in Example 6: (P = Q = 40, b = 40, \kappa = 0).</td>
</tr>
<tr>
<td>2.8</td>
<td>Numerical convergence to the equilibrium solution measured by distances for Example 7: (b = 16, k_{11} = 1.1, k_{12} = k_{21} = 0.15, \Delta t = 0.05, P = Q = 40).</td>
</tr>
<tr>
<td>2.9</td>
<td>Numerical convergence to the equilibrium solution measured by distances for Example 10: (b = 16, \gamma = 0.15, \Delta t = 0.05, P = Q = 40).</td>
</tr>
<tr>
<td>3.1</td>
<td>(L^\infty) and (L^2) error and order of accuracy of the (P^k) approximation for Example 1 on a uniform mesh of (N) cells: (b = 36, \Delta t = 0.001, ) final time (t = 0.1).</td>
</tr>
<tr>
<td>3.2</td>
<td>The mass and the relative entropy of the (P^3) approximation on a uniform mesh (N = 20, b = 36).</td>
</tr>
</tbody>
</table>
Table 3.3  The mass and the relative entropy of the $P^2$ approximation on a uniform mesh $P = Q = 10$. $b = 100, \kappa_{11} = 0, \kappa_{12} = -\kappa_{21} = 0.5$.  
Table 3.4  The mass and the relative entropy of the $P^2$ approximation on a uniform mesh $P = Q = 10$. $b = 100, \kappa_{11} = 0, \kappa_{12} = -\kappa_{21} = 0.5$.  
Table 3.5  The mass and the relative entropy of the $P^2$ approximation on a uniform mesh $P = Q = 10$. $b = 100, \alpha = 0$.  
Table 3.6  $L^\infty$ and $L^2$ error and order of accuracy of the $P^k$ approximation for Example 5 on a uniform mesh of $N$ cells. $\tilde{f}_0(r, \theta) = \left(1 - \frac{r^2}{\bar{r}}\right)^{\frac{2}{b}}, b = 10, \Delta t = 0.001$, final time $t = 0.4$.  
Table 4.1  The accuracy test of the $P^1$ approximation on a uniform mesh in the one-dimensional space. $\beta_0 = 2, \gamma = 0.75$, final time $t = 0.5$.  
Table 4.2  The accuracy test of the $P^2$ approximation on a uniform mesh in the one-dimensional space. $\beta_0 = 2, \beta_1 = 0.25, \gamma = 0.33$, final time $t = 0.1$.  
Table 4.3  The accuracy test of the $P^2$ approximation for the two-dimensional case. $(\beta_0, \beta_1) = (2, 0.25), \gamma = 0.3$, final time $t = 0.1$.  
Table 4.4  The effects of the parameter $\beta_0$ for the $P^2$ approximation in the one-dimensional space. $\beta_1 = 0.25, \gamma = 0.33$, final time $t = 0.1$.  
Table 4.5  The effects of the parameter $\gamma$ for the $P^2$ approximation in the one-dimensional space. $\beta_1 = 0.25$, final time $t = 0.1$.  
Table 4.6  The accuracy test of the $P^1$ approximation on a uniform mesh for the one-dimensional case. $b = 36, \beta_0 = 2$, final time $t = 0.1$.  
Table 4.7  The accuracy test of the $P^2$ approximation on a uniform mesh for the one-dimensional case. $b = 36, (\beta_0, \beta_1) = (2, 0.25)$, final time $t = 0.1$.  
Table 4.8  The entropy satisfying property of the $P^2$ approximation on a uniform mesh with the FENE spring potential. $N = 20, b = 36, (\beta_0, \beta_1) = (2, 0.25)$.  
Table 5.1  Error and order of accuracy of scheme (5.25) with the initial data (5.39) on a uniform mesh of $N$ cells, final time $t = 1.0$.  

84 84 85 86 117 118 118 119 120 121 121 122 148
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2</td>
<td>The entropy of the numerical solutions on a uniform mesh: $N = 80$</td>
</tr>
<tr>
<td>5.3</td>
<td>Error and order of accuracy of scheme (5.32)-(5.33) with initial data (5.39) on a uniform mesh of $N$ cells, final time $t = 1.0$</td>
</tr>
<tr>
<td>5.4</td>
<td>The energy of numerical solutions on a uniform mesh: $N = 80$</td>
</tr>
<tr>
<td>5.5</td>
<td>The discrete entropy on a uniform mesh: $N = 40 \times 40$</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

| Figure 2.1 | Diagram of the two-dimensional partition of $B$. | 30 |
| Figure 2.2 | $\tilde{f}_0(m) = 1 + \cos\left(\frac{2m\pi}{\sqrt{b} - \varepsilon}\right)$, $b = 16, \varepsilon = 0.01\sqrt{b}$. | 46 |
| Figure 2.3 | $\tilde{f}_0(m) = (b - |m|^2)^{\frac{1}{2}}$, $b = 16, k_{11} = 1.1, k_{12} = k_{21} = 0.15, \Delta t = 0.05$. | 49 |
| Figure 2.4 | $\tilde{f}_0(m) = \cos\left(3\pi |m|^2 b\right) + 1$, $b = 16, k_{11} = 1.1, k_{12} = k_{21} = 0.15, \Delta t = 0.05$. | 49 |
| Figure 2.5 | $b = 16, k_{11} = 1.1, k_{12} = k_{21} = 0.15, \Delta t = 0.05$, Left: $\tilde{f}_0(m) = (b - |m|^2)^{\frac{1}{2}}$. Right: $\tilde{f}_0(m) = \cos\left(3\pi |m|^2 b\right) + 1$. | 50 |
| Figure 2.6 | $\tilde{f}_0(m) = \delta_\varepsilon(m)$, $b = 16, \alpha = 1.1, \varepsilon = 2\Delta r, \beta = 2, P = Q = 40, \Delta t = 0.05$. | 51 |
| Figure 2.7 | $f_{eq}(m) = Z^{-1}_M(M(m), b = 16, P = Q = 40)$. | 52 |
| Figure 2.8 | The contours of $f_{i,j}^n$ at $t_n = 4$ where $\tilde{f}_0(m) = (b - |m|^2)^{\frac{1}{2}}$, $b = 16, P = Q = 40, \Delta t = 0.05$. | 53 |
| Figure 3.1 | Diagram of the two-dimensional partition of $B$. | 74 |
| Figure 3.2 | $b = 36, \Delta t = 0.1$. | 83 |
| Figure 3.3 | $\tilde{f}_0(r, \theta) = 1 - \frac{r^2}{\beta} \frac{36}{\beta} e^{\beta \left(0.2\cos(2\theta) - 0.7\sin(2\theta)\right)}, b = 100, \kappa_{11} = 0, \kappa_{12} = -\kappa_{21} = 0.5$. | 84 |
| Figure 3.4 | $b = 100, \alpha = 0.3$. | 88 |
| Figure 3.5 | $\tilde{f}_0(r, \theta) = \left(1 - \frac{r^2}{\beta}\right) \frac{36}{\beta} e^{\beta \left(0.5\cos(2\theta) - 0.7\sin(2\theta)\right)}, b = 100$, final time $t = 0.4$. | 88 |
| Figure 5.1 | For the figures in the first row, $u_0(x)$ is given by the random data; for the second row, $u_0(x)$ is given by the $\delta$-like function. $N = 80$. | 149 |
| Figure 5.2 | For the figures in the first row, $u_0(x)$ is given by the random data; for the second row, $u_0(x)$ is given by the $\delta$-like function. $N = 80$. | 151 |
| Figure 5.3 | Plot of $m(x) = 10 \left[0.25 - (x_1 - 0.5)^2 - (x_2 - 0.5)^2\right] e^{(x_1 - 0.5)^2-(x_2-0.5)^2} - 0.3$. | 152 |
Figure 5.4  Figures in the first row, $u_0(x)$ is given by the cos-like function; in the second row, $u_0(x)$ is given by the $\delta$-like function. $N = 20 \times 20, \lambda = 1 < \lambda^*$. 153

Figure 5.5  Figures in the first row, $u_0(x)$ is given by the cos-like function; in the second row, $u_0(x)$ is given by the $\delta$-like function. $N = 20 \times 20, \lambda = 1000 > \lambda^*$. 154

Figure 5.6  Plot of $m(x) = 0.5 \sin(\pi x_1) \sin(\pi x_2) + 0.6$. 155

Figure 5.7  The first row: $u$ component; the second row: $v$ component. $N = 40 \times 40$. 155
ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis.

First and foremost, Dr. Hailiang Liu for his guidance, patience and support throughout this research and the writing of this thesis. His insights and words of encouragement have often inspired me.

I would also like to thank my committee members for their efforts and contributions to this work: Dr. James Evans, Dr. Scott Hansen, Dr. Zhi Jian Wang and Dr. Jue Yan.

Last but not least, I want to thank all the colleagues, friends and staffs here in Department of Mathematics at Iowa State University.
Computation of Fokker–Planck equations with satisfying long time behavior is important in many applications. In this thesis, we design, analyze and implement entropy satisfying and maximum-principle-satisfying high-order numerical methods to solve the Fokker–Planck equation of the finitely extensible nonlinear elastic (FENE) dumbbell model for polymers, subject to homogeneous fluids, and the reaction-diffusion-advection equation arising in the evolution of biased dispersal of population dynamics. The design of each method is guided to satisfy three main properties, consisting of the nonnegativity principle, the mass conservation and the preservation of nonzero steady states. The relative entropy and the maximum principle are two powerful tools used to evaluate our methods, for instance, the steady state preservation can be ensured if the method is either entropy satisfying or maximum principle satisfying in the sense that the ratio of the solution to the equilibrium will stay in the same range as indicated by the initial data.

These schemes are constructed in several steps, including reformulation of the Fokker–Planck equation into its nonlogarithmic Landau form, spacial discretization by discontinuous Galerken (DG) methods and some Runge–Kutta time discretization. The special form of numerical fluxes motivated by those introduced in [H. Liu and J. Yan, Commun. Comput. Phys. 8(3), 2010, 541-564] is essential to incorporate desired properties into each scheme through choices of flux parameters.

In this thesis, we have obtained the following results.

1. For the Fokker–Planck equation of the FENE model, we propose an entropy satisfying conservative method which preserves all the three desired properties at both semidiscrete and discrete levels. This method is shown to be entropy satisfying in the sense that these schemes satisfy discrete entropy inequalities for both the physical entropy and the quadratic entropy. These ensure that the computed solution is a probability density, and the schemes are entropy
stable and preserve the equilibrium solutions.

2. We further develop an entropy satisfying DG method of arbitrary high order. Both semidiscrete and fully discrete methods are shown to satisfy two desired properties: mass conservation and entropy satisfying for the quadratic entropy, therefore preserving the equilibrium solutions. A positive numerical approximation is obtained with the same accuracy as the numerical solution through a reconstruction at the final time.

For both the finite volume scheme and the DG scheme we also prove the convergence of numerical solutions to the equilibrium solution as time tends to infinity. One- and two-dimensional numerical results are provided to demonstrate the good qualities of these schemes and effects of some canonical homogeneous flows.

3. We develop up to third-order accurate DG methods satisfying a strict maximum principle for a class of linear Fokker–Planck equations. A procedure is established to identify an effective test set in each computational cell to ensure the desired bounds of numerical averages during time evolution. This is achievable by properly choosing flux parameters and a positive decomposition of weighted cell averages. Based on this result, a scaling limiter for the DG method with Euler forward time discretization is proposed to solve both one- and multidimensional Fokker–Planck equations. As a consequence, the present scheme preserves steady states and provides a satisfying long time behavior. Numerical tests for the DG method are reported, with applications to polymer models with both Hookean and FENE potentials.

4. For Fokker–Planck equations with reaction such as the biased dispersal model in population dynamics, we develop entropy/energy stable finite difference schemes. For the numerical method to capture the long-time pattern of persistence or extinction, we use the relative entropy when the resource potential is logarithmic and explore the usual energy for other resource potentials. The present schemes are shown to satisfy three important properties of the continuous model for the population density: (i) positivity preserving, (ii) equilibrium preserving and (iii) entropy or energy satisfying. These ensure that our schemes provide a satisfying long-time behavior, thus revealing the desired dispersal pattern. Moreover, we present several numerical results which confirm the second-order accuracy for various resource potentials and underline the efficiency to preserve the large time asymptotic.
CHAPTER 1. GENERAL INTRODUCTION

Computation of Fokker–Planck equations with satisfying long time behavior is important in many applications and difficult in resolving solution structures induced by nonstandard forces. The goal of this thesis is to construct entropy satisfying and maximum-principle-satisfying high-order numerical methods to solve several significant Fokker–Planck equations.

1.1 General background

We consider Fokker–Planck equations of the following form
\[
\frac{\partial_t f}{\partial t} = \nabla_x \cdot \left( D \nabla_x f + F f \right), \quad x \in B, \quad t > 0,
\]  (1.1)
where \( D > 0 \) is the diffusion coefficient matrix, \( F \) is the underlying force, \( B \) is a bounded domain and \( f(t, x) \) is the probability distribution function (pdf) of the macroscopic variable.

There are many interpretations and derivations from Biology and other application areas that motivate (1.1); see Perthame (2007). For instance, it is a fundamental model in Chemistry at the molecular level. For active motions with oriented drift additionally to the Brownian motion, the equation is also called the drift-diffusion equation; in connection to the stochastic differential equations, it is called the Kolmogorov equation.

Our numerical study is motivated by the dumbbell models of polymers with finitely extensible nonlinear elastic (FENE) spring forces. This model couples the macroscopic incompressible Navier–Stokes equation for fluids and the microscopic Fokker–Planck equation for polymers. The FENE model is now widely used in numerical flow calculations to capture nonlinear rheological phenomena, both in the classical approach via a closed constitutive equation and in a modern approach in which the polymeric stress tensor is computed via Brownian dynamics.
In the case of the homogeneous fluids, the velocity gradient becomes a known matrix. The corresponding Fokker–Planck equation then decouples from the fluid equation and takes the form (1.1) with $D$ being an identity matrix and

$$F(x) = \frac{bx}{b - |x|^2} - 2\kappa x,$$

(1.2)

where $b$ denotes the maximum spring extension, the $d$-dimensional connector vector $x$ lies in a ball $B = B(0, \sqrt{b})$ and $\kappa$ is the velocity gradient matrix satisfying $Tr(\kappa) = 0$ due to the fluid incompressibility. In other words, the microscopic FENE model under homogeneous fluids and proper rescaling may be described by

$$\partial_t f = \nabla_x \cdot \left[ \nabla_x f + \left( \frac{bx}{b - |x|^2} - 2\kappa x \right) f \right], \quad x \in B, \ t > 0,$$

(1.3a)

$$\left[ \nabla_x f + \left( \frac{bx}{b - |x|^2} - 2\kappa x \right) f \right] \cdot \nu = 0, \quad x \in \partial B, t > 0,$$

(1.3b)

$$f(0, x) = f_0(x), \quad x \in B.$$

(1.3c)

The main difficulty in solving (1.3) is that $F$ as given in (1.2) is explosive at the boundary [Liu and Liu (2008), Liu and Shin (2012a)], which presents numerous challenges, both analytically and numerically. For theoretical results concerning the existence of solutions of the coupled system we refer to Liu and Shin (2012a), Masmoudi (2008), Zhang and Zhang (2006); see also the works Chupin (2009a,b, 2010) and the earlier works on this problem: Jourdain and Lelièvre (2003), Jourdain et al. (2004). For rigorous analysis of long-time asymptotics of the FENE model, see Jourdain et al. (2006); see Arnold et al. (2001) for entropy methods to study rate of convergence to equilibrium for Fokker–Planck type equations.

Another difficulty is that for general fluid configuration, the force may not be conservative, except for the irrotational fluid. For the irrotational fluid, $\kappa$ is symmetric. We then have

$$F = \nabla_x U, \quad U = -\frac{b}{2} \log(b - |x|^2) - x^T \kappa x,$$

leading to the standard form of the Fokker–Planck equation

$$\partial_t f = \nabla_x \cdot (\nabla_x f + \nabla_x U f), \quad x \in B, \ t > 0.$$

(1.4)
In spite of these difficulties, some elaborate numerical algorithms based on spectral methods were recently developed for the Fokker–Planck equation of the FENE model in Chauvière and Lozinski (2003, 2004a,b). A spectral Galerkin approximation was further introduced in Knezevic and Suli (2009) based on a weighted weak formulation for \( f(b - |x|^2)^{-\frac{b}{2}} \). An improved weighted formulation was proposed in Shen and Yu (2012) in terms of \( f(b - |x|^2)^{-\frac{s}{2}} \) for \( 1 < s \leq b \), leading to a different spectral Galerkin algorithm. We note that this weighted formulation was also used for specific values of \( s \) in Chauvière and Lozinski (2003, 2004a,b). The methods in Knezevic and Suli (2009) and Shen and Yu (2012) have provable stability results for certain weighted integrable initial data. However, these methods in general do not satisfy either the entropy or the maximum principle. We also point out that most numerical methods developed for the Fokker–Planck equation have been based on the form of (1.3a); see, for example, Ammar et al. (2006), Ammar et al. (2007), Fan (1985) and Warner (1972). In our current investigation we shall use the nonlogarithmic Landau formulation as detailed later in (1.9). Another powerful approach for kinetic PDEs is the use of moment closure approximation. Its main advantage is the reduction of the computational cost since with this method one chooses not to solve the kinetic equation in the full high dimensional phase space, instead, one seeks to solve the moment equation in physical space. For some special configuration solutions with small flow rates associated with the FENE model (1.3), the use of moment closure approximations has been investigated by several authors; see, e.g., Du et al. (2005), Hyon et al. (2008), Herrchen and Öttinger (1997), Samaey et al. (2011) and Wang et al. (2008).

In order to construct a high order numerical method with satisfying long time behavior, we recall some main properties of the solution of (1.1) which are the nonnegativity principle, the mass conservation and the existence of nonzero steady states, i.e.,

\[
\begin{align*}
\int_B f(t,x) \, dx &= \int_B f_0(x) \, dx \quad \forall t > 0, \\
(1.5) && \\
\int_B f(t,x) \, dx &= \int_B f_0(x) \, dx \quad \forall t > 0, \\
(1.6) && \\
There exist a family of steady states. \\
(1.7)
\end{align*}
\]

For \( F = \nabla_x U \), the steady states in (1.7) have the following explicit form,

\[
M(x) = Ce^{-U(x)} \text{ where } C \text{ is some constant.}
\]
In such a case, both property (1.5) and (1.7) are implied by the strict maximum principle, i.e., if
\[ c_1 = \min_{x \in B} \frac{f_0(x)}{M(x)}, \quad c_2 = \max_{x \in B} \frac{f_0(x)}{M(x)}, \quad (1.8) \]
then \( \frac{f(t,x)}{M(x)} \in [c_1, c_2] \) for any \( x \in B \) and \( t > 0 \). These properties are also naturally desired for numerical schemes solving (1.3). We shall develop such methods in this thesis.

Our methodology in constructing entropy satisfying schemes is through several steps: reformulation of the Fokker–Planck equation, the spatial discretization by the discontinuous Galerkin approximation and the Runge-Kutta high order time discretization. More precisely, we explore the nonlogarithmic Landau formulation of (1.4), namely, the reformulation in terms of \( g = \frac{f}{M} \),
\[ M \partial_t g = \nabla_x \cdot (M \nabla_x g). \quad (1.9) \]
Then the maximum principle (1.8) reduces to
\[ c_1 \leq g_0 \leq c_2 \implies c_1 \leq g(t,x) \leq c_2 \quad \forall t > 0, \quad (1.10) \]
while the mass conservation needs to be measured by \( \int_B M(x)g(t,x) \, dx = \text{constant} \). The quadratic relative entropy becomes a weighted \( L^2 \) energy: \( \int M g^2 \, dx \). Thus some standard Galerkin approximation can be applied. Throughout the thesis, we will illustrate the advantages of this reformulation and present several numerical methods that are able to preserve the main properties (1.5)-(1.7).

Regarding the spatial discretization by the DG approximation, we would like to review some recent related developments. The application of DG methods to hyperbolic problems has been quite successful; see, e.g., Reed and Hill (1973) for solving linear equations and Cockburn and Shu (1989), Cockburn et al. (1989), Cockburn et al. (1990) for solving nonlinear equations. However, the application of DG methods to diffusion problems has been a challenging task because of the subtle difficulty in defining an appropriate numerical flux for the solution gradient; see the earlier works Arnold (1982), Baker (1977), Wheeler (1978) using the interior penalty (IP) method. In the past decade there has been a renewed interest in developing DG methods to solve the diffusion problem, including the method originally proposed by Bassi and Rebay (1997) for compressible Navier–Stokes equations, its generalization called the local
discontinuous Galerkin (LDG) methods introduced in Cockburn and Shu (1998) and further studied in Castillo et al. (2000), Cockburn and Dawson (2000) and Cockburn et al. (2001), as well as the method introduced by Baumann and Oden (1999) and Oden et al. (1998). We refer to Arnold et al. (2002) for the unified analysis of DG methods for elliptic problems and background references for the IP methods. The direct discontinuous Galerkin (DDG) methods introduced in Liu and Yan (2009, 2010) adopt a different strategy, which is to solve the higher order PDE directly by the DG discretization with the special numerical flux for the solution gradient, yet without rewriting the equation into a first order system. There are other recent works sharing the direct feature, such as those in van Leer and Nomura (2005), Gassner et al. (2007) and Cheng and Shu (2007), all based on certain weak formulation derived from repeated integration by parts for the diffusion term. More general information about DG methods for elliptic, parabolic and hyperbolic partial differential equations can be found in the recent books and lectures notes [Hesthaven and Warburton (2007), Li (2006), Riviére (2008), Shu (2009)].

In our spatial discretization of (1.9), we follow the DDG methodology originated in Liu and Yan (2009, 2010). The special form of the DDG numerical fluxes plays a vital role for us to incorporate the desired properties of the pdf into the scheme formulation through careful analysis of flux parameters.

1.2 Thesis organization

The thesis is organized as follows. In Chapter 2, the FENE dumbbell model is investigated and we develop an entropy satisfying conservative method which preserves all the three properties of the pdf and satisfies the entropy inequality. A series of numerical tests demonstrate the good qualities of the schemes and the effects of some canonical flows. In Chapter 3, entropy satisfying DG methods are designed and analyzed. Numerical examples show the optimal order of accuracy and the satisfying long-time behavior. Chapter 4 is reserved for the maximum-principle-satisfying methods using the DDG framework. We propose a new decomposition of weighted cell averages and develop a procedure such that the algorithm can produce a maximum-principle-satisfying numerical solution without destroying the accuracy. In Chapter 5, we use the entropy/energy structure to design finite difference methods for the
reaction-diffusion-advection equations that arise in the evolution of biased dispersal of population dynamics. The content of each chapter is summarized in the following four sections.

1.3 Entropy satisfying conservative methods for finitely extensible nonlinear elastic dumbbell model

We consider the FENE model with fluid effects, for which (1.9) needs to be refined as

$$M \partial_t g = \nabla_x \cdot (M \nabla_x g - 2\kappa^a x M g), \quad (1.11)$$

where \(\kappa^a\) denotes the antisymmetric part of \(\kappa\) and \(M = (b - |x|^2)^{b/2} e^T \kappa^s x\) with \(\kappa^s\) being the symmetric part of \(\kappa\).

The reformulation (1.11) is no longer singular and suitable for using Galerkin approximations, even though \(M \partial_t g\) is degenerate at the boundary. It is also convenient to apply the relative entropy in the design of our numerical method, for

$$E(t) := \int_B \frac{f^2}{M} dx = \int_B M g^2 dx.$$

If \(\kappa\) is normal, it can be shown that the relative entropy satisfies the inequality:

$$E(t) + \int_0^t \int_B M |\nabla_x g|^2 dx d\tau \leq E(0) \quad \forall t > 0.$$

This entropy dissipation relation ensures that the relative entropy is decreasing in time and as time evolves the solution is expected to converge towards the equilibrium, i.e.,

$$\lim_{t \to \infty} f(t, x) = CM(x),$$

for some \(C > 0\).

Next we illustrate our idea using the one-dimensional case, for which (1.11) is the same as (1.9) since \(\kappa = 0\). Integration of (1.9) on each computational cell yields

$$\frac{d}{dt} \left( \frac{1}{h} \int_{I_j} M g dx \right) = \frac{1}{h} M \partial_x g \bigg|^{x_{j+\frac{1}{2}}}_{x_{j-\frac{1}{2}}},$$

where the cell is defined as \(I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]\) and \(h\) is the uniform mesh size, which naturally leads to a finite volume scheme: compute \(\{g_j\}\) which approximates \(\{\bar{g}_j := \frac{1}{h} \int_{I_j} g dx\}\) so that

$$M_j \frac{d}{dt} g_j = \frac{1}{h} \left( J_{j+\frac{1}{2}} - J_{j-\frac{1}{2}} \right),$$
where
\[ M_j = M(x_j) \] and the numerical flux \( J_{j+\frac{1}{2}} = M \partial_x g \bigg|_{x_{j+\frac{1}{2}}} = M_{j+\frac{1}{2}} \frac{g_{j+1} - g_j}{h} \).

And the approximation \( \{f_j\} \) is obtained by using \( f_j = M_j g_j \). The semidiscrete scheme is shown to satisfy all the three desired properties: positivity preserving, mass conservation and entropy satisfying in the sense that the discrete entropy inequality holds true. The fully discrete schemes with backward Euler temporal discretization is shown to have all three properties unconditionally for one-dimensional case and under certain Courant-Friedrichs-Lewy (CFL) condition for the two-dimensional case due to the antisymmetric part of \( \kappa \). More details can be found in Liu and Yu (2012a).

### 1.4 Entropy satisfying discontinuous Galerkin methods for Fokker–Planck equations

Next we develop high order schemes for Fokker–Planck equations that can satisfy the entropy inequality as well. We propose the following DG method for one space dimension: find \( g_h \) in the solution space \( V_h^k \) consisting of piecewise polynomials up to degree \( k \) such that
\[
\sum_{j=1}^{N} \int_{I_j} M \partial_t g_h v \, dx + A(g_h, v) = 0 \quad \forall v \in V_h^k, \tag{1.12}
\]
where the bilinear operator \( A \) is defined as
\[
A(g_h, v) = \sum_{j=1}^{N} \int_{I_j} M \partial_x g_h \partial_x v \, dx + \sum_{j=1}^{N-1} M \left( \partial_x g_h [v] + [g_h] \{ \partial_x v \} \right) \bigg|_{x_{j+\frac{1}{2}}}
\]
with the direct DG type numerical flux introduced in Liu and Yan (2009, 2010)
\[
\partial_x g = \frac{\beta_0}{h} [g] + \{ g \} + h \beta_1 [\partial_x^2 g]. \tag{1.13}
\]
Here \([g]\) denotes the jump at the cell interface and \( \{ g \} \) the average. The form (1.13) makes the numerical flux both consistent and conservative. The difficulty is how to choose the parameters \((\beta_0, \beta_1)\) such that it ensures the entropy stability and enforces the high order accuracy of the method. To this end, we define the discrete energy norm of \( v \in V_h^k \)
\[
\|v\|_E^2 = \sum_{j=1}^{N} \int_{I_j} M |\partial_x v|^2 \, dx + \sum_{j=1}^{N-1} \frac{\beta_0}{h} M [v]^2 \bigg|_{x_{j+\frac{1}{2}}} \tag{1.14}
\]
and the quantity
\[\Gamma_j := \Gamma(\beta_1, w_j) = \sup_{u \in P^{k-1}([-1,1]) \neq 0} \frac{(u(1) - 2\beta_1 \partial_\xi u(1))^2}{\int_{-1}^{1} w_j(\xi) u^2(\xi) \, d\xi}.\] (1.15)

Here the weight functions \(w_j(x)\) are defined as
\[w_j(\xi) = \min\{w_{jl}(\xi), w_{jr}(\xi)\} \quad \text{for } 1 \leq j \leq N - 1\]
with
\[w_{jl}(\xi) = M \left( x_j + \frac{h}{2} \xi \right) \quad \text{and} \quad w_{jr}(\xi) = M \left( x_{j+1} - \frac{h}{2} \xi \right) \quad \text{on } [-1, 1].\]

Our main result is the following. If on each cell interface \(x_{j+\frac{1}{2}}\) for \(1 \leq j \leq N - 1\),
\[\beta_0 > 2\Gamma_j M(x_{j+\frac{1}{2}}),\] (1.16)
then the semidiscrete scheme (1.12) conserves the mass and satisfies the entropy inequality:
\[\frac{d}{dt} E(t) \leq -\gamma \|g_h\|_{E}^2 \leq 0 \quad \text{for some } \gamma \in (0, 1).\] (1.17)

Then we propose a class of semiimplicit discrete DG schemes:
\[\sum_{j=1}^{N} \int_{I_j} M \frac{g^{n+1} - g^n}{\Delta t} v \, dx = -A(g^*, v) \quad \forall v \in V_h^k,\] (1.18)
where \(g^*\) is defined as
\[g^* = \eta g^{n+1} + (1 - \eta) g^n, \quad \frac{1}{2} \leq \eta \leq 1.\]

They are shown to conserve the mass and satisfy the entropy property. A positive numerical approximation is obtained through a reconstruction at the final time with the same accuracy as the numerical solution. Both one- and two-dimensional numerical results are provided to demonstrate the optimal order of accuracy and a satisfying long-time behavior, as well as effects of some canonical homogeneous flows in the FENE dumbbell model.

### 1.5 Maximum-principle-satisfying methods for Fokker–Planck equations

For the Fokker–Planck equations of the form
\[\partial_t f = \nabla_x \cdot (\nabla_x f + \nabla_x U f), \quad x \in B, \ t > 0,\]
the reformulation in terms of $g = \frac{f}{M}$ with $M(x) = e^{-U(x)}$ is
\[ M \partial_t g = \nabla_x \cdot (M \nabla_x g). \]

The numerical flux (1.13) makes the design of maximum-principle-satisfying high order schemes possible.

Recently a maximum-principle-satisfying framework has been established for scalar conservation laws in Zhang and Shu (2010), yet the key step of using the first order schemes as building blocks can not be applied to second order PDEs such as (1.9) in a straightforward manner. In spite of this difficulty, a non-conventional technique was introduced in Zhang et al. (2012) to design a high order maximum-principle-satisfying finite volume scheme for convection-diffusion equations. Yet, as pointed out in Zhang et al. (2013), it is not obvious how to generalize this non-conventional technique to DG methods. The maximum-principle-satisfying DG scheme on triangular meshes was subsequently proposed in Zhang et al. (2013), but the scheme is only second order accurate ($k = 1$).

Our main contribution is the third order accurate maximum-principle-satisfying schemes for one- and multidimensional Fokker–Planck equations, in the sense that the numerical approximation to $f$ never goes out of the range $[c_1, c_2]e^{-U}$ of the initial condition. Our scheme uses the simple forward Euler temporal discretization, allowing for easy and practical implementation and easy generalization from one to multiple dimensions. The scaling limiter introduced in Zhang and Shu (2010) is modified based on the weighted cell averages to control the maximum/minimum of the reconstructed polynomials.

We start with the forward Euler temporal discretization: to find $g_h \in V_h$ such that for any $v \in V_h$ and $I_j$,
\[
\int_{I_j} M \frac{g_h^{n+1} - g_h^n}{\Delta t} v \, dx = - \int_{I_j} M \partial_x g_h^n \partial_x v \, dx + M \left[ \frac{\partial_x g_h^n}{2} v + (g_h^n - \{g_h^n\}) \partial_x v \right]_{x_j - \frac{1}{2}}^{x_j + \frac{1}{2}},
\]  
(1.19)
where the numerical flux is defined as in (1.13). Particularly, we study the time evolution of the cell average, i.e., when $v \equiv 1$:
\[
\langle g_h^{n+1} \rangle_j = \langle g_h^n \rangle_j + \lambda h \left[ M \partial_x g_h^n \right]_{x_j - \frac{1}{2}}^{x_j + \frac{1}{2}},
\]  
(1.20)
where we used the notation

\[
\langle \phi \rangle_j := \frac{1}{h} \int_{I_j} M(x) \phi(x) \, dx
\]

and \( \lambda = \frac{\Delta t}{h^2} \) is the mesh ratio. Here parameters \((\beta_0, \beta_1)\) are essentially used to guarantee the existence of some controlled points in each cell so that the numerical solutions preserve the mass and positivity after time evolution and a reconstruction using the same controlled point values. More precisely, our procedure includes two crucial ingredients:

1. Decompose the weighted cell average of polynomials of degree \( k \) in terms of \( k+1 \) controlled points in each cell with positive coefficients. These points form a test set \( S_j \) in each cell \( I_j \), over which the decomposition appears as

\[
\langle \phi \rangle_j = \sum_{x^i \in S_j} \hat{\omega}^i \phi(x^i) \quad \forall \phi(x) \in P^k(I_j).
\]

2. Represent the numerical flux (1.13) in terms of solution values at the same controlled points from two neighboring cells in the following way

\[
h \hat{\partial}_x \phi|_{j+\frac{1}{2}} = \sum_{x^i \in S_{j+1}} \alpha^i_+ \phi(x^i) - \sum_{x^i \in S_j} \alpha^i_- \phi(x^i) \quad \forall \phi(x) \in P^k(I_j).
\]

The coefficients \( \hat{\omega}^i \), \( \alpha^i_+ \) and \( \alpha^i_- \) depend only on \( M(x) \) and \( k \) and can be made positive through choices of parameters \((\beta_0, \beta_1)\) for \( k \leq 2 \).

The above procedure when inserted into (1.20) enables us to show that, under a suitable CFL condition, the simple forward Euler will keep the property \( \langle g^n_h \rangle_j \in [c_1, c_2] \) and the validity of the maximum principle if we use the DG polynomials, thus maintaining uniform \( k+1 \) order accuracy. The proposed limiter based on the weighted cell average replaces the definition of maximum and minimum in each cell by those on a test set \( S_j \) of \( k+1 \) points, we can easily implement it for polynomials of degree \( k \).

The main conclusion drawn from our result is as follows: by applying the limiter or the simplified version which avoids the evaluation of extrema of polynomials to a DG scheme solving one- or multidimensional Fokker–Planck equations, with the time evolution by a SSP Runge-Kutta method, we obtain a third order accurate scheme with the strict maximum principle in the sense that the numerical solution \( f^n_h \) never goes out of the range \([c_1, c_2]e^{-U}\), where \( c_1 \) and \( c_2 \) are defined in (1.8).
1.6 Entropy/energy stable methods for evolutionary dispersal models

Using the entropy/energy structure, we develop second-order finite difference methods for a class of reaction-diffusion-advection equations arising in the evolution of biased dispersal of population dynamics. For one species, we study the following model

\[ \partial_t u = \nabla \cdot (\nabla u + u \nabla P) + \lambda u(m - u) \quad \text{in } \Omega \times (0, \infty), \]  

(1.21)

where \( P = P(m) \) which we call resource potential, reflects the movement tendency of the population. The time evolution is subjected to both the initial density \( u(x, 0) = u_0(x) \) and the zero flux boundary condition

\[ (\nabla u + u \nabla P) \cdot \nu = 0 \quad \text{on } \partial \Omega \times (0, \infty), \]  

(1.22)

where \( \nu \) is the outward normal vector on the boundary \( \partial \Omega \) which is assumed to be smooth.

The main theoretical result is about the existence of the threshold \( \lambda^* \).

**Theorem 1.6.1.** Suppose that \( m \in C^2(\bar{\Omega}) \) is positive somewhere in \( \Omega \) and \( P \) is smooth in \( m \). There exists a unique \( \lambda^* > 0 \) and a positive equilibrium solution to (1.21) if and only if

\[ \int_{\Omega} m e^{-P} \, dx < 0. \]

Moreover, \( \lambda^* = 0 \) if and only if \( \int_{\Omega} m e^{-P} \, dx \geq 0 \).

1. If \( 0 < \lambda \leq \lambda^* \), all nonnegative solutions of (1.21) decay toward zero as \( t \to \infty \).

2. If \( \lambda > \lambda^* \), the positive equilibrium is globally attractive among nonzero nonnegative solutions.

We shall design finite difference schemes to capture the large time pattern of solutions to (1.21) as stated in Theorem 1.6.1. In order to achieve this goal, following Liu and Yu (2012a), we rewrite (1.21) in terms of \( g = u/M \) with \( M = e^{-P} \):

\[ M \partial_t g + P g = 0 \quad \text{in } \Omega, \]  

(1.23)

where the operator \( P \) is defined as

\[ P g = -\nabla \cdot (M \nabla g) - \lambda M g(m - Mg). \]  

(1.24)
The initial condition is $g_0(x) = \frac{u_0(x)}{M(x)}$ and on the boundary $\partial \Omega$ we have $M \partial_n g = 0$.

For log potential $P(m) = -\log m$, the entropy is defined as

$$E[g] = \int_{\Omega} m \left[ g \log g + (1 - g) \right] \, dx.$$ 

Both the semidiscrete and fully discrete schemes are shown to preserve the positivity and the equilibrium solutions.

For other potential, we follow the standard gradient flow idea [Cantrell and Cosner (2004)] by using the functional

$$V[g] = \int_{\Omega} \left[ \frac{1}{2} M |\nabla g|^2 - G(x, g(x, t)) \right] \, dx,$$

where $G$ is chosen as

$$G(x, g) = \lambda \int_0^g F(s, m) \, ds = \frac{\lambda}{2} m(x) M(x) g^2 - \frac{\lambda}{3} M^2(x) g^3.$$

So (1.23) can be rewritten as

$$M \partial_t g = \nabla \cdot (M \nabla g) + \partial_t G.$$  

We develop a finite difference scheme which consists of two steps: prediction and correction. The scheme is proved to satisfy the energy inequality unconditionally and preserve the equilibrium solution under certain restriction on the time step.

For the two-species system,

$$\begin{cases} 
    u_t &= \alpha \nabla \cdot (\nabla u + u \nabla P) + u(m - u - v) \quad \text{in} \quad \Omega, \\
    v_t &= \beta \nabla \cdot (\nabla v + v \nabla Q) + v(m - u - v) \quad \text{in} \quad \Omega, \\
    \partial_n u + u \partial_n P &= \partial_n v + v \partial_n Q = 0 \quad \text{on} \quad \partial \Omega, \\
    u(x, 0) &= u_0(x), \quad v(x, 0) = v_0(x) \quad \text{in} \quad \Omega.
\end{cases}$$

We only consider the ideal free distribution where $m > 0$ on $\Omega$, $P(m) = -\log m + C$ and $Q(m) \neq -\log m + C$. Here $C$ is an arbitrary constant. For this particular case, $(m, 0)$ is the globally attractive equilibrium solution. Further details on this model can be found in Cantrell et al. (2010). The relative entropy $E[u, v]$ defined as

$$E[u, v] = \int_{\Omega} u \, dx + \int_{\Omega} v \, dx - \int_{\Omega} m(x) \log u \, dx$$
is nonincreasing in time. Our fully discrete scheme is able to satisfy the entropy inequality unconditionally and preserve the positivity under certain restriction on the time step.

Numerical tests show that all the above schemes are second-order accurate and provide a satisfying long-time behavior.
CHAPTER 2. AN ENTROPY SATISFYING CONSERVATIVE METHOD FOR THE FOKKER–PLANCK EQUATION OF FINITELY EXTENSIBLE NONLINEAR ELASTIC DUMBBELL MODEL

A paper published by SIAM Journal on Numerical Analysis

Hailiang Liu and Hui Yu

Abstract

In this paper, we propose an entropy satisfying conservative method to solve the Fokker–Planck equation of FENE dumbbell model for polymers, subject to homogeneous fluids. Both semidiscrete and fully discrete schemes satisfy all three desired properties: (i) mass conservation, (ii) positivity preserving, and (iii) entropy satisfying in the sense that these schemes satisfy discrete entropy inequalities for both the physical entropy and the quadratic entropy. These ensure that the computed solution is a probability density, and the schemes are entropy stable and preserve the equilibrium solutions. We also prove convergence of the numerical solution to the equilibrium solution as time becomes large. Zero-flux at boundary is naturally incorporated and boundary behavior is resolved sharply. Both one- and two-dimensional numerical results are provided to demonstrate the good qualities of the schemes and the effects of some canonical homogeneous flows.

2.1 Introduction

Dumbbell models with finitely extensible nonlinear elastic (FENE) spring forces are now widely used in numerical flow calculations to capture nonlinear rheological phenomena, both
in the classical approach via a closed constitutive equation and in a modern approach in which the polymeric stress tensor is computed via Brownian dynamics (BD) simulations; see Bird and West (1995), Laso and Öttinger (1993). For the dumbbell model the configuration probability density function (pdf) yields information on the probability of finding a dumbbell with a given configuration at a particular material point, hence solving the Fokker–Planck equation directly is desirable, as long as it is feasible [Owens and Phillips (2002)].

The original empirical FENE spring potential,

\[ \Psi(m) = -\frac{Hb_0}{2} \log \left( 1 - \frac{|m|^2}{b_0^2} \right), \]  

(2.1)

was first proposed by Warner in Warner (1972), where \( H \) is the spring constant and \( m \) is the \( d \)-dimensional connector vector of the beads with \( m \in B := B(0, \sqrt{b_0}) \), a ball in \( \mathbb{R}^d \) with radius \( \sqrt{b_0} \) denoting the maximum spring extension. It exhibits, for small extensions, the expected linear behavior and a finite length \( b_0 \) in the limit of an infinite force.

This paper is concerned with the numerical solution of the Fokker–Planck equation of the FENE dumbbell model for the pdf \( f = f(x, m, t) \),

\[ \partial_t f + (v \cdot \nabla_x)f + \nabla_m \cdot (\nabla_x v_m f) = \frac{2}{\zeta} \nabla_m \cdot (\nabla_m \Psi(m)f) + \frac{2k_BT}{\zeta} \Delta_m f, \]  

(2.2)

where \( x \in \mathbb{R}^d \) is the macroscopic Eulerian coordinate, \( v(x, t) \), the fluid velocity, is usually governed by the incompressible Navier–Stokes equation, \( \zeta \) is the friction coefficient of the dumbbell beads, \( T \) is the absolute temperature and \( k_B \) is the Boltzmann constant. We refer to Chapters 11 and 13 of Bird et al. (1987) for a comprehensive survey of the physical background and Degond and Liu (2009) for some augmented models with inertial forces.

Throughout this paper we consider only homogeneous flows. Therefore the velocity field of the fluid can be written as \( v = \kappa x \), where \( \kappa = \nabla v \) is independent of the position vector \( x \) in the fluid and has zero trace since we assume the fluid to be incompressible. Let the flow map be defined as

\[ \partial_t X(y; t) = v(X(y; t), t), \quad X(y; 0) = y. \]

Along the flow map, with a suitable scaling and \( b_0 \rightarrow b = \frac{Hb_0}{k_BT} \), we arrive at the following
equation for \( f(m, t) := f(X(y; t), m, t) \) for each fixed \( y \),
\[
\partial_t f = \frac{1}{2} \nabla_m \cdot \left[ \nabla_m f + \left( \frac{bm}{b-|m|^2} - 2\kappa m \right) f \right], \quad m \in B, \ t > 0, \quad (2.3a)
\]
\[
f(m, 0) = f_0(m), \quad m \in B, \quad (2.3b)
\]
\[
f(m, t) = o(b - |m|^2), \quad m \in \partial B, t > 0. \quad (2.3c)
\]

From now on \( \kappa \) is assumed to be a trace-free \( d \times d \) matrix, i.e., \( Tr(\kappa) = 0 \). Boundary requirement (2.3c) is imposed to ensure the existence and uniqueness of the weak solution to (2.3); see Liu and Shin (2012b).

The singularity of the Fokker–Planck equation near \( |m| = \sqrt{b} \) makes the boundary issue rather subtle [Liu and Liu (2008)] and presents numerous challenges, both analytically and numerically. These issues are particularly important in solving the coupled Navier–Stokes–Fokker–Planck system, in which the behavior of the polymer distribution near boundaries is of significance. Consequently, computing with sharp resolution and stability near boundaries is a major goal. On the other hand the pdf is the practically relevant solution [Liu and Shin (2012a)] for the underlying Fokker–Plack equation. It is therefore desirable to design a method which preserves three important properties of the pdf: constant integral (mass conservation), positivity preserving and entropy satisfying in the sense that entropy inequalities are satisfied at the discrete level. In this paper, we develop such a method.

A key concept in the design of our numerical method is the relative entropy. To illustrate the idea, we reformulate the Fokker–Planck equation (2.3a). If \( \kappa \) is normal in the sense that it commutes with its transpose, i.e., \( \kappa \kappa^T = \kappa^T \kappa \), one can verify that the equilibrium solution can be determined explicitly as
\[
M(m) = (b - |m|^2)^{\frac{3}{2}} \exp(m^T \kappa^s m), \quad (2.4)
\]
where \( \kappa^s \) is the symmetric part of \( \kappa \). Let \( \kappa^a \) be the antisymmetric part of \( \kappa \); then the Fokker–Planck equation can be rewritten as
\[
\partial_t f = \frac{1}{2} \nabla_m \cdot \left[ \nabla_m (M \nabla_m g - 2\kappa^a mf) \right], \quad f = gM. \quad (2.5)
\]

Using the zero flux boundary condition (2.6), it can be shown that the relative entropy
\[
E(t) := \int_B \frac{f^2}{M} \, dm = \int_B Mg^2 \, dm
\]
satisfies the following inequality
\[ E(t) + \int_0^t \int_B M|\nabla_m g|^2 \, dmd\tau \leq E(0) \quad \forall t > 0. \]

This entropy dissipation relation ensures that the relative entropy is decreasing in time, and as time evolves the solution is expected to converge towards the equilibrium, i.e.,
\[ \lim_{t \to \infty} f(t, m) = CM(m) \]
for some \( C > 0 \). One may also use the physical entropy defined by
\[ E_p(t) = \int_B f \log \left( \frac{f}{M} \right) \, dm, \]
which satisfies the following entropy dissipation equation
\[ E_p(t) + \int_0^t \int_B M \frac{|\nabla_m g|^2}{g} \, dmd\tau \leq E_p(0) \quad \forall t > 0. \]

Note that the physical entropy is bounded as long as \( E \) is bounded since
\[ E_p(t) \leq \int_B f \left( \frac{f}{M} - 1 \right) \, dm = E(t) - \int_B f \, dm = E(t) - \int_B f_0(m) \, dm. \]

For an initial density with \( E(0) < \infty \) it suffices to consider the quadratic entropy \( E(t) \), which is particularly convenient to use for higher order methods. As the first step, we shall design a finite volume scheme based on (2.5) and show positivity and stability properties in terms of the relative entropy. More precisely, for both semidiscrete and fully discrete schemes presented in this work, we are able to prove the entropy stability for both quadratic and physical entropy, based on which we also prove the long time convergence. Existence of positive solutions is established as well.

