A Well-Balanced Method for an Unstaggered Central Scheme

Yu-Chen Cheng

Advisor : Prof. Dr. Christian Klingenberg

Department of Mathematics

Julius-Maximilians-Universität Würzburg

November 2023

Acknowledgments

I would like to express my eternal gratitude to all the people who have ever helped me during my study in the University of Würzburg.

I would first like to thank my advisor, Prof. Dr. Christian Klingenberg, for his expert guidance and constant encouragement throughout the writing of this thesis. Without his enlightening instruction, I could not have completed this thesis. Also, I am extremely indebted to him for leading me to know the German life. I am absolutely delighted to have tasted various cakes and food.

I would particularly like to thank Prof. Dr. Rony Touma for taking the time to help me to sort out the issue and give me invaluable, illuminating comments for this thesis.

I must express my sincere thanks to Prof. Dr. Albert Chern for improving my programming skills that really saves me a lot of time.

I am also sincerely grateful to lovely Dr. Farah Kanbar for her conscientious tutoring. She is my saviour at the beginning of writing this thesis.

Besides, I also want to express my appreciation to the former and current members of Prof. Dr. Klingenberg's work group, especially Prof. Dr. Marlies Pirner, Sandra Warnecke, Claudius Birke, Kathrin Hellmuth and Lena Baumann. Special thank to Claudius Brike for the fruitful suggestion. With many heartfelt thanks to warm-hearted Kathrin Hellmuth for her friendly, patient and thoughtful assistance in these days.

In addition, I would like to thank my brother, Dr. Yu-Shiuan Cheng, for carefully reading the thesis draft, correcting my English grammar and giving me useful recommendations.

Finally, I would like to express my deepest gratitude to my parents for the financial support and their endless love, care and encouragement.

Abstract

In 1990, based on central scheme, Nessyahu and Tadmor introduced a second-order, high-resolution MUSCL scheme, (NT scheme), which was designed by approximating the average over Riemann fans which can avoid to solve Riemann problems. After ten years, in 2000, Kurganov and Tadmor advanced a modified NT scheme: A narrower cell is considered when evaluating the average over Riemann fans. It can avert to overestimate at the smooth region.

In this thesis, we propose a new MUSCL scheme for one and two dimensions by combining the ideas of the KT scheme and the Deviation method (a well-balanced finite volume method for hyperbolic balance laws), in which the difference between exact solution and a given stationary solution is used.

We develop a semi-discrete scheme from our new fully-discrete scheme and apply it to homogeneous scalar conservation laws to demonstrate it is almost nonoscillatory in the sense of satisfying TVD property in 1D and maximum principle in 2D for the difference constructed from the Deviation method.

We conclude the thesis by numerically applying our fully-discrete scheme to Euler equations with gravitational source term in 1D and 2D and presenting the results.

Zusammenfassung

Im Jahr 1990 führten Nessyahu und Tadmor auf der Grundlage des zentralen Schemas ein hochauflösendes MUSCL-Schema zweiter Ordnung ein, das auf der Annäherung an den Durchschnitt über Riemann-Fächer konzipiert wird, mit denen wir keine Riemann-Probleme lösen müssen. Nach zehn Jahren, im Jahr 2000 entwickelten Kurganov und Tadmor ein modifiziertes NT-Schema: Bei der Annäherung an den Durchschnitt über Riemann-Fächer wird eine schmalere Zelle verwendet. Es kann eine Überschätzung im glatten Bereich verhindern.

In dieser Masterarbeit schlagen wir ein neues MUSCL-Schema für eine und zwei Dimensionen vor, indem wir die Ideen des KT-Schemas und der Deviationsmethode kombinieren, bei der die Differenz zwischen der exakten Lösung und einer gegebenen stationären Lösung verwendet wird.

Und wir stellen ein semi-diskrete Schema aus unserem neuen voll- diskreten Schema vor, und wenden es für die homogenen Skalarerhaltungssätze an, um zu zeigen, dass es nahezu nicht oszillierend im Sinne der Erfüllung der TVD-Eigenschaft in 1D und des Maximalprinzips in 2D für die aus der Abweichungsmethode konstruierte Differenz ist.

Wir schließen die Masterarbeit ab, indem wir unser voll-diskrete Schema auf Euler-Gleichungen mit gravitativem Quellterm in 1D und 2D anwenden und die Ergebnisse präsentieren.

Contents

Chapter 1

Introduction

Consider the general balanced law

$$\partial_t q(x,t) + \nabla_x f(q(x,t)) = S(q(x,t)), \tag{1.1}$$

where $q(x,t) = (q_1(x,t), q_2(x,t), ..., q_N(x,t))^T$ is an N-vector of conserved quantities in the d-spatial variables $x = (x_1, x_2, ..., x_d)$, and $f(q) = (f^1, f^2, ..., f^d)$ is a nonlinear flux, and $S(q) = (s^1, s^2, ..., s^d)$ is a source term.

In this thesis, we are interested in the finite volume type method for solving the 1D and 2D Euler system with linear gravitational source term. In section 2.1.1 and 2.1.2, we introduce the derivations of Euler equations. And in section 2.1.3, the content of the hydrostatic equilibrium is an important key when we apply the Deviation method, which is described in section 2.2.2, to our new scheme.

Our scheme is inspired by Kurganov-Tadmor scheme (KT scheme) [3] and the well-balanced Deviation method [14]. The main idea of the KT scheme is to consider narrower control cells to approximate the averages over Riemann fans. In this way, overestimating the averages over the smooth regions can be avoided. The description of KT scheme is presented in section 2.2.1.

The Deviation method, as its name suggests, is a method considering the the deviation between the desired solution q and the target solution $\tilde{q}$. We illustrate this method and its well-balanced property in section 2.2.2.

We present a new scheme in section 3.1, which combines the idea of KT scheme and Deviation method, and prove the well-balanced property holds in section 3.2. In section 3.3, we construct the corresponding semi-discrete scheme and prove this semi-discrete scheme satisfies TVD property when applying to the modified homogeneous scalar conservation law.

Then we extend our scheme to two-dimension in section 4.1. To construct a new 2-dimensional scheme, we follow the structure of [7] and the 2-dimensional Deviation method in [4] and [15] . In [7] the authors developed a 2-dimensional modified KT-type scheme extended from 1-dimensional KT scheme [3]. In section 4.2, we introduce another structure of 2-dimensional KT-type scheme from [8] and [9], and derive a corresponding 2-dimensional semi-discrete scheme. Next, in section 4.3 we apply the new semi-discrete scheme to the modified homogeneous scalar conservation law, and prove the conservation form of this semi-scheme satisfies the maximum principle.

Finally, we apply our 1D and 2D schemes to four numerical tests and show the results of the comparison between our numerical solutions and the exact solutions (or compared solution from [4]) in chapter 5.

Chapter 2

Theoretical Basis

2.1 Mathematical Basics

2.1.1 Derivations of Euler Equations

The Euler equation is a particular important example of a hyperbolic system of conservation laws. It presents the variation of mass , energy, and energy as time or position changes.

In 1-dimension, the form of Euler system is as follows :

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho \\ \rho u \\ E\end{bmatrix} + \frac{\partial}{\partial x}\begin{bmatrix}\rho u \\ \rho u^2 + p \\ (E+p)u\end{bmatrix} = 0. \tag{2.1}$$

Here $\rho, u, \rho u$, and p are used to denote the fluid's density, velocity, momentum, and pressure, respectively. E is the total energy. We will give more details regarding energy E in the later section.

In 2-dimensions, the Euler equations take the form

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho \\ \rho u \\ \rho v \\ E\end{bmatrix} + \frac{\partial}{\partial x}\begin{bmatrix}\rho u \\ \rho u^2 + p \\ \rho uv \\ (E+p)u\end{bmatrix} + \frac{\partial}{\partial y}\begin{bmatrix}\rho v \\ \rho uv \\ \rho v^2 + p \\ (E+p)v\end{bmatrix} = 0, \tag{2.2}$$

where u and v denote the velocity in $x-$ and y direction, respectively.

In N-dimensions, a Euler system is composed of $m = N + 2$ equations. The derivations of these Euler equations will be illustrated later. The content in section 2.1 refer to [1], [2], [13], [17], and [18].

Before discussing the derivations of Euler equations, we need to introduce the so-called ***Reynolds transport theorem***, which is also known as a ***three-dimensional generalization of the Leibniz integral rule***, and plays an important role in the following derivations.

2.1.1.1 Reynolds transport theorem

The general form of the Reynolds transport theorem can be expressed as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega(t)} \varphi(t,x)\,dx = \int_{\Omega(t)} \frac{\partial}{\partial t}\varphi(t,x)\,dx + \int_{\partial\Omega(t)} \varphi(t,x)v(t,x)\cdot n(t,x)\,ds_x, \tag{2.3}$$

where $n(t,x)$ is the outward-pointing unit normal vector, x is the fixed position, $\Omega(t)$ is the time-dependent domain, $\partial\Omega(t)$ is the boundary of Ω, and v is the velocity vector. The function φ is a continuously differentiable function on Eulerian coordinates (t,x).

This theorem can be visualized as: the change of the quantity in a time-dependent domain with respect to the time is equal to the sum of the change of the quantity itself over time and the flux across the boundary, which contains outflow and inflow.

Furthermore, we apply the divergence theorem to the second term on the right-hand-side of (2.3),

$$\int_{\Omega(t)} \nabla\cdot(\varphi(t,x)v(t,x))\,dx = \int_{\partial\Omega(t)} \varphi(t,x)v(t,x)\cdot n(t,x)\,ds_x, \tag{2.4}$$

and the theorem is then rewritten as:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega(t)} \varphi(t,x)\,dx = \int_{\Omega(t)} \left[\frac{\partial}{\partial t}\varphi(t,x) + \nabla\cdot(\varphi(t,x)v(t,x))\right] dx. \tag{2.5}$$

A more thorough proof can be found in Chapter 5.4 of [1] and Chapter 7 of [2].

Remark. φ can be scalar-, vector- or tensor valued.

2.1.1.2 Conservation of Mass

The principle of mass conservation states that the domain $\Omega(t)$ contains the same value of mass for all time t; in other words, the rate of the change of mass in domain $\Omega(t)$ is equal to zero:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega(t)} \rho(t,x)\,dx = 0. \tag{2.6}$$

Applying the Reynolds transport theorem, it implies

$$\int_{\Omega(t)} \left[\frac{\partial}{\partial t}\rho(t,x) + \nabla\cdot(\rho(t,x)v(t,x))\right] dx = 0. \tag{2.7}$$

This equation holds for all $\Omega(t)$. Assuming ρ and v are sufficiently smooth, we obtain from (2.7) that

$$\partial_t \rho + \nabla\cdot(\rho v) = 0, \tag{2.8}$$

Eq.(2.8) is the differential form of the law of conservation of mass, which is also known as the ***continuity equation***.

Remark. In d- dimensions, velocity v is composed of velocity components from d directions; i.e.,

$$v = \sum_{1}^{d} v_i \tag{2.9}$$

2.1.1.3 Conservation of Momentum

To derive the conservation of moment we recall the Newton's second law of motion:

$$ma = F, \qquad \text{or} \qquad \frac{\mathrm{d}}{\mathrm{d}t}(mv) = F. \tag{2.10}$$

Now we formulate (2.10) with the continuum mechanical variables: density ρ, velocity v, outer unit normal n, body force per unit mass f, and surface traction vector $T^{(n)}$, which is also called stress vector, as follows,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega(t)} \rho(t,x)v(t,x)\,dx = \int_{\Omega(t)} \rho(t,x)f(t,x)\,dx + \int_{\partial\Omega(t)} T^{(n)}(t,x)\,ds_x. \tag{2.11}$$

Applying Reynolds transport theorem (2.5) to the left-hand-side on (2.11), it becomes

$$\begin{aligned}\int_{\Omega(t)} \frac{\partial}{\partial t}(\rho(t,x)v_i(t,x)) + \nabla\cdot(\rho(t,x)v_i(t,x)v(t,x))\;dx \\ = \int_{\Omega(t)} \rho(t,x)f_i(t,x)\;dx + \int_{\partial\Omega(t)} T_i^{(n)}(t,x)\;ds_x.\end{aligned} \tag{2.12}$$

According to the Cauchy's stress theorem (see section 3.2.4 of [13]), the traction vector $T^{(n)}$ can be expressed by the ***(Cauchy) stress tensor*** $\sigma(t,x)$ and the unit normal n,

$$T^{(n)} = \sigma\cdot n \qquad \text{or} \qquad T_i^{(n)} = \sum_{j=1}^{d}\sigma_{ij}n_j \tag{2.13}$$

Then, with the help of the divergence theorem, we can rewrite the total surface force as

$$\int_{\partial\Omega(t)} T^{(n)}(t,x)\, ds_x = \int_{\partial\Omega(t)} \sigma(t,x)\cdot n(t,x)\, ds_x = \int_{\Omega(t)} \nabla\cdot\sigma(t,x)\, dx, \tag{2.14}$$

Subsequently, (2.12) becomes

$$\begin{aligned}\int_{\Omega(t)} \left[\frac{\partial}{\partial t}(\rho(t,x)v_i(t,x)) + \nabla\cdot(\rho(t,x)v_i(t,x)v(t,x))\right]\, dx \\ = \int_{\Omega(t)} \rho(t,x)f_i(t,x)\, dx + \int_{\Omega(t)} (\nabla\cdot\sigma(t,x))_i\, dx,\end{aligned} \tag{2.15}$$

where

$$(\nabla\cdot\sigma(t,x))_i = \sum_{k=1}^{d} \partial_{x_k}\sigma_{ik}(t,x). \tag{2.16}$$

Moving all the terms on right-hand-side of (2.15) to left-hand-side, we obtain

$$\begin{aligned}\int_{\Omega(t)} \frac{\partial}{\partial t}\big(\rho(t,x)v_i(t,x)) + \nabla\cdot(\rho(t,x)v_i(t,x)v(t,x)\big) \\ -(\nabla\cdot\sigma(t,x))_i - \rho(t,x)f_i(t,x)\, dx = 0.\end{aligned} \tag{2.17}$$

Eq. (2.17) implies the formula of conservation of momentum:

$$\partial_t\rho v_i + \nabla\cdot(\rho v_i v) - (\nabla\cdot\sigma)_i = \rho f_i \tag{2.18}$$

Remark. In inviscid fluid, the stress tensor σ takes on the form:

$$\sigma_{ij} = -p\delta_{ij}, \tag{2.19}$$

with pressure $p > 0$ and ***Kronecker delta*** δ_{ij}.

2.1.1.4 Conservation of Energy

According to the work - energy principle, the change of the total energy of a system is equivalent to the work done by the body force on the system and the stress tensor on boundary. That is,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega(t)} E(t,x)\, dx = \int_{\Omega(t)} \rho(t,x)f(t,x)v(t,x)\, dx + \int_{\partial\Omega(t)} \sigma(t,x)n(t,x)\cdot v(t,x)\, ds_x. \tag{2.20}$$

Similar to the previous derivations, we apply the Reynolds transport theorem (2.5) on left-hand-side and the divergence theorem on the second term of right-hand-side of (2.20), then get

$$\begin{aligned}\int_{\Omega(t)} \frac{\partial}{\partial t}E(t,x) + \nabla \cdot (E(t,x)v(t,x))\,dx \\ = \int_{\Omega(t)} \rho(t,x)f(t,x)v(t,x)\,dx + \int_{\Omega(t)} \nabla \cdot \sigma(t,x)v(t,x)\,dx.\end{aligned} \tag{2.21}$$

Hence,

$$\int_{\Omega(t)} \frac{\partial}{\partial t}E(t,x) + \nabla \cdot (E(t,x)v(t,x) - \sigma(t,x)v(t,x)) - \rho(t,x)f(t,x)v(t,x)\;dx = 0, \tag{2.22}$$

and it implies

$$\partial_t E + \nabla \cdot ((E - \sigma)v) = \rho f v. \tag{2.23}$$

We can also write it as

$$\partial_t E + \nabla \cdot ((E + p)v) = \rho f v, \tag{2.24}$$

for the case of inviscid fluid by substituting $-p$ for σ.

Remark. The total Energy E includes the kinetic energy $E_{kin} = \frac{1}{2}\rho u^2$ and internal energy $\varepsilon = \rho\epsilon$, i.e., $E = \varepsilon + E_{kin}$. The variable ϵ is called ***specific internal energy***, meaning internal energy per mass, and is related to density and pressure. The functional relation between the three thermodynamical quantities, ϵ, ρ and p, is known as ***equation of states*** (EoS).

Now from conservation of mass, balance of momentum and balance of energy, we obtain three partial differential equations (2.8), (2.18) and (2.24) respectively, and the so-called Euler equation is comprised of these equations. Furthermore, if there is no body force f acting on the system, i.e., setting $f = 0$ to (2.18) and (2.24), then the equations become

$$\begin{aligned}\partial_t \rho + \nabla \cdot (\rho v) &= 0 \\ \partial_t \rho v + \nabla \cdot (\rho v^2 + p) &= 0 \\ \partial_t E + \nabla \cdot ((E + p)v) &= 0,\end{aligned} \tag{2.25}$$

which are called *compressible Euler equations.* Gathering the three quantities, mass ρ, momentum ρv, and energy E into a vector, we can write the equations in the form as (2.1) and (2.2) .

2.1.2 Euler Equations with Source Term

Considering the body force f, we rewrite the equations (2.8), (2.18) and (2.24) in terms of 1-dimension as the vector form, then obtain

$$\frac{\partial}{\partial t}\begin{bmatrix}\rho\\ \rho u\\ E\end{bmatrix}+\frac{\partial}{\partial x}\begin{bmatrix}\rho u\\ \rho u^2+p\\ (E+p)u\end{bmatrix}=\begin{bmatrix}0\\ \rho f\\ \rho f u\end{bmatrix}. \tag{2.26}$$

Since the second and the third component of the vector on the right-hand-side are sourced from the body force f, the right-hand-side is called ***source term***. There are some common body forces, for example, gravity, electric forces, magnetic forces. In this thesis, we focus on the Euler equations with gravitational source term.

For later discussion, we need to bring in some notations related to gravity. One is the gravitational acceleration, which is usually denoted as g, and the other related notation is the gravitational potential $\phi \in C^1(\mathbb{R}, \mathbb{R})$. The relation between the gravitational acceleration g and potential ϕ is

$$g = -\nabla\phi, \tag{2.27}$$

i.e., the gravitational g can be defined as the negative spatial derivative of a corresponding gravitational potential.

2.1.3 Hydrostatic Equilibrium

In fluid mechanics, hydrostatic equilibrium states that the condition of a fluid is stationary, that is to say, the fluid is at rest or the flow velocity maintains a constant. According to the Newton's law of motion, this condition occurs when the sum of force acting on the system is equal to zero.

Therefore, to present this in mathematics form, we set velocity $u(x,t) = 0$ for all x and t in the (2.26) with gravity, then the equations related to density ρ and energy E reduce to zero:

$$\begin{aligned}\partial_t\rho &= 0\\ \partial_t E &= 0,\end{aligned} \tag{2.28}$$

and the remaining equation becomes

$$\partial_x p = \rho g. \tag{2.29}$$

This equation (2.29) is called ***hydrostatic equation***. By solving this equation and together with the corresponding equation of state, we can find a ***hydrostatic solution***, which is usually expressed by ρ and p. Having a hydrostatic solution is one important key to develop our main scheme in this thesis.

2.2 Numerical Methods

2.2.1 Kurganov-Tadmor Scheme

Kurganov and Tadmor introduced a modification of Nessyahu-Tadmor scheme (NT scheme) [5] in 2000. The difference between these two scheme is that in NT scheme, the fixed width control cell $[x_j, x_{j+1}]$ is considered to approximate the average over Riemann fans (see figure 2.1), while a narrower control cell is adopted in KT scheme (see figure 2.2). The width of the control cell in KT scheme is determined by the local wave speed. In this section, we follow the steps in [3].

scheme.png

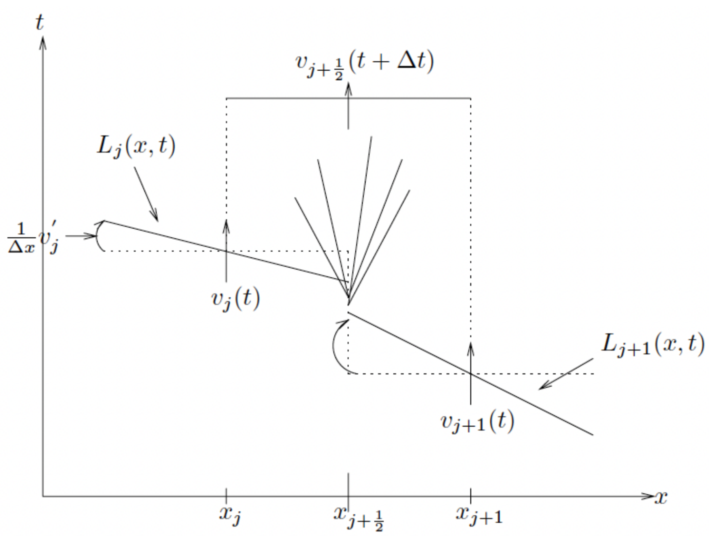
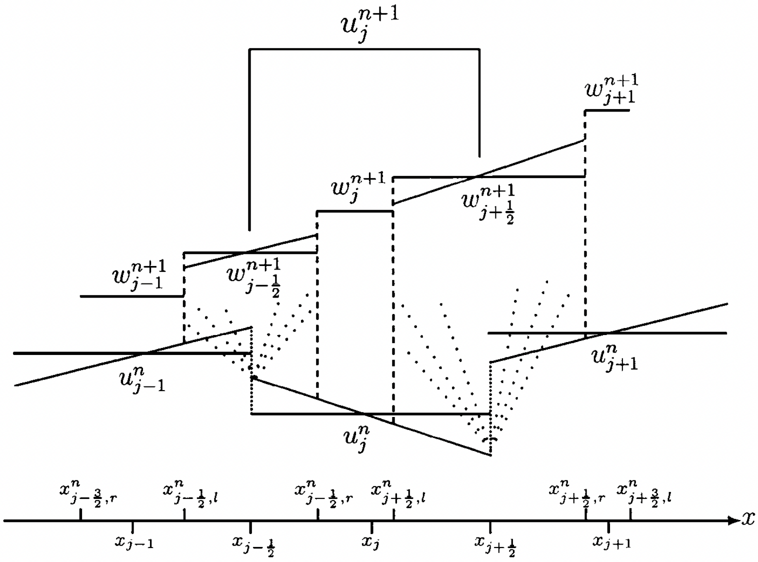


Figure 2.1: NT Scheme (Reference: Figure 2.3 in [5])



scheme.png

Figure 2.2: KT Scheme (Reference: Figure 3.2 in [3])

To illustrate the KT scheme, firstly we give the definition of the local wave speed $a^n_{j+\frac{1}{2}}$ at the cell boundary $x_{j+\frac{1}{2}}$ by

$$a^n_{j+\frac{1}{2}} = \max_{u \in C(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}})} \rho\left(\frac{\partial f}{\partial u}(u)\right), \tag{2.30}$$

where $u^-_{j+\frac{1}{2}} := u^n_j + \frac{\Delta x}{2}(u_x)^n_j$ and $u^+_{j+\frac{1}{2}} := u^n_{j+1} - \frac{\Delta x}{2}(u_x)^n_{j+1}$ are the corresponding left and right intermediate values of $u^n_{j+\frac{1}{2}}$ with the slope $(u_x)^n_j$, and $C(u^-_{j+\frac{1}{2}}, u^+_{j+\frac{1}{2}})$ is a curve in phase space connecting $u^-_{j+\frac{1}{2}}$ and $u^+_{j+\frac{1}{2}}$ via the Riemann fan. Here, $\rho(\frac{\partial f}{\partial u}(u))$ means the eigenvalues of the flux jacobian $\frac{\partial f(u)}{\partial u}$.

In most practical applications, the local wave speed can be simply evaluated by

$$a^n_{j+\frac{1}{2}} := \max\left\{\rho\Big(\frac{\partial f}{\partial u}(u^-_{j+\frac{1}{2}})\Big), \rho\Big(\frac{\partial f}{\partial u}(u^+_{j+\frac{1}{2}})\Big)\right\}. \tag{2.31}$$

Then we consider the interval $[x^n_{j+\frac{1}{2},l}, x^n_{j+\frac{1}{2},r}]$ with the definitions $x^n_{j+\frac{1}{2},l} := x_{j+\frac{1}{2}} - a^n_{j+\frac{1}{2}}\Delta t$ and $x^n_{j+\frac{1}{2},r} := x_{j+\frac{1}{2}} + a^n_{j+\frac{1}{2}}\Delta t$. Note that the points $x_{j+\frac{1}{2},l}$ and $x_{j+\frac{1}{2},r}$ split the staggered grid $[x_j, x_{j+1}]$ into smooth region and nonsmooth region. Hence, we should discuss their averages separately.

Proceeding with the evolution to the next time step t^{n+1}, the new cell-average of the nonsmooth region is approximated with a reconstruction $\widetilde{u}_j(x, t^n) = u^n_j +$

$(u_x)_j^n(x - x_j)$ by

$$\begin{aligned}
\overline{w}_{j+\frac{1}{2}}^{n+1} :=& \frac{1}{\Delta x_{j+\frac{1}{2}}^n} \int_{x_{j+\frac{1}{2},l}^n}^{x_{j+\frac{1}{2},r}^n} u(x, t^{n+1})\, dx \\
=& \frac{1}{\Delta x_{j+\frac{1}{2}}^n} \int_{x_{j+\frac{1}{2},l}^n}^{x_{j+\frac{1}{2},r}^n} u(x, t^n)\, dx \\
&\qquad - \frac{1}{\Delta x_{j+\frac{1}{2}}^n} \int_{t^n}^{t^{n+1}} [f(u(x_{j+\frac{1}{2},r}^n, t)) - f(u(x_{j+\frac{1}{2},l}^n, \tau))]\, dt \\
=& \frac{u_j^n + u_{j+1}^n}{2} + \frac{\Delta x - a_{j+\frac{1}{2}}^n \Delta t}{4}((u_x)_j^n - (u_x)_{j+1}^n) \\
&\qquad - \frac{1}{2a_{j+\frac{1}{2}}^n \Delta t} \int_{t^n}^{t^{n+1}} [f(u(x_{j+\frac{1}{2},r}^n, t)) - f(u(x_{j+\frac{1}{2},l}^n, t))]\, dt
\end{aligned} \tag{2.32}$$

where $\Delta x_{j+\frac{1}{2}}^n := x_{j+\frac{1}{2},r}^n - x_{j+\frac{1}{2},l}^n = 2a_{j+\frac{1}{2}}^n \Delta t$.

For the smooth region, the cell-average is evaluated by

$$\begin{aligned}
\overline{w}_j^{n+1} :=& \frac{1}{\Delta x_j^n} \int_{x_{j-\frac{1}{2},r}^n}^{x_{j+\frac{1}{2},l}^n} u(x, t^{n+1})\, dx \\
=& \frac{1}{\Delta x_j^n} \int_{x_{j-\frac{1}{2},r}^n}^{x_{j+\frac{1}{2},l}^n} u(x, t^n)\, dx - \frac{1}{\Delta x_j^n} \int_{t^n}^{t^{n+1}} [f(u(x_{j+\frac{1}{2},l}^n, \tau)) - f(u(x_{j-\frac{1}{2},r}^n, t))]\, dt \\
=& u_j^n + \frac{\Delta t}{2}(a_{j-\frac{1}{2}}^n - a_{j+\frac{1}{2}}^n)(u_x)_j^n \\
&\qquad - \frac{1}{\Delta x_j^n} \int_{t^n}^{t^{n+1}} [f(u(x_{j+\frac{1}{2},l}^n, t)) - f(u(x_{j-\frac{1}{2},r}^n, t))]\, dt
\end{aligned} \tag{2.33}$$

where $\Delta x_j^n := x_{j+\frac{1}{2},l}^n - x_{j-\frac{1}{2},r}^n = \Delta x - \Delta t(a_{j-\frac{1}{2}}^n + a_{j+\frac{1}{2}}^n)$.

