Numerical Methods for Hyperbolic Partial Differential Equations

Nguyen Thien Binh

Department of Mathematical Sciences
Graduate School of UNIST

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Numerical Methods for Hyperbolic Partial Differential Equations

A dissertation
submitted to the Graduate School of UNIST
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Nguyen Thien Binh

06.10.2015
Approved by

Jung

Advisor
Chang-Yeol Jung

ULSAN NATIONAL INSTITUTE OF
SCIENCE AND TECHNOLOGY

Numerical Methods for Hyperbolic Partial Differential Equations

Nguyen Thien Binh

This certifies that the dissertation of Nguyen Thien Binh is approved.

06.10.2015

Jung

Advisor: Chang-Yeol Jung

Pilwon Kim: Thesis Committee Member #1

Bongsuk Kwon

Bongsuk Kwon: Thesis Committee Member #2

Bongsoo Jang: Thesis Committee Member #3

Junseok Kim

Junseok Kim: Thesis Committee Member #4

I dedicate this dissertation to my parents

Abstract

In this dissertation, new numerical methods are proposed for different types of hyperbolic partial differential equations (PDEs). The objectives of these developments aim for the improvements in accuracy, robustness, efficiency, and reduction of the computational cost.

The dissertation consists of two parts. The first half discusses shock-capturing methods for nonlinear hyperbolic conservation laws, and proposes a new adaptive weighted essentially non-oscillatory WENO-θ scheme in the context of finite difference. Depending on the smoothness of the large stencils used in the reconstruction of the numerical flux, a parameter θ is set adaptively to switch the scheme between a $5th$-order upwind or $6th$-order central discretization. A new indicator τ^{θ} depending on parameter θ measures the smoothness of the large stencils in order to choose a smoother one for the reconstruction procedure. τ^{θ} is devised based on the possible highest-order variations of the reconstructing polynomials in an L^2 sense. In addition, a new set of smoothness indicators $\tilde{\beta}_k$'s of the sub-stencils is introduced. These are constructed in a central sense with respect to the Taylor expansions around point x_j. Numerical results show that the new scheme combines good properties of both $5th$-order upwind and $6th$-order central schemes. In particular, the new scheme captures discontinuities and resolves small-scaled structures much better than other $5th$-order schemes; overcomes the loss of resolution near some critical regions; and is able to maintain symmetry which are drawbacks detected in other $6th$-order central WENO schemes.

The second part extends the scope to hyperbolic PDEs with uncertainty, and semi-analytical methods using singular perturbation analysis for dispersive PDEs. For the former, a hybrid operator splitting method is developed for computation of the two-dimensional transverse magnetic Maxwell equations in media with multiple random interfaces. By projecting the solutions into random space using the Polynomial Chaos (PC) expansions, the deterministic and random parts of the solution are solved separately. The deterministic parts are then numerically approximated by the FDTD method with domain decomposition implemented on a staggered grid. Statistic quantities are obtained by the Monte Carlo sampling in the post-processing stage. Parallel computing is proposed for which the computational cost grows linearly with the number of random interfaces.

The last section deals with spectral methods for dispersive PDEs. The Korteweg-de Vries (KdV) equation is chosen as a prototype. By Fourier series, the PDE is transformed into a system of ODEs which is stiff, that is, there are rapid oscillatory modes for large wavenumbers. A new semi-analytical method is proposed to tackle the difficulty. The new method is based on the classical integrating factor (IF) and exponential time differencing (ETD) schemes. The idea is to approximate analytically the stiff parts by the so-called correctors and numerically the non-stiff parts by the IF and ETD methods. It turns out that rapid oscillations are well absorbed by our corrector method, yielding better accuracy in the numerical results. Due to the nonlinearity, all Fourier modes interact with each other, causing the computation of the correctors to be very costly. In order to overcome this, the correctors are recursively constructed to accurately capture the stiffness of the mode interactions.

Contents

List of Figures

List of Tables

1

Introduction

Hyperbolic partial differential equations are the governing equations of a wide range of problems which involve wave propagation or convective processes. These play important roles in many disciplines of engineering and science. Some of the most popular hyperbolic equations can be named, for example, the Euler equations in gas dynamics, conservation laws which have many applications in fluid dynamics, meteorology, astrophysics, or the magnetohydrodynamic (MHD) equations which take both electromagnetism and fluid dynamics into consideration. It is also worthy to mention the Maxwell equations, which are the fundamental equations of electrodynamics, to name but a few.

Despite the importance and various applications of hyperbolic equations, difficulties may arise as one seeks for their exact solutions, especially for the nonlinear problems. We take conservation laws for an instance. It is well-known that for these type of equations discontinuities, e.g., shocks or contacts, may develop in the solutions in time due to the collision of the characteristics when the solutions evolve in time. Hence the solutions break down in a classical sense since derivatives cannot be defined at these discontinuities. To remedy this, one can search for the solutions in a weaker sense, i.e., the integral form. But this approach may lead to other issues, e.g. uniqueness. In spite of all these difficulties, fundamental understandings on the behaviors of the solutions of hyperbolic problems have been established. This sheds light on an alternative approach: to approximate the solutions numerically. With leaping developments in computer sciences, this computational approach flourishes and is becoming a major tool in solving hyperbolic problems. More and more numerical methods have been developed and successfully implemented in real applications.

In terms of computational perspective, there are always questions of how one develops numerical schemes which are more accurate, stable, yet simpler in implementation. Besides, computational cost is also under a great deal of consideration. These questions are the main purpose of my research, which is reflected in this thesis.

The objectives of this thesis are to propose new numerical methods for some important partial differential equations which are of hyperbolic type, with the emphasis on the improvements to the accuracy, robustness, and the reduction of computational cost. The scope of the thesis is limited to the following particular problems: shock-capturing numerical schemes for nonlinear hyperbolic conservation laws in which a new 6th-order WENO-θ scheme is developed, numerical investigation of the evolution of waves in time for the 2D transverse Maxwell equations in random media resulting from the randomness in the parameters of the equations, and semi-analytical methods for stiff problems with an application to the dispersive Korteweg de Vries (KdV) equation by spectral methods.

The thesis consists of two parts. The first half discusses shock-capturing methods for conservation laws. For these problems, shocks may occur in the solutions when they evolve in time due to the facilitation of weak solutions. A good numerical scheme in this case must be able to capture sharply discontinuities or regions with high gradients in the solutions, yet prevents spurious oscillations from happening. Furthermore, since weak solutions are not unique, the numerical approximations must converge to the solutions which are physically meaningful. Shock-capturing methods are of this type. The term shock-capturing implies that these methods do not require the information on the exact locations or structures of the shocks in resolving the shocks, hence they are feasible for high-dimensional problems or the ones with complicated domains or structures. This first part is discussed in three chapters. In Chapter 2, we briefly review important properties of conservation laws which are often served as the analyzing tools in numerical analysis. We then review the literature on shock-capturing schemes, mainly on the scheme methodologies. The focus in this part is presented in Chapter 3 where WENO schemes are discussed. Here, we propose a new WENO-θ scheme which improves the accuracy order to sixth in smooth regions, using the same reconstruction stencils as those of other 5th-order upwind WENO schemes, e.g., WENO-JS or WENO-Z for a general case. We also show that the new scheme gives considerably more accurate and better resolution than other comparing 6th-order WENO methods. In Chapter 4, a variety of numerical tests are simulated to illustrate for the outperformance of the new scheme.

In the second part of the thesis, we extend the scope to discuss numerical schemes for other kinds of hyperbolic partial differential equations. In particular, the stochastic 2D Maxwell equations in electromagnetism and the KdV equation are approximated using our new proposed methods. Although categorized as hyperbolic type, these equations are different from conservation laws in the first part in terms of properties and applications, thus require different numerical techniques. The Maxwell equations are of 2nd-order in both space and time with the occurrence of uncertainties in the parameters; whereas for the KdV equation, there occurs a dispersive term in terms of a spatial derivative of degree three, which introduces a system of complex ODEs when spectral method is applied, in which the waves do not decay but oscillate rapidly as the wavenumber is large. For these problems, we will introduce new numerical schemes which reduce the computational costs as well as considerately increase the accuracy of the numerical solutions.

In Chapter 5, we study the cumulative distribution functions (CDFs) of the solutions of the 2D transverse Maxwell equations with uncertainties resulting from the presence of fluctuating multiple random interfaces. The CDFs evolve in time due to the randomness in the interfaces of the media. To handle the stochastic part, we represent the solutions in terms of polynomial chaos (PC) expansions and seek for the approximations of the PC modes, which are deterministic. For this, an FDTD scheme is implemented on a staggered grid with domain decomposition. The statistical quantities are then obtained in a post-processing stage using the Monte Carlo method. We suggest a new technique in computing the PC modes, which much reduces the computational cost from exponentially to linearly with respect to the dimensionality, i.e., the number of random parameters.

Chapter 6 deals with stiff problems, in particular, the dispersive KdV equation. By spectral methods, we can transform the PDE into a system of ODEs in which the unknowns are the Fourier coefficients. Unfortunately, this system is usually stiff due to the wide range of the wavenumbers. By singular perturbation analysis, we can derive the so-called correctors which are analytical approximations of the stiff parts in the solutions. Incorporating these correctors into the numerical methods for the non-stiff parts, we can obtain the approximations which are much more accurate than those obtained from conventional methods. Since a PDE can be transformed into a system of ODEs by spectral methods, we first discuss the ODE cases for the setting up of the numerical schemes and techniques. The KdV equation is then treated by spectral methods. The main issue for the latter stage is the interaction of all Fourier modes due to the nonlinear convective term, which needs the incorporation of all modes in computing the correctors, thus is very expensive. We handle this by proposing a cut-off approximating method in a recursive manner.

Conclusion and discussions on relevant future works are given in Chapter 7.

Part I

Shock-capturing Methods for Hyperbolic Conservation Laws

2

Overview of Theory and Numerical Methods for Nonlinear Conservation Laws

In this chapter, we briefly discuss theoretical aspects and numerical approaches for hyperbolic nonlinear conservation laws. We first have a review on important properties in the behavior of weak solutions. For complete discussions, we refer to good textbooks and lectures notes on conservation laws, for example, [127], [91], [11], [137], [96], to name but a few. Based on this analytic ground, we proceed to the numerical approaches. Here, the main concepts including conservative methods, convergence, stability in the sense of total variation, and the discrete entropy condition will be discussed. Finally, we present different numerical schemes in relation with the above mentioned concepts and briefly justify them. We organize the schemes following the flow of improvement in terms of accuracy and resolution by further relaxing the condition of the TV stability criterion, which plays roles as a mechanism in suppressing oscillations near discontinuities in the numerical approximations.

For the purpose of coherence, we mainly state results with a limitation of proofs, except for important ones. Instead, we will cite where interested readers can find these proofs. In addition, for simplicity, we mostly deal with the scalar case only when discussing the numerical schemes. In fact, this is the common approach for developing a numerical scheme for conservation laws. One first goes with the scalar case, and proceeds to check if the scheme works well in case of systems.

2.1 Theory of Nonlinear Hyperbolic Conservation Laws

Consider the following conservation law in 1D

$$\begin{cases} u_t + f(u)_x = 0, \quad -\infty < x < \infty,\ t > 0, \\ u(x,0) = u_0(x), \end{cases} \tag{2.1.1}$$

where $u(x,t)$ is some conserved quantity, $f(u)$ is called the flux function. Hereafter, we consider the case that f is convex, i.e., $f''(u) \geq 0$. We also assume that the initial condition $u_0(x) \in (L^\infty \cap L^1)((-\infty,\infty) \times (0,\infty))$ and has bounded total variation. The latter concept will be defined elsewhere below.

Eq. (2.1.1) can be written in a non-conservative form

$$\begin{cases} u_t + f'(u)u_x = 0, \\ u(x,0) = u_0(x). \end{cases} \tag{2.1.2}$$

The differential form (2.1.1) is equivalent to the following integral form. Let $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ be a control volume central at x_j with size $\Delta x = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$, and $t_{n+1} > t_n \geq 0$, we have that

$$\frac{d}{dt}\left(\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x,t)dx\right) = -\frac{f(u(x_{j+\frac{1}{2}},t)) - f(u(x_{j-\frac{1}{2}},t))}{\Delta x}. \tag{2.1.3}$$

Integrating Eq. (2.1.3) with respect to time from t_n to t_{n+1}, and multiplying with $\frac{1}{\Delta t}$ where $\Delta t = t_{n+1} - t_n$, we obtain that

$$\begin{aligned} &\frac{1}{\Delta t}\left[\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x,t_{n+1})dx - \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x,t_n)dx\right] = \\ &\quad -\frac{1}{\Delta x}\left[\frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}},t))dt - \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{j-\frac{1}{2}},t))dt\right], \end{aligned} \tag{2.1.4}$$

which states that the change in the total u over the control volume in time is equal to the difference of the incoming and outcoming of the flux through the interfaces $x_{j\pm\frac{1}{2}}$ over the time period. These forms are the fundamental equations for finite volume methods, which are discussed in the below sections.