For nonhomogeneous flows, which is the case when considering the coupled problem with the Navier–Stokes equation, we may apply the method developed in this paper using operator splitting. For instance, for each fixed \( m \), one may solve the transport equation
\[ \partial_t f + \nabla \cdot (vf) = 0 \]
with \( v \) obtained from solving the Navier–Stokes equation. With the obtained \( f \) as initial data, then one further solves the Fokker–Planck equation with \( \kappa = \nabla_x v(x, t) \). Note that this treatment using operator splitting techniques is a standard tool in fluid simulations; see Chauvière and Lozinski (2004a,b).
2.1.1 Related work

The regime of physical interest is $b > 2$, for which the boundary requirement (2.3c) was shown to be a sharp requirement for the solution to remain a probability density; see Liu and Shin (2012b). Moreover, this condition is equivalent to the zero flux boundary condition for $b > 2$ as shown in Liu and Shin (2012b),

$$\left[ \nabla_m f + \left( \frac{bm}{b - |m|^2} - 2\kappa m \right) f \right] \cdot m = 0, \quad m \in \partial B. \tag{2.6}$$

For theoretical results concerning the existence of solutions of the coupled system we refer to Liu and Shin (2012a), Masmoudi (2008), Zhang and Zhang (2006); see also the works Chupin (2009a,b, 2010) and the earlier works on this problem: Jourdain and Lelièvre (2003), Jourdain et al. (2004). For rigorous analysis of long-time asymptotics of the FENE model, see Jourdain et al. (2006); and see Arnold et al. (2001) for entropy methods to study rate of convergence to equilibrium for Fokker–Planck type equations.

For some special configuration solutions with small flow rates, the use of moment closure approximations has been investigated by several authors; see, e.g., Du et al. (2005), Hyon et al. (2008), Herrchen and Öttinger (1997), Samaey et al. (2011) and Wang et al. (2008). Most numerical methods developed for the Fokker–Planck equation have been based on the form of (2.3a); see, for example, Ammar et al. (2006, 2007), Fan (1985) and Warner (1972).

Some elaborate numerical algorithms based on spectral methods were recently developed for the Fokker–Planck equation of FENE model in Chauvière and Lozinski (2003, 2004a,b). A spectral Galerkin approximation was further introduced in Knezevic and Süli (2009) based on a weighted weak formulation for $f(b - |m|^2)^{-\frac{b}{2}}$. An improved weighted formulation was proposed in Shen and Yu (2012) in terms of $f(b - |m|^2)^{-\frac{s}{2}}$ for $1 < s \leq b$, leading to a different spectral Galerkin algorithm. We note that this weighted formulation was also used for specific values of $s$ in Chauvière and Lozinski (2003, 2004a,b) and was analyzed in Section 3.2 of Knezevic and Süli (2009). The methods in Knezevic and Süli (2009) and Shen and Yu (2012) have provable stability results for certain weighted integrable initial data. However, positivity of the numerical solution is not guaranteed.

Finally we comment on the concept of ‘entropy’ explored in numerical approximations.
There is a vast literature on entropic schemes for related equations including hyperbolic conservation laws and kinetic equations such as Fokker–Planck type equations. For the former, entropy dissipation at discrete level is often enforced through numerical viscosity so that physical relevant shocks, particularly exact stationary shocks, can be captured; see, e.g., Bouchut and Morales de Luna (2009), Botchorishvili et al. (2003), Perthame (1990). For the later, information carried by the pdf becomes less and less as time evolves; the probability density is expected to converge to the equilibrium solution in a closed system regardless of how initial data are distributed. The entropy dissipation in time is the underlying mechanism for this phenomenon. To ensure the entropy property at discrete levels, one often uses the logarithmic Landau form

\[ \partial_t f = \frac{1}{2} \nabla \cdot \left( f \nabla \log \frac{f}{M} \right). \]

For a nonlinear Fokker–Planck equation, it was shown in Buet et al. (2001) that the scheme based on some entropic averages makes the Landau form equivalent to the underlying equation at the discrete level. Another class of finite difference schemes for Fokker–Planck equations is due to Chang and Cooper; see Chang and Cooper (1970), Larsen et al. (1985). This method is based upon the requirement that the discrete Fokker–Planck operator possesses a quasi-equilibrium solution which agrees at the mesh points with a quasi-equilibrium solution of the analytic operator. For a linear Fokker–Planck equation the Chang–Cooper scheme is shown in Buet and Dellacherie (2010) to make the underlying equation equivalent to the nonlogarithmic Landau form

\[ \partial_t f = \frac{1}{2} \nabla \cdot \left( M \nabla \frac{f}{M} \right) \]

at the discrete level. In this paper we explore the nonlogarithmic Landau form subject to a nonsymmetric drift term (2.5). The novel features include (1) the equilibrium \( M \) has no positive lower bound but zero at the boundary, making the Landau formulation singular and numerical computations more difficult; (2) the force due to fluid effects is generally nonconservative, one has to consider a nonsymmetric perturbation upon the usual Landau formulation, which makes the study of long time convergence more interesting; and (3) the natural function space for \( f \) is \( ML^2(Mdm) \), which when \( M \) vanishes at the boundary is different from the usual weighted
space $L^2(Mdm)$ [Liu and Shin (2012b)], hence the corresponding Galerkin discretization is not standard.

2.1.2 Contents

This paper is organized as follows. In Section 2.2, we describe the formulation of our scheme for the one-dimensional case. Theoretical analysis for both semidiscrete and fully discrete schemes is provided. In Section 2.3, we generalize the schemes to two space dimensions. Implementation strategies and numerical results of both one and two dimensions are presented in Section 2.4. Finally, in Section 2.5, concluding remarks are given.

2.2 One-dimensional Fokker–Planck Equation

We begin by looking at the Fokker–Planck problem over the interval $B = (-\sqrt{b}, \sqrt{b})$ in one-dimensional space. In such a case $\kappa = 0$ because of the constraint $Tr(\kappa) = 0$; then the problem can be described as

$$\partial_t f = \frac{1}{2} \partial_m^2 f + \frac{1}{2} \partial_m \left( \frac{bm}{b-m^2} f \right), \quad m \in B, t > 0, \quad (2.7a)$$

$$f(m, 0) = f_0(m), \quad m \in B, \quad (2.7b)$$

$$\left( \partial_m f + \frac{bm}{b-m^2} f \right) \bigg|_{m=\pm\sqrt{b}} = 0, \quad t > 0. \quad (2.7c)$$

The associated equilibrium solution reduces to

$$M(m) = (b - m^2)^{\frac{1}{2}}, \quad m \in B,$$

and (2.7a) becomes

$$\partial_t f = \frac{1}{2} \partial_m \left( M \partial_m g \right), \quad \text{where} \quad g = \frac{f}{M}. \quad (2.8)$$

2.2.1 Semidiscrete scheme

Given a positive integer $N$, we partition the domain $(-\sqrt{b}, \sqrt{b})$ by defining the uniform mesh size $h = \frac{2\sqrt{b}}{N}$ and the cell center at

$$m_j = -\sqrt{b} + (j - \frac{1}{2})h, \quad 1 \leq j \leq N.$$
Notice that at two end points $M(m_{\frac{1}{2}}) = M(m_{N+\frac{1}{2}}) = 0$ and $M(m_{j+\frac{1}{2}}) > 0$ for $1 \leq j \leq N - 1$.

On each computational cell $I_j = [m_{j-\frac{1}{2}}, m_{j+\frac{1}{2}}]$, we define the cell average of $f$ as

$$\bar{f}_j(t) = \frac{1}{h} \int_{I_j} f(m, t) \, dm.$$  

Integration of (2.8) on $I_j$ yields

$$\frac{d}{dt} \bar{f}_j = \frac{1}{2h} \int_{I_j} \partial_m (M\partial_m g) \, dm = \frac{1}{2h} M\partial_m g \bigg|_{m_{j-\frac{1}{2}}}^{m_{j+\frac{1}{2}}}.$$  

Based on this formulation we derive a finite volume scheme to compute $\{f_j\}$ which approximates $\{\bar{f}_j\}$ by taking the numerical flux

$$J_{j+\frac{1}{2}} = \frac{M\partial_m g}{h} = M_{j+\frac{1}{2}} \frac{g_{j+1} - g_j}{h} \quad \text{for} \quad j = 1, \ldots, N - 1$$  

with $M_{j+\frac{1}{2}} := M(m_{j+\frac{1}{2}})$, $g_j(t) = \frac{f_j(t)}{M_j}$, where $M_j = M(m_j)$. We also set

$$J_{\frac{1}{2}} = J_{\frac{N}{2}} = 0$$  

(2.10)

to incorporate the zero flux at the boundary.

Then we obtain a semidiscrete scheme

$$\frac{d}{dt} f_1 = \frac{1}{2h} J_{\frac{1}{2}},$$

$$\frac{d}{dt} f_j = \frac{1}{2h} (J_{j+\frac{1}{2}} - J_{j-\frac{1}{2}}), \quad 2 \leq j \leq N - 1,$$

$$\frac{d}{dt} f_N = -\frac{1}{2h} J_{N-\frac{1}{2}},$$

subject to the initial data

$$f_j(0) = \frac{1}{h} \int_{I_j} f_0(m) \, dm, \quad j = 1, \ldots, N.$$  

**Theorem 2.2.1.** The semidiscrete scheme (2.11) satisfies the following properties:

1. **Conservation of mass:** $\sum_{j=1}^{N} f_j(t) h = \sum_{j=1}^{N} f_j(0) h = \int_B f_0(m) \, dm \quad \forall t > 0$.

2. **Positivity preserving:** for any $t > 0$, $f_j(t) \geq 0$ if $f_j(0) \geq 0$.

3. The relative entropy $E(t) = \sum_{j=1}^{N} \frac{f_j^2}{M_j} h$ is nonincreasing in time with

$$\frac{d}{dt} E(t) = -\frac{1}{h} \sum_{j=1}^{N-1} (g_{j+1} - g_j)^2 M_{j+\frac{1}{2}} \leq 0.$$
Proof. (1) Summing all equations in (2.11), we have
\[
\frac{d}{dt} \sum_{j=1}^{N} f_j(t) = \sum_{j=1}^{N} \frac{d}{dt} f_j(t) = 0.
\]
So
\[
\sum_{j=1}^{N} f_j(t)h = \sum_{j=1}^{N} f_j(0)h = \int_B f_0(m) \, dm.
\]

(2) Since \( M_j \) is independent of \( t \), we have \( \frac{d}{dt} f_j = M_j \frac{d}{dt} g_j \). The scheme (2.11) can be rewritten as
\[
\frac{d}{dt} g_1 = \frac{1}{2h^2 M_1} M_1^2 (g_2 - g_1),
\]
\[
\frac{d}{dt} g_j = \frac{1}{2h^2 M_j} [M_{j+\frac{1}{2}} (g_{j+1} - g_j) - M_{j-\frac{1}{2}} (g_j - g_{j-1})], \quad 2 \leq j \leq N - 1,
\]
\[
\frac{d}{dt} g_N = -\frac{1}{2h^2 M_N} M_{N-\frac{1}{2}} (g_N - g_{N-1}).
\]

From (1), we see that
\[
\sum_{j=1}^{N} M_j g_j(t) = \sum_{j=1}^{N} f_j(0) \quad \forall t > 0.
\]

Then all the trajectories of (2.12) remain on this hyperplane. We define a closed set on this hyperplane by
\[
\Sigma = \left\{ \vec{g} : g_j \geq 0, j = 1, \ldots, N, \text{ and } \sum_{j=1}^{N} M_j g_j = \sum_{j=1}^{N} f_j(0) \right\}.
\]

Let \( \vec{F}(\vec{g}) \) be the vector field defined by the right-hand side of (2.12); then
\[
\frac{d}{dt} \vec{g} = \frac{1}{2} \vec{F}(\vec{g}).
\]

It suffices to show that \( \Sigma \) is an invariant region of this system. This is indeed the case if the vector field \( \vec{F}(\vec{g}) \) points strictly into \( \Sigma \) on the boundary \( \partial \Sigma \); i.e., for any outward normal vector \( \vec{n} \) on any part of \( \partial \Sigma \),
\[
\vec{F}(\vec{g}) \cdot \vec{n} < 0.
\]

From (2.12), it follows that
\[
\vec{F}(\vec{g}) \cdot \vec{n} = \sum_{j=1}^{N-1} \frac{n_j}{h^2 M_j} M_{j+\frac{1}{2}} (g_{j+1} - g_j) - \sum_{j=2}^{N} \frac{n_j}{h^2 M_j} M_{j-\frac{1}{2}} (g_j - g_{j-1}),
\]
\[
= -\frac{1}{h^2} \sum_{j=1}^{N-1} \left( \frac{n_{j+1}}{M_{j+1}} - \frac{n_j}{M_j} \right) M_{j+\frac{1}{2}} (g_{j+1} - g_j).
\]
For each $\vec{g} \in \partial \Sigma$, we define the set of indices $S$ such that

$$S = \{1 \leq j \leq N : g_j = 0\},$$

which implies that $S \neq \emptyset$ for any $\vec{g}_b \in \partial \Sigma$. Then the outward normal vectors $\vec{n}$ at $\vec{g}_b$ are of the form

$$\vec{n} = (n_1, \ldots, n_N)^T \quad \text{with} \quad n_j = \begin{cases} -\alpha_j, & \text{if } j \in S, \\ M_j, & \text{if } j \notin S. \end{cases}$$

Furthermore, there exists a positive real number $\gamma$ such that $\vec{g}_b - \gamma \vec{n}$ is in the interior of $\Sigma$, which implies that $\alpha_j > 0, \quad j \in S,$

and

$$\sum_{j=1}^{N} M_j n_j = 0, \quad \text{i.e.,} \quad \sum_{j \in S} M_j \alpha_j = \sum_{j \notin S} M_j^2.$$

Now we look back at (2.14). Note that if $j, j+1 \in S$, then $g_j = g_{j+1} = 0$; if $j, j+1 \notin S$, then $\frac{n_{j+1}}{M_{j+1}} - \frac{n_j}{M_j} = 1 - 1 = 0$. Therefore the nonzero terms in (2.14) are only those with $j \in S, j+1 \notin S$ or $j \notin S, j+1 \in S$. Hence

$$\begin{align*}
\vec{F}(\vec{g}) \cdot \vec{n} &= -\frac{1}{h^2} \left( \sum_{j \in S} + \sum_{j \notin S} \frac{n_{j+1}}{M_{j+1}} - \frac{n_j}{M_j} \right) M_{j+\frac{1}{2}}(g_{j+1} - g_j) \\
&= -\frac{1}{h^2} \sum_{j \in S} \left( 1 + \frac{\alpha_j}{M_j} \right) M_{j+\frac{1}{2}} g_{j+1} - \frac{1}{h^2} \sum_{j \notin S} \left( 1 + \frac{\alpha_{j+1}}{M_{j+1}} \right) M_{j+\frac{1}{2}} g_j < 0.
\end{align*}$$

This leads to the conclusion that $g_j(t) \geq 0$ as long as $g_j(0) \in \Sigma$.

(3) We now show that the relative entropy $E(t)$ is nonincreasing. In fact,

$$\begin{align*}
\frac{d}{dt} \sum_{j=1}^{N} \frac{f_j^2}{M_j} h &= 2 \sum_{j=1}^{N} \frac{f_j}{M_j} \frac{d}{dt} f_j h = \sum_{j=1}^{N} g_j (J_{j+\frac{1}{2}} - J_{j-\frac{1}{2}}) \\
&= -\sum_{j=1}^{N-1} (g_{j+1} - g_j) J_{j+\frac{1}{2}} \\
&= -\frac{1}{h} \sum_{j=1}^{N-1} (g_{j+1} - g_j)^2 M_{j+\frac{1}{2}} \leq 0.
\end{align*}$$
Theorem 2.2.2. The physical entropy $E_p(t) = \sum_{j=1}^{N} f_j \log \left( \frac{f_j}{M_j} \right) h$ is nonincreasing in time. Moreover,
\[ \frac{d}{dt} E_p(t) = -\frac{1}{2h} \sum_{j=1}^{N-1} M_{j+\frac{1}{2}} \log \left( \frac{g_{j+1}}{g_j} \right) (g_{j+1} - g_j) \leq 0. \] (2.15)

Proof. A direct calculation using (2.11) and summation by parts gives
\[
\frac{d}{dt} \sum_{j=1}^{N} f_j \log \left( \frac{f_j}{M_j} \right) h = \sum_{j=1}^{N} \frac{d}{dt} f_j \left[ \log \left( \frac{f_j}{M_j} \right) + 1 \right] h
\]
\[= \frac{1}{2} \sum_{j=1}^{N} (J_{j+\frac{1}{2}} - J_{j-\frac{1}{2}}) (\log g_j + 1) = -\frac{1}{2} \sum_{j=1}^{N-1} J_{j+\frac{1}{2}} (\log g_{j+1} - \log g_j)
\]
\[= -\frac{1}{2h} \sum_{j=1}^{N-1} M_{j+\frac{1}{2}} (g_{j+1} - g_j) \log \left( \frac{g_{j+1}}{g_j} \right) \leq 0,
\]
where both (2.9) and $(X - Y)(\log X - \log Y) \geq 0$ have been used. \qed

We may also examine the large time behavior of $\bar{g}(t)$. Both positivity $\bar{g}(t) > 0$ and the constraint
\[
\sum_{j=1}^{N} g_j(t) M_j h = \int_B f_0(m) \, dm
\]
together ensure that $\bar{g}(t)$ will remain bounded for all time. Since (2.12) is an autonomous system, what happens as $t \to \infty$ is simple to describe. We summarize this result in the following theorem.

Theorem 2.2.3. Consider the semidiscrete scheme (2.11) subject to the initial data $f_j(0) > 0$ with $\sum_{j=1}^{N} f_j(0) h = \int_B f_0(m) \, dm$; then
\[
[f_1, f_2, \cdots, f_N]^T \to C[M_1, \cdots, M_N], \quad t \to \infty,
\]
where
\[C = \frac{\int_B f_0(m) \, dm}{\sum_{j=1}^{N} M_j h}.\] (2.16)

Proof. Define a functional $V(g)$ by
\[V(g) = \sum_{j=1}^{N} (g_j - C)^2 M_j
\]
and $\bar{g} = C(1, \cdots, 1)^T$. We see that $\bar{g} \in \Sigma$ as defined in (2.13), satisfying $F(\bar{g}) = 0$. A direct verification shows that $V(g) = E - C \int_B f_0(m) \, dm$, implying $\frac{d}{dt} V = \frac{d}{dt} E$. Hence $V$ satisfies the following:
\[
\begin{align*}
\bullet \quad V(g) > 0 & \text{ for any } g \neq \bar{g} \text{ (positive definite)}, \\
\bullet \quad \frac{d}{dt} V \leq 0 & \text{ for all } g \in \Sigma \text{ (negative semidefinite)}, \\
\bullet \quad \text{the set } \{ \frac{d}{dt} V = 0 \} \cap \Sigma & \text{ does not contain any trajectories of the ODE system (2.11) besides the trajectory } g(t) = \bar{g} \forall t > 0.
\end{align*}
\]

With these properties we can apply the Krasovskii–LaSalle principle to conclude that \( \lim_{t \to \infty} g(t) = \bar{g} \), which leads to the conclusion. It is left to verify the stated properties of \( V \). First two properties of \( V \) are easy to verify. We only verify the third property of \( V \): from (3) of Theorem 2.2.1 it follows that if \( \frac{d}{dt} E = 0 \), then 
\[
\sum_{j=1}^{N-1} (g_{j+1} - g_j)^2 M_{j+\frac{1}{2}} = 0,
\]
which ensures that \( g_j = \text{constant} \), while within \( \Sigma \), \( g = \bar{g} \) must hold. The proof is thus complete.

2.2.2 Fully discrete scheme

Let the time step be denoted by \( k \), and the mesh ratio \( \lambda = \frac{k}{2\pi} \). We apply the backward Euler method to the semidiscrete scheme (2.11) to get
\[
\begin{align*}
\frac{f^n_{j+1}}{f^n_{j+1}} &= f^n_{j} + \lambda M_{j+\frac{1}{2}} (g^n_{j+1} - g^n_{j}) , \\
\frac{f^n_{j}}{f^n_{j}} &= f^n_{j} + \lambda \left[ M_{j+\frac{1}{2}} (g^n_{j+1} - g^n_{j}) - M_{j-\frac{1}{2}} (g^n_{j} - g^n_{j-1}) \right], 2 \leq j \leq N - 1, (2.17) \\
\frac{f^n_{N}}{f^n_{N}} &= f^n_{N} - \lambda M_{N-\frac{1}{2}} (g^n_{N} - g^n_{N-1}).
\end{align*}
\]
Given \( \{f^n_j\} \), \( \{f^n_{j+1}\} \) can be obtained from \( f^n_{j+1} = Mg^n_j \) where \( \{g^n_{j+1}\} \) solves the following linear system:
\[
\begin{align*}
(M_1 + \lambda M_{\frac{3}{2}})g^n_{1} - \lambda M_{\frac{1}{2}} g^n_{2} &= f^n_1 , \\
-\lambda M_{\frac{1}{2}} g^n_{j-1} + [M_j + \lambda (M_{\frac{1}{2}} + M_{j-\frac{1}{2}})] g^n_j - \lambda M_{\frac{1}{2}} g^n_{j+1} &= f^n_j, 2 \leq j \leq N - 1, (2.18) \\
-\lambda M_{N-\frac{1}{2}} g^n_{N-1} + (M_N + \lambda M_{N-\frac{1}{2}}) g^n_{N} &= f^n_N.
\end{align*}
\]

**Theorem 2.2.4.** The fully discrete scheme (2.17) has a unique solution \( \{f^n_j\} \). Moreover, the solution satisfies the following properties:
(1) Conservation of mass: \[ \sum_{j=1}^{N} f_j^{n+1} h = \sum_{j=1}^{N} f_j^n h. \]

(2) Positivity. If \( f_j^n \geq 0 \), then \( f_j^{n+1} \geq 0 \).

(3) The relative entropy
\[ E^n = \sum_{j=1}^{N} \frac{(f_j^n)^2}{M_j} h \]

is nonincreasing. More precisely,
\[ E^{n+1} = E^n - \frac{k}{h} \sum_{j=1}^{N-1} (g_{j+1}^{n+1} - g_j^{n+1})^2 M_{j+\frac{1}{2}} - \sum_{j=1}^{N} \frac{(f_j^{n+1} - f_j^n)^2}{M_j} h. \] (2.19)

(4) \( f_j^n \) converges as \( n \to \infty \) with
\[ f_j^n \to CM_j, \]
where \( C \) is defined in (2.16).

Proof. First of all, we show the existence of a solution to (2.17). (2.18) is a linear system of
\[ A \bar{g}^{n+1} = \bar{f}^n, \]
where
\[ \bar{g}^{n+1} = (g_1^{n+1}, \ldots, g_N^{n+1})^T, \quad \bar{f}^n = (f_1^n, \ldots, f_N^n)^T. \]

From the fact that \( A \) is a strictly diagonally dominant matrix, it follows that there is a unique solution \( \bar{g}^{n+1} = A^{-1} \bar{f}^n \) for any \( \bar{f}^n \).

(1) Summing up the \( N \) equations in (2.17) gives
\[ \sum_{j=1}^{N} f_j^{n+1} h = \sum_{j=1}^{N} f_j^n h. \]

(2) Since \( M_j > 0, 1 \leq j \leq N \), we only need to prove that \( g_i^{n+1} \geq 0 \forall j \). It suffices to show that \( \min_{1 \leq j \leq N} g_j^{n+1} = g_i^{n+1} \geq 0 \). We only show the case \( 2 \leq i \leq N - 1 \), as the cases \( i = 1 \) and \( i = N \) are similar and simpler,
\[ f_i^n = -\lambda M_{i-\frac{1}{2}} g_i^{n+1} + [M_i + \lambda( M_{i+\frac{1}{2}} + M_{i-\frac{1}{2}} )] g_i^{n+1} - \lambda M_{i+\frac{1}{2}} g_{i+1}^{n+1} \]
\[ \leq -\lambda M_{i-\frac{1}{2}} g_i^{n+1} + [M_i + \lambda( M_{i+\frac{1}{2}} + M_{i-\frac{1}{2}} )] g_i^{n+1} - \lambda M_{i+\frac{1}{2}} g_i^{n+1} \]
\[ = M_i g_i^{n+1}. \]
Hence, \( g_i^{n+1} \geq M_i^{-1} f_i^n \geq 0 \).

(3) As for the relative entropy, we calculate

\[
\sum_{j=1}^{N} \left[ \frac{(f_j^{n+1})^2}{M_j} - \frac{(f_j^n)^2}{M_j} \right] = \sum_{j=1}^{N} \frac{(2f_j^{n+1} + f_j^n - f_j^{n+1}) (f_j^{n+1} - f_j^n)}{M_j}
\]

\[
= \frac{k}{h} \sum_{j=1}^{N} g_j^{n+1} (J_{j+\frac{1}{2}}^{n+1} - J_j^{n+1}) - \sum_{j=1}^{N} \frac{(f_j^{n+1} - f_j^n)^2}{M_j}
\]

\[
= - \frac{k}{h} \sum_{j=1}^{N-1} (g_j^{n+1} - g_j^{n+1})^2 M_{j+\frac{1}{2}} - \sum_{j=1}^{N} \frac{(f_j^{n+1} - f_j^n)^2}{M_j} \leq 0.
\]

This yields (2.19), which implies that the relative entropy is nonincreasing.

(4) Since \( E^n \) is nonincreasing and bounded from below, we have

\[
\lim_{n \to \infty} E^n = \inf \{E^n \}.
\]

Observe from (2.19) that \( E^n - E^{n+1} \) is a sum of nonnegative and bounded terms. When passing limit \( n \to \infty \) we conclude that each term must have zero as its limit, that is,

\[
\lim_{n \to \infty} (f_j^{n+1} - f_j^n)^2 = 0, \quad \lim_{n \to \infty} (g_j^{n+1} - g_j^{n+1})^2 = 0.
\]

(2.20)

The first relation in (2.20) tells that \( \vec{g}^n \) is a Cauchy sequence, which when combined with the completeness of \( \Sigma \) (a closed and bounded set in \( \mathbb{R}^N \)) ensures that \( \lim \vec{g}^n \) exists. The second relation in (2.20) infers that the limit must be \( \vec{g} \). The proof is complete.

\[\Box\]

**Theorem 2.2.5.** The physical entropy

\[
E_p^n = \sum_{j=1}^{N} f_j^n \log \left( \frac{f_j^n}{M_j} \right) h
\]

is nonincreasing. Moreover,

\[
E_p^{n+1} - E_p^n = - \sum_{j=1}^{N-1} \frac{k M_{j+\frac{1}{2}}}{2h} (g_j^{n+1} - g_j^{n+1})(\log g_j^{n+1} - \log g_j^{n+1}) + \sum_{j=1}^{N} h f_j^n \log \left( \frac{f_j^{n+1}}{f_j^n} \right) \leq 0.
\]

**Proof.** For physical entropy \( E_p^n = \sum_{j=1}^{N} f_j^n \log \left( \frac{f_j^n}{M_j} \right) h = \sum_{j=1}^{N} g_j^n \log \left( g_j^n \right) M_j h \), we calculate

\[
E_p^{n+1} - E_p^n = \sum_{j=1}^{N} h f_j^{n+1} \log g_j^{n+1} - \sum_{j=1}^{N} h f_j^n \log g_j^{n+1} + \sum_{j=1}^{N} h f_j^n \log g_j^{n+1} - \sum_{j=1}^{N} h f_j^n \log g_j^n
\]

\[
= \sum_{j=1}^{N} h (f_j^{n+1} - f_j^n) \log g_j^{n+1} + \sum_{j=1}^{N} f_j^n \log \left( \frac{f_j^{n+1}}{f_j^n} \right) h.
\]
Using scheme (2.17) and \( \log x \leq x - 1 \) for \( x > 0 \), we estimate

\[
E^{n+1}_p - E^n_p \leq \sum_{j=1}^N \frac{k}{2} (J_{j+\frac{1}{2}}^{n+1} - J_{j-\frac{1}{2}}^{n+1}) \log g_j^{n+1} + \sum_{j=1}^N h f_j^n \left( \frac{f_j^{n+1}}{f_j^n} - 1 \right)
\]

\[
= - \sum_{j=1}^{N-1} \frac{1}{2} (g_{j+1}^{n+1} - g_j^{n+1})(\log g_{j+1}^{n+1} - \log g_j^{n+1}) + \sum_{j=1}^N (f_j^{n+1} - f_j^n) h \leq 0
\]

for the first summation is nonnegative due to monotonicity of \( \log x \), and mass conservation implies that the second summation is zero.

\[
\square
\]

### 2.3 Extension to the multidimensional FENE model

#### 2.3.1 Reformulation

Let the matrix \( \kappa \) be decomposed into a sum of the symmetric part and the asymmetric part, i.e.,

\[
\kappa = \kappa^s + \kappa^a.
\]

Define \( M(m) \) as

\[
M(m) = (b - |m|^2)^{\frac{1}{2}} e^{m^T \kappa^s m}
\]

and \( g(m, t) = \frac{f(m, t)}{M(m)} \); then the Fokker–Planck equation (2.3a) can be rewritten as

\[
\partial_t f = \frac{1}{2} \nabla_m \cdot (M \nabla_m g - 2\kappa^a m f).
\]

**Lemma 2.3.1.** Let \( f \) be a solution to (2.22). If \( \kappa \) is normal, then \( M(m) \) is the equilibrium solution to (2.22). Moreover, the relative entropy \( E(t) = \int_B Mg^2 dm \) satisfies

\[
\frac{d}{dt} E(t) + \int_B M|\nabla_m g|^2 dm = 0.
\]

**Proof.** Using the zero flux condition in the evolution of \( E \) we find that

\[
\frac{d}{dt} E = \int_B 2g \partial_t f dm = \int_B g \nabla_m \cdot [M \nabla_m g - 2\kappa^a m f] dm = - \int_B M|\nabla_m g|^2 dm + 2 \int_B \nabla_m g \cdot \kappa^a m f dm.
\]
Let $B_r$ be a ball with radius $r < \sqrt{b}$; then using integration by parts we obtain

$$2 \int_{B_r} \nabla_m g \cdot \kappa^a m f \, dm = \int_{B_r} \nabla_m g^2 \cdot \kappa^a m M \, dm$$

$$= \int_{\partial B_r} g^2 M \kappa^a m \cdot \frac{m}{|m|} \, dS - \int_{B_r} g^2 \nabla_m \cdot (\kappa^a m M) \, dm$$

$$= \int_{B_r} g^2 \kappa^a m \cdot \nabla_m M \, dm,$$

which, in virtue of $\nabla M = (2\kappa^s m - \frac{bm}{b|m|^2})M$, reduces to

$$\int_{B_r} M g^2 m^T \kappa^a \kappa^a m \, dm = \frac{1}{4} \int_{B_r} M g^2 m^T (\kappa^T \kappa - \kappa \kappa^T) m \, dm.$$

This vanishes if $\kappa$ is normal. Let $r \to \sqrt{b}$ we obtain

$$\int_{B_r} \nabla_m g \cdot \kappa^a m f \, dm = 0,$$

hence the desired estimate (2.23) follows. \hfill \Box

**Remark 2.3.1.** If $\kappa$ is not normal, the above estimate can still be obtained if we replace $M$ by the equilibrium solution. But in such a case, an explicit expression of the equilibrium solution is not available. With $M$ defined above, we will have

$$0 \neq 2 \int_{B} \nabla_m g \cdot \kappa^a m f \, dm \leq \frac{1}{2} \int_{B} M |\nabla_m g|^2 \, dm + 2a^2 b \int_{B} M g^2 \, dm.$$

Hence

$$\frac{d}{dt} E + \frac{1}{2} \int_{B} M |\nabla_m g|^2 \, dm \leq 2a^2 b E,$$

leading to

$$E(t) \leq e^{2a^2 b t} E(0) \quad \text{for } t > 0.$$

In such a case, $E$ is no longer decreasing, though still bounded in finite time.

In the discretization to follow, we shall focus only on the two-dimensional case, for which $\kappa$ has the following form,

$$\kappa = \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & -k_{11} \end{pmatrix}$$

with

$$\kappa^a = \begin{pmatrix} 0 & k_{12}-k_{21} \\ k_{12}-k_{21} & 0 \end{pmatrix} = a \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

with $a = \frac{k_{12} - k_{21}}{2}$. 

and

\[ \kappa^s = \begin{pmatrix} k_{11} & \frac{k_{12}+k_{21}}{2} \\ \frac{k_{12}+k_{21}}{2} & -k_{11} \end{pmatrix}. \]

### 2.3.2 Discretization in \( m \in B \)

The domain \( B \) can be represented by \([0, \sqrt{b}] \times [0, 2\pi]\) in the polar coordinate system. Partition \( B \) into uniform rectangles

\[ K_{ij} = \{(r, \theta); r_{i-\frac{1}{2}} \leq r \leq r_{i+\frac{1}{2}}, \theta_{j-\frac{1}{2}} \leq \theta \leq \theta_{j+\frac{1}{2}}\}, \quad 1 \leq i \leq P, 1 \leq j \leq Q, \]

where

\[ r_{i+\frac{1}{2}} = i \Delta r, \quad \theta_{j+\frac{1}{2}} = j \Delta \theta \]

with steps of radius and angle

\[ \Delta r = \frac{\sqrt{b}}{P}, \quad \Delta \theta = \frac{2\pi}{Q}. \]

Let the cell average of \( f \) on \( K_{ij} \) be defined by

\[ \bar{f}_{i,j} = \frac{1}{|K_{ij}|} \int_{K_{ij}} f(m,t) \, dm, \]

![Figure 2.1 Diagram of the two-dimensional partition of \( B \).](image.png)
where $|K_{ij}| = \Delta \theta \Delta r_i r_i$ is the area of cell $K_{ij}$. Integrate (2.22) over $K_{ij}$ on both sides,

$$
\frac{d}{dt} \bar{r}_{i,j} = \frac{1}{2|K_{ij}|} \int_{K_{ij}} \nabla_m \cdot (M \nabla_m g - 2\kappa^a m_f) \, dm
$$

$$
= \frac{1}{2|K_{ij}|} \int_{\partial K_{ij}} (M \nabla_m g - 2\kappa^a m_f) \cdot \vec{\nu} \, ds,
$$

(2.24)

by the divergence theorem. Here $\vec{\nu}$ is the outward normal of the cell boundary $\partial K_{ij}$.

In order to derive a finite volume scheme, we use $f_{i,j} = g_{i,j}M_{i,j}$ as the numerical solution in $K_{ij}$ to approximate $\bar{f}_{i,j}$ and represent (2.24) in terms of $\{f_{i,j}\}$.

Because numerical representatives $f$ and $g$ are not defined on $\partial K_{ij}$, we need to define a numerical flux to represent $(M \nabla_m g - 2\kappa^a m_f) \cdot \vec{\nu}$ on $\partial K_{ij}$. To simplify the presentation, we introduce two difference operators,

$$
D_r g_{i,j} = \frac{g_{i+1,j} - g_{i,j}}{\Delta r}, \quad D_\theta g_{i,j} = \frac{g_{i,j+1} - g_{i,j}}{\Delta \theta}.
$$

There are four pieces within $\partial K_{ij}$, denoted by $\gamma_1, \gamma_2, \gamma_3$ and $\gamma_4$. On $\gamma_1 = \{(r, \theta) ; r = r_{i+\frac{1}{2}}, \theta_{j-\frac{1}{2}} \leq \theta \leq \theta_{j+\frac{1}{2}} \}$, we have

$$
\int_{\gamma_1} M \bar{\nabla}_m g \cdot \vec{\nu} \, ds = \int_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}} M(r_{i+\frac{1}{2}}, \theta) \partial_r g(r_{i+\frac{1}{2}}, \theta) r_{i+\frac{1}{2}} d\theta
$$

$$
= \int_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}} M(r_{i+\frac{1}{2}}, \theta) D_r g_{i,j} r_{i+\frac{1}{2}} d\theta
$$

$$
= \Delta \theta r_{i+\frac{1}{2}} M_{i+\frac{1}{2},j} D_r g_{i,j},
$$

where we use the midpoint rule for the integration in $\theta$ and $\partial_r g(r_{i+\frac{1}{2}}, \theta) = D_r g_{i,j}$. Similarly, on $\gamma_3 = \{(r, \theta) ; r = r_{i-\frac{1}{2}}, \theta_{j-\frac{1}{2}} \leq \theta \leq \theta_{j+\frac{1}{2}} \}$,

$$
\int_{\gamma_3} M \bar{\nabla}_m g \cdot \vec{\nu} \, ds = -\Delta \theta r_{i-\frac{1}{2}} M_{i-\frac{1}{2},j} D_r g_{i-1,j}.
$$

On $\gamma_2 = \{(r, \theta) ; r_{i-\frac{1}{2}} \leq r \leq r_{i+\frac{1}{2}}, \theta = \theta_{j+\frac{1}{2}} \}$, we have $\vec{\nu} = (-\sin \theta, \cos \theta)^T$ and $\nabla_m \cdot \vec{\nu} = \frac{1}{r} \frac{\partial}{\partial \theta}$, hence

$$
\int_{\gamma_2} M \bar{\nabla}_m g \cdot \vec{\nu} \, ds = \int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} M(r, \theta_{j+\frac{1}{2}}) \frac{\partial g(r, \theta_{j+\frac{1}{2}})}{r} dr
$$

$$
= \int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} M(r, \theta_{j+\frac{1}{2}}) \frac{\partial g_{i,j}}{r} dr
$$

$$
= \Delta r M_{i,j+\frac{1}{2}} D_\theta g_{i,j}.
$$
where we have taken \( \partial_{\theta g}(r, \theta_{j+\frac{1}{2}}) = D_{\theta g_{i,j}} \). Similarly, on \( \gamma_4 = \{(r, \theta); r_{i-\frac{1}{2}} \leq r \leq r_{i+\frac{1}{2}}, \theta = \theta_{j-\frac{1}{2}}\} \),
\[
\int_{\gamma_4} M \nabla g \cdot \bar{v} \, ds = -\frac{\Delta r M_{i,j-\frac{1}{2}}}{r_i} D_{\theta g_{i,j-1}}.
\]
For the antisymmetric part,
\[
2\kappa_a \cdot \bar{v} = 2a \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \cdot \bar{v} = \begin{cases} 0, & \text{on } \gamma_1, \\ -2ar, & \text{on } \gamma_2, \\ 0, & \text{on } \gamma_3, \\ 2ar, & \text{on } \gamma_4. \end{cases}
\]
It follows that
\[
\int_{\partial K_{ij}} (2\kappa_a m f) \cdot \bar{v} \, ds = -2a \int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} rf(r, \theta_{j+\frac{1}{2}}) \, dr + 2a \int_{r\frac{1}{2}}^{r_{i+\frac{1}{2}}} rf(r, \theta_{j-\frac{1}{2}}) \, dr.
\]
The numerical flux is chosen to be upwind,
\[
f(r, \theta_{j+\frac{1}{2}}) = \frac{1}{2} (f_{i,j+1} + f_{i,j}) + \frac{\text{sign}(a)}{2} (f_{i,j+1} - f_{i,j}).
\]
Hence
\[
\int_{\partial K_{ij}} (2\kappa_a m f) \cdot \bar{v} \, ds = -\Delta r_i [(a + |a|)(f_{i,j+1} - f_{i,j}) + (a - |a|)(f_{i,j} - f_{i,j-1})].
\]
Therefore we obtain the semidiscrete scheme
\[
\frac{d}{dt} f_{i,j} = \frac{r_{i+\frac{1}{2}} M_{i+\frac{1}{2},j}}{2\Delta r_i} D_{r g_{i,j}} - \frac{r_{i-\frac{1}{2}} M_{i-\frac{1}{2},j}}{2\Delta r_i} D_{r g_{i-1,j}} + \frac{M_{i,j+\frac{1}{2}}}{2\Delta \theta r_i^2} D_{\theta g_{i,j}} - \frac{M_{i,j-\frac{1}{2}}}{2\Delta \theta r_i^2} D_{\theta g_{i,j-1}}
\]
\[
+ \frac{1}{2} [(a + |a|) D_{\theta f_{i,j}} + (a - |a|) D_{\theta f_{i,j-1}}].
\]
(2.25)

In regards to (2.25), when \( i = 1, \gamma_3 \) is reduced to a point, so \( \frac{r_1 M_{i+\frac{1}{2},j}}{2\Delta r_i} D_{r g_{i,j}} \) is understood as 0; when \( i = P \), the zero flux gives that \( \frac{r^\frac{P+\frac{1}{2}}{M^\frac{P+\frac{1}{2}}{2}}}{D_{r g_{i,j}}} = 0 \). Due to the periodicity of \( f \) and \( M \) with respect to \( \theta \), we take
\[
f_{i,j} = f_{i,j+Q}, \quad M_{i,j} = M_{i,j+Q} \quad \text{for } 1 \leq j \leq Q.
\]
Thus (2.25) is well defined for \( 1 \leq i \leq P, 1 \leq j \leq Q \), which can be solved subject to the initial data
\[
f_{i,j}(0) = \frac{1}{|K_{ij}|} \int_{K_{ij}} f_0(m) \, dm.
\]
Theorem 2.3.2. Let \( \{ij\} = \{1 \leq i \leq P, 1 \leq j \leq Q\} \). The semidiscrete scheme (2.25) has the following properties:

1. \( \sum_{ij} f_{i,j}(t)|K_{ij}| = \sum_{ij} f_{i,j}(0)|K_{ij}| = \int_B f_0(m) \, dm. \)

2. Positivity. If \( f_{i,j}(0) \geq 0 \), then \( f_{i,j}(t) \geq 0 \forall t > 0 \).

3. The semidiscrete relative entropy, defined by
   \[ E(t) = \sum_{ij} \frac{f_{i,j}^2(t)}{M_{i,j}} |K_{ij}|, \]
   satisfies
   \[ E(t) \leq E(0), \quad t > 0 \]
   for \( \kappa \) normal and \( E(t) \leq e^{ct} E(0) \), for general \( \kappa \) with \( c > 0 \) dependent on \( b \).

Proof. (1) Summation of \( \frac{d}{dt} f_{i,j}(t)|K_{ij}| \) over \( \{ij\} \) in (2.25) gives
   \[ \frac{d}{dt} \sum_{ij} f_{i,j}(t)|K_{ij}| = \sum_{ij} \frac{d}{dt} f_{i,j}(t) \Delta \theta \Delta r_r i = 0. \]
   So
   \[ \sum_{ij} f_{i,j}(t)|K_{ij}| = \sum_{ij} f_{i,j}(0)|K_{ij}| = \int_B f_0(m) \, dm \quad \text{for all } t > 0. \]

(2) We arrange the solution \( \{g_{i,j}\} \) of (2.25) to be a vector \( \vec{g} = (g_{1,1}, g_{2,1}, \ldots, g_{P,1}, g_{1,2}, \ldots, g_{P,Q})^T \).

And rewrite (2.25) into the vector form
   \[ \frac{d}{dt} \vec{g} = \vec{F}(\vec{g}). \] (2.26)

Due to the mass conservation, we have
   \[ \sum_{ij} M_{ij} g_{i,j}|K_{ij}| = \sum_{ij} f_{i,j}(0)|K_{ij}| = \int_B f_0(m) \, dm \quad \forall t > 0. \]

Then all the trajectories of (2.26) remain on this hyperplane. We define a closed set \( \Sigma \) on this hyperplane by
   \[ \Sigma = \left\{ \vec{g} : g_{i,j} \geq 0, (i, j) \in \{ij\}, \text{ and } \sum_{ij} M_{ij} g_{i,j}|K_{ij}| = \int_B f_0(m) \, dm \right\}. \]
It suffices to show that \( \Sigma \) is an invariant region of the ODE system (2.26). Similar to the argument in one-dimensional case explored previously, we only need to prove that for any outward normal vector \( \tilde{n} \) on any part of the boundary of \( \Sigma \),

\[
\vec{F}(\vec{g}) \cdot \tilde{n} < 0, \quad \vec{g} \in \partial \Sigma.
\]

For each \( \vec{g} \in \partial \Sigma \), we define the set of indices \( S \) such that

\[
S = \{(i, j) \in \{ij\} : g_{i,j} = 0\},
\]

which implies that \( S \neq \emptyset \) for any \( \vec{g} \in \partial \Sigma \). Then the outward normal vectors \( \tilde{n} \) at \( \vec{g} \in \partial \Sigma \) are of the form

\[
\tilde{n} = (n_{ij}) \quad \text{with} \quad n_{ij} = \begin{cases} 
-|K_{ij}|\alpha_{ij}, & \text{if } (i, j) \in S, \\
|K_{ij}|M_{ij}, & \text{if } (i, j) \notin S.
\end{cases}
\]

Furthermore, \( \alpha_{ij} > 0 \) for \( (i, j) \in S \).

Here we assume \( a \geq 0 \). The proof is similar for the case of \( a < 0 \). Shifting the indices, we have

\[
\vec{F}(\vec{g}) \cdot \tilde{n} = -\sum_{ij} \frac{\Delta \theta r_{i+\frac{1}{2}} M_{i,\frac{1}{2},j}}{2} D_r g_{i,j} \left( \frac{n_{i+1,j}}{|K_{i+1,j}|M_{i+1,j}} - \frac{n_{ij}}{|K_{ij}|M_{ij}} \right)
\]

\[
- \sum_{ij} \frac{\Delta r M_{i,j+\frac{1}{2}}}{2r_i} D_{\theta} g_{i,j} \left( \frac{n_{i,j+1}}{|K_{i,j+1}|M_{i,j+1}} - \frac{n_{ij}}{|K_{ij}|M_{ij}} \right)
\]

\[
- \sum_{ij} a \Delta r r_i f_{i,j} \left( \frac{n_{ij}}{|K_{ij}|M_{ij}} - \frac{n_{i,j-1}}{|K_{i,j-1}|M_{i,j-1}} \right)
\]

\[
= -I - II - III.
\]

We only analyze I, since the discussion about II and III is analogous. If \( (i, j), (i + 1, j) \in S \), then \( g_{i,j} = g_{i+1,j} = 0 \), implying \( D_r g_{i,j} = 0 \); if \( (i, j), (i + 1, j) \notin S \), then \( \frac{n_{i+1,j}}{|K_{i+1,j}|M_{i+1,j}} - \frac{n_{ij}}{|K_{ij}|M_{ij}} = 1 - 1 = 0 \). Hence the nonzero terms in I are only those with \( (i, j) \in S, (i + 1, j) \notin S \) or \( (i, j) \notin S, (i + 1, j) \in S \). Therefore

\[
I = \sum_{(i,j)\in S} \sum_{(i+1,j)\notin S} \frac{\Delta \theta r_{i+1,\frac{1}{2}} M_{i+\frac{1}{2},j}}{2} D_r g_{i,j} \left( 1 + \frac{\alpha_{ij}}{M_{ij}} \right) + \sum_{(i,j)\notin S} \sum_{(i+1,j)\in S} \frac{\Delta \theta r_{i+\frac{1}{2}} M_{i+1,\frac{1}{2},j}}{2} D_r g_{i,j} \left( -\frac{\alpha_{i+1,j}}{M_{i+1,j}} - 1 \right).
\]
By the definition of \( S \), \( g_{i,j} = 0 \) in the first summation and \( g_{i+1,j} = 0 \) in the second summation. So

\[
I = \sum_{(i,j) \in S} \frac{\Delta \theta r_i + \frac{1}{2} M_i + \frac{i}{2} j}{2 \Delta r} g_{i+1,j} \left( 1 + \frac{\alpha_{ij}}{M_{ij}} \right) + \sum_{(i,j) \notin S, (i+1,j) \in S} \frac{\Delta \theta r_i + \frac{1}{2} M_i + \frac{i}{2} j}{2 \Delta r} \left( -g_{i,j} \right) \left( -\frac{\alpha_{i+1,j}}{M_{i+1,j}} - 1 \right)
\]

\[
> 0.
\]

With this, \( II > 0 \) and \( III > 0 \), we arrive at the conclusion that \( \tilde{F}(\tilde{\eta}) \cdot \tilde{n} < 0 \).