Applying the midpoint rule to approximate the flux integral in (2.32) and (2.33), we obtain the intermediate values

$$\begin{aligned}
\overline{w}_{j+\frac{1}{2}}^{n+1} &= \frac{u_j^n + u_{j+1}^n}{2} + \frac{\Delta x - a_{j+\frac{1}{2}}^n \Delta t}{4}((u_x)_j^n - (u_x)_{j+1}^n) \\
&\qquad - \frac{1}{2a_{j+\frac{1}{2}}^n}[f(u_{j+\frac{1}{2},r}^{n+\frac{1}{2}}) - f(u_{j+\frac{1}{2},l}^{n+\frac{1}{2}})], \\
\overline{w}_j^{n+1} &= u_j^n + \frac{\Delta t}{2}(a_{j-\frac{1}{2}}^n - a_{j+\frac{1}{2}}^n)(u_x)_j^n - \frac{\lambda}{1 - \lambda(a_{j-\frac{1}{2}}^n + a_{j+\frac{1}{2}}^n)}[f(u_{j+\frac{1}{2},l}^{n+\frac{1}{2}}) - f(u_{j-\frac{1}{2},r}^{n+\frac{1}{2}})].
\end{aligned} \tag{2.34}$$

The values of midpoints in (2.34) can be obtained from the Taylor expansion:

$$\begin{aligned}
u_{j+\frac{1}{2},l}^{n+\frac{1}{2}} &:= u_{j+\frac{1}{2},l}^{n} - \frac{\Delta t}{2} f(u_{j+\frac{1}{2},l}^{n})_x, \quad u_{j+\frac{1}{2},l}^{n} := u_j^n + \Delta x (u_x)_j^n \left(\frac{1}{2} - \lambda a_{j+\frac{1}{2}}^n\right), \\
u_{j+\frac{1}{2},r}^{n+\frac{1}{2}} &:= u_{j+\frac{1}{2},r}^{n} - \frac{\Delta t}{2} f(u_{j+\frac{1}{2},r}^{n})_x, \quad u_{j+\frac{1}{2},r}^{n} := u_{j+1}^n - \Delta x (u_x)_{j+1}^n \left(\frac{1}{2} - \lambda a_{j+\frac{1}{2}}^n\right),
\end{aligned} \tag{2.35}$$

Now we have the intermediate cell-averages $\overline{w}_{j+\frac{1}{2}}^{n+1}$ and $\overline{w}_j^{n+1}$ at time t^{n+1}, the next thing to do is to project back to the original uniform grid. That is to say, we have to approximate the average over the unstaggered cell $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$. To avoid the oscillation, we need a reconstruction of $w_{j+\frac{1}{2}}^{n+1}$, denoted by $\widetilde{W}_{j+\frac{1}{2}}^{n+1}$. Note that since the smooth part is smooth enough, we only need to reconstruct the nonsmooth part.

The piecewise-linear reconstruction is in the terms of

$$\widetilde{W}_{j+\frac{1}{2}}^{n+1} := \overline{w}_{j+\frac{1}{2}}^{n+1} + (u_x)_{j+\frac{1}{2}}^{n+1}(x - x_{j+\frac{1}{2}}) \qquad \forall x \in [x_{j+\frac{1}{2},l}^n, x_{j+\frac{1}{2},r}^n] \tag{2.36}$$

Finally, the new average over uniform grid at time t^{n+1} is evaluated by

$$\begin{aligned}
u_j^{n+1} =& \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \widetilde{W}(x, t^{n+1})\, dx \\
=& \lambda a_{j-\frac{1}{2}}^n \overline{w}_{j-\frac{1}{2}}^{n+1} + [1 - \lambda(a_{j-\frac{1}{2}}^n + a_{j+\frac{1}{2}}^n)] \overline{w}_j^{n+1} \\
& + \lambda a_{j+\frac{1}{2}}^n \overline{w}_{j+\frac{1}{2}}^{n+1} + \frac{\Delta x}{2} [(\lambda a_{j-\frac{1}{2}}^n)^2 (u_x)_{j-\frac{1}{2}}^{n+1} - (\lambda a_{j+\frac{1}{2}}^n)^2 (u_x)_{j+\frac{1}{2}}^{n+1}]
\end{aligned} \tag{2.37}$$

2.2.2 Deviation Method

In this section we introduce a modified well-balanced finite volume method for balance laws from [14] and [15]. The main ingredient in this method is the given target solution. Instead of the original solution, the difference between the target solution and the original solution is considered when approximating the averages over cells.

Consider the 1-dimensional system of hyperbolic balance laws

$$\partial_t q(x,t) + \partial_x f(q(x,t)) = s(q(x,t), x). \tag{2.38}$$

Let $\widetilde{q}$ be a continuous and sufficiently smooth solution (the target solution) of (2.38), i.e.,

$$\partial_t \widetilde{q}(x,t) + \partial_x f(\widetilde{q}(x,t)) = s(\widetilde{q}(x,t), x). \tag{2.39}$$

Next, subtracting (2.39) from (2.38) results in

$$\begin{aligned}\partial_t \Delta q(x,t) + \partial_x \Big(f(\widetilde{q}(x,t) + \Delta q(x,t)) - f(\widetilde{q}(x,t)) \Big) \\ = s(\widetilde{q}(x,t) + \Delta q(x,t), x) - s(\widetilde{q}(x,t), x),\end{aligned} \tag{2.40}$$

where $\Delta q := q - \widetilde{q}$ is defined by the deviation between the desired solution q and the target solution $\widetilde{q}$.

Define the cell-average as

$$\hat{q}_i(t) := \frac{1}{\Delta x_i} \int_{\Omega_i} q(x,t)\, dx. \tag{2.41}$$

Averaging (2.40) in Ω_i yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\Delta \hat{q}_i(t)) = -\frac{1}{\Delta x_i} \Big[\Big(f((\Delta q + \tilde{q})(x_{i+\frac{1}{2}}, t)) - f(\tilde{q}(x_{i+\frac{1}{2}}, t)) \Big) \\ - \Big(f((\Delta q + \tilde{q})(x_{i-\frac{1}{2}}, t)) - f(\tilde{q}(x_{i-\frac{1}{2}}, t)) \Big) \Big] \\ + \frac{1}{\Delta x_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} s((\Delta q + \tilde{q})(x,t), x, t) - s(\tilde{q}(x,t), x, t)\, dx,\end{aligned} \tag{2.42}$$

where $\Delta \hat{q}_i := \hat{q}_i - \hat{\tilde{q}}_i$ is the cell-averaged deviation.

Using the standard discretization techniques to (2.42) leads to the semi-discrete scheme

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(\Delta \hat{Q}_i(t)) = -\frac{1}{\Delta x_i} \Big[\Delta F \left(\Delta Q^L_{i+\frac{1}{2}}(t), \Delta Q^R_{i+\frac{1}{2}}(t), \tilde{q}(x_{i+\frac{1}{2}}) \right) \\ - \Delta F \left(\Delta Q^L_{i-\frac{1}{2}}(t), \Delta Q^R_{i-\frac{1}{2}}(t), \tilde{q}(x_{i-\frac{1}{2}}) \right) \Big] + \Delta S_i((\Delta Q)^{rec}_i, \tilde{q}, t),\end{aligned} \tag{2.43}$$

with

$$\Delta F(\Delta Q^L, \Delta Q^R, \tilde{q}) := F(\Delta Q^L + \tilde{q}, \Delta Q^R + \tilde{q}) - f(\tilde{q}), \tag{2.44}$$

where ΔQ_i is the approximation to the cell-averaged deviation $\Delta \hat{q}_i$, and F is a numerical flux function consistent with f, and the reconstructed function $(\Delta Q)^{rec}_i$ are obtained from a m-th order accurate consistent conservative reconstruction. The interface value ΔQ^L and ΔQ^R are computed from ΔQ^{rec}_i. The discretization of source term difference ΔS is defined as

$$\Delta S_i((\Delta Q)^{rec}_i, \tilde{q}, t) := S_i(\Delta q + \tilde{q}, t) - S_i(\tilde{q}, t), \tag{2.45}$$

where

$$S_i(q,t) = \frac{1}{\Delta x_i}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} s(q(x,t),x,t)dx + O(h^m) \tag{2.46}$$

is some m-th order accurate source term discretization.

Notice that the numerical result solved from the scheme (2.43) is just the deviation solution ΔQ. In order to obtain the desired numerical solution q, we need to add the target solution $\tilde{q}$ and the deviation ΔQ to transfer it back to q.

In the following, we demonstrate the well-Balanced property for this deviation method (consult [15]).

Theorem 2.2.1 *The semi-discrete scheme* (2.43) *maintains the zeros state, i.e. the initial conditions* $\Delta Q_i(t=0) = 0$ *for all* $i \in I$ *lead to* $\Delta Q_i(t) = 0$ *for all times* $t > 0$.

Proof. Assume $\Delta\hat{Q}_i = 0$ for all $i \in I$. Due to the consistency, the reconstruction

$$\Delta Q_i^{rec} = 0 \tag{2.47}$$

for all $i \in I$. And $\Delta Q^L_{i+\frac{1}{2}}$ and $\Delta Q^R_{i+\frac{1}{2}}$ are also equal to zero. Hence, the numerical flux difference is equal to

$$\begin{aligned}\Delta F(\Delta Q^L_{i+\frac{1}{2}}, \Delta Q^R_{i+\frac{1}{2}}, \tilde{q}_{i+\frac{1}{2}}) &= F(\Delta Q^L_{i+\frac{1}{2}} + \tilde{q}_{i+\frac{1}{2}}, \Delta Q^R_{i+\frac{1}{2}} + \tilde{q}_{i+\frac{1}{2}}) - f(\tilde{q}_{i+\frac{1}{2}}) \\ &= F(\tilde{q}_{i+\frac{1}{2}}, \tilde{q}_{i+\frac{1}{2}}) - f(\tilde{q}_{i+\frac{1}{2}})\end{aligned} \tag{2.48}$$

Since F is a numerical flux consistent with f, it has a property that

$$F(q,q) = f(q). \tag{2.49}$$

Therefore, (2.48) leads to

$$\Delta F(\Delta Q^L_{i+\frac{1}{2}}, \Delta Q^R_{i+\frac{1}{2}}, \tilde{q}_{i+\frac{1}{2}}) = f(\tilde{q}_{i+\frac{1}{2}}) - f(\tilde{q}_{i+\frac{1}{2}}) = 0 \tag{2.50}$$

Next, we discuss the source term in the scheme (2.43).

Because of $\Delta Q_i = 0$ and by the definition (2.45), the source term

$$\Delta S_i((\Delta Q)_i^{rec}, \tilde{q}, t) = \Delta S_i(0, \tilde{q}, t) = S_i(\tilde{q}, t) - S_i(\tilde{q}, t) = 0, \tag{2.51}$$

vanishes. Since both flux term and source term in scheme (2.43) are equivalent to zeros,

$$\frac{\mathrm{d}}{\mathrm{d}t}\Delta\hat{Q} = 0. \tag{2.52}$$

$\blacksquare$

We say the semi-discrete scheme (2.43) is a well-balanced method when it satisfies the following property: *If the initial condition $\hat{Q}_i(t = 0)$ is equal to the target solution $\tilde{q}_i(t = 0)$, then the approximation solution $\hat{Q}_i(t)$ is equal to the target solution for all time $t > 0$.*

Consider the definition of $\hat{Q}_i$,

$$\hat{Q}_i = \hat{\tilde{q}}_i + \Delta\hat{Q}_i, \tag{2.53}$$

then the property follows directly from Theorem 2.2.1.

Chapter 3

1D Scheme

Inspired by the Deviation method [14] and the KT scheme [3], we combine the ideas of them and construct a new scheme. In section 3.1 we detail how we construct this new fully-discrete scheme, and then prove it satisfies well-balanced property in section 3.2.

3.1 Description

To construct the new scheme, we firstly follow the steps in [4] to construct the modified balanced law which is similar to the description of Deviation method in section 2.2.2, and then apply the framework of KT scheme in section 2.2.1.

Consider the 1D balance law

$$\begin{cases} q_t + f(q)_x = S(q,x), \quad x \in \Omega \subset \mathbb{R}, t > 0 \\ q(x,0) = q_0(x). \end{cases} \tag{3.1}$$

Let $\tilde{q}$ be a given hydrostatic solution of (3.1). In other word, it satisfies

$$f(\tilde{q})_x = S(\tilde{q}, x). \tag{3.2}$$

Next, subtracting (3.2) from (3.1) yields

$$q_t + f(q)_x - f(\tilde{q})_x = S(q,x) - S(\tilde{q},x). \tag{3.3}$$

Define the deviation $\Delta q = q - \tilde{q}$. Applying $q = \Delta q + \tilde{q}$ to (3.3) leads to

$$(\Delta q + \tilde{q})_t + [f(\Delta q + \tilde{q}) - f(\tilde{q})]_x = S(\Delta q + \tilde{q}, x) - S(\tilde{q}, x) \tag{3.4}$$

Since $\tilde{q}$ is the stationary solution; i.e., $\frac{\partial q}{\partial t} = 0$, it implies

$$(\Delta q)_t + [f(\Delta q + \tilde{q}) - f(\tilde{q})]_x = S(\Delta q + \tilde{q}, x) - S(\tilde{q}, x) \tag{3.5}$$

If the source term $S(q, x)$ in (3.1) is a linear functional in terms of the conserved variables, then

$$S(\Delta q + \tilde{q}, x) - S(\tilde{q}, x) = S(\Delta q, x) \tag{3.6}$$

holds and (3.5) can be rewritten as

$$(\Delta q)_t + [f(\Delta q + \tilde{q}) - f(\tilde{q})]_x = S(\Delta q, x). \tag{3.7}$$

Lemma 3.1.1 *Consider the balanced law* (3.1) *and a given hydrostatic solution* $\tilde{q}$*. The deviation quantity* Δq *satisfied the modified balanced law* (3.7) *maintains the same local speed at interface as the original one in* (3.1).

Proof. Define

$$F(\Delta q) = f(\Delta q + \tilde{q}) - f(\tilde{q}). \tag{3.8}$$

We rewrite (3.7) as

$$(\Delta q)_t + F(\Delta q)_x = S(\Delta q, x). \tag{3.9}$$

Consider the flux jacobian of (3.9), $\frac{\partial F(\Delta q)}{\partial \Delta q}$. According to the definition of Δq,

$$\frac{\partial \Delta q}{\partial q} = \frac{\partial}{\partial q}(q - \tilde{q}) = 1. \tag{3.10}$$

Thus, the following equality

$$\frac{\partial F(\Delta q)}{\partial q} = \frac{\partial F(\Delta q)}{\partial \Delta q}\frac{\partial \Delta q}{\partial q} = \frac{\partial F(\Delta q)}{\partial \Delta q} \tag{3.11}$$

holds. By the definition of $F(\Delta q)$ in (3.8) ,

$$\frac{\partial F(\Delta q)}{\partial q} = \frac{\partial}{\partial q}\big(f(\Delta q + \tilde{q}) - f(\tilde{q})\big) = \frac{\partial}{\partial q} f(\Delta q + \tilde{q}) = \frac{\partial f(q)}{\partial q}. \tag{3.12}$$

holds. Hence, we obtain

$$\frac{\partial F(\Delta q)}{\partial \Delta q} = \frac{\partial f(q)}{\partial q}. \tag{3.13}$$

Since the local speeds at interface are determined by the eigenvalues of the flux jacobian, (3.13) shows that the local speeds maintain the same in (3.1). ∎

Now we proceed to apply the idea of KT scheme to the modified balanced law (3.7). We show this process in three steps: Reconstruction, Evolution, and Projection.

3.1.1 Reconstruction

To avoid the oscillation, a piecewise-linear reconstruction is needed when evaluating the average of $\Delta q(x,t^n)$ over the cells $C_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, and it is defined by

$$Q_j(x,t^n) = (\Delta q)_j^n + ((\Delta q)_x)_j^n (x - x_j), \tag{3.14}$$

where $((\Delta q)_x)_j^n$ is the numerical spatial derivative.
The so-called $MC-\theta$ limiter is a common choice to compute the numerical derivatives on (3.14) , and it is defined as follows,

$$((\Delta q)_x)_j^n = \text{minmod}\left[\theta\frac{(\Delta q)_{j+1}^n - (\Delta q)_j^n}{\Delta x}, \frac{(\Delta q)_{j+1}^n - (\Delta q)_{j-1}^n}{2\Delta x}, \theta\frac{(\Delta q)_j^n - (\Delta q)_{j-1}^n}{\Delta x}\right], \tag{3.15}$$

with $1 \le \theta \le 2$, while the minmod function is defined as

$$\text{minmod}(a,b,c) := \begin{cases} \text{sign}(a)\min\{|a|,|b|,|c|\}, & \text{if } \text{sign}(a) = \text{sign}(b) = \text{sign}(c) \\ 0, & \text{otherwise.} \end{cases} \tag{3.16}$$

This $MC-\theta$ will be adopted for the numerical tests in chapter 5.

3.1.2 Evolution

Next, we evolve the equation to the next time step $t^{n+1} = t + \Delta t$ and approximate the averages over the cells. Looking at the figure 2.2 again, the uniform cell is divided into unsmooth and smooth regions by the interpolations $x^n_{j-\frac{1}{2},r}$ and $x^n_{j+\frac{1}{2},l}$, which are defined in the same way as section 2.2.1 by that

$$x^n_{j+\frac{1}{2},l} := x_{j+\frac{1}{2}} - a^n_{j+\frac{1}{2}}\Delta t \qquad \text{and} \qquad x^n_{j+\frac{1}{2},r} := x_{j+\frac{1}{2}} + a^n_{j+\frac{1}{2}}\Delta t \tag{3.17}$$

with the wave speed $a^n_{j+\frac{1}{2}}$ defined in (2.30). For this reason, we consider two cases: in the first case, we discuss the average over the unsmooth interval $U^n_{j+\frac{1}{2}} = [x^n_{j+\frac{1}{2},l}, x^n_{j+\frac{1}{2},r}]$; in the second case, we discuss the average over the smooth interval $M^n_j = [x^n_{j-\frac{1}{2},r}, x^n_{j+\frac{1}{2},l}]$.
Case 1. Unsmooth region.

We firstly integrate (3.7) over $U^n_{j+\frac{1}{2}} \times [t^n, t^{n+1}]$,

$$\int_{t^n}^{t^n+1}\int_{U^n_{j+\frac{1}{2}}} (\Delta q)_t + [f(\Delta q + \tilde{q}) - f(\tilde{q})]_x \, dxdt = \int_{t^n}^{t^n+1}\int_{U^n_{j+\frac{1}{2}}} S(\Delta q, x)\, dxdt. \tag{3.18}$$

Applying the Green's theorem to the left-hand-side of (3.18) leads to

$$\oint_{\partial(UT)^n_{j+\frac{1}{2}}} \{[f(\Delta q+\tilde{q})-f(\tilde{q})]\,dt-\Delta q\,dx\} = \int_{t^n}^{t^{n+1}}\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} S(\Delta q,x)\,dxdt. \tag{3.19}$$

where $(UT)^n_{j+\frac{1}{2}} = U^n_{j+\frac{1}{2}}\times[t^n,t^{n+1}]$.

The (3.19) is then computed as follows,

$$\begin{aligned}
&\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}}\Big\{\Big[f(\Delta q(x,t^n)+\tilde{q}(x,t^n))-f(\tilde{q}(x,t^n))\Big]\,dt-\Delta q(x,t^n)\,dx\Big\}\\
&\quad+\int_{t_n}^{t_{n+1}}\Big\{\Big[f(\Delta q(x^n_{j+\frac{1}{2},r},t)+\tilde{q}(x^n_{j+\frac{1}{2},r},t))-f(\tilde{q}(x^n_{j+\frac{1}{2},r},t))\Big]\,dt\\
&\qquad\qquad-\Delta q(x^n_{j+\frac{1}{2},r},t)\,dx\Big\}\\
&\quad+\int_{x^n_{j+\frac{1}{2},r}}^{x^n_{j+\frac{1}{2},l}}\Big\{\Big[f(\Delta q(x,t^{n+1})+\tilde{q}(x,t^{n+1}))-f(\tilde{q}(x,t^{n+1}))\Big]\,dt\\
&\qquad\qquad-\Delta q(x,t^{n+1})\,dx\Big\}\\
&\quad+\int_{t_{n+1}}^{t_n}\Big\{\Big[f(\Delta q(x^n_{j+\frac{1}{2},l},t)+\tilde{q}(x^n_{j+\frac{1}{2},l},t))-f(\tilde{q}(x^n_{j+\frac{1}{2},l},t))\Big]\,dt\\
&\qquad\qquad-\Delta q(x^n_{j+\frac{1}{2},l},t)\,dx\Big\}\\
=&\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}}\Big\{\Big[f(\Delta q(x,t^n)+\tilde{q}(x,t^n))-f(\tilde{q}(x,t^n))\Big]\,dt-\Delta q(x,t^n)\,dx\Big\}\\
&\quad+\int_{t_n}^{t_{n+1}}\Big\{\Big[f(\Delta q(x^n_{j+\frac{1}{2},r},t)+\tilde{q}(x^n_{j+\frac{1}{2},r},t))-f(\tilde{q}(x^n_{j+\frac{1}{2},r},t))\Big]\,dt\\
&\qquad\qquad-\Delta q(x_{j+\frac{1}{2},r},t)\,dx\Big\}\\
&\quad-\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}}\Big\{\Big[f(\Delta q(x,t^{n+1})+\tilde{q}(x,t^{n+1}))-f(\tilde{q}(x,t^{n+1}))\Big]\,dt\\
&\qquad\qquad-\Delta q(x,t^{n+1})\,dx\Big\}\\
&\quad-\int_{t_n}^{t_{n+1}}\Big\{\Big[f(\Delta q(x^n_{j+\frac{1}{2},l},t)+\tilde{q}(x^n_{j+\frac{1}{2},l},t))-f(\tilde{q}(x^n_{j+\frac{1}{2},l},t))\Big]\,dt\\
&\qquad\qquad-\Delta q(x^n_{j+\frac{1}{2},l},t)\,dx\Big\}\\
=&-\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}}\Delta q(x,t^n)\,dx\\
&\quad+\int_{t_n}^{t_{n+1}}\Big[f(\Delta q(x^n_{j+\frac{1}{2},r},t)+\tilde{q}(x^n_{j+\frac{1}{2},r},t))-f(\tilde{q}(x^n_{j+\frac{1}{2},r},t))\Big]\,dt
\end{aligned}$$

$$
\begin{aligned}
&+\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} \Delta q(x,t^{n+1})\,dx \\
&-\int_{t_n}^{t_{n+1}} \left[f(\Delta q(x^n_{j+\frac{1}{2},l},t)+\tilde{q}(x^n_{j+\frac{1}{2},l},t))-f(\tilde{q}(x^n_{j+\frac{1}{2},l},t))\right]dt \\
=&\int_{t^n}^{t^{n+1}}\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} S(\Delta q,x)\,dxdt. \qquad (3.20)
\end{aligned}
$$

Consider the last equality on (3.20). Leaving $\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} \Delta q(x,t^{n+1})\,dx$ on the left-hand-side and moving other terms to the right-hand-side, the equation is reformulated as

$$
\begin{aligned}
&\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} \Delta q(x,t^{n+1})dx \\
&\quad=\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} \Delta q(x,t^{n})dx \\
&\quad-\int_{t_n}^{t_{n+1}}[f(\Delta q(x^n_{j+\frac{1}{2},r},t)+\tilde{q}(x^n_{j+\frac{1}{2},r},t))-f(\tilde{q}(x^n_{j+\frac{1}{2},r},t))]dt \qquad (3.21)\\
&\quad+\int_{t_n}^{t_{n+1}}[f(\Delta q(x^n_{j+\frac{1}{2},l},t)+\tilde{q}(x^n_{j+\frac{1}{2},l},t))-f(\tilde{q}(x^n_{j+\frac{1}{2},l},t))]dt \\
&\quad+\int_{t^n}^{t^{n+1}}\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} S(\Delta q,x)\,dxdt.
\end{aligned}
$$

Define $\Delta x^n_{j+\frac{1}{2}}=x^n_{j+\frac{1}{2},r}-x^n_{j+\frac{1}{2},l}=2a^n_{j+\frac{1}{2}}\Delta t$, and denote $f(\Delta q(x^n_{j+\frac{1}{2},r},t)+\tilde{q}(x^n_{j+\frac{1}{2},r},t))$ by $f((\Delta q+\tilde{q})(x^n_{j+\frac{1}{2},r},t))$. In order to get the average over the the interval $U^n_{j+\frac{1}{2}}$, denoted as $\overline{w}^{n+1}_{j+\frac{1}{2}}$, we divide the size $\Delta x^n_{j+\frac{1}{2}}$ into (3.21)

$$
\begin{aligned}
\overline{w}^{n+1}_{j+\frac{1}{2}} :=&\frac{1}{\Delta x^n_{j+\frac{1}{2}}}\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} \Delta q(x,t^{n+1})dx \\
=&\frac{1}{\Delta x^n_{j+\frac{1}{2}}}\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} \Delta q(x,t^{n})dx \\
&-\frac{1}{\Delta x^n_{j+\frac{1}{2}}}\int_{t_n}^{t_{n+1}}[f((\Delta q+\tilde{q})(x^n_{j+\frac{1}{2},r},t))-f(\tilde{q}(x^n_{j+\frac{1}{2},r},t))] \\
&+\frac{1}{\Delta x^n_{j+\frac{1}{2}}}\int_{t_n}^{t_{n+1}}[f((\Delta q+\tilde{q})(x^n_{j+\frac{1}{2},l},t))-f(\tilde{q}(x^n_{j+\frac{1}{2},l},t))]dt \qquad (3.22)\\
&+\frac{1}{\Delta x^n_{j+\frac{1}{2}}}\int_{t^n}^{t^{n+1}}\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} S(\Delta q,x)\,dxdt \\
=&\frac{1}{\Delta x^n_{j+\frac{1}{2}}}\left[\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} \Delta q(x,t^{n})dx+F_U+S_U\right],
\end{aligned}
$$

where

$$\begin{aligned} F_U = &-\int_{t_n}^{t_{n+1}} [f((\Delta q + \tilde{q})(x^n_{j+\frac{1}{2},r}, t)) - f(\tilde{q}(x^n_{j+\frac{1}{2},r}, t))]dt \\ &+\int_{t_n}^{t_{n+1}} [f(\Delta q + \tilde{q})(x^n_{j+\frac{1}{2},l}, t)) - f(\tilde{q}(x^n_{j+\frac{1}{2},l}, t))]dt, \end{aligned} \tag{3.23}$$

and

$$S_U = \int_{t^n}^{t^{n+1}} \int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} S(\Delta q, x)\, dxdt \tag{3.24}$$

Now we consider the first term on the last line of (3.22). According to the mid-point rule and the conservation property of the reconstruction Q , we have the following approximations

$$\int_{x_j}^{x_{j+1}} \Delta q(x, t^n)dx := \Delta x \Delta q(x_{j+\frac{1}{2}}, t^n), \tag{3.25}$$

$$\int_{x_j}^{x_{j+1}} \Delta q(x, t^n)dx = \int_{x_j}^{x_{j+1}} Q(x, t^n)dx := \Delta x Q_j(x_{j+\frac{1}{2}}, t^n). \tag{3.26}$$

Then we obtain

$$\begin{aligned} &\frac{1}{\Delta x^n_{j+\frac{1}{2}}} \int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} \Delta q(x, t^n)dx \\ &= \frac{1}{\Delta x^n_{j+\frac{1}{2}}} \left[\int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2}}} \Delta q(x, t^n)dx + \int_{x^n_{j+\frac{1}{2}}}^{x^n_{j+\frac{1}{2},r}} \Delta q(x, t^n)dx \right] \\ &= \frac{1}{\Delta x^n_{j+\frac{1}{2}}} \left[\frac{\Delta x^n_{j+\frac{1}{2}}}{2} Q_j(x^n_{j+\frac{1}{2},lm}, t^n) + \frac{\Delta x^n_{j+\frac{1}{2}}}{2} Q_{j+1}(x^n_{j+\frac{1}{2},rm}, t^n) \right] \\ &= \frac{1}{\Delta x^n_{j+\frac{1}{2}}} \left[\frac{\Delta x^n_{j+\frac{1}{2}}}{2} \Bigg((\Delta q)^n_j + (x^n_{j+\frac{1}{2},lm} - x_j)((\Delta q)_x)^n_j) \right. \\ &\qquad \left. + (\Delta q)^n_{j+1} + (x^n_{j+\frac{1}{2},rm} - x_{j+1})((\Delta q)_x)^n_{j+1}) \Bigg) \right] \\ &= \frac{1}{\Delta x^n_{j+\frac{1}{2}}} \left[\frac{\Delta x^n_{j+\frac{1}{2}}}{2} \Bigg((\Delta q)^n_j + (\frac{\Delta x}{2} - \frac{\Delta x^n_{j+\frac{1}{2}}}{4})((\Delta q)_x)^n_j) \right. \\ &\qquad \left. + (\Delta q)^n_{j+1} - (\frac{\Delta x}{2} - \frac{\Delta x^n_{j+\frac{1}{2}}}{4})((\Delta q)_x)^n_{j+1}) \Bigg) \right] \\ &= \frac{(\Delta q)^n_j + (\Delta q)^n_{j+1}}{2} + \frac{\Delta x - \frac{\Delta x^n_{j+\frac{1}{2}}}{2}}{4} \Bigg[((\Delta q)_x)^n_j - ((\Delta q)_x)^n_{j+1} \Bigg], \end{aligned} \tag{3.27}$$

where $x^n_{j+\frac{1}{2},lm}$ denote the midpoint between $x^n_{j+\frac{1}{2},l}$ and $x^n_{j+\frac{1}{2}}$, and $x^n_{j+\frac{1}{2},rm}$ denote the midpoint between $x^n_{j+\frac{1}{2}}$ and $x^n_{j+\frac{1}{2},r}$.