2.1.1 Characteristics Method

Supposing that there is a time $t^* > 0$ such that u is a solution of Eq. (2.1.1) in a classical sense, i.e., u satisfies (2.1.1) pointwise, for all $t \in (0, t^*)$. We want to find a family of curves along which the solution is conserved, that is,

$$\left\{\Upsilon : x = x(s) \middle| \frac{du(x(s),s)}{ds} = 0\right\}. \tag{2.1.5}$$

For such curves exist, they are called characteristics of u. Since u is classical for $t \in (0, t^*)$, choosing $s \equiv t$ we have that

$$\frac{du(x(t),t)}{dt} = u_t + x'(t)u_x = u_t + f'(u)u_x = 0, \tag{2.1.6}$$

thus the characteristic curves are

$$\begin{cases} x'(t) = f'(u), \\ x(0) = x_0, \end{cases} \tag{2.1.7}$$

which gives

$$x(t) = x_0 + f'(u)t. \tag{2.1.8}$$

Hence, the solution, if exists, has the following form

$$u(x,t) = u_0(x - f'(u)t). \tag{2.1.9}$$

The classical solution implies that its characteristics do not collide, for $t \in (0, t^*)$. The threshold time t^* at which collision occurs can be determined by applying the Implicit function theorem (see, e.g., [26]) to Eq. (2.1.9).

Unfortunately, there are cases in which there does not exist any $t^* > 0$. In other words, the characteristics cross out initially, for example, the shock case with Riemann initial data where $u_L > u_R$, u_L, u_R are constants. For these cases, classical solutions do not make sense. Hence, we seek for a weaker sense in which the solutions can be defined, which are called weak solutions.

2.1.2 Integral (Weak) Solutions and Their Properties

Multiplying Eq. (2.1.1) with a smooth and compactly supported test function $\phi(x,t)$, and integrating over the whole space, by integration by parts, we have that

$$\int_0^\infty \int_{-\infty}^\infty [u\phi_t + f(u)\phi_x] dxdt + \int_{-\infty}^\infty u_0(x)\phi(x,0)dx = 0. \tag{2.1.10}$$

Definition 2.1.1. (Weak solution) A solution $u \in L^\infty((-\infty,\infty)\times(0,\infty))$ is called a weak solution of the conservation law (2.1.1) if it satisfies the integral form (2.1.10) for any smooth and compactly supported test function ϕ.

We note that unlike the differential form (2.1.1) where u is at least $\mathcal{C}^1$ so that the equation makes sense, there is no such condition required for the weak solution, i.e., the one satisfying Eq. (2.1.10) since the differentiation has been switched to the smooth test function. Hence, it is expected that there are discontinuities present in a integral (weak) solution. Moreover, integral solutions are not unique. We next discuss the condition for which discontinuities must satisfy.

Lemma 2.1.1. *(Rankine-Hugoniot condition) ([26]) Let $\Omega \subset (-\infty, \infty) \times (0, \infty)$ be some open region on which u is $\mathcal{C}^1$ on either side Ω_L and Ω_R of a smooth curve C.*
Then, if u is a weak solution of Eq. (2.1.1), it satisfies the following jump condition,

$$[f] = s[u], \tag{2.1.11}$$

where $[g] = g_L - g_R$ denotes the jump of g over the curve C. s is called the propagation speed of the discontinuity.

The non-uniqueness of weak solutions is illustrated by the following example.

Example 2.1.1. Consider Burgers' equation supplemented with Riemann initial condition

$$\begin{cases} u_t + \left(\dfrac{u^2}{2}\right)_x = 0, \\ u_0(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{if } x > 0. \end{cases} \end{cases} \tag{2.1.12}$$

Here, the flux function $f(u) = \frac{u^2}{2}$ is convex.

Applying the Rankine-Hugoniot condition, we can obtain the following weak solution

$$u(x,t) = \begin{cases} 0, & \text{if } x/t < \frac{1}{2}, \\ 1, & \text{if } x/t > \frac{1}{2}. \end{cases} \tag{2.1.13}$$

However, Eq. (2.1.1) also accepts the following similarity solution

$$u(x,t) = \begin{cases} 0, & \text{if } x/t < 0, \\ x/t, & \text{if } 0 < x/t < 1, \\ 1, & \text{if } x/t > 1. \end{cases} \tag{2.1.14}$$

Both solutions satisfy Eq. (2.1.1) in a weak sense. In fact, it can be shown that the solution (2.1.13) is a non-physical one since it does not depend continuously on the initial data (characteristics span out from the shock). Hence, in order to judge a weak solution which is physically relevant, we need another criterion, i.e., the entropy condition.

Definition 2.1.2. (Entropy pair) A smooth function $U(u)$ is said to be an entropy function if it satisfies the following conditions,

i. U is a convex function of u, i.e., $U_{xx} > 0$,

ii. There is a function $F(u)$, which is called an entropy flux, such that

$$U_u f_u = F_u. \tag{2.1.15}$$

Definition 2.1.3. (Viscosity vanishing solution) A physically relevant weak solution, or entropy solution, is defined as the solution of the viscosity vanishing equation, i.e., as $\varepsilon \to 0$,

$$u_t + f(u)_x = \varepsilon u_{xx}, \quad \varepsilon > 0. \tag{2.1.16}$$

It can be shown that an entropy solution satisfies the following entropy condition. See, for examples, [55], [9].

Lemma 2.1.2. *(Entropy condition)*

i. Any smooth solution u of Eq. (2.1.1) satisfies

$$U(u)_t + F(u)_x = 0. \tag{2.1.17}$$

ii. The viscosity vanishing solution of Eq. (2.1.16) satisfies the following entropy condition,

$$U(u)_t + F(u)_x \leq 0, \tag{2.1.18}$$

in a weak sense, i.e., for any nonnegative smooth test function ϕ with compact support,

$$-\int_0^\infty \int_{-\infty}^\infty [U\phi_t + F\phi_x]dxdt - \int_{-\infty}^\infty U_0\phi(x,0)dx \leq 0, \tag{2.1.19}$$

where $U_0 = U(u_0(x))$.

Writing condition (2.1.19) for each interval $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, for all j's in the time slab

$[t_n, t_{n+1})$, we obtain that

$$\begin{aligned}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} U(u(x,t_{n+1}))dx \leq& \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} U(u(x,t_n))dx \\ &+ \int_{t_n}^{t_{n+1}} F(u(x_{j-\frac{1}{2}},t))dt - \int_{t_n}^{t_{n+1}} F(u(x_{j+\frac{1}{2}},t))dt.\end{aligned} \tag{2.1.20}$$

Condition (2.1.20) is often used as a discrete entropy condition for a numerical scheme. (See the Godonov method below.)

Remark 2.1.1.

i. An example of entropy pair for the scalar case is the Kružkov pair (see [83], [137])

$$U(u;c) = |u-c|, \quad F(u;c) = sgn(u-c)(f(u)-f(c)), \tag{2.1.21}$$

for any constant $c \in \mathbb{R}$. Here, $sgn(x) = \dfrac{x}{|x|}$ for $x \neq 0$, and $sgn(x) = 0$ for $x = 0$. Note that Kružkov's entropy pair are symmetric in both arguments.

ii. Hereafter, we only consider the conservation law (2.1.1) which possesses such an entropy pair defined in definition 2.1.2.

We now list some key properties of entropy weak solutions.

Proposition 2.1.1. *(see [24], [46], or [137]) Let $u^{1,2}(x,t)$'s be the entropy weak solutions of Eq. (2.1.1) corresponding to initial data $u_0^1(x)$ and $u_0^2(x)$, respectively, that satisfy the entropy conditions defined in lemma 2.1.2. Then, the following properties hold for all $t > 0$,*

1. *(L^1-contraction)*

$$\|u^2(\cdot,t) - u^1(\cdot,t)\|_{L^1(\mathbb{R})} \leq \|u_0^1 - u_0^2\|_{L^1(\mathbb{R})}. \tag{2.1.22}$$

2. *(Monotonicity preserving)*

$$u_0(x_2) \geq u_0(x_1) \quad \Rightarrow \quad u(x_2,t) \geq u(x_1,t). \tag{2.1.23}$$

3. *(TV-bound) for $t_2 \geq t_1$,*

$$TV(u(\cdot,t_2)) \leq TV(u(\cdot,t_1)), \tag{2.1.24}$$

where TV of a function is defined in Eq. (2.2.24) below.

Remark 2.1.2.

i. L^1-contractive property is often mimicked in a discrete sense for monotone schemes. In addition, it immediately implies the uniqueness of the entropy solution.

ii. Monotonicity preserving property implies that there are no new local maximum or minimum created in the solution of the conservation law, and that the maximal (minimal) values do not increase (decrease) when the solution evolves in time. These play a key role in designing the limiters to control oscillations in TVD methods discussed below.

2.1.3 Characteristic Decomposition

We now extend the scalar conservation law to a system case. Here, we discuss the linearized case only. In particular, we present the technique of characteristic decomposition. We note that this is sufficient in terms of numerical aspects. As mentioned above, a reasonable numerical approach for conservation laws is starting with the scalar case Eq. (2.1.1), and then proceed further to the system. For the latter, the nonlinearity is usually linearized around the cell interface $x_{j+\frac{1}{2}}$, e.g., the Roe averaging (see [118]). Then, the approximations are accomplished in a characteristic space through the characteristic decomposition. For complete discussions on systems of conservation laws, we refer to the monograph of Lax ([91]), the book of Smoller ([127]), or the lecture notes [137], [9].

We consider the system

$$\begin{cases} \mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0, \\ \mathbf{u}(x,t) = \mathbf{u}_0(x). \end{cases} \tag{2.1.25}$$

Here, $\mathbf{u}(x,t) : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}^m$, and $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^m$ is the flux function.

Linearizing system (2.1.25) about some constant state $\bar{\mathbf{u}}$, and writing in a non-conservative form, we obtain that

$$\begin{cases} \mathbf{u}_t + A(\bar{\mathbf{u}})\mathbf{u}_x = 0, \\ \mathbf{u}(x,t) = \mathbf{u}_0(x), \end{cases} \tag{2.1.26}$$

where $A : \mathbb{R}^m \to \mathbb{R}^m$ is the Jacobian matrix of the flux function $\mathbf{f}$ at the state $\bar{\mathbf{u}}$. The system (2.1.26) is called hyperbolic assuming that the eigenvalues λ_k's, $k = 1, \ldots, m$, are all real and finite, and strictly hyperbolic if these eigenvalues are all distinct; and the right eigenvectors $\mathbf{r}_k$'s constitute a complete set. We denote $R = [\mathbf{r}_k]$ and $L = [\mathbf{l}_k]$ the matrices of right column and left row eigenvectors of A, respectively. Here, after a standard normalization, we have that

$$\mathbf{l}_i \cdot \mathbf{r}_j = \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{otherwise.} \end{cases} \tag{2.1.27}$$

Thanks to the hyperbolicity, matrix A has a similarity transformation. That is,

$$LAR = \Lambda, \tag{2.1.28}$$

where $\Lambda = [\lambda_k]$ is the eigenvalue matrix. Note that Λ is diagonal.

Left multiplying Eq. (2.1.26) with L, with a note of Eq. (2.1.28) we obtain that

$$\begin{cases} \mathbf{w}_t + \Lambda \mathbf{w}_x = 0, \\ \mathbf{w}(x,t) = \mathbf{w}_0(x), \end{cases} \tag{2.1.29}$$

where $\mathbf{w} = L\mathbf{u}$ is called the characteristic variable and $\mathbf{w}_0 = L\mathbf{u}_0$. Notice that system (2.1.29) is now decoupled, and each characteristic component $w_k(x,t)$ can be solved exactly,

$$w_k(x,t) = w_{0k}(x - \lambda_k t) = \sum_{j=1}^{m} L_{kj} u_{0j}(x - \lambda_k t); \tag{2.1.30}$$

then the solution $\mathbf{u}$ can be retrieved by the relation

$$\mathbf{u}(x,t) = R\mathbf{w}. \tag{2.1.31}$$

We note that with the characteristic decomposition, simple waves are now decoupled. Hence there do not exist cases where these waves collide with each other, e.g. the collision of two shocks. See the characteristic projection for ENO and WENO schemes below.

We move on to the discussion of numerical aspects for nonlinear conservation laws.

2.2 Numerical Methods for Conservation Laws

Before proceeding, we set up the notations for later discussions.

2.2.1 Notations

Definition 2.2.1. (Grid discretizations)

i. Spatial grid: $\ldots < x_{j-1} < x_j < x_{j+1} < \ldots$ with $x_j = j\Delta x$, $j = \ldots, -1, 0, 1, \ldots$, where the grid size Δx is uniform for simplicity.

ii. Grid interval: $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ central at x_j with interfaces $x_{j\pm\frac{1}{2}} = x_j \pm \dfrac{\Delta x}{2}$.

iii. Temporal grid: $0 = t_0 < t_1 < \ldots < t_n < \ldots$ with $t_n = n\Delta t$.

iv. We denote the ratio $\sigma = \dfrac{\Delta t}{\Delta x} = \mathcal{O}(1)$ as $\Delta x \to 0$.

v. Approximating stencils: let p and q be some natural numbers. We denote $S_j = \{I_{j-p}, \ldots, I_{j+q}\}$ or $S_{j+\frac{1}{2}} = \{x_{j-p+\frac{1}{2}}, \ldots, x_{j+q+\frac{1}{2}}\}$ the $(p+q+1)$-point stencils over which a polynomial of degree $(p+q)$ is reconstructed or interpolated, respectively.