(3) Next, we show that \( E(t) \) remains bounded for any \( t > 0 \). For definiteness, we assume \( a > 0 \).

\[
\frac{d}{dt} E(t) = \sum_{ij} \frac{2 f_{i,j}}{M_{i,j}} \frac{d f_{i,j}}{dt} |K_{i,j}| = \sum_{ij} 2g_{i,j} \frac{d f_{i,j}}{dt} \Delta \theta \Delta r_i \\
= \Delta \theta \sum_{ij} g_{i,j} (r_i + \frac{1}{2} M_i + \frac{i}{2} j D_r g_{i,j} - r_i - \frac{1}{2} M_i - \frac{i}{2} j D_r g_{i-1,j}) \\
+ \Delta r \sum_{ij} \frac{g_{i,j}}{r_i} (M_i + \frac{1}{2} D_\theta g_{i,j} - M_i - \frac{1}{2} D_\theta g_{i,j-1}) + 2a \Delta r \sum_{ij} g_{i,j} r_i (f_{i,j+1} - f_{i,j}) \\
= \Delta \theta I + \Delta r II + 2a \Delta r III.
\]

By shifting the indices in \( i \) and using \( r_i = 0, M_{i+\frac{1}{2},j} = 0 \), we have

\[
I = - \sum_{1 \leq i \leq P-1, 1 \leq j \leq Q} \Delta r_i + \frac{1}{2} M_i + \frac{i}{2} j (D_r g_{i,j})^2 = -\Delta r \sum_{ij} r_i + \frac{1}{2} M_i + \frac{i}{2} j (D_r g_{i,j})^2.
\]

Similarly, shifting the indices in \( j \) gives

\[
II = - \sum_{1 \leq i \leq P, 1 \leq j \leq Q-1} \frac{\Delta \theta M_i + \frac{1}{2} j}{r_i} (D_\theta g_{i,j})^2 + \sum_{1 \leq i \leq P} \frac{M_i + Q + \frac{1}{2}}{r_i} g_{i,Q} D_\theta g_{i,Q} - \sum_{1 \leq i \leq P} \frac{M_i + \frac{1}{2}}{r_i} g_{i,0} D_\theta g_{i,0}
\]

\[
= - \Delta \theta \sum_{ij} \frac{M_i + \frac{1}{2}}{r_i} (D_\theta g_{i,j})^2.
\]

Here we have used \( M_{i+\frac{1}{2}} = M_{i+\frac{1}{2}, Q+\frac{1}{2}}, g_{i+1,0} = g_{i,Q} \) and \( g_{i,0} = g_{i,Q} \).

Summation by parts in \( j \) gives

\[
III = - \sum_{ij} (g_{i,j+1} - g_{i,j}) r_i f_{i,j+1}
\]

\[
= - \sum_{ij} r_i M_{i,j+1} (g_{i,j+1} - g_{i,j})
\]

\[
= - \frac{1}{2} \sum_{ij} r_i M_{i,j+1} (g_{i,j+1} - g_{i,j})^2 + \frac{1}{2} \sum_{ij} r_i g_{i,j}^2 (M_{i,j+1} - M_{i,j}).
\]
In the two-dimensional case, $\kappa$ is normal if and only if $\kappa$ is either symmetric, i.e., $a = 0$, or antisymmetric, i.e., $M_{i,j+1} = M_{i,j}$. In either case we have

$$2a\Delta r \text{II} \leq 0,$$

hence $\frac{d}{dt} E(t) \leq 0$.

For general matrix $\kappa$, we have

$$\frac{d}{dt} E(t) \leq -D(t) + a\Delta r \sum_{ij} r_i g_{i,j}^2 (M_{i,j+1} - M_{i,j}),$$

where

$$D(t) = \Delta \theta \Delta r \sum_{ij} r_{i+rac{1}{2}} M_{i,j+rac{1}{2}} (D_r g_{i,j})^2 + \Delta \theta \Delta r \sum_{ij} \frac{M_{i,j+rac{1}{2}}}{r_i} (D_\theta g_{i,j})^2.$$

For $A = \max_{ij} \frac{|M_{i,j+1} - M_{i,j}|}{\Delta \theta M_{i,j}}$,

$$\frac{d}{dt} E(t) \leq -D(t) + \frac{aA}{\beta} E(t).$$

By Gronwall's inequality,

$$E(t) \leq e^{\frac{aA}{\beta} t} E(0) - \int_0^t D(\tau) e^{\frac{aA}{\beta} (t-\tau)} d\tau.$$

\[\square\]

**Theorem 2.3.3.** If $\kappa$ is normal, then the physical entropy $E_p(t) = \sum_{ij} f_{i,j} \log(g_{i,j})|K_{ij}|$ is nonincreasing in time, satisfying

$$\frac{d}{dt} E_p(t) \leq -\frac{\Delta \theta}{2} \sum_{ij} r_{i+rac{1}{2}} M_{i,j+rac{1}{2}} \log \left( \frac{g_{i,j+1}}{g_{i,j}} \right) D_r g_{i,j} - \frac{\Delta r}{2} \sum_{ij} \frac{M_{i,j+rac{1}{2}}}{r_i} \log \left( \frac{g_{i,j+1}}{g_{i,j}} \right) D_\theta g_{i,j} \leq 0. \tag{2.27}$$

**Proof.** By mass conservation we obtain

$$\frac{d}{dt} E_p(t) = \sum_{ij} \frac{d}{dt} f_{i,j} (\log g_{i,j} + 1)|K_{ij}|$$

$$= \sum_{ij} \frac{d}{dt} f_{i,j} \log g_{i,j}|K_{ij}| =: I + II + III,$$
where, in virtue of the monotonicity property of the log function,
\[
I = \sum_{ij} \left( \frac{r_{i+\frac{1}{2}}}{2\Delta r_i} M_{i+\frac{1}{2},j} D_r g_{i,j} - \frac{r_{i-\frac{1}{2}}}{2\Delta r_i} M_{i-\frac{1}{2},j} D_r g_{i,j-1} \right) \log g_{i,j} |K_{ij}|
\]
\[
= -\frac{\Delta \theta}{2} \sum_{ij} r_{i+\frac{1}{2}} M_{i+\frac{1}{2},j} D_r g_{i,j} (\log g_{i+1,j} - \log g_{i,j}) \leq 0,
\]
\[
II = \sum_{ij} \left( \frac{M_{i,j+\frac{1}{2}}}{2\Delta \theta r_i^2} D_\theta g_{i,j} - \frac{M_{i,j-\frac{1}{2}}}{2\Delta \theta r_i^2} D_\theta g_{i,j-1} \right) \log g_{i,j} |K_{ij}|
\]
\[
= -\frac{\Delta \theta}{2} \sum_{ij} M_{i,j+\frac{1}{2}} r_i D_\theta g_{i,j} (\log g_{i,j+1} - \log g_{i,j}) \leq 0
\]
and III = \sum_{ij} a D_\theta f_{i,j} \log g_{i,j} |K_{ij}|, which corresponds to the case \( a \geq 0 \). The case \( a < 0 \) can be treated in a similar fashion. By summation by parts in \( j \) we have
\[
III = \sum_{ij} a \Delta r_i (f_{i,j+1} - f_{i,j}) \log g_{i,j}
\]
\[
= \sum_{ij} a \Delta r_i f_{i,j} \log \frac{g_{i,j-1}}{g_{i,j}}.
\]
If \( \kappa \) is symmetric, then III = 0. If \( \kappa \) is antisymmetric, then \( M_{i,j} = M_{i,j-1} \). Therefore,
\[
III = \sum_{ij} a \Delta r_i f_{i,j} \log \frac{f_{i,j-1}}{f_{i,j}}
\]
\[
\leq \sum_{ij} a \Delta r_i f_{i,j} \left( \frac{f_{i,j-1}}{f_{i,j}} - 1 \right)
\]
\[
= \frac{a}{\Delta \theta} \sum_{ij} |K_{ij}| (f_{i,j-1} - f_{i,j}) = 0
\]
due to the conservation of mass.

Similar to the one-dimensional case we can show the long time convergence of solutions of the semidiscrete system.

**Theorem 2.3.4.** Consider the semidiscrete scheme (2.25) subject to the initial data \( f_{i,j}(0) > 0 \) with \( \sum f_{i,j}(0) |K_{ij}| = \int_B f_0(m) \, dm \). If \( \kappa \) is normal, then
\[
f_{i,j}(t) \to CM_{i,j} \quad \text{as } t \to \infty,
\]
where
\[
C = \frac{\int_B f_0(m) \, dm}{\sum_{ij} M_{i,j} |K_{ij}|},
\]
Proof. Define a functional $V(g)$ by

$$V(g) = \sum_{ij} (g_{i,j} - C)^2 M_{i,j} |K_{i,j}|$$

and $\bar{g}_{i,j} = C$. We see that $\bar{g} \in \Sigma$ with

$$\Sigma := \{ g : g_{i,j} \geq 0, \sum_{ij} g_{i,j} M_{i,j} |K_{i,j}| = \int_B f_0(m) \, dm \}$$

is the equilibrium solution. A direct verification shows that $V(g) = E(t) - C \int_B f_0(m) \, dm$, implying that $\frac{d}{dt} V = \frac{d}{dt} E$. Hence $V$ satisfies the following

- $V(g) > 0$ for any $g \neq \bar{g}$ (positive definite),
- $\frac{d}{dt} V \leq 0$ for all $g$ (negative semidefinite),
- the set $\{ \frac{d}{dt} V = 0 \} \cap \Sigma$ does not contain any trajectories of the ODE system besides the trajectory $g(t) = \bar{g} \forall t > 0$.

With these properties we can apply the Krasovskii–LaSalle principle to conclude that $\lim_{t \to \infty} g(t) = \bar{g}$, which leads to the conclusion. We only verify the third property of $V$: from (3) of Theorem 2.3.2 it follows that

$$\frac{d}{dt} V = - \sum_{ij} |K_{i,j}| \frac{r_i + \frac{1}{2}}{r_i} M_{i,j+1} (D_r g_{i,j})^2 - \sum_{ij} |K_{i,j}| \frac{M_{i,j+1} + \frac{1}{2}}{r_i^2} (D_\theta g_{i,j})^2$$

$$- a \Delta r \sum_{ij} r_i M_{i,j+1} (g_{i,j+1} - g_{i,j})^2.$$

If $\frac{d}{dt} V = 0$, then each term in the sum on the right side must vanish, that is

$$D_r g_{i,j} = 0, \quad D_\theta g_{i,j} = 0,$$

which ensures that $g_{i,j} = \text{constant}$, while within $\Sigma$, $g = \bar{g}$ must hold. The proof is thus complete.

\[\square\]

2.3.3 Time discretization

We apply the backward Euler method to (2.25), but treating the asymmetric part explicitly,

\[
\frac{f_{i,j}^{n+1} - f_{i,j}^n}{\Delta t} = \frac{r_i + \frac{1}{2}}{2 \Delta r r_i} M_{i+\frac{1}{2}j} D_r g_{i,j}^{n+1} - \frac{r_i - \frac{1}{2}}{2 \Delta r r_i} M_{i-\frac{1}{2}j} D_r g_{i,j}^{n+1} + \frac{M_{i,j+1} + \frac{1}{2}}{2 \Delta \theta r_i^2} D_\theta g_{i,j}^{n+1} - \frac{M_{i,j-1} + \frac{1}{2}}{2 \Delta \theta r_i^2} D_\theta g_{i,j-1}^{n+1}
\]

\[+ \frac{1}{2} [(a + |a|) D_\theta f_{i,j}^n + (a - |a|) D_\theta f_{i,j-1}^n], \quad (2.29)\]
with \( f_{i,j}^0 = f_{i,j}(0) \). We assume that \( \Delta t \) satisfies the CFL condition

\[
\frac{|a| \Delta t}{\Delta \theta} \leq 1. \tag{2.30}
\]

**Theorem 2.3.5.** The discrete scheme (2.29) with (2.30) satisfies the following properties:

1. \( \sum_{ij} f_{i,j}^{n+1} |K_{ij}| = \sum_{ij} f_{i,j}^0 |K_{ij}| \forall n \in \mathbb{N} \).

2. If the initial data \( f_{i,j}^0 \geq 0 \), then \( f_{i,j}^n \geq 0 \forall n \in \mathbb{N} \).

3. The discrete relative entropy

\[
E^n = \sum_{ij} \frac{(f_{i,j}^n)^2}{M_{i,j}} |K_{ij}|
\]

satisfies \( E^{n+1} \leq E^n \) for \( a = 0 \).

4. \( f_{i,j}^n \) converges as \( n \to \infty \) with

\[ f_{i,j}^n \to CM_{i,j}, \]

where \( C \) is defined in (2.28).

**Proof.** (1) Multiply (2.29) by \( |K_{ij}| \) and sum over \( \{ij\} \) so that

\[
\frac{1}{\Delta t} \left( \sum_{ij} f_{i,j}^{n+1} |K_{ij}| - \sum_{ij} f_{i,j}^n |K_{ij}| \right) = 0.
\]

Therefore

\[
\sum_{ij} f_{i,j}^{n+1} |K_{ij}| = \sum_{ij} f_{i,j}^n |K_{ij}| = \ldots = \sum_{ij} f_{i,j}^0 |K_{ij}|.
\]

(2) Rewrite the scheme (2.29) in terms of \( g_{i,j}^n \) as follows:

\[
- \Delta t \frac{r_{i-1/2} M_{i-1/2,j} g_{i-1,j}^{n+1}}{2(\Delta r)^2 r_i} - \Delta t \frac{r_{i+1/2} M_{i+1/2,j} g_{i+1,j}^{n+1}}{2(\Delta r)^2 r_i} - \Delta t \frac{M_{i,j-1/2} g_{i,j-1}^{n+1}}{2(\Delta \theta)^2 r_i^2} g_{i,j-1} - \Delta t \frac{M_{i,j+1/2} g_{i,j+1}^{n+1}}{2(\Delta \theta)^2 r_i^2} g_{i,j+1} + (M_{i,j} - \cdots) g_{i,j}^{n+1}
\]

\[
= - \frac{(a - |a|) \Delta t M_{i,j-1} g_{i,j-1} g_{i,j}^{n+1}}{2 \Delta \theta} + \left( 1 - \frac{|a| \Delta t}{\Delta \theta} \right) M_{i,j} g_{i,j}^{n+1} + \frac{(a + |a|) \Delta t M_{i,j+1} g_{i,j+1}^{n+1}}{2 \Delta \theta} g_{i,j+1}, \tag{2.31}
\]

where \( \cdots \) is the sum of the coefficients of the first four terms on the left-hand side. The CFL condition (2.30) ensures that the right-hand side of (2.31) is nonnegative. Note that the
coefficient matrix of (2.31) is diagonally dominated. A similar argument to that in the one-dimensional case can be applied here to prove that \( \{g^{n+1}_{i,j}\} \) are nonnegative. It follows that \( \{f^{n+1}_{i,j}\} \) are nonnegative.

(3) We calculate the change of entropy in one time step,

\[
E^{n+1} - E^n = \sum_{ij} \frac{(2f^{n+1}_{i,j} - f^n_{i,j} + f^n_{i+\frac{1}{2},j})(f^{n+1}_{i,j} - f^n_{i,j})}{M_{i,j}} |K_{ij}|
\]

\[
= 2 \sum_{ij} g^{n+1}_{i,j} (f^{n+1}_{i,j} - f^n_{i,j}) |K_{ij}| - \sum_{ij} \frac{(f^{n+1}_{i,j} - f^n_{i,j})^2}{M_{i,j}} |K_{ij}|
\]

\[
= 2(I + II) - \sum_{ij} \frac{(f^{n+1}_{i,j} - f^n_{i,j})^2}{M_{i,j}} |K_{ij}|
\]

where

\[
I = \sum_{ij} g^{n+1}_{i,j} \left( \frac{r_{i+\frac{1}{2}} M_{i+\frac{1}{2},j} D_r g^{n+1}_{i,j} - r_{i-\frac{1}{2}} M_{i-\frac{1}{2},j} D_r g^{n+1}_{i-1,j}}{2\Delta r r_i} \right) |K_{ij}| \Delta t
\]

\[
= \frac{\Delta \theta \Delta t}{2} \left( \sum_{ij} g^{n+1}_{i,j} r_{i+\frac{1}{2}} M_{i+\frac{1}{2},j} D_r g^{n+1}_{i,j} - \sum_{ij} g^{n+1}_{i-1,j} r_{i-\frac{1}{2}} M_{i-\frac{1}{2},j} D_r g^{n+1}_{i,j} \right)
\]

\[
= - \frac{\Delta \theta \Delta r \Delta t}{2} \sum_{ij} r_{i+\frac{1}{2}} M_{i+\frac{1}{2},j} (D_r g^{n+1}_{i,j})^2 \leq 0,
\]

and similarly, by shifting the index in \( j \), we have

\[
II = \sum_{ij} g^{n+1}_{i,j} \left( \frac{M_{i+\frac{1}{2},j+\frac{1}{2}} D_\theta g^{n+1}_{i,j} - M_{i,j+\frac{1}{2}} D_\theta g^{n+1}_{i,j-\frac{1}{2}}}{2\Delta \theta r^2_i} \right) |K_{ij}| \Delta t
\]

\[
= - \frac{\Delta \theta \Delta r \Delta t}{2} \sum_{ij} r_{i+\frac{1}{2}} M_{i+\frac{1}{2},j} (D_\theta g^{n+1}_{i,j})^2 \leq 0.
\]

So \( E^{n+1} \leq E^n \).

(4) Since \( E^n \) is nonincreasing and bounded from below, we have

\[
\lim_{n \to \infty} E^n = \inf \{ E^n \}.
\]

Observe from analysis of (3) that \( E^n - E^{n+1} \) is a sum of nonnegative and bounded terms. When passing limit \( n \to \infty \) we conclude that each term must have zero as its limit, that is

\[
\lim_{n \to \infty} (f^{n+1}_{i,j} - f^n_{i,j})^2 = 0, \quad \lim_{n \to \infty} [(D_\theta g^{n+1}_{i,j})^2 + (D_r g^{n+1}_{i,j})^2] = 0.
\]

(2.32)
The first relation in (2.32) says that $\vec{g}^n$ is a Cauchy sequence, which when combined with the completeness of $\Sigma$ (a closed and bounded set in $\mathbb{R}^{PQ}$) ensures that $\lim_{n \to \infty} \vec{g}^n$ exists. The second relation in (2.32) infers that the limit must be $\vec{g}$. The proof is complete.

**Theorem 2.3.6.** For symmetric $\kappa$, the physical entropy

$$E^n_p = \sum_{ij} f^n_{i,j} \log \frac{f^n_{i,j}}{M_{i,j}} |K_{ij}|$$

is nonincreasing. Moreover,

$$E^{n+1}_p - E^n_p = - \sum_{ij} \log g^{n+1}_{i,j} (g^{n+1}_{i,j} - g^n_{i,j}) |M_{i,j}||K_{ij}| + \sum_{ij} f^n_{i,j} \log \frac{f^{n+1}_{i,j}}{f^n_{i,j}} |K_{ij}| \leq 0.$$

**Proof.** We calculate

$$E^{n+1}_p - E^n_p = \sum_{ij} \left( f^{n+1}_{i,j} \log \frac{f^{n+1}_{i,j}}{M_{i,j}} - f^n_{i,j} \log \frac{f^n_{i,j}}{M_{i,j}} + f^n_{i,j} \log \frac{f^{n+1}_{i,j}}{M_{i,j}} - f^n_{i,j} \log \frac{f^n_{i,j}}{M_{i,j}} \right) |K_{ij}|$$

$$= \sum_{ij} \left[ (f^{n+1}_{i,j} - f^n_{i,j}) \log g^{n+1}_{i,j} + f^n_{i,j} \log \frac{f^{n+1}_{i,j}}{f^n_{i,j}} \right] |K_{ij}|.$$

The second sum is nonpositive since

$$\sum_{ij} f^n_{i,j} \log \frac{f^{n+1}_{i,j}}{f^n_{i,j}} |K_{ij}| \leq \sum_{ij} f^n_{i,j} \left( \frac{f^{n+1}_{i,j}}{f^n_{i,j}} - 1 \right) |K_{ij}| = 0$$

for mass is conserved at each time step. The first sum when recalling the fully discrete scheme may be expressed as I+II with

$$I = \sum_{ij} \Delta t \left( \frac{r_{i+\frac{1}{2}} M_{i+\frac{1}{2},j} D_r g^{n+1}_{i,j}}{2 \Delta r r_i} - \frac{r_{i-\frac{1}{2}} M_{i-\frac{1}{2},j} D_r g^{n+1}_{i,j}}{2 \Delta r r_i} \right) \log g^{n+1}_{i,j} |K_{ij}|$$

$$= - \frac{\Delta t \Delta \theta}{2} \sum_{ij} r_{i+\frac{1}{2}} M_{i+\frac{1}{2},j} D_r g^{n+1}_{i,j} (\log g^{n+1}_{i+1,j} - \log g^{n+1}_{i,j}) \leq 0$$

and

$$II = \sum_{ij} \Delta t \left( \frac{M_{i,j+\frac{1}{2}} D_\theta g^{n+1}_{i,j}}{2 \Delta \theta r_i^2} - \frac{M_{i,j-\frac{1}{2}} D_\theta g^{n+1}_{i,j}}{2 \Delta \theta r_i^2} \right) \log g^{n+1}_{i,j} |K_{ij}|$$

$$= - \frac{\Delta t \Delta r}{2} \sum_{ij} \frac{M_{i,j+\frac{1}{2}}}{r_i} D_\theta g^{n+1}_{i,j} (\log g^{n+1}_{i,j+1} - \log g^{n+1}_{i,j}) \leq 0.$$

These together make the proof complete.
Remark 2.3.2. If the drift term corresponding to the antisymmetric part is made implicit in time discretization, the entropy dissipation relations also hold at fully discrete level when $\kappa$ is normal (for both the physical entropy and the quadratic entropy). But such an implicit treatment does not guarantee the positivity preserving property.

2.4 Numerical implementation and results

Implementation strategies. For the one-dimensional case, we apply the tridiagonal matrix algorithm (also known as the Thomas algorithm) to scheme (2.18). The computation cost is $O(N)$.

For the two-dimensional case, we use a direct method to solve the linear system $Ax = b$ with a sparse $N \times N$ coefficient matrix with $N = PQ$. If the final time $t$ is a multiple of the time step $\Delta t$, the coefficient matrix is the same for each time step. So we only need to compute the LU decomposition once. Furthermore, for large $N$, the sparsity of coefficient matrix reduces the complexity significantly, which is about $O(P^3Q)$. Solving the decomposed system $LUx = b$ costs $O(N^2)$. So the total complexity is $O(N^2)$.

For $\kappa = 0$ or antisymmetric, $M$ is independent of $\theta$, we use the Fourier method in $\theta$ to reduce the computational cost. More precisely, we express the solution as

$$g_{i,j} = \sum_{l=1}^{Q} \hat{g}_{i,l} e^{-\hat{z}(j-1)(l-1)\Delta \theta}, \quad \hat{z} = \sqrt{-1},$$

with its inverse

$$\hat{g}_{i,l} = \frac{1}{Q} \sum_{j=1}^{Q} g_{i,j} e^{\hat{z}(j-1)(l-1)\Delta \theta}.$$  

For each $l$, we obtain a linear system of $(\hat{g}_{1,l}^{n+1}, \ldots, \hat{g}_{P,l}^{n+1})^T$. The Fourier transform and the inverse Fourier transform need $O(PQ^2)$ operations. And for each time step, the computational cost of solving $Q$ linear systems is $O(QP)$, since they all have a tridiagonal coefficient matrix. So the total complexity is $O(PQ^2)$, which with complexity $O(N^{1.5})$ is clearly faster than the direct solver described above.
Numerical tests. We now present our numerical results to demonstrate (i) accuracy of the schemes, (ii) the capacity to capture equilibrium solutions and the large time behavior of the solution, and (iii) the effects of some typical homogeneous flows.

Denote the initial function without normalization by \( \tilde{f}_0(m) \) and the normalized initial data by \( f_0(m) = Z^{-1} \tilde{f}_0(m) \), where \( Z \) is a normalization factor defined by

\[
Z = \int_B \tilde{f}_0(m) \, dm.
\]

We also denote \( Z_M = \int_B M(m) \, dm \).

2.4.1 One-dimensional tests

Denote the numerical solution by \( f_j^n \) and the exact solution by \( f(m_j, t_n) \).

**Definition 1.** \( L^1 \) error is given by

\[
\sum_{j=1}^{N} |f_j^n - f(m_j, t_n)| h,
\]

and \( L^\infty \) error is given by

\[
\max_{1 \leq j \leq N} |f_j^n - f(m_j, t_n)|.
\]

When the exact solution is not available, we replace \( f(m_j, t_n) \) by a reference solution to compute the errors.

2.4.1.1 Accuracy

We illustrate accuracy of scheme (2.18) with several choices of initial data.

**Example 1.** In this example, we consider four kinds of initial data.

(i) \( \tilde{f}_0(m) = (b - m^2)^{ab} \) with \( \alpha = \frac{1}{4}, \frac{1}{2}, \frac{3}{2} \),

(ii) the distance function \( \tilde{f}_0(m) = \sqrt{b} - |m| \),

(iii) the characteristic function \( \tilde{f}_0(m) = \chi_{[-\sqrt{b} + \varepsilon, \sqrt{b} - \varepsilon]} \), \( 0 < \varepsilon < \sqrt{b} \), and

(iv) a cosine function \( \tilde{f}_0(m) = 1 + \cos \left( \frac{2\pi m}{\sqrt{b} - \varepsilon} + \pi \right), 0 < \varepsilon < \sqrt{b} \).
Table 2.1 Error and order of accuracy for **Example 1** on a uniform mesh of \( N \) cells: \( b = 16 \), \( \Delta t = 0.1 \), final time \( t = 1.8 \).

<table>
<thead>
<tr>
<th>( f_0(m) )</th>
<th>((b - m^2)^{\frac{1}{2}})</th>
<th>((b - m^2)^{\frac{3}{2}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>( L^1 ) Error</td>
<td>( L^\infty ) Error</td>
</tr>
<tr>
<td>20</td>
<td>8.0174E-02</td>
<td>4.8757E-02</td>
</tr>
<tr>
<td>40</td>
<td>3.9997E-02</td>
<td>1.003</td>
</tr>
<tr>
<td>80</td>
<td>1.9987E-02</td>
<td>1.001</td>
</tr>
<tr>
<td>160</td>
<td>9.9923E-03</td>
<td>1.000</td>
</tr>
<tr>
<td>320</td>
<td>4.9960E-03</td>
<td>1.000</td>
</tr>
<tr>
<td>640</td>
<td>2.4980E-03</td>
<td>1.000</td>
</tr>
</tbody>
</table>

We take the numerical solution with \( N = 2560 \) as the reference solution. Table 2.1 shows the results from the above initial data when \( b = 16 \).

**Example 2.** We consider the same initial data as in Example 1 but with \( b = 50 \). The results are given in Table 2.2.

### 2.4.1.2 Large time behavior

The normalized equilibrium solution of the Fokker–Planck equation is

\[
 f_{eq}(m) = Z^{-1}_M M(m).
\]

We define the distance of the solution from the equilibrium as

\[
 \max_{1 \leq j \leq N} |f_j^n - f_{eq}(m_j)|.
\]
Table 2.2 Error and order of accuracy for **Example 2** on a uniform mesh of \(N\) cells: \(b = 50, \Delta t = 0.1\), final time \(t = 1.8\).

| \(f_0(m)\) | \((b - m^2)^{\frac{1}{2}}\) | \(\sqrt[b]{\frac{1}{b}} - |m|\) |
|----------------|-------------------------------|---------------------------------|
| \(N\) | \(L^1\) Error | Order | \(L^\infty\) Error | Order | \(L^1\) Error | Order | \(L^\infty\) Error | Order |
| 20 | 1.3517E-01 | 7.5263E-02 | 1.4470E-01 | 8.5089E-02 |
| 40 | 6.7112E-02 | 1.011 | 3.7720E-02 | 1.040 |
| 80 | 3.3499E-02 | 1.002 | 1.866E-02 | 1.003 |
| 160 | 1.6742E-02 | 1.001 | 9.0516E-03 | 1.000 |
| 320 | 8.3703E-03 | 1.000 | 4.2262E-03 | 1.000 |
| 640 | 4.1850E-03 | 1.000 | 2.1337E-03 | 1.000 |

**Example 3.** Take (iv) in Example 1 as the initial data and let \(b = 16, \varepsilon = 0.01\sqrt{b}\). The numerical solutions at \(t = 0, 1.0, 1.8\) are plotted in Figure 2.2, which indicate a fast convergence to the equilibrium state. In Table 2.3 we see that the distance from the equilibrium solution is decreasing. This confirms that the solution converges to the equilibrium solution \(f_{eq}\) as time increases.

### 2.4.1.3 Relative entropy

Now we test the relative entropy of the numerical solutions. The scaled discrete entropy is defined as

\[
\sum_{j=1}^{N} \frac{(f^n_j)^2}{Z_j^{-1} M_j} h.
\]
\[ \tilde{f}_0(m) = 1 + \cos \left( \frac{2m\pi}{\sqrt{b-\varepsilon}} + \pi \right), \quad b = 16, \varepsilon = 0.01\sqrt{b} \]

Table 2.3 Numerical convergence to the equilibrium solution measured by distances for Example 3: \( b = 16, \varepsilon = 0.01\sqrt{b} \) and \( N = 160 \).

<table>
<thead>
<tr>
<th>( \tilde{f}_0(m) )</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ 1 + \cos \left( \frac{2m\pi}{\sqrt{b-\varepsilon}} + \pi \right) ]</td>
<td>2.0962E-01</td>
<td>1.6390E-02</td>
<td>4.2450E-03</td>
<td>1.0971E-03</td>
<td>2.8348E-04</td>
</tr>
</tbody>
</table>

Example 4. We test the time evolution of the relative entropy by using the initial data (i) - (iv) from Example 1. Table 2.4 shows that the relative entropy is nonincreasing.

2.4.2 Two-dimensional tests

Denote the numerical solution by \( f^n_{i,j} \) and the exact solution by \( f(r_i, \theta_j, t_n) \).

Definition 2. \( L^1 \) error is given by

\[ \sum_{ij} |f^n_{i,j} - f(r_i, \theta_j, t_n)|K_{ij} \]

and \( L^\infty \) error is given by

\[ \max_{ij} |f^n_{i,j} - f(r_i, \theta_j, t_n)|. \]
Again when the exact solution is not available, we replace $f(r_i, \theta_j, t_n)$ by a reference solution to compute the errors.

The scaled discrete relative entropy is defined by

$$\sum_{ij} \left( \frac{f^n_{i,j}}{Z^{-1}_M M_{i,j}} \right)^2 |K_{ij}|$$

and the distance from the equilibrium solution by

$$\max_{ij} |f^n_{i,j} - f_{eq}(r_i, \theta_j)|.$$ 

### 2.4.2.1 Accuracy test

We test the two-dimensional accuracy also with several choices of initial data.

**Example 5.** In this test, we consider the two-dimensional problem with $\kappa = 0, b = 40$ and two types of initial data:

(i) $\tilde{f}_0(m) = (b - |m|^2)^{\alpha b}$ with $\alpha = \frac{1}{4}, \frac{1}{2}, \frac{3}{2}$, and

(ii) $\tilde{f}_0(m) = \cos (3\pi \frac{|m|^2}{b}) + 1$.

The results are given in Table 2.5.

In Table 2.6, we choose a symmetric $\kappa$ with different values of $b$ and let $\tilde{f}_0(m) = M(m)$. In this particular case, we know that the exact solution is independent of $t$, which is given by $f_{eq}(m) = Z^{-1}_M M(m)$. 

### Table 2.4  Relative entropy in Example 4: $b = 16, N = 640$ and $\Delta t = 0.1$.

| $t$ | $f_0(m)$ | $(b - m^2)^{\frac{1}{2}}$ | $(b - m^2) \frac{4}{3}$ | $\sqrt{b} - |m|$ | $\frac{\chi_{\sqrt{b}+\epsilon, \sqrt{b}-\epsilon}}{\epsilon}$ | $1 + \cos \left( \frac{2\pi m \sqrt{b} \pm \pi}{\epsilon} \right)$ |
|-----|----------|----------------|-----------------|----------------|----------------|------------------|
| 0   | 1.8141   | 1              | 1.3105          | 1.0129E+12     | 3446.32        | 5280.76          |
| 0.2 | 1.2122   | 1              | 1.1704          | 8.3964         | 35.0358        | 14.8988          |
| 0.6 | 1.0601   | 1              | 1.0575          | 1.5095         | 2.9207         | 3.0020           |
| 1.0 | 1.0207   | 1              | 1.0204          | 1.1435         | 1.4865         | 1.6093           |
| 1.4 | 1.0074   | 1              | 1.0074          | 1.0488         | 1.1608         | 1.2111           |
| 1.8 | 1.0027   | 1              | 1.0027          | 1.0174         | 1.0569         | 1.0755           |
Table 2.5  Error and order of accuracy for **Example 5**: $b = 40$, $\kappa = 0$, $\Delta t = 0.05$, final time $t = 4$ and the reference solution is given by $P = Q = 320$. 

<table>
<thead>
<tr>
<th>$f_0(m)$</th>
<th>$(b - m^2)^{\frac{3}{2}}$</th>
<th>$(b - m^2)^{\frac{3}{2}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P = Q$</td>
<td>$L^1$ Error</td>
<td>Order</td>
</tr>
<tr>
<td>20</td>
<td>1.0210E-01</td>
<td>0.996</td>
</tr>
<tr>
<td>40</td>
<td>5.1202E-02</td>
<td>0.996</td>
</tr>
<tr>
<td>80</td>
<td>2.5621E-02</td>
<td>0.999</td>
</tr>
</tbody>
</table>

Table 2.6  Error and order of accuracy for **Example 5**: $k_{11} = 0.5$, $k_{12} = k_{21} = 0.15$, $\Delta t = 0.05$, final time $t = 4$. 

<table>
<thead>
<tr>
<th>$f_0(m)$</th>
<th>$M(m)$ with $b = 40$</th>
<th>$M(m)$ with $b = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P = Q$</td>
<td>$L^1$ Error</td>
<td>Order</td>
</tr>
<tr>
<td>10</td>
<td>3.2228E-01</td>
<td>0.996</td>
</tr>
<tr>
<td>20</td>
<td>1.6176E-01</td>
<td>0.996</td>
</tr>
<tr>
<td>40</td>
<td>8.1719E-02</td>
<td>0.995</td>
</tr>
</tbody>
</table>

### 2.4.2.2 Entropy decreasing and large time behavior

**Example 6.** Consider the initial data (i) and (ii) in Example 5 with symmetric $\kappa$, i.e., $a = 0$. The Fokker–Planck equation has an equilibrium solution $f_{eq}(m) = Z_M^{-1} M(m)$ whose relative entropy is 1. Table 2.7 shows that the relative entropy is nonincreasing and converges to 1.

Table 2.7  Relative entropy in **Example 6**: $P = Q = 40$, $b = 40$, $\kappa = 0$. 

| $\tilde{f}_0(m)$ | $(b - m^2)^{\frac{3}{2}}$ | $(b - m^2)^{\frac{3}{2}}$ | $\cos \left(3 \pi \frac{|m|^2}{b} \right) + 1$ |
|-----------------|-----------------|-----------------|-----------------|
| $t$             | $(b - m^2)^{\frac{3}{2}}$ | $(b - m^2)^{\frac{3}{2}}$ | $\cos \left(3 \pi \frac{|m|^2}{b} \right) + 1$ |
| 1               | 1.06563         | 1               | 1.04837         | 9.25247         |
| 2               | 1.00596         | 1               | 1.00448         | 1.3167          |
| 3               | 1.00056         | 1               | 1.00042         | 1.02827         |
| 4               | 1.00005         | 1               | 1.00004         | 1.00267         |
Especially, in the second column where we take the equilibrium solution as the initial data, the relative entropy stays the same.

**Example 7.** Let \( a = 0 \), i.e., \( k_{12} = k_{21} \). A comparison of solution behavior for two different initial data but with the same \( b = 16 \) is plotted in Figure 2.3 and 2.4. Moreover, Table 2.8 shows that solutions in these two tests converge to the equilibrium solution.

**Figure 2.3** \( \tilde{f}_0(m) = (b - |m|^2)^{\frac{1}{2}}, b = 16, k_{11} = 1.1, k_{12} = k_{21} = 0.15, \Delta t = 0.05 \).

**Figure 2.4** \( \tilde{f}_0(m) = \cos \left( \frac{3\pi |m|^2}{b} \right) + 1, b = 16, k_{11} = 1.1, k_{12} = k_{21} = 0.15, \Delta t = 0.05 \).

**Table 2.8** Numerical convergence to the equilibrium solution measured by distances for **Example 7**: \( b = 16, k_{11} = 1.1, k_{12} = k_{21} = 0.15, \Delta t = 0.05, P = Q = 40 \).

<table>
<thead>
<tr>
<th>( \tilde{f}_0(m) )</th>
<th>( t )</th>
<th>3</th>
<th>6</th>
<th>9</th>
<th>12</th>
<th>15</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td>( (b -</td>
<td>m</td>
<td>^2)^{\frac{1}{2}} )</td>
<td></td>
<td>8.80349E-02</td>
<td>1.42082E-02</td>
<td>2.27856E-03</td>
<td>3.75317E-04</td>
</tr>
<tr>
<td>( \cos \left( \frac{3\pi</td>
<td>m</td>
<td>^2}{b} \right) + 1 )</td>
<td></td>
<td>4.64376E-02</td>
<td>7.56072E-03</td>
<td>1.30858E-03</td>
<td>3.12548E-04</td>
</tr>
</tbody>
</table>
2.4.2.3 Positivity preserving

Another feature of our scheme is positivity preserving. Let
\[ f_{n}^{\text{min}} = \min_{i,j} f_{i,j}^{n} \]
denote the minimum of numerical solutions over all computational cells at \( t = t_n \). In two numerical tests presented in Figure 2.3 and 2.4, we obtain \( f_{n}^{\text{min}} = 0 \) at all time steps tested. Positivity may also be observed through visualizing the numerical solution from a different angle. We see in Figure 2.5 that the solutions displayed in (d) of Figure 2.3 and 2.4 are strictly nonnegative.

![Figure 2.5](image)

(a) \( t=3 \)  
(b) \( t=3 \)

Figure 2.5  \( b = 16, k_{11} = 1.1, k_{12} = k_{21} = 0.15, \Delta t = 0.05 \), Left: \( \tilde{f}_0(m) = (b - |m|^2)^{0.5} \). Right: \( \tilde{f}_0(m) = \cos \left( 3\pi \frac{|m|^2}{b} \right) + 1 \).

2.4.2.4 Flow effects

Let \( (x, y) \) be the macroscopic Eulerian coordinate and \( \nabla = \left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right)^T \), associated with a fluid velocity field \( \tilde{v}(x, y) \).

**Example 8.** Simple extensional flow.

We consider a homogeneous planar strain flow with the velocity field
\[ \tilde{v} = (\alpha x, -\alpha y), \]
where $\alpha$ is the extensional rate. Then the velocity gradient tensor is

$$\kappa = \nabla \vec{v} = \begin{pmatrix} \alpha & 0 \\ 0 & -\alpha \end{pmatrix}.$$  

This flow is irrotational and forms a strain flow. With this extensional flow, we consider the initial data with four separate peaks, defined by $\tilde{f}_0(m) = \delta_\varepsilon(m)$, where

$$\delta_\varepsilon(m) = \begin{cases} 
\frac{\cos\left(\frac{\pi(m_1 - m_{10})}{\varepsilon}\right) + 1}{2\varepsilon}, & |m_1 - m_{10}| \leq \varepsilon \text{ and } |m_2 - m_{20}| \leq \varepsilon, \\
0, & \text{elsewhere},
\end{cases}$$

where $(m_{10}, m_{20}) \in \{(\pm \beta, 0), (0, \pm \beta)\}$ and $\varepsilon < \beta < \sqrt{b} - \varepsilon$.

Note that in such a case the normalized equilibrium solution is

$$f(m) = Z_M^{-1} M(m), \quad M(m) = (b - |m|^2)^\frac{b}{2} e^{\alpha(x^2 - y^2)}.$$ 

The solutions at different times are plotted in Figure 2.6. In these tests we can see that the proposed method can well capture the equilibrium solutions for extensional flows.

![Figure 2.6](attachment:image.png)

Figure 2.6 $\tilde{f}_0(m) = \delta_\varepsilon(m), b = 16, \alpha = 1.1, \varepsilon = 2\Delta r, \beta = 2, P = Q = 40, \Delta t = 0.05$.

The contours in Figure 2.7 show how the equilibrium solution $f_{eq}(m) = Z_M^{-1} M(m)$ changes with respect to $\alpha$. Observe that the two peaks of the equilibrium solution move away from each other as $\alpha$ gets larger. For large $\alpha$ one expects to see sharp peaks near boundary, with an amplification factor $e^{\alpha b}$ of the profile $(b - |m|^2)^\frac{b}{2}$. Due to low order of our scheme, its performance when $\alpha b$ becomes large tends less satisfactory. A higher order extension of the present method constitutes a future publication.
Example 9. Steady state shear flow.

The steady state shear flow has the velocity field

\[ \mathbf{v} = (\gamma y, 0), \]

where \( \gamma \) is a constant shear rate, and the velocity gradient tensor is

\[ \kappa = \nabla \mathbf{v} = \begin{pmatrix} 0 & \gamma \\ 0 & 0 \end{pmatrix}. \]

Let \( \gamma = 0.1, 0.3, 0.5, 1.0, 2.0 \). Figure 2.8 gives the contour plots of \( f^n_{ij} \) at \( t_n = 4 \), from which the shear effects are clearly seen. Note that since for shear flow, \( \kappa \) is not normal, we do not expect the scheme to capture the large time behavior of the solution.

Example 10. A vortex. A typical vortex has the velocity field

\[ \mathbf{v} = (-\gamma y, \gamma x), \]

with velocity gradient tensor

\[ \kappa = \nabla \mathbf{v} = \begin{pmatrix} 0 & \gamma \\ -\gamma & 0 \end{pmatrix}. \]

Note that \( \kappa \) is not symmetric but normal, i.e., \( \kappa^T \kappa = \kappa \kappa^T \), hence \( f_{eq}(m) = Z^{-1}_M M(m) \) is still an equilibrium solution. In addition, \( \kappa^s = 0 \) in this case, so \( M(m) = (b - |m|^2)^\frac{b}{2} \). Table 2.9 shows the convergence to \( f_{eq} \) as \( t \) increases.
Figure 2.8  The contours of \( f_{i,j}^n \) at \( t_n = 4 \) where \( \tilde{f}_0(m) = (b - |m|^2)^{\frac{k}{2}}, b = 16, P = Q = 40, \Delta t = 0.05 \).

2.5 Conclusion

In this paper, we have investigated the Fokker–Planck equation which is of bead-spring type FENE dumbbell model for polymers, with our focus on the development of an entropy satisfying numerical method for the Fokker–Planck equation subject to zero flux on the boundary. We constructed simple and easy-to-implement conservative schemes which preserve equilibrium solutions and proved that they satisfy all three desired properties of the pdf, i.e., constant integral (mass conservation), positivity preserving and entropy satisfying for \( \kappa \) normal. We also proved the long time convergence to the equilibrium solution at discrete levels. The goal of our future work is to extend the numerical method and analytical results herein to a higher order discontinuous Galerkin method.
Table 2.9  Numerical convergence to the equilibrium solution measured by distances for Example 10: $b = 16, \gamma = 0.15, \Delta t = 0.05, P = Q = 40$.

<table>
<thead>
<tr>
<th>$\tilde{f}_0(m)$</th>
<th>$t$</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_e(m)$</td>
<td></td>
<td>8.39533E-02</td>
<td>2.44483E-02</td>
<td>2.25787E-03</td>
<td>2.14324E-04</td>
<td>2.03809E-05</td>
</tr>
</tbody>
</table>

**Acknowledgments**

The authors thank the referees who provided valuable comments resulting in improvements in this paper. This research was supported by the National Science Foundation under Kinetic FRG grant DMS07-57227.
CHAPTER 3. THE ENTROPY SATISFYING DISCONTINUOUS GALERKIN METHOD FOR FOKKER–PLANCK EQUATIONS, WITH APPLICATIONS TO THE FINITELY EXTENSIBLE NONLINEAR ELASTIC DUMBBELL MODEL

A paper submitted to SIAM

Hailiang Liu and Hui Yu

Abstract

Computation of Fokker–Planck equations with satisfying long time behavior is important in many applications and difficult in resolving solution structures induced by nonstandard forces. Entropy satisfying conservative methods are proven to be powerful to ensure both equilibrium preserving and mass conservation properties at the discrete level. Following Liu and Yu (2012a), we present entropy satisfying discontinuous Galerkin methods to solve the Fokker–Planck equation of the finitely extensible nonlinear elastic dumbbell model for polymers, subject to homogeneous fluids. Both semidiscrete and fully discrete methods satisfy two desired properties: mass conservation and entropy satisfying in the sense that these schemes are shown to satisfy discrete entropy inequalities for the quadratic entropy. These ensure that the schemes are entropy satisfying and preserve the equilibrium solutions. It is also proved the convergence of numerical solutions to the equilibrium solution as time becomes large. Zero flux at the boundary is naturally incorporated and boundary behavior is resolved sharply. A positive numerical approximation is obtained with the same accuracy as the numerical solution through a reconstruction at the final time. Both one- and two-dimensional numerical results are
provided to demonstrate the good qualities of the schemes, as well as effects of some canonical homogeneous flows.

3.1 Introduction

This paper is the continuation of our project, initiated in Liu and Yu (2012a), of developing entropy satisfying numerical methods to solve the Fokker–Planck equation of the finitely extensible nonlinear elastic (FENE) dumbbell model. Let \( f = f(x,m,t) : \mathbb{R}^d \times B \times \mathbb{R}^+ \to \mathbb{R} \) denote the probability density function, then the dimensionless FENE model is

\[
\partial_t f + (v \cdot \nabla_x)f = \frac{1}{2D_e} \nabla_m \cdot (\nabla_m f - Ff),
\]

where \( m \) is the \( d \)-dimensional connector vector of the beads, \( D_e \) is a dimensionless number called the Deborah number, \( v(x,t) \) is the fluid velocity and \( F \) is the force field defined by

\[
F(m) = 2D_e \nabla_x v m - \frac{bm}{b - |m|^2}, \quad m \in B := \{|m|^2 \leq b\},
\]

which indicates that the FENE model takes into account the finite extensibility of the polymer chain through an important explosive force as \( m \) tends to \( \sqrt{b} \), the maximum spring extension.