Next, we consider the flux function F_U. Applying the midpoint rule to F_U, we get

$$
\begin{aligned}
F_U = \Delta t \Big[& -f((\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},r} + \tilde{q}_{j+\frac{1}{2},r}) + f((\tilde{q})_{j+\frac{1}{2},r}) \\
& + f((\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},l} + \tilde{q}_{j+\frac{1}{2},l}) - f((\tilde{q})_{j+\frac{1}{2},l}) \Big].
\end{aligned} \tag{3.28}
$$

The midpoint values $(\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},l}$ and $(\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},r}$ can be approximated by the Taylor expansion:

$$
\begin{aligned}
(\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},l} &:= (\Delta q)^{n}_{j+\frac{1}{2},l} + \frac{\Delta t}{2}((\Delta q)_t)^{n}_{j+\frac{1}{2},l} \\
&= (\Delta q)^{n}_{j+\frac{1}{2},l} + \frac{\Delta t}{2}\left[-\big[(f(\Delta q + \tilde{q}) - f(\tilde{q}))_x\big]_{(x^n_{j+\frac{1}{2},l}, t^n)} + S((\Delta q)_{j+\frac{1}{2},l}, t^n) \right],
\end{aligned} \tag{3.29}
$$

$$
\begin{aligned}
(\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},r} &:= (\Delta q)^{n}_{j+\frac{1}{2},r} + \frac{\Delta t}{2}((\Delta q)_t)^{n}_{j+\frac{1}{2},r} \\
&= (\Delta q)^{n}_{j+\frac{1}{2},r} + \frac{\Delta t}{2}\left[-\big[(f(\Delta q + \tilde{q}) - f(\tilde{q}))_x\big]_{(x^n_{j+\frac{1}{2},r}, t^n)} + S((\Delta q)_{j+\frac{1}{2},r}, t^n) \right],
\end{aligned} \tag{3.30}
$$

and

$$
\begin{aligned}
(\Delta q)^{n}_{j+\frac{1}{2},l} &:= (\Delta q)^{n}_{j} + \Delta x((\Delta q)_x)^{n}_{j}\left[\frac{1}{2} - \lambda a^n_{j+\frac{1}{2}}\right], \\
(\Delta q)^{n}_{j+\frac{1}{2},r} &:= (\Delta q)^{n}_{j+1} - \Delta x((\Delta q)_x)^{n}_{j+1}\left[\frac{1}{2} - \lambda a^n_{j+\frac{1}{2}}\right],
\end{aligned} \tag{3.31}
$$

with the mesh ratio $\lambda = \frac{\Delta t}{\Delta x}$.

The last approximation of the source term S_U is simply obtained by using the midpoint rule on time and the trapezoidal rule on space,

$$
S_U := \Delta t \int_{x^n_{j+\frac{1}{2},l}}^{x^n_{j+\frac{1}{2},r}} S(\Delta q^{n+\frac{1}{2}}, x)dx := \Delta t \Delta x^n_{j+\frac{1}{2}} \left[\frac{S^{n+\frac{1}{2}}_{j+\frac{1}{2},l} + S^{n+\frac{1}{2}}_{j+\frac{1}{2},r}}{2} \right]. \tag{3.32}
$$

Equipped with the average at t^n (3.27), the approximation of the flux term (3.28), and the source term (3.32), the intermediate average over the interval $U^n_{j+\frac{1}{2}}$ is equal to

$$
\begin{aligned}
\overline{w}^{n+1}_{j+\frac{1}{2}} = \frac{(\Delta q)^n_j + (\Delta q)^n_{j+1}}{2} &+ \frac{\Delta x - \frac{\Delta x^n_{j+\frac{1}{2}}}{2}}{4}\Big[((\Delta q)_x)^n_j - ((\Delta q)_x)^n_{j+1}\Big] \\
&+ \frac{\Delta t}{\Delta x^n_{j+\frac{1}{2}}}\Big[- f((\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},r} + \tilde{q}_{j+\frac{1}{2},r}) + f((\tilde{q})_{j+\frac{1}{2},r}) \\
&\qquad + f((\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},l} + \tilde{q}_{j+\frac{1}{2},l}) - f((\tilde{q})_{j+\frac{1}{2},l})\Big] \\
&+ \Delta t\left[\frac{S^{n+\frac{1}{2}}_{j+\frac{1}{2},l} + S^{n+\frac{1}{2}}_{j+\frac{1}{2},r}}{2}\right].
\end{aligned}
\tag{3.33}
$$

Case 2. Smooth region.

We consider the interval M^n_j. In a similar way, integrating (3.7) over $M^n_j \times [t^n, t^{n+1}]$, we obtain

$$
\oint_{\partial M^n_j} \{[f(\Delta q + \tilde{q}) - f(\tilde{q})]\, dt - \Delta q\, dx\} = \int_{t^n}^{t^{n+1}} \int_{x^n_{j-\frac{1}{2},r}}^{x^n_{j+\frac{1}{2},l}} S(\Delta q, x)\, dxdt, \tag{3.34}
$$

Define $\Delta x^n_j = x^n_{j+\frac{1}{2},l} - x^n_{j-\frac{1}{2},r} = \Delta x - \Delta t(a^n_{j-\frac{1}{2}} - a^n_{j+\frac{1}{2}})$. Then using the same calculation steps as in case 1, the average over M^n_j denoted as $\overline{w}^{n+1}_j$ is approximated by

$$
\begin{aligned}
\overline{w}^{n+1}_j &:= \frac{1}{\Delta x^n_j}\int_{x^n_{j-\frac{1}{2},r}}^{x^n_{j+\frac{1}{2},l}} \Delta q(x, t^{n+1})dx \\
&= \frac{1}{\Delta x^n_j}\left[\int_{x^n_{j-\frac{1}{2},r}}^{x^n_{j+\frac{1}{2},l}} \Delta q(x, t^n)dx + F_M + S_M\right] \\
&= \frac{1}{\Delta x^n_j}\left[\Delta x^n_j Q_j(x^n_{j,m}, t^n) + F_M + S_M\right] \\
&= \left[(\Delta q)^n_j + (x^n_{j,m} - x_j)((\Delta q)_x)^n_j\right] + \frac{1}{\Delta x^n_j}\left[F_M + S_M\right] \\
&= (\Delta q)^n_j + \frac{\Delta x^n_{j-\frac{1}{2}} - \Delta x^n_{j+\frac{1}{2}}}{4}((\Delta q)_x)^n_j + \frac{1}{\Delta x^n_j}\left[F_M + S_M\right],
\end{aligned}
\tag{3.35}
$$

where $x^n_{j,m}$ denotes the midpoint between $x^n_{j-\frac{1}{2},r}$ and $x^n_{j+\frac{1}{2},l}$, and

$$
\begin{aligned}
F_M = &-\int_{t_n}^{t_{n+1}} [f((\Delta q + \tilde{q})(x^n_{j+\frac{1}{2},l}, t)) - f(\tilde{q}(x^n_{j+\frac{1}{2},l}, t))]\, dt \\
&+ \int_{t_n}^{t_{n+1}} [f(\Delta q + \tilde{q})(x^n_{j-\frac{1}{2},r}, t)) - f(\tilde{q}(x^n_{j-\frac{1}{2},r}, t))]\, dt,
\end{aligned}
\tag{3.36}
$$

and

$$
S_M = \int_{t^n}^{t^{n+1}} \int_{x^n_{j-\frac{1}{2},r}}^{x^n_{j+\frac{1}{2},l}} S(\Delta q, x)\, dxdt. \tag{3.37}
$$

Likewise, the approximation of the flux term F_M and the source term S_M are obtained with the help of the Taylor expansion, the midpoint rule and the trapezoidal rule. Hence,

$$\begin{aligned} F_M = \Delta t \Big[& -f((\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},l} + \tilde{q}_{i+\frac{1}{2},l}) + f(\tilde{q}_{j+\frac{1}{2},l}) \\ & + f((\Delta q)^{n+\frac{1}{2}}_{j-\frac{1}{2},r} + \tilde{q}_{j-\frac{1}{2},r}) - f(\tilde{q}_{j-\frac{1}{2},r}) \Big] \end{aligned} \tag{3.38}$$

and

$$S_M := \Delta t \int_{x_{j-\frac{1}{2},r}}^{x_{j+\frac{1}{2},l}} S(\Delta q^{n+\frac{1}{2}}, x)dx := \Delta t \Delta x_j^n \left[\frac{S^{n+\frac{1}{2}}_{j-\frac{1}{2},r} + S^{n+\frac{1}{2}}_{j+\frac{1}{2},l}}{2} \right] \tag{3.39}$$

Equipped with (3.35), (3.38) and (3.39), the intermediate average over the interval M_j^n is equal to

$$\begin{aligned} \overline{w}_j^{n+1} = (\Delta q)_j^n & + \frac{\Delta x^n_{j-\frac{1}{2}} - \Delta x^n_{j+\frac{1}{2}}}{4} ((\Delta q)_x)_j^n \\ & + \frac{\Delta t}{\Delta x_j^n} \Big[- f((\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},l} + \tilde{q}_{i+\frac{1}{2},l}) + f(\tilde{q}_{j+\frac{1}{2},l}) \\ & \qquad + f((\Delta q)^{n+\frac{1}{2}}_{j-\frac{1}{2},r} + \tilde{q}_{j-\frac{1}{2},r}) - f(\tilde{q}_{j-\frac{1}{2},r}) \Big] \\ & + \Delta t \left[\frac{S^{n+\frac{1}{2}}_{j-\frac{1}{2},r} + S^{n+\frac{1}{2}}_{j+\frac{1}{2},l}}{2} \right]. \end{aligned} \tag{3.40}$$

3.1.3 Projection

Finally, the last procedure is to predict back to the average over the original uniform cell $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$. Due to the same reason that we want to avoid oscillations, an appropriate piecewise reconstruction is needed for the unsmooth region, and it takes the form that

$$\widetilde{W}_{j+\frac{1}{2}}(x, t^{n+1}) := \overline{w}^{n+1}_{j+\frac{1}{2}} + (x - x_{j+\frac{1}{2}})(w_x)^{n+1}_{j+\frac{1}{2}}, \qquad \forall x \in [x^n_{j+\frac{1}{2},l}, x^n_{j+\frac{1}{2},r}], \tag{3.41}$$

where the slope w_x is given by

$$(w_x)^{n+1}_{j+\frac{1}{2}} = \operatorname{minmod}\Big(\theta \frac{w^{n+1}_{j+\frac{1}{2}} - w^{n+1}_j}{x_{j+\frac{1}{2}} - x^n_{j,m}}, \frac{w^{n+1}_{j+1} - w^{n+1}_j}{x^n_{j+1,m} - x^n_{j,m}}, \theta \frac{w^{n+1}_{j+1} - w^{n+1}_{j+\frac{1}{2}}}{x^n_{j+1,m} - x_{j+\frac{1}{2}}}\Big), \quad 1 \le \theta \le 2. \tag{3.42}$$

Consequently, the average over $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ at time t^{n+1} denoted as $(\Delta q)_j^{n+1}$ is obtained after a simple computation,

$$
\begin{aligned}
(\Delta q)_j^{n+1} &:= \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \overline{w}\,dx \\
&= \frac{1}{\Delta x}\left[\int_{x_{j-\frac{1}{2}}}^{x^n_{j-\frac{1}{2},r}} \overline{w}^{n+1}_{j-\frac{1}{2}} + \int_{x^n_{j-\frac{1}{2},r}}^{x^n_{j+\frac{1}{2},l}} \overline{w}^n_j + \int_{x^n_{j+\frac{1}{2},l}}^{x_{j+\frac{1}{2}}} \overline{w}^{n+1}_{j+\frac{1}{2}}\right] dx \\
&= \frac{1}{\Delta x}\left[\int_{x_{j-\frac{1}{2}}}^{x^n_{j-\frac{1}{2},r}} \widetilde{W}^{n+1}_{j-\frac{1}{2}} + \int_{x^n_{j-\frac{1}{2},r}}^{x^n_{j+\frac{1}{2},l}} \overline{w}^n_j + \int_{x^n_{j+\frac{1}{2},l}}^{x_{j+\frac{1}{2}}} \widetilde{W}^{n+1}_{j+\frac{1}{2}}\right] dx \\
&= \frac{1}{\Delta x}\left[\frac{\Delta x^n_{j-\frac{1}{2}}}{2}\left(\overline{w}^{n+1}_{j-\frac{1}{2}} + (x^n_{j-\frac{1}{2},rm} - x_{j-\frac{1}{2}})(w_x)^{n+1}_{j-\frac{1}{2}}\right)\right. \\
&\qquad + \left(\Delta x - \frac{\Delta x^n_{j-\frac{1}{2}} + \Delta x^n_{j+\frac{1}{2}}}{2}\right)\overline{w}^{n+1}_j \\
&\qquad \left. + \frac{\Delta x^n_{j+\frac{1}{2}}}{2}\left(\overline{w}^{n+1}_{j+\frac{1}{2}} + (x^n_{j+\frac{1}{2},lm} - x_{j+\frac{1}{2}})(w_x)^{n+1}_{j+\frac{1}{2}}\right)\right] \\
&= \frac{1}{\Delta x}\left[\frac{\Delta x^n_{j-\frac{1}{2}}}{2}\left(\overline{w}^{n+1}_{j-\frac{1}{2}} + \frac{\Delta x^n_{j-\frac{1}{2}}}{4}(w_x)^{n+1}_{j-\frac{1}{2}}\right)\right. \\
&\qquad + \left(\Delta x - \frac{\Delta x^n_{j-\frac{1}{2}} + \Delta x^n_{j+\frac{1}{2}}}{2}\right)\overline{w}^{n+1}_j \\
&\qquad \left. + \frac{\Delta x^n_{j+\frac{1}{2}}}{2}\left(\overline{w}^{n+1}_{j+\frac{1}{2}} - \frac{\Delta x^n_{j+\frac{1}{2}}}{4}(w_x)^{n+1}_{j+\frac{1}{2}}\right)\right] \\
&= \frac{1}{\Delta x}\left[a^n_{j-\frac{1}{2}}\Delta t\left(\overline{w}^{n+1}_{j-\frac{1}{2}} + \frac{a^n_{j-\frac{1}{2}}\Delta t}{2}(w_x)^{n+1}_{j-\frac{1}{2}}\right)\right. \\
&\qquad + \left(\Delta x - (a^n_{j-\frac{1}{2}} + a^n_{j+\frac{1}{2}})\Delta t\right)\overline{w}^{n+1}_j \\
&\qquad \left. + a^n_{j+\frac{1}{2}}\Delta t\left(\overline{w}^{n+1}_{j+\frac{1}{2}} - \frac{a^n_{j+\frac{1}{2}}\Delta t}{2}(w_x)^{n+1}_{j+\frac{1}{2}}\right)\right].
\end{aligned} \tag{3.43}
$$

To get the desired numerical solution q, we just need to add the given stationary solution $\tilde{q}$,

$$q_j^{n+1} = (\Delta q)_j^{n+1} + \tilde{q}_j. \tag{3.44}$$

3.2 Well-Balanced Property

In this section we demonstrate that the new scheme presented in the previous section satisfies the well-balanced property, which means the numerical solution

$q_j, \forall j$, maintain stationary, i.e.,

$$q_j(t=0) = \tilde{q}_j(t=0) \quad \Rightarrow \quad q_j(t) = \tilde{q}_j(t), \quad \forall t. \tag{3.45}$$

By the definition of the deviation Δq, the initial condition $q_j^0 = \tilde{q}_j^0$ implies $(\Delta q)_j^0 = 0$. Subsequently, both $(\Delta q)^0_{j+\frac{1}{2},r}$ and $(\Delta q)^0_{j+\frac{1}{2},l}$ in (3.31) are equal to zero. This leads to that $(\Delta q)^{0+\frac{1}{2}}_{j+\frac{1}{2},r} = 0$ and $(\Delta q)^{0+\frac{1}{2}}_{j+\frac{1}{2},l} = 0$. Then the flux approximations in (3.28) and (3.38) show that

$$F_U^0 = \Delta t \left[-f(\tilde{q}_{j+\frac{1}{2},r}) + f(\tilde{q}_{j+\frac{1}{2},r}) + f(\tilde{q}_{j+\frac{1}{2},l}) - f(\tilde{q}_{j+\frac{1}{2},l}) \right] = 0, \tag{3.46}$$

$$F_M^0 = \Delta t \left[-f(\tilde{q}_{j+\frac{1}{2},l}) + f(\tilde{q}_{j+\frac{1}{2},l}) + f(\tilde{q}_{j-\frac{1}{2},r}) - f(\tilde{q}_{j-\frac{1}{2},r}) \right] = 0, \tag{3.47}$$

and the estimated value of the source term S_U^0 and S_M^0 also vanish. Hence, the intermediate averages in (3.33) and (3.40) reduce to

$$\Delta w^{0+1}_{j+\frac{1}{2}} = 0. \tag{3.48}$$

$$\Delta w^{0+1}_{j} = 0. \tag{3.49}$$

Then the deviation $(\Delta q)_j(t=1)$ is equivalent to zero. Since $(\Delta q)_j^1 = 0$, we repeat the above derivations and obtain $(\Delta q)_j^2 = 0$. Once again, $(\Delta q)_j^3 = 0$ is then proved. Consequently, it means that, for all time steps, the approximation $(\Delta q)_j$ is equal to zero.

Since $(\Delta q)_j(t) = 0, \forall t$, it implies

$$q_j(t) = (\Delta q)_j(t) + \tilde{q}(t) = \tilde{q}(t), \quad \forall t. \tag{3.50}$$

Thus, we conclude that the numerical solution $q_j(t)$ remains stationary and prove that our scheme satisfies the well-balance property.

3.3 TVD Property

In this section, we construct a semi-discrete scheme from the fully-discrete scheme (3.43) without the source term, and then prove the TVD property of the semi-discrete scheme applied to the homogeneous scalar conservation laws.

Inspired by [3], in order to construct a semi-discrete scheme, we firstly compute the value of $\frac{(\Delta q)_j^{n+1} - (\Delta q)_j^n}{\Delta t}$, and then let $\Delta t \to 0$, i.e.,

$$\frac{d}{dt}(\Delta q)_j(t) = \lim_{\Delta t \to 0} \frac{(\Delta q)_j^{n+1} - (\Delta q)_j^n}{\Delta t}. \tag{3.51}$$

Substituting the result in (3.43) for $(\Delta q)_j^{n+1}$, and setting $O(\lambda)$ to denote the collection of all terms being proportional to λ yields

$$
\begin{aligned}
&\frac{(\Delta q)_j^{n+1} - (\Delta q)_j^n}{\Delta t} = \\
&= \frac{a^n_{j-\frac{1}{2}}}{\Delta x}\overline{w}^{n+1}_{j-\frac{1}{2}} + \left(\frac{1}{\Delta t} - \frac{a^n_{j-\frac{1}{2}} + a^n_{j+\frac{1}{2}}}{\Delta x}\right)\overline{w}^{n+1}_j + \frac{a^n_{j+\frac{1}{2}}}{\Delta x}\overline{w}^{n+1}_{j+\frac{1}{2}} - \frac{1}{\Delta t}(\Delta q)_j^n + O(\lambda) \\
&= \frac{a^n_{j-\frac{1}{2}}}{\Delta x}\Bigg[\frac{(\Delta q)^n_{j-1} + (\Delta q)^n_j}{2} + \frac{\Delta x - a^n_{j-\frac{1}{2}}\Delta t}{4}\left(((\Delta q)_x)^n_{j-1} - ((\Delta q)_x)^n_j\right) \\
&\quad - \frac{\Delta t}{2a^n_{j-\frac{1}{2}}\Delta t}\left[f(q^{n+\frac{1}{2}}_{j-\frac{1}{2},r} + \tilde{q}_{j-\frac{1}{2},r}) - f(\tilde{q}_{j-\frac{1}{2},r}) - f(q^{n+\frac{1}{2}}_{j-\frac{1}{2},l} + \tilde{q}_{j-\frac{1}{2},l}) + f((\tilde{q})_{j-\frac{1}{2},l})\right]\Bigg] \\
&\quad + \left(\frac{1}{\Delta t} - \frac{a^n_{j-\frac{1}{2}} + a^n_{j+\frac{1}{2}}}{\Delta x}\right)\times \\
&\qquad \Bigg[(\Delta q)^n_j + \frac{a^n_{j-\frac{1}{2}}\Delta t - a^n_{j+\frac{1}{2}}\Delta t}{2}((\Delta q)_x)^n_j \\
&\qquad - \frac{\Delta t}{\Delta x - \Delta t(a^n_{j-\frac{1}{2}} + a^n_{j+\frac{1}{2}})}\Bigg[f(q^{n+\frac{1}{2}}_{j+\frac{1}{2},l} + \tilde{q}_{j+\frac{1}{2},l}) - f(\tilde{q}_{j+\frac{1}{2},l}) \\
&\qquad\qquad - f(q^{n+\frac{1}{2}}_{j-\frac{1}{2},r} + \tilde{q}_{j-\frac{1}{2},r}) + f(\tilde{q}_{j-\frac{1}{2},r})\Bigg]\Bigg] \\
&\quad + \frac{a^n_{j+\frac{1}{2}}}{\Delta x}\Bigg[\frac{(\Delta q)^n_j + (\Delta q)^n_{j+1}}{2} + \frac{\Delta x - a^n_{j+\frac{1}{2}}\Delta t}{4}\left(((\Delta q)_x)^n_j - ((\Delta q)_x)^n_{j+1}\right) \\
&\quad - \frac{\Delta t}{2a^n_{j+\frac{1}{2}}\Delta t}\left[f(q^{n+\frac{1}{2}}_{j+\frac{1}{2},r} + \tilde{q}_{j+\frac{1}{2},r}) - f(\tilde{q}_{j+\frac{1}{2},r}) - f(q^{n+\frac{1}{2}}_{j+\frac{1}{2},l} + \tilde{q}_{j+\frac{1}{2},l}) + f(\tilde{q}_{j+\frac{1}{2},l})\right]\Bigg] \\
&\quad - \frac{1}{\Delta t}(\Delta q)^n_j + O(\lambda) \\
&= \frac{a^n_{j-\frac{1}{2}}}{2\Delta x}((\Delta q)^n_{j-1} + (\Delta q)^n_j) + \frac{a^n_{j-\frac{1}{2}}}{4}\left(((\Delta q)_x)^n_{j-1} - ((\Delta q)_x)^n_j\right) \\
&\qquad - \frac{(a^n_{j-\frac{1}{2}})^2}{4}\lambda\left(((\Delta q)_x)^n_{j-1} - ((\Delta q)_x)^n_j\right) \\
&\quad - \frac{1}{2\Delta x}\left[f(q^{n+\frac{1}{2}}_{j-\frac{1}{2},r} + \tilde{q}_{j-\frac{1}{2},r}) - f(\tilde{q}_{j-\frac{1}{2},r}) - f(q^{n+\frac{1}{2}}_{j-\frac{1}{2},l} + \tilde{q}_{j-\frac{1}{2},l}) + f(\tilde{q}_{j-\frac{1}{2},l})\right] \\
&\quad + \frac{1}{\Delta t}(\Delta q)^n_j - \frac{a^n_{j-\frac{1}{2}} + a^n_{j+\frac{1}{2}}}{\Delta x}(\Delta q)^n_j + \frac{a^n_{j-\frac{1}{2}} - a^n_{j+\frac{1}{2}}}{2}((\Delta q)_x)^n_j \\
&\qquad - \frac{(a^n_{j-\frac{1}{2}})^2 - (a^n_{j+\frac{1}{2}})^2}{2}\lambda(\Delta q_x)^n_j \\
&\quad - \frac{1}{\Delta x}\left[f(q^{n+\frac{1}{2}}_{j+\frac{1}{2},l} + \tilde{q}_{j+\frac{1}{2},l}) - f(\tilde{q}_{j+\frac{1}{2},l}) - f(q^{n+\frac{1}{2}}_{j-\frac{1}{2},r} + \tilde{q}_{j-\frac{1}{2},r}) + f(\tilde{q}_{j-\frac{1}{2},r})\right] \\
&\quad + \frac{a^n_{i+\frac{1}{2}}}{2\Delta x}((\Delta q)^n_j + (\Delta q)^n_{j+1}) + \frac{a^n_{j+\frac{1}{2}}}{4}\left(((\Delta q)_x)^n_j - ((\Delta q)_x)^n_{j+1}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\qquad\qquad\qquad\qquad\qquad\qquad -\frac{(a^n_{j+\frac{1}{2}})^2}{4}\lambda\left(((\Delta q)_x)^n_j-((\Delta q)_x)^n_{j+1}\right)\\
&-\frac{1}{2\Delta x}\left[f(q^{n+\frac{1}{2}}_{j+\frac{1}{2},r}+\tilde{q}_{j+\frac{1}{2},r})-f(\tilde{q}_{j+\frac{1}{2},r})-f(q^{n+\frac{1}{2}}_{j+\frac{1}{2},l}+\tilde{q}_{j+\frac{1}{2},l})+f(\tilde{q}_{j+\frac{1}{2},l})\right]\\
&-\frac{1}{\Delta t}(\Delta q)^n_j+O(\lambda)\\
=&-\frac{1}{2\Delta x}\Big[f(q^{n+\frac{1}{2}}_{j+\frac{1}{2},r}+\tilde{q}_{j+\frac{1}{2},r})-f(\tilde{q}_{j+\frac{1}{2},r})+f(q^{n+\frac{1}{2}}_{j+\frac{1}{2},l}+\tilde{q}_{j+\frac{1}{2},l})-f(\tilde{q}_{j+\frac{1}{2},l})\\
&\qquad -f(q^{n+\frac{1}{2}}_{j-\frac{1}{2},r}+\tilde{q}_{j-\frac{1}{2},r})+f(\tilde{q}_{j-\frac{1}{2},r})-f(q^{n+\frac{1}{2}}_{j-\frac{1}{2},l}+\tilde{q}_{j-\frac{1}{2},l})+f(\tilde{q}_{j-\frac{1}{2},l})\Big]\\
&+\frac{a^n_{j+\frac{1}{2}}}{2\Delta x}\left[((\Delta q)^n_{j+1}-\frac{\Delta x}{2}((\Delta q)_x)^n_{j+1})-((\Delta q)^n_j+\frac{\Delta x}{2}((\Delta q)_x)^n_j)\right]\\
&-\frac{a^n_{j-\frac{1}{2}}}{2\Delta x}\left[((\Delta q)^n_j-\frac{\Delta x}{2}((\Delta q)_x)^n_j)-((\Delta q)^n_{j-1}+\frac{\Delta x}{2}((\Delta q)_x)^n_{j-1})\right]+O(\lambda)
\end{aligned}
\tag{3.52}
$$

By the definition of the midpoint values in (3.29), (3.30) and (3.31), we have

$$
\begin{aligned}
&(\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},r}\\
&=(\Delta q)^n_{j+1}-\Delta x((\Delta q)_x)^n_{j+1}(\frac{1}{2}-\lambda a^n_{j+\frac{1}{2}})-\frac{\Delta t}{2}[f((\Delta q)^n_{j+\frac{1}{2},r}+\tilde{q}_{j+\frac{1}{2},r})-f(\tilde{q}_{j+\frac{1}{2},r})]\\
&(\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},l}\\
&=(\Delta q)^n_j+\Delta x((\Delta q)_x)^n_j(\frac{1}{2}-\lambda a^n_{j+\frac{1}{2}})-\frac{\Delta t}{2}[f((\Delta q)^n_{j+\frac{1}{2},l}+\tilde{q}_{j+\frac{1}{2},l})-f(\tilde{q}_{j+\frac{1}{2},l})].
\end{aligned}
\tag{3.53}
$$

Applying the Taylor expansion to $\tilde{q}_{j+\frac{1}{2},r}$ and $\tilde{q}_{j+\frac{1}{2},l}$, we obtain

$$
\tilde{q}_{j+\frac{1}{2},r}:=\tilde{q}_{j+\frac{1}{2}}+a^n_{j+\frac{1}{2}}\Delta t(\tilde{q}_x)_{j+\frac{1}{2}},\quad \text{and}\quad \tilde{q}_{j+\frac{1}{2},l}:=\tilde{q}_{j+\frac{1}{2}}-a^n_{j+\frac{1}{2}}\Delta t(\tilde{q}_x)_{j+\frac{1}{2}}. \tag{3.54}
$$

As $\Delta t\to 0$, the midvalues approach

$$
\begin{aligned}
(\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},r}&\to(\Delta q)_{j+1}(t)-\frac{\Delta x}{2}((\Delta q)_x)^n_{j+1}=:(\Delta q)^+_{j+\frac{1}{2}}(t),\\
(\Delta q)^{n+\frac{1}{2}}_{j+\frac{1}{2},l}&\to(\Delta q)_j(t)+\frac{\Delta x}{2}((\Delta q)_x)^n_j=:(\Delta q)^-_{j+\frac{1}{2}}(t),\\
\tilde{q}_{j+\frac{1}{2},r}&\to\tilde{q}_{j+\frac{1}{2}},\\
\tilde{q}_{j+\frac{1}{2},l}&\to\tilde{q}_{j+\frac{1}{2}}.
\end{aligned}
\tag{3.55}
$$

Hence, letting $\Delta t \to 0$ in (3.52), the semi-discrete scheme takes form as

$$\begin{aligned}\frac{d}{dt}(\Delta q)_j(t) &= \lim_{\Delta t\to 0}\frac{(\Delta q)_j^{n+1}-(\Delta q)_j^n}{\Delta t}\\ &= -\frac{1}{2\Delta x}\Big[\Big(f((\Delta q)^-_{j+\frac12}(t)+\tilde q_{j+\frac12}(t))-f(\tilde q_{j+\frac12}(t))\\ &\qquad + f((\Delta q)^+_{j+\frac12}(t)+\tilde q_{j+\frac12}(t))-f(\tilde q_{j+\frac12}(t))\Big)\\ &\quad -\Big(f((\Delta q)^-_{j-\frac12}(t)+\tilde q_{j-\frac12}(t))-f(\tilde q_{j-\frac12}(t))\\ &\qquad + f((\Delta q)^+_{j-\frac12}(t)+\tilde q_{j-\frac12}(t))-f(\tilde q_{j-\frac12}(t))\Big)\Big]\\ &\quad +\frac{1}{2\Delta x}\Big[a^n_{j+\frac12}((\Delta q)^+_{j+\frac12}(t)-(\Delta q)^-_{j+\frac12}(t))\\ &\qquad - a^n_{j-\frac12}((\Delta q)^+_{j-\frac12}(t)-(\Delta q)^-_{j-\frac12}(t))\Big]\end{aligned} \tag{3.56}$$

Then we reformulate (3.56) as the conservation form

$$\frac{d}{dt}(\Delta q)_j(t) = -\frac{H_{j+\frac12}(t)-H_{j-\frac12}(t)}{\Delta x}, \tag{3.57}$$

with the numerical flux

$$\begin{aligned}&H_{j+\frac12}(t) :=\\ &\quad \frac12\left[F((\Delta q)^-_{j+\frac12})(t)+F((\Delta q)^+_{j+\frac12})(t)\right]-\frac{a_{j+\frac12}(t)}{2}\left((\Delta q)^+_{j+\frac12}(t)-(\Delta q)^-_{j+\frac12}(t)\right),\end{aligned} \tag{3.58}$$

where

$$F((\Delta q)^{\mp}_{j+\frac12}) := f((\Delta q)^{\mp}_{j+\frac12}+\tilde q_{j+\frac12})-f(\tilde q_{j+\frac12}). \tag{3.59}$$

To prove the semi-discrete scheme satisfies the TVD property, we introduce the following lemma 3.3.1 from [6]. The proof of this lemma can be found in the example 2.4 in [6].