Definition 2.2.2. (Numerical approximations)

i. Pointwise approximation: $v_j^n \approx u(x_j, t_n)$.

ii. Average approximation: $\bar{v}_j^n \approx \bar{u}(x_j, t_n) = \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t_n)dx$.

iii. Collection of $\{\bar{v}_j^n\} = \bar{v}^n$.

iv. Time-dependent pointwise approximation: $v_j(t) \approx u(x_j, t)$.

v. Time-dependent average approximation: $\bar{v}_j(t) \approx \bar{u}(x_j, t)$.

vi. Approximating function over a stencil S_j: $v_{S_j}^n(x) \approx u(x, t_n)\chi_{S_j}(x)$. Here, $\chi_\Omega(y)$ is the standard indicator function,

$$\chi_\Omega(y) = \begin{cases} 1, & \text{if } y \in \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2.1}$$

vii. Approximating functions:

- At time t_n: $v_{\Delta x}^n(x) = \sum_j v_{S_j}^n(x)\chi_{S_j}(x)$. If $S_j \equiv I_j$ then $v_{S_j}^n(x) \equiv \bar{v}_j^n$, $v_{\Delta x}^n(x) = \bar{v}_{\Delta x}^n$.
- Over the whole domain: $v_{\Delta x}(x,t) = \sum_n \sum_j v_{S_j}^n(x)\chi_{S_j \times [t_n, t_{n+1})}(x,t)$. Similarly as above, we have $\bar{v}_{\Delta x}(x,t)$ if $S_j \equiv I_j$.

Definition 2.2.3. (Numerical flux) Let p, q be natural numbers. We denote the numerical flux function $\hat{f}_{j+\frac{1}{2}}$ as

$$\hat{f}_{j+\frac{1}{2}} = \hat{f}(\bar{v}_{j-p+1}^n, \ldots, \bar{v}_{j+q}^n; x_{j+\frac{1}{2}}) \approx \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(u(x_{j+\frac{1}{2}}, t))dt. \tag{2.2.2}$$

Notice that $\hat{f}_{j+\frac{1}{2}}$ consists of $(p+q)$ arguments.

2.2.2 Conservative Methods

We recall that a conservation law can be expressed either in conservative form Eq. (2.1.1) or non-conservative Eq. (2.1.2). It is advisable that one should not use the latter form for the discretization purpose due to the convergence to some wrong weak solution. The below example illustrates for this fact.

Example 2.2.1. Consider Burgers' equation written in a non-conservative form and supplemented with Riemann initial data,

$$\begin{cases} u_t + uu_x = 0, \\ u_0(x) = \begin{cases} 1, & \text{if } x < 0, \\ 0, & \text{if } x > 0, \end{cases} \end{cases} \tag{2.2.3}$$

whose exact solution is the propagation of the shock to the right at speed $s = \frac{1}{2}$, by the Rankine-Hugoniot condition (2.1.11), i.e.,

$$u(x,t) = \begin{cases} 1, & \text{if } x/t < \dfrac{1}{2}, \\ 0, & \text{if } x/t > \dfrac{1}{2}. \end{cases} \tag{2.2.4}$$

Approximating Eq. (2.2.3) by a first-order upwind method gives us,

$$v_j^{n+1} = v_j^n - \sigma v_j^n (v_j^n - v_{j-1}^n). \tag{2.2.5}$$

For $j < 0$, $v_j^n = v_{j-1}^n = 1$; for $j > 0$, $v_j^n = v_{j-1}^n = 0$; for $j = 0$, $v_0^n = 0$, $v_{-1}^n = 1$; thus overall we have that

$$v_j^{n+1} = v_j^n = \ldots = v_j^0 = u_0(x_j), \quad \forall n, j, \tag{2.2.6}$$

which is stationary and totally wrong comparing with the exact one (2.2.4).

Definition 2.2.4. (Conservative method) A numerical method is said to be conservative if it discretizes the integral form, i.e., Eq. (2.1.4), of the conservation law (2.1.1). A conservative method has the following form

$$\bar{v}_j^{n+1} = \bar{v}_j^n - \sigma(\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}), \tag{2.2.7}$$

where the numerical flux $\hat{f}_{j+\frac{1}{2}}$ defined in Eq. (2.2.2) and is,

i. consistent with the analytic flux, i.e.,

$$\hat{f}(u, \ldots, u) = f(u), \tag{2.2.8}$$

ii. Lipschitz continuous in all of its arguments, that is, there is a uniform constant L such that, $\forall j$'s,

$$\left|\hat{f}(\bar{v}^n_{j-p+1},\ldots,\bar{v}^n_{j+q};x_{j+\frac{1}{2}})-\hat{f}(u(x_j,t_n),\ldots,u(x_j,t_n))\right|\leq L\max_{-p+1\leq k\leq q}\left|\bar{v}^n_{j+k}-u(x_j,t_n)\right|. \tag{2.2.9}$$

The advantage of a conservative method is shown in the below Lax-Wendroff theorem ([98], [55]) which guarantees the convergence to some weak solution of Eq. (2.1.1), in case of convergence.

Theorem 2.2.1. *(Lax-Wendroff) Let $\{v_{\Delta x}\}$ be an approximating sequence with respect to Δx to conservation law (2.1.1) obtained from a conservative method (2.2.7). Let $\Omega=(-\infty,\infty)\times(0,\infty)$. We assume that the following conditions hold for $\{v_{\Delta x}\}$:*

i. (Uniform boundedness) There is $M>0$ constant such that

$$\|v_{\Delta x}\|_{L^\infty(\Omega)}\leq M,\quad \textit{for all } \Delta x, \tag{2.2.10}$$

ii. (Pointwise convergence) There is a function $u(x,t)$ such that

$$\lim_{\Delta x\to 0}v_{\Delta x}(x,t)=u(x,t)\quad \textit{a.e., in } \Omega. \tag{2.2.11}$$

Then, $u(x,t)$ is a weak solution of Eq. (2.1.1).

Proof. We need the following lemma for the proof. It is the summation by parts which is a discrete analogue of the integration by parts, and is a useful technique in numerical analysis.

Lemma 2.2.1. *(Summation by parts)*

$$\sum_{k=N+1}^{M}f_k(g_k-g_{k-1})=f_Mg_M-f_Ng_N-\sum_{k=N+1}^{M}(f_k-f_{k-1})g_{k-1}. \tag{2.2.12}$$

The proof of the lemma is simple. One just needs to re-arrange the indexes and (2.2.12) follows immediately.

Let $\varphi(x,t)$ be any smooth function which vanishes as t and $|x|$ are large. Restrict it to grid points $\varphi^n_j=\varphi(x_j,t_n)$. Multiplying Eq. (2.2.7) with φ^n_j, summing over the whole domain, and rearranging terms we obtain that

$$\begin{aligned} 0 &= \sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty}\left[\frac{\bar{v}_j^{n+1}-\bar{v}_j^n}{\Delta t}-\frac{\hat{f}_{j+\frac{1}{2}}-\hat{f}_{j-\frac{1}{2}}}{\Delta x}\right]\varphi_j^n \\ &= \sum_{j=-\infty}^{\infty}\varphi_j^0\bar{v}_j^0+\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty}\frac{\varphi_j^{n+1}-\varphi_j^n}{\Delta t}\bar{v}_j^n+\sum_{n=0}^{\infty}\sum_{j=-\infty}^{\infty}\frac{\varphi_{j+1}^n-\varphi_j^n}{\Delta x}\hat{f}_{j+\frac{1}{2}}, \end{aligned} \tag{2.2.13}$$

where the last equality is obtained by summation by parts Eq. (2.2.12).

Multiplying Eq. (2.2.13) with $\Delta x\Delta t$ and letting $\Delta x \to 0$. We note that $\Delta t = \mathcal{O}(\Delta x) \to 0$ as $\Delta x \to 0$. Changing $\sum$ to $\int$, we obtain that

$$\begin{aligned} 0 = &\lim_{\Delta x\to 0}\int_{-\infty}^{\infty}\varphi(x,0)v_{\Delta x}(x,0)dx+\lim_{\Delta x\to 0}\int_0^{\infty}\int_{-\infty}^{\infty}\frac{\varphi(x,t+\Delta t)-\varphi(x,t)}{\Delta t}v_{\Delta x}(x,t)dxdt \\ &+\lim_{\Delta x\to 0}\int_0^{\infty}\int_{-\infty}^{\infty}\frac{\varphi(x+\Delta x,t)-\varphi(x,t)}{\Delta x}\hat{f}_{j+\frac{1}{2}}dxdt, \end{aligned} \tag{2.2.14}$$

where

$$\hat{f}_{j+\frac{1}{2}} = \hat{f}(v_{\Delta x}(x+(1-p)\Delta x,t),\dots,v_{\Delta x}(x+q\Delta x,t)). \tag{2.2.15}$$

It suffices to show that:

$$\lim_{\Delta x\to 0}\int_0^{\infty}\int_{-\infty}^{\infty}\frac{\varphi(x,t+\Delta t)-\varphi(x,t)}{\Delta t}v_{\Delta x}(x,t)dxdt = \int_0^{\infty}\int_{-\infty}^{\infty}\varphi_t(x,t)u(x,t)dxdt; \tag{2.2.16}$$

$$\lim_{\Delta x\to 0}\int_{-\infty}^{\infty}\varphi(x,0)v_{\Delta x}(x,0)dx = \int_{-\infty}^{\infty}\varphi(x,0)u_0(x)dx; \tag{2.2.17}$$

$$\lim_{\Delta x\to 0}\int_0^{\infty}\int_{-\infty}^{\infty}\frac{\varphi(x+\Delta x,t)-\varphi(x,t)}{\Delta x}\hat{f}_{j+\frac{1}{2}}dxdt = \int_0^{\infty}\int_{-\infty}^{\infty}\varphi_x(x,t)f(u(x,t))dxdt. \tag{2.2.18}$$

We first have that

$$\begin{aligned} &\left|v_{\Delta x}(x,t)\frac{\varphi(x,t+\Delta t)-\varphi(x,t)}{\Delta t}-u(x,t)\varphi_t(x,t)\right| \\ &\qquad\le |v_{\Delta x}(x,t)|\left|\frac{\varphi(x,t+\Delta t)-\varphi(x,t)}{\Delta t}-\varphi_t(x,t)\right|+|v_{\Delta x}(x,t)-u(x,t)|\,|\varphi_t(x,t)| \\ &\qquad< M\varepsilon_1+\tilde{M}\varepsilon_2, \end{aligned} \tag{2.2.19}$$

for sufficiently small Δx with $\|\varphi_t\|_{L^\infty(\Omega)} \le \tilde{M} < \infty$, thanks to conditions (2.2.10) and (2.2.11) and the fact that φ is smooth. Thus

$$\lim_{\Delta x \to 0} v_{\Delta x}(x,t) \frac{\varphi(x,t+\Delta t) - \varphi(x,t)}{\Delta t} = u(x,t)\varphi_t(x,t) \quad a.e. \tag{2.2.20}$$

We also note that

$$\left\| v_{\Delta x} \frac{\varphi(\cdot,\cdot+\Delta t) - \varphi(\cdot,\cdot)}{\Delta t} \right\|_{L^\infty(\Omega)} \le C M \tilde{M}, \tag{2.2.21}$$

for all Δx and some constant C which can be derived from a Taylor expansion. Hence, by Lebesgue's convergence theorem, we deduce Eq. (2.2.16).

The proofs of Eqs. (2.2.17) and (2.2.18) follow similarly. For the latter, we need the following pointwise convergence.

By Lipschitz condition (2.2.9), we have that

$$\begin{aligned} \left| \hat{f}\ (v_{\Delta x}(x_j + (1-p)\Delta x, t_n), \ldots, v_{\Delta x}(x_j + q\Delta x, t_n)) - \hat{f}(u(x_j,t_n), \ldots, u(x_j,t_n)) \right| \\ \le L \max_{-p+1 \le k \le q} \left| v_{\Delta x}(x_{j+k}, t_n) - u(x_j,t_n) \right|, \end{aligned} \tag{2.2.22}$$

which implies that

$$\lim_{\Delta x \to 0} \hat{f}_{j+\frac{1}{2}} = \hat{f}(u(x,t), \ldots, u(x,t)) = f(u(x,t)) \quad a.e. \tag{2.2.23}$$

since $v_{\Delta x}(x,t) \to u(x,t)$ as $\Delta x \to 0$ *a.e.* in Ω. The last equality is due to the consistency (2.2.8). □

2.2.3 Total Variation Stability, Convergence, and Discrete Entropy Condition

The Lax-Wendroff theorem guarantees that a conservative method, if converges, will converge to a weak solution, but says nothing about criteria for such a method to converge. It turns out that the convergence depends heavily on the concept of compactness of bounded functions whose total variations (TV) are also bounded. This stems from Helly's selection principle below. Firstly, we define the total variation of a function.