In this work, we consider only homogeneous fluids with the velocity field \( v = \kappa x \), where \( \kappa = \nabla v \) is independent of the position vector \( x \) and has zero trace for incompressible fluids. A reformulation of equation (3.1) along the flow map and taking \( D_e = 1 \), subject to both initial data and zero boundary flux, lead to the following problem

\[
\begin{align*}
\partial_t f &= \frac{1}{2} \nabla_m \cdot \left[ \nabla_m f + \left( \frac{bm}{b - |m|^2} - 2\kappa m \right) f \right], \quad m \in B, \quad (3.2a) \\
n(m, 0) &= f_0(m), \quad m \in B, \quad (3.2b) \\
\left[ \nabla_m f + \left( \frac{bm}{b - |m|^2} - 2\kappa m \right) f \right] \cdot m &= 0, \quad m \in \partial B, \quad (3.2c)
\end{align*}
\]

which we shall study in this work with attention on the case of \( b > 2 \). Indeed, if \( b < 2 \), the FENE model yields many solutions unless a further requirement on the boundary behavior of \( f \) is imposed; see Liu and Shin (2012a,b). A general discussion of this problem and background references are given in the introduction of Liu and Yu (2012a).
Our interest is to develop higher order entropy satisfying methods for Fokker–Planck equations subject to various forces. Problem (3.2) is of particular interest since it involves two special difficulties: (i) nonconservative force due to the fluid coupling and (ii) the singularity of the Fokker–Planck equation near $|m| = \sqrt{b}$. These together present numerous challenges, both analytically and numerically. The boundary singularity issue is particularly important in the presence of fluid coupling, in which the behavior of the polymer distribution near the boundary is of significance. Consequently, computing with sharp resolution and stability near the boundary is a major goal. In order to achieve this goal, three main properties of the solution of the Fokker–Planck equations are naturally desired for numerical schemes solving the equation, including the nonnegativity principle, the mass conservation and the existence of nonzero steady states (equilibrium). For Fokker–Planck type equations, it is often a challenge to find a proper formulation and an appropriate discretization to satisfy all these natural constraints; see the lecture notes (Le Bris and Lelièvre, 2011, Section 4.3).

Our idea is to reformulate the Fokker–Planck equation (3.2a) as

$$\partial_t f = \frac{1}{2} \nabla_m \cdot \left( M \nabla_m \left( \frac{f}{M} \right) - 2\kappa^a m f \right), \quad (3.3)$$

where $M(m) = (b - |m|^2)^{\frac{b}{2}} \exp(m^T \kappa^s m)$ with $\kappa = \kappa^s + \kappa^a$ as the decomposition of symmetric and antisymmetric parts (see Section 3.2). With this formulation, we may use the relative entropy to test the stability of the model as well as its numerical approximation. For any convex function $H : \mathbb{R} \to \mathbb{R}$, using the zero flux boundary condition (3.2c), we have

$$\frac{d}{dt} \int_B MH \left( \frac{f}{M} \right) \, dm = -\frac{1}{2} \int_B MH'' \left( \frac{f}{M} \right) \left| \nabla_m \frac{f}{M} \right|^2 \, dm \leq 0, \quad (3.4)$$

provided $\kappa$ is normal so that $CM$ is the equilibrium for some constant $C$. One may take $H(g) = g^2$ to obtain a weighted $L^2$ estimate or $H(g) = g \log g$ to bound the physical entropy. We refer to Arnold et al. (2001) for principles of entropy methods for Fokker–Planck equations.

The entropy satisfying property will be tested using $H = g^2$ in (3.4), which is particularly convenient for higher order Galerkin methods.

As for the configurational discretization, we explore higher order discontinuous Galerkin (DG) methods based on formulation (3.3), while the finite volume method introduced in Liu
and Yu (2012a), satisfying all three desired properties, may be viewed as a first order DG method. The DG method is a finite element method using a completely discontinuous piecewise polynomial space for the numerical solution and the test functions. One main advantage of the DG method was the flexibility afforded by local approximation spaces combined with the suitable design of numerical fluxes crossing cell interfaces. In this work we follow the methodology of the Direct Discontinuous Galerkin (DDG) method. The main feature of the DDG schemes proposed in Liu and Yan (2009, 2010) for diffusion problems lies in numerical flux choices for the solution gradient, which involves higher order derivatives evaluated crossing interfaces, motivated by a trace formula for the derivatives of the heat solution [Liu and Yan (2009)].

However, the DDG numerical flux given in Liu and Yan (2009, 2010) cannot be directly used, due to several novel features of (3.3): (1) the equilibrium $M$ has no positive lower bound, but zero at the boundary, making (3.3) singular and numerical computations more difficult; (2) the force due to fluid effects is generally nonconservative so that one has to consider a nonsymmetric perturbation upon the usual Landau formulation for Fokker–Planck equations, which makes the study of long time convergence more interesting; (3) the natural function space for $f$ is $ML^2(Mdm)$, which when $M$ vanishes at the boundary is different from the usual weighted space $L^2(Mdm)$ [Liu and Shin (2012a,b)].

More precisely, this work involves several steps: we first apply the DG discretization to (3.3) with $f/M$ replaced by $g$, including interface corrections determined by a DDG type numerical flux, in one-dimensional case, of the form

$$\bar{\partial}_m g = \frac{\beta_0}{h} [g] + \partial_m g + h\beta_1 [\partial^2_m g];$$

we then identify a key quantity $\Gamma(\beta_1)$ such that when $\beta_0 > 2\Gamma(\beta_1) M$, the resulting semidiscrete DG scheme satisfies the discrete version of entropy inequality (3.4) with $H(g) = g^2$. This ensures that the scheme preserves the equilibrium state and has the capacity of capturing the large time behavior of solutions correctly. We further present an approach to characterize $\Gamma$ in terms of $M$, which shows that the above flux satisfying $\beta_0 > 2\Gamma(\beta_1) M$ becomes adaptive in terms of both interface location and the size of the elongation parameter $b$. This adaptive
feature brings several advantages: (i) it has the ability of handling a wide range of $b$, particularly for more realistic case when $b$ is large and (ii) when $\kappa$ is normal, the method captures the long time behavior of solutions very well as predicted by the analysis and evidenced by the two dimensional tests on both the simple extensional flow and the vortex flow. In addition, for both semidiscrete and fully discrete DG schemes, we prove that numerical solutions converge towards the equilibrium state as time becomes large, i.e.,

$$\lim_{t \to \infty} f(t, m) = CM(m) \quad \text{where} \quad C = \int_B f_0 \, dm.$$

For general $\kappa$ such as the shear flow, $M$ is no longer an equilibrium solution, the equilibrium preserving property is not guaranteed. However, the method is still entropy stable for finite time; see Theorem 3.4.4 in Section 3.4. Indeed the numerical results show quite good convergence for short time simulations; see Example 5 in Section 3.5. We also provide an approach to construct a nonnegative numerical approximation from the obtained numerical solution at final time, without destroying the accuracy of the numerical solution.

We should point out that for nonhomogeneous flows which is the case when considering the coupled system with the Navier–Stokes equation, we may apply the method developed in this paper using operator splitting. For instance, for each fixed $m$, one may solve the transport equation

$$\partial_t f + \nabla \cdot (vf) = 0,$$

with $v$ obtained from solving the Navier–Stokes equation. With the obtained $f$ as initial data, then one further solves the Fokker–Planck equation with $\kappa = \nabla_x v(x, t)$.

### 3.1.1 Related work

The regime under consideration is $b > 2$, for which the zero flux boundary condition (3.2c) is shown to be equivalent to the sharp requirement for the solution to remain a probability density [Liu and Shin (2012b)]:

$$f = o(b - |m|^2) \quad \text{on} \ \partial B. \quad (3.5)$$

For theoretical results concerning the existence of solutions of the coupled system we refer to Liu and Shin (2012b), Masmoudi (2008), Zhang and Zhang (2006); see also the works Chupin
(2009a,b, 2010) and the earlier works on this problem: Jourdain and Lelièvre (2003), Jourdain et al. (2004). See Jourdain et al. (2006), Arnold et al. (2010) for rigorous analysis of long-time asymptotics of the FENE model and Arnold et al. (2001) for entropy methods to study rate of convergence to the equilibrium for Fokker–Planck type equations. For some special configuration solutions with small flow rates, the use of moment closure approximations has been investigated by several authors; see, e.g., Du et al. (2005), Hyon et al. (2008), Herrchen and Öttinger (1997), Samaey et al. (2011), Wang et al. (2008). Most numerical methods developed for the Fokker–Planck equation have been based on the form of (3.2a); see, for example, Ammar et al. (2006, 2007), Fan (1985) and Warner (1972). Some elaborate numerical algorithms based on spectral methods were recently developed for the Fokker–Planck equation of the FENE model in Chauvière and Lozinski (2003, 2004a,b). A spectral Galerkin approximation was further introduced in Knezevic and Süli (2009) based on a weighted weak formulation for \( f(m,t)(b - |m|^2)^{-\frac{4}{3}} \). An improved weighted formulation was proposed in Shen and Yu (2012) in terms of \( f(m,t)(b - |m|^2)^{-s} \) for \( 1 < s \leq \frac{5}{2} \), leading to a different spectral Galerkin algorithm. We note that this weighted formulation was also used for specific values of \( s \) in Chauvière and Lozinski (2003, 2004a,b) and was analyzed in Section 3.2 of Knezevic and Süli (2009). The methods in Knezevic and Süli (2009) and Shen and Yu (2012) have provable stability results for certain weighted integrable initial data.

The application of DG methods to hyperbolic problems has been quite successful; see, e.g., Reed and Hill (1973) for solving linear equations and Cockburn and Shu (1989), Cockburn et al. (1989), Cockburn et al. (1990) for solving nonlinear equations. However, the application of the DG method to diffusion problems has been a challenging task because of the subtle difficulty in defining an appropriate numerical flux for the solution gradient; see the earlier works Arnold (1982), Baker (1977), Wheeler (1978) using the interior penalty (IP) method. In the past decade there has been a renewed interest in developing DG methods to solve the diffusion problem, including the method originally proposed by Bassi and Rebay (1997) for compressible Navier–Stokes equations, its generalization called the local discontinuous Galerkin (LDG) methods introduced in Cockburn and Shu (1998) and further studied in Castillo et al. (2000), Cockburn and Dawson (2000) and Cockburn et al. (2001), as well as the method
introduced by Baumann and Oden (1999) and Oden et al. (1998). We refer to Arnold et al.
(2002) for the unified analysis of DG methods for elliptic problems and background references
for the IP methods. The direct discontinuous Galerkin (DDG) methods introduced in Liu and
Yan (2009, 2010) adopt a different strategy, which is to solve the higher order PDE directly
by the DG discretization with the special numerical flux for the solution gradient, yet without
rewriting the equation into a first order system. There are other recent works sharing the direct
feature, such as those in van Leer and Nomura (2005), Gassner et al. (2007) and Cheng and
Shu (2007), all based on certain weak formulation derived from repeated integration by parts
for the diffusion term. More general information about DG methods for elliptic, parabolic and
hyperbolic partial differential equations can be found in the recent books and lectures notes
[Hesthaven and Warburton (2007), Li (2006), Riviére (2008), Shu (2009)].

Finally we comment on the concept of entropy explored in numerical approximations. There
is a vast literature on entropic schemes for related equations including hyperbolic conservation
laws and kinetic equations such as Fokker–Planck type equations. For Fokker–Planck equations
subject to conservative forces such as $F = -\nabla U$ with $M = e^{-U}$, one often uses the logarithmic
Landau form,

$$\partial_t f = \frac{1}{2} \nabla \cdot \left( f \nabla \log \frac{f}{M} \right),$$

to ensure the entropy property at discrete levels; see, e.g., Buet et al. (2001). Another class
of finite difference schemes for Fokker–Planck equations is due to Chang–Cooper; see Chang
and Cooper (1970) and Larsen et al. (1985). This method is based upon the requirement that
the discrete Fokker–Planck operator possesses a quasi-equilibrium solution which agrees at the
mesh points with a quasi-equilibrium solution of the analytic operator. For a linear Fokker–
Planck equation the Chang–Cooper scheme is shown in Buet and Dellacherie (2010) to make
the underlying equation equivalent to the nonlogarithmic Landau form

$$\partial_t f = \frac{1}{2} \nabla \cdot \left( M \nabla \frac{f}{M} \right)$$

at the discrete level.
3.1.2 Contents

This paper develops high order DG methods for problem (3.2). In Section 3.2, we briefly state the reformulation and associated entropy dissipation inequality, which are important for the design of the DG method. In Section 3.3, we describe our scheme for the one-dimensional case. Theoretical analysis for both semidiscrete and fully discrete schemes is provided. In Section 3.4, we generalize the scheme to two space dimensions. Numerical results of both one and two dimensions are presented in Section 3.5. Finally, in Section 3.6, concluding remarks are given.

3.2 Reformulation

We consider the Fokker–Planck equation (3.2a) with $Tr(\kappa) = 0$. If $\kappa$ is normal in the sense that it commutes with its transpose, i.e., $\kappa \kappa^T = \kappa^T \kappa$, one can verify that the equilibrium solution can be determined explicitly as $CM(m)$ where

$$M(m) = (b - |m|^2)^{\frac{1}{2}} \exp(m^T \kappa^s m). \tag{3.6}$$

Here $\kappa^s$ is the symmetric part of $\kappa$ and $\kappa^a = \kappa - \kappa^s$ is the antisymmetric part. The Fokker–Planck equation can be rewritten as (3.3), i.e.,

$$\partial_t f = \frac{1}{2} \nabla_m \cdot \left( M \nabla_m \left( \frac{f}{M} \right) - 2 \kappa^a m f \right). \tag{3.7}$$

Defining a new function $g = \frac{f}{M}$, then (3.7) can be written as

$$2M \partial_t g = \nabla_m \cdot \left[ M \nabla_m g - 2M \kappa^a mg \right]. \tag{3.8}$$

We state the following fact verified in Liu and Yu (2012a) for $\kappa$ normal.

**Lemma 3.2.1.** [LY] Let $f$ be the solution to (3.2). If $\kappa$ is normal, then the associated equilibrium solution $f_{eq}(m) = CM(m)$ where

$$C = \frac{\int_B f_0 dm}{\int_B M dm}.$$  

Moreover, the relative entropy $E(t) = \int_B g^2 M dm$ satisfies

$$\frac{d}{dt} E(t) + \int_B M |\nabla_m g|^2 dm = 0. \tag{3.9}$$
We will design a class of DG methods to compute the numerical solution $g$ to (3.8) and then obtain $f$ by using $f = Mg$ such that the entropy dissipation inequality (3.9) will be satisfied at the discrete setting.

### 3.3 One-dimensional Fokker–Planck Equation

In the one-dimensional case, we have $\kappa = 0$ and the Fokker–Planck problem for (3.8) becomes

\begin{align*}
2M \partial_t g &= \partial_m (M \partial_m g) \quad \text{in } B \times (0, \infty), \\
g(m, 0) &= g_0(m) \quad \text{in } B, \\
M \partial_m g &= 0 \quad \text{on } \partial B \times (0, \infty),
\end{align*}

where $B = [-\sqrt{b}, \sqrt{b}]$ and $M(m)$ reduces to

$$M(m) = (b - m^2)^{\frac{b}{2}}.$$

#### 3.3.1 Semidiscrete DG

We first discretize the equation with respect to $m$. Let

$$-\sqrt{b} = m_\frac{1}{2} < m_\frac{3}{2} < \cdots < m_\frac{N-1}{2} < m_\frac{N+1}{2} = \sqrt{b}$$

be a partition of the domain $B$ into subintervals $I_j = [m_{j-\frac{1}{2}}, m_{j+\frac{1}{2}}]$ of length $h_j = m_{j+\frac{1}{2}} - m_{j-\frac{1}{2}}$, $j = 1, \ldots, N$. The center of $I_j$ is defined as

$$m_j = m_{j-\frac{1}{2}} + \frac{1}{2} h_j.$$

Notice that $M(m_\frac{1}{2}) = M(m_{N+\frac{1}{2}}) = 0$ and $M(m_{j+\frac{1}{2}}) > 0$ for $1 \leq j \leq N - 1$. The weak formulation of (3.10) on $I_j$ is

$$\int_{I_j} M \partial_t g v dm = - \int_{I_j} M \partial_m g \partial_m v dm + M \partial_m g v \bigg|_{\partial I_j} \quad \forall v \in H^1(M dm).$$

Let the solution space $V_h$ be defined as

$$V_h = \{ v : v|_{I_j} \in P^k(I_j), \quad j = 1, \cdots, N \},$$
where $P^k(I_j)$ is the space of polynomials of degree up to $k$ on $I_j$. When the weak formulation (3.11) is enforced into the solution space $V_h$, interface corrections must be added to form a stable method. We thus propose the following DG method: find $g_h \in V_h$ such that for $1 \leq j \leq N$,

$$
\int_{I_j} M \partial_t g_h v \, dm = - \int_{I_j} M \partial_m g_h \partial_m v \, dm + M[\overline{\partial_m g_h v} + (g_h - \hat{g}_h)\partial_m v]\bigg|_{m_{j-\frac{1}{2}}}^{m_{j+\frac{1}{2}}} \quad \forall v \in V_h, (3.12)
$$

subject to the initial data $g_h(m,0)$ generated by the weighted $L^2$ projection

$$
\int_{I_j} M [g_h(m,0) - g_0(m)] v(m) \, dm = 0 \quad \forall v \in V_h. \quad (3.13)
$$

The boundary contribution at $m_{\frac{1}{2}}$ and $m_{N+\frac{1}{2}}$ are taken to be zero to incorporate the zero flux condition. The numerical fluxes are single-valued functions defined only on the interfaces $m_{j+\frac{1}{2}}$ and should be designed to ensure the entropy stability. Hereafter we assume uniform mesh size $h$ only for simplicity. We follow the DDG method in Liu and Yan (2010) and take

$$
\overline{\partial_m g_h} = \frac{\beta_0}{h} [g_h] + \overline{\partial_m g_h} + \beta_1 h [\partial^2_m g_h] \quad \text{and} \quad \hat{g}_h = \overline{g_h}, \quad (3.14)
$$

where the parameters $(\beta_0, \beta_1)$ may vary with $j$. The notations in (3.14) are defined as

$$
[g] = g^+ - g^- \quad \text{and} \quad \overline{g} = \frac{g^+ + g^-}{2},
$$

where $g^-$ and $g^+$ are the values of $g$ at $m_{j+\frac{1}{2}}$ from the left and right cells respectively.

The global DG formulation is obtained by summing (3.12) over all cells

$$
\sum_{j=1}^N \int_{I_j} M \partial_t g_h v \, dm + A(g_h, v) = 0 \quad \forall v \in V_h, \quad (3.15)
$$

where the bilinear operator $A$ is defined as

$$
A(g_h, v) = \sum_{j=1}^N \int_{I_j} M \partial_m g_h \partial_m v \, dm + \sum_{j=1}^{N-1} M(\overline{\partial_m g_h} [v] + [g_h]\overline{\partial_m v})\bigg|_{m_{j+\frac{1}{2}}}^{m_{j+\frac{1}{2}}}.
$$

The algorithm is well defined once the parameters $\beta_0$ and $\beta_1$ are wisely chosen.

### 3.3.2 Numerical flux and coercivity of $A$

The numerical flux at the cell interface $m_{j+\frac{1}{2}}$ is designed such that it depends only on the left and right polynomials and that it (i) is consistent with $\partial_m g$ when $g$ is smooth, (ii)
is conservative in the sense that the flux is single valued on $m_{j+\frac{1}{2}}$, (iii) ensures the entropy stability, and (iv) enforces the high order accuracy of the method. The form (3.14) makes the numerical flux adopted in (3.12) both consistent and conservative. In this section, we will provide a detailed discussion on how to choose the parameters $(\beta_0, \beta_1)$ to ensure the coercivity of $A$ and hence the entropy satisfying property.

To this end, we define the discrete energy norm of $v \in V_h$

$$\|v\|_E^2 = \sum_{j=1}^{N} \int_{I_j} M|\partial_m v|^2 \, dm + \frac{N-1}{h} \sum_{j=1}^{N-1} M|v|^2\Big|_{m_{j+\frac{1}{2}}}$$

(3.16)

and the quantity

$$\Gamma_j := \Gamma(\beta_1, w_j) = \sup_{u \in P^{k-1}([-1,1]) \setminus \{0\}} \frac{(u(1) - 2\beta_1 \partial_x u(1))^2}{\int_{-1}^{1} w_j(x) u^2(x) \, dx}.$$

(3.17)

Here the weight functions $w_j(x)$ are defined as

$$w_j(x) = \min\{w_{jl}(x), w_{jr}(x)\} \quad \text{for } 1 \leq j \leq N - 1$$

with

$$w_{jl}(x) = M\left(m_j + \frac{h}{2} x\right) \quad \text{and} \quad w_{jr}(x) = M\left(m_{j+1} - \frac{h}{2} x\right) \quad \text{on } [-1,1].$$

Our main result in this section is the following.

**Theorem 3.3.1. (Coercivity)** If on each cell interface $m_{j+\frac{1}{2}}$ for $1 \leq j \leq N - 1$,

$$\beta_0 > 2\Gamma_j M_{j+\frac{1}{2}},$$

(3.18)

where $M_{j+\frac{1}{2}} = M(m_{j+\frac{1}{2}})$, then there exists $\gamma \in (0,1)$ such that

$$A(v, v) \geq \gamma \|v\|^2_E \quad \forall v \in V_h.$$

(3.19)

To prove this result, we need the following lemma.

**Lemma 3.3.2.** For any $v \in V_h$,

$$\left(2\beta_0 v + \beta_1 h[\partial_m^2 v]\right)^2\Big|_{m_{j+\frac{1}{2}}} \leq \frac{4\Gamma_j}{h} \left(\int_{I_j} + \int_{I_{j+1}}\right) M|\partial_m v|^2 \, dm.$$  

(3.20)
Proof. Note that, for any $u_l, u_r \in P^{k-1}([-1, 1])$, we have
\[
(u(1) - 2 \beta_1 \partial_x u(1))^2 \leq \Gamma \int_{-1}^{1} w_j(x)u^2(x) dx \leq 2 \Gamma \left( \int_{-1}^{1} w_{jt}(x)u_l^2(x) dx + \int_{-1}^{1} w_{jr}(x)u_r^2(x) dx \right),
\]
(3.21)
where $u = u_l + u_r$. It is clear that $\Gamma$ depends on $\beta_1$, the weight $w_j$ and the polynomial degree $k$.

This when applied to any $v \in V_h$ with the change of variables
\[
x = \frac{m - m_j}{h/2} \text{ on } I_j \quad \text{and} \quad x = \frac{m_{j+1} - m}{h/2} \text{ on } I_{j+1}
\]
gives (3.20).

Using this key estimate we are able to establish the coercivity of the bilinear operator $A$.

Proof of Theorem 3.3.1.

In fact,
\[
A(v, v) = \sum_{j=1}^{N} \int_{I_j} M|\partial_m v|^2 \, dm + \sum_{j=1}^{N-1} M[v] \left( \frac{\beta_0}{h} [v] + 2 \partial_m v + \beta_1 h [\partial_m^2 v] \right) \bigg|_{m_{j+\frac{1}{2}}} \\
\geq \sum_{j=1}^{N} \int_{I_j} M|\partial_m v|^2 \, dm + \sum_{j=1}^{N-1} M \left[ \frac{\beta_0 - \alpha}{h} [v]^2 - \frac{h}{4\alpha} (2 \partial_m v + \beta_1 h [\partial_m^2 v])^2 \right] \bigg|_{m_{j+\frac{1}{2}}}, \quad (3.22)
\]
where $\alpha$ can be any positive constant.

Note that there exists $\gamma \in (0, 1)$ such that
\[
\beta_0 = \frac{2 \Gamma_j M_{j+\frac{1}{2}}}{1-\gamma}.
\]
Let $\alpha = \frac{2 \Gamma_j M_{j+\frac{1}{2}}}{1-\gamma}$ and it follows that $\beta_0 - \alpha \geq \gamma \beta_0$. Then in virtue of (3.20) we have
\[
A(v, v) \geq \sum_{j=1}^{N} \int_{I_j} M|\partial_m v|^2 \, dm + \gamma \sum_{j=1}^{N-1} \frac{\beta_0}{h} M[v]^2 \bigg|_{m_{j+\frac{1}{2}}} - \frac{1 - \gamma}{2} \sum_{j=1}^{N-1} \left( \int_{I_j} + \int_{I_{j+1}} \right) M|\partial_m v|^2 \, dm \\
\geq \gamma \sum_{j=1}^{N} \int_{I_j} M|\partial_m v|^2 \, dm + \gamma \sum_{j=1}^{N-1} \frac{\beta_0}{h} M[v]^2 \bigg|_{m_{j+\frac{1}{2}}} \\
= \gamma \|v\|^2_{E}.
\]
\[
\square
\]
It is clear that \( \Gamma_j \) plays an important role in the determination of \( \beta_0 \). We proceed to give some refined characterization of \( \Gamma_j \).

**Lemma 3.3.3.** It holds

\[
\Gamma_j = \rho(H^{-\frac{1}{2}}OH^{-\frac{1}{2}}),
\]

(3.23)

where \( \rho(\cdot) \) denotes the spectral radius of a matrix and the two matrices \( O = (O_{\mu\nu}) \) and \( H = (H_{\mu\nu}) \) are defined by

\[
O_{\mu\nu} = [1 - 2\beta_1 (\mu - 1)] [1 - 2\beta_1 (\nu - 1)], \quad 1 \leq \mu, \nu \leq k,
\]

and

\[
H_{\mu\nu} = \int_{-1}^{1} w(x)\phi_{\mu}(x)\phi_{\nu}(x) \, dx = \int_{-1}^{1} w(x)x^{\mu+\nu-2} \, dx, \quad 1 \leq \mu, \nu \leq k.
\]

**Proof.** If the basis \( \{\phi_{\mu}\}_{\mu=1}^{k} \) of \( P^{k-1}([-1, 1]) \) is taken as \( \{x^{\mu-1}\}_{\mu=1}^{k} \), then \( u \) can be represented as \( u(x) = \sum_{\mu=1}^{k} a_{\mu} x^{\mu-1} = \vec{a}^T \phi(x) \). It follows that

\[
(u(1) - 2\beta_1 \partial_x u(1))^2 = \vec{a}^T O \vec{a} \quad \text{and} \quad u^2(x) = \vec{a}^T \phi(x) \phi^T(x) \vec{a}.
\]

Moreover,

\[
\vec{a}^T H \vec{a} = \int_{-1}^{1} w_j(x) u(x)^2 \, dx > 0 \quad \forall \vec{a} \neq 0
\]

implies that \( H \) is both symmetric and positive definite. So is \( H^{\pm \frac{1}{2}} \). Let \( y = H^{\frac{1}{2}} \vec{a} \) and we have

\[
\vec{a}^T O \vec{a} = y^T H^{-\frac{1}{2}} O H^{-\frac{1}{2}} y \leq \rho(H^{-\frac{1}{2}} O H^{-\frac{1}{2}}) \|y\|^2 = \rho(H^{-\frac{1}{2}} O H^{-\frac{1}{2}}) \vec{a}^T H \vec{a}.
\]

This yields (3.23) as desired. \[\square\]

### 3.3.3 Entropy stability

Our DG scheme (3.15) then has the following properties.

**Theorem 3.3.4.** The semidiscrete DDG (3.15) with (3.18) satisfies the following properties:

1. Conservation of mass:

\[
\sum_{j=1}^{N} \int_{I_j} f_h(m, t) \, dm = \int_{B} f_0(m) \, dm \quad \forall t \geq 0.
\]
2. The semidiscrete relative entropy

\[ E(t) = \sum_{j=1}^{N} \int_{I_j} M g_h^2(m, t) \, dm \]

is nonincreasing in time. More precisely, there exists \( \gamma \in (0, 1) \) such that

\[ \frac{d}{dt} E(t) \leq -\gamma \| g_h \|_E^2 \leq 0. \]  
(3.24)

Moreover,

\[ E(t) \leq E(0) \leq \int_B M g_0^2(m) \, dm \quad \forall t \geq 0. \]

**Proof.** (1) Since the numerical flux is single valued at each cell interface, we can take \( v = 1 \) and sum over all cells to obtain the desired property.

(2) Notice that if we take \( v = g_h \), then (3.15) gives that

\[ \frac{d}{dt} E(t) = 2 \int_B M \partial_t g_h g_h \, dm = -A(g_h, g_h), \]

which when using the coercivity estimate in Theorem 3.3.1 gives (3.24).

From (3.24) it follows that

\[ E(t) \leq E(0) = \sum_{j=1}^{N} \int_{I_j} M g_h^2(m, 0) \, dm. \]

Set \( v(m) = g_h(m, 0) \) in (3.13) and we have

\[ \int_{I_j} M g_h^2(m, 0) \, dm = \int_{I_j} M g_0(m) g_h(m, 0) \, dm \leq \left( \int_{I_j} M g_0^2(m) \, dm \right)^{1/2} \left( \int_{I_j} M g_h^2(m, 0) \, dm \right)^{1/2}, \]

which leads to \( \int_{I_j} M g_h^2(m, 0) \, dm \leq \int_{I_j} M g_0^2(m) \, dm \), hence \( E(t) \leq \int_B M g_0^2(m) \, dm. \) \( \square \)

We may also examine the large time behavior of \( g_h \).

**Theorem 3.3.5.** Consider the semidiscrete scheme (3.15). The numerical solution \( f_h(m, t) = M(m)g_h(m, t) \) converges to \( CM(m) \) as \( t \to \infty \), where

\[ C = \frac{\int_B f_0(m) \, dm}{\int_B M(m) \, dm}. \]
(3.25)
Proof. Define a functional \( V[g_h] \) by
\[
V[g_h] = \sum_{j=1}^{N} \int_{I_j} (g_h - C)^2 M \, dm
\]
and a set of functions \( \Sigma \) by
\[
\Sigma = \left\{ g_h \in V_h^N; \sum_{j=1}^{N} \int_{I_j} M g_h \, dm = \int_B f_0(m) \, dm \right\}.
\]
Let \( \bar{g}(x) \equiv C \). We see that \( \bar{g} \in \Sigma \) and \( \frac{d}{dt} \bar{g} = 0 \).

A direct verification shows that
\[
V[g_h] = E(t) - C \int_B f_0(m) \, dm,
\]
which implies that \( \frac{d}{dt} V = \frac{d}{dt} E \). Hence \( V \) satisfies the following:

- \( V[g_h] > 0 \) for any \( g_h \in \Sigma \setminus \{ \bar{g} \} \)(positive definite),
- \( \frac{d}{dt} V[g_h] \leq 0 \) for all \( g_h \in \Sigma \)(negative semidefinite),
- the set \( \{ \frac{d}{dt} V(g_h) = 0 \} \cap \Sigma \) does not contain any trajectories of the ODE system (3.15) besides the trajectory \( g_h(m,t) = \bar{g} \forall t > 0 \).

With these properties we can apply the Krasovskii–LaSalle principle to conclude that
\[
\lim_{t \to \infty} g_h(m,t) = \bar{g},
\]
which leads to the conclusion. It is left to verify the stated properties of \( V \). First two properties of \( V \) are easy to verify. We only verify the third property: from (2) of Theorem 3.3.4 it follows that if \( \frac{d}{dt} E(t) = 0 \), then
\[
\|g_h\|_E^2 = \sum_{j=1}^{N} \int_{I_j} M |\partial_m g_h|^2 \, dm + \sum_{j=1}^{N-1} \frac{\beta_0}{h} M[g_h]^2 \bigg|_{j + \frac{1}{2}} = 0.
\]
From the first summation, we have \( \partial_m g_h = 0 \) on all computational cells. Therefore \( g_h \) must be a constant on each cell. The second summation gives that \( [g_h] = 0 \) on all the cell interfaces. They both ensure that \( g_h = \text{constant} \) on \( B \). While within \( \Sigma \), \( g_h = \bar{g} \) must hold. The proof is thus complete. \( \square \)
3.3.4 Fully discrete DG

Let $g^n(m)$ denote the numerical solution $g_h$ at $t = t_n$ and $\Delta t$ the time step. We take the following fully discrete DG scheme:

$$2 \sum_{j=1}^{N} \int_{I_j} Mg^{n+1}_j - g^n_i \Delta t v dm = -A(g^*, v) \quad \forall v \in V_h. \quad (3.26)$$

Here $g^*$ is defined as

$$g^* = \xi g^{n+1} + (1 - \xi) g^n, \quad \frac{1}{2} \leq \xi \leq 1.$$

Given the cell representative $f^n$, $f^{n+1}$ can be obtained from the transformation $f^{n+1} = Mg^{n+1}$.

If $\xi = 1$, (3.26) is a backward difference DG approximation. If $\xi = \frac{1}{2}$, (3.26) yields the Crank-Nicolson DG approximation.

**Theorem 3.3.6.** The fully discrete scheme (3.26) has a unique solution $\{g^n\}$. Moreover, the solution satisfies the following properties:

1. Conservation of mass: $\sum_{j=1}^{N} \int_{I_j} f^{n+1} dm = \int_B f_0 dm$.

2. The relative entropy

$$E^n = \sum_{j=1}^{N} \int_{I_j} M(g^n)^2 dm$$

satisfies the following inequality:

$$E^{n+1} \leq E^n - \Delta t \gamma\|g^*\|_E^2 - (2\xi - 1) \sum_{j=1}^{N} \int_{I_j} M(g^{n+1} - g^n)^2 dm \leq E^n. \quad (3.27)$$

**Proof.** By showing the uniqueness of the solution, we can obtain the existence of the solution.

Assume there are two solutions to the linear system, $g_1^{n+1}$ and $g_2^{n+1}$. Then we have

$$2 \sum_{j=1}^{N} \int_{I_j} Mg_1^{n+1} v dm - 2 \sum_{j=1}^{N} \int_{I_j} Mg_2^{n+1} v dm = -\Delta t A(\xi(g_1^{n+1} - g_2^{n+1}), v).$$

Define $e = g_1^{n+1} - g_2^{n+1}$ and let $v = e$. So the equation above gives that

$$2 \sum_{j=1}^{N} \int_{I_j} Me^2 dm = -\Delta t \xi A(e, e).$$
The left-hand side of the equation is nonnegative, however the right-hand side is nonpositive. Therefore

\[ 2 \sum_{j=1}^{N} \int_{I_j} M e^2 \, dm = 0 \]

and it follows that \( e = 0 \), which gives the uniqueness of the solution.

(1) Taking \( v = 1 \) in (3.26) gives

\[ \sum_{j=1}^{N} \int_{I_j} f^{n+1} \, dm = \sum_{j=1}^{N} \int_{I_j} f^{n} \, dm. \]

According to the initial projection, we have the mass conservation.

(2) Using \( v = g^* \) and

\[ 2g^* = (g^{n+1} + g^{n}) + (2\xi - 1)(g^{n+1} - g^{n}) \]

in (3.26), we obtain

\[ E^{n+1} - E^{n} = -\Delta t A(g^*, g^*) - (2\xi - 1) \sum_{j=1}^{N} \int_{I_j} M(g^{n+1} - g^{n})^2 \, dm. \] (3.28)

Hence with \( \xi \in [\frac{1}{2}, 1] \) and the coercivity of \( A \), we have

\[ E^{n+1} \leq E^{n}. \]

Moreover, we have

**Theorem 3.3.7.** Let \( f^n \) be the numerical solution to the fully discrete DG scheme (3.26) with \( \frac{1}{2} < \xi \leq 1 \). Then it converges as \( n \to \infty \) with

\[ f^n \to CM, \]

where \( C \) is defined in (3.25).

**Proof.** Since \( E^n \) is nonincreasing and bounded from below, we have

\[ \lim_{n \to \infty} E^n = \inf \{E^n\}. \]
Observe from (3.28) that $E^{n+1} - E^n$ is a sum of nonpositive and bounded terms. When passing the limit $n \to \infty$ we conclude that each term must have zero as its limit, that is,

$$
\lim_{n \to \infty} \sum_{j=1}^{N} \int_{I_j} M |\partial_m g^*|^2 \, dm = 0,
$$

$$
\lim_{n \to \infty} \sum_{j=1}^{N-1} \frac{\beta_0}{h} M |g^*|^2 |_{j+\frac{1}{2}} = 0,
$$

(3.29)

$$
\lim_{n \to \infty} \sum_{j=1}^{N} \int_{I_j} M (g^{n+1} - g^n)^2 \, dm = 0.
$$

The first and third relations in (3.29) tell that \{g^n_h\} has a limit and the limit must be constant in each computational cell. The second relation in (3.29) infers that the limit must be $C$ over the whole domain $B$. The proof is complete. □

### 3.3.5 A positive approximation

Since $g^n_h(m)$ is a high order polynomial approximation of $g(m,t_n)$, which is nonnegative, then $g^n$ should be nearly nonnegative, if it is not nonnegative everywhere over the whole domain. Therefore it is reasonable to construct a nonnegative approximation from the obtained numerical solution at the final time. For example, we simply take

$$
\tilde{g}^n(m) = \begin{cases} 
g^n(m), & \text{if } g^n(m) \geq 0; \\
0, & \text{otherwise.} 
\end{cases}
$$

(3.30)

In fact, $\tilde{g}^n$ does maintain the same accuracy of the numerical solution.

**Theorem 3.3.8.** The nonnegative polynomial $\tilde{g}^n$ defined by (3.30) gives the same order of accuracy as the numerical solution $g^n$ in any $L^p$ norm ($1 \leq p \leq \infty$).

**Proof.** Let $g^n^-$ denote the negative part of $g^n$. Then

$$
|\tilde{g}^n - g(\cdot,t_n)| = |\tilde{g}^n - g^n + g^n - g(\cdot,t_n)|
\leq |\tilde{g}^n - g^n| + |g^n - g(\cdot,t_n)|
\leq |g^n^-| + |g^n - g(\cdot,t_n)|.
$$

Since $g(m,t_n)$ and $g^n^-$ are both nonnegative, we have

$$
0 \leq g^n^-(m) \leq g(m,t_n) + g^n^-(m) = g(m,t_n) - (-g^n^-)(m) = g(m,t_n) - g^n(m).
$$
It follows that
\[ |\tilde{g}^n - g(\cdot, t_n)| \leq 2|g^n - g(\cdot, t_n)| \quad \text{pointwise in each cell.} \]

At interfaces, the numerical solution is taken as the average of values of \( g^n \) from two neighboring cells. With this choice the above estimate remains valid.

Finally we come to a point to show how fully discrete solutions \( g^n \) are represented: if we take the basis function \( \{ \phi_l \}_{l=1}^{N(k+1)} \), then
\[ g^n = \sum_{l=1}^{(k+1)N} a^n_l \phi_l, \]
where \( \{ \phi_l \}_{l=(j-1)(k+1)+1}^{j(k+1)} \) is the basis of \( P^k \) on the cell \( I_j \) and zero in any other cell. Problem (3.26) is then equivalent to a linear system with an unknown vector \( \bar{a}^n = (a^n_1, \cdots, a^n_{N(k+1)})^T \)
\[ (W - \Delta t \xi F)\bar{a}^{n+1} = (W + \Delta t(1 - \xi) F)\bar{a}^n, \]
where
\[ W = 2 \sum_{j=1}^{N} \int_{I_j} M \phi \phi^T \, dm, \quad F = -A(\phi^T, \phi). \]

### 3.4 Extension to the multidimensional FENE Model

In the discretization to follow, we shall focus only on the two-dimensional case, for which \( \kappa \) has the following form,
\[ \kappa = \begin{pmatrix} \kappa_{11} & \kappa_{12} \\ \kappa_{21} & -\kappa_{11} \end{pmatrix} \]
with
\[ \kappa^s = \begin{pmatrix} \kappa_{11} & \frac{\kappa_{12} + \kappa_{21}}{2} \\ \frac{\kappa_{12} + \kappa_{21}}{2} & -\kappa_{11} \end{pmatrix} \]
and
\[ \kappa^o = \begin{pmatrix} 0 & \frac{\kappa_{12} - \kappa_{21}}{2} \\ -\frac{\kappa_{12} - \kappa_{21}}{2} & 0 \end{pmatrix} = a \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \text{ with } a = \frac{\kappa_{12} - \kappa_{21}}{2}. \]
Notice that when \( \kappa \) is not normal, \( M(m) \) is no longer the equilibrium solution. However we still use the functional
\[ E(t) = \int_B Mg^2 \, dm \]
to measure the stability of the method. Our goal is to design a higher order numerical method which satisfies the discrete entropy inequality for $\kappa$ normal and is entropy stable in the sense of $E(t) \leq C(t)E(0)$ for general $\kappa$.

3.4.1 Discretization

The domain $B$ can be represented by $[0, \sqrt{b}] \times [0, 2\pi]$ in the polar coordinate system. Partition $B$ into uniform rectangles

$$K_{i,j} = \{(r, \theta); r \in R_i, \theta \in \Theta_j\}, \quad 1 \leq i \leq P, \quad 1 \leq j \leq Q,$$

where

$$r_{i+\frac{1}{2}} = i \triangle r, \quad \theta_{j+\frac{1}{2}} = j \triangle \theta, \quad R_i = [r_{i-\frac{1}{2}}, r_{i+\frac{1}{2}}], \quad \Theta_j = [\theta_{j-\frac{1}{2}}, \theta_{j+\frac{1}{2}}]$$

with steps of radius and angle

$$\triangle r = \frac{\sqrt{b}}{P}, \quad \triangle \theta = \frac{2\pi}{Q}.$$

Figure 3.1 Diagram of the two-dimensional partition of $B$. 
The DG scheme is as follows: for any function $v \in V_h$, find $g_h \in V_h$ such that
\[
2 \int_{K_{i,j}} M \partial_t g_h v \, dm = - \int_{K_{i,j}} M(\nabla m g_h - 2\kappa_m^a m g_h) \cdot \nabla_m v \, dm - \int_{\partial K_{i,j}} 2M \hat{g}_h v \kappa_m \cdot \hat{v} \, ds
\] (3.31)
\[
+ \int_{\partial K_{i,j}} M \left[ \nabla_m g_h + (g_h - \bar{g}_h) \nabla_m v \right] \cdot \hat{v} \, ds,
\]
where $\hat{v}$ is the outward normal vector on $\partial K_{i,j}$.

We use a weighted $L^2$ projection to prepare the initial data for the scheme: obtain a piecewise polynomial $g_h(m,0) \in V_h$ such that in each cell $K_{i,j}$,
\[
\int_{K_{i,j}} M g_h(m,0) v(m) \, dm = \int_{K_{i,j}} M g_0(m) v(m) \, dm \quad \forall v \in V_h.
\]

Using the polar coordinates in the last term of (3.31), the scheme can be written as
\[
2 \int_{K_{i,j}} M \partial_t g_h v \, dm = - \int_{K_{i,j}} M(\nabla m g_h - 2\kappa_m^a m g_h) \cdot \nabla_m v \, dm - 2 \int_{\partial K_{i,j}} M \hat{g}_h v \kappa_m \cdot \hat{v} \, ds
\] (3.32)
\[
+ \int_{\Theta_j} \frac{M}{r} \left[ \partial_r g_h v + (g_h - \bar{g}_h) \partial_r v \right]_{\theta_{j+\frac{1}{2}}}^{\theta_{j-\frac{1}{2}}} \, d\theta
\]
\[
+ \int_{R_i} \frac{M}{r} \left[ \partial_\theta g_h v + (g_h - \bar{g}_h) \partial_\theta v \right]_{\theta_{j+\frac{1}{2}}}^{\theta_{j-\frac{1}{2}}} \, dr.
\]

As to the numerical flux, we choose
\[
\hat{\partial_r g}_h = \frac{\beta_r}{\Delta r} [g_h] + \partial_r g_h + \beta_1 r \Delta r [\partial_r^2 g_h], \quad \hat{\partial_\theta g}_h = \frac{\beta_\theta}{\Delta \theta} [g_h] + \partial_\theta g_h + \beta_1 \Delta \theta [\partial_\theta^2 g_h].
\] (3.32)

And for $\hat{g}_h$, we have two options.

1. The central flux:
\[
\hat{g}_h = \bar{g}_h.
\] (3.33)

2. The upwinding flux:
\[
\hat{g}_h = \begin{cases} 
  g_h^+ & \text{if } a > 0, \\
  g_h^- & \text{if } a < 0.
\end{cases}
\] (3.34)

At $r_{\frac{1}{2}}$ and $r_{P+\frac{1}{2}}$, we enforce that $g_h^+ = g_h^-$. This ensures the zero boundary contribution and incorporates the zero flux boundary condition. Also it is natural to define the second flux in (3.32) periodically.
Using the notation \( \{ij\} = \{1 \leq i \leq P, 1 \leq j \leq Q\} \), the primal form is given by

\[
2 \sum_{ij} \int_{K_{i,j}} M \partial_t g_h v \, dm + A(g_h, v) = 0 \quad \forall v \in V_h, \tag{3.35}
\]

where

\[
A(g_h, v) = \sum_{ij} \int_{K_{i,j}} M \nabla_m g_h \cdot \nabla_m v \, dm \\
- 2 \sum_{ij} \int_{K_{i,j}} M \kappa^a m g_h \cdot \nabla_m v \, dm + 2 \sum_{ij} \int_{\partial K_{i,j}} M \bar{g}_h \kappa^a m \cdot \bar{v} \, ds \\
+ \sum_{1 \leq i \leq P-1 \atop 1 \leq j \leq Q} \int_{\Theta_j} r M \left( \partial_r g_h[v] + [g_h] \partial_r v \right) \left|_{r_{i+1/2}} \right. \, d\theta \\
+ \sum_{ij} \int_{R_i} M \left( \partial_{\theta} g_h[v] + [g_h] \partial_{\theta} v \right) \left|_{\theta_{j+1/2}} \right. \, dr.
\]

### 3.4.2 Numerical flux and coercivity of \( A \)

The results from the one-dimensional case can be applied to \( r \) and \( \theta \) respectively. We define the discrete energy norms of \( v \in V_h \) in \( r \) and \( \theta \) directions as follows:

\[
\|v(\cdot, \theta)\|_r^2 = \sum_{i=1}^P \int_{R_i} r M |\partial_r v|^2 \, dr + \sum_{i=1}^{P-1} \frac{\beta_{\partial_r}}{\Delta r} r M |v|_{r_{i+1/2}}^2, \\
\|v(r, \cdot)\|_\theta^2 = \sum_{j=1}^Q \int_{\Theta_j} M |\partial_{\theta} v|^2 \, d\theta + \sum_{j=1}^{Q-1} \frac{\beta_{\partial_{\theta}}}{\Delta \theta} M |v|_{\theta_{j+1/2}}^2.
\]

We introduce two quantities, one at the interface \( r_{i+1/2} \) along the \( r \) direction:

\[
\Gamma_r(\beta_{1r}, w_r) := \sup_{u \in P^{k-1}([-1,1]) \atop u \neq 0} \frac{(u(1) - 2\beta_{1r} \partial_x u(1))^2}{\int_{-1}^1 w_r(x) u^2(x) \, dx},
\]

the other one at the interface \( \theta_{j+1/2} \) along the \( \theta \) direction:

\[
\Gamma_\theta(\beta_{1\theta}, w_\theta) := \sup_{u \in P^{k-1}([-1,1]) \atop u \neq 0} \frac{(u(1) - 2\beta_{1\theta} \partial_x u(1))^2}{\int_{-1}^1 w_\theta(x) u^2(x) \, dx}.
\]

Here the weight functions \( w_r \) for \( i = 1, \ldots, P-1 \) are defined by

\[
w_r(x) = \min \left\{ \left( r_i + \frac{\Delta r}{2} x \right) M \left( r_i + \frac{\Delta r}{2} x, \theta \right), \left( r_{i+1} - \frac{\Delta r}{2} x \right) M \left( r_{i+1} - \frac{\Delta r}{2} x, \theta \right) \right\},
\]
and \( w_j \) for \( j = 1, \ldots, Q \) by

\[
w_j(x) = \min \left\{ M \left( r, \theta_j + \frac{\Delta \theta}{2} x \right), M \left( r, \theta_{j+1} - \frac{\Delta \theta}{2} x \right) \right\}.
\]

Now we are able to establish the coercivity of the bilinear operator \( A \).