Lemma 3.3.1 (consult example 2.4 in[6]) *Consider the class of generalized MUSCL scheme,*

$$\frac{d}{dt}u_j(t) = -\frac{1}{\Delta x}\left[h_{j+\frac12}-h_{j-\frac12}\right], \tag{3.60}$$

with the E-flux $h_{j+\frac12}$*,*

$$h_{j+\frac12} = h^E\left(u_j+\frac{\Delta x}{2}(u_x)_j, u_{j+1}-\frac{\Delta x}{2}(u_x)_{j+1}\right). \tag{3.61}$$

The scheme (3.60) *is TVD if*

$$\left| \frac{\Delta x (u_x)_{j-1}}{\Delta u_j - \Delta u_{j-1}} \right| \leq 2, \quad \left| \frac{\Delta x (u_x)_{j+1}}{\Delta u_{j+1} - \Delta u_j} \right| \leq 2. \tag{3.62}$$

Theorem 3.3.2 (TVD of semi-discrete scheme, consult Thm 4.1 of [3]) *Consider the semi-discrete scheme* (3.57) *and assume that the numerical spatial derivatives* $(\Delta q_x)_j(t)$ *are chosen as* (3.15)*. Then the scheme satisfies the TVD property.*

Proof. The scheme (3.57) can be viewed as a general MUSCL scheme with the Rusanov type E-flux,

$$H_{j+\frac{1}{2}} = H^{Rus}\left((\Delta q)^{+}_{j+\frac{1}{2}}(t), (\Delta q)^{-}_{j+\frac{1}{2}}(t) \right). \tag{3.63}$$

With the choice of derivatives (3.15) and the definition of minmod function (3.16), we have

$$\begin{aligned} |((\Delta q)_x)_{j-1}| &\leq \left| \theta \frac{(\Delta q)_j - (\Delta q)_{j-1}}{\Delta x} \right|, \\ |((\Delta q)_x)_{j+1}| &\leq \left| \theta \frac{(\Delta q)_{j+1} - (\Delta q)_j}{\Delta x} \right|. \end{aligned} \tag{3.64}$$

The inequalities (3.64) imply that

$$\begin{aligned} \left| \frac{\Delta x ((\Delta q)_x)_{j-1}}{(\Delta q)_j - (\Delta q)_{j-1}} \right| &\leq |\theta| \leq 2, \\ \left| \frac{\Delta x ((\Delta q)_x)_{j+1}}{(\Delta q)_{j+1} - (\Delta q)_j} \right| &\leq |\theta| \leq 2, \end{aligned} \tag{3.65}$$

which satisfy the condition (3.62) in lemma 3.3.1. Hence, the TVD property holds. ∎

Notice that in the above theorem 3.3.2, we only prove the TVD property for the deviation Δq. To conclude the TVD property holds for the solution $q(= \Delta q + \tilde{q})$ need extraordinary steps.

Chapter 4

2D Scheme

4.1 Description

In the previous chapter we showed a new scheme by using the combination of the Deviation method 2.2.2 and the KT scheme 2.2.1. In this section we use the same idea to construct a 2-dimensional scheme. Similarly, we firstly refer to [4] to derive a modified 2-dimensional well-balanced law with the Deviation method, and then follow the 2-dimensional KT-type scheme in [7] to construct a new 2-dimensional fully-discrete scheme. The KT-type scheme in [7] consider the more specific wave speeds to define the regions over Riemann fans.

Consider the 2-dimensional balance laws

$$\begin{cases} q_t + f(q)_x + g(q)_y = S(q, x, y), \quad (x, y) \in \Omega \subset \mathbb{R}^2, t > 0 \\ q(x, y, 0) = q_0(x, y), \end{cases} \tag{4.1}$$

where $f(q)$ and $g(q)$ are the fluxes in x- and y- directions ans S is the source term.

Assume $\tilde{q}$ is a given stationary solution of (4.1) Then it satisfies

$$f(\tilde{q})_x + g(\tilde{q})_y = S(\tilde{q}, x, y). \tag{4.2}$$

Define the deviation $\Delta q = q - \tilde{q}$. Applying $q = \Delta q + \tilde{q}$ to (4.1), we obtain

$$(\Delta q + \tilde{q})_t + f(\Delta q + \tilde{q})_x + g(\Delta q + \tilde{q})_y = S(\Delta q + \tilde{q}). \tag{4.3}$$

Since $\tilde{q}$ is a stationary solution, (4.3) reduces to

$$\Delta q_t + f(\Delta q + \tilde{q})_x + g(\Delta q + \tilde{q})_y = S(\Delta q + \tilde{q}). \tag{4.4}$$

Then we subtract (4.2) from (4.4), and assume that the source term $S(\Delta q+\tilde{q})$ in (4.1) is a linear functional in terms of the conserved variables, we obtain

$$\begin{aligned}(\Delta q)_t+[f(\Delta q-\tilde{q})-f(\tilde{q})]_x+[g(\Delta q+\tilde{q})-g(\tilde{q})]_y&=S(\Delta q+\tilde{q},x,y)-S(\tilde{q},x,y)\\&=S(\Delta q,x,y).\end{aligned} \tag{4.5}$$

Now we can apply the idea of KT scheme. Referring to the introduction in [7], we also separate our derivation into three steps: Reconstruction, Evolution, and Projection.

4.1.1 Reconstruction

Consider the control cell $C_{j,k}=[x_{j-\frac{1}{2}},x_{j+\frac{1}{2}}]\times[y_{k-\frac{1}{2}},y_{k+\frac{1}{2}}]$ for all j,k. We define a piecewise-linear reconstruction Q,

$$Q_{j,k}(x,y,t^n)=(\Delta q)^n_{j,k}+(x-x_j)((\Delta q)_x)^n_{j,k}+(y-y_k)((\Delta q)_y)^n_{j,k},\quad \forall(x,y)\in C_{i,j} \tag{4.6}$$

where $(\Delta q)_x$ and $(\Delta q)_y$ are the x- and y derivatives of Δq .

We adopt the $(MC-\theta)$ limiter to evaluate the values of the slopes $(\Delta q)_x$ and $(\Delta q)_y$. The $(MC-\theta)$ limiter is defined as

$$\begin{aligned}((\Delta q)_x)^n_{j,k}&=\text{minmod}\left(\theta\Delta^+_x(\Delta q)^n_{j,k},\ \Delta^0_x(\Delta q)^n_{j,k},\ \theta\Delta^-_x(\Delta q)^n_{j,k}\right),\\((\Delta q)_y)^n_{j,k}&=\text{minmod}\left(\theta\Delta^+_y(\Delta q)^n_{j,k},\ \Delta^0_y(\Delta q)^n_{j,k},\ \theta\Delta^-_y(\Delta q)^n_{j,k}\right),\quad 1\le\theta\le 2\end{aligned} \tag{4.7}$$

where $\Delta^{\pm}_x$, Δ^0_x, $\Delta^{\pm}_y$, and Δ^0_y are the standard divided difference operators:

$$\begin{aligned}\Delta^+_x(\cdot)_{j,k}&:=\frac{(\cdot)_{j+1,k}-(\cdot)_{j,k}}{\Delta x},\Delta^0_x(\cdot)_{j,k}:=\frac{(\cdot)_{j+1,k}-(\cdot)_{j-1,k}}{2\Delta x},\Delta^-_x(\cdot)_{j,k}:=\frac{(\cdot)_{j,k}-(\cdot)_{j-1,k}}{\Delta x},\\\Delta^+_y(\cdot)_{j,k}&:=\frac{(\cdot)_{j,k+1}-(\cdot)_{j,k}}{\Delta y},\Delta^0_y(\cdot)_{j,k}:=\frac{(\cdot)_{j,k+1}-(\cdot)_{j,k-1}}{2\Delta y},\Delta^-_y(\cdot)_{j,k}:=\frac{(\cdot)_{j,k}-(\cdot)_{j,k-1}}{\Delta y}.\end{aligned} \tag{4.8}$$

4.1.2 Evolution

In this step, the wave speed, which is the most important material of the KT scheme, is used to define the regions of the subdomains when we consider the new cell-averages. The local wave speed denoted by $a^{\pm}_{j+\frac{1}{2},k}$ and $b^{\pm}_{j,k+\frac{1}{2}}$ in x- and y-direction, respectively, are determined by the eigenvalues of flux jacobian:

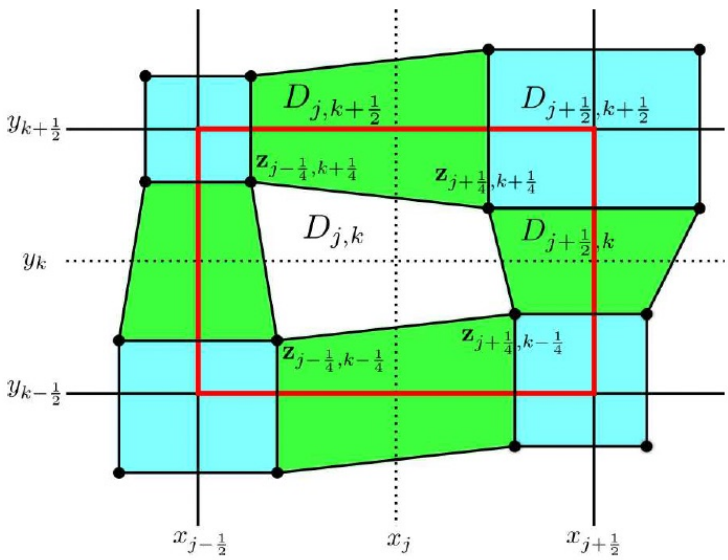
$$
\begin{aligned}
a^{+}_{j+\frac{1}{2},k} &:= \max\left\{\lambda_N\left(\frac{\partial}{\partial q}f(q^{+}_{j+\frac{1}{2},k})\right),\ \lambda_N\left(\frac{\partial}{\partial q}f(q^{-}_{j+\frac{1}{2},k})\right),\ \epsilon\right\},\\
a^{-}_{j+\frac{1}{2},k} &:= \min\left\{\lambda_1\left(\frac{\partial}{\partial q}f(q^{+}_{j+\frac{1}{2},k})\right),\ \lambda_1\left(\frac{\partial}{\partial q}f(q^{-}_{j+\frac{1}{2},k})\right),\ -\epsilon\right\},\\
b^{+}_{j,k+\frac{1}{2}} &:= \max\left\{\lambda_N\left(\frac{\partial}{\partial q}g(q^{+}_{j,k+\frac{1}{2}})\right),\ \lambda_N\left(\frac{\partial}{\partial q}g(q^{-}_{j,k+\frac{1}{2}})\right),\ \epsilon\right\},\\
b^{-}_{j,k+\frac{1}{2}} &:= \min\left\{\lambda_1\left(\frac{\partial}{\partial q}g(q^{+}_{j,k+\frac{1}{2}})\right),\ \lambda_1\left(\frac{\partial}{\partial q}g(q^{-}_{j,k+\frac{1}{2}})\right),\ -\epsilon\right\},
\end{aligned}
\tag{4.9}
$$

where $\lambda_1 < \lambda_2 < \cdots < \lambda_N$ are the N eigenvalues of the corresponding Jacobians, and ϵ is a small positive number and

$$
\begin{aligned}
q^{-}_{j+\frac{1}{2},k} &:= q^n_{j,k} + \frac{\Delta x}{2}(q_x)^n_{j,k}, \quad & q^{+}_{j+\frac{1}{2},k} &:= q^n_{j+1,k} - \frac{\Delta x}{2}(q_x)^n_{j+1,k},\\
q^{-}_{j,k+\frac{1}{2}} &:= q^n_{j,k} + \frac{\Delta y}{2}(q_y)^n_{j,k}, \quad & q^{+}_{j,k+\frac{1}{2}} &:= q^n_{j,k+1} - \frac{\Delta y}{2}(q_y)^n_{j,k+1}.
\end{aligned}
\tag{4.10}
$$

Remark. In the definition of wave speed in (4.9), the small positive number ϵ is used to avoid the corresponding quadrilateral subdomain degenerate to triangle when the wave speed a or b is equal to zero.

In a similar way to the 1-dimensional scheme, with the wave speed we split the control domain $C_{j,k}$ into the unsmooth side subdomain $D_{j,k+\frac{1}{2}}$, $D_{j+\frac{1}{2},k}$ and the corner subdomain $D_{j+\frac{1}{2},k+\frac{1}{2}}$, and the smooth central subdomain $D_{j,k}$. This set-up is shown in figure 4.1.

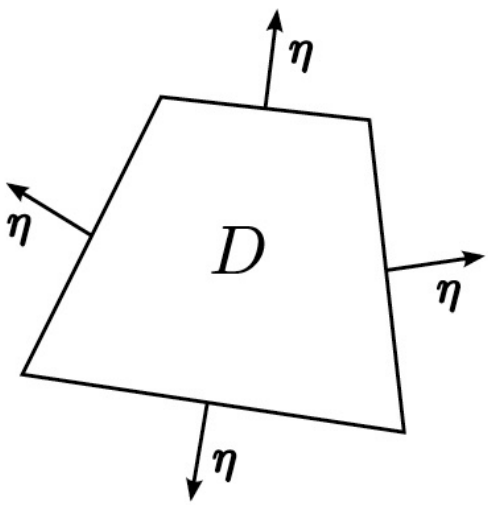


set-up.jpeg

Figure 4.1: 2D set-up (Reference: Figure 1 in [7])

The vertices $z_{j\pm\frac{1}{4},k\pm\frac{1}{4}}$ of these subdomains are computed by

$$\begin{aligned}
z_{j+\frac{1}{4},k+\frac{1}{4}} &:= (x_{j+\frac{1}{2}} + \Delta t^n \min(a^-_{j+\frac{1}{2},k}, a^-_{j+\frac{1}{2},k+1}), y_{k+\frac{1}{2}} + \Delta t^n \min(b^-_{j,k+\frac{1}{2}}, b^-_{j+1,k+\frac{1}{2}})),\\
z_{j-\frac{1}{4},k+\frac{1}{4}} &:= (x_{j-\frac{1}{2}} + \Delta t^n \max(a^+_{j-\frac{1}{2},k}, a^+_{j-\frac{1}{2},k+1}), y_{k+\frac{1}{2}} + \Delta t^n \min(b^-_{j,k+\frac{1}{2}}, b^-_{j-1,k+\frac{1}{2}})),\\
z_{j-\frac{1}{4},k-\frac{1}{4}} &:= (x_{j-\frac{1}{2}} + \Delta t^n \max(a^+_{j-\frac{1}{2},k}, a^+_{j-\frac{1}{2},k-1}), y_{k-\frac{1}{2}} + \Delta t^n \max(b^+_{j,k-\frac{1}{2}}, b^+_{j-1,k-\frac{1}{2}})),\\
z_{j+\frac{1}{4},k-\frac{1}{4}} &:= (x_{j+\frac{1}{2}} + \Delta t^n \min(a^-_{j+\frac{1}{2},k}, a^-_{j+\frac{1}{2},k-1}), y_{k-\frac{1}{2}} + \Delta t^n \max(b^+_{j,k-\frac{1}{2}}, b^+_{j+1,k-\frac{1}{2}})).
\end{aligned} \tag{4.11}$$



D.jpeg

Figure 4.2: General quadrilateral D with normal η. (Reference: Figure 2 in [7])

Now we evolve the system from t^n to t^{n+1} by integrating the modified balance law (4.5) over the subdomains $D_{j,k}-$, $D_{j+\frac{1}{2},k}-$, $D_{j,k+\frac{1}{2}}-$, and $D_{j+\frac{1}{2},k+\frac{1}{2}} \times [t^n, t^{n+1}]$. Considering a general quadrilateral D as shown in figure 4.2, , the integration over $D \times [t^n, t^{n+1}]$ is as follows,

$$\begin{aligned}
\int_{t^n}^{t^n+1} \iint_D (\Delta q(x,y,t))_t + [f(\Delta q + \tilde{q}) - f(\tilde{q})]_x + [g(\Delta q + \tilde{q}) - g(\tilde{q})]_y \, dxdydt \\
= \int_{t^n}^{t^n+1} \iint_D S(\Delta q, x, y) \, dxdydt.
\end{aligned} \tag{4.12}$$

It implies

$$\begin{aligned}\iint_D \Delta q(x,y,t^{n+1})\,dxdy - \iint_D \Delta q(x,y,t^n)\,dxdy \\ + \int_{t^n}^{t^n+1} \iint_D [f(\Delta q+\tilde{q}) - f(\tilde{q})]_x + [g(\Delta q+\tilde{q}) - g(\tilde{q})]_y\,dxdydt \\ = \int_{t^n}^{t^{n+1}} \iint_D S(\Delta q, x, y)\,dxdydt.\end{aligned} \tag{4.13}$$

Moving the second and third integration on the left-hand-side of (4.13) to the right-hand-side, and dividing the area $|D|$ into (4.13), we obtain

$$\begin{aligned}\frac{1}{|D|}\iint_D \Delta q(x,y,t^{n+1})\,dxdy \\ = \frac{1}{|D|}\iint_D \Delta q(x,y,t^n)\,dxdy \\ -\frac{1}{|D|}\int_{t^n}^{t^n+1} \iint_D [f(\Delta q+\tilde{q}) - f(\tilde{q})]_x + [g(\Delta q+\tilde{q}) - g(\tilde{q})]_y\,dxdydt \\ +\frac{1}{|D|}\int_{t^n}^{t^{n+1}} \iint_D S(\Delta q, x, y)\,dxdydt.\end{aligned} \tag{4.14}$$

The left-hand-side of (4.14) is the new cell-average over D at t^{n+1}, which is denoted by $\overline{w}_D^{n+1}$ in the following. Due to the conservation property of the reconstruction Q, we have the relation

$$\overline{(\Delta q)_D^n} := \frac{1}{|D|}\iint_D \Delta q(x,y,t^n)\,dxdy = \frac{1}{|D|}\iint_D Q(x,y,t^n)\,dxdy, \tag{4.15}$$

where $\overline{(\Delta q)_D^n}$ defines the cell-average over D at t^n. Applying the divergence theorem to the flux integration, we rewrite the (4.14) in the form of

$$\begin{aligned}\overline{w}_D^{n+1} =& \overline{(\Delta q)_D^n} \\ &- \frac{1}{|D|}\int_{t^n}^{t^{n+1}} \oint_{\partial D} \eta^x[f(\Delta q+\tilde{q}) - f(\tilde{q})] + \eta^y[g(\Delta q+\tilde{q}) - g(\tilde{q})]\,dxdydt \\ &+ \frac{1}{|D|}\int_{t^n}^{t^{n+1}} \iint_D S(\Delta q, x, y)\,dxdydt\end{aligned} \tag{4.16}$$

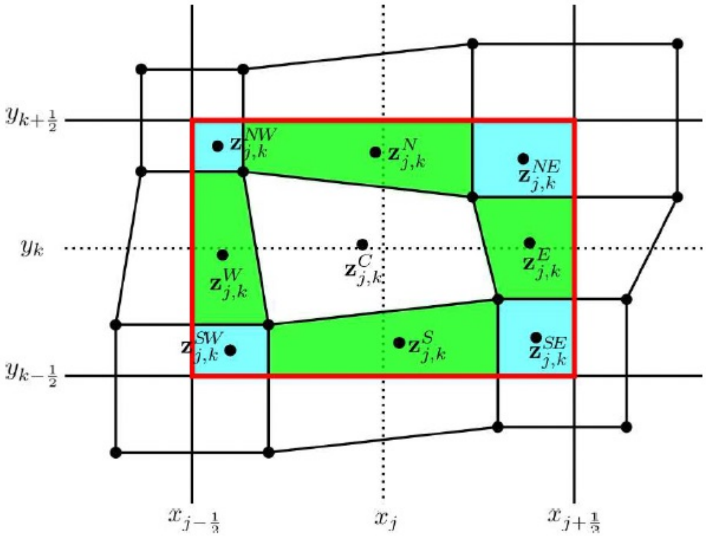
Next, we discuss the approximations of $\overline{(\Delta q)_D^n}$, the flux integral, and the source integral on the right-hand-side of (4.16) separately.

For the following discussion, we use the notations $C_{j,k}^I$, $I \in \{NW, N, NE, W, C, E,$

$SW, S, SE\}$, to denote the nine subdomains inside the red frame of figure 4.1. The definition of $C^I_{j,k}$ are

$$\begin{aligned}
&C^{NW}_{j,k} := C_{j,k} \cap D_{j-\frac{1}{2},k+\frac{1}{2}}, && C^{N}_{j,k} := C_{j,k} \cap D_{j,k+\frac{1}{2}}, && C^{NE}_{j,k} := C_{j,k} \cap D_{j+\frac{1}{2},k+\frac{1}{2}},\\
&C^{W}_{j,k} := C_{j,k} \cap D_{j-\frac{1}{2},k}, && C^{C}_{j,k} := D_{j,k}, && C^{E}_{j,k} := C_{j,k} \cap D_{j+\frac{1}{2},k},\\
&C^{SW}_{j,k} := C_{j,k} \cap D_{j-\frac{1}{2},k-\frac{1}{2}}, && C^{S}_{j,k} := C_{j,k} \cap D_{j,k-\frac{1}{2}}, && C^{SE}_{j,k} := C_{j,k} \cap D_{j+\frac{1}{2},k-\frac{1}{2}},
\end{aligned} \tag{4.17}$$

and their center of mass are denoted by $z^I_{j,k} := (x^I_{j,k}, y^I_{j,k})$. Figure 4.3 outlines these nine subdomains with their center of mass. The detailed computation of the coordinates of the center of mass can be found in Appendix A of [7].



with D.jpeg

Figure 4.3: Nine subdomains of the original control cell $C_{j,k}$ and their center of mass. (Reference: figure 3 in [7])

We first discuss the estimate of the cell-average $\overline{(\Delta q)^n_D}$. Due to the equality (4.15), we obtain,

$$\begin{aligned}
\overline{(\Delta q)^n_{D_{j,k}}} &= \frac{1}{|D_{j,k}|}\iint_{D_{j,k}} Q(x,y,t^n)dxdy\\
&= Q(z^C_{j,k}, t^n)\\
&= \Delta q^n_{j,k} + (z^{C,x}_{j,k} - x_j)(\Delta q_x)^n_{j,k} + (z^{C,y}_{j,k} - y_k)(\Delta q_y)^n_{j,k},\\
\overline{(\Delta q)^n_{D_{j+\frac{1}{2},k}}} &= \frac{1}{|D_{j+\frac{1}{2},k}|}\iint_{D_{j+\frac{1}{2},k}} Q(x,y,t^n)dxdy
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{|D_{j+\frac{1}{2},k}|}\left[\iint_{C^E_{j,k}} Q(x,y,t^n)dxdy + \iint_{C^W_{j+1,k}} Q(x,y,t^n)dxdy\right] \\
&= \frac{1}{|D_{j+\frac{1}{2},k}|}\left[|C^E_{j,k}|Q(z^E_{j,k},t^n) + |C^W_{j+1,k}|Q(z^W_{j+1,k},t^n)\right] \\
&= \frac{1}{|D_{j+\frac{1}{2},k}|}\Big[|C^E_{j,k}|\left(\Delta q^n_{j,k} + (z^{E,x}_{j,k} - x_j)(\Delta q_x)^n_{j,k} + (z^{E,y}_{j,k} - y_k)(\Delta q_y)^n_{j,k}\right) \\
&\qquad + |C^W_{j+1,k}|\Big(\Delta q^n_{j+1,k} + (z^{W,x}_{j+1,k} - x_{j+1})(\Delta q_x)^n_{j+1,k} \\
&\qquad\qquad + (z^{W,y}_{j+1,k} - y_k)(\Delta q_y)^n_{j+1,k}\Big)\Big]
\end{aligned}
$$

$$
\begin{aligned}
\overline{(\Delta q)^n_{D_{j,k+\frac{1}{2}}}} &= \frac{1}{|D_{j,k+\frac{1}{2}}|}\iint_{D_{j,k+\frac{1}{2}}} Q(x,y,t^n)dxdy \\
&= \frac{1}{|D_{j,k+\frac{1}{2}}|}\left[\iint_{C^N_{j,k}} Q(x,y,t^n)dxdy + \iint_{C^S_{j,k+1}} Q(x,y,t^n)dxdy\right] \\
&= \frac{1}{|D_{j,k+\frac{1}{2}}|}\left[|C^N_{j,k}|Q(z^N_{j,k},t^n) + |C^S_{j,k+1}|Q(z^S_{j,k+1},t^n)\right] \\
&= \frac{1}{|D_{j,k+\frac{1}{2}}|}\Big[|C^N_{j,k}|\left(\Delta q^n_{j,k} + (z^{N,x}_{j,k} - x_j)(\Delta q_x)^n_{j,k} + (z^{N,y}_{j,k} - y_k)(\Delta q_y)^n_{j,k}\right) \\
&\qquad + |C^S_{j,k+1}|\Big(\Delta q^n_{j,k+1} + (z^{S,x}_{j,k+1} - x_j)(\Delta q_x)^n_{j,k+1} \\
&\qquad\qquad + (z^{S,y}_{j,k+1} - y_{k+1})(\Delta q_y)^n_{j,k+1}\Big)\Big]
\end{aligned}
$$

$$
\begin{aligned}
\overline{(\Delta q)^n_{D_{j+\frac{1}{2},k+\frac{1}{2}}}} &= \frac{1}{|D_{j+\frac{1}{2},k+\frac{1}{2}}|}\iint_{D_{j+\frac{1}{2},k+\frac{1}{2}}} Q(x,y,t^n)dxdy \\
&= \frac{1}{|D_{j+\frac{1}{2},k+\frac{1}{2}}|}\Big[\iint_{C^{NE}_{j,k}} Q(x,y,t^n)dxdy + \iint_{C^{NW}_{j+1,k}} Q(x,y,t^n)dxdy \\
&\qquad + \iint_{C^{SE}_{j,k+1}} Q(x,y,t^n)dxdy + \iint^{SW}_{j+1,k+1} Q(x,y,t^n)dxdy\Big] \\
&= \frac{1}{|D_{j+\frac{1}{2},k+\frac{1}{2}}|}\Big[|C^{NE}_{j,k}|Q(z^{NE}_{j,k},t^n) + |C^{NW}_{j+1,k}|Q(z^{NW}_{j+1,k},t^n) \\
&\qquad + |C^{SE}_{j,k+1}|Q(z^{SE}_{j,k+1},t^n) + |C^{SW}_{j+1,k+1}|Q(z^{SW}_{j+1,k+1},t^n)\Big] \\
&= \frac{1}{|D_{j+\frac{1}{2},k+\frac{1}{2}}|}\times \\
&\quad \Big[|C^{NE}_{j,k}|\left(\Delta q^n_{j,k} + (z^{NE,x}_{j,k} - x_j)(\Delta q_x)^n_{j,k} + (z^{NE,y}_{j,k} - y_k)(\Delta q_y)^n_{j,k}\right) \\
&\quad + |C^{NW}_{j+1,k}|\Big(\Delta q^n_{j+1,k} + (z^{NW,x}_{j+1,k} - x_{j+1})(\Delta q_x)^n_{j+1,k} \\
&\qquad\qquad + (z^{NW,y}_{j+1,k} - y_k)(\Delta q_y)^n_{j+1,k}\Big) \\
&\quad + |C^{SE}_{j,k+1}|\Big(\Delta q^n_{j,k+1} + (z^{SE,x}_{j,k+1} - x_j)(\Delta q_x)^n_{j,k+1}
\end{aligned}
$$

$$
\begin{aligned}
&\qquad\qquad\qquad + (z^{SE,y}_{j,k+1} - y_{k+1})(\Delta q_y)^n_{j,k+1}\Big) \\
&+ |C^{SW}_{j+1,k+1}|\Big(\Delta q^n_{j+1,k+1} + (z^{SW,x}_{j+1,k+1} - x_{j+1})(\Delta q_x)^n_{j+1,k+1} \\
&\qquad\qquad\qquad + (z^{SW,y}_{j+1,k+1} - y_{k+1})(\Delta q_y)^n_{j+1,k+1}\Big)\Big]
\end{aligned}
\tag{4.18}
$$

Next, we evaluate the flux integral on (4.16). We need to integrate the fluxes along the four edges of D, and in order to present formula for these, we separate them into two cases. As shown in figure 4.4, in case 1, we consider the flux integral over the right and left edges, and on the other hand, we consider the flux integral over top and bottom edges in case 2.