Definition 2.2.5. (Total variation)

i. Let $\Pi_N = \{0, 1, \ldots, j, \ldots, N-1, N\}$ be a set of natural indexes. We define the total variation of a function $g(x)$ as below,

$$TV(g) = \sup_{\Pi_N} \sum_{j=1}^{N} |g(x_j) - g(x_{j-1})|, \tag{2.2.24}$$

where the grid $-\infty = x_0 < \ldots < x_j < \ldots < x_N = \infty$ and the supremum is taken over all Π_N's.

ii. A function $g(x)$ is of bounded total variation if its TV is finite.

Lemma 2.2.2. *(Helly's selection principle) (see [103]) Let an infinite sequence of functions $\{g_k\}$ with $g_k(x) : [x_a, x_b] \to \mathbb{R}$. Assuming that the sequence is uniformly bounded and of bounded total variation, i.e., there exist constants M_1, M_2 such that, for all k's,*

i. $\|g_k\|_\infty \leq M_1$,

ii. $TV(g_k) \leq M_2$.

Then there is a subsequence $\{g_{k_j}\} \subset \{g_k\}$ and a function $g(x) : [x_a, x_b] \to \mathbb{R}$ such that

$$\lim_{j\to\infty} g_{k_j} = g(x), \quad a.e. \tag{2.2.25}$$

We note that since the domain is bounded, and the functions are uniformly bounded, the pointwise convergence (2.2.25) implies convergence in L^1 norm by Lebesgue's convergence theorem.

We are now ready to discuss the condition for a sequence of approximations of Eq. (2.1.1) in the form (2.2.7) to converge. It is called total variation stability.

Definition 2.2.6. (Total variation stability) An approximation $v_{\Delta x}(x, t)$ is said to be total variation stable (TV-stable) if there is $M > 0$ constant such that

$$TV(v_{\Delta x}(\cdot, t)) \leq M, \quad \forall \Delta x, \Delta t. \tag{2.2.26}$$

The uniform bound in space in terms of total variation implies a uniform bound in time, as stated in the following lemma (see [96]).

Lemma 2.2.3. *(Uniform boundedness in time) There is a uniform $\tilde{M} > 0$ such that, for fixed Δx and all n's,*

$$\|v_{\Delta x}(\cdot, t_{n+1}) - v_{\Delta x}(\cdot, t_n)\|_{L^1} \leq \tilde{M} \Delta t. \tag{2.2.27}$$

Proof. For all n's, we have by Eq. (2.2.7)

$$\bar{v}_j^{n+1} - \bar{v}_j^n = -\sigma(\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}), \tag{2.2.28}$$

where $\sigma = \frac{\Delta t}{\Delta x}$.

Summing over all j's and multiplying both sides with Δx, we obtain that

$$\|v_{\Delta x}(\cdot, t_{n+1}) - v_{\Delta x}(\cdot, t_n)\|_{L^1} \leq \sigma \sum_{j=-\infty}^{\infty} |\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}|. \tag{2.2.29}$$

Since $\hat{f}$ is Lipschitz continuous, we have that

$$\begin{aligned} |\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}| &\leq L \max_{-p\leq l\leq q} |\bar{v}_{j+l}^n - \bar{v}_{j+l-1}^n| \\ &\leq L \sum_{l=-p}^{q} |\bar{v}_{j+l}^n - \bar{v}_{j+l-1}^n|. \end{aligned} \tag{2.2.30}$$

Hence,

$$\begin{aligned} \|v_{\Delta x}(\cdot, t_{n+1}) - v_{\Delta x}(\cdot, t_n)\|_{L^1} &\leq \Delta t L \sum_{l=-p}^{q} \sum_{j=-\infty}^{\infty} |\bar{v}_{j+l}^n - \bar{v}_{j+l-1}^n| \\ &\leq \Delta t L \sum_{l=-p}^{q} TV(v_{\Delta x}(\cdot, t_n)) \\ &\leq \Delta t L(p+q+1)M, \end{aligned} \tag{2.2.31}$$

thanks to the uniform bound (2.2.26).

Choosing $\tilde{M} := L(p+q+1)$ completes the proof. □

Theorem 2.2.2. *Suppose that a sequence $\{v_{\Delta x}(x,t)\}$ with respect to Δx obtained from a conservative method and is TV-stable. Then, for any bounded domain $\Omega \subset \mathbb{R}$, and a finite time $0 < T < \infty$, there are $u(x,t)$ and $\{v_{\Delta x_k}\} \subset \{v_{\Delta x}\}$ such that,*

$$\lim_{\Delta x_k \to 0} \|v_{\Delta x_k}(\cdot, t) - u(\cdot, t)\|_{L^1(\Omega)} = 0, \quad \textit{uniformly for all } t \in [0,T]. \tag{2.2.32}$$

Proof. (sketch of proof) (see [36], [54], or [127] for the Lax-Friedrichs scheme)

i. For any fixed time $t_n \in [0,T]$, since Ω is bounded, and $\{v_{\Delta x}\}$ is TV-stable, i.e., there is

$M > 0$ such that condition (2.2.26) holds, it can be checked that

$$\|v_{\Delta x}(\cdot, t_n)\|_{L^\infty} \leq \tilde{M}, \tag{2.2.33}$$

for some constant $\tilde{M}$ dependent on M. Then, by Helly's selection principle, there is an L^1-convergent subsequence of $\{v_{\Delta x_n}\} \subset \{v_{\Delta x}\}$ at time t_n.

ii. Since the set $\{t_n\} \in \mathbb{Q}$ is countable, by a diagonal argument, we can then select another subsequence from all $\{v_{\Delta x_n}\}$'s which converges in L^1 for all $t_n \in [0, T]$. We denote this subsequence $\{v_{\Delta x_k}\}$.

iii. We prove that $\{v_{\Delta x_k}\}$ converges for all $t \in [0, T]$. Since $\{t_n\}$ is dense in $\mathbb{R}$, $\forall t \in [0, T]$, there is an index n such that $t_n \leq t \leq t_{n+1}$. Let $v_{\Delta x_{k_i}}, v_{\Delta x_{k_j}} \in \{v_{\Delta x_k}\}$, we have that

$$\begin{aligned}\int_\Omega |v_{\Delta x_{k_i}}(x,t) - v_{\Delta x_{k_j}}(x,t)|dx \leq & \int_\Omega |v_{\Delta x_{k_i}}(x,t) - v_{\Delta x_{k_i}}(x,t_n)|dx \\ &+ \int_\Omega |v_{\Delta x_{k_i}}(x,t_n) - v_{\Delta x_{k_j}}(x,t_n)|dx \\ &+ \int_\Omega |v_{\Delta x_{k_j}}(x,t) - v_{\Delta x_{k_j}}(x,t_n)|dx.\end{aligned} \tag{2.2.34}$$

The second term on the RHS converges pointwise at t_n in space since $\{v_{\Delta x_k}\}$ is convergent. The other two, thanks to the inequality (2.2.27), are uniformly bounded by $\mathcal{O}(\Delta t) = \mathcal{O}(\Delta x) \to 0$. Hence, $v_{\Delta x_k}(x,t)$ converges uniformly for all $t \in [0, T]$, which completes the proof.

□

We immediately have the following corollaries.

Corollary 2.2.1. *(Weak solution) The limit function $u(x,t)$ obtained from Theorem 2.2.2 is a weak solution of conservation law (2.1.1).*

Proof. Since $v_{\Delta x}(x,t)$ is obtained from a conservative method, and converges by theorem 2.2.2, by Lax-Wendroff theorem 2.2.1, the limit $u(x,t)$ is a weak solution. □

Lemma 2.2.4. *(Uniqueness)([55]) Suppose $v_{\Delta x}(x,t)$ converges in the sense of theorem 2.2.2. We further assume that $v_{\Delta x}(x,t)$ satisfies the discrete entropy condition*

$$U_j^{n+1} \leq U_j^n - \sigma(\hat{F}_{j+\frac{1}{2}} - \hat{F}_{j-\frac{1}{2}}), \tag{2.2.35}$$

where

$$U_j^n = U(\bar{v}_j^n), \quad \hat{F}_{j+\frac{1}{2}} = \hat{F}(\bar{v}_{j-p+1}^n, \ldots, \bar{v}_{j+q}^n; x_{j+\frac{1}{2}}), \tag{2.2.36}$$

are the entropy pair defined in definition 2.1.2, $\hat{F}_{j+\frac{1}{2}}$ is the numerical entropy flux function, consistent with $F(u)$ and is Lipschitz continuous in the sense of Eqs. (2.2.8) and (2.2.9).

Then, $v_{\Delta x}(x,t)$ converges to a physically admissible weak solution, which is unique.

Proof. The proof follows that of the Lax-Wendroff theorem with $v_{\Delta x}$ and $\hat{f}_{j+\frac{1}{2}}$ replaced by $U(v_{\Delta x})$ and $\hat{F}_{j+\frac{1}{2}}$, respectively, and the equality by an inequality. □

We now proceed to discuss conservative schemes which satisfies the TV-stable condition (2.2.26). In particular, we will mention total variation diminishing (TVD) schemes originated from the work of Harten in [46] and his celebrated essentially non-oscillatory (ENO) schemes ([51]). For for former, the uniform bound M is taken to be $TV(u_0)$.

2.3 Total Variation Diminishing (TVD) Schemes

Definition 2.3.1. (TVD schemes) (see [46]) A conservative scheme

$$\bar{v}_j^{n+1} = \bar{v}_j^n - \sigma(\hat{f}(v_{j-p+1}^n, \ldots, v_{j+q}^n; x_{j+\frac{1}{2}}) - \hat{f}(v_{j-p}^n, \ldots, v_{j+q-1}^n; x_{j-\frac{1}{2}})), \tag{2.3.1}$$

or in its operator form,

$$v_{\Delta x}^{n+1} = \mathcal{L}(v_{\Delta x}^n), \tag{2.3.2}$$

is called TVD if its total variation is non-increasing in time. That is,

$$TV(v_{\Delta x}^{n+1}) \leq TV(v_{\Delta x}^n) \leq \ldots \leq TV(v_{\Delta x}^0) = TV(u_0). \tag{2.3.3}$$

The following lemma sets the conditions for such a conservative scheme (2.3.1) to be TVD.

Lemma 2.3.1. *(TVD conditions) ([46]) Suppose that a conservative scheme can be written in the form*

$$\bar{v}_j^{n+1} = \bar{v}_j^n + C(\bar{v}_{j+1}^n - \bar{v}_j^n) - D(\bar{v}_j^n - \bar{v}_{j-1}^n), \tag{2.3.4}$$

where the coefficients may depend on $v_{\Delta x}^n$.

Then the scheme is TVD providing that the following conditions hold,

$$\begin{cases} C \geq 0, \quad D \geq 0, \\ C + D \leq 1. \end{cases} \tag{2.3.5}$$

Proof. We check that

$$\begin{aligned}
TV(v_{\Delta x}^{n+1}) &= \sum_j |\bar{v}_j^{n+1} - \bar{v}_{j-1}^{n+1}| \\
&= \sum_j |\bar{v}_j^n + C(\bar{v}_{j+1}^n - \bar{v}_j^n) - D(\bar{v}_j^n - \bar{v}_{j-1}^n) - \bar{v}_{j-1}^n - C(\bar{v}_j^n - \bar{v}_{j-1}^n) + D(\bar{v}_{j-1}^n - \bar{v}_{j-2}^n)| \\
&= \sum_j |(1 - C - D)(\bar{v}_j^n - \bar{v}_{j-1}^n) + C(\bar{v}_{j+1}^n - \bar{v}_j^n) + D(\bar{v}_{j-1}^n - \bar{v}_{j-2}^n)| \\
&\leq (1 - C - D) \sum_j |\bar{v}_j^n - \bar{v}_{j-1}^n| + C \sum_j |\bar{v}_{j+1}^n - \bar{v}_j^n| + D \sum_j |\bar{v}_{j-1}^n - \bar{v}_{j-2}^n| \\
&= TV(v_{\Delta x}^n).
\end{aligned} \tag{2.3.6}$$

The inequality is thanks to condition (2.3.5). The last equality is obtained by rearranging the indexes in the last two terms of the inequality. □

Below we list the properties of TVD schemes. These are the analogues of those given in proposition 2.1.1. The proofs can be found in the indicated references.

Proposition 2.3.1. *For a TVD scheme, the following properties hold.*

i. A TVD scheme is monotonicity preserving. (See [46].)

ii. (Maximum principle) (see [111]) Let $m = \min_j v_{\Delta x}^0 = \min_j u_0(x_j)$, $M = \max_j v_{\Delta x}^0 = \max_j u_0(x_j)$*. Then, for any time* $t_n > 0$ *and each* j,

$$m \leq \bar{v}_j^n \leq M. \tag{2.3.7}$$

iii. A TVD scheme is at most first-order accurate at non-sonic critical points of the solution u*, that is, where* $f'(u) = 0$*. (See [111].)*

iv. A 2D TVD scheme is at most first-order accurate. (See [39].)