**Theorem 3.4.1. (Coercivity)** With

\[
\beta_{0r} > 2r_{i+1} \max_{\theta_j} M(r_{i+1}, \theta) \quad \text{and} \quad \beta_{0\theta} > 2 \max_{r_i} \Gamma_{\theta} M(r, \theta_{j+1}),
\]

there exists \( \gamma \in (0, 1) \) such that

\[
A(v, v) \geq \gamma \left( \sum_{j=1}^{Q} \int_{\Theta_j} \|v\|^2 d\theta + \sum_{i=1}^{P} \int_{R_i} \frac{1}{r} \|v\|^2 r dr \right) - \frac{b}{4} \| I \kappa - \kappa I \| \sum_{i,j} \int_{K_{i,j}} M v^2 dm \quad \forall v \in V_h.
\]

(3.37)

In order to prove this theorem, we need the following lemma.

**Lemma 3.4.2.** If \( (\beta_{0r}, \beta_{1r}) \) and \( (\beta_{0\theta}, \beta_{1\theta}) \) satisfy (3.36), then there exists \( \gamma \in (0, 1) \) such that for any \( v \in V_h \),

\[
\sum_{i=1}^{P} \int_{R_i} r M |\partial_r v|^2 dr + \sum_{i=1}^{P-1} r M \left( \frac{\beta_{0r}}{\Delta r} [v] - \frac{\Delta r}{4 \alpha} (2 \partial_r v + \beta_{1r} \Delta r [\partial_r^2 v]) \right) \bigg|_{r \frac{i+1}{2}} \geq \gamma \| g_h \|^2_{r}.
\]

(3.38)

\[
\sum_{j=1}^{Q} \int_{\Theta_j} M |\partial_\theta v|^2 d\theta + \sum_{j=1}^{Q-1} M \left( \frac{\beta_{0\theta}}{\Delta \theta} [v] - \frac{\Delta \theta}{4 \alpha} (2 \partial_\theta v + \beta_{1\theta} \Delta \theta [\partial_\theta^2 v]) \right) \bigg|_{\theta \frac{j+1}{2}} \geq \gamma \| v \|^2_{\theta}.
\]

(3.39)

**Proof.** We only discuss (3.38) along the \( r \) direction, since the argument for (3.39) is analogous.

The definition of \( \Gamma_r \) gives that

\[
(2 \partial_r v + \beta_{1r} \Delta r [\partial_r^2 v]) \bigg|_{r \frac{i+1}{2}} \leq \frac{4 \Gamma_r}{\Delta r} \left( \int_{R_i} r M |\partial_r v|^2 dr + \int_{R_{i+1}} r M |\partial_r v|^2 dr \right).
\]

(3.40)

Using the Young's inequality, we have

\[
\sum_{i=1}^{P-1} r M \left( \frac{\beta_{0r}}{\Delta r} [v] - \frac{\Delta r}{4 \alpha} (2 \partial_r v + \beta_{1r} \Delta r [\partial_r^2 v]) \right) \bigg|_{r \frac{i+1}{2}} \geq \sum_{i=1}^{P-1} r M \left( \frac{\beta_{0r} - \alpha}{\Delta r} [v] - \frac{\Delta r}{4 \alpha} (2 \partial_r v + \beta_{1r} \Delta r [\partial_r^2 v]) \right) \bigg|_{r \frac{i+1}{2}}
\]

\[
\geq \sum_{i=1}^{P-1} r M \left( \frac{\beta_{0r} - \alpha}{\Delta r} [v] \right) \bigg|_{r \frac{i+1}{2}}
\]

\[
= \frac{\Gamma_r r_{i+1} M (r_{i+1}, \theta)}{\alpha} \left( \int_{R_i} r M |\partial_r v|^2 dr + \int_{R_{i+1}} r M |\partial_r v|^2 dr \right).
\]
Note that there exists $\gamma \in (0, 1)$ such that

$$\beta_0 = \frac{2\Gamma r_{i+\frac{1}{2}} M(r_{i+\frac{1}{2}}, \theta)}{(1-\gamma)^2}.$$ 

Then let $\alpha = \frac{2\Gamma r_{i+\frac{1}{2}} M(r_{i+\frac{1}{2}}, \theta)}{1-\gamma}$ and we have

$$\beta_0 - \alpha = \gamma \beta_0 \quad \text{and} \quad \frac{\Gamma r_{i+\frac{1}{2}} M(r_{i+\frac{1}{2}}, \theta)}{\alpha} = \frac{1-\gamma}{2}.$$ 

Therefore

$$\sum_{i=1}^{P} \int_{R_i} r M |\partial_r v|^2 \, dr + \sum_{i=1}^{P-1} r M (\partial_r v[v] + [v] \partial_r v) \bigg|_{r_{i+\frac{1}{2}}} \geq \gamma \|v\|_r^2.$$ 

Now we can complete the proof of the coercivity of $A$.

**Proof of Theorem 3.4.1.**

Let $A(v, v) = I + II + III + IV$ corresponding to each line in the definition of $A$ with $g_h$ replaced by $v$. By the change of variables, we have

$$I = \sum_{i,j} \int_{K_{i,j}} M \left( r |\partial_r v|^2 + \frac{1}{r} |\partial_\theta v|^2 \right) \, dr \, d\theta.$$ 

Combining this with the estimates (3.38) and (3.39), we have

$$A(v, v) \geq \gamma \left( \sum_{j=1}^{Q} \int_{\Theta_j} \|v\|_r^2 \, d\theta + \sum_{i=1}^{P} \int_{R_i} \frac{1}{r} \|v\|_r^2 \, dr \right) + II.$$ 

Next, we only need to analyze II. Using integration by parts, we have

$$II = -2 \sum_{i,j} \int_{K_{i,j}} M \kappa^a m \cdot \nabla_m v \, dm + 2 \sum_{i,j} \int_{\partial K_{i,j}} M \widehat{v} v \kappa^a m \cdot \vec{v} \, ds$$

$$= -\sum_{i,j} \int_{K_{i,j}} M \kappa^a m \cdot \nabla_m v^2 \, dm + 2 \sum_{i,j} \int_{\partial K_{i,j}} M \widehat{v} v \kappa^a m \cdot \vec{v} \, ds$$

$$= \sum_{i,j} \int_{K_{i,j}} v^2 \nabla \cdot (M \kappa^a m) \, dm + \sum_{i,j} \int_{\partial K_{i,j}} M (2\widehat{v} - v) v \kappa^a m \cdot \vec{v} \, ds$$

$$= V + VI.$$
Because $\nabla mM = (2\kappa^* m - \frac{bm}{b - |m|^2})M$, we can infer that
\[
\nabla \cdot (M \kappa^a m) = 2m^T \kappa^* \kappa^a m M = \frac{1}{4} m^T (\kappa^T \kappa - \kappa \kappa^T) mM.
\] (3.41)

Particularly, for $\kappa$ normal, we have $\nabla \cdot (M \kappa^a m) = 0$. Hence
\[
V \geq -\frac{b}{4} \| \kappa^T \kappa - \kappa \kappa^T \| \sum_{ij} \int_{K_{i,j}} M v^2 \, dm.
\]

Also notice that $\kappa^a m \cdot \vec{\nu} = 0$ on $\gamma_1$ and $\gamma_3$ as shown in Figure 3.1. So the boundary term VI in II becomes
\[
VI = \sum_{ij} \int_{\gamma_2} M (2\bar{v} - v^-) v^- \kappa^a m \cdot \vec{\nu} \, ds + \sum_{ij} \int_{\gamma_4} M (2\bar{v} - v^+) v^+ \kappa^a m \cdot \vec{\nu} \, ds.
\]

Moreover, the outward normal vector $\vec{\nu}$ on $\gamma_4$ of $K_{i,j}$ and $\gamma_2$ of $K_{i,j-1}$ point in opposite directions. With this fact and the periodicity in $\theta$, we can shift the line integration on $\gamma_4$ and get
\[
VI = \sum_{ij} \int_{\gamma_2} M (2\bar{v} - v^-) v^- \kappa^a m \cdot \vec{\nu} \, ds - \sum_{ij} \int_{\gamma_2} M (2\bar{v} - v^+) v^+ \kappa^a m \cdot \vec{\nu} \, ds
\]
\[
= -4a \sum_{ij} \int_{R_i} M r (\bar{v} - \bar{v}) [v] \, dr.
\]

For $\bar{v}$, if we take (3.33), then $VI = 0$. If we take (3.34), then we always have
\[
VI = 2|a| \sum_{ij} \int_{R_i} M r [v]^2 \, dr \geq 0.
\]

Hence either option ensures the coercivity of $A$. \qed

For computational purpose, we further estimate $\Gamma_r$ and $\Gamma_\theta$.

**Lemma 3.4.3.** If the basis $\{\phi_\mu\}_{\mu=1}^{k-1}$ of $P^{k-1}([-1, 1])$ is taken as $\{x^{\mu-1}\}_{\mu=1}^k$, then
\[
\Gamma_r(\beta_{1r}, w_r) = \rho(H_r^{-\frac{1}{2}} O_r H_r^{-\frac{1}{2}})
\] (3.42)

and
\[
\Gamma_\theta(\beta_{1\theta}, w_\theta) = \rho(H_\theta^{-\frac{1}{2}} O_\theta H_\theta^{-\frac{1}{2}}),
\] (3.43)

where $\rho$ denotes the largest eigenvalue of the underlying matrix and the four matrices $O_r = (O_{r\mu\nu})$, $H_r = (H_{r\mu\nu})$, $O_\theta = (O_{\theta\mu\nu})$ and $H_\theta = (H_{\theta\mu\nu})$ are defined by
\[
O_{r\mu\nu} = (1 - 2\beta_{1r}(\mu - 1))(1 - 2\beta_{1r}(\nu - 1)), 1 \leq \mu, \nu \leq k,
\]
\[
H_{r_{\mu\nu}} = \int_{-1}^{1} w_r(x)x^{\mu+\nu-2} \, dx, \quad 1 \leq \mu, \nu \leq k,
\]
\[
O_{\theta_{\mu\nu}} = (1 - 2\beta_1\theta(\mu - 1))(1 - 2\beta_1\theta(\nu - 1)), \quad 1 \leq \mu, \nu \leq k,
\]
and
\[
H_{\theta_{\mu\nu}} = \int_{-1}^{1} w_\theta(x)x^{\mu+\nu-2} \, dx, \quad 1 \leq \mu, \nu \leq k.
\]

Proof. We prove the conclusion about \(\Gamma_r\). Represent \(u\) as
\[
u = \sum_{\mu=1}^{k} a_\mu x^{\mu-1} = \vec{a}^T \phi(x). \]
Then
\[
(u(1) - 2\beta_1 \partial_x u(1))^2 = \vec{a}^T O_r \vec{a} \quad \text{and} \quad u^2(x) = \vec{a}^T \phi(x) \phi^T(x) \vec{a}.
\]
Moreover,
\[
\vec{a}^T H_r \vec{a} = \int_{-1}^{1} w_r(x)u(x)^2 \, dx > 0 \quad \forall \vec{a} \neq 0
\]
implying that \(H_r\) is both symmetric and positive definite. So is \(H_r^{1/2}\). Let \(y = H_r^{1/2} \vec{a}\) and we have
\[
\vec{a}^T O_r \vec{a} = y^T H_r^{-\frac{1}{2}} O_r H_r^{-\frac{1}{2}} y \leq \rho(H_r^{-\frac{1}{2}} O_r H_r^{-\frac{1}{2}}) \|y\|^2 = \rho(H_r^{-\frac{1}{2}} O_r H_r^{-\frac{1}{2}}) \vec{a}^T H_r \vec{a}.
\]
The conclusion about \(\Gamma_{\theta}\) can be proved in a similar manner. \(\square\)

3.4.3 Entropy stability

Our DG scheme (3.35) then has the following properties.

Theorem 3.4.4. Consider the semidiscrete DG scheme (3.35) with (3.36). It satisfies the following properties:

1. Conservation of mass:
\[
\sum_{ij} \int_{K_{i,j}} f_h(m, t) \, dm = \int_B f_0(m) \, dm \quad \forall t \geq 0.
\]
2. For \(\kappa\) normal, the semidiscrete relative entropy
\[
E(t) = \sum_{ij} \int_{K_{i,j}} M g_h^2 \, dm
\]
is nonincreasing in time. More precisely, there exists \(\gamma \in (0, 1)\) such that
\[
\frac{d}{dt} E(t) \leq -\gamma \left( \|g_h\|_\theta^2 + \|g_h\|_\theta^2 \right) \leq 0.
\]

(3.44)
3. The scheme is entropy stable in the sense that
\[ E(t) \leq \exp\left(\frac{bT}{4}\|\kappa^T\kappa - \kappa\kappa^T\|\right) E(0) \quad \text{for } t \in [0, T], \]

Moreover, \( E(0) \leq \int_B M g_0^2(m) \, dm \).

Proof. (1) Taking the test function \( v = 1 \) both in the scheme (3.35) and the initial projection, we have the mass conservation.

(2) Let \( v = g_h \) and then apply the coercivity of \( A \). We are able to prove that \( E(t) \) is nonincreasing in time.

(3) For general \( \kappa \), the estimate (3.37) gives
\[ \frac{d}{dt} E(t) \leq -\gamma (\|g_h\|_r^2 + \|g_h\|_\theta^2) + \frac{b}{4} \|\kappa^T\kappa - \kappa\kappa^T\| E(t), \]
which upon integration in time yields the desired stability estimate. As in the one-dimensional case, the weighted \( L^2 \) projection ensures that \( E(0) \leq \int_B M g_0^2(m) \, dm \).

3.5 Numerical tests

In this section, we provide numerical results to demonstrate (i) the accuracy of the schemes and (ii) the capacity to capture solution features for large times.

We denote the initial function without normalization by \( \tilde{f}_0(m) \). And the numerical results are obtained after normalizing \( \tilde{f}_0 \) and \( M \) so that the mass of \( f \) and the relative entropy of the equilibrium solution will be 1.

3.5.1 One-dimensional tests

Denote the numerical solution at \( t = t_n \) by \( f^n(m) = M(m) g^n(m) \) and the exact solution by \( f(m, t_n) \). When the exact solution is not available, we replace \( f(m, t_n) \) by a reference solution \( f^n_{ref}(m) \) obtained from \( N_{ref} \) computational cells. Here \( N_{ref} \) is much larger than \( N \).

We compute the \( L^\infty \) and \( L^2 \) errors in the following ways.

1. \( L^\infty \) error is given by
\[ \max_{m_c \in S} |f^n(m_c) - f^n_{ref}(m_c)|, \] where \( S \) is the set of the cell centers for \( f^n_{ref} \).
2. $L^2$ error is given by
\[
\sum_{I_{ref}} \| f^n(\cdot) - f^n_{ref}(\cdot) \|_{L^2(I_{ref})}, \quad \text{where } I_{ref} \text{ runs over all the reference cells.}
\]

**Example 1.** (Accuracy) We illustrate the accuracy of the DG scheme (3.26) with $\xi = 1$ subject to the following initial data
\[
\tilde{f}_0(m) = \left(1 - \frac{m^2}{b}\right)^{\frac{3}{4}b}.
\]

We take the numerical solution with $N = 640$ as the reference solution. Table 3.1 shows both error and the order of accuracy of the $P^k$ elements. We observe that the numerical scheme is able to achieve the optimal order up to degree $k = 3$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$(\beta_0, \beta_1)$</th>
<th>$N = 5$</th>
<th>$N = 10$</th>
<th>$N = 20$</th>
<th>$N = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$L^\infty$</td>
<td>error</td>
<td>error</td>
<td>order</td>
<td>error</td>
</tr>
<tr>
<td>0</td>
<td>(1,0)</td>
<td>9.102e-02</td>
<td>6.593e-02</td>
<td>0.465</td>
<td>3.161e-02</td>
</tr>
<tr>
<td></td>
<td>$L^2$</td>
<td>6.797e-02</td>
<td>3.621e-02</td>
<td>0.908</td>
<td>1.854e-02</td>
</tr>
<tr>
<td>1</td>
<td>(3,0)</td>
<td>3.246e-02</td>
<td>9.783e-03</td>
<td>1.730</td>
<td>2.504e-03</td>
</tr>
<tr>
<td></td>
<td>$L^\infty$</td>
<td>1.524e-02</td>
<td>4.296e-03</td>
<td>1.827</td>
<td>1.102e-03</td>
</tr>
<tr>
<td></td>
<td>$L^2$</td>
<td>1.938e-03</td>
<td>3.962e-04</td>
<td>2.324</td>
<td>4.984e-05</td>
</tr>
<tr>
<td>2</td>
<td>(3, 0.1)</td>
<td>2.480e-03</td>
<td>5.661e-04</td>
<td>2.131</td>
<td>6.945e-05</td>
</tr>
<tr>
<td></td>
<td>$L^\infty$</td>
<td>1.938e-03</td>
<td>3.962e-04</td>
<td>2.324</td>
<td>4.984e-05</td>
</tr>
<tr>
<td></td>
<td>$L^2$</td>
<td>3.944e-04</td>
<td>2.076e-05</td>
<td>4.248</td>
<td>1.056e-06</td>
</tr>
</tbody>
</table>

**Example 2.** (Entropy satisfying property) In this example, we take the initial data:
\[
\tilde{f}_0(m) = \begin{cases} 
\cos \left(3\pi \frac{m}{\sqrt{b-\eta}}\right) + 1, & \text{if } |m| \leq \sqrt{b-\eta}, \\
0, & \text{if } |m| > \sqrt{b-\eta},
\end{cases}
\]
where $\eta$ is a constant in $(0, \sqrt{b})$. The use of $\eta$ is to ensure that $g_0 = \frac{\tilde{f}_0}{\mu}$ has a compact support. Otherwise it could induce a large error to the initial projection. For $\eta = \frac{\sqrt{b}}{3}$, Table 3.2 shows that the numerical solutions preserve the mass and get very close to the discrete steady state before $t = 800$ and Figure 3.2 illustrates how the numerical solution approaches the steady state as time evolves. Here we have taken $(\beta_0, \beta_1) = (3.5, 0.5)$ and $\Delta t = 0.1$. 
Table 3.2  The mass and the relative entropy of the $P^3$ approximation on a uniform mesh $N = 20. b = 36.$

<table>
<thead>
<tr>
<th>$t$</th>
<th>0</th>
<th>1</th>
<th>10</th>
<th>50</th>
<th>100</th>
<th>800</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Entropy</td>
<td>3029.250</td>
<td>2447.520</td>
<td>561.094</td>
<td>18.948</td>
<td>3.193</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Figure 3.2  $b = 36, \Delta t = 0.1.$

3.5.2 Two-dimensional tests

For two-dimensional problems, we will demonstrate the entropy satisfying property of the fully implicit time discretization of (3.35) with fluxes (3.32) and (3.33). Since $(\beta_0, \beta_1)$ depend on both $\kappa$ and $b$, we will give the choice of these pairs for each example to follow. All these choices satisfy either the sufficient condition (3.36) or something weaker than (3.36).

**Example 3.** (Equilibrium preserving) In this two dimensional test, we consider $\kappa$ antisymmetric, hence normal, and take the equilibrium $M(m)$ as the initial data. Table 3.3 shows that the third order DG method has the capacity of preserving the equilibrium in time. Here $(\beta_{0r}, \beta_{1r}) = (\beta_{0g}, \beta_{1g}) = (6, 0.5)$ and $\Delta t = 0.1$.

**Example 4.** (Entropy satisfying property) We test the entropy satisfying property of the scheme by two fluid velocities where $\kappa$ is normal. The first test is on a vortex with $\kappa$ being
Table 3.3 The mass and the relative entropy of the $P^2$ approximation on a uniform mesh $P = Q = 10$. $b = 100, \kappa_{11} = 0, \kappa_{12} = -\kappa_{21} = 0.5$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>0</th>
<th>1</th>
<th>10</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Entropy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

antisymmetric and therefore normal. Figure 3.3 shows that the numerical solution approaches the equilibrium $f_{eq}(m)$ as time evolves. In this test, $(\beta_0, \beta_1) = (\beta_0, \beta_1) = (15, 0.35)$ and $\Delta t = 0.1$. Indeed, we can observe this entropy satisfying property quantitatively from Table 3.4.

![Figure 3.3](image)

Figure 3.3 $\tilde{f}_0(r, \theta) = \left(1 - \frac{r^2}{b}\right)^{\frac{10}{3} e^{-\frac{r^2}{b}(0.2 \cos(2\theta) - 0.7 \sin(2\theta))}}, b = 100, \kappa_{11} = 0, \kappa_{12} = -\kappa_{21} = 0.5.$

Table 3.4 The mass and the relative entropy of the $P^2$ approximation on a uniform mesh $P = Q = 10$. $b = 100, \kappa_{11} = 0, \kappa_{12} = -\kappa_{21} = 0.5$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>100</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Entropy</td>
<td>686.634</td>
<td>2.135</td>
<td>1.098</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

The second test is on a simple extensional flow with the velocity field

\[ \vec{v} = (\alpha x, -\alpha y), \]
where $\alpha$ denotes the extensional rate. Then the velocity gradient tensor is

$$\kappa = \nabla \vec{v} = \begin{pmatrix} \alpha & 0 \\ 0 & -\alpha \end{pmatrix}.$$ 

Figure 3.4 shows that the numerical solution approaches the equilibrium solution $f_{eq}(m)$ as time evolves. The initial data is taken as

$$\tilde{f}_0(r, \theta) = \begin{cases} \cos \left( 3\pi \frac{r}{\sqrt{b} - \eta} \right) + 1.0, & \text{if } r \leq \sqrt{b} - \eta, \\ 0, & \text{elsewhere}. \end{cases}$$

While the entropy satisfying property is given quantitatively in Table 3.5 from the same initial data. Here we have taken $(\beta_0, \beta_1) = (\beta_{0 \theta}, \beta_{1 \theta}) = (15, 0.3)$ and $\Delta t = 0.1$.

Table 3.5 The mass and the relative entropy of the $P^2$ approximation on a uniform mesh $P = Q = 10$. $b = 100$, $\alpha = 0.3$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>0</th>
<th>1</th>
<th>10</th>
<th>100</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mass</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Entropy</td>
<td>1180.36</td>
<td>2.53658</td>
<td>1.001</td>
<td>1.001</td>
<td>1.001</td>
</tr>
</tbody>
</table>

**Example 5.** (The simple shear flow) In this example, we test the short time numerical performance for a simple shear flow, where the gradient of the velocity is

$$\kappa = \nabla \vec{v} = \begin{pmatrix} 0 & \gamma \\ 0 & 0 \end{pmatrix}.$$ 

Here we show the numerical convergence by taking $b = 10$. See Table 3.6 and the results are obtained with $(\beta_{0r}, \beta_{1r}) = (\beta_{0 \theta}, \beta_{1 \theta}) = (6, 0)$ and $\Delta t = 0.1$.

Finally we point out that if $b$ becomes larger, $M$ tends to be flatter near the boundary, making it more difficult to evaluate $g = \frac{f}{M}$. Nevertheless, we present a numerical comparison in terms of shear rates for $b = 100$. Figure 3.5 shows that when the shear rate $\gamma$ gets larger, the two concentrations of $f$ are stretched apart further. Here we have taken $k = 1$, $(\beta_{0r}, \beta_{1r}) = (\beta_{0 \theta}, \beta_{1 \theta}) = (6, 0)$ and $\Delta t = 0.01$. 
Table 3.6 \( L^\infty \) and \( L^2 \) error and order of accuracy of the \( P^k \) approximation for Example 5 on a uniform mesh of \( N \) cells. \( \tilde{f}_0(r, \theta) = \left(1 - \frac{r^2}{b}\right)^{\frac{5}{4}} b, \ b = 10, \ \Delta t = 0.001, \ \text{final time} \ t = 0.4. \)

<table>
<thead>
<tr>
<th>( k )</th>
<th>( (\beta_0, \beta_1) )</th>
<th>( N = 5 )</th>
<th>( N = 10 )</th>
<th>( N = 20 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 )</td>
<td>( (1,0) )</td>
<td>( L^\infty )</td>
<td>4.610e-01</td>
<td>2.069e-01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( L^2 )</td>
<td>1.015e-01</td>
<td>5.014e-02</td>
</tr>
<tr>
<td>( 1 )</td>
<td>( (3,0) )</td>
<td>( L^\infty )</td>
<td>1.153e-01</td>
<td>2.751e-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( L^2 )</td>
<td>2.702e-02</td>
<td>7.668e-03</td>
</tr>
<tr>
<td>( 2 )</td>
<td>( (8, 0.1) )</td>
<td>( L^\infty )</td>
<td>2.607e-02</td>
<td>6.681e-03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( L^2 )</td>
<td>5.105e-03</td>
<td>6.373e-04</td>
</tr>
</tbody>
</table>

3.6 Concluding remarks

We have investigated the Fokker–Planck equation which is of bead-spring type FENE dumbbell model for polymers, with our focus on the development of the entropy satisfying discontinuous Galerkin (ESDG) method for the model subject to the zero flux boundary condition. We constructed simple a and easy-to-implement ESDG schemes which preserve equilibrium solutions. Both semidiscrete and fully discrete methods are proven rigorously to satisfy the desired properties: mass conservation and entropy satisfying in the sense that these schemes satisfy discrete entropy inequalities for the quadratic entropy. We also proved the convergence of numerical solutions to the equilibrium solution as time becomes large. Numerical examples are given to illustrate the accuracy and capability of the methods. The main advantage of using the DDG type of numerical fluxes is that there is room for choosing proper paris of \((\beta_0, \beta_1)\) so that the desired properties such as the entropy satisfying property and mass conservation are still preserved for DG methods of an arbitrary order.

A positive numerical approximation of the same accuracy as the numerical solution is obtained through a reconstruction at the final time without destroying the solution accuracy. Further investigation of positivity preserving methods for this problem is more involved and left for future work.
Acknowledgments

This research was supported by the National Science Foundation under grant DMS09-07963.
Figure 3.4 $b = 100, \alpha = 0.3$.

Figure 3.5 $\tilde{f}_0(r, \theta) = \left(1 - \frac{r^2}{b}\right)^{\frac{3b}{4}} e^{\frac{r^2}{b}(0.5 \cos(2\theta) - 0.7 \sin(2\theta))}$, $b = 100$, final time $t = 0.4$. 
CHAPTER 4. MAXIMUM-PRINCIPLE-SATISFYING THIRD-ORDER DISCONTINUOUS GALERKIN SCHEMES FOR FOKKER–PLANCK EQUATIONS

Hailiang Liu and Hui Yu

Abstract

We design and analyze up to third order accurate discontinuous Galerkin (DG) methods satisfying a strict maximum principle for Fokker–Planck equations. A procedure is established to identify an effective test set in each computational cell to ensure the desired positivity of numerical averages during time evolution. This is achievable mainly by using the two parameters in the numerical flux and a novel decomposition of weighted cell averages. Based on this result, a scaling limiter for the DG method with first order Euler forward time discretization is proposed to solve the one-dimensional Fokker–Planck equations. Strong stability preserving high order time discretizations will keep the positivity principle. It is straightforward to extend the method to two and higher dimensions on rectangular meshes. We also show that a modified limiter can preserve the strict maximum principle for DG schemes solving Fokker–Planck equations. As a consequence, the present scheme preserves steady states and provides a satisfying long-time behavior. Numerical tests for the DG method are reported, with applications to polymer models with both Hookean and FENE potentials.

4.1 Introduction

In this paper we are interested in constructing high order accurate schemes for solving Fokker–Planck equations. For a given potential $U : \mathbb{R}^d \to \mathbb{R}$, the probability density function
pdf solves the following initial value problem

\[
\begin{cases}
\partial_t f = \nabla_x \cdot (\nabla_x f + \nabla_x U f), & x \in B, \ t > 0, \\
\partial_\nu f + \partial_\nu U f|_{\partial B} = 0, & t > 0, \\
f(0, x) = f_0(x), & x \in B,
\end{cases}
\]

(4.1)

where \(x\) is the configuration variable in a bounded domain \(B\) satisfying the zero-flux boundary condition and \(\nu\) is the outward normal vector on the boundary \(\partial B\). There are many interpretations and derivations from Biology and other application areas that motivate the Fokker–Planck equation [Perthame (2007)]. For instance, for active motions with the velocity \(u = -\nabla U\) additionally to the Brownian motion, the equation is also called the drift-diffusion equation; in connection to the stochastic differential equations, it is called the Kolmogorov equation; it is a fundamental model in Chemistry at the molecular level.

The main properties of the solution to (4.1) are the nonnegativity principle, the mass conservation and the existence of nonzero steady state, i.e.,

\[
f_0 \geq 0 \implies f \geq 0 \quad \forall t > 0,
\]

\[
\int_B f(t, x) \, dx = \int_B f_0(x) \, dx \quad \forall t > 0,
\]

\[
Ce^{-U} \text{ are steady states for some } C \in \mathbb{R}.
\]

Both property (4.2) and (4.4) are implied by the strict maximum principle, i.e., if

\[
c_1 = \min \left( f_0 e^U \right), c_2 = \max \left( f_0 e^U \right),
\]

then \(f(t, x)e^U \in [c_1, c_2]\) for any \(x \in B\) and \(t > 0\). These properties are also naturally desired for numerical schemes solving (4.1). In this paper, we develop such a method.

This paper is also the continuation of our project, initiated in Liu and Yu (2012a) and followed by Liu and Yu (2012b), of developing high order entropy satisfying numerical methods for the FENE dumbbell model of polymers. The microscopic FENE model under homogeneous fluids and proper rescaling may be described by

\[
\begin{cases}
\partial_t f = \nabla_x \cdot \left[ \nabla_x f + \left( \frac{bx^2}{b - |x|^2} - 2\kappa x \right) f \right], & x \in B, \ t > 0, \\
\left[ \nabla_x f + \left( \frac{bx^2}{b - |x|^2} - 2\kappa x \right) f \right] \cdot \nu = 0, & x \in \partial B, t > 0, \\
f(0, x) = f_0(x), & x \in B,
\end{cases}
\]

(4.6)
where $b$ denotes the maximum spring extension, the $d$-dimensional connector vector $x$ lies in a ball $B = B(0, \sqrt{b})$ and $\kappa$ is the velocity gradient matrix with $\text{Tr}(\kappa) = 0$. The main difficulty in solving (4.6) is that the equation is singular at the boundary [Liu and Liu (2008), Liu and Shin (2012a)], which presents numerous challenges, both analytically and numerically. A general discussion of this problem and background references are given in the introduction of Liu and Yu (2012b).

If $\kappa$ is symmetric, (4.6) can be written as (4.1) with the modified FENE spring potential [Warner (1972)],

$$U(x) = -\frac{b}{2} \log(1 - \frac{|x|^2}{b}) - x^T \kappa x.$$  \hfill (4.7)

In the limit of $b \to \infty$, it reduces to

$$U(x) = \frac{|x|^2}{2} - x^T \kappa x,$$  \hfill (4.8)

which corresponds to the well-known Hookean potential.

It is very difficult to obtain a high order accurate scheme satisfying a strict maximum principle in the sense that the numerical solution never goes out of the range $[c_1, c_2]e^{-U}$ for (4.1). Our approach, following Liu and Yu (2012b), is to explore the following reformulation in terms of $g = f/M$ with the equilibrium $M = e^{-U}$,

$$M \partial_t g = \nabla_x \cdot (M \nabla_x g),$$  \hfill (4.9)

for which the maximum principle (4.5) reduces to

$$c_1 \leq g_0 \leq c_2 \implies c_1 \leq g(t, x) \leq c_2 \quad \forall t > 0,$$  \hfill (4.10)

while the mass conservation needs to be measured by $\int_B M(x) g(t, x) \, dx = \text{constant}$. An entropy satisfying DG method for the Fokker–Planck equation (4.6) was developed in Liu and Yu (2012b) using the relative entropy based on (4.9). The developed method ensures that the steady state is preserved at the discrete level. The focus of this paper is to obtain a discrete version of (4.10) so that the resulting DG scheme will also preserve the steady state and provide a satisfying long-time behavior.

Successful high order numerical schemes for solving (4.1) include finite volume schemes in Bessemoulin-Chatard and Filbet (2012) and Liu and Yu (2012a), entropy satisfying DG
methods in Liu and Yu (2012b), spectral Galerkin methods in Chauvière and Lozinski (2003, 2004a,b), Knezevic and Süli (2009) and Shen and Yu (2012) for the Fokker–Planck equation of FENE model. Although these schemes are nonlinearly stable in numerical experiments and some of them can be proved to be entropy stable, they do not in general satisfy a strict maximum principle or the schemes are at most second order.

Recently a maximum-principle-satisfying framework has been established for scalar conservation laws in Zhang and Shu (2010). The main idea in their work is to find a sufficient condition to preserve the positivity of the cell averages by repeated convex combinations, namely,

1. Use first order schemes which can keep the positivity as building blocks. In the high order spatial discretization with forward Euler, cell averages at the next time step can be written as a convex combination of formal first order schemes, thus will keep the positivity of the cell average provided a certain sufficient condition is satisfied.

2. A simple scaling limiter can enforce the sufficient condition without destroying accuracy and conservation.

3. Use strong stability preserving (SSP) [Gottlieb et al. (2001)] high order time discretizations which are convex combinations of forward Euler. Then it suffices to find a way to preserve the positivity for the forward Euler time discretization.

Thus DG methods with this simple scaling limiter satisfy the maximum principle for scalar conservation laws. Unfortunately, Step 1 can not be applied to second order PDEs such as (4.9) in a straightforward manner. In spite of this difficulty, a non-conventional technique was introduced in Zhang et al. (2012) to design a high order maximum-principle-satisfying finite volume scheme for convection-diffusion equations. Yet, as pointed out in Zhang et al. (2013), it is not obvious how to generalize this non-conventional technique to DG methods. The maximum-principle-satisfying DG scheme on triangular meshes was subsequently proposed in Zhang et al. (2013), but the scheme is only second order accurate ($k = 1$).

In this paper, we develop up to third order accurate maximum-principle-satisfying schemes for one- and multidimensional Fokker–Planck equations, in the sense that the numerical solution never goes out of the range $[c_1, c_2]e^{-U}$ of the initial condition. Our scheme uses the simple Euler
forward, allowing for easy and practical implementation and easy generalization from one- to multidimensions. The scaling limiter introduced in Zhang and Shu (2010) is modified based on the weighted cell averages to control the maximum/minimum of the reconstruction polynomials.

The major difficulty to construct a maximum-principle-satisfying scheme is to maintain the property that the weighted cell average remains in \([c_1, c_2]e^{-U}\) during the time evolution, without destroying accuracy. The novelty of this work is to establish a procedure with some special spatial discretization of (4.9) to satisfy the maximum principle.

Let us illustrate the idea using the simplest one-dimensional equation \(M \partial_t g = \partial_x (M \partial_x g)\).

Integrate this equation on \(I_j\) to obtain

\[
\frac{d}{dt} \langle g \rangle_j := \frac{d}{dt} \frac{1}{h} \int_{I_j} M g(t, x) \, dx = \frac{1}{h} \left[ (M \partial_x g)(t, x_{j+\frac{1}{2}}) - (M \partial_x g)(t, x_{j-\frac{1}{2}}) \right],
\]

where \(h = |I_j| = |[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]|.\) So a conservative finite volume scheme with Euler forward time discretization has the form

\[
\langle g^{n+1} \rangle_j = \langle g^n \rangle_j - \Delta t \left( M_{j+\frac{1}{2}} \widehat{\partial_x g}^{n}_{j+\frac{1}{2}} - M_{j-\frac{1}{2}} \widehat{\partial_x g}^{n}_{j-\frac{1}{2}} \right), \quad (4.11)
\]

where \(\widehat{\partial_x g}_{j+\frac{1}{2}}\) is an approximation to \(\partial_x g\) at \(x_{j+\frac{1}{2}}\). The monotonicity with respect to selected point values seems to be achievable only for first order approximations; see Section 4.2.1. In this work we are able to achieve the third order accuracy while still preserving the positivity mainly because we use special numerical fluxes which, in one-dimensional case, are of the form

\[
\widehat{\partial_x g} = \frac{\beta_0}{h} [g] + \{\partial_x g\} + \beta_1 h [\partial_x^2 g], \quad (4.12)
\]

where \([\cdot]\) denotes the jump of \(g\), and \(\{\cdot\}\) the average of \(g\) crossing the interface. This flux formulation is inspired by the numerical flux introduced in Liu and Yan (2009, 2010) for solving the diffusion problem by the direct discontinuous Galerkin method, which is shown \(L^2\) stable for some choices of parameters \((\beta_0, \beta_1)\).

In the present work parameters \((\beta_0, \beta_1)\) are essentially used to guarantee the existence of some controlled points in each cell so that the numerical solutions preserve the mass and positivity after time evolution and a reconstruction using the same controlled point values.

More precisely, our procedure includes two crucial ingredients:
1. Decompose the weighted cell average of polynomials of degree $k$ in terms of $k+1$ controlled points in each cell with positive coefficients. These points form a test set $S_j$ in each cell $I_j$, over which the decomposition appears as

$$\langle \phi \rangle_j := \frac{1}{h} \int_{I_j} M(x) \phi(x) \, dx = \sum_{x^i \in S_j} \hat{\omega}^i \phi(x^i) \quad \forall \phi(x) \in P^k(I_j).$$

2. Represent the numerical flux (4.12) in terms of solution values at the same controlled points from two neighboring cells in the following way

$$h \hat{\partial}_x \phi \big|_{j + \frac{1}{2}} = \sum_{x^i \in S_{j+1}} \alpha^i_+ \phi(x^i) - \sum_{x^i \in S_j} \alpha^i_- \phi(x^i) \quad \forall \phi(x) \in P^k(I_j).$$

The coefficients $\hat{\omega}^i$, $\alpha^i_+$ and $\alpha^i_-$ depend only on $M(x)$ and $k$ and can be made positive through choices of parameters $(\beta_0, \beta_1)$ for $k \leq 2$.

The above procedure when inserted into (4.11) enables us to show that, under a suitable CFL condition, the simple Euler forward will keep the property $\langle g^n \rangle_j \in [c_1, c_2]$ and the validity of the maximum principle if we use the DG polynomials, thus maintaining uniform $k + 1$ order accuracy. The proposed limiter based on the weighted cell average replaces the definition of maximum and minimum in each cell by those on a test set $S_j$ of $k + 1$ points, so we can easily implement it for polynomials of degree $k$.

The main conclusion of this paper is as follows: by applying the limiter or the simplified version which avoids the evaluation of extrema of polynomials, to a DG scheme solving one- or multidimensional Fokker–Planck equations, with the time evolution by a SSP Runge-Kutta method, we obtain a third order accurate scheme solving (4.1) with the strict maximum principle in the sense that the numerical solution never goes out of the range $[c_1, c_2]e^{-U}$, where the $c_1$ and $c_2$ are defined in (4.5).

The paper is organized as follows: we first describe our DG scheme to solve the nonlogarithmic Landau formulation (4.9) in Section 4.2. In Section 4.3, we prove the maximum principle for up to third schemes in one space dimension. In Section 4.4, we provide a straightforward extension to two space dimensions on rectangular meshes for the third order scheme. Section 4.5 contains an implementation algorithm using the linear scaling limiter. In Section 4.6, numerical tests for the DG method will be reported, including examples from the heat equation,
the Hookean and FENE model in dumbbell models. Concluding remarks are given in Section 4.7.

4.2 DG formulation

Following Liu and Yu (2012a), we reformulate the Fokker–Planck equation (4.1) by finding \(f = Mg\) with \(g\) satisfying

\[
M \partial_t g = \nabla_x \cdot (M \nabla_x g), \quad x \in B, \ t > 0,
\]

\[
M \partial_n g|_{\partial B} = 0, \quad t > 0,
\]

\[
g(0, x) = \frac{f_0(x)}{M(x)}, \quad x \in B,
\]

where \(M(x) = e^{-U(x)}\) is an equilibrium solution. Let \(B\) be partitioned into nonoverlapping uniform rectangular cells \(I_\alpha\) with \(\bigcup I_\alpha = B \subset \mathbb{R}^d\), where \(\alpha\) is the \(d\)-dimensional index. We define the finite element space \(V^k_h\) as

\[
V^k_h = \{v \in L^2(B) : v|_{I_\alpha} \in P^k(I_\alpha), \ \cup I_\alpha = B\},
\]

where \(P^k(I_\alpha)\) denotes the space of polynomials of degree up to \(k\) on \(I_\alpha\).

The semidiscrete DG scheme is to find \(g_h \in V^k_h\) such that for any \(I_\alpha \in B\) and \(v \in V^k_h\),

\[
\int_{I_\alpha} M \partial_t g_h v \, dx = -\int_{I_\alpha} M \nabla_x g_h \cdot \nabla_x v \, dx + \int_{\partial I_\alpha} M \left( \hat{\partial}_n g_h v + (g_h - \{g_h\}) \partial_n v \right) \, ds,
\]

where \(\nu\) is the outward normal direction on the boundary \(\partial I_\alpha\) and \(\{g_h\}\) denotes the average of the trace of numerical solutions on the interface from within the cell and the neighboring cell. The ‘hat’ term is the numerical flux to be chosen and the boundary contributions are taken to be zero to incorporate the zero flux condition. The initial data is generated by the weighted \(L^2\) projection

\[
\int_{I_\alpha} M(g_h(0, x) - g(0, x)) v \, dx = 0 \quad \forall v \in V^k_h \text{ and } I_\alpha \in B.
\]

The numerical solution to the original Fokker–Planck equation is then obtained by \(f_h = Mg_h\).

Crucial for the \(L^\infty\) stability as well as for the accuracy of the DG method is the choice of the numerical flux \(\hat{\partial}_n g_h\) defined at the cell interfaces \(\partial I_\alpha\). The guiding principle is that the flux is chosen in such a way that it depends only on the neighboring polynomials and that
it (i) is consistent with $\partial_\nu g$ when $g$ is smooth, (ii) is conservative in the sense that the flux is single valued on $\partial_\alpha I$, (iii) ensures the positivity preserving property and (iv) enforces the high order accuracy of the method. Numerical fluxes will be given below for both one- and two-dimensional problems.

### 4.2.1 One-dimensional case

We begin with the one-dimensional case for $B = [0, 1]$, partitioned by $\bigcup_{j=1}^{N} I_j$ with $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ and a uniform mesh size $h = \frac{1}{N}$ where

$$x_{j+\frac{1}{2}} = jh$$

and the center of the cell is $x_j = \left(j - \frac{1}{2}\right)h$, $1 \leq j \leq N$.

The direct DG method is as follows: find $g_h \in V^k$ such that for any $v \in V^k$ and $I_j$,

$$\int_{I_j} M \partial_t g_h v \, dx = - \int_{I_j} M \partial_x g_h \partial_x v \, dx + M \left( \hat{\partial_x g_h} v + (g_h - \{g_h\}) \partial_x v \right) \bigg|_{x_j+\frac{1}{2}}^{x_{j+\frac{1}{2}}}.$$  \hspace{1cm} (4.16)

The boundary contribution at $j = \frac{1}{2}$ and $j = N + \frac{1}{2}$ are taken to be zero to incorporate the zero flux condition. We denote by $g^+$ and $g^-$ the value of $g$ at a cell interface from the right cell and from the left cell, respectively. The jump of these two values, $g^+ - g^-$, is denoted by $[g]$. The numerical flux is chosen as follows

$$\hat{\partial_x g_h} = \frac{\beta_0}{h}[g_h] + \{\partial_x g_h\} + \beta_1 h[\partial_x^2 g_h].$$  \hspace{1cm} (4.17)

The form (4.17) makes the numerical flux adopted in (4.16) both consistent and conservative. The algorithm is well defined once the parameters $(\beta_0, \beta_1)$ are chosen.

### 4.2.2 Two-dimensional case

We now formulate a DDG method for multidimensional problems. Here we present schemes for only two-dimensional case with $B = [0, 1]^2$. The two-dimensional equation becomes

$$M \partial_t g = \partial_x (M \partial_x g) + \partial_y (M \partial_y g), \quad (x, y) \in B \subset \mathbb{R}^2,$$

subject to the initial condition $g(0, x, y) = g_0(x, y)$ and zero flux boundary conditions

$$M \partial_x g \bigg|_{x=0,1} = 0, \quad M \partial_y g \bigg|_{y=0,1} = 0.$$
For simplicity, we adopt a uniform rectangular mesh where $B = \bigcup I_{i,j}$ with

$$I_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}],$$

where the mesh sizes $\Delta x = \frac{1}{P}$, $\Delta y = \frac{1}{Q}$ and $1 \leq i \leq P, 1 \leq j \leq Q$.

Then the semidiscrete DDG scheme is

$$\int_{I_{i,j}} M \partial_t g_h v \, dx \, dy = - \int_{I_{i,j}} M \nabla g_h \cdot \nabla v \, dx \, dy + \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} M \left[ \frac{\partial_x g_h v + (g_h - \{g_h\}) \partial_x v}{\partial x} \right] \left|_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \right| \, dy + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} M \left[ \frac{\partial_y g_h v + (g_h - \{g_h\}) \partial_y v}{\partial y} \right] \left|_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \right| \, dx,$$

where

$$\frac{\partial_x g_h}{\partial x} \bigg|_{(x_{j+\frac{1}{2}}, y)} = \frac{\beta_0}{\Delta x} [g_h] + \{\partial_x g_h\} + \beta_1 \Delta x [\partial_x^2 g_h], \quad (4.19)$$

$$\frac{\partial_y g_h}{\partial y} \bigg|_{(x, y_{j+\frac{1}{2}})} = \frac{\beta_0}{\Delta y} [g_h] + \{\partial_y g_h\} + \beta_1 \Delta [\partial_y^2 g_h]. \quad (4.20)$$

Zero flux is chosen at the boundary $\partial B$ to incorporate the boundary condition.

### 4.2.3 Nonrectangular domain

For the dumbbell models, we shall use polar coordinates to design the numerical method.