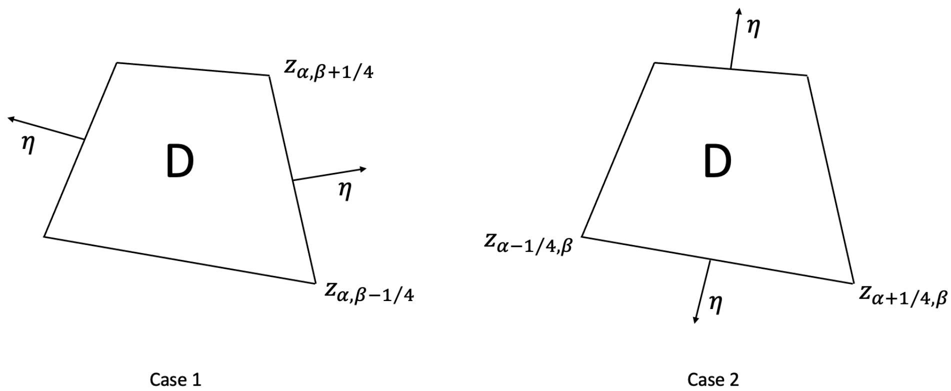


Figure 4.4: A general quadrilateral with normal η split into two cases.

Case 1.

For the nodes $z_{\alpha,\beta-\frac{1}{4}}$ and $z_{\alpha,\beta+\frac{1}{4}}$, where $\alpha = j + \frac{1}{4}$ or $j - \frac{1}{4}$ and $\beta = k$ or $k + \frac{1}{2}$, we define the flux across these two nodes by

$$
\begin{aligned}
H_{\alpha,\beta} := \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} \int_{z_{\alpha,\beta-\frac{1}{4}}}^{z_{\alpha,\beta+\frac{1}{4}}} \Big[&\eta^x_{\alpha,\beta}\left(f(\Delta q + \tilde{q}) - f(\tilde{q})\right) \\
&+ \eta^y_{\alpha,\beta}\left(g(\Delta q + \tilde{q}) - g(\tilde{q})\right)\Big]\, ds dt
\end{aligned}
\tag{4.19}
$$

Applying the midpoint rule in time and the trapezoidal rule in space to (4.19), we obtain

$$
\begin{aligned}
H_{\alpha,\beta} &= \int_{z_{\alpha,\beta-\frac{1}{4}}}^{z_{\alpha,\beta+\frac{1}{4}}} \left[\eta^x_{\alpha,\beta}\left(f(\Delta q^{n+\frac{1}{2}}+\tilde{q})-f(\tilde{q})\right)+\eta^y_{\alpha,\beta}\left(g(\Delta q^{n+\frac{1}{2}}+\tilde{q})-g(\tilde{q})\right)\right]ds \\
&= \frac{|z_{\alpha,\beta+\frac{1}{4}}-z_{\alpha,\beta-\frac{1}{4}}|}{2}\Big\{\eta^x_{\alpha,\beta}\Big[f(\Delta q^{n+\frac{1}{2}}_{\alpha,\beta-\frac{1}{4}}+\tilde{q}_{\alpha,\beta-\frac{1}{4}})-f(\tilde{q}_{\alpha,\beta-\frac{1}{4}}) \\
&\qquad +f(\Delta q^{n+\frac{1}{2}}_{\alpha,\beta+\frac{1}{4}}+\tilde{q}_{\alpha,\beta+\frac{1}{4}})-f(\tilde{q}_{\alpha,\beta+\frac{1}{4}})\Big] \\
&\qquad +\eta^y_{\alpha,\beta}\Big[g(\Delta q^{n+\frac{1}{2}}_{\alpha,\beta-\frac{1}{4}}+\tilde{q}_{\alpha,\beta-\frac{1}{4}})-g(\tilde{q}_{\alpha,\beta-\frac{1}{4}}) \\
&\qquad +g(\Delta q^{n+\frac{1}{2}}_{\alpha,\beta+\frac{1}{4}}+\tilde{q}_{\alpha,\beta+\frac{1}{4}})-g(\tilde{q}_{\alpha,\beta+\frac{1}{4}})\Big]\Big\}
\end{aligned}
\tag{4.20}
$$

where the unit normal vectors $\eta_{\alpha,\beta}$ are

$$
\eta_{\alpha,k} = \frac{(y_{\alpha,k+\frac{1}{4}}-y_{\alpha,k-\frac{1}{4}}, x_{\alpha,k+\frac{1}{4}}-x_{\alpha,k-\frac{1}{4}})}{|z_{\alpha,\beta+\frac{1}{4}}-z_{\alpha,\beta-\frac{1}{4}}|} \quad \text{and} \quad \eta_{\alpha,k+\frac{1}{2}} = (1,0). \tag{4.21}
$$

Case 2.

For the nodes $z_{\alpha-\frac{1}{4},\beta}$ and $z_{\alpha+\frac{1}{4},\beta}$, where $\alpha = j$ or $j+\frac{1}{2}$ and $\beta = k+\frac{1}{4}$ or $k-\frac{1}{4}$, we define the flux across these two nodes by

$$
\begin{aligned}
H_{\alpha,\beta} := \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}\int_{z_{\alpha-\frac{1}{4},\beta}}^{z_{\alpha+\frac{1}{4},\beta}} \Big[&\eta^x_{\alpha,\beta}\left(f(\Delta q+\tilde{q})-f(\tilde{q})\right) \\
&+\eta^y_{\alpha,\beta}\left(g(\Delta q+\tilde{q})-g(\tilde{q})\right)\Big]\,dsdt
\end{aligned}
\tag{4.22}
$$

Likewise, applying the midpoint rule in time and the trapezoidal rule in space to (4.22) results in

$$
\begin{aligned}
H_{\alpha,\beta} &= \int_{z_{\alpha-\frac{1}{4},\beta}}^{z_{\alpha+\frac{1}{4},\beta}} \left[\eta^x_{\alpha,\beta}\left(f(\Delta q^{n+\frac{1}{2}}+\tilde{q})-f(\tilde{q})\right)+\eta^y_{\alpha,\beta}\left(g(\Delta q^{n+\frac{1}{2}}+\tilde{q})-g(\tilde{q})\right)\right]ds \\
&= \frac{|z_{\alpha+\frac{1}{4},\beta}-z_{\alpha-\frac{1}{4},\beta}|}{2}\Big\{\eta^x_{\alpha,\beta}\Big[f(\Delta q^{n+\frac{1}{2}}_{\alpha-\frac{1}{4},\beta}+\tilde{q}_{\alpha-\frac{1}{4},\beta})-f(\tilde{q}_{\alpha-\frac{1}{4},\beta}) \\
&\qquad +f(\Delta q^{n+\frac{1}{2}}_{\alpha+\frac{1}{4},\beta}+\tilde{q}_{\alpha+\frac{1}{4},\beta})-f(\tilde{q}_{\alpha+\frac{1}{4},\beta})\Big] \\
&\qquad +\eta^y_{\alpha,\beta}\Big[g(\Delta q^{n+\frac{1}{2}}_{\alpha-\frac{1}{4},\beta}+\tilde{q}_{\alpha-\frac{1}{4},\beta})-g(\tilde{q}_{\alpha-\frac{1}{4},\beta}) \\
&\qquad +g(\Delta q^{n+\frac{1}{2}}_{\alpha+\frac{1}{4},\beta}+\tilde{q}_{\alpha+\frac{1}{4},\beta})-g(\tilde{q}_{\alpha+\frac{1}{4},\beta})\Big]\Big\}
\end{aligned}
\tag{4.23}
$$

where the unit normal vectors $\eta_{\alpha,\beta}$ are

$$
\eta_{j,\beta} = \frac{(y_{j-\frac{1}{4},\beta}-y_{j+\frac{1}{4},\beta}, x_{j+\frac{1}{4},\beta}-x_{j-\frac{1}{4},\beta})}{|z_{\alpha+\frac{1}{4},\beta}-z_{\alpha-\frac{1}{4},\beta}|} \quad \text{and} \quad \eta_{j+\frac{1}{2},\beta} = (0,1). \tag{4.24}
$$

To obtain the solution at the points $z_{j\pm\frac{1}{4},k\pm\frac{1}{4}}$,we apply the Taylor expansion twice (once in time and once in space), and then get the second-order approximation

$$
\begin{aligned}
\Delta q^{n+\frac{1}{2}}_{j\pm\frac{1}{4},k\pm\frac{1}{4}} &:= \Delta q^{n}_{j\pm\frac{1}{4},k\pm\frac{1}{4}} + \frac{\Delta t}{2}((\Delta q)_t)^n_{j\pm\frac{1}{4},k\pm\frac{1}{4}} \\
&= \Delta q^{n}_{j\pm\frac{1}{4},k\pm\frac{1}{4}} - \frac{\Delta t}{2}\Big[[(f(\Delta q+\tilde{q}) - f(\tilde{q}))_x]^n_{j\pm\frac{1}{4},k\pm\frac{1}{4}} \\
&\qquad + [(g(\Delta q+\tilde{q}) - g(\tilde{q}))_y]^n_{j\pm\frac{1}{4},k\pm\frac{1}{4}}\Big] + \frac{\Delta t}{2} S(\Delta q)^n_{j\pm\frac{1}{4},k\pm\frac{1}{4}} \\
&:= \Delta q^{n}_{j\pm\frac{1}{4},k\pm\frac{1}{4}} - \frac{\Delta t}{2}\Big[[(f(\Delta q+\tilde{q}) - f(\tilde{q}))_x]^n_{j,k} + [(g(\Delta q+\tilde{q}) - g(\tilde{q}))_y]^n_{j,k}\Big] \\
&\qquad + \frac{\Delta t}{2} S(\Delta q)^n_{j\pm\frac{1}{4},k\pm\frac{1}{4}},
\end{aligned}
\tag{4.25}
$$

where the values of $\Delta q^n_{j\pm\frac{1}{4},k\pm\frac{1}{4}}$ are estimated by the Taylor expansion:

$$
\Delta q^n_{j\pm\frac{1}{4},k\pm\frac{1}{4}} := \Delta q^n_{j,k} + (x_{j\pm\frac{1}{4},k\pm\frac{1}{4}} - x_j)(\Delta q_x)^n_{j,k} + (y_{j\pm\frac{1}{4},k\pm\frac{1}{4}} - y_k)(\Delta q_y)^n_{j,k}, \tag{4.26}
$$

and the slopes $(F(\Delta q)_x)^n_{j,k} := ((f(\Delta q+\tilde{q}) - f(\tilde{q}))_x)^n_{j,k}$ and $(G(\Delta q)_y)^n_{j,k} := ((g(\Delta q+\tilde{q}) - g(\tilde{q}))_y)^n_{j,k}$ are obtained by using the $(MC-\theta)$ limiter:

$$
\begin{aligned}
(F(\Delta q)_x)^n_{j,k} &= \operatorname{minmod}\left(\theta\Delta^+_x\left(F(\Delta q^n_{j,k})\right),\, \Delta^0_x\left(F(\Delta q^n_{j,k})\right),\, \theta\Delta^-_x\left(F(\Delta q^n_{j,k})\right)\right), \\
(G(\Delta q)_y)^n_{j,k} &= \operatorname{minmod}\left(\theta\Delta^+_y\left(G(\Delta q^n_{j,k})\right),\, \Delta^0_y\left(G(\Delta q^n_{j,k})\right),\, \theta\Delta^-_y\left(G(\Delta q^n_{j,k})\right)\right), \\
&\qquad \theta \in [1,2]
\end{aligned}
\tag{4.27}
$$

Equipped with (4.20), (4.21), (4.23) and (4.24), the flux integral on (4.16) results in

$$
\begin{aligned}
\frac{1}{|D_{\alpha,\beta}|}\int_{t^n}^{t^{n+1}} \oint_{\partial D_{\alpha,\beta}} \eta^x[f(\Delta q+\tilde{q}) - f(\tilde{q})] + \eta^y[g(\Delta q+\tilde{q}) - g(\tilde{q})]dxdydt \\
= \frac{\Delta t^n}{|D_{\alpha,\beta}|}\left[H_{\alpha+\frac{1}{4},\beta} - H_{\alpha-\frac{1}{4},\beta} + H_{\alpha,\beta+\frac{1}{4}} - H_{\alpha,\beta-\frac{1}{4}}\right]
\end{aligned}
\tag{4.28}
$$

Finally we consider the last integral on (4.16). To evaluate the average of the source term, we use the midpoint rule and the value at center of mass $\Delta q(z^I_{j,k}, t)$

to define the approximation of the source term:

$$\begin{aligned}
&\frac{1}{|D_{j,k}|}\int_{t^n}^{t^{n+1}}\iint_{D_{j,k}} S(\Delta q(x,y,t))dxdydt \\
&:= \frac{\Delta t}{|D_{j,k}|}\iint_{D_{j,k}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy \\
&= \frac{\Delta t}{|D_{j,k}|}|D_{j,k}|S(\Delta q(z^C_{j,k},t^{n+\frac{1}{2}})) \\
&=: \Delta t S_{D_{j,k}}(\Delta q)
\end{aligned} \tag{4.29}$$

$$\begin{aligned}
&\frac{1}{|D_{j+\frac{1}{2},k}|}\int_{t^n}^{t^{n+1}}\iint_{D_{j+\frac{1}{2},k}} S(\Delta q(x,y,t))dxdydt \\
&:= \frac{\Delta t}{|D_{j+\frac{1}{2},k}|}\iint_{D_{j+\frac{1}{2},k}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy \\
&= \frac{\Delta t}{|D_{j+\frac{1}{2},k}|}\left[\iint_{C^E_{j,k}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy + \iint_{C^W_{j+1,k}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy\right] \\
&= \frac{\Delta t}{|D_{j+\frac{1}{2},k}|}\left[|C^E_{j,k}|S(\Delta q(z^E_{j,k},t^{n+\frac{1}{2}})) + |C^W_{j+1,k}|S(\Delta q(z^W_{j+1,k},t^{n+\frac{1}{2}}))\right] \\
&=: \Delta t S_{D_{j+\frac{1}{2},k}}(\Delta q)
\end{aligned} \tag{4.30}$$

$$\begin{aligned}
&\frac{1}{|D_{j,k+\frac{1}{2}}|}\int_{t^n}^{t^{n+1}}\iint_{D_{j,k+\frac{1}{2}}} S(\Delta q(x,y,t))dxdydt \\
&:= \frac{\Delta t}{|D_{j,k+\frac{1}{2}}|}\iint_{D_{j,k+\frac{1}{2}}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy \\
&= \frac{\Delta t}{|D_{j,k+\frac{1}{2}}|}\left[\iint_{C^N_{j,k}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy + \iint_{C^S_{j,k+1}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy\right] \\
&= \frac{\Delta t}{|D_{j,k+\frac{1}{2}}|}\left[|C^N_{j,k}|S(\Delta q(z^N_{j,k},t^{n+\frac{1}{2}})) + |C^S_{j,k+1}|S(\Delta q(z^S_{j,k+1},t^{n+\frac{1}{2}}))\right] \\
&=: \Delta t S_{D_{j,k+\frac{1}{2}}}(\Delta q)
\end{aligned} \tag{4.31}$$

$$
\begin{aligned}
&\frac{1}{|D_{j+\frac{1}{2},k+\frac{1}{2}}|}\int_{t^n}^{t^{n+1}}\iint_{D_{j+\frac{1}{2},k+\frac{1}{2}}} S(\Delta q(x,y,t))dxdydt \\
&:= \frac{\Delta t}{|D_{j+\frac{1}{2},k+\frac{1}{2}}|}\iint_{D_{j+\frac{1}{2},k+\frac{1}{2}}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy \\
&= \frac{\Delta t}{|D_{j+\frac{1}{2},k+\frac{1}{2}}|}[\iint_{C^{NE}_{j,k}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy + \iint_{C^{NW}_{j+1,k}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy \\
&\qquad + \iint_{C^{SE}_{j,k+1}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy + \iint_{C^{SW}_{j+1,k+1}} S(\Delta q(x,y,t^{n+\frac{1}{2}}))dxdy] \\
&= \frac{\Delta t}{|D_{j+\frac{1}{2},k+\frac{1}{2}}}[|C^{NE}_{j,k}|S(\Delta q(z^{NE}_{j,k},t^{n+\frac{1}{2}})) + |C^{NW}_{j+1,k}|S(\Delta q(z^{NW}_{j+1,k},t^{n+\frac{1}{2}})) \\
&\qquad + |C^{SE}_{j,k+1}|S(\Delta q(z^{SE}_{j,k+1},t^{n+\frac{1}{2}})) + |C^{SW}_{j+1,k+1}|S(\Delta q(z^{SW}_{j+1,k+1},t^{n+\frac{1}{2}}))] \\
&=: \Delta t S_{D_{j+\frac{1}{2},k+\frac{1}{2}}}(\Delta q)
\end{aligned}
\tag{4.32}
$$

Similar to the approximation (4.25), the value of $\Delta q(z^I_{j,k}, t^{n+\frac{1}{2}})$ is given by the Taylor expansion,

$$
\begin{aligned}
\Delta q^{n+\frac{1}{2}}_{z^I_{j,k}} &:= \Delta q^n_{z^I_{j,k}} + \frac{\Delta t}{2}(\Delta q_t)^n_{z^I_{j,k}} \\
&= \Delta q^n_{z^I_{j,k}} - \frac{\Delta t}{2}\Big[[(f(\Delta q+\tilde{q}) - f(\tilde{q}))_x]^n_{z^I_{j,k}} \\
&\qquad + [(g(\Delta q+\tilde{q}) - g(\tilde{q}))_y]^n_{z^I_{j,k}}\Big] + \frac{\Delta t}{2}S(\Delta q)^n_{z^I_{j,k}} \\
&:= \Delta q^n_{z^I_{j,k}} - \frac{\Delta t}{2}\Big[[(f(\Delta q+\tilde{q}) - f(\tilde{q}))_x]^n_{j,k} + [(g(\Delta q+\tilde{q}) - g(\tilde{q}))_y]^n_{j,k}\Big] \\
&\qquad + \frac{\Delta t}{2}S(\Delta q)^n_{z^I_{j,k}}
\end{aligned}
\tag{4.33}
$$

To summarize this evolution step, we substitute the results of the cell-average $\overline{(\Delta q)^n_D}$ (4.18), the flux integral (4.28), and the source integral (4.29)-(4.32) for (4.16),

$$
\begin{aligned}
\overline{w}^{n+1}_{D_{\alpha,\beta}} = \overline{\Delta q^n_{D_{\alpha,\beta}}} - \frac{\Delta t^n}{|D_{\alpha,\beta}|}\Big[H_{\alpha+\frac{1}{4},\beta} - H_{\alpha-\frac{1}{4},\beta} + H_{\alpha,\beta+\frac{1}{4}} - H_{\alpha,\beta-\frac{1}{4}}\Big] \\
+ \Delta t S_{D_{\alpha,\beta}}(\Delta q),
\end{aligned}
\tag{4.34}
$$

where $\alpha = j$ or $j \pm \frac{1}{2}$ and $\beta = k$ or $k \pm \frac{1}{2}$.

4.1.3 Projection

At this final procedure, we project the intermediate solutions $\overline{w}^{n+1}_D$ in (4.34) back onto the original uniform cell $C_{j,k}$. In order to smooth the solution in this step, we need to define a piecewise-linear reconstruction $\widetilde{W}^{n+1}(x, y)$. Then we can integrate it over the original cell $C_{j,k}$ to obtain the new cell-average at t^{n+1},

$$(\Delta q)^{n+1}_{j,k} = \frac{1}{\Delta x \Delta y} \iint_{C_{j,k}} \widetilde{W}^{n+1}(x, y)\, dxdy. \tag{4.35}$$

We discuss the reconstruction $\widetilde{W}^{n+1}$ over smooth and unsmooth domains separately. Firstly, considering the central smooth subdomain $D_{j,k}$, since $D_{j,k} \subset C_{j,k}$, the intermediate solution $\overline{w}^{n+1}_{D_{j,k}}$ is smooth enough and does not need a reconstruction. Thus,

$$\widetilde{W}^{n+1}(x, y) = \overline{w}^{n+1}_{D_{j,k}} \quad \text{for } (x, y) \in D_{j,k}. \tag{4.36}$$

Next, we consider the unsmooth subdomains $D_{\alpha,\beta}$ with $(\alpha, \beta) = (j + \frac{1}{2}, k)$, $(j + \frac{1}{2}, k + \frac{1}{2})$, or $(j, k + \frac{1}{2})$, and define the reconstruction of $D_{\alpha,\beta}$ by

$$\widetilde{W}^{n+1}_{D_{\alpha,\beta}} = \overline{w}^{n+1}_{D_{\alpha,\beta}} + (x - z^{n,x}_{D_{\alpha,\beta}})(w_x)^{n+1}_{D_{\alpha,\beta}} + (y - z^{n,y}_{D_{\alpha,\beta}})(w_y)^{n+1}_{D_{\alpha,\beta}} \tag{4.37}$$

where $z^n_{D_{\alpha,\beta}}$ denote the center of the mass of domain $D_{\alpha,\beta}$, and the spatial derivatives are determined by

$$\begin{aligned}
(w_x)^{n+1}_{D_{\alpha,\beta}} &= \text{minmod}\Big(\theta \frac{\overline{w}^{n+1}_{D_{\alpha,\beta}} - \overline{w}^{n+1}_{D_{\alpha-\frac{1}{2},\beta}}}{z^{n,x}_{D_{\alpha,\beta}} - z^{n,x}_{D_{\alpha-\frac{1}{2},\beta}}}, \frac{\overline{w}^{n+1}_{D_{\alpha+\frac{1}{2},\beta}} - \overline{w}^{n+1}_{D_{\alpha-\frac{1}{2},\beta}}}{z^{n,x}_{D_{\alpha+\frac{1}{2},\beta}} - z^{n,x}_{D_{\alpha-\frac{1}{2},\beta}}}, \theta \frac{\overline{w}^{n+1}_{D_{\alpha+\frac{1}{2},\beta}} - \overline{w}^{n+1}_{D_{\alpha,\beta}}}{z^{n,x}_{D_{\alpha+\frac{1}{2},\beta}} - z^{n,x}_{D_{\alpha,\beta}}}\Big) \\
(w_y)^{n+1}_{D_{\alpha,\beta}} &= \text{minmod}\Big(\theta \frac{\overline{w}^{n+1}_{D_{\alpha,\beta}} - \overline{w}^{n+1}_{D_{\alpha,\beta-\frac{1}{2}}}}{z^{n,y}_{D_{\alpha,\beta}} - z^{n,y}_{D_{\alpha,\beta-\frac{1}{2}}}}, \frac{\overline{w}^{n+1}_{D_{\alpha,\beta+\frac{1}{2}}} - \overline{w}^{n+1}_{D_{\alpha,\beta-\frac{1}{2}}}}{z^{n,y}_{D_{\alpha,\beta+\frac{1}{2}}} - z^{n,y}_{D_{\alpha,\beta-\frac{1}{2}}}}, \theta \frac{\overline{w}^{n+1}_{D_{\alpha,\beta+\frac{1}{2}}} - \overline{w}^{n+1}_{D_{\alpha,\beta}}}{z^{n,y}_{D_{\alpha,\beta+\frac{1}{2}}} - z^{n,y}_{D_{\alpha,\beta}}}\Big), \\
& \qquad 1 \le \theta \le 2.
\end{aligned} \tag{4.38}$$

Finally, with the above reconstruction $\widetilde{W}^{n+1}(x, y)$, we can complete the approximation of integral on the right-hand-side of (4.35) and obtain our new cell-average at t^{n+1},

$$\begin{aligned}
(\Delta q)^{n+1}_{j,k} = \frac{1}{\Delta x \Delta y} \Big[&|D_{j,k}| \overline{w}^{n+1}_{D_{j,k}} + |C^E_{j,k}| W^{n+1}_{j+\frac{1}{2},k} + |C^{NE}_{j,k}| W^{n+1}_{j+\frac{1}{2},k+\frac{1}{2}} \\
&+ |C^N_{j,k}| W^{n+1}_{j,k+\frac{1}{2}} + |C^{NW}_{j,k}| W^{n+1}_{j-\frac{1}{2},k+\frac{1}{2}} + |C^W_{j,k}| W^{n+1}_{j-\frac{1}{2},k} \\
&+ |C^{SW}_{j,k}| W^{n+1}_{j-\frac{1}{2},k-\frac{1}{2}} + |C^S_{j,k}| W^{n+1}_{j,k-\frac{1}{2}} + |C^{SE}_{j,k}| W^{n+1}_{j+\frac{1}{2},k-\frac{1}{2}} \Big].
\end{aligned} \tag{4.39}$$

In order to get the desired solution $q^{n+1}_{j,k}$, we add the hydrostatic solution $\tilde{q}_{j,k}$ and the computational solution $(\Delta q)^{n+1}_{j,k}$,

$$q^{n+1}_{j,k} = (\Delta q)^{n+1}_{j,k} + \tilde{q}_{j,k}. \tag{4.40}$$

4.2 2D Semi-discrete scheme

In this section, we construct a 2D semi-discrete scheme from another KT-type fully-discrete scheme. In section 4.1, we split the control cell to nine subdomains to consider the values at the interfaces, and the nodes of the subdomains are decided by the wave speed $a^{\pm}_{j+\frac{1}{2},k}$ and $b^{\pm}_{j,k+\frac{1}{2}}$ in the x- and y direction. Likewise, the wave speed are again used to decide the ranges of the subdomains in this section; nevertheless, compared to section 4.1, the rectangle subdomains are considered in the following scheme. The reason that we choose the rectangle subdomains instead of the trapezoid is that the area of the rectangle is easier to be formalized and it helps us for computation.

In order to construct the semi-discrete scheme, we firstly build the fully-discrete scheme. We follow the three-step structure as before: Reconstruction, Evolution, Projection. In this section, we refer to [8], [9] and [16].

Reconstruction

In this step, we use the same piecewise linear interpolant in (4.6) , and the same approximations of the derivatives at x- and y direction in (4.7)-(4.8).