Remark 2.3.1.

i. Properties (*i.*) and (*ii.*) assures no oscillations in the approximation in a sense that the operator $\mathcal{L}$ in Eq. (2.3.2) is monotonicity preserving (see proposition 2.1.1).

ii. TVD condition (2.3.5) is crucial for preventing oscillations from happening, but is strict and makes the scheme badly performs near critical points, i.e., at most first-order (see properties (*iii.*) and (*iv.*) and [112] for high-order TVD schemes). It will be relaxed in ENO and WENO schemes. See the below section.

In the following sections, we mention commonly discussed TVD schemes, starting from the first-order ones. We discuss in detail the Godonov scheme, which is the fundamental approximation for all the others in the upwind approach. The term “upwind” refers to the incorporation of the characteristic information of the solution in the numerical discretizations. Here, we note that another approach based on central discretizations stands alone as an active research field till the present. Central schemes stem from the first-order Lax-Friedrichs and was pioneered in the paper of Nessyahu and Tadmor ([106]) for hyperbolic conservation laws on a staggered grid. We refer to, e.g., [10], [94], or [117] for high-order central schemes. We then proceed to the treatments on the coefficients C and D in definition 2.3.1 so that the accuracy is improved to second order, except at critical points as indicated in proposition 2.3.1. Here, the essential idea is to design such limiters that no oscillations occur in the numerical approximations. Typical limiters will also be listed. We limit the discussion to only the case of scalar conservation laws.

2.3.1 First-order Schemes

For these schemes, the approximation of the flux at the interface $x_{j+\frac{1}{2}}$ involves the information of the solution at two grid points x_j and x_{j+1}. Thus Eq. (2.3.1) is written as

$$\bar{v}_j^{n+1} = \bar{v}_j^n - \sigma(\hat{f}(\bar{v}_j^n, \bar{v}_{j+1}^n; x_{j+\frac{1}{2}}) - \hat{f}(\bar{v}_{j-1}^n, \bar{v}_j^n; x_{j-\frac{1}{2}})). \tag{2.3.8}$$

2.3.1.1 Godonov's Scheme

Godonov in [37] (see also [55], [147]) proposed a method to approximate $u(x_{j+\frac{1}{2}}, t_n)$ by exactly solving the Riemann problem of the piecewise-constant approximation $v_{\Delta x}(x,t)$. That is,

$$\hat{f}(\bar{v}_j^n, \bar{v}_{j+1}^n; x_{j+\frac{1}{2}}) = \hat{f}^G(\bar{v}_j^n, \bar{v}_{j+1}^n) = \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(\tilde{v}(x_{j+\frac{1}{2}}, t))dt, \tag{2.3.9}$$

where $\tilde{v}$ is the exact solution of the Riemann problem

$$\begin{cases} \tilde{v}_t + f(\tilde{v})_x = 0, \quad x_j < x < x_{j+1},\ t_n < t < t_{n+1}, \\ \tilde{v}(x,t_n) = \begin{cases} \bar{v}_j^n =: v_L, & \text{if } x < x_{j+\frac{1}{2}}, \\ \bar{v}_{j+1}^n =: v_R, & \text{if } x > x_{j+\frac{1}{2}}, \end{cases} \end{cases} \tag{2.3.10}$$

Since v_L, v_R are constants, problem (2.3.10) can be solved exactly as follows, assuming that $f''(\tilde{v}) \geq 0$, and that there are no interactions of waves, i.e., we limit their distance of propagation at most half of the interval,

$$\max_{\tilde{v}} |f'(\tilde{v})|\Delta t \leq \frac{1}{2}\Delta x. \tag{2.3.11}$$

Eq. (2.3.11) is often referred as the CFL condition.

- *Case 1: $v_L > v_R$ (shock-wave)*

$$\tilde{v}(x,t) = \begin{cases} v_L, & x/t < s, \\ v_R, & x/t > s, \end{cases} \tag{2.3.12}$$

where

$$s = \frac{f(v_R) - f(v_L)}{v_R - v_L}, \tag{2.3.13}$$

is the propagation speed of the shock by the Rankine-Hugoniot condition.

- *Case 2: $v_L \leq v_R$ (rarefaction-wave)*

$$\tilde{v}(x,t) = \begin{cases} v_L, & x/t < f'(v_L), \\ f^{-1}(x/t), & f'(v_L) \leq x/t \leq f'(v_R), \\ v_R, & x/t > f'(v_R). \end{cases} \tag{2.3.14}$$

Thanks to condition (2.3.11), there are no interactions of waves, thus $\tilde{v}(x_{j+\frac{1}{2}}, t)$ remains constant for all $t_n \leq t < t_{n+1}$. The Godonov's flux is then correspondingly written compactly as below (see [109]),

$$\begin{aligned} \hat{f}^G(\bar{v}_j^n, \bar{v}_{j+1}^n) &= \frac{1}{\Delta t}\int_{t_n}^{t_{n+1}} f(\tilde{v}(x_{j+\frac{1}{2}}, t))dt = f(\tilde{v}(x_{j+\frac{1}{2}}, t)) \\ &= \begin{cases} \min_{\bar{v}_j^n \leq \tilde{v} \leq \bar{v}_{j+1}^n} f(\tilde{v}), & \text{if } \bar{v}_j^n \leq \bar{v}_{j+1}^n, \\ \max_{\bar{v}_{j+1}^n \leq \tilde{v} \leq \bar{v}_j^n} f(\tilde{v}), & \text{if } \bar{v}_j^n > \bar{v}_{j+1}^n. \end{cases} \end{aligned} \tag{2.3.15}$$

We check that Godonov's scheme is conservative, TVD, and satisfies the discrete entropy condition (2.2.35).

Proposition 2.3.2. *(Godonov's scheme) The Godonov scheme with the flux defined in Eq. (2.3.15) is conservative, TVD, and converges to the unique physical weak solution.*

Proof. i. (Conservativeness) It suffices to show that the Godonov flux (2.3.15) is consistent and Lipschitz continuous. The former property follows directly from the definition of the flux. For the latter, since $f(u)$ is $\mathcal{C}^2$, by the mean value theorem there exists a constant L such that, for any u_1, u_2 we have that

$$|f(u_1) - f(u_2)| \leq L|u_1 - u_2|. \tag{2.3.16}$$

Suppose that $\bar{v}_j^n \leq \bar{v}_{j+1}^n$, then

$$\hat{f}^G(\bar{v}_j^n, \bar{v}_{j+1}^n) = \min_{\bar{v}_j^n \leq \tilde{v} \leq \bar{v}_{j+1}^n} f(\tilde{v}) =: f(\hat{v}), \tag{2.3.17}$$

for some $\hat{v} \in [\bar{v}_j^n, \bar{v}_{j+1}^n]$.

Thus,

$$|\hat{f}^G(\bar{v}_j^n, \bar{v}_{j+1}^n) - f(u)| = |f(\hat{v}) - f(u))| \leq L_1|\hat{v} - u|, \tag{2.3.18}$$

for sufficiently small $|\hat{v}-u|$. Similarly, for $\bar{v}_j^n > \bar{v}_{j+1}^n$, there exists L_2 such that Eq. (2.3.18) holds. Choosing $L = \max(L_1, L_2)$ proves relation (2.3.16).

ii. (TVD'ness) Comparing the conservative form (2.2.7) and the TVD one (2.3.4), we deduce that

$$\frac{\Delta x}{\Delta t}C = -\frac{\hat{f}^G_{j+\frac{1}{2}} - f(\bar{v}_j^n)}{\bar{v}_{j+1}^n - \bar{v}_j^n} \geq 0, \quad \frac{\Delta x}{\Delta t}D = -\frac{\hat{f}^G_{j-\frac{1}{2}} - f(\bar{v}_j^n)}{\bar{v}_j^n - \bar{v}_{j-1}^n} \geq 0, \tag{2.3.19}$$

by the definition (2.3.15) of the Godonov flux (2.3.15). We next check that

$$C + D = \frac{\Delta t}{\Delta x}\left[\frac{f(\bar{v}_j^n) - \hat{f}^G_{j+\frac{1}{2}}}{\bar{v}_{j+1}^n - \bar{v}_j^n} + \frac{f(\bar{v}_j^n) - \hat{f}^G_{j-\frac{1}{2}}}{\bar{v}_j^n - \bar{v}_{j-1}^n}\right] \leq 1, \tag{2.3.20}$$

thanks to the CFL condition (2.3.11). Hence, Godonov's scheme is TVD.

iii. (Entropy condition) Since Godonov's scheme solves the Riemann problem (2.3.10) exactly, the following entropy condition holds for $\tilde{v}$,

$$\frac{\partial}{\partial t}U(\tilde{v}(x,t)) + \frac{\partial}{\partial t}F(\tilde{v}(x,t)) \leq 0, \tag{2.3.21}$$

where (U, F) is an entropy pair.

Integrating (2.3.21) over $\Omega := I_j \times [t_n, t_{n+1})$, we obtain that

$$\begin{aligned} &\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} U(\tilde{v}(x,t_{n+1}))dx - \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} U(\tilde{v}(x,t_n))dx \\ &\qquad + \frac{1}{\Delta x}\int_{t_n}^{t_{n+1}} F(\tilde{v}(x_{j+\frac{1}{2}},\tau))d\tau - \frac{1}{\Delta x}\int_{t_n}^{t_{n+1}} F(\tilde{v}(x_{j-\frac{1}{2}},\tau))d\tau \leq 0. \end{aligned} \tag{2.3.22}$$

Thanks to condition (2.3.11), $\tilde{v}$ is constant along its characteristic, i.e., for all $(x,t) \in \Omega$,

$$\begin{cases} \tilde{v}(x,t_n) = \bar{v}_j^n, \\ \tilde{v}(x_{j\pm\frac{1}{2}},t) = \tilde{v}(x_{j\pm\frac{1}{2}},t_n), \end{cases} \tag{2.3.23}$$

we deduce that

$$\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} U(\tilde{v}(x,t_{n+1}))dx \leq U(\bar{v}_j^n) - \sigma(F(\tilde{v}(x_{j+\frac{1}{2}},t_n)) - F(\tilde{v}(x_{j-\frac{1}{2}},t_n))). \tag{2.3.24}$$

Since $U(\tilde{v}(x,t))$ is convex, by Jensen's inequality, we have that

$$U(\bar{v}_j^{n+1}) = U\left(\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \tilde{v}(x,t_{n+1})dx\right) \leq \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} U(\tilde{v}(x,t_{n+1}))dx; \tag{2.3.25}$$

hence,

$$U(\bar{v}_j^{n+1}) \leq U(\bar{v}_j^n) - \sigma(F(\tilde{v}(x_{j+\frac{1}{2}},t_n)) - F(\tilde{v}(x_{j-\frac{1}{2}},t_n)), \tag{2.3.26}$$

which is the discrete entropy condition (2.2.35). The uniqueness then follows immediately.

□

2.3.1.2 Other Schemes

The Godonov scheme, due to the exact solving of a Riemann problem in order to approximate $\bar{v}_{\Delta x}(x_{j+\frac{1}{2}}, t_n)$, is costly and complicated, especially when dealing with a system case. For the the latter, an iterative solver must be facilitated. For example, see [138] for the solver of Riemann problems for the Euler system. The below schemes, except for the Lax-Friedrichs, are derivations of the Godonov one where $\bar{v}_{\Delta x}(x_{j+\frac{1}{2}}, t_n)$ is approximated from the Riemann problem (2.3.10).

- *Lax-Friedrichs' scheme:*

$$\hat{f}^{LF}(\bar{v}_j^n, \bar{v}_{j+1}^n) = \frac{1}{2}(\bar{v}_j^n + \bar{v}_{j+1}^n) - \alpha(\bar{v}_{j+1}^n - \bar{v}_j^n), \tag{2.3.27}$$

  where $\alpha = \max_j |f'(\bar{v}_j^n)|$.