For instance when $d = 2$, we change the variables $(x, y)$ to $(r, \theta)$ by

$$r = \sqrt{x^2 + y^2} \quad \text{and} \quad \theta = \arctan \left( \frac{y}{x} \right).$$

Then the domain $B(0, \sqrt{b})$ becomes a rectangle $[0, \sqrt{b}] \times [0, 2\pi]$. And (4.13) becomes

$$r M \partial_t g = \partial_r (r M \partial_r g) + \partial_\theta \left( \frac{M}{r} \partial_\theta g \right).$$
we obtain the following DG scheme:

\[
\int_{I_{i,j}} r M \partial_t g_h v \, drd\theta = - \int_{I_{i,j}} M \left( r \partial_r g_h \partial_r v + \frac{1}{r} \partial_\theta g_h \partial_\theta v \right) \, drd\theta
\]  

(4.21)

\[
+ \int_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}} r M \left[ \partial_r g_h v + (g_h - \{g_h\}) \partial_r v \right] \, d\theta
\]  

(4.21)

\[
+ \int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} M \left[ \partial_\theta g_h v + (g_h - \{g_h\}) \partial_\theta v \right] \, dr
\]  

We impose zero flux boundary condition at \( r_{\frac{1}{2}} \) and \( r_{P+\frac{1}{2}} \) and periodic boundary condition for \( \theta_{\frac{1}{2}} \) and \( \theta_{Q+\frac{1}{2}} \).

### 4.3 Positivity-preserving schemes

We consider the Euler forward temporal discretization of (4.16)

\[
\int_{I_j} M \frac{g_{h}^{n+1} - g_{h}^{n}}{\Delta t} v \, dx = - \int_{I_j} M \partial_x g_{h}^{n+1} v \, dx + M \left[ \partial_x g_{h}^{n} v + (g_{h}^{n} - \{g_{h}^{n}\}) \partial_x v \right]_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}, \quad 1 \leq j \leq N,
\]  

(4.22)

where \( g_{h}^{n} \) is the approximation to \( g(t^n, x) \). We introduce the following notation

\[
\langle \phi \rangle_j = \frac{1}{2} \int_{-1}^{1} M(x_{j} + \frac{h}{2}\xi)\phi(\xi)\,d\xi \quad \text{where } \phi \in P^k([-1, 1]).
\]  

(4.23)

With this notation we have

\[
\langle g_{h} \rangle_j = \frac{1}{h} \int_{I_j} M(x) g_{h}(x) \, dx.
\]

Taking \( v|_{I_j} = 1 \) in (4.22), we have

\[
\langle g_{h}^{n+1} \rangle_j = \langle g_{h}^{n} \rangle_j + \lambda h \, M \partial_x g_{h}^{n} \bigg|_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}}.
\]  

(4.24)

where \( \lambda = \frac{\Delta t}{h^2} \) is the mesh ratio. Assuming that \( \langle g_{h}^{n} \rangle_j \geq 0 \) for all \( j \)'s, we would like to derive some sufficient conditions such that \( \langle g_{h}^{n+1} \rangle_j \geq 0 \) under certain Courant-Friedrich-Lewy (CFL) conditions on \( \lambda \).
4.3.1 The first order scheme

In Liu and Yu (2012a), we showed the unconditionally positivity preserving property for the implicit time discretization. For the explicit time discretization, positivity preserving property holds only under suitable CFL conditions on the mesh ratio $\lambda$, as illustrated below.

When $k = 0$, $g_h$ is a constant on each cell $I_j$ which is exactly $g_j = \langle g_h \rangle_j$. We take the numerical flux

$$\overline{\partial_x g_h} \bigg|_{x_{j+\frac{1}{2}}} = \frac{g_{j+1} - g_j}{h},$$

so that the first order scheme has the form

$$\frac{1}{h} \int_{I_1} M \, dx g_1^{n+1} = \frac{1}{h} \int_{I_1} M \, dx g_1^n + \lambda M_{\frac{3}{2}} (g_2^n - g_1^n),$$

$$\frac{1}{h} \int_{I_j} M \, dx g_j^{n+1} = \frac{1}{h} \int_{I_j} M \, dx g_j^n + \lambda \left[ M_{j+\frac{1}{2}} (g_{j+1}^n - g_j^n) - M_{j-\frac{1}{2}} (g_j^n - g_{j-1}^n) \right], \quad (4.25)$$

$$2 \leq j \leq N - 1,$$

$$\frac{1}{h} \int_{I_N} M \, dx g_N^{n+1} = \frac{1}{h} \int_{I_N} M \, dx g_N^n - \lambda M_{N-\frac{1}{2}} (g_N^n - g_{N-1}^n).$$

It follows that

$$\frac{1}{h} \int_{I_1} M \, dx g_1^{n+1} = \left( \frac{1}{h} \int_{I_1} M \, dx - \lambda M_{\frac{3}{2}} \right) g_1^n + \lambda M_{\frac{3}{2}} g_2^n,$$

$$\frac{1}{h} \int_{I_j} M \, dx g_j^{n+1} = \left[ \frac{1}{h} \int_{I_j} M \, dx - \lambda (M_{j+\frac{1}{2}} + M_{j-\frac{1}{2}}) \right] g_j^n + \lambda M_{j+\frac{1}{2}} g_{j+1}^n + \lambda M_{j-\frac{1}{2}} g_{j-1}^n,$$

$$2 \leq j \leq N - 1,$$

$$\frac{1}{h} \int_{I_N} M \, dx g_N^{n+1} = \left( \frac{1}{h} \int_{I_N} M \, dx - \lambda M_{N-\frac{1}{2}} \right) g_N^n + \lambda M_{N-\frac{1}{2}} g_{N-1}^n.$$

We, therefore, immediately have the strict nonnegativity principle.

Theorem 4.3.1. Consider the fully discrete scheme (4.25).

If $c_1 \leq g_j^n \leq c_2$, then $c_1 \leq g_j^{n+1} \leq c_2$,

provided

$$\lambda \leq \min \left\{ \frac{1}{h} \int_{I_1} M \, dx \, M_{\frac{3}{2}}, \min_{2 \leq j \leq N-1} \frac{1}{h} \int_{I_j} M \, dx \, (M_{j+\frac{1}{2}} + M_{j-\frac{1}{2}}), \frac{1}{h} \int_{I_N} M \, dx \, M_{N-\frac{1}{2}} \right\}.$$
4.3.2 The second order schemes

For higher order schemes, it is known difficult to achieve the nonnegativity by mere restrictions on the mesh ratio, if not impossible. Our idea is to realize the positivity preserving property by identifying a range of pairs \((\beta_0, \beta_1)\) together with the CFL condition. This strategy works if in addition we assume \(g^n_h(x)\) is positive on \(I_j\). The latter requirement can be realized by reconstructing \(\tilde{g}^n_h\) from \(g^n_h\) using \(\langle g^n_h \rangle_j\) such that \(\tilde{g}^n_h\) is nonnegative pointwise and maintains the same accuracy as \(g^n_h\).

For the second order scheme \((k = 1)\), we have the following result.

**Theorem 4.3.2.** \((k = 1)\) The scheme \((4.24)\) with \(\beta_0 \geq 1\) is positivity preserving, namely, \(\langle g^{n+1}_h \rangle_j \geq 0\) if \(g^n_h(x) \geq 0\) on the set \(S_j\) where

\[
S_j = x_j + \frac{h}{2}\{-\gamma, \gamma\}
\]

for \(\gamma\) satisfying

\[
\max_{1 \leq j \leq N} \frac{|\langle \xi \rangle_j|}{\langle 1 \rangle_j} < \gamma \leq 1 \quad \text{and} \quad \gamma \geq |1 - \beta_0^{-1}|, \tag{4.26}
\]

under the CFL condition \(\lambda \leq \lambda_0\) defined below in \((4.30)\).

**Proof.** For any \(p \in P^1([-1, 1])\) and \(\gamma \in [-1, 1] \setminus \{0\}\), we have

\[
p(\xi) = \frac{p(-\gamma) + p(\gamma)}{2} + \frac{p(\gamma) - p(-\gamma)}{2\gamma} \xi. \tag{4.27}
\]

Hence the weighted average becomes

\[
\langle p \rangle = \frac{\langle \gamma - \xi \rangle}{2\gamma} p(-\gamma) + \frac{\langle \gamma + \xi \rangle}{2\gamma} p(\gamma).
\]

Let \(p(\xi) = g_h(x_j + \frac{h}{2}\xi) = g_h(x)|_{I_j}\) for \(\xi \in [-1, 1]\). Then we have

\[
\langle g_h \rangle_j = \hat{\omega}_j^1 p(-\gamma) + \hat{\omega}_j^2 p(\gamma), \tag{4.28}
\]

where

\[
\hat{\omega}_j^1 = \frac{\gamma \langle 1 \rangle_j - \langle \xi \rangle_j}{2\gamma} \quad \text{and} \quad \hat{\omega}_j^2 = \frac{\gamma \langle 1 \rangle_j + \langle \xi \rangle_j}{2\gamma}.
\]
are positive for $\gamma$ satisfying the first inequality in (4.26). Moreover, $\beta_0 \geq 1$ ensures the existence of $\gamma$ such that both inequalities in (4.26) hold.

We next express the numerical flux in terms of solution values over the test set $S_j$. Set also $p_\pm(\xi) := g_h(x_{j+1} + \frac{\xi}{2})$ for $\xi \in [-1, 1]$. A direct calculation using (4.27) gives

$$h\partial_x g_h\bigg|_{x_j + \frac{1}{2}} = \beta_0 \left( g_h^+(x_{j+\frac{1}{2}}) - g_h^-(x_{j+\frac{1}{2}}) \right) + \frac{h}{2} \left( \partial_x g_h^+(x_{j+\frac{1}{2}}) + \partial_x g_h^-(x_{j+\frac{1}{2}}) \right)$$

$$= (\beta_0 p_+(-1) + \partial_x p_+(-1)) - (\beta_0 p(1) - \partial_x p(1))$$

$$= \alpha_+ p_+(-\gamma) + \alpha_- p_+(\gamma) - (\alpha_- p(-\gamma) + \alpha_+ p(\gamma)), \quad (4.29)$$

where

$$\alpha_\pm = \frac{\beta_0}{2} \pm \frac{\beta_0 - 1}{2\gamma}.$$ 

Notice that with $\gamma$ satisfying (4.26), we have

$$\alpha_\pm = \frac{\beta_0}{2\gamma} (\gamma \pm (1 - \beta_0^{-1})) \geq 0.$$ 

Then substitution of (4.28) and (4.29) into (4.24) gives

$$\langle g_h^{n+1} \rangle_j = \langle g_h^n \rangle_j + \lambda \left( Mh\partial_x g_h^n\bigg|_{x_{j+\frac{1}{2}}} - Mh\partial_x g_h^n\bigg|_{x_{j-\frac{1}{2}}} \right)$$

$$\quad = \left[ \hat{\omega}_j^1 - \lambda (\alpha_+ M_{j-\frac{1}{2}} + \alpha_- M_{j+\frac{1}{2}}) \right] p(-\gamma) + \left[ \hat{\omega}_j^2 - \lambda (\alpha_- M_{j-\frac{1}{2}} + \alpha_+ M_{j+\frac{1}{2}}) \right] p(\gamma)$$

$$\quad \quad + \lambda M_{j+\frac{1}{2}} [\alpha_+ p_+(-\gamma) + \alpha_- p_+(\gamma)] + \lambda M_{j-\frac{1}{2}} [\alpha_- p_+(-\gamma) + \alpha_+ p_-(\gamma)].$$

For $j = 1$ or $N$, we will not have the terms involving $M_{1/2}$ or $M_{N+1/2}$ due to the zero flux boundary condition. Hence $\langle g_h^{n+1} \rangle_j \geq 0$ as long as $g_h^n \geq 0$ on $S_j$ and $\lambda \leq \lambda_0$ with

$$\lambda_0 := \min_{1 \leq j \leq N} \left\{ \frac{\hat{\omega}_j^1}{\alpha_+ M_{j-\frac{1}{2}} + \alpha_- M_{j+\frac{1}{2}}}, \frac{\hat{\omega}_j^2}{\alpha_- M_{j-\frac{1}{2}} + \alpha_+ M_{j+\frac{1}{2}}} \right\}.$$

Upon simplification,

$$\lambda_0 = \frac{1}{\beta_0} \min_{1 \leq j \leq N} \frac{\langle \gamma \neq \xi \rangle_j}{(\gamma \pm 1 \mp \beta_0^{-1}) M_{j-\frac{1}{2}} + (\gamma \mp 1 \pm \beta_0^{-1}) M_{j+\frac{1}{2}}} \quad (4.30)$$

Remark 4.3.1. From (4.26) we see that one can always set $\gamma = 1$, however such a choice would require $g_h^n(x) \geq 0$ for all $x$. In practice, we prefer to choose a smaller $\gamma$ so that positivity on a subset is sufficient to update our positivity preserving scheme. This point will become clearer after we introduce the reconstruction in Section 4.3.4 later.
4.3.3 The third order schemes

In the case \( k = 2 \), the use of nontrivial \( \beta_1 \) is essential so that \( (g^h_{n+1})_j \) can be expressed as a nondecreasing function of solution values on some test sets in the previous time step, which when combined with the scheme consistency ensures the positivity property. We first present a way of decomposition for the weighted average of polynomials of degree 2.

For any \( p \in P^2([-1, 1]) \) and \( \gamma \in (-1, 1) \), the unique interpolation of \( p \) at three points \( \{-1, \gamma, 1\} \) gives the following
\[
p(\xi) = \omega^1 p(-1) + \omega^2 p(\gamma) + \omega^3 p(1),
\]
where
\[
\omega^1 = \frac{\gamma - \xi(1 + \gamma) + \xi^2}{2(1 + \gamma)},
\]
\[
\omega^2 = \frac{1 - \xi^2}{1 - \gamma^2},
\]
\[
\omega^3 = \frac{-\gamma + \xi(1 - \gamma) + \xi^2}{2(1 - \gamma)}.
\]
This gives the following identity for the weighted average,
\[
\langle p \rangle = \tilde{\omega}^1 p(-1) + \tilde{\omega}^2 p(\gamma) + \tilde{\omega}^3 p(1),
\]
where \( \tilde{\omega}^i = \langle \omega^i \rangle \) are
\[
\tilde{\omega}^1 = \frac{\langle \gamma - \xi(1 + \gamma) + \xi^2 \rangle}{2(1 + \gamma)},
\]
\[
\tilde{\omega}^2 = \frac{\langle 1 - \xi^2 \rangle}{1 - \gamma^2},
\]
\[
\tilde{\omega}^3 = \frac{\langle -\gamma + \xi(1 - \gamma) + \xi^2 \rangle}{2(1 - \gamma)}.
\]
We use the following notation
\[
a = \frac{\langle \xi - \xi^2 \rangle}{\langle 1 - \xi \rangle}, \quad b = \frac{\langle \xi + \xi^2 \rangle}{\langle 1 + \xi \rangle}.
\]
Note that we shall use \( \langle \cdot \rangle_j \) later when distinguishing the weighted average by each cell becomes necessary. Regarding the positivity of the coefficients given in (4.34), we have the following result.
Lemma 4.3.3. $\hat{\omega}^i > 0$ for $i = 1, 2, 3$ if and only if

$$\gamma \in (a, b)$$

where $a, b$ satisfy $-1 < a < b < 1$.

Proof. We first show the interval $(a, b)$ is within $(-1, 1)$ and not empty. Since $\langle \xi^2 \rangle < 1$ and $\langle 1 \pm \xi \rangle > 0$, we have

$$a = \frac{\langle \xi - \xi^2 \rangle}{\langle 1 - \xi \rangle} > \frac{\langle \xi - 1 \rangle}{\langle 1 - \xi \rangle} = -1, \quad b = \frac{\langle \xi + \xi^2 \rangle}{\langle 1 + \xi \rangle} < \frac{\langle \xi + 1 \rangle}{\langle 1 + \xi \rangle} = 1.$$  

Therefore $(a, b) \subset [-1, 1]$. Note also

$$b - a = \frac{\langle \xi + \xi^2 \rangle}{\langle 1 + \xi \rangle} - \frac{\langle \xi - \xi^2 \rangle}{\langle 1 - \xi \rangle} = 2 \frac{\langle 1 \rangle \langle \xi^2 \rangle - \langle \xi \rangle^2}{\langle 1 - \xi \rangle \cdot \langle 1 + \xi \rangle}. \quad (4.36)$$

Using the Hölder inequality,

$$\langle \xi \rangle^2 = \left( \frac{1}{2} \int_{-1}^{1} \xi M \, d\xi \right)^2 < \left( \frac{1}{2} \int_{-1}^{1} M \, d\xi \right) \left( \frac{1}{2} \int_{-1}^{1} \xi^2 M \, d\xi \right) = \langle 1 \rangle \langle \xi^2 \rangle.$$  

Hence $b - a > 0$. Next we show positivity of $\hat{\omega}^i$: from (4.34) it follows that $\hat{\omega}^2 > 0$ unconditionally and

$$2(1 + \gamma)\hat{\omega}^1 = (\gamma - a)(1 - \xi) > 0, \quad 2(1 - \gamma)\hat{\omega}^3 = (b - \gamma)(1 + \xi) > 0.$$  

The above positive decomposition enables us to obtain the following result for the third-order scheme where $k = 2$.

Theorem 4.3.4. $(k = 2)$ The scheme (4.24) with

$$\frac{1}{8} \leq \beta_1 \leq \frac{1}{4} \quad \text{and} \quad \beta_0 \geq 1 \quad (4.37)$$

is positivity preserving, namely, $\langle g_h^{n+1} \rangle_j \geq 0$ if $g_h^n(x)$ is nonnegative on the set $S_j$’s where

$$S_j = x_j + \frac{h}{2} \{ -1, \gamma, 1 \}.$$
with $\gamma$ satisfying

$$a_j < \gamma < b_j \quad \text{and} \quad |\gamma| \leq 8\beta_1 - 1, \quad (4.38)$$

under the CFL condition

$$\lambda \leq \lambda_0, \quad (4.39)$$

for some $\lambda_0$ defined in (4.45) below.

**Proof.** We present the proof in three steps:

1. **Weighted integral decomposition.** Let

   $$p(\xi) = gh(x_j + \frac{h}{2}\xi) = gh(x)|_{I_j}, \quad \xi \in [-1, 1].$$

   We have

   $$(gh)_j = \hat{\omega}_j^1 p(-1) + \hat{\omega}_j^2 p(\gamma) + \hat{\omega}_j^3 p(1), \quad (4.40)$$

   where $\hat{\omega}_j^i$ are given in (4.34) with the weight $M(x)|_{I_j} = M(x_j + \frac{h}{2}\xi)$ and they are positive since

   $$a_j < \gamma < b_j, \quad j = 1, \ldots, N.$$

2. **Flux representation.** In order to express the numerical flux in terms of solution values over the set $S_j$, we need the following derivatives from (4.31):

   $$p'(\xi) = \left(\frac{\xi}{1+\gamma} - \frac{1}{2}\right) p(-1) - 2\frac{\xi}{1+\gamma} p(\gamma) + \left(\frac{\xi}{1+\gamma} + \frac{1}{2}\right) p(1),$$

   $$p''(\xi) = \frac{1}{1+\gamma} p(-1) - 2\frac{1}{1+\gamma} p(\gamma) + \frac{1}{1+\gamma} p(1). \quad (4.41)$$

   Set $p_{\pm}(\xi) := gh(x_{j\pm1} + \frac{h}{2}\xi)$ for $\xi \in [-1, 1]$. A direct calculation gives

   $$h \left. \frac{\partial_x gh}{x_{j+\frac{1}{2}}} \right|_{x_{j+\frac{1}{2}}} = \beta_0 [gh] + h \{ \partial_x g_h \} + \beta_1 h^2 [\partial_x^2 g_h]|_{x_{j+\frac{1}{2}}}$$

   $$= \beta_0 (p_+(-1) - p(1)) + (\partial_\xi p_+(-1) + \partial_\xi p(1)) + 4\beta_1 (\partial_\xi^2 p_+(-1) - \partial_\xi^2 p(1))$$

   $$= \beta_0 p_+(-1) + \partial_\xi p_+(-1) + 4\beta_1 \partial_\xi^2 p_+(-1) - \left[ \beta_0 p(1) - \partial_\xi p(1) + 4\beta_1 \partial_\xi^2 p(1) \right].$$

   Using (4.41), we have

   $$\beta_0 p(1) - \partial_\xi p(1) + 4\beta_1 \partial_\xi^2 p(1) = \alpha_1(\gamma) p(-1) + \alpha_2(\gamma) p(\gamma) + \alpha_3(\gamma) p(1),$$
where

\[
\alpha_1(\gamma) = \frac{8\beta_1 - 1 + \gamma}{2(1 + \gamma)}, \quad \alpha_2(\gamma) = 2\frac{1 - 4\beta_1}{1 - \gamma^2}, \quad \alpha_3(\gamma) = \beta_0 + \frac{8\beta_1 - 3 + \gamma}{2(1 - \gamma)}.
\]  

(4.42)

Similarly, we have

\[
\beta_0 p_+(-1) + \partial_{t}p_+(-1) + 4\beta_1 \partial_{x}^2 p_+(-1) = \alpha_3(-\gamma)p_+(-1) + \alpha_2(-\gamma)p_+(\gamma) + \alpha_1(-\gamma)p_+(1).
\]

It follows that

\[
\beta_0 p_+(-1) + \partial_{t}p_+(-1) + 4\beta_1 \partial_{x}^2 p_+(-1) = \alpha_3(-\gamma)p_+(-1) + \alpha_2(-\gamma)p_+(\gamma) + \alpha_1(-\gamma)p_+(1)
\]

(4.43)

\[- [\alpha_1(\gamma)p(-1) + \alpha_2(\gamma)p(\gamma) + \alpha_3(\gamma)p(1)].
\]

It is easy to verify that (4.37) ensures \(\alpha_i(\pm \gamma) \geq 0\) for \(i = 1, 2\). For \(\alpha_3(\gamma)\) we have

\[
\alpha_3(\gamma) = \beta_0 + \frac{8\beta_1 - 3 + \gamma}{2(1 - \gamma)} \geq 1 + \frac{8\beta_1 - 3 + \gamma}{2(1 - \gamma)} = \frac{8\beta_1 - 1 - \gamma}{2(1 - \gamma)} \geq 0.
\]

In a similar manner, we can verify that \(\alpha_3(-\gamma) \geq 0\).

**Step 3.** Monotonicity under some CFL condition. We now substitute (4.40) and (4.43) into (4.24) to obtain

\[
\langle g_h^{n+1} \rangle_j = \langle g_h^n \rangle_j + \lambda \left( Mh \partial_x g_h^n \right)_{x_j + \frac{\lambda}{2}} - Mh \partial_x \hat{g}_h^n_{x_j - \frac{\lambda}{2}} \right)
\]

(4.44)

\[= \left[ \hat{\omega}_j^1 - \lambda \left( \alpha_3(-\gamma)M_j - \frac{1}{2} + \alpha_1(\gamma)M_j + \frac{1}{2} \right) \right] p(-1)
\]

\[+ \left[ \hat{\omega}_j^2 - \lambda \left( \alpha_2(-\gamma)M_j - \frac{1}{2} + \alpha_2(\gamma)M_j + \frac{1}{2} \right) \right] p(\gamma)
\]

\[+ \left[ \hat{\omega}_j^3 - \lambda \left( \alpha_1(-\gamma)M_j - \frac{1}{2} + \alpha_3(\gamma)M_j + \frac{1}{2} \right) \right] p(1)
\]

\[+ \lambda M_{j - \frac{1}{2}} \left[ \alpha_3(-\gamma)p_{-1} + \alpha_2(-\gamma)p_+(\gamma) + \alpha_1(-\gamma)p_+(1) \right]
\]

\[+ \lambda M_{j - \frac{1}{2}} \left[ \alpha_1(\gamma)p_{-1} + \alpha_2(\gamma)p_{-1} + \alpha_3(\gamma)p(-1) \right].
\]

For \(j = 1 \text{ or } N\), using the zero flux boundary condition, we will not have the terms involving \(M_{\frac{1}{2}}\) or \(M_{\frac{N-1}{2}}\). Hence \(\langle g_h^{n+1} \rangle_j \geq 0\) as long as \(g_h^n \geq 0\) on \(S_j\) and \(\lambda \leq \lambda_0\) with \(\lambda_0\) being

\[
\min \left\{ \frac{\hat{\omega}_j^1}{\alpha_3(-\gamma)M_j + \alpha_1(\gamma)M_j + \frac{1}{2}}, \frac{\hat{\omega}_j^2}{\alpha_2(-\gamma)M_j + \alpha_2(\gamma)M_j + \frac{1}{2}}, \frac{\hat{\omega}_j^3}{\alpha_1(-\gamma)M_j + \alpha_3(\gamma)M_j + \frac{1}{2}} \right\}.
\]
Using the fact $\hat{\omega}_j^3(\gamma) = \hat{\omega}_j^1(-\gamma)$ and formulas for $\alpha_2(\gamma), \hat{\omega}_j^2$, we have

$$\lambda_0 = \min_{1 \leq j \leq N} \left\{ \frac{\hat{\omega}_j^1(\pm\gamma)}{\alpha_3(\mp\gamma)M_{j-\frac{1}{2}} + \alpha_1(\pm\gamma)M_{j+\frac{1}{2}}} \cdot \frac{\langle 1 - \xi^2 \rangle_j}{2(1 - 4\beta_1)(M_{j-\frac{1}{2}} + M_{j+\frac{1}{2}})} \right\}. \quad (4.45)$$

**Remark 4.3.2.** The above positivity preserving property indicates that the use of $\beta_1$ ensures the existence of $\gamma$ satisfying both inequalities in (4.53). In particular, the option $\beta_1 = \frac{1}{4}$ works for any mesh and will be tested in our numerical simulation later.

**Remark 4.3.3.** For a fixed mesh, it may be necessary to choose different $\gamma$ for each cell since $\bigcap_{j=1}^N (a_j, b_j)$ can be empty.

**Remark 4.3.4.** Indeed, Theorem 4.3.4 remains valid if we use the following test set

$$S_j = x_j + \frac{h}{2} \{-1, \gamma_j, 1\}$$

with $\gamma_j$ satisfying

$$a_j < \gamma_j < b_j \quad \text{and} \quad |\gamma_j| \leq 8\beta_1 - 1. \quad (4.46)$$

In the proof, we need to track the dependence of coefficients on the $\gamma_j$'s. In such a case, the cell average (4.24) is given by

$$\langle g_{h+1} \rangle_j = \left[ \hat{\omega}_j^1 - \lambda \left( \frac{\alpha_3(-\gamma_j)M_{j-\frac{1}{2}} + \alpha_1(\gamma_j)M_{j+\frac{1}{2}}}{\alpha_3(\mp\gamma)M_{j-\frac{1}{2}} + \alpha_1(\pm\gamma)M_{j+\frac{1}{2}}} \right) \right] p(-1) + \left[ \hat{\omega}_j^2 - \lambda \left( \frac{\alpha_2(-\gamma_j)M_{j-\frac{1}{2}} + \alpha_2(\gamma_j)M_{j+\frac{1}{2}}}{\alpha_3(\mp\gamma)M_{j-\frac{1}{2}} + \alpha_1(\pm\gamma)M_{j+\frac{1}{2}}} \right) \right] p(\gamma_j) + \left[ \hat{\omega}_j^3 - \lambda \left( \frac{\alpha_1(-\gamma_j)M_{j-\frac{1}{2}} + \alpha_3(\gamma_j)M_{j+\frac{1}{2}}}{\alpha_3(\mp\gamma)M_{j-\frac{1}{2}} + \alpha_1(\pm\gamma)M_{j+\frac{1}{2}}} \right) \right] p(1) + \lambda M_{j+\frac{1}{2}} \left[ \alpha_3(-\gamma_{j+1})p_+(1) + \alpha_2(-\gamma_{j+1})p_+(\gamma_{j+1}) + \alpha_1(-\gamma_{j+1})p_+(-1) \right] + \lambda M_{j-\frac{1}{2}} \left[ \alpha_3(\gamma_{j-1})p_-(1) + \alpha_2(\gamma_{j-1})p_-(\gamma_{j-1}) + \alpha_3(\gamma_{j-1})p_-(-1) \right].$$

The coefficients of $\{p(-1), p(\gamma_j), p(1)\}$ being nonnegative gives the modified CFL number $\lambda_0$.

**Remark 4.3.5.** The CFL conditions are much more restrictive than the commonly used ones. Nevertheless, they are sufficient conditions rather than necessary to preserve the positivity of solutions. Therefore, in practice, these CFL conditions are strictly enforced only in the case the positivity preserving property is violated.
4.3.4 A scaling limiter

Theorem 4.3.2 and 4.3.4 tell us that for the scheme (4.22), we need to modify \( g_h^n \) such that it is nonnegative on \( S_j \). To begin with, we will show an approach so that the modified \( g_h^n \geq 0 \) pointwise.

Let \( g_h \in P^k(I_j) \) be an approximation of a smooth function \( g(x) \geq 0 \) with the weighted cell average defined by

\[
\bar{g}_j := \frac{\int_{I_j} M g_h \, dx}{\int_{I_j} M \, dx}.
\]

Following the idea of scaling limiter by Zhang and Shu (2010), we define the scaled polynomial by

\[
\tilde{g}_h(x) = \eta (g_h(x) - \bar{g}_j) + \bar{g}_j, \quad \eta = \min \left\{ 1, \frac{\bar{g}_j}{\bar{g}_j - \zeta_j} \right\}
\]

with

\[
\zeta_j = \min_{x \in I_j} g_h(x).
\]

It is easy to check that the cell average of \( \tilde{g}_h \) is still \( \bar{g}_j \) and \( \tilde{g}_h \geq 0 \) in \( I_j \). Following Zhang and Shu (2010), we have the next lemma.

**Lemma 4.3.5.** If \( \bar{g}_j > 0 \), then the modified polynomial is as accurate as \( g_h \) in the following sense

\[
|\tilde{g}_h(x) - g_h(x)| \leq C_k \|g_h - g\|_{\infty} \quad \forall x \in I_j,
\]

where \( C_k \) is a constant depending on the polynomial degree \( k \) only.

**Proof.** We only need to consider the case when \( \eta = \frac{\bar{g}_j}{\bar{g}_j - \zeta_j} \), i.e., \( \min_{x \in I_j} g_h(x) < 0 \).

\[
|\tilde{g}_h(x) - g_h(x)| = |(\eta - 1)(\bar{g}_j - g_h(x))|
\]

\[
= \left| \min_{x \in I_j} g_h(x) \right| \left| \frac{\bar{g}_j - \min_{x \in I_j} g_h(x)}{\bar{g}_j - g_h(x)} \right| |g_h(x)|
\]

\[
\leq \frac{\max_{x \in I_j} |\bar{g}_j - g_h(x)|}{|\bar{g}_j - \min_{x \in I_j} g_h(x)|} \max_{x \in I_j} \left| \min_{x \in I_j} g_h(x) \right| |g_h(x)|
\]

\[
\leq \frac{\max_{x \in I_j} |\bar{g}_j - g_h(x)|}{\max_{x \in I_j} (\bar{g}_j - g_h(x))} \|g_h - g\|_{\infty}.
\]
Let $p(\xi) = \bar{g}_j - g_h \left( \frac{h}{2} \xi + x_j \right)$ for $\xi \in I = [-1, 1]$. It suffices to show the boundedness of

$$\max_{\xi \in I} \frac{|p(\xi)|}{\max_{\xi \in I} p(\xi)},$$

where $P_0^k$ denotes the finite dimensional linear space which consists of all the polynomials in $P^k(I)$ that satisfy

$$\int_I M \left( \frac{h}{2} \xi + x_j \right) p(\xi) d\xi = 0.$$

Let $\phi = (\phi_1, \ldots, \phi_{k+1})^T$ be the basis of $P_0^k$. Then for any $p \in P_0^k$, there is a unique vector $a \in \mathbb{R}^{k+1}$ such that $p(\xi) = a^T \phi(\xi)$. Using this expression, we have

$$\sup_{a \neq 0} \max_{\xi \in I} \frac{|a^T \phi(\xi)|}{\max_{\xi \in I} |a^T \phi(\xi)|} = \sup_{\|a\|_2 = 1} \max_{\xi \in I} \frac{|a^T \phi(\xi)|}{\|\phi(\xi)\|_2} \leq \frac{\inf_{\|a\|_2 = 1} F(a)}{\inf_{\|a\|_2 = 1} \|\phi(\xi)\|_2},$$

where

$$F(a) = \max_{\xi \in I} a^T \phi(\xi) \quad \text{for } a \in S^k = \{a \in \mathbb{R}^{k+1} : \|a\|_2 = 1\}.$$ 

Notice that $F(a) = 0$ if and only if $a = 0$. Therefore the minimum of $F(a)$ exists and is positive as long as we can show $F(a)$ is a continuous function on the sphere $S^k$. Let $a, b$ be any two vectors in $S^k$. Then using $a = (a - b) + b$, we have

$$\max_{\xi \in I} a^T \phi(\xi) \leq \max_{\xi \in I} (a^T - b^T) \phi(\xi) + \max_{\xi \in I} b^T \phi(\xi).$$

Therefore

$$|F(a) - F(b)| = \max_{\xi \in I} a^T \phi(\xi) - \max_{\xi \in I} b^T \phi(\xi) \leq \max_{\xi \in I} |(a^T - b^T) \phi(\xi)| \leq \|a - b\|_2 \max_{\xi \in I} \|\phi(\xi)\|_2.$$ 

The continuity of $F(a)$ on $S^k$ implies that $\min_{a \in S^k} F(a) > 0$. Hence

$$\sup_{p \in P_0^k \atop p \neq 0} \max_{\xi \in I} \frac{|p(\xi)|}{\max_{\xi \in I} p(\xi)} \leq C_k := \frac{\min_{\|a\|_2 = 1} F(a)}{\max_{\|a\|_2 = 1} \|\phi(\xi)\|_2}.$$ (4.50)
where $C_k$ depends only on $k$. The proof of (4.49) is now complete.

**Remark 4.3.6.** Since we only need to control the values at the points in $S_j$, we could replace (4.48) by

$$
\zeta_j = \min_{x \in S_j} g_h(x)
$$

and the limiter (4.47) with (4.51) is sufficient to ensure

$$
\tilde{g}_h(x) \geq 0 \quad \forall x \in S_j.
$$

Furthermore, Lemma 4.3.5 remains valid with this less restrictive limiter, i.e., we have

$$
|\tilde{g}_h(x) - g_h(x)| \leq C_k\|g_h - g\|_{\infty} \quad \forall x \in I_j,
$$

where $C_k$ is still given by (4.50), yet with $F(a) = \max_{\xi \in \{-1,\gamma,1\}} a^T \phi(\xi)$.

### 4.3.5 The maximum-principle-satisfying property

It is known that the nonnegativity principle is implied by the following maximum principle for the Fokker–Planck problem:

if $c_1 \leq g_0(x) \leq c_2 \forall x \in B$, then $c_1 \leq g(x,t) \leq c_2 \forall t > 0$, which in terms of $f$ is equivalent to (4.5). The following result shows that the scheme (4.24) is also bound preserving under the same sufficient conditions. More precisely, we have the following

**Theorem 4.3.6.** (k = 2) The scheme (4.24) with

$$
\frac{1}{8} \leq \beta_1 \leq \frac{1}{4} \quad \text{and} \quad \beta_0 \geq 1
$$

is bound preserving, namely, $c_1 \leq \tilde{g}_j^{n+1} \leq c_2$ if $g_h^n(x)$ is in $[c_1, c_2]$ on the set $S_j$‘s where

$$
S_j = x_j + \frac{h}{2} \{-1, \gamma, 1\}
$$

with $\gamma$ satisfying

$$
a_j < \gamma < b_j \quad \text{and} \quad |\gamma| \leq 8\beta_1 - 1,
$$

(4.53)
under the CFL condition

\[ \lambda \leq \lambda_0, \tag{4.54} \]

for some \( \lambda_0 \) defined in (4.45).

**Proof.** Note that (4.44) is nondecreasing in the point values \( p(\pm 1), p(\gamma), p_{\pm}(\pm 1), p_{\pm}(\gamma) \), hence when these values are replaced with the lower and upper bounds \( c_1 \) and \( c_2 \) respectively, we have

\[ c_1 \sum_{i=1}^{3} \hat{\omega}^i_j \leq \langle g_{h}^{n+1} \rangle_j \leq c_2 \sum_{i=1}^{3} \hat{\omega}^i_j, \]

since the terms with \( \alpha_i \)'s are cancelled out. Moreover the sum of \( \hat{\omega} \)'s is \( \langle 1 \rangle_j \). Therefore

\[ c_1 \langle 1 \rangle_j \leq \langle g_{h}^{n+1} \rangle_j \leq c_2 \langle 1 \rangle_j. \]

Therefore, \( c_1 \leq \bar{g}_{j}^{n+1} \leq c_2 \).

This result tells us that for the scheme (4.22), we need to modify \( g_{h}^{n} \) such that it is in \([c_1, c_2]\) on \( S_j \). We can use the following scaling limiter

\[ \bar{g}_{h}(x) = \eta (g_{h}(x) - \bar{g}_{j}) + \bar{g}_{j}, \quad \eta = \min \left\{ 1, \left| \frac{g_{j} - c_1}{\bar{g}_{j} - \zeta_j} \right|, \left| \frac{c_2 - \bar{g}_{j}}{\bar{g}_{j} - \zeta_j} \right| \right\} \tag{4.55} \]

with \( \bar{g}_{j} = \langle g_{h} \rangle_j \) and

\[ \zeta_j = \min_{x \in S_j} g_{h}(x), \quad \bar{\zeta}_j = \max_{x \in S_j} g_{h}(x). \tag{4.56} \]

Then the modified polynomial \( \bar{g}_{h}(x) \) satisfies

\[ c_1 \leq \bar{g}_{h}(x) \leq c_2 \quad \text{and} \quad \langle \bar{g}_{h} \rangle_j = \bar{g}_{j} \langle 1 \rangle_j. \]

Moreover, it can be shown (following the proof of Lemma 4.3.5) that if \( c_1 \leq \bar{g}_{j} \leq c_2 \), then the above scaling limiter does not destroy the accuracy. Therefore, we get the revised scheme of (4.24),

\[ \langle g_{h}^{n+1} \rangle_j = \langle g_{h}^{n} \rangle_j + \lambda h M \hat{\partial}_x g_{h}^{n} \big|_{x_j + \frac{1}{2}}. \tag{4.57} \]

A detailed implementation algorithm will be given in Section 4.5.
4.4 Two-dimensional extensions

4.4.1 Rectangular meshes

We consider the Euler forward temporal discretization of (4.18)

\[
\int_{I_{i,j}} M g_h^{n+1} v \, dx \, dy = \int_{I_{i,j}} M g_h^n v \, dx \, dy - \Delta t \int_{I_{i,j}} M \nabla g_h^n \cdot \nabla v \, dx \, dy
\]

(4.58)

\[
+ \Delta t \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} M \left[ \partial_x g_h^n v + (g_h^n - \{g_h^n\}) \partial_x v \right] \, dy
\]

\[
+ \Delta t \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} M \left[ \partial_y g_h^n v + (g_h^n - \{g_h^n\}) \partial_y v \right] \, dx
\]

We introduce the following notation

\[
\langle \phi \rangle_{ij} = \frac{1}{4} \int_{I_{i,j}} M \left( x_i + \frac{\Delta x}{2}, y_j + \frac{\Delta y}{2} \right) \phi(x, y) \, dx \, dy
\]

(4.59)

With this notation, we have

\[
\langle g_h^n \rangle_{ij} = \frac{1}{\Delta x \Delta y} \int_{I_{i,j}} M g_h^n \, dx \, dy
\]

(4.60)

where \( f \) denotes the average integral. We obtain the cell average update from (4.58) with \( v = 1 \),

\[
\langle g_h^{n+1} \rangle_{ij} = \langle g_h^n \rangle_{ij} + \lambda_x \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \Delta x M \partial_x g_h^n \, dy
\]

(4.61)

where \( \lambda_x = \frac{\Delta x}{\Delta x^2} \) and \( \lambda_y = \frac{\Delta y}{\Delta y^2} \). Let \( \lambda = \lambda_x + \lambda_y \) and decompose \( \langle g_h^n \rangle_{ij} \) as

\[
\langle g_h^n \rangle_{ij} = \frac{\lambda_x}{\lambda} \langle g_h^n \rangle_{ij} + \frac{\lambda_y}{\lambda} \langle g_h^n \rangle_{ij},
\]

so that (4.61) can be rewritten as

\[
\langle g_h^{n+1} \rangle_{ij} = \frac{\lambda_x}{\lambda} \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} H_1(y) \, dy + \frac{\lambda_y}{\lambda} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} H_2(x) \, dx,
\]

(4.62)

where

\[
H_1(y) = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} M g_h^n \, dx + \lambda \Delta x \left. M \partial_x g_h^n \right|_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}},
\]

(4.63)

\[
H_2(x) = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} M g_h^n \, dy + \lambda \Delta y \left. M \partial_y g_h^n \right|_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}
\]

(4.64)
The two integrals in (4.62) can be approximated by quadratures with sufficient accuracy. Let us assume that we use a Gauss quadrature with \( L \geq k+2 \) points, which has accuracy of at least \( O(h^{k+2}) \). Let

\[
S_x^i = \{ x_{\sigma}^i, \sigma = 1, \ldots, L \} \quad \text{and} \quad S_y^j = \{ y_{\sigma}^j, \sigma = 1, \ldots, L \} \tag{4.65}
\]
denote the quadrature points on \([x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]\) and \([y_{j-\frac{1}{2}},y_{j+\frac{1}{2}}]\), respectively. The subscript \( \sigma \) will denote the index of the Gauss quadrature points and \( \omega^{\sigma} \)'s are the quadrature weights at the quadrature points so that

\[
\sum_{\sigma=1}^{L} \omega^{\sigma} = 1.
\]

Using the quadrature rule on the right-hand side of (4.62), we obtain the following scheme

\[
\langle g_{n+1}^h \rangle_{ij} = \frac{\lambda_x}{\lambda} \sum_{\sigma=1}^{L} \omega^{\sigma} H_1(y_{\sigma}^j) + \frac{\lambda_y}{\lambda} \sum_{\sigma=1}^{L} \omega^{\sigma} H_2(x_{\sigma}^i). \tag{4.66}
\]

Applying the one-dimensional result in Theorem 4.3.4 to both \( H_1(y_{\sigma}^j) \) and \( H_2(x_{\sigma}^i) \), we can establish the positivity preserving result for the two-dimensional case. Here we only show the case when \( k = 2 \). Let

\[
\hat{S}_x^i = x_i + \frac{\Delta x}{2} \{-1, \gamma^x, 1\} \quad \text{and} \quad \hat{S}_y^j = y_j + \frac{\Delta y}{2} \{-1, \gamma^y, 1\} \tag{4.67}
\]
denote the test set on \([x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}]\) and \([y_{j-\frac{1}{2}},y_{j+\frac{1}{2}}]\), respectively, with \( \gamma^x, \gamma^y \) satisfying

\[
\frac{\langle \xi - \xi^2 \rangle_i (y_{\sigma}^j)}{(1 - \xi)_i} < \gamma^x < \frac{\langle \xi + \xi^2 \rangle_i (y_{\sigma}^j)}{(1 + \xi)_i}, \quad |\gamma^x| \leq 8\beta_1 - 1, \tag{4.68}
\]

\[
\frac{\langle \eta - \eta^2 \rangle_j (x_{\sigma}^i)}{(1 - \eta)_j} < \gamma^y < \frac{\langle \eta + \eta^2 \rangle_j (x_{\sigma}^i)}{(1 + \eta)_j}, \quad |\gamma^y| \leq 8\beta_1 - 1.
\]

Here we have used the notation

\[
\langle \phi(\xi) \rangle_i(y) = \int_{-1}^{1} M \left( x_i + \frac{\Delta x}{2} \xi, y \right) \phi(\xi) d\xi, \quad \langle \phi(\eta) \rangle_j(x) = \int_{-1}^{1} M \left( x, y_j + \frac{\Delta y}{2} \eta \right) \phi(\eta) d\eta.
\]

We want to find sufficient conditions for the scheme (4.58) to satisfy \( \langle g_{n+1}^h \rangle_{ij} \geq 0 \). We use \( \otimes \) to denote the tensor product and define

\[
S_{ij} = (S_x^i \otimes S_y^j) \cup (\hat{S}_x^i \otimes S_y^j). \tag{4.69}
\]

We can now state our two-dimensional result.
Theorem 4.4.1. \((k = 2)\) Consider a two-dimensional scheme \((4.66)\) satisfied by the weighted cell averages of the DG method \((4.58)\) on rectangular meshes, associated with the approximation DG polynomials \(g^n_{ij}(x, y)\) of degree \(k\), with \((\beta_0, \beta_1)\) chosen so that

\[
\frac{1}{8} \leq \beta_1 \leq \frac{1}{4} \quad \text{and} \quad \beta_0 \geq 1.
\]

If \(g^n_{ij}(x, y) \geq 0\) for all \((x, y) \in S_{ij}\), then \((g^{n+1}_{h})_{ij} \geq 0\) under the CFL condition

\[
\lambda < \lambda_0,
\]

where \(\lambda_0\) is given in \((4.70)\) below.

Proof. It is easy to check that \((g^{n+1}_{h})_{ij}\) in \((4.62)\) is a convex combination of \(H_1(y^n_j)\) and \(H_2(x^n_i)\) for \(\sigma = 1, \ldots, L\); hence \((g^{n+1}_{h})_{ij} \geq 0\) if

\[
H_1(y^n_j) \geq 0 \quad \text{and} \quad H_2(x^n_i) \geq 0, \quad \sigma = 1, \ldots, L.
\]

Applying the one-dimensional result obtained in Theorem 4.3.4 to \(H_1(y^n_j)\), we obtain that for each quadrature point \(y \in S^y_j, H_1(y) \geq 0\) if \(g^n_h(x, y) \geq 0\) on the test set \(S^x_i\) and \(\lambda \leq \lambda_0^x\) with

\[
\lambda_0^x = \min_{ij} \min_{1 \leq \sigma \leq L} \left\{ \frac{(\pm \gamma^x \mp \xi(1 + \gamma^x) + \xi^2)_j(y^n_j)}{2(1 + \gamma^x)(\alpha_3(\mp \gamma^x)M(x_{i-\frac{1}{2}}, y^n_j) + \alpha_1(\pm \gamma^x)M(x_{i+\frac{1}{2}}, y^n_j))}, \frac{(1 - \xi^2)_j(y^n_j)}{2(1 - 4\beta_1) \left[ M(x_{i-\frac{1}{2}}, y^n_j) + M(x_{i+\frac{1}{2}}, y^n_j) \right]} \right\},
\]

where we have used \((4.34)\) for \(\dot{\omega}, (4.42)\) for \(\alpha_i, (4.45)\) and notation \((4.68)\). In an entirely similar manner, we obtain that for each quadrature point \(x \in S^x_i, H_2(x) \geq 0\) if \(g^n_h(x, y) \geq 0\) on the test set \(S^y_j\) and \(\lambda \leq \lambda_0^y\) with

\[
\lambda_0^y = \min_{ij} \min_{1 \leq \sigma \leq L} \left\{ \frac{(\pm \gamma^y \mp \eta(1 + \gamma^y) + \eta^2)_j(x^n_i)}{2(1 + \gamma^y)(\alpha_3(\mp \gamma^y)M(x^n_i, y_{j-\frac{1}{2}}) + \alpha_1(\pm \gamma^y)M(x^n_i, y_{j+\frac{1}{2}}))}, \frac{(1 - \eta^2)_j(x^n_i)}{2(1 - 4\beta_1) \left[ M(x^n_i, y_{j-\frac{1}{2}}) + M(x^n_i, y_{j+\frac{1}{2}}) \right]} \right\},
\]

The proof is thus complete if we take

\[
\lambda_0 = \min\{\lambda_0^x, \lambda_0^y\}. \quad \text{(4.70)}
\]
To enforce the condition in Theorem 4.4.1, we can use the following scaling limiter similar to the one-dimensional case. For all \( i \) and \( j \), assume the cell averages \( \langle g_h \rangle_{ij} \geq 0 \) and define the weighted cell average by

\[
\bar{g}_{ij} := \frac{\int_{I_{ij}} M g_h \, dx \, dy}{\int_{I_{ij}} M \, dx \, dy}.
\]

We use the modified polynomial \( \tilde{g}_h(x, y) \) instead of \( g_h(x, y) \),

\[
\tilde{g}_h(x, y) = \eta \left( g_h(x, y) - \bar{g}_{ij} \right) + \bar{g}_{ij} \quad \text{with} \quad \eta = \min \left\{ 1, \frac{\bar{g}_{ij} - \zeta_{ij}}{\bar{g}_{ij}} \right\},
\]

where \( \zeta_{ij} = \min_{(x, y) \in S_{ij}} g_h(x, y) \). It is also straightforward to prove the high order accuracy of this limiter following the proof of Lemma 4.3.5. For the maximum principle to be satisfied, we need a modified limiter (4.71) with

\[
\eta = \min \left\{ 1, \frac{\bar{g}_{ij} - c_1}{\bar{g}_{ij} - \zeta_{ij}}, \frac{c_2 - \bar{g}_{ij}}{\bar{g}_{ij} - \zeta_{ij}} \right\},
\]

where \( \zeta_{ij} = \min_{(x, y) \in S_{ij}} g_h(x, y) \) and \( \bar{\zeta}_{ij} = \max_{(x, y) \in S_{ij}} g_h(x, y) \).