Evolution

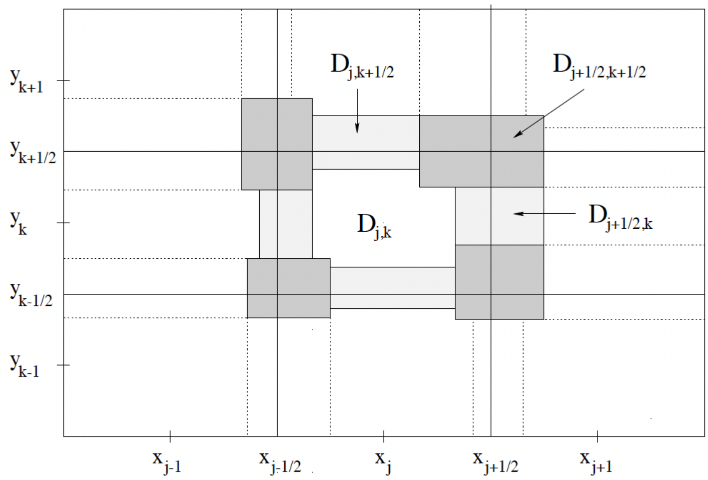
Here, we adopt the same approximation in (4.10) to compute the reconstructed values at the interfaces,

$$\begin{aligned} q^{-}_{j+\frac{1}{2},k} &:= q^{n}_{j,k} + \frac{\Delta x}{2}(q_x)^{n}_{j,k}, \quad q^{+}_{j+\frac{1}{2},k} := q^{n}_{j+1,k} - \frac{\Delta x}{2}(q_x)^{n}_{j+1,k}, \\ q^{-}_{j,k+\frac{1}{2}} &:= q^{n}_{j,k} + \frac{\Delta y}{2}(q_y)^{n}_{j,k}, \quad q^{+}_{j,k+\frac{1}{2}} := q^{n}_{j,k+1} - \frac{\Delta y}{2}(q_y)^{n}_{j,k+1}, \end{aligned} \tag{4.41}$$

and the similar definition of the wave speeds used in (4.9),

$$\begin{aligned} a^{+}_{j+\frac{1}{2},k} &:= max\left\{\lambda_N\left(\frac{\partial}{\partial q}f(q^{+}_{j+\frac{1}{2},k})\right),\ \lambda_N\left(\frac{\partial}{\partial q}f(q^{-}_{j+\frac{1}{2},k})\right),\ 0\right\}, \\ a^{-}_{j+\frac{1}{2},k} &:= min\left\{\lambda_1\left(\frac{\partial}{\partial q}f(q^{+}_{j+\frac{1}{2},k})\right),\ \lambda_1\left(\frac{\partial}{\partial q}f(q^{-}_{j+\frac{1}{2},k})\right),\ 0\right\}, \\ b^{+}_{j,k+\frac{1}{2}} &:= max\left\{\lambda_N\left(\frac{\partial}{\partial q}g(q^{+}_{j,k+\frac{1}{2}})\right),\ \lambda_N\left(\frac{\partial}{\partial q}g(q^{-}_{j,k+\frac{1}{2}})\right),\ 0\right\}, \\ b^{-}_{j,k+\frac{1}{2}} &:= min\left\{\lambda_1\left(\frac{\partial}{\partial q}g(q^{+}_{j,k+\frac{1}{2}})\right),\ \lambda_1\left(\frac{\partial}{\partial q}g(q^{-}_{j,k+\frac{1}{2}})\right),\ 0\right\}. \end{aligned} \tag{4.42}$$

As we mentioned earlier, to estimate the values over the Riemann fans, we split the control cell to nine nonuniform rectangle subdomains, outlined in figure 4.5.



subdomain.png

Figure 4.5: 2D set-up (Reference: Figure 3.2 in [9])

The ranges of the subdomains are defined by

$$\begin{aligned}
D_{j+\frac{1}{2},k} &:= [x_{j+\frac{1}{2}} + a^-_{j+\frac{1}{2},k}\Delta t, x_{j+\frac{1}{2}} + a^+_{j+\frac{1}{2},k}\Delta t] \\
&\qquad \times [y_{k-\frac{1}{2}} + B^+_{j+\frac{1}{2},k-\frac{1}{2}}\Delta t, y_{k+\frac{1}{2}} + B^-_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t], \\
D_{j,k+\frac{1}{2}} &:= [x_{j-\frac{1}{2}} + A^+_{j-\frac{1}{2},k+\frac{1}{2}}\Delta t, x_{j+\frac{1}{2}} + A^-_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t] \\
&\qquad \times [y_{k+\frac{1}{2}} + b^-_{j,k+\frac{1}{2}}\Delta t, y_{k+\frac{1}{2}} + b^+_{j,k+\frac{1}{2}}\Delta t], \\
D_{j+\frac{1}{2},k+\frac{1}{2}} &:= [x_{j+\frac{1}{2}} + A^-_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t, x_{j+\frac{1}{2}} + A^+_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t] \\
&\qquad \times [y_{k+\frac{1}{2}} + B^-_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t, y_{k+\frac{1}{2}} + B^+_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t], \\
D_{j,k} &:= [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \times [y_{k-\frac{1}{2}}, y_{k+\frac{1}{2}}] \setminus \bigcup_{\pm} [D_{j,k\pm\frac{1}{2}} \cup D_{j\pm\frac{1}{2},k} \cup D_{j\pm\frac{1}{2},k\pm\frac{1}{2}}],
\end{aligned} \tag{4.43}$$

where

$$\begin{aligned}
A^+_{j+\frac{1}{2},k+\frac{1}{2}} &:= \max\left\{a^+_{j+\frac{1}{2},k}, a^+_{j+\frac{1}{2},k+1}\right\}, \qquad & B^+_{j+\frac{1}{2},k+\frac{1}{2}} &:= \max\left\{b^+_{j,k+\frac{1}{2}}, b^+_{j+1,k+\frac{1}{2}}\right\}, \\
A^-_{j+\frac{1}{2},k+\frac{1}{2}} &:= \max\left\{a^-_{j+\frac{1}{2},k}, a^-_{j+\frac{1}{2},k+1}\right\}, \qquad & B^-_{j+\frac{1}{2},k+\frac{1}{2}} &:= \max\left\{b^-_{j,k+\frac{1}{2}}, b^-_{j+1,k+\frac{1}{2}}\right\}.
\end{aligned} \tag{4.44}$$

Next, applying (4.14) to subdomain $D_{j+\frac{1}{2},k}-$, $D_{j,k+\frac{1}{2}}-$, $D_{j+\frac{1}{2},k+\frac{1}{2}}-$ and $D_{j,k}$ $\times$

$[t^n, t^{n+1}]$ to compute the interpolation at time t^{n+1}, it yields

$$
\begin{aligned}
\overline{w}^{n+1}_{j+\frac{1}{2},k} = \frac{1}{|D_{j+\frac{1}{2},k}|} \Bigg[& \int_{D_{j+\frac{1}{2},k}} Q(x,y,t^n)\,dxdy \\
& - \int_{t^n}^{t^{n+1}} \int_{y_{k-\frac{1}{2}}+B^+_{j+\frac{1}{2},k-\frac{1}{2}}\Delta t}^{y_{k+\frac{1}{2}}+B^-_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t} [f(\Delta q+\tilde{q})-f(\tilde{q})]_{x=x_{j+\frac{1}{2}}+a^-_{j+\frac{1}{2},k}\Delta t}^{x_{j+\frac{1}{2}}+a^+_{j+\frac{1}{2},k}\Delta t}\,dydt \\
& - \int_{t^n}^{t^{n+1}} \int_{x_{j+\frac{1}{2}}+a^-_{j+\frac{1}{2},k}\Delta t}^{x_{j+\frac{1}{2}}+a^+_{j+\frac{1}{2},k}\Delta t} [g(\Delta q+\tilde{q})-g(\tilde{q})]_{y=y_{k-\frac{1}{2}}+B^+_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t}^{y_{j+\frac{1}{2}}+B^-_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t}\,dxdt \\
& + \int_{t^n}^{t^{n+1}} \iint_{D_{j+\frac{1}{2},k}} S(\Delta q)\,dxdydt \Bigg],
\end{aligned} \tag{4.45}
$$

$$
\begin{aligned}
\overline{w}^{n+1}_{j,k+\frac{1}{2}} = \frac{1}{|D_{j,k+\frac{1}{2}}|} \Bigg[& \int_{D_{j,k+\frac{1}{2}}} Q(x,y,t^n)\,dxdy \\
& - \int_{t^n}^{t^{n+1}} \int_{y_{k+\frac{1}{2}}+b^-_{j,k+\frac{1}{2}}\Delta t}^{y_{k+\frac{1}{2}}+b^+_{j,k+\frac{1}{2}}\Delta t} [f(\Delta q+\tilde{q})-f(\tilde{q})]_{x=x_{j-\frac{1}{2}}+A^+_{j-\frac{1}{2},k+\frac{1}{2}}\Delta t}^{x_{j+\frac{1}{2}}+A^-_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t}\,dydt \\
& - \int_{t^n}^{t^{n+1}} \int_{x_{j-\frac{1}{2}}+A^+_{j-\frac{1}{2},k+\frac{1}{2}}\Delta t}^{x_{j+\frac{1}{2}}+A^-_{j+\frac{1}{2},k+\frac{1}{2}}\Delta t} [g(\Delta q+\tilde{q})-g(\tilde{q})]_{y=_{k+\frac{1}{2}}+b^-_{j,k+\frac{1}{2}}\Delta t}^{y_{k+\frac{1}{2}}+b^+_{j,k+\frac{1}{2}}\Delta t}\,dxdt \\
& + \int_{t^n}^{t^{n+1}} \iint_{D_{j,k+\frac{1}{2}}} S(\Delta q)\,dxdydt \Bigg],
\end{aligned} \tag{4.46}
$$

$$
\begin{aligned}
\overline{w}^{n+1}_{j+\frac{1}{2},k+\frac{1}{2}} = \frac{1}{|D_{j+\frac{1}{2},k+\frac{1}{2}}|} \Bigg[& \int_{D_{j+\frac{1}{2},k+\frac{1}{2}}} Q(x,y,t^n)\,dxdy \\
& - \int_{t^n}^{t^{n+1}} \iint_{D_{j+\frac{1}{2},k+\frac{1}{2}}} [f(\Delta q+\tilde{q})-f(\tilde{q})]_x\,dxdydt \\
& - \int_{t^n}^{t^{n+1}} \iint_{D_{j+\frac{1}{2},k+\frac{1}{2}}} [g(\Delta q+\tilde{q})-g(\tilde{q})]_y\,dxdydt \\
& + \int_{t^n}^{t^{n+1}} \iint_{D_{j+\frac{1}{2},k+\frac{1}{2}}} S(\Delta q)\,dxdydt \Bigg],
\end{aligned} \tag{4.47}
$$

$$
\begin{aligned}
\overline{w}_{j,k}^{n+1} = \frac{1}{|D_{j,k}|} \Bigg[& \int_{D_{j,k}} Q(x,y,t^n)\,dxdy \\
& - \int_{t^n}^{t^{n+1}} \iint_{D_{j,k}} [f(\Delta q + \tilde{q}) - f(\tilde{q})]_x \,dxdydt \\
& - \int_{t^n}^{t^{n+1}} \iint_{D_{j,k}} [g(\Delta q + \tilde{q}) - g(\tilde{q})]_y \,dxdydt \\
& + \int_{t^n}^{t^{n+1}} \iint_{D_{j,k}} S(\Delta q)\,dxdydt \Bigg].
\end{aligned} \tag{4.48}
$$

Projection

Finally, we project the intermediate cell-average back onto the uniform cell $C_{j,k} = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \times [y_{k-\frac{1}{2}}, y_{k+\frac{1}{2}}]$. Then we obtain

$$
\begin{aligned}
(\Delta q)_{j,k}^{n+1} = \frac{1}{\Delta x \Delta y} \Bigg[& \iint_{C_{j,k}^{E(W)}} \widetilde{W}_{j\pm\frac{1}{2},k}^{n+1}(x,y)dxdy + \iint_{C_{j,k}^{N(S)}} \widetilde{W}_{j,k\pm\frac{1}{2}}^{n+1}(x,y)dxdy \\
& + \iint_{C_{j,k}^{NE(NW)(SW)(SE)}} \widetilde{W}_{j\pm\frac{1}{2},k\pm\frac{1}{2}}^{n+1}(x,y)dxdy + \iint_{D_{j,k}} \overline{w}_{j,k}^{n+1}(x,y)dxdy \Bigg]
\end{aligned} \tag{4.49}
$$

where subdomains $C_{j,k}^{I}$, $I = \{E, W, N, S, NE, NW, SW, SE\}$ are defined in (4.17). Here the reconstruction $\widetilde{W}^{n+1}(x,y)$ is defined similarly to the (4.37) that

$$
\widetilde{W}_{j+\frac{1}{2},k}^{n+1} = \overline{w}_{j+\frac{1}{2},k}^{n+1} + (x - x_{j+\frac{1}{2},k}^{n+1})(w_x)_{j+\frac{1}{2},k}^{n+1} + (y - x_{j+\frac{1}{2},k}^{n+1})(w_y)_{j+\frac{1}{2},k}^{n+1}, \tag{4.50}
$$

where

$$
\begin{aligned}
x_{j+\frac{1}{2},k}^{n+1} &:= x_{j+\frac{1}{2}} + \frac{a_{j+\frac{1}{2},k}^{-} + a_{j+\frac{1}{2},k}^{+}}{2}\Delta t, \\
y_{j+\frac{1}{2},k}^{n+1} &:= y_k + \frac{B_{j+\frac{1}{2},k-\frac{1}{2}}^{+} + B_{j+\frac{1}{2},k+\frac{1}{2}}^{-}}{2}\Delta t,
\end{aligned} \tag{4.51}
$$

are the coordinates of the center of $D_{j+\frac{1}{2},k}$, and $(w_x)_{j+\frac{1}{2},k}^{n+1}$ and $(w_y)_{j+\frac{1}{2},k}^{n+1}$ are the x- and y derivatives. The reconstruction $\widetilde{W}_{j,k+\frac{1}{2}}^{n+1}$ and $\widetilde{W}_{j+\frac{1}{2},k+\frac{1}{2}}^{n+1}$ are defined in the similar way. Due to that $\overline{w}_{j,k}^{n+1}$ is smooth enough, we have $\widetilde{W}_{j,k}^{n+1} = \overline{w}_{j,k}^{n+1}$.

Next, we proceed to the construction of semi-discrete scheme.

Semi-discrete scheme

With the definition of (4.49), to obtain the semi-discrete scheme we compute that

$$
\begin{aligned}
\frac{d(\Delta q)_{j,k}(t)}{dt} &= \lim_{\Delta t \to 0} \frac{\Delta q^{n+1}_{j,k} - \Delta q^n_{j,k}}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left[\frac{1}{\Delta x \Delta y} \left(\iint_{C^{E(W)}_{j,k}} \widetilde{W}^{n+1}_{j\pm\frac{1}{2},k}(x,y)dxdy \right.\right. \\
&+ \iint_{C^{N(S)}_{j,k}} \widetilde{W}^{n+1}_{j,k\pm\frac{1}{2}}(x,y)dxdy + \iint_{C^{NE(NW)(SE)(SW)}_{j,k}} \widetilde{W}^{n+1}_{j\pm\frac{1}{2},k\pm\frac{1}{2}}(x,y)dxdy \\
&+ \left.\left. \iint_{D_{j,k}} \overline{w}^{n+1}_{j,k}(x,y)dxdy \right) - \Delta q^n \right]
\end{aligned}
\tag{4.52}
$$

Due to the conservation property, the last integral on the right-hand-side of (4.52) has the relation that

$$
\iint_{D_{j,k}} \overline{w}^{n+1}_{j,k}(x,y)dxdy = |D_{j,k}|\overline{w}^{n+1}_{j,k}. \tag{4.53}
$$

For $S = (j \pm \frac{1}{2}, k), (j, k \pm \frac{1}{2}), (j \pm \frac{1}{2}, k \pm \frac{1}{2})$, according to the approximation of $\widetilde{W}^{n+1}_S(x,y)$ in (4.50) and assuming the spatial derivatives are bounded independently of Δt, the relation between $\widetilde{W}^{n+1}_S$ and $\overline{w}^{n+1}_S$ is that

$$
\widetilde{W}^{n+1}_S(x,y) = \overline{w}^{n+1}_S + O(\Delta t) \tag{4.54}
$$

Since the areas of the domain $C^{E(W)}_{j,k}$, $C^{N(S)}_{j,k}$, and $C^{NE(NW)(SE)(SW)}_{j,k}$ are evaluated by

$$
\begin{aligned}
|C^{E(W)}_{j,k}| &= \Delta t \Delta y (\mp a^{\mp}_{j\pm\frac{1}{2},k}) + O((\Delta t)^2), \\
|C^{N(S)}_{j,k}| &= \Delta t \Delta x (\mp b^{\mp}_{j,k\pm\frac{1}{2}}) + O((\Delta t)^2), \\
|C^{NE(NW)(SE)(SW)}_{j,k}| &= O((\Delta t)^2),
\end{aligned}
\tag{4.55}
$$

the relation between $\widetilde{W}^{n+1}_S$ and $\overline{w}^{n+1}_S$ can be written as

$$
\begin{aligned}
&\iint_{C^{E(W)}_{j,k}} \widetilde{W}_{j\pm\frac{1}{2},k}dxdy = |C^{E(W)}_{j,k}|\overline{w}^{n+1}_{j\pm\frac{1}{2},k} + O((\Delta t)^2), \\
&\iint_{C^{N(S)}_{j,k}} \widetilde{W}_{j,k\pm\frac{1}{2}}dxdy = |C^{N(S)}_{j,k}|\overline{w}^{n+1}_{j,k\pm\frac{1}{2}} + O((\Delta t)^2), \\
&\iint_{C^{NE(NW)(SE)(SW)}_{j,k}} \widetilde{W}^{n+1}_{j\pm\frac{1}{2},k\pm\frac{1}{2}}dxdy = O((\Delta t)^2).
\end{aligned}
\tag{4.56}
$$

As $\Delta t \to 0$, the values at the corner vanish, because the area are proportional to $(\Delta t)^2$. Thus, (4.52) reduces to

$$\frac{d(\Delta q)_{j,k}(t)}{dt} = \lim_{\Delta t\to 0} \frac{1}{\Delta t}\left[\frac{1}{\Delta x\Delta y}\left(|C^{E(W)}_{j,k}|\overline{w}^{n+1}_{j\pm\frac{1}{2},k} + |C^{N(S)}_{j,k}|\overline{w}^{n+1}_{j,k\pm\frac{1}{2}} + |D_{j,k}|\overline{w}^{n+1}_{j,k}\right) - \Delta q^n_{j,k}\right] \tag{4.57}$$

Applying the areas of $C^{E(W)}_{j,k}$ and $C^{N(S)}_{j,k}$ to the first and the second term on the right-hand-side of (4.57) respectively, we obtain that

$$\begin{aligned}
\lim_{\Delta t\to 0} \frac{|C^{E(W)}_{j,k}|}{\Delta t\Delta x\Delta y}\overline{w}^{n+1}_{j\pm\frac{1}{2},k} &= -\frac{a^{\mp}_{j\pm\frac{1}{2},k}}{\Delta x}\lim_{\Delta t\to 0}\overline{w}^{n+1}_{j\pm\frac{1}{2},k},\\
\lim_{\Delta t\to 0} \frac{|C^{N(S)}_{j,k}|}{\Delta t\Delta x\Delta y}\overline{w}^{n+1}_{j,k\pm\frac{1}{2}} &= -\frac{b^{\mp}_{j,k\pm\frac{1}{2}}}{\Delta y}\lim_{\Delta t\to 0}\overline{w}^{n+1}_{j,k\pm\frac{1}{2}}.
\end{aligned} \tag{4.58}$$

Then using the approximations of $\overline{w}^{n+1}_{j+\frac{1}{2},k}$ and $\overline{w}^{n+1}_{j,k+\frac{1}{2}}$ in (4.45) and (4.46) to substitute the values in (4.58) results in

$$\begin{aligned}
\lim_{\Delta t\to 0} \frac{|C^{E(W)}_{j,k}|}{\Delta t\Delta x\Delta y}\overline{w}^{n+1}_{j\pm\frac{1}{2},k} = &-\frac{1}{(a^{+}_{j\pm\frac{1}{2},k} - a^{-}_{j\pm\frac{1}{2},k})\Delta x}\times\\
&\left[a^{-}_{j\pm\frac{1}{2},k}a^{+}_{j\pm\frac{1}{2},k}(\Delta q)^{W(E)}_{j\pm 1,k} - (a^{\mp}_{j\pm\frac{1}{2},k})^2(\Delta q)^{E(W)}_{j,k}\right]\\
&+\frac{a^{\mp}_{j\pm\frac{1}{2},k}}{(a^{+}_{j\pm\frac{1}{2},k} - a^{-}_{j\pm\frac{1}{2},k})\Delta x}\left[F((\Delta q)^{W(E)}_{j\pm 1,k}) - F((\Delta q)^{E(W)}_{j,k})\right],\\
\lim_{\Delta t\to 0} \frac{|C^{N(S)}_{j,k}|}{\Delta t\Delta x\Delta y}\overline{w}^{n+1}_{j,k\pm\frac{1}{2}} = &-\frac{1}{(b^{+}_{j,k\pm\frac{1}{2}} - b^{-}_{j,k\pm\frac{1}{2}})\Delta y}\times\\
&\left[b^{-}_{j,k\pm\frac{1}{2}}b^{+}_{j,k\pm\frac{1}{2}}(\Delta q)^{S(N)}_{j,k\pm 1} - (b^{\mp}_{j,k\pm\frac{1}{2}})^2(\Delta q)^{N(S)}_{j,k}\right]\\
&+\frac{b^{\mp}_{j,k\pm\frac{1}{2}}}{(b^{+}_{j,k\pm\frac{1}{2}} - b^{-}_{j,k\pm\frac{1}{2}})\Delta y}\left[G((\Delta q)^{S(N)}_{j,k\pm 1}) - G((\Delta q)^{N(S)}_{j,k})\right],
\end{aligned} \tag{4.59}$$

where

$$\begin{aligned}
F((\Delta q)^{W}_{j+1,k}) &= f((\Delta q)^{W}_{j+1,k} + \tilde{q}_{j+\frac{1}{2},k}) - f(\tilde{q}_{j+\frac{1}{2},k}),\\
F((\Delta q)^{E}_{j,k}) &= f((\Delta q)^{E}_{j,k} + \tilde{q}_{j+\frac{1}{2},k}) - f(\tilde{q}_{j+\frac{1}{2},k}),\\
G((\Delta q)^{S}_{j,k+1}) &= g((\Delta q)^{S}_{j,k+1} + \tilde{q}_{j,k+\frac{1}{2}}) - g(\tilde{q}_{j,k+\frac{1}{2}}),\\
G((\Delta q)^{N}_{j,k}) &= g((\Delta q)^{N}_{j,k} + \tilde{q}_{j,k+\frac{1}{2}}) - g(\tilde{q}_{j,k+\frac{1}{2}}),
\end{aligned} \tag{4.60}$$

and

$$\begin{aligned} F((\Delta q)^E_{j-1,k}) &= f((\Delta q)^E_{j-1,k} + \tilde{q}_{j-\frac{1}{2},k}) - f(\tilde{q}_{j-\frac{1}{2},k}), \\ F((\Delta q)^W_{j,k}) &= f((\Delta q)^W_{j,k} + \tilde{q}_{j-\frac{1}{2},k}) - f(\tilde{q}_{j-\frac{1}{2},k}), \\ G((\Delta q)^N_{j,k-1}) &= g((\Delta q)^N_{j,k-1} + \tilde{q}_{j,k-\frac{1}{2}}) - g(\tilde{q}_{j,k-\frac{1}{2}}), \\ G((\Delta q)^S_{j,k}) &= g((\Delta q)^S_{j,k} + \tilde{q}_{j,k-\frac{1}{2}}) - g(\tilde{q}_{j,k-\frac{1}{2}}). \end{aligned} \tag{4.61}$$

Here, the notation $(\Delta q)^I$ with $I = \{E, W, N, S\}$ is defined similarly to (4.41) by

$$\begin{aligned} (\Delta q)^E_{j,k} &:= (\Delta q)^n_{j,k} + \frac{\Delta x}{2}((\Delta q)_x)^n_{j,k}, \quad (\Delta q)^W_{j,k} := (\Delta q)^n_{j,k} - \frac{\Delta x}{2}((\Delta q)_x)^n_{j,k}, \\ (\Delta q)^N_{j,k} &:= (\Delta q)^n_{j,k} + \frac{\Delta y}{2}((\Delta q)_y)^n_{j,k}, \quad (\Delta q)^S_{j,k} := (\Delta q)^n_{j,k} - \frac{\Delta y}{2}((\Delta q)_y)^n_{j,k}, \end{aligned} \tag{4.62}$$

with the $MC - \theta$ limiter

$$\begin{aligned} &((\Delta q)_x)^n_{j,k} \\ &:= \text{minmod}\left(\theta\frac{(\Delta q)^n_{j,k} - (\Delta q)^n_{j-1,k}}{\Delta x}, \frac{(\Delta q)^n_{j+1,k} - (\Delta q)^n_{j-1,k}}{2\Delta x}, \theta\frac{(\Delta q)^n_{j+1,k} - (\Delta q)^n_{j,k}}{\Delta x}\right), \\ &((\Delta q)_x)^n_{j,k} \\ &:= \text{minmod}\left(\theta\frac{(\Delta q)^n_{j,k} - (\Delta q)^n_{j,k-1}}{\Delta y}, \frac{(\Delta q)^n_{j,k+1} - (\Delta q)^n_{j,k-1}}{2\Delta y}, \theta\frac{(\Delta q)^n_{j,k+1} - (\Delta q)^n_{j,k}}{\Delta y}\right), \end{aligned} \tag{4.63}$$

and $1 \leq \theta \leq 2$. The detailed computation of (4.59) is similar to the derivation in section 3.3 of [8].

Next, we consider the rest terms on the right-hand-side of (4.57). The sub-domain $D_{j,k}$ can be regarded as a rectangle when $\Delta t \to 0$, up to small corners of a negligible size $O((\Delta t)^2)$. Applying (4.48) to the rest terms yields that

$$\begin{aligned} &\lim_{\Delta t \to 0} \frac{1}{\Delta t}\left[\frac{1}{\Delta x \Delta y}|D_{j,k}|\overline{w}^{n+1}_{j,k} - \Delta q^n_{j,k}\right] = \\ &\quad \left[\frac{a^-_{j+\frac{1}{2},k}}{\Delta x}(\Delta q)^E_{j,k} + \frac{b^-_{j,k+\frac{1}{2}}}{\Delta y}(\Delta q)^N_{j,k} - \frac{a^+_{j-\frac{1}{2},k}}{\Delta x}(\Delta q)^W_{j,k} - \frac{b^+_{j,k-\frac{1}{2}}}{\Delta y}(\Delta q)^S_{j,k}\right] \\ &\quad - \frac{1}{\Delta x}\left[F((\Delta q)^E_{j,k}) - F((\Delta q)^W_{j,k})\right] - \frac{1}{\Delta y}\left[G((\Delta q)^N_{j,k}) - G((\Delta q)^S_{j,k})\right] \\ &\quad + S((\Delta q)_{j,k})] \end{aligned} \tag{4.64}$$

Finally, combining (4.59) and (4.64), the semi-discrete scheme takes form of

$$\frac{d}{dt}(\Delta q)_{j,k}(t) = -\frac{H^x_{j+\frac{1}{2},k}(t) - H^x_{j-\frac{1}{2},k}(t)}{\Delta x} - \frac{H^y_{j,k+\frac{1}{2}}(t) - H^y_{j,k-\frac{1}{2}}(t)}{\Delta y} + S((\Delta q)_{j,k}(t)) \tag{4.65}$$

with the numerical fluxes

$$
\begin{aligned}
H^x_{j+\frac{1}{2},k} =& \frac{a^+_{j+\frac{1}{2},k} F((\Delta q)^E_{j,k}) - a^-_{j+\frac{1}{2},k} F((\Delta q)^W_{j+1,k})}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}} \\
&+ \frac{a^+_{j+\frac{1}{2},k} a^-_{j+\frac{1}{2},k}}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}} \left[(\Delta q)^W_{j+1,k} - (\Delta q)^E_{j,k} \right], \\
H^y_{j,k+\frac{1}{2}} =& \frac{b^+_{j,k+\frac{1}{2}} G((\Delta q)^N_{j,k}) - b^-_{j,k+\frac{1}{2}} G((\Delta q)^S_{j,k+1})}{b^+_{j,k+\frac{1}{2}} - b^-_{j,k+\frac{1}{2}}} \\
&+ \frac{b^+_{j,k+\frac{1}{2}} b^-_{j,k+\frac{1}{2}}}{b^+_{j,k+\frac{1}{2}} - b^-_{j,k+\frac{1}{2}}} \left[(\Delta q)^S_{j,k+1} - (\Delta q)^N_{j,k} \right].
\end{aligned}
\tag{4.66}
$$

4.3 Maximum Principle

In this section, we discuss the maximum principle of the 2D semi-discrete scheme (4.65) applied to homogeneous scalar conservation laws,

$$
\frac{d}{dt}(\Delta q)_{j,k}(t) = -\frac{H^x_{j+\frac{1}{2},k}(t) - H^x_{j-\frac{1}{2},k}(t)}{\Delta x} - \frac{H^y_{j,k+\frac{1}{2}}(t) - H^y_{j,k-\frac{1}{2}}(t)}{\Delta y}, \tag{4.67}
$$

with the numerical fluxes (4.66). For the following theorem we refer to [3], [9] and [11].