- *Engquist-Osher's scheme:* (see [28])

$$\hat{f}^{EO}(\bar{v}_j^n, \bar{v}_{j+1}^n) = \int_0^{\bar{v}_{j+1}^n} \min(f'(s), 0)ds + \int_0^{\bar{v}_j^n} \max(f'(s), 0)ds + f(0). \tag{2.3.28}$$

- *Roe's scheme:* ([118])

$$\hat{f}^{R}(\bar{v}_j^n, \bar{v}_{j+1}^n) = \begin{cases} f(\bar{v}_j^n), & \text{if } s > 0, \\ f(\bar{v}_{j+1}^n), & \text{if } s < 0, \end{cases} \tag{2.3.29}$$

  where

$$s = \frac{f(\bar{v}_{j+1}^n) - f(\bar{v}_j^n)}{\bar{v}_{j+1}^n - \bar{v}_j^n}. \tag{2.3.30}$$

Remark 2.3.2.

i. The Godonov flux (2.3.15) is also applicable for a non-convex flux function $f(u)$. (See [109].)

ii. The Engquist-Osher flux and Roe flux are approximations of the Godonov one in a sense of using only rarefaction-waves or shock-waves, respectively. (See [147], [128], [111], [56], etc.)

iii. Roe's scheme is especially efficient for a system case, e.g., the Euler equations ([118]), thanks to its simplicity. A tradeoff is that the scheme does admit nonphysical weak solutions, i.e., expansion shock-waves (see [55]), due to the use of shock-waves only in defining the flux. A remedy for this drawback is called an entropy fix, for example (see [92]),

$$\hat{f}^{RF}(\bar{v}_j^n, \bar{v}_{j+1}^n) = \begin{cases} f(\bar{v}_j^n), & \text{if } s \geq 0, \\ f(\bar{v}_j^{n+1}), & \text{if } s \leq 0, \\ \hat{f}^{LF}(\bar{v}_j^n, \bar{v}_{j+1}^n), & \end{cases} \tag{2.3.31}$$

where $\min(\bar{v}_j^n, \bar{v}_{j+1}^n) \leq \tilde{v} \leq \max(\bar{v}_j^n, \bar{v}_{j+1}^n)$.

iv. Except for the Roe flux, the other ones are called monotone fluxes, i.e., for all j's,

$$\frac{\partial \hat{f}}{\partial \bar{v}_j^n} \geq 0, \quad \frac{\partial \hat{f}}{\partial \bar{v}_{j+1}^n} \leq 0, \tag{2.3.32}$$

and are consistent and Lipschitz continuous; thus satisfy the discrete entropy condition (2.2.35). See [21] for detailed discussions on monotone fluxes and schemes. See also [109] for E schemes whose fluxes satisfy the entropy condition.

v. It is worthy to mention the Godonov theorem ([37]) which states that all monotone schemes are at most first-order. This implies all linear schemes applying to conservation laws are also at most first-order (see [46]). Hence, in order to improve the accuracy, one must design schemes whose coefficients depend on the solution themselves. This leads to the concept of "limiters" which play roles as an adapting mechanism of high- and low-order fluxes, or a controller of the slope of stencil-wise approximating polynomials so that condition (2.3.5) holds, i.e., no oscillations occur in the approximation. We discuss these approaches in the below section.

2.3.2 High-resolution TVD Schemes

Definition 2.3.2. (High resolution schemes) A numerical scheme is said to be of high resolution if its accuracy is higher than first-order. The term *resolution* implies how well the scheme resolves discontinuities in the numerical approximation.

There are three approaches in improving the accuracy and resolution of a scheme. These include the modified-equation-based, the flux-limiting, and slope-limiting methods. We recall that a modified equation is obtained by substituting the exact solution $u(x, t)$ into the difference equation of the approximation $v_{\Delta x}(x, t)$. Hence, an error term will appear which gives information on the behavior of the scheme. In this section, we briefly mention the main ideas underlying the latter two approaches, i.e., the flux and slope limiting. The schemes which depend on modified equations are limited to low orders due to its complication in manipulating the modified equations and are omitted here. Interested readers may refer to, e.g., [46] and the references therein for more detail.

A main difficulty when designing a high-resolution method is how one handles the dissipation and dispersion in the scheme. The former is responsible for smoothing (thus smearing) out all discontinuities or high gradients appearing in the numerical approximation, which in turn guarantees no oscillations. On the other hand, dispersion improves the resolution of the scheme, i.e., giving sharp capturing of discontinuities or small-scaled structure like vortices, but allows oscillations to occur. A typical technique to control the amount of dispersion of the scheme is through the so-called "limiters" which we will discuss below. We follow the presentation given in [96]. See also [128] for a complete discussion.

For simplicity, we discuss the schemes for a linear scalar case, i.e,

$$u_t + (au)_x = 0, \quad a > 0. \tag{2.3.33}$$

2.3.2.1 Flux-limiting Schemes

The underlying idea is to use a low-order (first-order) flux which is dissipative to handle discontinuities while in smooth regions, a high-order (second-order or higher) flux is employed for improving the formal accuracy order. The approach originated from the multi-step flux corrected transport (FCT) scheme by Boris and Book (see, e.g., [7], [138]) and later adapted for one-step methods and improved in, for examples, [119], [23], and the references therein.

In seeking for such a scheme, we write the numerical flux in the following adaptive form,

$$\hat{f}_{j+\frac{1}{2}} = \hat{f}^1_{j+\frac{1}{2}} + \phi(r_j)(\hat{f}^2_{j+\frac{1}{2}} + \hat{f}^1_{j+\frac{1}{2}}), \tag{2.3.34}$$

where $\hat{f}^1_{j+\frac{1}{2}}$ and $\hat{f}^2_{j+\frac{1}{2}}$ are the numerical fluxes of a $1st$- and $2nd$-order, respectively. ϕ is called a limiter depending on r which is defined as follows,

$$r_j = \frac{\bar{v}^n_j - \bar{v}^n_{j-1}}{\bar{v}^n_{j+1} - \bar{v}^n_j}. \tag{2.3.35}$$

We point out that r measures how smooth the solution is around x_j in a sense that it is considered smooth for r_j close to 1 and nonsmooth when r is far from 1. The principle to design ϕ is the TVD condition (2.3.5).

If we take the $2nd$-order Lax-Wendroff flux (see [96], [97])

$$\hat{f}^2_{j+\frac{1}{2}} = a\bar{v}^n_j + \frac{1}{2}a(1 - a\sigma)(\bar{v}^n_{j+1} - \bar{v}^n_j), \tag{2.3.36}$$

and the $1st$-order upwind one

$$\hat{f}^1_{j+\frac{1}{2}} = a\bar{v}^n_j, \tag{2.3.37}$$

substituted into Eq. (2.3.34), we obtain that

$$\hat{f}_{j+\frac{1}{2}} = a\bar{v}^n_j + \phi_j a(1 - a\sigma)(\bar{v}^n_{j+1} - \bar{v}^n_j), \tag{2.3.38}$$

where $\phi_j = \phi(r_j)$. We recall that $\sigma = \frac{\Delta t}{\Delta x}$.

Comparing with the TVD form (2.3.4), we deduce that

$$\begin{cases} C = a\sigma\left[1 + \frac{1}{2}(1 - a\sigma)\left(\frac{\phi_j}{r_j} - \phi_{j-1}\right)\right], \\ D = 0. \end{cases} \tag{2.3.39}$$

In order that the TVD property of the scheme is acquired, we require that

$$0 < C \leq 1; \tag{2.3.40}$$

that is, for all j's (see [128]),

$$\begin{cases} a\sigma \leq 1, \\ \left|\dfrac{\phi_j}{r_j} - \phi_{j-1}\right| \leq 2, \end{cases} \tag{2.3.41}$$

where the first is the normal CFL condition, while the second one is required from the TVD property. These are further reduced to

$$\begin{cases} \phi(r) = 0, \quad \text{for } r \leq 0, \\ 0 \leq \left(\dfrac{\phi(r)}{r}, \phi(r)\right) \leq 2. \end{cases} \tag{2.3.42}$$

It is shown in [144] that any $2nd$-order scheme relying on the stencil $\{u_{j-2}, u_{j-1}, u_j, u_{j+1}\}$ is a convex combination of the Lax-Wendroff (LW) and Beam-Warming (BW) scheme, i.e.,

$$\phi(r) = (1 - \theta(r))\phi^{LW}(r) + \theta(r)\phi^{BW}(r), \tag{2.3.43}$$

where

$$|\theta(r)| \leq 1, \quad \phi^{LW}(r) = 1, \quad \phi^{BW}(r) = r. \tag{2.3.44}$$

Thus,

$$\phi(r) = 1 + \theta(r)(r - 1). \tag{2.3.45}$$

Combining condition (2.3.45) with the TVD restriction (2.3.42), Sweby ([128]) indicated a region in which the limiter $\phi(r)$ must belong to so that the scheme is TVD. We refer to Figs. 1 and 2 in [128] for the visualization of the TVD region and limiters. We note that only Van Leer's limiter is continuous. See also [96], [97].

- Van Leer's limiter ([144]):

$$\phi^{VL}(r) = \frac{|r| + r}{|r| + 1}. \tag{2.3.46}$$

- Minmod limiter ([129]):

$$\phi^{MM}(r) = \text{minmod}(1, r), \tag{2.3.47}$$

where the *minmod* function is defined as follows,

$$\text{minmod}(a, b) = \begin{cases} |a|, & \text{for } ab > 0, |a| < |b|, \\ |b|, & \text{for } ab > 0, |a| \geq |b|, \\ 0, & \text{for } ab \leq 0. \end{cases} \tag{2.3.48}$$

- Superbee limiter ([120]):

$$\phi^{SB}(r) = \begin{cases} \text{maxmod}(1, r), & \text{for } \frac{1}{2} \leq r \leq 2, \\ 2\text{minmod}(1, r), & \text{for } r < \frac{1}{2} \text{ or } r > 2. \end{cases} \tag{2.3.49}$$

Here, the *maxmod* function is similar to the *minmod* except for choosing the larger modulus of between a and b if they are of the same sign.

- Chakravarthy and Osher's limiter ([23]):

$$\phi^{CO}(r) = \text{minmod}(r, \psi), \tag{2.3.50}$$

for some $1 \leq \psi \leq 2$.

2.3.2.2 Slope-limiting Schemes

The idea for slope-limiting schemes is to improve the Godonov method by using a more accurate stencil-wise approximating polynomials instead of a piecewise constant one. That is, we consider $v_{\Delta x}(x, t_n)$ as follows,

$$v^n_{\Delta x}(x) = \sum_j v^n_{S_j}(x) \chi_{S_j}(x), \tag{2.3.51}$$

where $v^n_{S_j}(x)$ is, say a linear polynomial, defined over the stencil $S_j \equiv I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$, i.e.,

$$v^n_{S_j}(x) = v^n_{I_j}(x) = \bar{v}^n_j + \gamma_j(x - x_j), \quad x_{j-\frac{1}{2}} < x \leq x_{j+\frac{1}{2}}, \tag{2.3.52}$$

for some slope γ_j. It is easy to check that, for any finite slope γ_j, we have that

$$\bar{v}_j^n = \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} (\bar{v}_j^n + \gamma_j(x - x_j))dx. \tag{2.3.53}$$

Thus the approximation (2.3.51) is conservative. The second order of the scheme is confirmed by observing that if we let

$$\gamma_j = \frac{\bar{v}_{j+1}^n - \bar{v}_j^n}{\Delta x}, \tag{2.3.54}$$

then Eq. (2.3.51) is an approximation obtained from the $2nd$-order Lax-Wendroff scheme.

In order to satisfy the TVD condition (2.3.5) which prevents oscillations from generating, we need to limit the slope γ_j in some sense. Hence, the schemes are called slope-limiting. Good examples for these schemes are the MUSCL by Van Leer (see [144], [145], [146], [147], [110]), or the quadratic PPM scheme by Colella Woodward (see [25]). The latter scheme uses a quadratic reconstructing polynomial instead of the linear one in Eq. (2.3.52).

Since the equation is linear with $a > 0$, we can solve for the solution exactly at t_{n+1}, i.e.,

$$v_{\Delta x}^{n+1}(x) = v_{\Delta x}^n(x - a\Delta t). \tag{2.3.55}$$

Thus for each interval I_j, we have that

$$v_{I_j}^{n+1}(x) = v_{I_{j-1}}^n(x)\chi_{[x_{j-\frac{1}{2}}, x_{j-\frac{1}{2}}+a\Delta t)}(x) + v_{I_j}^n(x)\chi_{[x_{j-\frac{1}{2}}+a\Delta t, x_{j+\frac{1}{2}}]}(x). \tag{2.3.56}$$

Integrating $v_{I_j}^{n+1}(x)$ over I_j, we obtain that

$$\bar{v}_j^{n+1} = \bar{v}_j^n - a\sigma(\bar{v}_j^n - \bar{v}_{j-1}^n) - \frac{1}{2}a\sigma(1 - a\sigma)(\gamma_j - \gamma_{j-1})\Delta x, \tag{2.3.57}$$

for which the flux is,

$$\hat{f}_{j+\frac{1}{2}} = a\bar{v}_j^n + \frac{1}{2}a(1 - a\sigma)\gamma_j\Delta x. \tag{2.3.58}$$

Setting

$$\gamma_j = \left(\frac{\bar{v}_{j+1}^n - \bar{v}_j^n}{\Delta x}\right)\phi_j, \tag{2.3.59}$$

gives the flux in Eq. (2.3.58) to be the same as that of a flux-limiting scheme, i.e., Eq. (2.3.34). Hence, the limiters given in Eqs. (2.3.46) - (2.3.50) can be applied in this case. This shows a close relation between the two approaches.

For a nonlinear conservation law, i.e., a replaced by $f'(u)$, the ideas are essential the same although the implementation is more complicated and tedious. We refer readers to, e.g., [128], [96], [97], or [138], etc.