4.4.2 Nonrectangular meshes

In the case that the domain is a disk, we use polar coordinates, with which the mesh is still rectangular. We consider the Euler forward temporal discretization of (4.21)

\[
\int_{I_{i,j}} r M g_{h}^{n+1} v \, dr \, d\theta = \int_{I_{i,j}} r M g_{h}^{n} v \, dr \, d\theta - \int_{I_{i,j}} M \left( r \partial_r g_{h}^{n} \partial_r v + \frac{1}{r} \partial_\theta g_{h}^{n} \partial_\theta v \right) \, dr \, d\theta
\]

\[
+ \int_{r_{j-\frac{1}{2}}}^{r_{j+\frac{1}{2}}} M \left[ \partial_r g_{h}^{n} v + (g_{h}^{n} - \{g_{h}^{n}\}) \partial_r v \right] \, dr
\]

\[
+ \int_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}} M \left[ \partial_\theta g_{h}^{n} v + (g_{h}^{n} - \{g_{h}^{n}\}) \partial_\theta v \right] \, d\theta
\]

Similarly, we introduce the following notation

\[
\langle \phi \rangle_{ij} = \frac{1}{4} \int_{-1}^{1} \int_{-1}^{1} r M \left( r_i + \frac{\Delta r}{2} \xi, \theta_j + \frac{\Delta \theta}{2} \eta \right) \phi(\xi, \eta) \, d\xi \, d\eta \quad \text{where} \quad \phi \in P^k([-1, 1]^2).
\]

With this notation, we have

\[
\langle g_{h}^{n} \rangle_{ij} = \frac{1}{\Delta r \Delta \theta} \int_{I_{i,j}} r M g_{h}^{n} \, dr \, d\theta = \int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} \int_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}} r M g_{h}^{n} \, dr \, d\theta,
\]
\[ \langle g^{n+1}_h \rangle_{ij} = \langle g^n_h \rangle_{ij} + \lambda_r \int_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}} \Delta r M \partial_r g^n_h d\theta \bigg|_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}} + \lambda_\theta \int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} \Delta \theta M \partial_\theta g^n_h dr \bigg|_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}}, \] (4.74)

where \( \lambda_r = \frac{\Delta t}{(\Delta r)^2} \) and \( \lambda_\theta = \frac{\Delta t}{(\Delta \theta)^2} \). Let \( \lambda = \lambda_r + \lambda_\theta \) and we decompose \( \langle g^n_h \rangle_{ij} \) as

\[ \langle g^n_h \rangle_{ij} = \frac{\lambda_r}{\lambda} \langle g^n_h \rangle_{ij} + \frac{\lambda_\theta}{\lambda} \langle g^n_h \rangle_{ij}, \]

Hence (4.74) can be rewritten as

\[ \langle g^{n+1}_h \rangle_{ij} = \frac{\lambda_r}{\lambda} \int_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}} H_1(\theta) d\theta + \frac{\lambda_\theta}{\lambda} \int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} r H_2(r) dr, \] (4.75)

where

\[ H_1(\theta) = \int_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}} (r M) g^n_h dr + \lambda \Delta r (r M) \partial_r g^n_h \bigg|_{r_{i-\frac{1}{2}}}^{r_{i+\frac{1}{2}}}, \] (4.76)

\[ H_2(r) = \int_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}} M g^n_h d\theta + \frac{\lambda}{r^2} \Delta \theta M \partial_\theta g^n_h \bigg|_{\theta_{j-\frac{1}{2}}}^{\theta_{j+\frac{1}{2}}}. \] (4.77)

Notice that the weight function for \( H_1(\theta) \) is \( r M \) and the CFL condition derived from \( H_2(r) \) will be modified due to the term \( \frac{1}{r^2} \). The rest of the analysis is analogous to that for the rectangular meshes in Section 4.4.1, so details are omitted.

### 4.5 Implementation details

The fact that we only require \( g^n_h \) is nonnegative at certain points can reduce the computational cost in a great deal. To illustrate, we only present the algorithm for the one-dimensional case. Instead of finding the minimum of \( g_h \) on the whole computational cell \( I_j \), we take the minimum only on the test set \( S_j \).

Given the weighted \( L^2 \) projection \( g^0_h \) computed from the initial data \( g_0(x) \), the algorithm is stated below:

1. **Reconstruction.**

Check the point values of \( g^n_h \) on the test set \( S_j \).
If one of them is negative,

reconstruct \( \tilde{g}^n_h \) using the formula (4.47) and (4.51) and set \( g^n_h = \tilde{g}^n_h \).

2. Evolution.

Use the scheme (4.22) to compute \( g^{n+1}_h \).

This algorithm with forward Euler time discretization can be extended to higher order ODE solvers. Following Zhang and Shu (2010), we can apply the strong stability preserving Runge-Kutta method for higher order time discretizations, which are a convex linear combination of the forward Euler. The desired positivity preserving property is ensured under a suitable CFL condition.

To maintain the maximum-principle-satisfying property, we need modify the reconstruction step in the following manner.

1. Reconstruction.

Check the point values of \( g^n_h \) on the test set \( S_j \).

If one of them goes outside of \([c_1, c_2]\),

reconstruct \( \tilde{g}^n_h \) using the formula (4.55) and (4.56) and set \( g^n_h = \tilde{g}^n_h \).

4.6 Numerical tests

In this section, we will demonstrate the accuracy of the proposed numerical schemes, discuss the effects of the parameters \((\beta_0, \beta_1)\) and \(\gamma\) and show the entropy satisfying property of numerical solutions. We test on two types of problems, one is with the zero potential, such as in the scalar diffusion equation; the other is with the Hookean or the FENE spring potential arising in the dumbbell model of polymers.
4.6.1 Accuracy tests

Let \( f(t^n, x) \) and \( f_h^n(x) = M(x)g_h^n(x) \) be the exact solution and numerical approximations, respectively. We define the \( L^\infty \) and \( L^2 \) errors for \( d = 1 \) in the following way:

\[
\| f_h^n - f(t^n, \cdot) \|_{L^\infty} = \max_{1 \leq j \leq N} | f_h^n(x_j) - f(t^n, x_j) |.
\]

\[
\| f_h^n - f(t^n, \cdot) \|_{L^2} = \left( \sum_{j=1}^{N} \| f_h^n - f(t^n, \cdot) \|_{L^2}^2 \right)^{\frac{1}{2}}.
\]

These definitions may be extended to multidimensional cases in straightforward manner. When the exact solution is unavailable, we take a numerical solution with refined mesh as the reference solution. For the Gaussian quadrature rule, we choose \( L = 16 \) through all the examples.

Example 1. The heat equation.

Consider the heat equation \( f_t = f_{xx} \) which corresponds to the trivial equilibrium \( M(x) = 1 \), subject to the initial data

\[
f_0(x) = 1 + \cos(\pi x), \quad x \in [0, 1],
\]

and zero flux \( \partial_x f = 0 \) at \( x = 0 \) and \( x = 1 \). The exact solution is

\[
f(t, x) = 1 + e^{-\pi^2 t} \cos(\pi x).
\]

Table 4.1 The accuracy test of the \( P^1 \) approximation on a uniform mesh in the one-dimensional space. \( \beta_0 = 2, \gamma = 0.75 \), final time \( t = 0.5 \).

<table>
<thead>
<tr>
<th>( N )</th>
<th>( L^\infty ) error</th>
<th>order</th>
<th>( L^2 ) error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.010e-03</td>
<td>-</td>
<td>6.890e-04</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>3.022e-04</td>
<td>1.741</td>
<td>1.976e-04</td>
<td>1.802</td>
</tr>
<tr>
<td>20</td>
<td>7.928e-05</td>
<td>1.930</td>
<td>5.130e-05</td>
<td>1.945</td>
</tr>
<tr>
<td>40</td>
<td>2.019e-05</td>
<td>1.974</td>
<td>1.304e-05</td>
<td>1.977</td>
</tr>
<tr>
<td>80</td>
<td>5.193e-06</td>
<td>1.959</td>
<td>3.359e-06</td>
<td>1.956</td>
</tr>
<tr>
<td>160</td>
<td>1.431e-06</td>
<td>1.860</td>
<td>9.328e-07</td>
<td>1.848</td>
</tr>
<tr>
<td>320</td>
<td>4.891e-07</td>
<td>1.549</td>
<td>3.259e-07</td>
<td>1.517</td>
</tr>
</tbody>
</table>
Table 4.2 The accuracy test of the $P^2$ approximation on a uniform mesh in the one-dimensional space. $\beta_0 = 2, \beta_1 = 0.25, \gamma = 0.33$, final time $t = 0.1$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$L^\infty$ error</th>
<th>order</th>
<th>$L^2$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.121e-03</td>
<td>-</td>
<td>3.173e-03</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>9.367e-05</td>
<td>3.758</td>
<td>4.246e-04</td>
<td>2.901</td>
</tr>
<tr>
<td>20</td>
<td>9.813e-06</td>
<td>3.255</td>
<td>5.411e-05</td>
<td>2.972</td>
</tr>
<tr>
<td>40</td>
<td>1.568e-06</td>
<td>2.646</td>
<td>6.838e-06</td>
<td>2.984</td>
</tr>
<tr>
<td>80</td>
<td>3.362e-07</td>
<td>2.222</td>
<td>8.776e-07</td>
<td>2.962</td>
</tr>
<tr>
<td>160</td>
<td>8.098e-08</td>
<td>2.054</td>
<td>1.441e-07</td>
<td>2.607</td>
</tr>
<tr>
<td>320</td>
<td>1.981e-08</td>
<td>2.031</td>
<td>2.166e-08</td>
<td>2.733</td>
</tr>
</tbody>
</table>

Table 4.3 The accuracy test of the $P^2$ approximation for the two-dimensional case. $(\beta_0, \beta_1) = (2, 0.25), \gamma = 0.3$, final time $t = 0.1$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$L^\infty$ error</th>
<th>order</th>
<th>$L^2$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1.042e-03</td>
<td>-</td>
<td>2.237e-03</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>7.536e-05</td>
<td>3.789</td>
<td>2.993e-04</td>
<td>2.902</td>
</tr>
<tr>
<td>20</td>
<td>6.558e-06</td>
<td>3.522</td>
<td>3.812e-05</td>
<td>2.973</td>
</tr>
<tr>
<td>40</td>
<td>2.113e-06</td>
<td>1.634</td>
<td>4.885e-06</td>
<td>2.964</td>
</tr>
</tbody>
</table>

Table 4.1 and 4.2 show the optimal convergence of the scheme (4.22) for $k = 1, 2$. Table 4.3 shows the optimal order of convergence of the scheme (4.58) with the initial data

$$f_0(x, y) = 2 + \cos(\pi x) + \cos(\pi y).$$

The exact solution is

$$f(t, x, y) = 2 + e^{-\pi^2 t} [\cos(\pi x) + \cos(\pi y)].$$

**Example 2. Effects of the parameters $(\beta_0, \beta_1)$ and $\gamma$.** For the heat equation, we have in (4.53),

$$a_j = -\frac{1}{3} \quad \text{and} \quad b_j = \frac{1}{3}. \quad (4.53)$$

The restriction (4.53) on $\gamma$ is thus reduced to

$$|\gamma| \leq \min\left\{ \frac{1}{3}, 8\beta_1 - 1 \right\}. \quad (4.78)$$
This implies that for \( \frac{1}{8} \leq \beta_1 \leq \frac{1}{6} \), we need take \(|\gamma| \leq 8\beta_1 - 1\). But for \( \frac{1}{6} \leq \beta_1 \leq \frac{1}{4} \), one may take any \( \gamma \) as long as \(|\gamma| \leq \frac{1}{3}\).

Table 4.4  The effects of the parameter \( \beta_0 \) for the \( P^2 \) approximation in the one-dimensional space. \( \beta_1 = 0.25, \gamma = 0.33, \) final time \( t = 0.1 \).

<table>
<thead>
<tr>
<th>( \beta_0 )</th>
<th>( N )</th>
<th>( L^\infty ) error</th>
<th>order</th>
<th>( L^2 ) error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.000000</td>
<td>5</td>
<td>1.186e-02</td>
<td>-</td>
<td>4.144e-02</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>3.041e-03</td>
<td>1.964</td>
<td>2.061e-02</td>
<td>1.007</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>7.649e-04</td>
<td>1.991</td>
<td>1.029e-02</td>
<td>1.002</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>1.915e-04</td>
<td>1.998</td>
<td>5.146e-03</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>4.790e-05</td>
<td>1.999</td>
<td>2.573e-03</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 4.4 shows the results with the same \( \gamma = 0.33 \) but different \( \beta_0 \). We observe that \( \beta_0 \) needs to be larger than a threshold to ensure the optimal order.

We observe that \( \lambda_0 \) will decrease when \( \beta_0 \) is increasing. So we prefer to use smaller \( \beta_0 \). Table 4.5 shows that (1) when \( \gamma = 0 \), we need to enforce a more restrictive \( \Delta t \) so that the scheme can give a satisfying performance, namely, the optimal order of accuracy since \( 1e-06 \) is much smaller than the time step given by the derived CFL condition in Theorem 4.3.4; and (2) increasing \( \gamma \) can make the scheme march at a larger time step so that the computation will be faster.

We remark that \( \beta_1 \) is the key quantity that determines the range of \( \gamma \). If \( \gamma \) is already chosen satisfying (4.78), then \( \beta_1 \) can be any number satisfying

\[
\frac{1}{6} \leq \beta_1 \leq \frac{1}{4}.
\]

Our numerical tests indicate that using different \( \beta_1 \) in this range does not seem to affect the numerical performance in any noticeable way. However, for the same chosen \( \gamma \), using larger \( \beta_0 \) does improve the numerical accuracy.
Table 4.5  The effects of the parameter $\gamma$ for the $P^2$ approximation in the one-dimensional space. $\beta_1 = 0.25$, final time $t = 0.1$.

| $\beta_0 = 2, \gamma = 0, \Delta t = 1e-06$ | $N$ | $L^\infty$ error | order | $L^2$ error | order |
|---------------------------------------------|-----|------------------|-------|--------------|--|--|
| 5                                           | 1.040e-03 | -               | 3.163e-03 | -           | |
| 10                                          | 7.535e-05 | 3.787           | 4.233e-04 | 2.902       | |
| 20                                          | 6.558e-06 | 3.522           | 5.391e-05 | 2.973       | |
| 40                                          | 2.113e-06 | 1.634           | 6.908e-06 | 2.964       | |

| $\beta_0 = 2, \gamma = 0, \Delta t = 0.9h^2\lambda_0$ | $N$ | $L^\infty$ error | order | $L^2$ error | order |
|--------------------------------------------------------|-----|------------------|-------|--------------|--|--|
| 5                                                      | 7.347e-01 | -               | 4.862e-01 | -           | |
| 10                                                     | 4.754e-01 | 0.628           | 3.934e-01 | 0.305       | |
| 20                                                     | 4.959e-01 | -0.061          | 4.097e-01 | -0.058      | |
| 40                                                     | 5.830e-01 | -0.233          | 4.173e-01 | -0.026      | |

| $\beta_0 = 2, |\gamma|= 0.33, \Delta t = 0.9h^2\lambda_0$ | $N$ | $L^\infty$ error | order | $L^2$ error | order |
|-----------------------------------------------------|-----|------------------|-------|--------------|--|--|
| 5                                                   | 1.121e-03 | -               | 3.173e-03 | -           | |
| 10                                                  | 9.367e-05 | 3.581           | 4.246e-04 | 2.901       | |
| 20                                                  | 9.813e-06 | 3.255           | 5.411e-05 | 2.972       | |
| 40                                                  | 1.568e-06 | 2.646           | 6.838e-06 | 2.984       | |

**Example 3.** The dumbbell model for polymers. We consider the FENE dumbbell polymer model (4.6) which is problem (4.1) with the FENE potential (4.7).

We will test this problem in the one-dimensional case. The FENE potential is

$$U(x) = -\frac{b}{2} \log(b - |x|^2), \quad x \in B(0, \sqrt{b}).$$

We also test the Hookean potential

$$U(x) = \frac{|x|^2}{2}.$$

In order to compare with the FENE case, this test will also set in the ball $B(0, \sqrt{b})$, subject to the zero flux boundary condition.

We take the following initial data

$$f_0(x) = \left(1 - \frac{|x|^2}{b}\right)^\frac{3b}{4}, \quad x \in [-\sqrt{b}, \sqrt{b}].$$

Table 4.6 and 4.7 show the optimal order of the convergence with $k = 1, 2$ for the one-dimensional dumbbell model.
Table 4.6 The accuracy test of the $P^1$ approximation on a uniform mesh for the one-dimensional case. $b = 36, \beta_0 = 2$, final time $t = 0.1$.

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>$L^\infty$ error</th>
<th>order</th>
<th>$L^2$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>FENE</td>
<td>5</td>
<td>3.172e-02</td>
<td>-</td>
<td>3.938e-02</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>6.315e-03</td>
<td>2.329</td>
<td>9.670e-03</td>
<td>2.026</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>2.821e-03</td>
<td>1.162</td>
<td>2.806e-03</td>
<td>1.785</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>8.808e-04</td>
<td>1.679</td>
<td>7.146e-04</td>
<td>1.973</td>
</tr>
</tbody>
</table>

Table 4.7 The accuracy test of the $P^2$ approximation on a uniform mesh for the one-dimensional case. $b = 36, (\beta_0, \beta_1) = (2, 0.25)$, final time $t = 0.1$.

<table>
<thead>
<tr>
<th></th>
<th>$N$</th>
<th>$L^\infty$ error</th>
<th>order</th>
<th>$L^2$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>FENE</td>
<td>5</td>
<td>2.141e-02</td>
<td>-</td>
<td>1.232e-01</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>5.517e-04</td>
<td>5.278</td>
<td>2.790e-03</td>
<td>5.464</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>7.899e-05</td>
<td>2.804</td>
<td>5.194e-04</td>
<td>2.425</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>7.375e-06</td>
<td>3.421</td>
<td>7.075e-05</td>
<td>2.876</td>
</tr>
</tbody>
</table>

From Table 4.8, we observe that the discrete relative entropy

$$E_h(t^n) = \sum_j \int_{I_j} M(g^n_h)^2 \, dx$$

is decreasing to 1, which indicates the convergence of the numerical solution to the equilibrium.

4.7 Concluding remarks

In this paper, we have proposed up to third order DG schemes which can be proven maximum-principle-satisfying for linear Fokker–Planck equations. We also show an extension
Table 4.8  The entropy satisfying property of the $P^2$ approximation on a uniform mesh with the FENE spring potential. $N = 20, b = 36, (\beta_0, \beta_1) = (2, 0.25)$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>0</th>
<th>0.1</th>
<th>1</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_h(t)$</td>
<td>1.16263</td>
<td>1.10463</td>
<td>1.00267</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

to two dimensions. Moreover, the same idea applies to the two-dimensional Polymer models on a disk. We have tested both second and third order DG schemes and clearly observed the strict maximum principle in all these tests. Even though the CFL condition derived to preserve maximum principle is very small compared to the ones for conventional finite volume schemes, we emphasize that it is not a necessary condition. To save computational costs, one can strictly enforce the CFL conditions only when a precalculation with a usual time step produces overshoot or undershoot.

The scheme in this paper can not be extended to fourth or even higher order in a straightforward way. Even higher order maximum-principle-satisfying schemes for solving Fokker–Planck equations will be explored in the future.

Acknowledgments

This research was supported by the National Science Foundation under grant DMS09-07963.
CHAPTER 5. ENTROPY/ENERGY STABLE SCHEMES FOR EVOLUTIONARY DISPERsal MODELS

A paper submitted to JCP

Hailiang Liu and Hui Yu

Abstract

In this paper we propose some entropy/energy stable finite difference schemes for the reaction-diffusion-advection equation arising in the evolution of biased dispersal of population dynamics. The peculiar feature of these active dispersal models is that the transient solution converges to the stable steady state when time goes to infinity. For the numerical method to capture the long-time pattern of persistence or extinction, we use the relative entropy when the resource potential is logarithmic and explore the usual energy for other resource potentials. The present schemes are shown to satisfy three important properties of the continuous model for the population density: (i) positivity preserving, (ii) equilibrium preserving and (iii) entropy or energy satisfying. These ensure that our schemes provide a satisfying long-time behavior, thus revealing the desired dispersal pattern. Moreover, we present several numerical results which confirm the second-order accuracy for various resource potentials and underline the efficiency to preserve the large time asymptotic.

5.1 Introduction

This work is concerned with the numerical approximation of a class of reaction-diffusion-advection equations arising in the evolution of biased dispersal of population dynamics, with
emphasis on exploring the entropy/energy structure of the dispersal model so that the resulting methods provide a satisfying long-time behavior.

5.1.1 Mathematical formulations

Reaction-diffusion equations have been widely used to model the biological problems [Okubo (1980), Murray (2003)]. One of the well known examples is the logistic reaction-diffusion model for the population growth with random dispersal,

$$\partial_t u = \Delta u + \lambda u(m - u) \quad \text{in } \Omega \times (0, \infty),$$

(5.1)

where the population inhabiting a bounded domain $\Omega \subset \mathbb{R}^d$ has density $u(x,t)$ at location $x$ and time $t$ and local growth rate $m(x)$. The parameter $\lambda > 0$ is the reverse of the dispersal rate.

If the environment is heterogeneous, i.e., $m(x)$ is not a constant, then the population may have a tendency to move toward resources in addition to the random movement. The model may be upgraded to the following form

$$\partial_t u = \nabla \cdot (\nabla u + u\nabla P) + \lambda u(m - u) \quad \text{in } \Omega \times (0, \infty),$$

(5.2)

where $P = P(m)$, which we call resource potential, reflects the movement tendency of the population. The time evolution is subjected to both the initial density $u(x,0) = u_0(x)$ and the zero flux boundary condition

$$(\nabla u + u\nabla P) \cdot \nu = 0 \quad \text{on } \partial \Omega \times (0, \infty),$$

(5.3)

where $\nu$ is the outward normal vector on the boundary $\partial \Omega$ which is assumed to be smooth. Several dispersal strategies have been studied in literature, for example $P = -\alpha m$ in Belgacem and Cosner (1995), Cosner and Lou (2003), Cantrell et al. (2006), Chen et al. (2008) and Bezuglyy and Lou (2010), and the obtained results may apply to a more general reaction than

$$F(u,m) = u(m - u),$$

as long as it satisfies $F(0,m) = 0$ and $F(+\infty,m) < 0$. With the logistical reaction term or its variation, the biologically relevant solutions to (5.2) are nonnegative and ultimately bounded.
It follows from the regularity theory of parabolic partial differential equations that in the state space bounded orbits are pre-compact, hence the semiflow will have a compact attractor and the \( \omega \)-limit set of any initial state \( u_0(x) \) will be a compact invariant set. In other words, as time evolves the solution of (5.2) is expected to approach some stable patterns, independent of the choice of initial density; see, e.g., Cantrell and Cosner (2004). The main result reviewed in Section 5.2 may be stated as follows.

**Theorem 5.1.1.** Suppose that \( m \in C^2(\bar{\Omega}) \) is positive somewhere in \( \Omega \) and \( P \) is smooth in \( m \). There exists a unique \( \lambda^* > 0 \) and a positive equilibrium solution to (5.2) if and only if

\[
\int_{\Omega} me^{-P} dx < 0.
\]

Moreover, \( \lambda^* = 0 \) if and only if \( \int_{\Omega} me^{-P} dx \geq 0 \).

1. If \( 0 < \lambda \leq \lambda^* \), all nonnegative solutions of (5.2) decay toward zero as \( t \to \infty \).

2. If \( \lambda > \lambda^* \), the positive equilibrium is globally attractive among nonzero nonnegative solutions.

Of special interest is the ideal free dispersal strategy determined by \( P \) such that at an equilibrium \( u_{eq}(x) \) [Stephen Cantrell et al. (2007); Cantrell et al. (2008, 2010)], both

\[
\nabla \cdot (\nabla u_{eq} + u_{eq} \nabla P) = 0 \quad \text{and} \quad m - u_{eq} = 0 \quad \text{in } \Omega.
\]

This will hold if \( P = -\log m \) (unique up to some constant). In such a setting, the species \( u \) can perfectly match the environmental resource. The role of ideal free distributions has been well recognized in literature. Moreover, the quantity

\[
E[u] = \int_{\Omega} \left[ u \log \frac{u}{m} + m - u \right] dx \quad (5.4)
\]

is nonincreasing in time. This corresponds to some physical entropy relative to the equilibrium state \( u_{eq}(x) = m(x) \), called relative entropy. For general resource potential \( P \), the quantity

\[
V[u] = \int_{\Omega} \left[ \frac{1}{2} e^{-P} |\nabla (ue^P)|^2 - \lambda e^P \int_{\Omega} F(\xi, m) d\xi \right] dx \quad (5.5)
\]
is nonincreasing in time. Moreover, equation (5.2) can be rewritten as a gradient flow,
\[
\partial_t u = -e^{-p \frac{\delta V}{\delta u}},
\]  
where \( \frac{\delta V}{\delta u} \) denotes the standard variational derivative of \( V \) with respect to \( u \) and the functional \( V \) often corresponds to some free energy of the underlying physics which thus is called the energy.

In the context of two competing species, the evolutionary dynamics of conditional dispersal becomes much more complex; see Dockery et al. (1998) and Chen et al. (2008). A typical model may be described as
\[
\begin{aligned}
\frac{\partial u}{\partial t} &= \alpha \nabla \cdot (\nabla u + u \nabla P) + u(m - u - v) \quad \text{in} \quad \Omega, \\
\frac{\partial v}{\partial t} &= \beta \nabla \cdot (\nabla v + v \nabla Q) + v(m - u - v) \quad \text{in} \quad \Omega, \\
\partial_\nu u + u \partial_\nu P &= \partial_\nu v + v \partial_\nu Q = 0 \quad \text{on} \quad \partial \Omega, \\
u(x, 0) &= u_0(x), \quad v(x, 0) = v_0(x) \quad \text{in} \quad \Omega.
\end{aligned}
\]  
(5.7)
The two nonnegative constants \( \alpha \) and \( \beta \) represent the dispersal rates of two species respectively. An interesting question is whether there is any strategy in system (5.7) that is global evolutionarily stable, in the sense that the semitrivial steady state \((m, 0)\) is always locally stable. This is the case if the dispersal strategy is ideal free at \((m, 0)\), i.e., \( P(m) = -\log m + C \) and \( Q(m) \neq P(m) \); see Cantrell et al. (2010) and Gejji et al. (2011). Moreover, the following functional
\[
E[u, v] = \int_\Omega u \, dx + \int_\Omega v \, dx - \int_\Omega m(x) \log u \, dx
\]  
(5.8)
is nonincreasing in time.

For other results on dispersal strategies in heterogeneous landscapes, we refer to Chen et al. (2008), Lam and Ni (2010) and Lam (2011) for the study of aggregation profiles in terms of the resource distribution and Coville et al. (2008) and Kao et al. (2010) for nonlocal dispersal strategies.

The aim of this paper is to give reliable numerical schemes for (5.2) and (5.7) from the perspective of providing a satisfying long time behavior. A key fact is that they both admit certain entropy/energy structure and we demand our numerical schemes to satisfy the entropy/energy decreasing property in the discrete setting.
In recent years, such energy preservation or dissipation numerical schemes have drawn much interest and been extensively studied for several PDEs, since they are more likely to give a better performance in the long time simulation; see Furihata (1999), Carrillo et al. (2008), Masmoudi (2008), Matsuo and Yamaguchi (2009), Celledoni et al. (2012) and Liu and Yu (2012a). To the best of our knowledge, however, no entropy stable scheme for (5.2) or (5.7) has yet appeared in the literature. Though, due to the physical and mathematical significance of advection-diffusion-reaction equations, numerical methods have been developed and analyzed through decades; see, e.g., Bermejo and El Amrani (2003), Bermejo and Carpio (2008), Lubich and Ostermann (1995), Lanser and Verwer (1999) and the recent monograph Hundsdorfer and Verwer (2003).

5.1.2 Main steps of this paper

In Section 5.2 mathematical theory and properties regarding the target equation (5.2) are summarized; we particularly highlight the energy/entropy structure, as well as the large time pattern formation in terms of the parameter $\lambda$. The key idea is to reformulate (5.2) and (5.7) into the nonlogarithmic Landau form in terms of $g = ue^P$ and $h = ve^Q$, following the strategy in Liu and Yu (2012a) for the Fokker–Planck equation of FENE dumbbell polymers.

Section 5.3 is devoted to the case with the logarithmic potential and the proposed scheme is presented. The schemes at both semidiscrete and fully discrete levels are shown to satisfy three important properties of the continuous model for the population density: (i) positivity preserving, (ii) equilibrium preserving and (iii) entropy satisfying. The schemes presented in Section 5.4 apply to the case with general resource potential and are extensible to handle more general reaction terms. In that section we first present a semidiscrete scheme with discretization only in time. This scheme is proved to be uniquely solvable provided the time step satisfies a bound depending on the growth rate $m$ and the numerical solution at the previous time step. The scheme preserves the equilibrium solution with numerical energy being nonincreasing in time for arbitrary time step. In addition, the scheme can be expressed as a second-order prediction-correction method and the prediction is shown to be in the same range of the numerical solution at the previous time step. As a consequence, we prove that the fully discrete scheme satisfies
both desired properties: energy satisfying and equilibrium preserving. We also propose an entropy stable scheme in Section 5.5 to simulate the ideal free distribution for two competing species. Section 5.6 is devoted to numerical tests of proposed schemes. Finally some concluding remarks are presented in Section 5.7.

5.2 Model reformulation and mathematical properties

In order to reveal the dynamic picture of (5.2), following Liu and Yu (2012a), we rewrite (5.2) in terms of \( g = u/M \) with \( M = e^{-P} \):

\[
M \partial_t g + P g = 0 \quad \text{in } \Omega, \tag{5.9}
\]

where the operator \( P \) is defined as

\[
P g = -\nabla \cdot (M \nabla g) - \lambda Mg(m - Mg). \tag{5.10}
\]

The initial condition is \( g_0(x) = \frac{u_0(x)}{M(x)} \) and on the boundary \( \partial \Omega \) we have \( M \partial_n g = 0 \). The reaction term becomes \( F(g, m) = Mg(m - Mg) \).

5.2.1 Entropy for log potential \( P(m) = -\log m \).

In such a case \( M = e^{-P} = m \) and (5.9) reduces to

\[
m \partial_t g = \nabla \cdot (m \nabla g) + \lambda m^2 g(1 - g) \quad \text{in } \Omega. \tag{5.11}
\]

The logistic reaction term ensures that any solution remains nonnegative and bounded, i.e.,

\[
0 \leq g(x, t) \leq \max\{1, \|g_0\|_{\infty}\} \quad \forall (x, t) \in \Omega \times [0, \infty).
\]

As \( t \to \infty \), \( g \) will converge to the stable equilibrium solution \( \tilde{g} \equiv 1 \), independent of the choice of the initial data. This motivates us to explore the entropy which might have some biological significance.

The entropy defined in (5.4) becomes

\[
E[g] = \int_{\Omega} m[g \log g + (1 - g)] dx.
\]
As a convex functional of $g$, it defines the entropy structure of (5.11):

$$m\partial_t g = \nabla \cdot \left( mg \nabla \left( m^{-1} \frac{\delta E}{\delta g} \right) \right) + \lambda F(g, m),$$

where the variational derivative $\frac{\delta E}{\delta g} = m \log g$. The reaction term $F(g, m) = m^2 g(1 - g)$ satisfies

$$F \cdot \frac{\delta E}{\delta g} \leq 0$$

for nonnegative $g$. Using integration by parts, we have

$$\frac{d}{dt} E = -\int_{\Omega} m g \left\| \nabla \left( m^{-1} \frac{\delta E}{\delta g} \right) \right\|^2 dx + \lambda \int_{\Omega} m^{-1} F \cdot \frac{\delta E}{\delta g} dx \leq 0,$$

where the zero flux boundary condition has been taken into account. We shall propose a numerical method to satisfy this entropy property.

5.2.2 Energy for all other potential

In general, the evolution may lead to extinction or persistence of the species and the concept of entropy is no longer adequate to control the reaction term. We shall follow the standard gradient flow idea [Cantrell and Cosner (2004)] by using the functional defined in (5.5),

$$V[g] = \int_{\Omega} \left[ \frac{1}{2} M |\nabla g|^2 - G(x, g(x, t)) \right] dx, \quad (5.12)$$

where $G$ is chosen as

$$G(x, g) = \lambda \int_{0}^{g} F(s, m) ds = \frac{\lambda}{2} m(x) M(x) g^2 - \frac{\lambda}{3} M^2(x) g^3.$$  

So (5.9) can be rewritten as

$$M\partial_t g = \nabla \cdot (M \nabla g) + \partial G.$$  

(5.13)

The above energy also defines the gradient flow structure of (5.13):

$$M\partial_t g = -\frac{\delta V}{\delta g},$$

where the zero flux has been used in the variational derivative. A formal calculation gives

$$\frac{d}{dt} V = \int_{\Omega} \frac{\delta V}{\delta g} \partial_t g dx = -\int_{\Omega} M(\partial_t g)^2 dx \leq 0.$$  

(5.14)

We shall propose a numerical method to satisfy this energy dissipation inequality in the discrete setting.
5.2.3 Large time pattern formation

The nonnegative equilibrium solution \( \bar{g}(x) \), if it exists, solves

\[
P \bar{g} = 0 \text{ in } \Omega \quad \text{and} \quad M \partial_{\nu} \bar{g} |_{\partial \Omega} = 0. \tag{5.15}
\]

Let \( g = \bar{g} + \varepsilon w \) in (5.9) and then send \( \varepsilon \to 0 \) to obtain the linearized problem

\[
M \partial_t w = \nabla \cdot (M \nabla w) + \lambda Mw(m - 2M \bar{g}) \text{ in } \Omega \quad \text{and} \quad M \partial_{\nu} w |_{\partial \Omega} = 0. \tag{5.16}
\]

It is known that if \( \bar{g} \) is linearly stable, then \( \bar{g} \) is stable; for a proof see, e.g., Theorem 11.22 in Smoller (1983). The linear stability of \( \bar{g} \) is determined by the spectrum of the linearized operator of \( P \). For \( \bar{g} = 0 \), it boils down to the eigenvalue problem:

\[
\nabla \cdot (M \nabla \phi) + \lambda m M \phi = \sigma M \phi \text{ in } \Omega \quad \text{and} \quad M \partial_{\nu} \phi |_{\partial \Omega} = 0. \tag{5.17}
\]

When \( \sigma < 0 \), \( 0 \) is stable; when \( \sigma > 0 \), \( 0 \) is unstable. Hence \( \sigma = 0 \) is a threshold value, corresponding to a special parameter \( \lambda^* \) which satisfies

\[
\nabla \cdot (M \nabla \psi) + \lambda^* m M \psi = 0 \text{ in } \Omega \quad \text{and} \quad M \partial_{\nu} \psi |_{\partial \Omega} = 0. \tag{5.18}
\]

In what follows we shall denote the weighted \( L^2 \) norm by \( \|v\|_M \) with

\[
\|v\|^2_M = \int_{\Omega} M v^2 \, dx.
\]

We now summarize the basic results for (5.9), (5.15)-(5.18). These results are known for some specific choices of \( P \) such as \( P = -\alpha m \) [Belgacem and Cosner (1995) and Cantrell and Cosner (2004)]. For the sake of completeness, we present here the main results for (5.9) and highlight the key points in proofs.

**Theorem 5.2.1.** Problem (5.18) has a unique \( \lambda^* > 0 \) characterized by a positive eigenfunction if and only if

\[
\int_{\Omega} m M \, dx < 0.
\]

Moreover, \( \lambda^* = 0 \) if and only if \( \int_{\Omega} m M \, dx \geq 0 \).
Proof. The existence of \((\lambda^*, \psi)\) with \(\psi > 0\) in \(\Omega\) and \(M = 1\) is given in Brown and Lin (1980). In a similar manner, one can prove the existence of \((\lambda^*, \psi)\) with \(\psi > 0\) and nonconstant \(M > 0\). Also, \(\lambda^*\) can be expressed by the variational formulation

\[
\lambda^* = \inf_{v \in S} \frac{\|\nabla v\|_M^2}{\int_{\Omega} MMv^2 \, dx},
\]

where

\[
S = \{ v \in H^1 : M\partial_\nu v = 0 \text{ on } \partial\Omega \text{ and } \int_{\Omega} MMv^2 \, dx > 0 \}.
\]

In fact, (5.18) yields

\[
\lambda^* = \frac{\|\nabla \psi\|_M^2}{\int_{\Omega} MM\psi^2 \, dx} \geq 0. \tag{5.19}
\]

Divide both sides of (5.18) by \(\psi\) and integrate over \(\Omega\) to obtain

\[
\lambda^* \int_{\Omega} MM \, dx = -\int_{\Omega} \frac{M\|\nabla \psi\|_M^2}{\psi^2} \, dx \leq 0. \tag{5.20}
\]

For \(\int_{\Omega} MM \, dx < 0\), (5.20) implies that \(\lambda^* > 0\). For \(\int_{\Omega} MM \, dx = 0\), (5.20) implies that \(\|\nabla \psi\|_M^2 = 0\). Therefore \(\lambda^* = 0\) by (5.19). For \(\int_{\Omega} MM \, dx > 0\), we must have \(\lambda^* = 0\) by (5.20). \(\square\)

Let \((\sigma^*, \phi)\) denote the pair of the principle eigenvalue and eigenfunction of (5.17). We thus have \(\phi > 0\) in \(\Omega\); see Theorem 11.10 in Smoller (1983). The following result relates the sign of \(\sigma^*\) to the relative size of \(\lambda\) in terms of \(\lambda^*\).

**Theorem 5.2.2.** \(\text{sign} (\sigma^*) = \text{sign} (\lambda - \lambda^*)\).

**Proof.** Rewrite (5.17) as

\[
\nabla \cdot (M \nabla \phi) + \lambda \left( m - \frac{\sigma^*}{\lambda} \right) M\phi = 0, \quad M\partial_\nu \phi|_{\partial\Omega} = 0. \tag{5.21}
\]

Let \(\tilde{m} = m - \frac{\sigma^*}{\lambda}\). Then \(\lambda\) is determined by

\[
\lambda = \inf_{\tilde{S}} \frac{\|\nabla v\|_{\tilde{M}}^2}{\int_{\Omega} \tilde{m}Mv^2 \, dx},
\]

where \(\tilde{S}\) is a modification of \(S\) with \(m\) replaced by \(\tilde{m}\). If \(\sigma^* = 0\), then \(\phi = \psi\) and \(\lambda = \lambda^*\) by the simpleness of the principle eigenvalue. If \(\sigma^* > 0\), then \(\tilde{m} < m\) and

\[
0 < \int_{\Omega} \tilde{m}M\phi^2 \, dx < \int_{\Omega} MM\phi^2 \, dx.
\]
Therefore
\[ \lambda = \frac{\|\nabla \phi\|_M^2}{\int_{\Omega} m M \phi^2 \, dx} > \frac{\|\nabla \phi\|_M^2}{\int_{\Omega} m M \phi^2 \, dx} \geq \lambda^*. \]

If \( \sigma^* < 0 \), then \( \tilde{m} > m \) and
\[ \int_{\Omega} \tilde{m} M \psi^2 \, dx > \int_{\Omega} m M \psi^2 \, dx > 0. \]
Hence,
\[ \lambda \leq \frac{\|\nabla \psi\|_M^2}{\int_{\Omega} \tilde{m} M \psi^2 \, dx} < \frac{\|\nabla \psi\|_M^2}{\int_{\Omega} m M \psi^2 \, dx} = \lambda^*. \]

The result about the equilibrium solution \( \tilde{g} \) is the following.

**Theorem 5.2.3.** Suppose that \( m \in C^2(\overline{\Omega}) \) is positive somewhere in \( \Omega \) and \( P \) is bounded.

1. If \( 0 < \lambda \leq \lambda^* \), (5.15) only has a zero solution.

2. If \( \lambda > \lambda^* \), (5.15) has a zero solution and a unique positive solution.

**Proof.** (1) For \( 0 < \lambda \leq \lambda^* \), multiply (5.15) by \( \tilde{g} \) to get
\[ \lambda \int_{\Omega} m M \tilde{g}^2 \, dx = \int_{\Omega} M |\nabla \tilde{g}|^2 \, dx + \lambda \int_{\Omega} M^2 \tilde{g}^2 \, dx \geq 0. \]
By the definition of \( \lambda^* \), we have
\[ \lambda \int_{\Omega} m M \tilde{g}^2 \, dx \geq \lambda^* \int_{\Omega} m M \tilde{g}^2 \, dx + \lambda \int_{\Omega} M^2 \tilde{g}^2 \, dx. \]
It follows that
\[ \lambda \int_{\Omega} M^2 \tilde{g}^2 \, dx \leq (\lambda - \lambda^*) \int_{\Omega} m M \tilde{g}^2 \, dx \leq 0. \]
Therefore \( \tilde{g} \equiv 0 \) on \( \Omega \).

(2) One can verify that any constant \( \tilde{g} \) greater than \( \frac{m(x)}{M(x)} \) is a supersolution of (5.15).

We normalize the positive eigenfunction \( \phi \) such that \( \|\phi\|_\infty = 1 \). Then \( \tilde{g} = \varepsilon \phi \) with
\[ 0 < \varepsilon \leq \min \left\{ \frac{\sigma^*}{\lambda \|M\|_\infty}, \tilde{g} \right\} \]
is a subsolution of (5.15) because
\[ P_g = \varepsilon M \phi (\varepsilon \lambda M \phi - \sigma^*) \leq \varepsilon M \phi (\varepsilon \lambda M - \sigma^*) \leq 0 \text{ in } \Omega \quad \text{and} \quad M \partial_\nu g|_{\partial \Omega} = 0. \]

Moreover,
\[ 0 < g \leq \bar{g} \quad \text{and} \quad \frac{\partial}{\partial g} \left( \frac{\lambda F(g,m)}{g} \right) = -\lambda M^2 < 0. \]

So the conditions of Theorem 3.4 in Chapter 3 of Pao (1992) are satisfied and it ensures that there exists a unique solution \( \bar{g} \) of (5.15) such that \( \underline{g} \leq \bar{g} \leq \bar{g} \).

The above claimed uniqueness is actually restricted to the situation that any two solutions, if \( \bar{g}_1 \leq \bar{g}_2 \), must coincide. For the general case, we define the minimum of two arbitrary positive solutions as
\[ g^*(x) = \min\{\bar{g}_1(x), \bar{g}_2(x)\}. \]

A slight modification of the argument in Lemma 1.5 of Dancer and Sweers (1989) when applied to (5.15) (care is needed for treating the Neumann boundary condition) enables us to show that \( g^* \) is a weak supersolution. Subtracting the weak formulations of \( g^* \) and \( \bar{g}_i \) against the test function taken as \( \bar{g}_i \) and \( g^* \) respectively, we have
\[ \lambda \int_\Omega M^2 g^*(\bar{g}_i - g^*) \, dx \leq 0 \quad \text{for } i = 1, 2, \]
which implies that \( g^* = \bar{g}_i \) in \( \Omega \), hence \( \bar{g}_1 = \bar{g}_2 \).

The above results when combined with the nonincreasing property of the energy \( V \) defined in (5.12) enable us to predict persistence or extinction of the evolving species; see Cantrell and Cosner (2004) for more background details.

**Theorem 5.2.4.** Under the same assumptions as in Theorem 5.2.3, we have:

1. If \( 0 < \lambda \leq \lambda^* \), all nonnegative solutions of (5.9) decay toward zero as \( t \to \infty \).

2. If \( \lambda > \lambda^* \), the positive equilibrium is globally attractive among nonzero nonnegative solutions.
Proof. We first prove the convergence to the steady state. The discussion on $V$ in Section 5.2.2 shows that
\[
\frac{d}{dt}V = - \int_\Omega M(\partial_t g)^2 \, dx \leq 0.
\] (5.22)
Thus $V$ is nonincreasing along the orbit starting at $g_0(x)$. The boundedness and precompactness of the orbits imply that $V[g]$ is bounded from below, so there exists $V_{eq} = \lim_{t \to \infty} V[g]$. If $w$ is in the $\omega$-limit set of $g_0(x)$, then $V[w] = V_{eq}$. Let $w^*(x,t)$ be the orbit starting at some point $w$ in the $\omega$-limit set of $g_0(x)$. Since that set is invariant, the entire orbit $w^*(x,t)$ must belong to it, so $V[w^*] = V_{eq}$. In view of (5.22), this is possible if and only if $\partial_t w^* = 0$. Therefore, $w$ must be an equilibrium of (5.9). Since $w$ is an arbitrary element of the $\omega$-limit set of an arbitrary initial point $g_0$, the $\omega$-limit set for the semiflow generated by (5.9) must consist of all the equilibria, which shows the global attractiveness of stable equilibrium solutions.