Theorem 4.3.1 (Maximum principle) *Consider the modified scalar conservation law*

$$
(\Delta q)_t + [f(\Delta q + \tilde{q}) - f(\tilde{q})]_x + [g(\Delta q + \tilde{q}) - g(\tilde{q})]_y = 0. \tag{4.68}
$$

and the forward Euler time discretization of the 2-dimensional scheme (4.67)

$$
(\Delta q)^{n+1}_{j,k} = (\Delta q)^n_{j,k} - \frac{\Delta t^n}{\Delta x}\left[H^x_{j+\frac{1}{2},k}(t) - H^x_{j-\frac{1}{2},k}(t)\right] - \frac{\Delta t^n}{\Delta x}\left[H^y_{j,k+\frac{1}{2}}(t) - H^y_{j,k-\frac{1}{2}}(t)\right] \tag{4.69}
$$

with the numerical fluxes (4.66) *and the minmod limiter* (4.63).
Assume the following CFL condition holds:

$$
\max\left(\frac{\Delta t^n}{\Delta x} \max_{\Delta q} |F'(\Delta q)|, \frac{\Delta t^n}{\Delta y} \max_{\Delta q} |G'(\Delta q)| \right) \leq \frac{1}{8}, \tag{4.70}
$$

then the fully-discrete scheme (4.69) *satisfies the maximum principle*

$$
\max_{j,k}\{(\Delta q)^{n+1}_{j,k}\} \leq \max_{j,k}\{(\Delta q)^n_{j,k}\}. \tag{4.71}
$$

Proof. We begin with the explicit form of (4.69),

$$\begin{aligned}(\Delta q)^{n+1}_{j,k} = (\Delta q)^n_{j,k} &- \frac{\lambda^n}{(a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k})}\left[a^+_{j+\frac{1}{2},k}F((\Delta q)^E_{j,k}) - a^-_{j+\frac{1}{2},k}F((\Delta q)^W_{j+1,k})\right]\\
&+ \frac{\lambda^n}{(a^+_{j-\frac{1}{2},k} - a^-_{j-\frac{1}{2},k})}\left[a^+_{j-\frac{1}{2},k}F((\Delta q)^E_{j-1,k}) - a^-_{j-\frac{1}{2},k}F((\Delta q)^W_{j,k})\right]\\
&- \frac{\lambda^n(a^+_{j+\frac{1}{2},k}a^-_{j+\frac{1}{2},k})}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}}\left[(\Delta q)^W_{j+1,k} - (\Delta q)^E_{j,k}\right]\\
&+ \frac{\lambda^n(a^+_{j-\frac{1}{2},k}a^-_{j-\frac{1}{2},k})}{a^+_{j-\frac{1}{2},k} - a^-_{j-\frac{1}{2},k}}\left[(\Delta q)^W_{j,k} - (\Delta q)^E_{j-1,k}\right]\\
&- \frac{\mu^n}{(b^+_{j,k+\frac{1}{2}} - b^-_{j,k+\frac{1}{2}})}\left[b^+_{j,k+\frac{1}{2}}G((\Delta q)^N_{j,k}) - b^-_{j,k+\frac{1}{2}}G((\Delta q)^S_{j,k+1})\right]\\
&+ \frac{\mu^n}{(b^+_{j,k-\frac{1}{2}} - b^-_{j,k-\frac{1}{2}})}\left[b^+_{j,k-\frac{1}{2}}G((\Delta q)^N_{j,k-1}) - b^-_{j,k-\frac{1}{2}}G((\Delta q)^S_{j,k})\right]\\
&- \frac{\mu^n(b^+_{j,k+\frac{1}{2}}b^-_{j,k+\frac{1}{2}})}{b^+_{j,k+\frac{1}{2}} - b^-_{j,k+\frac{1}{2}}}\left[(\Delta q)^S_{j,k+1} - (\Delta q)^N_{j,k}\right]\\
&+ \frac{\mu^n(b^+_{j,k-\frac{1}{2}}b^-_{j,k-\frac{1}{2}})}{b^+_{j,k-\frac{1}{2}} - b^-_{j,k-\frac{1}{2}}}\left[(\Delta q)^S_{j,k} - (\Delta q)^N_{j,k-1}\right].\end{aligned} \tag{4.72}$$

All the terms on the right-hand of (4.72) are taken at the time step t^n. By the definition (4.62), we have the equality

$$(\Delta q)^n_{j,k} = \frac{(\Delta q)^E_{j,k} + (\Delta q)^W_{j,k} + (\Delta q)^N_{j,k} + (\Delta q)^S_{j,k}}{4}. \tag{4.73}$$

Then we substitute (4.73) for $(\Delta q)^n_{j,k}$ and adjust the other terms in (4.72),

$$\begin{aligned}(\Delta q)^{n+1}_{j,k} =& \frac{(\Delta q)^E_{j,k} + (\Delta q)^W_{j,k} + (\Delta q)^N_{j,k} + (\Delta q)^S_{j,k}}{4}\\
&+ \lambda^n\left[\frac{a^-_{j+\frac{1}{2},k}}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}}\left[F((\Delta q)^W_{j+1,k}) - F((\Delta q)^E_{j,k})\right] - F((\Delta q)^E_{j,k})\right]\\
&- \lambda^n\left[\frac{a^+_{j-\frac{1}{2},k}}{a^+_{j-\frac{1}{2},k} - a^-_{j-\frac{1}{2},k}}\left[F((\Delta q)^W_{j,k}) - F((\Delta q)^E_{j-1,k})\right] - F((\Delta q)^W_{j,k})\right]\\
&- \lambda^n\frac{(a^+_{j+\frac{1}{2},k}a^-_{j+\frac{1}{2},k})}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}}\left[(\Delta q)^W_{j+1,k} - (\Delta q)^E_{j,k}\right]\\
&+ \lambda^n\frac{(a^+_{j-\frac{1}{2},k}a^-_{j-\frac{1}{2},k})}{a^+_{j-\frac{1}{2},k} - a^-_{j-\frac{1}{2},k}}\left[(\Delta q)^W_{j,k} - (\Delta q)^E_{j-1,k}\right]\end{aligned}$$

$$
\begin{aligned}
&+\mu^n\left[\frac{b^-_{j,k+\frac{1}{2}}}{b^+_{j,k+\frac{1}{2}}-b^-_{j,k+\frac{1}{2}}}\Big[G((\Delta q)^S_{j,k+1})-G((\Delta q)^N_{j,k})\Big]-G((\Delta q)^N_{j,k})\right]\\
&-\mu^n\left[\frac{b^+_{j,k-\frac{1}{2}}}{b^+_{j,k-\frac{1}{2}}-b^-_{j,k-\frac{1}{2}}}\Big[G((\Delta q)^S_{j,k})+G((\Delta q)^N_{j,k-1})\Big]-G((\Delta q)^S_{j,k})\right]\\
&-\mu^n\frac{(b^+_{j,k+\frac{1}{2}}b^-_{j,k+\frac{1}{2}})}{b^+_{j,k+\frac{1}{2}}-b^-_{j,k+\frac{1}{2}}}\Big[(\Delta q)^S_{j,k+1}-(\Delta q)^N_{j,k}\Big]\\
&+\mu^n\frac{(b^+_{j,k-\frac{1}{2}}b^-_{j,k-\frac{1}{2}})}{b^+_{j,k-\frac{1}{2}}-b^-_{j,k-\frac{1}{2}}}\Big[(\Delta q)^S_{j,k}-(\Delta q)^N_{j,k-1}\Big].
\end{aligned}
\tag{4.74}
$$

To simplify notations, we use the abbreviations used in [3]

$$
\begin{aligned}
\Delta^x_{j+\frac{1}{2},k}(\Delta q) &:= (\Delta q)^W_{j+1,k}(t^n)-(\Delta q)^E_{j,k}(t^n),\\
\Delta^x_{j,k}F &:= F((\Delta q)^E_{j,k})-F((\Delta q)^W_{j,k})\\
\Delta^x_{j,k}G &:= G((\Delta q)^N_{j,k})-G((\Delta q)^S_{j,k}).
\end{aligned}
\tag{4.75}
$$

Then we rewrite the (4.74) as

$$
\begin{aligned}
(\Delta q)^{n+1}_{j,k} =&\frac{(\Delta q)^E_{j,k}+(\Delta q)^W_{j,k}+(\Delta q)^N_{j,k}+(\Delta q)^S_{j,k}}{4}\\
&+\lambda^n\left[\frac{a^-_{j+\frac{1}{2},k}}{a^+_{j+\frac{1}{2},k}-a^-_{j+\frac{1}{2},k}}\frac{\Delta^x_{j+\frac{1}{2},k}F}{\Delta^x_{j+\frac{1}{2},k}(\Delta q)}\Big[(\Delta q)^W_{j+1,k}-(\Delta q)^E_{j,k}\Big]\right.\\
&\qquad-\frac{\Delta^x_{j,k}F}{\Delta^x_{j,k}(\Delta q)}\Big[(\Delta q)^E_{j,k}-(\Delta q)^W_{j,k}\Big]\\
&\qquad\left.-\frac{a^+_{j-\frac{1}{2},k}}{a^+_{j-\frac{1}{2},k}-a^-_{j-\frac{1}{2},k}}\frac{\Delta^x_{j-\frac{1}{2},k}F}{\Delta^x_{j-\frac{1}{2},k}(\Delta q)}\Big[(\Delta q)^W_{j,k}-(\Delta q)^E_{j-1,k}\Big]\right]\\
&-\lambda^n\frac{(a^+_{j+\frac{1}{2},k}a^-_{j+\frac{1}{2},k})}{a^+_{j+\frac{1}{2},k}-a^-_{j+\frac{1}{2},k}}\Big[(\Delta q)^W_{j+1,k}-(\Delta q)^E_{j,k}\Big]\\
&+\lambda^n\frac{(a^+_{j-\frac{1}{2},k}a^-_{j-\frac{1}{2},k})}{a^+_{j-\frac{1}{2},k}-a^-_{j-\frac{1}{2},k}}\Big[(\Delta q)^W_{j,k}-(\Delta q)^E_{j-1,k}\Big]\\
&+\mu^n\left[\frac{b^-_{j,k+\frac{1}{2}}}{b^+_{j,k+\frac{1}{2}}-b^-_{j,k+\frac{1}{2}}}\frac{\Delta^y_{j,k+\frac{1}{2}}G}{\Delta^y_{j,k+\frac{1}{2}}(\Delta q)}\Big[(\Delta q)^S_{j,k+1}-(\Delta q)^N_{j,k}\Big]\right.\\
&\qquad-\frac{\Delta^y_{j,k}G}{\Delta^y_{j,k}(\Delta q)}\Big[(\Delta q)^N_{j,k}-(\Delta q)^S_{j,k}\Big]\\
&\qquad\left.-\frac{b^+_{j,k-\frac{1}{2}}}{b^+_{j,k-\frac{1}{2}}-b^-_{j,k-\frac{1}{2}}}\frac{\Delta^y_{j,k-\frac{1}{2}}G}{\Delta^y_{j,k-\frac{1}{2}}(\Delta q)}\Big[(\Delta q)^S_{j,k}-(\Delta q)^N_{j,k-1}\Big]\right]\\
&-\mu^n\frac{(b^+_{j,k+\frac{1}{2}}b^-_{j,k+\frac{1}{2}})}{b^+_{j,k+\frac{1}{2}}-b^-_{j,k+\frac{1}{2}}}\Big[(\Delta q)^S_{j,k+1}-(\Delta q)^N_{j,k}\Big]
\end{aligned}
$$

$$+\mu^n \frac{(b^+_{j,k-\frac{1}{2}} b^-_{j,k-\frac{1}{2}})}{b^+_{j,k-\frac{1}{2}} - b^-_{j,k-\frac{1}{2}}} \left[(\Delta q)^S_{j,k} - (\Delta q)^N_{j,k-1}\right] \tag{4.76}$$

Collecting the coefficients of $(\Delta q)^W_{j+1,k}$, $(\Delta q)^{E(W)}_{j,k}$, $(\Delta q)^E_{j-1,k}$, and $(\Delta q)^S_{j,k+1}$, $(\Delta q)^{N(S)}_{j,k}$, $(\Delta q)^N_{j,k-1}$,

$$\begin{aligned}
(\Delta q)^{n+1}_{j,k} =& \lambda^n \left(\frac{a^-_{j+\frac{1}{2},k}}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}} \frac{\Delta^x_{j+\frac{1}{2},k} F}{\Delta^x_{j+\frac{1}{2},k}(\Delta q)} - \frac{a^+_{j+\frac{1}{2},k} a^-_{j+\frac{1}{2},k}}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}} \right) (\Delta q)^W_{j+1,k} \\
&+ \left[\frac{1}{4} - \lambda^n \left(\frac{a^-_{j+\frac{1}{2},k}}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}} \frac{\Delta^x_{j+\frac{1}{2},k} F}{\Delta^x_{j+\frac{1}{2},k}(\Delta q)} + \frac{\Delta^x_{j,k} F}{\Delta^x_{j,k}(\Delta q)} \right.\right. \\
&\qquad \left.\left. - \frac{a^+_{j+\frac{1}{2},k} a^-_{j+\frac{1}{2},k}}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}} \right) \right] (\Delta q)^E_{j,k} \\
&+ \left[\frac{1}{4} + \lambda^n \left(\frac{\Delta^x_{j,k} F}{\Delta^x_{j,k}(\Delta q)} - \frac{a^+_{j-\frac{1}{2},k}}{a^+_{j-\frac{1}{2},k} - a^-_{j-\frac{1}{2},k}} \frac{\Delta^x_{j-\frac{1}{2},k} F}{\Delta^x_{j-\frac{1}{2},k}(\Delta q)} \right.\right. \\
&\qquad \left.\left. + \frac{a^+_{j-\frac{1}{2},k} a^-_{j-\frac{1}{2},k}}{a^+_{j-\frac{1}{2},k} - a^-_{j-\frac{1}{2},k}} \right) \right] (\Delta q)^W_{j,k} \\
&+ \lambda^n \left(\frac{a^+_{j-\frac{1}{2},k}}{a^+_{j-\frac{1}{2},k} - a^-_{j-\frac{1}{2},k}} \frac{\Delta^x_{j-\frac{1}{2},k} F}{\Delta^x_{j-\frac{1}{2},k}(\Delta q)} - \frac{a^+_{j-\frac{1}{2},k} a^-_{j-\frac{1}{2},k}}{a^+_{j-\frac{1}{2},k} - a^-_{j-\frac{1}{2},k}} \right) (\Delta q)^E_{j-1,k} \\
&+ \mu^n \left(\frac{b^-_{j,k+\frac{1}{2}}}{b^+_{j,k+\frac{1}{2}} - b^-_{j,k+\frac{1}{2}}} \frac{\Delta^y_{j,k+\frac{1}{2}} G}{\Delta^y_{j,k+\frac{1}{2}}(\Delta q)} - \frac{b^+_{j,k+\frac{1}{2}} b^-_{j,k+\frac{1}{2}}}{b^+_{j,k+\frac{1}{2}} - b^-_{j,k+\frac{1}{2}}} \right) (\Delta q)^S_{j,k+1} \\
&+ \left[\frac{1}{4} - \mu^n \left(\frac{b^-_{j,k+\frac{1}{2}}}{b^+_{j,k+\frac{1}{2}} - b^-_{j,k+\frac{1}{2}}} \frac{\Delta^y_{j,k+\frac{1}{2}} G}{\Delta^y_{j,k+\frac{1}{2}}(\Delta q)} + \frac{\Delta^y_{j,k} G}{\Delta^y_{j,k}(\Delta q)} \right.\right. \\
&\qquad \left.\left. - \frac{b^+_{j,k+\frac{1}{2}} b^-_{j,k+\frac{1}{2}}}{b^+_{j,k+\frac{1}{2}} - b^-_{j,k+\frac{1}{2}}} \right) \right] (\Delta q)^N_{j,k} \\
&+ \left[\frac{1}{4} + \mu^n \left(\frac{\Delta^y_{j,k} G}{\Delta^y_{j,k}(\Delta q)} - \frac{b^+_{j,k-\frac{1}{2}}}{b^+_{j,k-\frac{1}{2}} - b^-_{j,k-\frac{1}{2}}} \frac{\Delta^y_{j,k-\frac{1}{2}} G}{\Delta^y_{j,k-\frac{1}{2}}(\Delta q)} \right.\right. \\
&\qquad \left.\left. + \frac{b^+_{j,k-\frac{1}{2}} b^-_{j,k-\frac{1}{2}}}{b^+_{j,k-\frac{1}{2}} - b^-_{j,k-\frac{1}{2}}} \right) \right] (\Delta q)^S_{j,k} \\
&+ \mu^n \left(\frac{b^+_{j,k-\frac{1}{2}}}{b^+_{j,k-\frac{1}{2}} - b^-_{j,k-\frac{1}{2}}} \frac{\Delta^y_{j,k-\frac{1}{2}} G}{\Delta^y_{j,k-\frac{1}{2}}(\Delta q)} - \frac{b^+_{j,k-\frac{1}{2}} b^-_{j,k-\frac{1}{2}}}{b^+_{j,k-\frac{1}{2}} - b^-_{j,k-\frac{1}{2}}} \right) (\Delta q)^N_{j,k-1}.
\end{aligned} \tag{4.77}$$

Next, we discuss the coefficients. Due to the lemma 3.1.1 and the fact that $a^\pm_{j+\frac{1}{2},k}$ is the maximal speed on its direction, which is determined from the flux derivatives, (consult the definition (4.42)), we obtain the following inequalities:

$$\begin{aligned} a^+_{j+\frac{1}{2},k} &\geq 0 \qquad \text{and} \qquad \left|\frac{\Delta^x_{j+\frac{1}{2},k}F}{\Delta^x_{j+\frac{1}{2},k}(\Delta q)}\right| \leq a^+_{j+\frac{1}{2},k}, \\ a^-_{j+\frac{1}{2},k} &\leq 0, \qquad \text{and} \qquad \left|\frac{\Delta^x_{j+\frac{1}{2},k}F}{\Delta^x_{j+\frac{1}{2},k}(\Delta q)}\right| \leq -a^-_{j+\frac{1}{2},k}. \end{aligned} \tag{4.78}$$

Hence, the rearranged coefficients of $(\Delta q)^W_{j+1,k}$ and $(\Delta q)^E_{j-1,k}$ are non-negative,

$$\begin{aligned} &\underbrace{\lambda^n \frac{(-a^-_{j+\frac{1}{2},k})}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}}}_{\geq 0} \underbrace{\left[a^+_{j+\frac{1}{2},k} - \frac{\Delta^x_{j+\frac{1}{2},k}F}{\Delta^x_{j+\frac{1}{2},k}(\Delta q)}\right]}_{\geq 0} \geq 0 \\ &\underbrace{\lambda^n \frac{a^+_{j-\frac{1}{2},k}}{a^+_{j-\frac{1}{2},k} - a^-_{j-\frac{1}{2},k}}}_{\geq 0} \underbrace{\left[\frac{\Delta^x_{j-\frac{1}{2},k}F}{\Delta^x_{j-\frac{1}{2},k}(\Delta q)} - a^-_{j-\frac{1}{2},k}\right]}_{\geq 0} \geq 0. \end{aligned} \tag{4.79}$$

The coefficients of $(\Delta q)^E_{j,k}$ and $(\Delta q)^W_{j,k}$ are also non-negative due to the CFL assumption (4.70). By the assumption (4.70), we have

$$\lambda^n a^+_{j+\frac{1}{2},k} \leq \frac{1}{8} \quad \text{and} \quad -\frac{1}{8} \leq \lambda^n a^-_{j+\frac{1}{2},k}. \tag{4.80}$$

The above inequalities imply

$$\begin{aligned} & \lambda^n a^+_{j+\frac{1}{2},k} - \lambda^n a^-_{j+\frac{1}{2},k} \leq \frac{1}{4} \\ \Rightarrow \quad & 4 \leq \frac{1}{\lambda^n a^+_{j+\frac{1}{2},k} - \lambda^n a^-_{j+\frac{1}{2},k}} \\ \Rightarrow \quad & -\frac{1}{2} \leq \frac{\lambda^n a^-_{j+\frac{1}{2},k}}{\lambda^n a^+_{j+\frac{1}{2},k} - \lambda^n a^-_{j+\frac{1}{2},k}} \\ \Rightarrow \quad & \frac{-\lambda^n a^-_{j+\frac{1}{2},k}}{\lambda^n a^+_{j+\frac{1}{2},k} - \lambda^n a^-_{j+\frac{1}{2},k}} \leq \frac{1}{2} \end{aligned} \tag{4.81}$$

Hence, the coefficient of $(\Delta q)^E_{j,k}$ is non-negative because of

$$\frac{1}{4} + \underbrace{\frac{(-a^-_{j+\frac{1}{2},k})}{a^+_{j+\frac{1}{2},k} - a^-_{j+\frac{1}{2},k}}}_{\leq \frac{1}{2}} \left(\underbrace{\lambda^n \frac{\Delta^x_{j+\frac{1}{2},k}F}{\Delta^x_{j+\frac{1}{2},k}(\Delta q)}}_{\geq -\frac{1}{8}} \underbrace{-\lambda^n a^+_{j+\frac{1}{2},k}}_{\geq -\frac{1}{8}} \right) \underbrace{-\lambda^n \frac{\Delta^x_{j,k}F}{\Delta^x_{j,k}(\Delta q)}}_{\geq -\frac{1}{8}} \geq 0. \tag{4.82}$$

By the same way, the coefficient of $(\Delta q)^W_{j,k}$ can be proved to be non-negative, and the other four coefficients as well. Since all the coefficients are non-negative and the sum of the coefficients are equal to 1, the combination on the right-hand-side of (4.77) is a convex combination. Hence,

$$(\Delta q)^{n+1}_{j,k} \leq \max\Big((\Delta q)^{W}_{j+1,k}, (\Delta q)^{E(W)}_{j,k}, (\Delta q)^{E}_{j-1,k}, (\Delta q)^{S}_{j,k+1}, (\Delta q)^{N(S)}_{j,k}, (\Delta q)^{N}_{j,k-1}\Big).$$

Because of the definition of the intermediate values $(\Delta q)^{E,W,S,N}$ and the choice of the derivative in (4.62), these intermediate values satisfy the local maximum principle, (consult Theorem 1 of [11]),

$$\max\Big((\Delta q)^{W}_{j+1,k}, (\Delta q)^{E(W)}_{j,k}, (\Delta q)^{E}_{j-1,k}, (\Delta q)^{S}_{j,k+1}, (\Delta q)^{N(S)}_{j,k}, (\Delta q)^{N}_{j,k-1}\Big) \leq \max_{j,k}((\Delta q)^{n}_{j,k}).$$

Then the maximum principle: $\max_{j,k}(\Delta q)^{n+1}_{j,k} \leq \max_{j,k}(\Delta q)^{n}_{j,k}$ holds. ■

With the help of Corollary 5.1 and Corollary 5.2 in [3], we can prove the semi-discrete (4.67) satisfies maximum principle.

Notice that, since the maximum function is nonlinear, we can not prove the maximum principle also holds for $q(= \Delta q + \tilde{q})$ directly from the above theorem. It needs some extraordinary approaches.

Chapter 5

Numerical Results

In this chapter, we present the results of a number of numerical experiments in the Euler system with gravitational source term by employing our designed well-balanced schemes. In all the tests, the parameter of the $(MC-\theta)$ limiter is set as $\theta = 1.5$. With the help of the CFL condition, the computation of the time-step is defined by

$$\begin{aligned} &\Delta t = \text{CFL} \cdot \frac{\Delta x}{\max(|\lambda_k|)}, \qquad \text{and} \\ &\Delta t = \text{CFL} \cdot \min\{\frac{\Delta x}{\max\limits_{j,k}(a^+_{j+\frac{1}{2},k}, -a^-_{j+\frac{1}{2},k})}, \frac{\Delta y}{\max\limits_{j,k}(b^+_{j,k+\frac{1}{2}}, -b^-_{j,k+\frac{1}{2}})}\} \end{aligned} \tag{5.1}$$

for one-dimensional examples in section 5.1 and two-dimensional examples in section 5.2, respectively. Here, λ_k denote the eigenvalues of the flux jacobian $\frac{\partial f(q)}{\partial q}$ and the CFL number is taken to be 0.485 in 1D experiments and 0.45 in 2D experiments.

In section 5.1.2, 5.1.4, 5.2.2 and section 5.2.4, we compare our solutions with the solutions from [4]. The scheme described in [4] is the combination of the NT scheme and the Deviation method.

5.1 1D Model

In this section, we test our one-dimensional scheme derived in section 3.1 and consider the 1D Euler system with gravitational source term as follows,

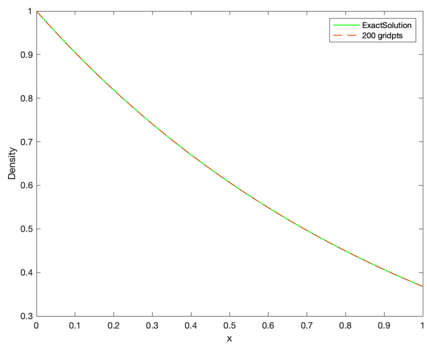
$$\begin{cases} q_t + f(q)_x = S(q, x), \quad x \in \Omega \subset \mathbb{R}, t > 0 \\ q(x, 0) = q_0(x), \end{cases} \tag{5.2}$$

where

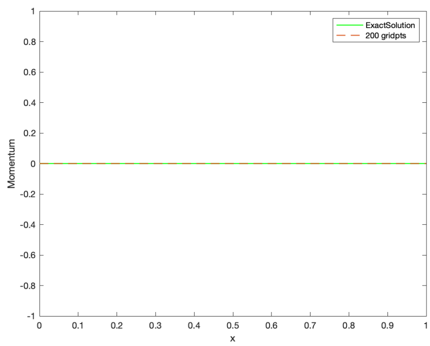
$$q = \begin{pmatrix} \rho \\ \rho u \\ E \end{pmatrix}, \quad f(q) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ (E+p)u \end{pmatrix}, \quad S(u) = \begin{pmatrix} 0 \\ -\rho\phi_x \\ -\rho u\phi_x \end{pmatrix}. \tag{5.3}$$

The given function $\phi = \phi(x)$ is the gravitational field and the ration of specific heats γ is suggested to be 1.4 for an ideal gas.

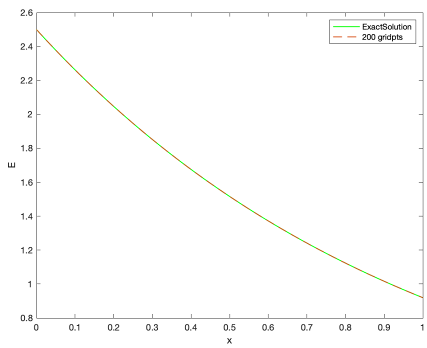
5.1.1 Isothermal Equilibrium



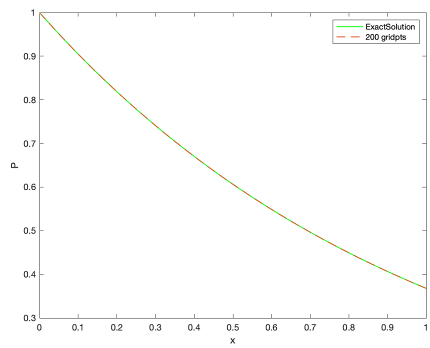
(a) Density



(b) Momentum



(c) Energy



(d) Pressure

Figure 5.1: Results of 1D isothermal equilibrium

In the first experiment, we consider the steady isothermal state with a linear gravitational field $\phi_x = g = 1$ (see [10]). The isothermal equilibrium state is given by

$$
\begin{aligned}
\rho(x) &= \rho_0 \exp(-\frac{\rho_0 g}{p_0} x), \\
u(x) &= 0, \\
p(x) &= p_0 \exp(-\frac{\rho_0 g}{p_0} x).
\end{aligned}
\tag{5.4}
$$

We set $\rho_0 = 1$ and $p_0 = 1$. The chosen stationary solution $\tilde{q}$ is the isothermal equilibrium state. The solution is computed on the 200 grid points of the interval $[0, 1]$ with outflow boundary condition until the final time $t = 0.25$ and compared with the exact solution. The figure 5.1 depicts the results of the density, momentum, energy, and pressure.

5.1.2 Isothermal Equilibrium with Perturbation

The second experiment is related to the first one. We add a small perturbation to the initial pressure, which is given by

$$
p(x) = p_0 \exp(-\frac{\rho_0 g}{p_0} x) + \eta \exp(-100 \frac{\rho_0 g}{p_0} (x - 0.5)^2), \tag{5.5}
$$

with $\eta = 0.001$. The figure 5.2 shows the comparison of the initial perturbation at $t = 0$, final perturbation at $t = 0.25$ on 200 grid points from our scheme, and the final perturbation at $t = 0.25$ on 200 grid points from the compared solution in [4].

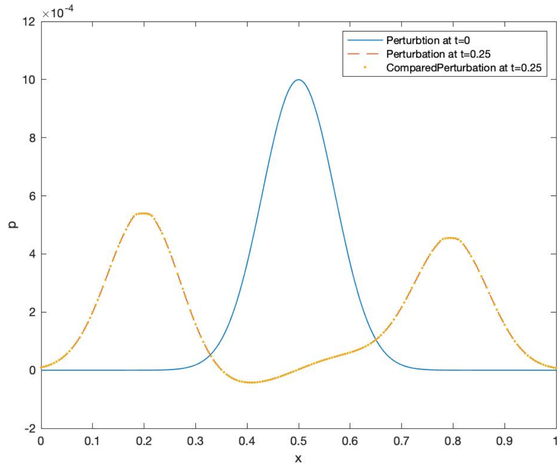


Figure 5.2: Initial perturbation at t=0 compared to the perturbation at the final time t=0.25 and the perturbation from compared solution.

We use L_1-norm to compute the loss of the density, pressure and the energy, and report the losses and the convergence rates in table 5.1. Here, we set the solution obtained from our scheme on 25600 grid points as the reference solution.

N	ρ L_1-error	rate	p L_1-error	rate	E L_1-error	rate
200	3.3030E-06	-	4.4358E-06	-	1.1091E-05	-
400	1.4317E-06	1.21	1.9702E-06	1.17	4.9260E-0.6	1.17
800	5.2586E-07	1.44	7.3033E-07	1.43	1.8260E-06	1.43
1600	8.4609E-08	2.64	1.1739E-07	2.64	2.9351E-07	2.64

Table 5.1: 1D isothermal equilibrium with perturbation: L_1-errors and convergence rates.

5.1.3 Moving Equilibrium

In this experiment, we are interested in preserving the following moving equilibrium state with a nonlinear gravitational field $\phi(x) = \exp(-\exp(x)+\gamma(\exp(-\gamma x)))$,

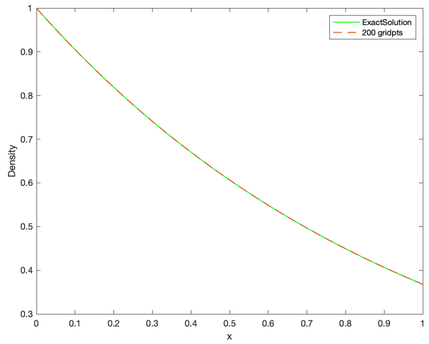
$$\begin{aligned}
\rho(x) &= \rho_0 \exp(-\frac{\rho_0 g}{p_0}x), \\
u(x) &= \exp(x), \\
p(x) &= \exp(-\frac{\rho_0 g}{p_0}x)^{\gamma}.
\end{aligned} \tag{5.6}$$

with $\rho_0 = 1$ and $p_0 = 1$. And we set the initial condition by

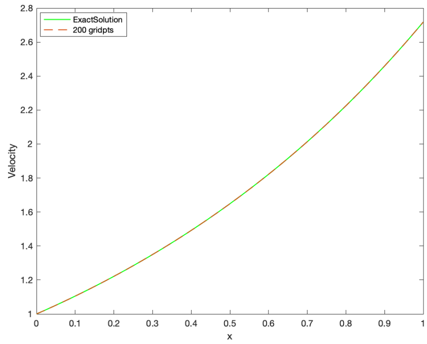
$$\begin{aligned}
\rho(x) &= \exp(-x), \\
u(x) &= \exp(x), \\
p(x) &= \exp(-\gamma x).
\end{aligned} \tag{5.7}$$

The detailed introduction of the initial condition can be found in [12]. The stationary solution used here is the equilibrium state itself and the chosen boundary condition is same as the one in the first experiment.