Remark 2.3.3.

i. All TVD schemes have poor performance around critical points so that condition (2.3.5) holds, which in turn implies the monotonicity preserving property to prevent oscillations from happening. We take the *minmod* limiter for example. In the regions where the solution is monotone, the *minmod* chooses the linear approximating polynomial $v^n_{I_j}(x)$ in Eq. (2.3.52) with a slope of smaller modulus (see Eq. (2.3.59)), but near a critical point where $(\bar{v}^n_j - \bar{v}^n_{j-1})$ and $(\bar{v}^n_{j+1} - \bar{v}^n_j)$ have opposite signs, thus $\text{minmod}(\phi_j) = 0$, and so $v^n_{I_j}(x) = a\bar{v}^n_j$ which is just first-order. This is indeed the nature of TVD schemes where they are at most first-order at critical points. (See proposition 2.3.1.)

ii. In order to improve the accuracy at critical points, one must relax condition (2.3.5). In fact, comparing (2.3.5) and the TV-stability condition (2.2.26), the uniform constant M is not necessarily the TV of the initial data. This relaxation is given in ENO and WENO schemes. In the next section, we briefly mention the principles of ENO schemes, whereas the WENO will be discussed in detail in the next chapter.

2.4 Essentially Non-Oscillatory (ENO) Schemes

As indicated in the concluding remark in the previous section, the main drawback of TVD schemes is that it is necessary for the schemes to reduce the accuracy to first-order, no matter what order the schemes can achieve in smooth regions. Attempts to overcome this difficulty mostly involves how to relax the TVD condition (2.3.5) yet the TV stability (2.2.26) still holds, at least in some sense. One of the approaches was proposed by Shu in [123] where he constructed the total variation bounded (TVB) schemes, in which the total variation of the numerical approximation is not diminishing in time, but instead there remains a uniform bound dependent on the final time T. The author claimed that the scheme is uniformly high-order, including critical regions. In [59], [60], and later [51], Harten et al. suggested a new concept in designing a shock-capturing scheme. They named these essentially non-oscillatory (ENO) schemes. The non-oscillatory property of the numerical approximation does not depend on limiters as that of TVD schemes, but on the choice of the smoothest of the stencils for the reconstruction procedure. Here, the smoothness of a stencil is determined in the sense of undivided Newton's difference, which we will discuss below.

2.4.1 Semi-discrete Discretization

Approximating the 1D scalar conservation law (2.1.1) by a semi-discrete form, i.e., we only discretize with respect to space and let the time be continuous,

$$\frac{d\bar{v}_j(t)}{dt} = -\frac{\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}}{\Delta x}, \tag{2.4.1}$$

which is a discrete form of Eq. (2.1.3). The numerical flux

$$\hat{f}_{j+\frac{1}{2}} = \hat{f}(v^-_{j+\frac{1}{2}}, v^+_{j+\frac{1}{2}}) \approx f(u(x_{j+\frac{1}{2}}, t^k)), \quad t_n \leq t^k < t_{n+1}. \tag{2.4.2}$$

Here, we have abused the notation. The numerical flux (2.4.2) is different from that defined in Eq. (2.2.2) since not the time integration but the flux is approximated directly. In addition, $v^\pm_{j+\frac{1}{2}}$ is time-dependent for a time integrator must be employed together with the ENO spatial discretization. For example, we can use the typical third-order TVD RK3 method (see the next chapter about WENO schemes). For simplicity, we suppress the time dependence of $v^\pm_{j+\frac{1}{2}}$, which will be discussed below.

The scheme (2.4.1) is said to have r-order of accuracy if its spatial discretization, e.g., ENO type, is r-order. It is noted that for high-order schemes, in general, the accuracy of the time integrator is less than that of the spatial discretization method. Nevertheless, we can choose an appropriate time step Δt so that a uniform order is achieved in both time and space. Let r_t and r_x be the accuracy orders of the time integrator and spatial discretization, respectively. The truncation errors of the scheme (2.4.1) are

$$\frac{d\bar{v}_j(t)}{dt} = -\frac{\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}}{\Delta x} = -\left.\frac{\partial f}{\partial x}\right|_{x=x_j} + \mathcal{O}(\Delta t^{r_t}) + \mathcal{O}(\Delta x^{r_x}). \tag{2.4.3}$$

Choosing the time step such that

$$\Delta t = \min(\Delta t_{CFL}, \Delta x^{r_x/r_t}), \tag{2.4.4}$$

where the first argument is restrained from the CFL condition. Substituting in Eq. (2.4.3), we obtain a uniform Δx^{r_x} truncation error.

We note that $v^\pm_{j+\frac{1}{2}} = v^\pm(x_{j+\frac{1}{2}})$ where $v^\pm(x)$ are the reconstruction polynomials of the solution $u(x, t^k)$ for fixed time t^k on either side of interface $x_{j+\frac{1}{2}}$, respectively. The procedure is called reconstruction since a pointwise value is approximated from the averaged ones. Here, we again abuse the notations by not mentioning the time dependence of $\bar{v}_j$'s with the implication that these are given information from some fixed time step t^k. In terms of notations,

$\hat{f}(v^-_{j+\frac{1}{2}}, v^+_{j+\frac{1}{2}})$ is consistent with $\hat{f}(\bar{v}_{j-p}, \ldots, \bar{v}_{j+q}; x_{j+\frac{1}{2}})$ for

$$\begin{cases} v^-_{j+\frac{1}{2}} = v^-(x_{j+\frac{1}{2}}; \bar{v}^n_{j-p}, \ldots, \bar{v}^n_{j+q-1}), \\ v^+_{j+\frac{1}{2}} = v^+(x_{j+\frac{1}{2}}; \bar{v}^n_{j-p+1}, \ldots, \bar{v}^n_{j+q}). \end{cases} \tag{2.4.5}$$

We notice that in general, $v^-_{j+\frac{1}{2}} \neq v^+_{j+\frac{1}{2}}$ since $v^-(x) \neq v^+(x)$.

2.4.2 Reconstruction

For simplicity, we discuss the 1D scalar case. Let $S_j := \{I_{j-p}, \ldots, I_{j+q-1}\}$ be the reconstruction stencil with $p+q = r$ points. We remind that the global reconstruction function is denoted as follows at some fixed time t^k,

$$v_{\Delta x}(x) = \sum_j v_{S_j}(x)\chi_{S_j}(x). \tag{2.4.6}$$

Comparing Eq. (2.4.6) with (2.4.5), we have that $v^-(x) = v_{S_j}(x)$ and $v^+(x) = v_{S_{j+1}}(x)$.

An ENO reconstruction $v_{\Delta x}(x)$ with $v_{S_j}(x)$ be a polynomial of degree at most $(r-1)$ satisfies the following conditions ([51], [124]),

i. *Order of accuracy:*

$$v_{\Delta x}(x) = v(x) + \mathcal{O}(\Delta x^r), \tag{2.4.7}$$

where $v(x)$ is smooth over some neighborhood of x. Here, $v(x)$ is as follows, for all j's,

$$\bar{v}_j = \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v(x)dx. \tag{2.4.8}$$

Note that we are given only $\{\bar{v}_j\}$, not $v(x)$.

ii. *Conservation:* for all j's,

$$\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v_{\Delta x}(x)dx = \bar{v}_j. \tag{2.4.9}$$

iii. *Monotonicity:* $v_{\Delta x}(x)$ is monotone in any interval which contains a discontinuity.

iv. *ENO property:*

$$TV(v_{\Delta x}) \leq TV(v) + \mathcal{O}(\Delta x^r). \tag{2.4.10}$$

The last inequality implies that $v_{\Delta x}$ is reconstructed in such a way that no oscillations of order $\mathcal{O}(1)$ occur near discontinuities, but indeed allows those in the order of the truncation error. This follows from properties (i) and (ii) (see [124], [60]).

ENO reconstruction is based on the primitive function $V'(x) = v(x)$, and is constructed in a hierarchical approach in which a point either from the left or the right of the current stencil is added to form a new one. The basic stencil is chosen to be $S_0 := \{x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}\}$. We note that, thanks to property $(iii.)$ above, it is admissible for S_0 to contain a discontinuity. A point is added in such a way that the modulus of the undivided Newton difference of V over the new stencil is smaller. For example, starting from S_0, we compute $V[x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ and $V[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}]$. We note that an undivided Newton's difference has the recursive formula

$$g[x_0, \ldots, x_k] = g[x_1, \ldots, x_k] - g[x_0, \ldots, x_{k-1}], \tag{2.4.11}$$

where $g[x_0] = g(x_0)$.

We then choose $S_1 = S_0 \cup \{x_{j-\frac{3}{2}}\}$ if

$$\left|V[x_{j-\frac{3}{2}}, x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]\right| \leq \left|V[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}, x_{j+\frac{3}{2}}]\right|; \tag{2.4.12}$$

or $S_1 = S_0 \cup \{x_{j+\frac{3}{2}}\}$ otherwise.

The procedure continues until the designed accuracy order is achieved.

We note that the undivided Newton's difference is closely related to the smoothness of a function given appropriate smoothness. That is, for a polynomial $g(x)$ of degree k over the stencil $\{x_0, \ldots, x_k\}$, we have that

$$g[x_0, \ldots, x_k] = \frac{1}{k!}\frac{d^k g(\xi)}{dx^k}\Delta x^k = \mathcal{O}(\Delta x^k), \quad \text{for some } \xi \in [x_0, x_k]; \tag{2.4.13}$$

whereas

$$g[x_0, \ldots, x_k] = \mathcal{O}(1), \tag{2.4.14}$$

in case $g(x)$ is a function containing a discontinuity. This property can be checked directly from observing that the polynomial $g(x)$ can be represented in terms of undivided Newton's differences as follows,

$$g(x) = \sum_{l=0}^{k} g[x_0, \ldots, x_l]\Delta x^l \prod_{m=0}^{k-1}(x - x_m). \tag{2.4.15}$$

Differentiating $g(x)$ k times, we obtain relation (2.4.13) immediately.

Eq. (2.4.13) implies that our choice of stencils in the ENO reconstruction in Eq (2.4.12) always results in a smoother stencil. Proofs of properties (*i*) - (*iv*) for ENO reconstruction can be found in [60], or [124]. Details on the reconstruction algorithms, ENO schemes in terms of finite volume, finite difference, and numerical illustrations are given in [124].

2.4.3 Numerical Flux

By reconstruction, we can obtain approximations on the left and right of $x_{j+\frac{1}{2}}$, i.e., $v^{\pm}_{j+\frac{1}{2}} = v^{\pm}(x_{j+\frac{1}{2}})$. The issue now is how one can approximate $\hat{f}(v^{-}_{j+\frac{1}{2}}, v^{+}_{j+\frac{1}{2}})$. As suggested in the section of the Godonov scheme, we seek for $\tilde{v}(x,t)$ such that

$$\hat{f}(v^{-}_{j+\frac{1}{2}}, v^{+}_{j+\frac{1}{2}}) = f(\tilde{v}(x_{j+\frac{1}{2}})), \tag{2.4.16}$$

where $\tilde{v}$ is the exact or approximate solution of the generalized Riemann problem,

$$\begin{cases} \tilde{v}_t + f(\tilde{v})_x = 0, \quad x \in S_j,\ t_n \le t < t_{n+1}, \\ \tilde{v}(x,t_n) = \begin{cases} v^{-}(x), & \text{for } x < x_{j+\frac{1}{2}}, \\ v^{+}(x), & \text{for } x > x_{j+\frac{1}{2}}. \end{cases} \end{cases} \tag{2.4.17}$$

We note that since the initial data is dependent on $x \in S_j$, it is not a trivial task in solving Eq. (2.4.17). A closer observation at the numerical flux reveals that we do not need $\tilde{v}$ for all x, but at $x_{j+\frac{1}{2}}$ only. Denoting $v_L := v^{-}_{j+\frac{1}{2}}$, $v_R := v^{+}_{j+\frac{1}{2}}$, problem (2.4.17) reduces to exactly Eq. (2.3.10) whose solution is explicit. Hence, the monotone fluxes given in Eqs. (2.3.15), (2.3.27), (2.3.28), or the Roe's one with entropy fix Eq. (2.3.31) can be employed for ENO schemes, with $\bar{v}^n_j$ and $\bar{v}^n_{j+1}$ replaced by $v^{-}_{j+\frac{1}{2}}$, $v^{+}_{j+\frac{1}{2}}$, respectively.

2.4.4 Characteristic Projection

In this subsection, we consider the hyperbolic conservation law

$$\begin{cases} \mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0, \quad -\infty < x < \infty,\ t > 0, \\ \mathbf{u}(x,0) = \mathbf{u}_0(x). \end{cases} \tag{2.4.18}$$

For ENO schemes, we need the assumption that there are a finite number of discontinuities in the solution so that for sufficiently small grid size Δx, we can always choose such a smooth stencil for the reconstruction. Unfortunately, this condition is not always satisfied. For example, if there are two shocks propagating in time with different directions, there is a certain point of time when these shocks collide. A typical shock collision problem can be found in the the paper of Woodward and Colella ([151]). See also numerical test 5 in the next chapter. Thus no matter how small Δx is, there does not exist enough grid points for the reconstruction, which leads to oscillations. This implies that ENO discretizations of the component-wise variable $\mathbf{u}$ is not a good choice.