We now distinguish two cases in terms of $\lambda$. The claimed results follow from the fact that zero equilibrium solution is stable for $0 < \lambda \leq \lambda^*$ and unstable for $\lambda > \lambda^*$, as inferred from Theorem 5.2.2 and 5.2.3.

\[\square\]

5.3 Entropy stable method for one-species model

5.3.1 Semidiscrete scheme

We present only the one-dimensional case. Generalizations to the multidimensional space are straightforward for tensor product grids and the results remain valid without modifications. Given an integer $N$, we partition the domain $\Omega := [a,b]$ by using a uniform mesh $\Delta x$, where
\[
\Delta x = \frac{b - a}{N} \quad \text{and} \quad x_j = a + (j - \frac{1}{2})\Delta x.
\]
Note that $x_{\frac{1}{2}} = a$ and $x_{N+\frac{1}{2}} = b$. Let $g_j(t)$ be the approximation of $g(x_j,t)$. We propose the following semidiscrete scheme: for $1 \leq j \leq N$,
\[
m_j \frac{d}{dt}g_j = D_-(m_{j+\frac{1}{2}}D_+g_j) + \lambda m_j^2g_j(1 - g_j) \quad \text{and} \quad g_j(0) = g_0(x_j), \tag{5.23}
\]
where
\[
m_j = m(x_j), \quad m_{j+\frac{1}{2}} = m(x_{j+\frac{1}{2}})
\]
and the two difference operators are

$$D_u \rightarrow u_j = \frac{u_j - u_{j-1}}{\Delta x}, \quad D_u \rightarrow u_j = \frac{u_{j+1} - u_j}{\Delta x}. $$

In order to incorporate the zero flux boundary condition, we force

$$m \frac{1}{2} D_+ g_0 = m_{N+\frac{1}{2}} D_+ g_N = 0. \quad (5.24)$$

**Theorem 5.3.1.** The semidiscrete scheme (5.23) possesses three properties:

1. If \( g_j(0) \geq 0 \) \( \forall j \), then
   \[
   0 \leq g_j(t) \leq K \quad \forall j \text{ and } \forall t > 0,
   \]
   where \( K = \max \{1, \max_{1 \leq j \leq N} g_j(0)\} \).

2. The semidiscrete entropy
   \[
   E_h(t) = \sum_{j=1}^{N} \left( \sum_{j=1}^{N} m_j \left[ g_j \log g_j + (1 - g_j) \right] \Delta x \right)
   \]
   is nonincreasing as time evolves.

3. It preserves the equilibrium solution \( \bar{g} \equiv 1 \); i.e., if \( g_j(0) = 1 \), then
   \[
   g_j(t) = 1 \quad \forall t > 0, \quad 1 \leq j \leq N.
   \]

**Proof.** (1) Define

\[
\Sigma = \{ \bar{g}: g_j \in [0, K], 1 \leq j \leq N \} \subset \mathbb{R}^N.
\]

The claimed bounds follow if we show that \( \Sigma \) is an invariant region. Since \( \bar{0} \) is an equilibrium solution of the ODE system (5.23), it suffices to prove that

\[
m \frac{d\bar{g}}{dt} \cdot \nu < 0 \quad \forall \bar{g} \in \partial \Sigma \setminus \{\bar{0}\},
\]

where \( \nu \) is the outward normal vector at \( \bar{g} \). For each \( \bar{g} \), we define two sets of indices \( s_0 \) and \( s_K \) such that

\[
s_0 = \{1 \leq j \leq N; g_j = 0\}, \quad s_K = \{1 \leq j \leq N; g_j = K\}.
\]

Note that \( s_0 \) and \( s_K \) cannot be empty at the same time. So there are only two cases:
(a) $s_0 \neq \emptyset$. Then there exists $j$ such that
\[ g_j = 0, \quad g_{j+1} > 0 \quad \text{or} \quad g_j > 0, \quad g_{j+1} = 0, \]
and
\[ \nu_j = \begin{cases} -\alpha_j < 0, & \text{if } j \in s_0, \\ 0, & \text{if } j \notin s_0. \end{cases} \]
Therefore,
\[ m \frac{d}{dt} \bar{g} \cdot \nu = -\sum_{j \in s_0} \frac{m_{j+\frac{1}{2}} g_{j+1} + m_{j-\frac{1}{2}} g_j}{(\Delta x)^2} \alpha_j < 0. \]

(b) $s_0 = \emptyset$; then $s_K \neq \emptyset$. The normal vector $\nu$ is in the following form
\[ \nu_j = \begin{cases} \alpha_j > 0, & \text{if } j \in s_K, \\ 0, & \text{if } j \notin s_K. \end{cases} \]
It follows that
\[ m \frac{d}{dt} \bar{g} \cdot \nu = \sum_{j \in s_K} D_-(m_{j+\frac{1}{2}} D_+ g_j) + \lambda m_j^2 g_j (1 - g_j) < 0. \]
So $\Sigma$ is an invariant region. In other words, the scheme preserves the positivity of the solution and $g_j \leq K \forall j$.

(2) The time derivative of the entropy $E_h(t)$ is given by
\[ \frac{dE_h(t)}{dt} = \sum_{j=1}^{N} m_j \frac{d}{dt} g_j \log g_j \Delta x \]
\[ = \sum_{j=1}^{N} \left[ D_-(m_{j+\frac{1}{2}} D_+ g_j) + \lambda m_j^2 g_j (1 - g_j) \right] \log g_j \Delta x \]
\[ = -\sum_{j=1}^{N-1} m_{j+\frac{1}{2}} D_+ g_j (\log g_{j+1} - \log g_j) - \lambda \sum_{j=1}^{N} m_j^2 g_j (g_j - 1) \log g_j \Delta x \]
\[ \leq 0, \]
where we used the zero flux boundary condition and $(X - Y) \log \frac{X}{Y} \geq 0$ for $X, Y > 0$.

(3) If $g_j(0) = 1$ for $1 \leq j \leq N$, then $E_h(0) = 0$. So $E_h(t) \leq 0$ by (2). However $E_h(t) \geq 0 \forall t \geq 0$. Hence $E_h(t) = 0$, implying that $g_j$ must be 1 for $1 \leq j \leq N$. \qed
5.3.2 Fully discrete scheme

To preserve both positivity and entropy property at fully discrete level, we propose the following time discretization:

\[ m_j \frac{g_j^{n+1} - g_j^n}{\Delta t} = D_-(m_{j+\frac{1}{2}}D_+g_j^{n+1}) + \lambda m_j^2 g_j^n(1 - g_j^{n+1}) \text{ for } 1 \leq j \leq N, \tag{5.25} \]

subject to initial data

\[ g_j^0 = g_0(x_j), \quad 1 \leq j \leq N, \]

and the zero flux boundary condition

\[ m_{\frac{1}{2}}D_+g_0^{n+1} = m_{N+\frac{1}{2}}D_+g_N^{n+1} = 0. \]

The scheme enjoys the following properties.

**Theorem 5.3.2.** The fully discrete scheme (5.25) is unconditionally stable in the sense that:

1. If \( g_j^0 \geq 0 \forall j \), then for any \( n \geq 1 \),

\[ 0 \leq g_j^n \leq K, \quad 1 \leq j \leq N, \]

where \( K = \max\{1, \max_{1 \leq j \leq N} g_j^0\} \).

2. The fully discrete entropy

\[ E^n = \sum_{j=1}^{N} m_j [g_j^n \log g_j^n + (1 - g_j^n)] \Delta x \tag{5.26} \]

is nonincreasing, i.e., \( E^{n+1} \leq E^n \forall n \geq 0 \).

3. It preserves the equilibrium solution \( \tilde{g} \equiv 1 \); i.e., if \( g_j^0 = 1 \), then \( g_j^n = 1 \forall n \geq 1 \).

**Proof.** (1) Rearranging (5.25), we have

\[ - \Delta t m_{j-\frac{1}{2}} g_{j-1}^{n+1} + [\Delta t(m_{j-\frac{1}{2}} + m_{j+\frac{1}{2}}) + (\Delta x)^2 m_j(1 + \lambda \Delta tm_j g_j^n)]g_j^{n+1} - \Delta t m_{j+\frac{1}{2}} g_j^{n+1} \]

\[ = (\Delta x)^2 m_j(1 + \Delta t \lambda m_j) g_j^n. \tag{5.27} \]
This is a tridiagonal linear system with the coefficient matrix being diagonally dominant. So \( \{g_j^{n+1}\} \) is nonnegative, as long as \( \{g_j^n\} \) is nonnegative. It also implies the existence and uniqueness of the numerical solution.

As to the upper bound, it suffices to show that if \( g_j^n \leq K \) \( \forall j \), then \( g_j^{n+1} \leq K \) \( \forall j \). Let \( g_i^{n+1} = \max_{1 \leq j \leq N} g_j^{n+1} \). Then the \( i \)-th row of the linear system gives

\[
(\Delta x)^2 m_i (1 + \Delta t \lambda m_i) g_i^n \geq (\Delta x)^2 m_i (1 + \Delta t \lambda m_i g_i^n) g_i^{n+1}.
\]

Therefore

\[
g_i^{n+1} - g_i^n \leq \Delta t \lambda m_i g_i^n (1 - g_i^{n+1}).
\]

This implies that \( g_i^{n+1} \leq \max\{1, g_i^n\} \). By the definition of \( K \), we always have

\[
g_j^{n+1} \leq K, \quad 1 \leq j \leq N.
\]

(2) A simple calculation gives

\[
E^{n+1} - E^n = \sum_{j=1}^{N} m_j \left[ g_j^{n+1} \log g_j^{n+1} - g_j^n \log g_j^n - (g_j^{n+1} - g_j^n) \right] \Delta x
= \sum_{j=1}^{N} m_j \left[ (g_j^{n+1} - g_j^n) \log g_j^{n+1} + g_j^n \log \frac{g_j^{n+1}}{g_j^n} - (g_j^{n+1} - g_j^n) \right] \Delta x.
\]

Because \( m_j > 0 \), \( g_j^n \geq 0 \) and \( \log X \leq X - 1 \) for \( X > 0 \), we have

\[
E^{n+1} - E^n \leq \sum_{j=1}^{N} m_j \left[ (g_j^{n+1} - g_j^n) \log g_j^{n+1} + g_j^n \left( \frac{g_j^{n+1}}{g_j^n} - 1 \right) - (g_j^{n+1} - g_j^n) \right] \Delta x
= \sum_{j=1}^{N} m_j (g_j^{n+1} - g_j^n) \log g_j^{n+1} \Delta x
= -\Delta t \sum_{j=1}^{N-1} m_{j+\frac{1}{2}} (g_{j+1}^{n+1} - g_j^{n+1}) \log \frac{g_{j+1}^{n+1}}{g_j^{n+1}} - \Delta t \Delta x \lambda \sum_{j=1}^{N} m_j^2 g_j^n (g_j^{n+1} - 1) \log g_j^{n+1}
\leq 0.
\]

Here we have used the facts that \( m_1 D_+ g_0^{n+1} = m_{N+\frac{1}{2}} D_+ g_N^{n+1} = 0 \) and \( (X - Y) \log \frac{X}{Y} \geq 0 \) for \( X, Y > 0 \).

(3) Suppose \( g_j^n = 1 \). We sum (5.27) over \( j \) and end up with

\[
\sum_{j=1}^{N} m_j (1 + \Delta t \lambda m_j) (g_j^{n+1} - 1) = 0.
\]
From (1) we know that $0 \leq g_{j}^{n+1} \leq 1$. Therefore $g_{j}^{n+1} = 1$ for all $j$’s. So the scheme preserves the equilibrium solution.

5.4 Energy stable method for one-species system

5.4.1 Semidiscrete scheme

Let $g^{n}(x)$ be an approximation of $g(x, t_{n})$ and $G^{n} = G(x, g^{n})$. The initial data is given by $g^{0}(x) = g_{0}(x)$. We first discretize (5.13) in time to obtain

$$
M \frac{g^{n+1} - g^{n}}{\Delta t} = \nabla \cdot (M \nabla g^{*}) + \frac{G^{n+1} - G^{n}}{g^{n+1} - g^{n}}
$$

(5.28)

with

$$
g^{*} = \frac{g^{n+1} + g^{n}}{2}.
$$

Notice that

$$
\frac{G^{n+1} - G^{n}}{g^{n+1} - g^{n}} = \lambda Mg^{*}(m - Mg^{*}) - \frac{\lambda}{3} M^{2}(g^{*} - g^{n})^{2}.
$$

So (5.28) can be rewritten as a prediction-correction scheme:

1. Given $g^{n}$, compute the intermediate solution $g^{*}$ which solves

$$
\begin{cases}
2M \frac{g^{*} - g^{n}}{\Delta t} + Pg^{*} + \frac{\lambda}{3} M^{2}(g^{*} - g^{n})^{2} = 0 & \text{in } \Omega, \\
M \partial_{\nu} g^{*} = 0 & \text{on } \partial \Omega.
\end{cases}
$$

(5.29)

2. Obtain $g^{n+1}$ from $g^{n}$ and $g^{*}$:

$$
g^{n+1} = 2g^{*} - g^{n}.
$$

(5.30)

**Theorem 5.4.1.** Given $g^{n} \geq 0$ in $\Omega$, if

$$
\Delta t < \frac{6}{\lambda(3\bar{m} + 2\bar{M}K^{n})}
$$

(5.31)

with $\bar{m} = \max_{x \in \Omega} m(x)$ and $\bar{M} = \max_{x \in \Omega} M(x)$, then there exists a unique solution to (5.29) satisfying

$$
0 \leq g^{*} \leq K^{n} \text{ in } \Omega,
$$

where $K^{n} = \max \{ \| \frac{m}{M} \|_{\infty}, \| g^{n} \|_{\infty} \}$. 
Proof. Let the differential operator in (5.29) be denoted by $L$. Then $K^n$ is a supersolution of (5.29) since

$$LK^n = 2M \frac{K^n - g^n}{\Delta t} + \lambda M^2 K^n \left( K^n - \frac{m}{M} \right) + \frac{\lambda}{3} M^2 (K^n - g^n)^2 \geq 0.$$  

Meanwhile, 0 is a subsolution of (5.29) since condition (5.31) yields

$$L0 = 2 \left( \frac{\lambda M g^n}{6} - \frac{1}{\Delta t} \right) M g^n \leq 0.$$  

Thus the super-sub solution approach [Pao (1992)] yields the existence of $g^*$ and

$$0 \leq g^* \leq K^n \text{ in } \Omega.$$  

Next we prove the uniqueness of the solution. Suppose there are two solutions $g^*_1$ and $g^*_2$ with $g^*_1 \neq g^*_2$. Then $w := g^*_1 - g^*_2$ solves

$$\frac{2}{\Delta t} M w = \nabla \cdot (M \nabla w) + \lambda M M w - \frac{4}{3} \lambda M^2 (g^*_1 + g^*_2) w + \frac{2}{3} \lambda M^2 g^n w.$$  

Multiply by $w$ and integrate over $\Omega$ to deduce

$$\frac{2}{\Delta t} \|w\|_M^2 = -\|\nabla w\|_M^2 + \lambda \int_{\Omega} m M w^2 \, dx - \frac{4}{3} \lambda \int_{\Omega} M^2 (g^*_1 + g^*_2) w^2 \, dx + \frac{2}{3} \lambda \int_{\Omega} M^2 g^n w^2 \, dx \leq -\|\nabla w\|_M^2 + \frac{\lambda}{3} (3\bar{m} + 2\bar{M} K^n) \|w\|_M^2.$$  

It follows that

$$\|\nabla w\|_M^2 + \left[ \frac{2}{\Delta t} - \frac{\lambda (3\bar{m} + 2\bar{M} K^n)}{3} \right] \|w\|_M \leq 0.$$  

Due to the choice of $\Delta t$ (5.31), we must have $\|w\|_M^2 = \|\nabla w\|_M^2 = 0$, i.e., $w = 0$ on $\Omega$.  

The semidiscrete energy $V^n$ is defined as

$$V^n = \int_{\Omega} \left[ \frac{1}{2} M |\nabla g^n|^2 - G^n \right] \, dx.$$  

**Theorem 5.4.2.** The semidiscrete two-step scheme (5.29)-(5.30) has the following properties:

1. The functional $V^n$ is nonincreasing unconditionally, i.e.,

$$V^{n+1} \leq V^n \quad \forall n \geq 0.$$
2. If (5.31) is met, then the scheme preserves the equilibrium solution.

Proof. (1) We first prove that $V^n$ is nonincreasing in $n$:

$$V^{n+1} - V^n = \int_\Omega \left[ \frac{1}{2} M (|\nabla g^{n+1}|^2 - |\nabla g^n|^2) - (G^{n+1} - G^n) \right] \, dx$$

$$= - \int_\Omega \left[ \nabla \cdot (M \nabla g^n) (g^{n+1} - g^n) + (G^{n+1} - G^n) \right] \, dx$$

$$= - \int_\Omega \frac{M}{\Delta t} (g^{n+1} - g^n)^2 \, dx \leq 0.$$

Again, the zero flux $M \partial_\nu g^n |_{\partial \Omega} = 0$ has been used.

(2) Now suppose $g^n$ achieves the steady state $\tilde{g}$. Then $g^n$ satisfies the equation

$$\mathcal{P} g^n = 0.$$

Using (5.29) and letting $w = g^* - g^n$, we have

$$\frac{2}{\Delta t} M w = \nabla \cdot (M \nabla w) + \lambda m M w - \lambda M^2 (g^* + g^n) w - \frac{\lambda}{3} M^2 w^2.$$

Multiply the above equation by $w$ and integrate on $\Omega$:

$$\frac{2}{\Delta t} \|w\|^2_M \leq -\|\nabla w\|^2_M + \frac{\lambda}{3} (3\bar{m} + 2\bar{M} K^n) \|w\|^2_M.$$

It follows that

$$\|\nabla w\|^2_M + \left[ \frac{2}{\Delta t} - \frac{\lambda (3\bar{m} + 2\bar{M} K^n)}{3} \right] \|w\|^2_M \leq 0.$$

Due to the condition on $\Delta t$, we must have $\|w\|^2_M = \|\nabla w\|^2_M = 0$. Therefore, $g^* = g^n$ and then $g^{n+1} = g^n$; i.e., the scheme preserves the equilibrium solution. \qed

5.4.2 Fully discrete scheme

We apply the central difference in space to the operator $\mathcal{P}$ to get $\mathcal{P}_h$. In the one-dimensional setting, it can be simply written as

$$\mathcal{P}_h g_i = -D_- (M_{i+\frac{1}{2}} D_+ g_i) - \lambda M_i g_i (m_i - M_i g_i), \quad 1 \leq i \leq N.$$

Then given the initial data $g^0_i = g_0(x_i)$, the fully discrete scheme consists of two steps:
(1) Compute the intermediate solution $\vec{g}^*$ which solves
\[ 2M_i \frac{g_i^*-g_i^n}{\Delta t} + \mathcal{P}_h g_i^* + \frac{\lambda}{3} M_i^2 (g_i^* - g_i^n)^2 = 0, \quad 1 \leq i \leq N \] (5.32)
with $M \frac{1}{2} D + g_0^* = M_{N+\frac{1}{2}} D + g_N^* = 0$.

(2) Obtain $\vec{g}^{n+1}$ from $\vec{g}^n$ and $\vec{g}^*$:
\[ \vec{g}^{n+1} = 2\vec{g}^* - \vec{g}^n. \] (5.33)

The fully discrete energy $V^n$ is defined as
\[ V^n = \sum_{i=1}^{N-1} \left[ \frac{1}{2} M_i + \frac{1}{2} (D + g_i^n)^2 - G(x_i, g_i^n) \right] \Delta x. \]

**Theorem 5.4.3.** Let $\vec{g}^n$ be the solution of the fully discrete scheme (5.32)-(5.33). Then

1. The functional $V^n$ is nonincreasing, i.e.,
\[ V^{n+1} \leq V^n \quad \forall n \geq 0. \]

2. If $\Delta t$ satisfies (5.31), then the scheme preserves the equilibrium solution $\{\vec{g}_i\}$ that satisfies
\[ \mathcal{P}_h \vec{g}_i = 0, \quad 0 \leq i \leq N. \]

**Proof.** (1) Using summation by parts, we have
\[ V^{n+1} - V^n = \sum_{i=1}^{N-1} M_i \frac{g_i^{n+1} - g_i^n}{\Delta t} \Delta x - \sum_{i=1}^{N} \left[ G(x_i, g_i^{n+1}) - G(x_i, g_i^n) \right] \Delta x \]
\[ = - \sum_{i=1}^{N} D_-(M_i + \frac{1}{2} D + g_i^n)^2(g_i^{n+1} - g_i^n) \Delta x - \sum_{i=1}^{N} \left[ G(x_i, g_i^{n+1}) - G(x_i, g_i^n) \right] \Delta x. \]

Note that (5.32) can be rewritten as
\[ M_i \frac{g_i^{n+1} - g_i^n}{\Delta t} = D_-(M_i + \frac{1}{2} D + g_i^n)^2 + \frac{G_i^{n+1} - G_i^n}{g_i^{n+1} - g_i^n}. \]

Therefore
\[ V^{n+1} - V^n = - \frac{\Delta x}{\Delta t} \sum_{i=1}^{N} M_i (g_i^{n+1} - g_i^n)^2 \leq 0. \]

(2) Suppose at time $t_n$, it achieves the equilibrium state, i.e., $\mathcal{P}_h g_i^n = 0$ for $1 \leq i \leq N$. Then $w_i = g_i^* - g_i^n$ solves
\[ \frac{2}{\Delta t} M_i w_i = D_-(M_i + \frac{1}{2} D + w_i) + \lambda M_i^2 w_i - \lambda M_i^2 (g_i^* + g_i^n) w_i + \frac{\lambda}{3} M_i^2 w_i^n. \]
Multiply by $w_i$ and sum over $i$:

$$\frac{2}{\Delta t} \sum_{i=1}^{N} M_i w_i^2 \leq \frac{\lambda}{3}(3\bar{m} + 2\bar{M}K^n) \sum_{i=1}^{N} M_i w_i^2.$$  

The condition (5.31) implies that

$$\sum_{i=1}^{N} M_i w_i^2 = 0.$$  

Hence $g_i^{*} = g_i^n$, $1 \leq i \leq N$.  

**Remark 5.4.1.** Results in above two theorems remain valid for the multidimensional setting.

We conclude this section by briefly discussing some implementation issues.

Since the fully discrete scheme is nonlinear in $g^*$, we shall compute $g^*$ by an efficient iteration. Define a vector function $\vec{F}(\vec{g}^*)$ such that

$$F_i(\vec{g}^*) := 2M_ig_i^* - g_i^n \Delta t + P_h g_i^* + \frac{\lambda}{3} M_i^2 (g_i^* - g_i^n)^2, \quad 1 \leq i \leq N.$$  

Then solving for (5.32) is equivalent to finding the root of $\vec{F}(\vec{g}^*)$. Let $J_F(\vec{g}^*)$ denote the Jacobian matrix of $\vec{F}$. We use the Newton’s method with the initial guess produced from the following one-step scheme:

$$M_i g_i^{n+1} - g_i^n \Delta t = D_-(M_{i+\frac{1}{2}} D_i g_i^{n+1}) + \lambda M_i g_i^n (m_i - M_i g_i^{n+1}).$$  

(5.34)

By a simple analysis of this linear system, we can show that $g_i^{n+1} \geq 0$ if $g_i^n \geq 0 \forall i$.

Given $\vec{g}^n$, the algorithm of solving for $\vec{g}^{n+1}$ can be summarized as follows:

**Algorithm**

1. Initialization. Use the scheme (5.34) to get $\vec{g}_0^{n+1}$. Then set

$$\vec{g}_0 = \frac{\vec{g}^n + \vec{g}_0^{n+1}}{2}.$$  

2. Iteration. Let $\vec{g}_k^n$ denote the approximation to $\vec{g}^n$ at the $k$-th iteration.

   Do $k = 0, 1, 2, \ldots$

   Solve $J_F(\vec{g}_k^n)\vec{x} = -\vec{F}(\vec{g}_k^n)$.
\[ \vec{g}_{k+1}^* = \vec{x} + \vec{g}_k^*. \]

Check \textit{error} = \| \vec{g}_{k+1}^* - \vec{g}_k^* \|. If \textit{error} is small enough, stop.

EndDo

3. \( \vec{g}^{n+1} = 2\vec{g}_{k+1}^* - \vec{g}^n \).

### 5.5 Entropy stable method for two-species system

The two-species system is described as follows.

\[
\begin{align*}
\begin{cases}
    u_t &= \alpha \nabla \cdot (\nabla u + u \nabla P) + u(m - u - v) \quad \text{in } \Omega, \\
    v_t &= \beta \nabla \cdot (\nabla v + v \nabla Q) + v(m - u - v) \quad \text{in } \Omega, \\
    \partial_\nu u + u \partial_\nu P &= \partial_\nu v + v \partial_\nu Q = 0 \quad \text{on } \partial \Omega, \\
    u(x,0) &= u_0(x), \quad v(x,0) = v_0(x) \quad \text{in } \Omega.
\end{cases}
\end{align*}
\] (5.35)

We only consider the ideal free distribution where \( m > 0 \) on \( \Omega \), \( P(m) = -\log m + C \), and \( Q(m) \neq P(m) \). Here \( C \) is an arbitrary constant. For this particular case, \((m,0)\) is the globally attractive equilibrium solution. Further details on this model can be found in Cantrell et al. (2010). For the relative entropy \( E[u,v] \), we have

\[
\frac{dE}{dt} = \int_\Omega \left( u_t + v_t - \frac{m}{u} u_t \right) dx
\]

\[
= -\alpha \int_\Omega \frac{|m \nabla u - u \nabla m|^2}{mu^2} dx - \int_\Omega (m - u - v)^2 dx \leq 0.
\]

Here the zero boundary flux has been applied.

Define \( g = \frac{u}{m}, \quad h = \frac{v}{W} \) with \( W = e^{-Q} \) and \( F(g,h,m) = m - mg - Wh \). (5.35) reduces to

\[
\begin{align*}
\begin{cases}
    mg_t &= \alpha \nabla \cdot (m \nabla g) + mgF(g,h,m) \quad \text{in } \Omega, \\
    Wh_t &= \beta \nabla \cdot (W \nabla h) + WhF(g,h,m) \quad \text{in } \Omega, \\
    m\partial_\nu g &= W \partial_\nu h = 0 \quad \text{on } \partial \Omega, \\
    g(x,0) &= g_0(x), \quad h(x,0) = h_0(x) \quad \text{in } \Omega.
\end{cases}
\end{align*}
\] (5.36)

We shall develop an entropy stable scheme based on this reformulation.
5.5.1 Semidiscrete scheme

We use the central difference in spatial discretization to obtain, for \(1 \leq j \leq N\),

\[
\begin{align*}
  m_j \frac{dg_j}{dt} &= \alpha D_-(m_{j+\frac{1}{2}} D_+ g_j) + m_j g_j F_j, \\
  W_j \frac{dh_j}{dt} &= \beta D_-(W_{j+\frac{1}{2}} D_+ h_j) + W_j h_j F_j.
\end{align*}
\] (5.37)

Set \(m_{\frac{1}{2}} D_+ g_0 = m_{N+\frac{1}{2}} D_+ g_N = W_{\frac{1}{2}} D_+ h_0 = W_{N+\frac{1}{2}} D_+ h_N = 0\) to incorporate the zero boundary flux. And the initial data is \(g_j(0) = g_0(x_j), h_j(0) = h_0(x_j)\).

For the semidiscrete entropy

\[
E_h(t) = \Delta x \sum_{j=1}^{N} [m_j g_j + W_j h_j - m_j \log(m_j g_j)],
\]

we have

\[
\frac{dE_h}{dt} = \Delta x \sum_{j=1}^{N} \left[ m_j \frac{dg_j}{dt} + W_j \frac{dh_j}{dt} - \frac{m_j}{g_j} \frac{dg_j}{dt} \right] = -\frac{\alpha}{\Delta x} \sum_{j=1}^{N} \frac{m_{j+\frac{1}{2}} (g_{j+1} - g_j)^2}{g_j g_{j+1}} - \Delta x \sum_{j=1}^{N} F_j^2 \leq 0,
\]

where the zero flux has been used.

A natural question is whether this functional suffices to enforce the equilibrium preserving property. To see this, we assume the evolution becomes steady at certain time \(t\); then

\[
\frac{dg_j}{dt} = \frac{dh_j}{dt} = 0 \quad \forall j
\]

and \(\frac{dE_h}{dt} = 0\) ensures that

\[
g_{j+1} - g_j = 0 \quad \text{and} \quad F_j = 0 \quad \forall j.
\]

Therefore, \(g_j \equiv c\) with \(c\) being some constant and \(h_j = \frac{(1-c)m_j}{W_j}\). If \(c = 1\), then \(\frac{dg_j}{dt} = \frac{dh_j}{dt} = 0 \forall j\) and the system achieves the steady state. Otherwise, we have \(\frac{dg_j}{dt} = 0\) for all \(j\)'s and

\[
W_j \frac{dh_j}{dt} = \beta(1-c) D_-(W_{j+\frac{1}{2}} D_+ m_j W_j) \neq 0 \quad \text{for some } j.
\]

The system will continue to evolve in time and approach the equilibrium solution \((m,0)\) eventually.
5.5.2 Fully discrete scheme

We apply backward Euler method to (5.37): for \( n \geq 0 \) and \( 1 \leq j \leq N \),
\[
m_j \frac{g_j^{n+1} - g_j^n}{\Delta t} = \alpha D_- (m_{j+\frac{1}{2}} D_g g_j^{n+1}) + m_j g_j^{n+1} F_j^{n+1}, \tag{5.38a}
\]
\[
W_j \frac{h_j^{n+1} - h_j^n}{\Delta t} = \beta D_- (W_{j+\frac{1}{2}} D_h h_j^{n+1}) + W_j h_j^{n+1} F_j^{n+1}. \tag{5.38b}
\]
The initial data is given by \( g_j^0 = g_0(x_j), h_j^0 = h_0(x_j) \). And the zero boundary flux is enforced by setting
\[
m_{\frac{1}{2}} D_g g_0^n = m_{N+\frac{1}{2}} D_g g_N^n = W_{\frac{1}{2}} D_h h_0^n = W_{N+\frac{1}{2}} D_h h_N^n = 0 \quad \forall n \geq 0.
\]

**Theorem 5.5.1.** The fully discrete scheme (5.38) possesses the following properties:

1. If \( g_j^n > 0 \) and \( h_j^n \geq 0, 1 \leq j \leq N \), then
   \[
g_j^{n+1} > 0 \quad \text{and} \quad h_j^{n+1} \geq 0 \quad \forall n \geq 0, \quad 1 \leq j \leq N,
   \]
   provided \( \Delta t < \frac{1}{\|m\|_{\infty}} \).

2. The fully discrete entropy
   \[
   E^n = \Delta x \sum_{j=1}^{N} \left[ m_j g_j^n + W_j h_j^n - m_j \log(m_j g_j^n) \right]
   \]
   is nonincreasing unconditionally, i.e., \( E^{n+1} \leq E^n \).

**Proof.** (1) Notice that the numerical solution \( \{\vec{g}^{n+1}, \vec{h}^{n+1}\} \) depends on \( \Delta t \) continuously. If we increase \( \Delta t \) continuously from zero up to \( \frac{1}{\|m\|_{\infty}} \), no component of \( \vec{g}^{n+1} \) can become negative without passing through zero since \( \vec{g} > 0 \). Assume for some \( \Delta t \), we have \( \vec{g}_i^{n+1} = \min_{1\leq j} g_j^{n+1} = 0 \).

Then equation (5.38a) gives
\[
-\frac{g_i^n}{\Delta t} \geq 0,
\]
which contradicts \( g_i^n > 0 \). So we must have \( \vec{g}_i^{n+1} > 0 \), hence \( g_j^{n+1} > 0 \ \forall j \).

We apply the similar argument to \( \vec{h}^{n+1} \). Assume for some \( \Delta t \), we have \( \vec{h}_i^{n+1} = \min_{1\leq j} h_j^{n+1} = -\varepsilon \), where \( \varepsilon \) is a sufficiently small number. Then equation (5.38b) gives
\[
\left( \frac{1}{\Delta t} - m_i + m_i g_i^{n+1} - W_i \varepsilon \right) \varepsilon \leq \frac{h_i^n}{\Delta t} \leq 0
\]
\[
\frac{1}{\Delta t} - m_i \leq -m_i g_i^{n+1} + W_i \varepsilon \leq 0,
\]
since \( g_i^{n+1} > 0 \), \( W_i \) is bounded and \( \varepsilon \) is sufficiently small. This contradicts the choice of \( \Delta t \). Therefore, we must have \( h_i^{n+1} \geq 0. \)

(2) For the fully discrete entropy \( E^n \), we have

\[
E^{n+1} - E^n \leq -\frac{\alpha \Delta t}{\Delta x} \sum_{j=1}^{N} \frac{M_{j+\frac{1}{2}} (g_{j+1}^{n+1} - g_j^n)^2}{g_j^n g_{j+1}^{n+1}} - \Delta t \Delta x \sum_{j=1}^{N} (F_j^{n+1})^2 \leq 0.
\]

\( \square \)

5.6 Numerical Tests

We provide numerical results to demonstrate the accuracy of the schemes and the capacity to capture solution features for large times.

Denote the numerical solution at \((x_j, t_n)\) by \( u_j^n \) and the exact solution by \( u(x_j, t_n) \). Define the \( L^\infty \) error as

\[
\max_{1 \leq j \leq N} |u_j^n - u(x_j, t_n)|
\]

and the \( L^1 \) error as

\[
\sum_{j=1}^{N} |u_j^n - u(x_j, t_n)| h.
\]

When the exact solution is not available, we replace \( u(x_j, t_n) \) by a reference solution to compute the errors.

5.6.1 One-dimensional tests when \( P(m) = -\log m \)

Consider the problem with \( \Omega = [0, 1] \) and

\[
m(x) = 10x^2(x - 1)^2(x - 0.4005)^2.
\]

Note that \( m(0) = m(1) = 0 \), so equation (5.11) becomes degenerate on \( \partial \Omega \). We first illustrate the accuracy of scheme (5.25) with the continuous initial data:

\[
u_0(x) = \frac{\sin(100x) + 1}{2}.
\]

Table 5.1 shows that the scheme is of second order in space. Here the reference solution is taken as the numerical solution with \( N = 1280 \).
Table 5.1 Error and order of accuracy of scheme (5.25) with the initial data (5.39) on a uniform mesh of $N$ cells, final time $t = 1.0$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$L^\infty$ error</th>
<th>order</th>
<th>$L^1$ error</th>
<th>order</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2.352E-03</td>
<td>-</td>
<td>1.175E-03</td>
<td>-</td>
</tr>
<tr>
<td>40</td>
<td>2.601E-04</td>
<td>3.176</td>
<td>1.305E-04</td>
<td>3.170</td>
</tr>
<tr>
<td>80</td>
<td>5.763E-05</td>
<td>2.174</td>
<td>2.821E-05</td>
<td>2.210</td>
</tr>
<tr>
<td>160</td>
<td>1.524E-05</td>
<td>1.919</td>
<td>6.800E-06</td>
<td>2.053</td>
</tr>
</tbody>
</table>

Table 5.2 The entropy of the numerical solutions on a uniform mesh: $N = 80$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>$E(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.770E+00</td>
</tr>
<tr>
<td>5</td>
<td>1.534E-01</td>
</tr>
<tr>
<td>10</td>
<td>5.937E-02</td>
</tr>
<tr>
<td>50</td>
<td>2.702E-03</td>
</tr>
<tr>
<td>100</td>
<td>3.719E-04</td>
</tr>
<tr>
<td>200</td>
<td>1.905E-05</td>
</tr>
<tr>
<td>400</td>
<td>9.530E-08</td>
</tr>
</tbody>
</table>

The entropy decreasing property is shown in Table 5.2, where the relative entropy (5.26) is decreasing to 0 as $t$ increases, indicating the numerical convergence to the equilibrium solution $m(x)$ as time becomes large.

We also test the numerical convergence to the equilibrium solution $m(x)$ by two numerical examples shown in Figure 5.1. The initial data are respectively given by random data and the $\delta$-like function

$$u_0(x) = \begin{cases} 
10 \left[ \cos(20\pi(x-0.8)) + 1 \right] & \text{if } |x-0.8| \leq 0.05; \\
0 & \text{elsewhere.}
\end{cases}$$

5.6.2 One-dimensional tests when $P(m) = -\alpha m$

Here $\alpha \geq 0$ is a parameter. We consider the growth rate

$$m(x) = 10x^2(x-1)^2(x-0.4005)^2 - 0.02,$$

which is positive for some $x \in \Omega$, and set up $\alpha = 40$ and $\lambda = 0.1$ such that

$$\int_\Omega m(x) \, dx \approx -0.0048 < 0 \quad \text{and} \quad \int_\Omega m(x)e^{-P} \, dx \approx 0.0045 > 0.$$

In such a setting we have $\lambda^* = 0$: the solution will converge to the positive equilibrium other than $m(x)$, regardless of the size of $\lambda$ as well as the initial density.
Figure 5.1 For the figures in the first row, $u_0(x)$ is given by the random data; for the second row, $u_0(x)$ is given by the $\delta$-like function. $N = 80$.

We first illustrate the accuracy of scheme (5.32)-(5.33) with Newton’s iteration. The initial data is taken as (5.39), again using the numerical solution with $N = 1280$ as the reference solution. The results in Table 5.3 show that the scheme is of the second order in space.

We display in Table 5.4 the discrete energy $V$ at different times, using the same initial data (5.39).

Figure 5.2 shows that solutions with different initial data converge to the same positive equilibrium solution, as predicted by the theory. Moreover, we observe that the equilibrium solution concentrates at the local positive maximum of $m(x)$. This is consistent with the theoretical result obtained in Theorem 4 of Bezuglyy and Lou (2010).
Table 5.3  Error and order of accuracy of scheme (5.32)-(5.33) with initial data (5.39) on a uniform mesh of \( N \) cells, final time \( t = 1.0 \).

<table>
<thead>
<tr>
<th>( u_0(x) )</th>
<th>( \frac{\sin(100x)+1}{2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>( L^1 ) error</td>
</tr>
<tr>
<td>10</td>
<td>2.55E-03</td>
</tr>
<tr>
<td>20</td>
<td>1.16E-03</td>
</tr>
<tr>
<td>40</td>
<td>1.23E-04</td>
</tr>
<tr>
<td>80</td>
<td>2.66E-05</td>
</tr>
<tr>
<td>160</td>
<td>6.37E-06</td>
</tr>
</tbody>
</table>

Table 5.4  The energy of numerical solutions on a uniform mesh: \( N = 80 \).

<table>
<thead>
<tr>
<th>( t_n )</th>
<th>( V^n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7.387E+02</td>
</tr>
<tr>
<td>5</td>
<td>4.405E-04</td>
</tr>
<tr>
<td>10</td>
<td>7.523E-05</td>
</tr>
<tr>
<td>50</td>
<td>6.435E-07</td>
</tr>
<tr>
<td>100</td>
<td>4.125E-08</td>
</tr>
<tr>
<td>200</td>
<td>-1.522E-08</td>
</tr>
</tbody>
</table>

5.6.3  Two-dimensional tests with \( P(m) = -\alpha m^2 \)

We consider the resource

\[
m(x) = 10 \left[ 0.25 - \left( x_1 - 0.5 \right)^2 - \left( x_2 - 0.5 \right)^2 \right] e^{(x_1-0.5)^2-(x_2-0.5)^2} - 0.3
\]

and set \( \alpha = 4 \) so that

\[
\int_{\Omega} m(x) \, dx \approx -0.1198 < 0 \quad \text{and} \quad \int_{\Omega} m(x)e^{-P} \, dx \approx -0.1590 < 0.
\]

In such a setting we have \( \lambda^* > 0 \): the solution will converge to the positive equilibrium if \( \lambda > \lambda^* \) and decay to zero if \( \lambda \leq \lambda^* \), regardless of the choice of initial density.

We test scheme (5.32)-(5.33) with two initial data:

(i) the cos-like function \( u_0(x) = \cos(3\pi|x|^2) + 1 \);

(ii) the \( \delta \)-like function concentrating at \( (x_0i, x_0j) \) for \( i,j = 1,2 \) with \( x_{01} = 0.2 \) and \( x_{02} = 0.8 \).

\[
u_0(x) = \begin{cases} 
25 \left[ \cos(10\pi(x_1 - x_{0i})) + 1 \right] \left[ \cos(10\pi(x_2 - x_{0j})) + 1 \right] & \text{if } |x_i - x_{0j}| \leq 0.1 \text{ and } i,j = 1,2; \\
0 & \text{elsewhere.}
\end{cases}
\]
Figure 5.2 For the figures in the first row, $u_0(x)$ is given by the random data; for the second row, $u_0(x)$ is given by the $\delta$-like function. $N = 80$.

From Figure 5.4, we observe that when $t = 100$, the numerical solutions is of order $O(10^{-7})$, approaching zero. In other words, the species will extinct eventually, consistent with the theoretical result stated in Theorem 5.1.1.

The approach toward the positive equilibrium is shown with two examples in Figure 5.5, corresponding to the persistence of the species as predicted by Theorem 5.1.1. Moreover, the energy $V \approx -0.13439$ when $t = 0.5$ and 100 for both of the initial data. These tests indicate that the steady state is asymptotically approached in time and well preserved.

5.6.4 Two-dimensional test for the two-species system

We consider the two species system with resource potentials $P(m) = -\log m$ and $Q(m) = -m$, where the resource $m(x)$ is given by

$$m(x) = 0.5 \sin(\pi x_1) \sin(\pi x_2) + 0.6, \quad x \in \Omega = [0, 2.5]^2.$$
Figure 5.3 Plot of \( m(x) = 10 \left[ 0.25 - (x_1 - 0.5)^2 - (x_2 - 0.5)^2 \right] e^{(x_1-0.5)^2-(x_2-0.5)^2 - 0.3} \).

We set \((\alpha, \beta) = (5, 1)\) and take initial value as

\[
\begin{align*}
u_0(x) & = 1 - 2v_0(x), \\
v_0(x) & = \begin{cases} 
\frac{1}{4} \sin^2(4\pi x_1) \sin^2(4\pi x_2) & \text{if } 1 \leq x_1, x_2 \leq 1.5, \\
0 & \text{elsewhere}.
\end{cases}
\end{align*}
\]

Visually, we observe the approach toward the equilibrium solution \((m, 0)\) through the numerical result shown in Figure 5.7.

Quantitatively, Table 5.5 shows that the entropy is decreasing as time evolves and indicates the asymptotic approach toward the equilibrium state, while we note that \( E \approx 5.34191 \) at the steady state.

Table 5.5 The discrete entropy on a uniform mesh: \( N = 40 \times 40 \).

<table>
<thead>
<tr>
<th>( t_n )</th>
<th>0</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E^n )</td>
<td>6.26436</td>
<td>5.80962</td>
<td>5.66577</td>
<td>5.57095</td>
<td>5.35172</td>
</tr>
</tbody>
</table>

5.7 Concluding remarks

We have developed finite difference schemes for a class of reaction-diffusion-advection equations arising in the evolution of biased dispersal of population dynamics, including the single
species model and two competing species system. The schemes are shown to preserve both positivity and equilibrium, satisfying either entropy or energy dissipation structure. These ensure that the numerical solution provides a satisfying long-time behavior, thus revealing the desired dispersal pattern. Numerical examples are shown to illustrate the second-order accuracy for various resource potentials and underline the efficiency to preserve the large-time asymptotic. We also proved that the unique existence of the numerical solutions for the proposed schemes.

Our future work would develop higher order methods based on the proposed semidiscrete formulations. For dispersal models with multispecies, the dynamics become much more complex and further numerical investigation into multispecies dispersal models deserves serious consideration.
Figure 5.5 Figures in the first row, $u_0(x)$ is given by the cos-like function; in the second row, $u_0(x)$ is given by the $\delta$-like function. $N = 20 \times 20, \lambda = 1000 > \lambda^*$. 

Acknowledgments

This research was partially supported by the National Science Foundation under grant DMS07-57227 and the Mathematical Research Network KI-net grant.
Figure 5.6  Plot of $m(x) = 0.5 \sin(\pi x_1) \sin(\pi x_2) + 0.6$.

Figure 5.7  The first row: $u$ component; the second row: $v$ component. $N = 40 \times 40$. 
CHAPTER 6. SUMMARY AND DISCUSSION

6.1 General conclusion

We have proposed entropy satisfying and maximum-principle-satisfying high-order numerical methods for solving Fokker–Planck type equations, including the Fokker–Planck equation of the finitely extensible nonlinear elastic (FENE) dumbbell model for polymers, subject to homogeneous fluids, and the reaction-diffusion-advection equation arising in the evolution of biased dispersal of population dynamics. The main feature of our method is its capacity of capturing the long time asymptotic behavior. This is ensured by enforcing the scheme to satisfy several main properties, including the nonnegativity principle, the mass conservation and the preservation of nonzero steady states.

One main methodology is to use the nonlogarithmic Landau form of the Fokker–Planck equation which is expressed in terms of the relative ratio of the pdf $f$ and the equilibrium $M$. There are several advantages of using this reformulation: (1) relative quadratic entropy becomes the usual weighted $L^2$ energy so that standard weighted Galerkin approximation can be applied; (2) the maximum principle of $f$ in the range of $[c_1,c_2]M$ reduces to the usual maximum principle for $f/M$ in the range of $[c_1,c_2]$ and (3) the difficulty caused by boundary singularity in the FENE model is well dealt with by this reformulation.

Our numerical results on the entropy satisfying DG schemes for the FENE model include both one- and two-dimensional examples, in terms of accuracy, long asymptotic measured by the relative entropy and the effects of some canonical homogeneous flows.

It is known a rather difficult problem for a high order method to satisfy the maximum principle in solving Fokker–Planck equations. To overcome the difficulty, we have developed a positive decomposition of weighted cell averages. This decomposition together with the
particular numerical flux and a modified scaling limiter enables us to develop up to third-order DG schemes on rectangular meshes that satisfy the strict maximum principle.

For the evolutionary dispersal models, we have developed second-order finite difference methods for one-species problem with general resource potential and two-species system with the ideal free distribution. The methods are proved to be capable of preserving the positivity and converging to the stable steady state.

6.2 Future work

As an important property of linear Fokker-Planck equations, positivity preservation is always the challenging aspect of numerical methods. Our first-order entropy satisfying conservative methods show this good quality both analytically and numerically. Using a particularly chosen numerical flux, we develop up to third-order schemes on rectangular meshes that not only preserve the positivity but also satisfy the maximum principle. We plan to extend the developed third-order DG method to unstructured meshes and possibly explore even higher order positivity-preserving methods.

For the FENE model, the boundary singularity makes the linear system derived from the numerical schemes ill-conditioned. We plan to use adaptive meshes to address this problem.

Finally, the rigorous error estimate for our DG schemes is still missing. One main difficulty in the case of solving the FENE model is that the weight function $M$ vanishes at the boundary. Another difficulty is the loss of order of accuracy due to involved high order derivatives in the numerical flux. We plan to obtain the optimal error estimate for our numerical methods.
BIBLIOGRAPHY


Furihata, D. (1999). Finite difference schemes for $\frac{\partial u}{\partial t} = (\frac{\partial}{\partial x})^{\alpha} \frac{\delta G}{\delta u}$ that inherit energy conservation or dissipation property. *Journal of Computational Physics*, 156(1):181–205.