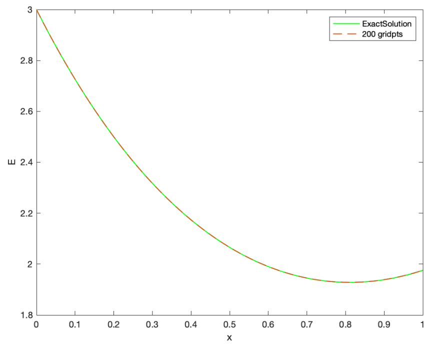
We compute the solution in the interval $[0, 1]$ on 200 grid points until the final time $t = 10$ and compare it to the exact solution. The obtained result is shown in figure 5.3.



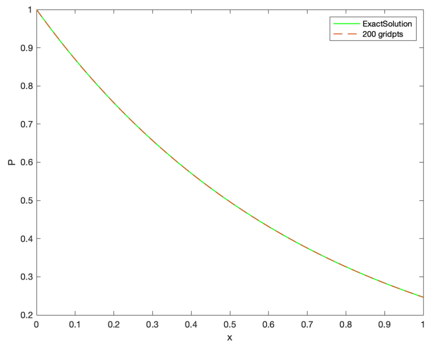
(a) Density



(b) Velocity



(c) Energy



(d) Pressure

Figure 5.3: Results of 1D moving equilibrium.

5.1.4 The Shock Tube Problem

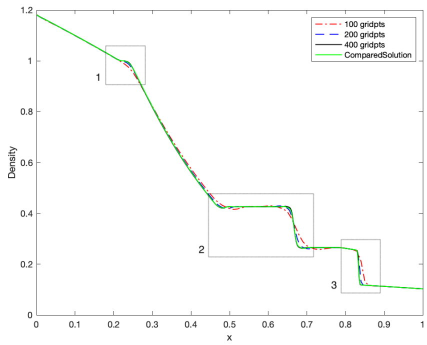
In this example, we consider the shock tube problem with the definition from [10], and compare with the solution from [4]. The initial state of the shock tube problem is given by

$$\begin{aligned} \rho(x) &= \begin{cases} 1, & \text{if } x \le 0.5, \\ 0.125, & \text{otherwise}, \end{cases} \\ u(x) &= 0, \\ p(x) &= \begin{cases} 1, & \text{if } x \le 0.5, \\ 0.1, & \text{otherwise}. \end{cases} \end{aligned} \tag{5.8}$$

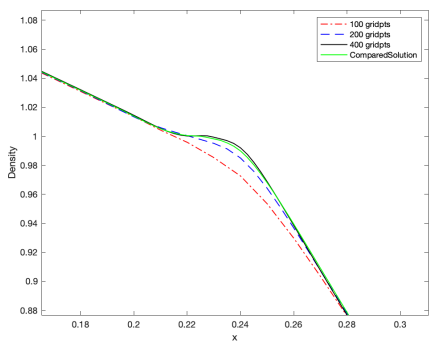
and the gravitational field is defined with $\phi_x = g = 1$. The stationary solution $\tilde{q}$ considered in this case is the isothermal equilibrium.

We use the computational interval $[0, 1]$ with the reflecting boundary condition and compute the solution on 100, 200, 400 grid points until the final time $t = 0.2$.

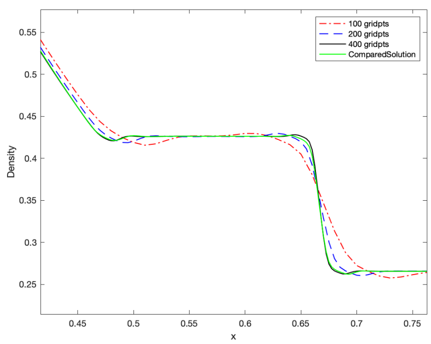
The compared solution is obtained from [4] and computed on 400 grid points. In figure 5.4, we show the the results of the density and zoom in on the shocks. In figure 5.5, we show the results of velocity, energy and pressure.



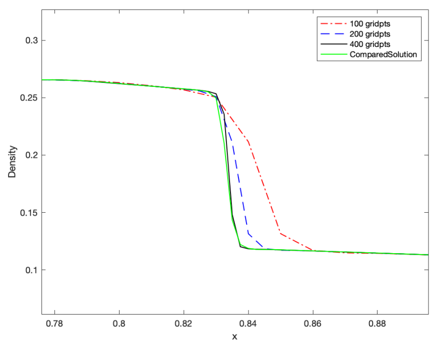
(a) Density



(b) Zoom in on the first block

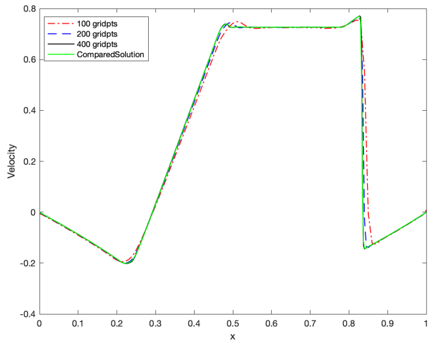


(c) Zoom in on the second block

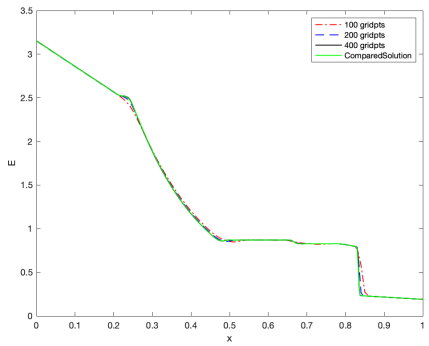


(d) Zoom in on the third block

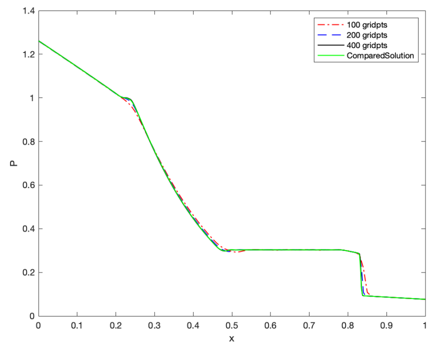
Figure 5.4: Results of 1D shock tube problem.



(a) Velocity



(b) Energy



(c) Pressure

Figure 5.5: Results of 1D shock tube problem: velocity, energy and pressure.

5.2 2D Model

In this section, we test our 2-dimensional scheme derived in section 4.1 and consider the 2D Euler system with the gravitational source term as follows,

$$\begin{cases} q_t + f(q)_x + g(q)_y = S(q,x), & x \in \Omega \subset \mathbb{R}, t > 0 \\ q(x,y,0) = q_0(x,y), \end{cases} \tag{5.9}$$

where

$$q = \begin{pmatrix} \rho \\ \rho u_1 \\ \rho u_2 \\ E \end{pmatrix}, \quad f(q) = \begin{pmatrix} \rho u_1 \\ \rho u_1^2 + p \\ \rho u_1 u_2 \\ (E+p)u_1 \end{pmatrix}, \quad g(q) = \begin{pmatrix} \rho u_2 \\ \rho u_1 u_2 \\ \rho u_2^2 + p \\ (E+p)u_2 \end{pmatrix},$$

and

$$S(u) = \begin{pmatrix} 0 \\ -\rho\phi_x \\ -\rho\phi_y \\ -\rho u_1\phi_x - \rho u_2\phi_y \end{pmatrix}.$$

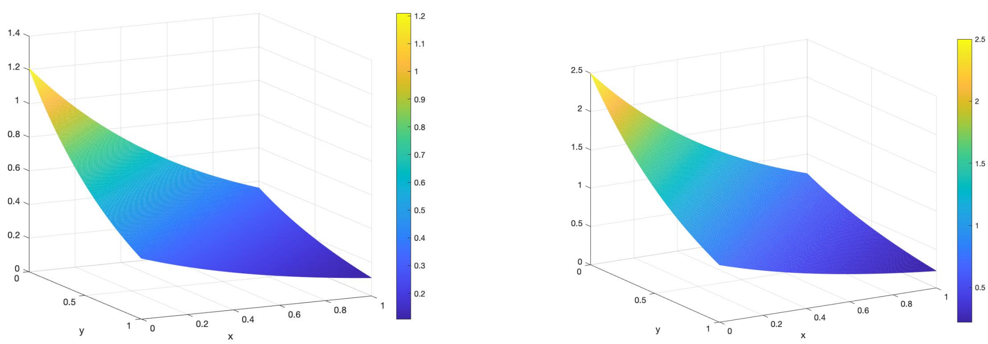
The given function $\phi = \phi(x, y)$ is the gravitational field and the ration of specific heats γ is suggested to be 1.4 for an ideal gas. The following tests are extended from the the 1D experiments , thus we adopt the same stationary solutions $\tilde{q}$ for the corresponding experiments. For the boundary conditions, we also use the same settings as in 1D.

5.2.1 Isothermal Equilibrium

The 2D isothermal equilibrium state is given by

$$\begin{aligned} \rho(x, y) &= \rho_0 \exp(-\frac{\rho_0}{p_0}(\phi_x x + \phi_y y)), \\ u_1(x, y) &= 0, \\ u_2(x, y) &= 0, \\ p(x, y) &= p_0 \exp(-\frac{\rho_0}{p_0}(\phi_x x + \phi_y y)). \end{aligned} \tag{5.10}$$

Here, we set $\rho_0 = 1.21$ and $p_0 = 1$. The linear gravitational potential is given by $\phi_x = 1$ and $\phi_y = 1$. The solution is computed on the 200×200 grid points of the interval $[0, 1]^2$ until the final time $t = 0.25$ and presented in figure 5.6.



(a) Density (b) Energy

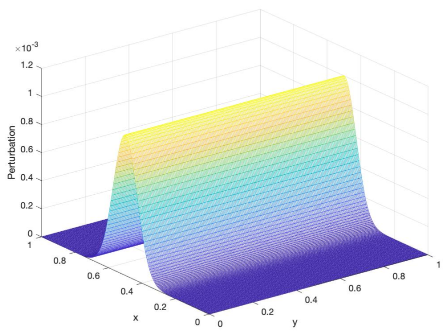
Figure 5.6: Results of 2D isothermal equilibrium

5.2.2 Isothermal Equilibrium with Perturbation

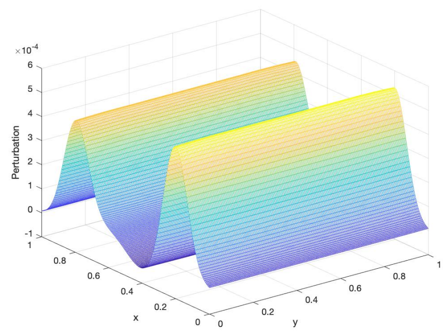
Similar to the 1D case, this example is extended from the previous experiment that we add a small perturbation along $x-$ or y axis to the initial pressure. For the case along x axis, the initial state is set as

$$\begin{aligned}
\rho(x,y) &= \exp(-x), \\
u_1(x,y) &= 0, \\
u_2(x,y) &= 0, \\
p(x,y) &= \exp(-x) + \eta \exp(-100(x-0.5)^2).
\end{aligned} \tag{5.11}$$

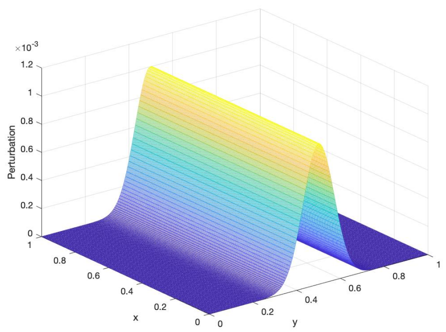
For the case along y axis, the initial data is defined similarly. The figure 5.7 shows the perturbation along x- and y axis respectively on 200×200 grid points at the initial time $t = 0$ and the final time $t = 0.25$. The comparison of cross section of perturbation between the initial time and the final time is presented in figure 5.8.



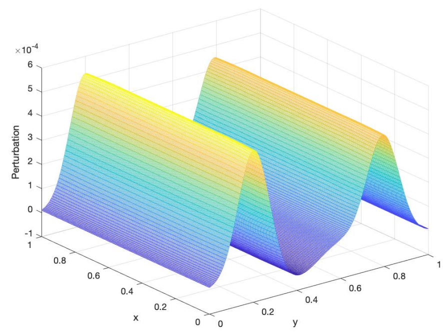
(a) Initially along x



(b) at t=0.25 along x

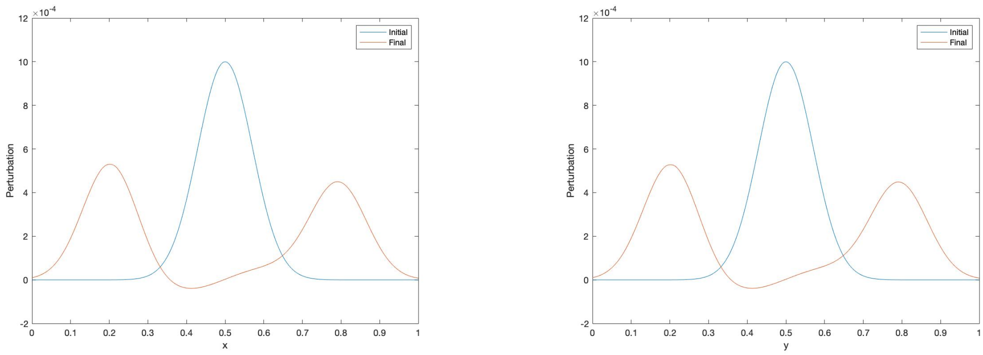


(c) Initially along y



(d) at t=0.25 along y

Figure 5.7: Results of 2D unidirectional equilibrium perturbation.



(a) along x-axis (b) along y-axis

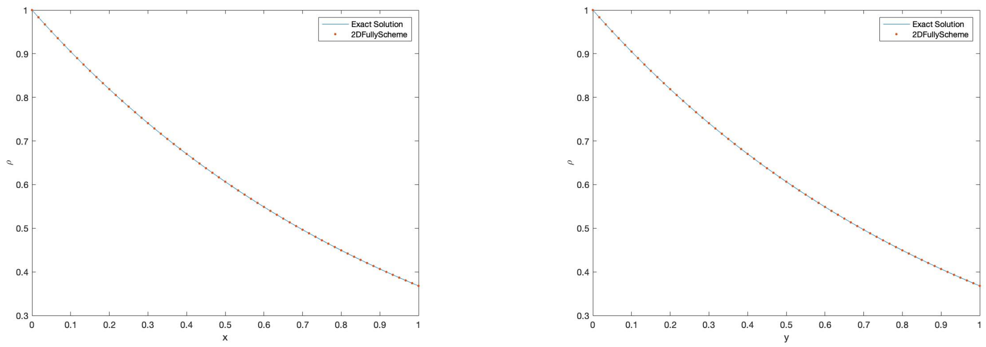
Figure 5.8: Results of 2D unidirectional equilibrium perturbation:Initial perturbation at t=0 and the later perturbation at the final time t=0.25.

5.2.3 Moving Equilibrium

The initial state of this experiment is defined by

$$
\begin{aligned}
\rho(x,y) &= \rho_0 \exp(-\frac{\rho_0 g}{p_0}(x+y)),\\
u_1(x,y) &= \exp(x+y),\\
u_2(x,y) &= \exp(x+y),\\
p(x,y) &= \exp(-\frac{\rho_0 g}{p_0}(x+y))^{\gamma}.
\end{aligned}
\tag{5.12}
$$

with $\rho_0 = 1$, $p_0 = 1$, and $g = 1$. The considered nonlinear gravitational potential is $\phi(x,y) = \exp(x+y)(-\exp(x+y) + \gamma(\exp(-\gamma(x+y))))$.
We consider the experiments along x- and y axis separately as the previous one. We compute the solution on 60×10 grid points along x axis and on 10×60 grid points along y axis until the final time $t = 0.25$. The figure 5.9 shows the results of cross section along $x-$ and along y axis compared to the exact solution.



(a) along x-axis (b) along y-axis

Figure 5.9: Results of 2D moving equilibrium at the final time t=0.25 compared to the exact solution.

5.2.4 Shock Tube Problem

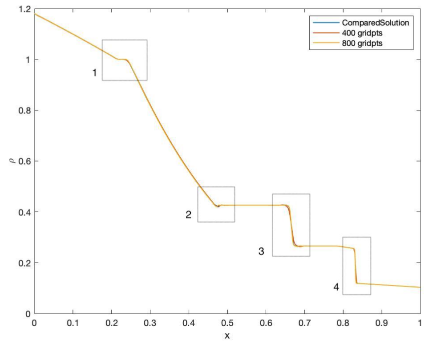
In a similar way with 1D case, the initial data of the shock tube problem along x axis is given by

$$
\begin{aligned}
\rho(x,y) &= \begin{cases} 1, & \text{if } x \le 0.5, \\ 0.125, & \text{otherwise,} \end{cases} \\
u_1(x,y) &= 0 = u_2(x,y), \\
p(x,y) &= \begin{cases} 1, & \text{if } x \le 0.5, \\ 0.1, & \text{otherwise.} \end{cases}
\end{aligned} \tag{5.13}
$$

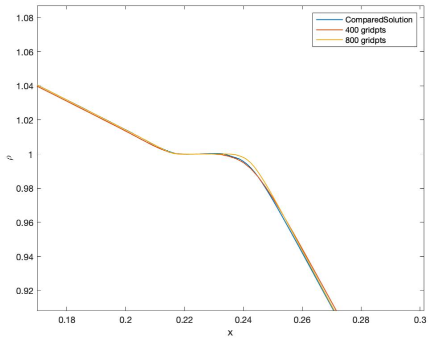
In figures 5.10 and 5.11, we show the solutions on 400×10 and 800×10 grid points until the final time $t = 0.2$, and the compared solution on 800×10 grid points from [4]. In figure 5.12, we compare the solution of the density on 200×10 grid points obtained from the scheme with the reconstruction (4.37) at the projection step to the solution on 200×10 grid points without reconstruction, i.e., we consider the reconstruction on nonsmooth subdomains as

$$\widetilde{W}^{n+1}_{D_{\alpha,\beta}} = \overline{w}^{n+1}_{D_{\alpha,\beta}}, \quad (x,y) \in D_{\alpha,\beta},$$

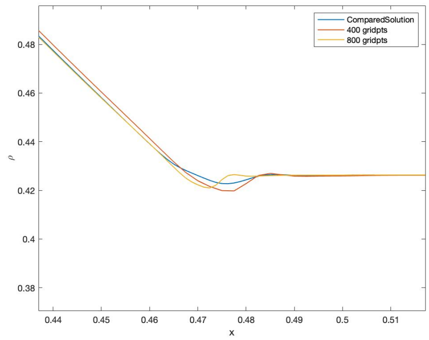
where $D_{\alpha,\beta} = \{D_{j+\frac{1}{2},k}, D_{j+\frac{1}{2},k+\frac{1}{2}}, D_{j,k+\frac{1}{2}}\}$.



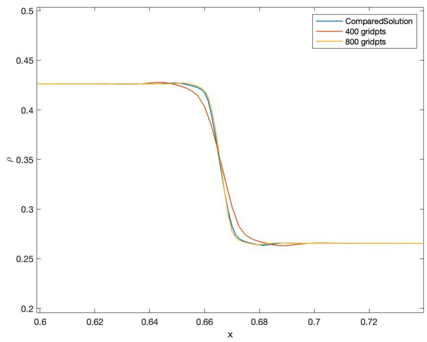
(a) Density



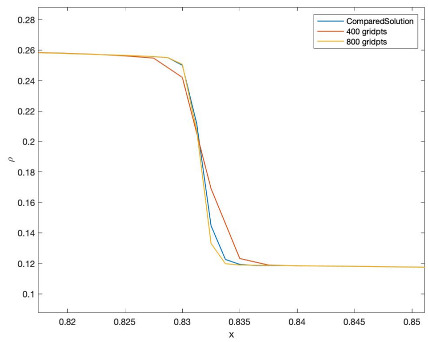
(b) Zoom in on the first block



(c) Zoom in on the second block

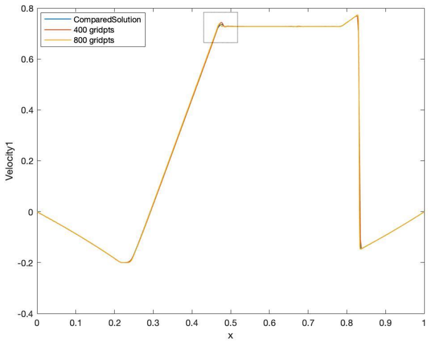


(d) Zoom in on the third block

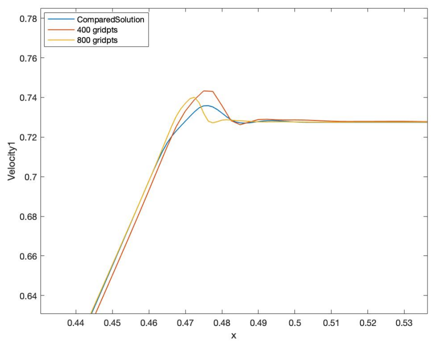


(e) Zoom in on the third block

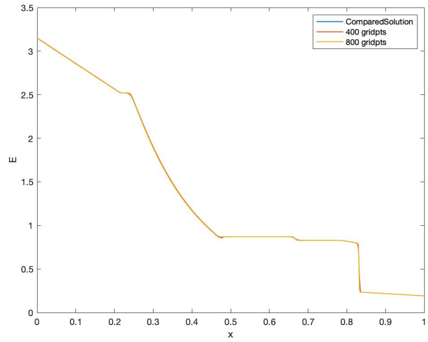
Figure 5.10: Results of 2D shock tube problem along x: density and zoom in on the shock.



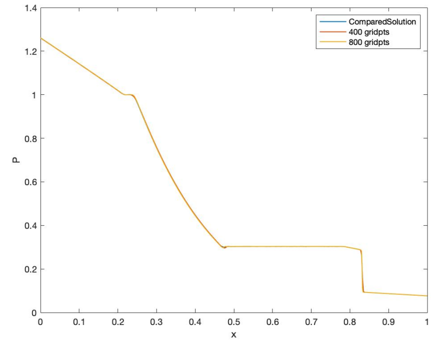
(a) velocity



(b) Zoom in on the block in the figure (a)

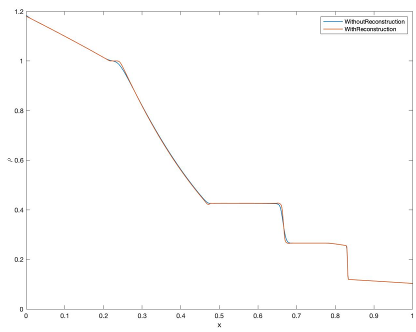


(c) Energy



(d) Pressure

Figure 5.11: Results of 2D shock tube problem along x: velocity, zoom in on the figure of velocity, energy and pressure.



(a) Density

Figure 5.12: Results of 2D shock tube problem along x: the solution without reconstruction and the solution with reconstruction.

Chapter 6

Conclusion and Outlook

In this thesis, we presented a brief introduction of the Euler equations and two numerical schemes: the central KT scheme and the Deviation Method. Based on these two schemes, we constructed a new second-order well-balanced scheme for one and two dimensions. Then we developed a semi-discrete scheme from the fully-discrete scheme. The semi-discrete scheme is potentially nonoscillatory in the sense of satisfying the TVD property in one-dimensional modified balance laws (3.7) and the maximum principle in two-dimensional modified balance laws (4.5). Since the total variation and maximum function are not linear functions, we can not guarantee the scheme remains nonoscillatory after the last step in which we add the computational solution Δq and the stationary solution $\tilde{q}$ to obtain the desired solution q.

The last point in the thesis is the numerical results of our new fully-discrete scheme applied to Euler equations with gravitational source term. The solutions obtained from our scheme look almost perfect, although in the shock tube problem there are some oscillations at the shocks when the number of grid points is not large enough. According to the last figure 5.12 in chapter 5, if we can find a more suitable reconstruction for the projection step, then it might be able to get better results without oscillation. In [16], the authors provided a different suggestion for the reconstruction at projection step, and it would be worth an attempt for further studies.

Besides, according to [3], this type of semi-discrete scheme has more advantages when applied to the convection-diffusion equations, in comparison with fully-discrete NT scheme. Applying our semi-discrete scheme to convection-diffusion equations might be a worthwhile challenge in the future.

Bibliography

[1] Eck, C., Garcke, H., & Knabner, P. (2017). Mathematical Modeling. In Springer Undergraduate Mathematics Series. Springer International Publishing.

[2] Lai, W. M., Rubin, D., & Krempl, E (1994). Introduction to Continuum Mechanics. Butterworth-Heinemann.

[3] Kurganov, A., & Tadmor, E. (2000). New High-Resolution Central Schemes for Nonlinear Conservation Laws and Convection–Diffusion Equations. In Journal of Computational Physics (Vol. 160, Issue 1, pp. 241–282).

[4] Kanbar, F., Touma, R., & Klingenberg, C. (2020). Well-balanced central schemes for the one and two-dimensional Euler systems with gravity. In Applied Numerical Mathematics (Vol. 156, pp. 608–626).

[5] Nessyahu, H., & Tadmor, E. (1990). Non-oscillatory central differencing for hyperbolic conservation laws. In Journal of Computational Physics (Vol. 87, Issue 2, pp. 408–463).

[6] Tadmor, E. (1988). Convenient Total Variation Diminishing Conditions for Nonlinear Difference Schemes. In SIAM Journal on Numerical Analysis (Vol. 25, No. 5, pp. 1002–1014.)

[7] Kurganov, A., Prugger, M., & Wu, T. (2017). Second-Order Fully Discrete Central-Upwind Scheme for Two-Dimensional Hyperbolic Systems of Conservation Laws. In SIAM Journal on Scientific Computing (Vol. 39, Issue 3, pp. A947–A965).

[8] Kurganov, A., & Petrova, G. (2001). A third-order semi-discrete genuinely multidimensional central scheme for hyperbolic conservation laws and related problems. In Numerische Mathematik (Vol. 88, Issue 4, pp. 683–729).

[9] Kurganov, A., Noelle, S., & Petrova, G. (2001). Semidiscrete Central-Upwind Schemes for Hyperbolic Conservation Laws and Hamilton–Jacobi Equations. In SIAM Journal on Scientific Computing (Vol. 23, Issue 3, pp. 707–740).

[10] Touma, R., Koley, U., & Klingenberg, C. (2016). Well-Balanced Unstaggered Central Schemes for the Euler Equations with Gravitation. In SIAM Journal on Scientific Computing (Vol. 38, Issue 5, pp. B773–B807).

[11] Jiang, G.-S., & Tadmor, E. (1998). Nonoscillatory Central Schemes for Multidimensional Hyperbolic Conservation Laws. In SIAM Journal on Scientific Computing (Vol. 19, Issue 6, pp. 1892–1917).

[12] Veiga, M. H. (2019). Capturing Near-Equilibrium Solutions: A Comparison between High-Order Discontinuous Galerkin Methods and Well-Balanced Schemes. In Communications in Computational Physics (Vol. 26, Issue 1, pp. 1–34).

[13] Irgens, F. (2008). Continuum Mechanics. Springer Berlin Heidelberg.

[14] Berberich, J. P., Chandrashekar, P., & Klingenberg, C. (2021). High order well-balanced finite volume methods for multi-dimensional systems of hyperbolic balance laws. In Computers & Fluids (Vol. 219).

[15] Berberich, J. P. (2020). Fluids in Gravitational Fields – Well-Balanced Modifications for Astrophysical Finite-Volume Codes. [Doctoral dissertation, University of Würzburg].

[16] Kurganov, A., & Lin, C.-T. (2007). On the Reduction of Numerical Dissipation in Central-Upwind Schemes. In Communications in Computational Physics (Vol. 2, No. 1, pp. 141-163).

[17] Berberich, J. P. (2017). A Well-Balanced Low Mach Numerical Scheme to Solve Euler Equations with Gravity for Astrophysical Applications. [Master's thesis, University of Würzburg].

[18] Chorin, A. J., & Marsden, J. E. (1993). A Mathematical Introduction to Fluid Mechanics. In Texts in Applied Mathematics. Springer New York.

Appendix

Algorithm

Algorithm 1 The main structure of the fully-discrete scheme in section 4.1

Input: number of grid points; final time; initial data; hydrostatic solution $\widetilde{q}$

Output: solution q

```
define size of original computation cell Δx, Δy
compute difference Δq between initial data and hydrostatic solution
add ghost cells to grids of Δq and q̃
determine wave speeds (Algorithm 2)
determine size of next time step Δt
set t = Δt
while t < final time do
    compute vertices of subdomains (See eq.(4.11) )
    compute center of mass (See Appendix A in [7])
    compute area of subdomains (See Appendix A in [7])
    compute minmod-limiter of Reconstruction (See Reconstruction step 4.1.1)
    compute intermediate values in Evolution step (Algorithm 3)
    compute the new cell average from intermediate values (See Projection step 4.1.3)
    determine wave speed for next time step (Algorithm 2)
    compute size of next time step Δt
    set t = t + Δt
    add ghost cells to grids of new average
end while
```

compute solution q by $\Delta q + \widetilde{q}$

Algorithm 2 Determine wave speeds.

compute the MC-θ limiter of q
compute values near interface (See eq.(4.10))
find the corresponding eigenvalues of above values
compare those eigenvalues to determine wave speeds

Algorithm 3 Evolution step 4.1.2.

compute average at t^n (See eq.(4.18))
compute numerical fluxes (See eq.(4.20), (4.23))
compute the average of source term (See eq.(4.29)-(4.31))