To circumvent the situation, it is suggested in [51] that ENO discretizations can be employed in the characteristic fields. As demonstrated in the previous section about characteristic decomposition, there is only one simple wave, either shock, rarefaction, or contact discontinuity, which propagates in time in each characteristic field. Thus the case that two shocks may collide is eliminated, and there is enough space to choose an appropriate smooth stencil for the reconstruction. A difficulty is for a nonlinear case, these characteristic fields continuously depends on the solution itself. To overcome this, we seek for the numerical solution of not the system (2.4.18) but the linearized one,

$$\mathbf{u}_t + A(\mathbf{u}_{j+\frac{1}{2}})\mathbf{u}_x = 0, \tag{2.4.19}$$

where $A(\mathbf{u}_{j+\frac{1}{2}})$ is the average (in some sense) Jacobian of the flux $\mathbf{f}(\mathbf{u})$. A good candidate is Roe's averaging one $\tilde{A}(\mathbf{u}_j, \mathbf{u}_{j+1})$, which satisfies the following conditions

i. $\tilde{A}(\mathbf{u}_j, \mathbf{u}_{j+1})$ is hyperbolic, i.e., the eigenvalues are real and eigenvectors are linearly independent.

ii. $\tilde{A}(\mathbf{u}_j, \mathbf{u}_{j+1})$ is consistent with $A(\mathbf{u})$, i.e., $\tilde{A}(\mathbf{u}, \mathbf{u}) = A(\mathbf{u})$.

iii. The following relation holds

$$\mathbf{f}(\mathbf{u}_j) - \mathbf{f}(\mathbf{u}_{j+1}) = \tilde{A}(\mathbf{u}_j, \mathbf{u}_{j+1})(\mathbf{u}_j - \mathbf{u}_{j+1}). \tag{2.4.20}$$

The last property mimics the Rankine-Hugoniot condition which ensures the conservation of the system in a sense that a shock wave is correctly recognized. In fact, by Eq. (2.4.20), the shock propagation speed is an eigenvalue of $\tilde{A}$. There exists a closed form of $\tilde{A}$ for the Euler systems. See [118], [138] for detail.

For ENO schemes in characteristic fields, we first seek for the eigenvalues, left and right eigenvectors of the Roe average $\tilde{A}(\mathbf{u}_j, \mathbf{u}_{j+1})$. We then project the variable $\mathbf{u}$ to the characteristic fields by left multiplying (2.4.19) with L the matrix of left eigenvectors to obtain the characteristic variable $\mathbf{w}$. ENO discretization is applied to $\mathbf{w}$, then the approximation is projected back the the physical field by multiplying with R the matrix of right eigenvectors. For a detailed algorithm, we refer to [124].

3

WENO and the Adaptive WENO-θ Schemes

In the previous chapter, we have presented numerical approaches to conservation laws so that oscillations near discontinuities can be controlled in some sense. For TVD schemes, this goal is achieved through limiters, whereas for ENO schemes, a stencil-wise polynomial is reconstructed on the smoothest stencil so that it is essentially non-oscillatory. The smoothness depends on a comparison of the undivided Newton differences. This allows ENO schemes to achieve much higher order of accuracy than the TVD ones, which usually have $2nd$-order.

The limitation of ENO schemes is that it is costly and hard to analyze due to lots of "if" statements, and somehow inefficient since only one of the stencil candidates is chosen, i.e., the smoothest one. For an rth-order ENO scheme, only the smoothest stencil is chosen among r candidates to approximate the numerical flux. The smoothness of the solution on each stencil is determined by an indicator of smoothness. Later on, Liu, Osher, and Chan ([92]) upgraded ENO schemes and introduced the Weighted ENO (WENO) by combining all stencil candidates (hereafter sub-stencils) in the numerical flux approximation. Here, a nonlinear weight is assigned to each sub-stencil to control its contribution in the procedure. WENO schemes maintain the essentially non-oscillatory property of the ENO near discontinuities and outperform the latter in smooth regions where the accuracy order is increased to $(r+1)$th-order if r sub-stencils are used. Consequently, Jiang and Shu (see [78], also [122], and the review [125]) constructed WENO schemes in the framework of finite difference and further improved the order to $(2r-1)$th in smooth regions by introducing a new class of smoothness indicators. Hereafter, we denote WENO-JS for the 5th-order finite difference WENO developed in [78]. In [12], [135] higher order than 5th-order WENO schemes are given.

Since the introduction of WENO, many improvements and derivatives of the schemes have been developed and introduced. Henrick et al. in [50] carefully analyzed the sufficient conditions of the nonlinear weights and found that WENO-JS does not achieve the designed 5th-order but reduces to only 3rd-order in case the first and third derivatives of the flux are simultaneously zero (e.g., $f'(x_j) = 0$ but $f'''(x_j) \neq 0$ for the scalar case of Eq. (3.1.1) below). They then suggested an improved version which is called mapped WENO, abbreviated by WENO-M. By using a mapping on the nonlinear weights, WENO-M satisfies the sufficient condition on which WENO-JS fails and obtains optimal order near simple smooth extrema. In a different approach on the construction of the nonlinear weights, in [8] Borges et al. introduced the 5th-order WENO-Z scheme. Here, the authors also measured the smoothness of the large stencil which comprises all sub-stencils and incorporated this in devising the new smoothness indicators and nonlinear weights. It was proven numerically that WENO-Z is less dissipative than WENO-JS and more efficient than WENO-M, respectively. It was also checked that WENO-Z attains 4th-order near simple smooth extrema comparing with 3rd-order of WENO-JS. For higher order WENO-Z schemes, we refer readers to [13]. Another approach to improve WENO schemes is the new designs of the smoothness indicators. In [53], L^1-norm based smoothness indicators are suggested, and the ones devised from Lagrange interpolation polynomials are given in [31], and [35]. See also [34] for a new mapped WENO scheme.

3.1 Overview

3.1.1 Finite Difference Schemes

For simplicity, we consider Eq. (2.1.1) in a scalar case. We write the equation as below,

$$\begin{cases} u_t + f(u)_x = 0, \quad x \in [x_l, x_r], \\ u(x,0) = u_0(x). \end{cases} \tag{3.1.1}$$

For simplicity, periodic boundary conditions, i.e., $u(x_l, t) = u(x_r, t)$, are applied.

We denote $h(x)$ the flux function defined as follows,

$$f(u(x,\cdot)) = \bar{h}(x) := \frac{1}{\Delta x} \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} h(y)dy. \tag{3.1.2}$$

Evaluating Eq. (3.1.1) at grid point x_j, we obtain the semi-discrete form as follows,

$$\frac{du_j}{dt} = -\frac{\partial f}{\partial x}\bigg|_{x=x_j} = -\frac{h(x_{j+\frac{1}{2}}) - h(x_{j-\frac{1}{2}})}{\Delta x} =: \mathcal{L}(u). \tag{3.1.3}$$

It is noticed that Eq. (3.1.3) is exact since there are no approximating errors in the formula.

Approximating Eq. (3.1.3) by

$$\frac{dv_j(t)}{dt} = -\frac{\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}}{\Delta x} =: \hat{\mathcal{L}}(v), \tag{3.1.4}$$

where the numerical flux $\hat{f}(x)$ is an approximation of the flux $h(x)$. We notice that $v_j(t)$ here is a pointwise approximation of $u(x_j, t)$. Hence, the scheme is finite difference.

We say that a scheme is of pth order of accuracy providing that

$$\hat{\mathcal{L}}(v) = -f(u(x,t))_x\big|_{x=x_j} + \mathcal{O}(\Delta x^p). \tag{3.1.5}$$

3.1.2 Time Integration

We first mention about time advancing for Eq. (3.1.3). Following [43] and the references therein, for all below WENO schemes, we apply the 3rd-order TVD Runge-Kutta method as below. For TVD, we mean that the time integrator follows the same property in the previous chapter. Here, Δt is the time step satisfying some proper CFL condition.

$$\begin{aligned} v^{(1)} &= v^n + \Delta t\hat{\mathcal{L}}(v^n), \\ v^{(2)} &= \frac{3}{4}v^n + \frac{1}{4}v^{(1)} + \frac{1}{4}\Delta t\hat{\mathcal{L}}(v^{(1)}), \\ v^{n+1} &= \frac{1}{3}v^n + \frac{2}{3}v^{(2)} + \frac{2}{3}\Delta t\hat{\mathcal{L}}(v^{(2)}), \end{aligned} \tag{3.1.6}$$

where $\hat{\mathcal{L}}(u)$ obtained from some method is an approximation of the spatial operator $\mathcal{L}(u)$. In particular, see the below WENO discretizations $\hat{\mathcal{L}}^5(u)$ in Eq. (3.2.13) where $\hat{f}^5_{j+\frac{1}{2}}$ follows Eq. (3.2.14) and $\hat{\mathcal{L}}^6(u)$ in Eq. (3.3.5) where $\hat{f}^6_{j+\frac{1}{2}}$ is defined in Eq. (3.3.6) with γ_k's replaced by ω_k's.

We now proceed to the discussions on the spatial discretizations.

3.2 5th-order Upwind WENO Schemes

3.2.1 Reconstruction

We notice that for simplicity, we can assume that $f'(u) \geq 0$ over the whole computational domain. In case there is a change in signs of $f'(u)$, a flux splitting technique is invoked. We discuss this in remark 3.2.1 below.

Originally, WENO schemes were constructed in the context of finite volume (see [92]). Thanks to lemma 3.1 given in [125], the schemes can be transformed into finite difference through relation (3.1.2). The $\bar{h}_j$ is called an average value of the numerical flux $h(x)$ over the interval I_j. We then seek for an approximating polynomial $\hat{f}^5(x)$ of degree four of $h(x)$ as below

$$h(x) \approx \hat{f}^5(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4, \tag{3.2.1}$$

over the large stencil $S^5 = \{x_{j-2}, x_{j-1}, x_j, x_{j+1}, x_{j+2}\}$. We note that S^5 is chosen biased to the left with respect to the point $x_{j+\frac{1}{2}}$ for the stability purpose. Hence, the scheme is in an upwind sense.

Substituting the polynomial in Eq. (3.2.1) into Eq. (3.1.2) and evaluating at

$$x_k = (k-j)\Delta x, \quad k = j-2, \dots, j+2, \ \text{ around } x_j = 0, \tag{3.2.2}$$

we obtain that

$$\begin{cases} \frac{1}{\Delta x}\int_{-\frac{5\Delta x}{2}}^{-\frac{3\Delta x}{2}} (a_0 + a_1 y + a_2 y^2 + a_3 y^3 + a_4 y^4) dy = f_{j-2}, \\ \frac{1}{\Delta x}\int_{-\frac{3\Delta x}{2}}^{-\frac{\Delta x}{2}} (a_0 + a_1 y + a_2 y^2 + a_3 y^3 + a_4 y^4) dy = f_{j-1}, \\ \frac{1}{\Delta x}\int_{-\frac{\Delta x}{2}}^{\frac{\Delta x}{2}} (a_0 + a_1 y + a_2 y^2 + a_3 y^3 + a_4 y^4) dy = f_{j}, \\ \frac{1}{\Delta x}\int_{\frac{\Delta x}{2}}^{\frac{3\Delta x}{2}} (a_0 + a_1 y + a_2 y^2 + a_3 y^3 + a_4 y^4) dy = f_{j+1}, \\ \frac{1}{\Delta x}\int_{\frac{3\Delta x}{2}}^{\frac{5\Delta x}{2}} (a_0 + a_1 y + a_2 y^2 + a_3 y^3 + a_4 y^4) dy = f_{j+2}. \end{cases} \tag{3.2.3}$$

We then obtain the Vandermonde matrix of the coefficients a_k's as follows,

$$M = \begin{bmatrix} 1 & -2\Delta x & \frac{49}{12}\Delta x^2 & -\frac{17}{2}\Delta x^3 & \frac{1441}{80}\Delta x^4 \\ 1 & -\Delta x & \frac{13}{12}\Delta x^2 & -\frac{5}{4}\Delta x^3 & \frac{121}{80}\Delta x^4 \\ 1 & 0 & \frac{1}{12}\Delta x^2 & 0 & \frac{1}{80}\Delta x^4 \\ 1 & \Delta x & \frac{13}{12}\Delta x^2 & \frac{5}{4}\Delta x^3 & \frac{121}{80}\Delta x^4 \\ 1 & 2\Delta x & \frac{49}{12}\Delta x^2 & \frac{17}{2}\Delta x^3 & \frac{1441}{80}\Delta x^4 \end{bmatrix}. \tag{3.2.4}$$

By solving the linear system

$$M\hat{\mathbf{c}} = \mathbf{f}, \tag{3.2.5}$$

where $\hat{\mathbf{c}}$ is the column vector of the unknown coefficients, $\mathbf{f}$ is the grid values of the flux function $f(u)$ given in the right-hand side of Eq. (3.2.3), the reconstruction polynomial $\hat{f}^5(x)$ is uniquely determined with the coefficients given as follows,

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} \frac{3}{640}f_{j-2} - \frac{29}{480}f_{j-1} + \frac{1067}{960}f_j - \frac{29}{480}f_{j+1} + \frac{3}{640}f_{j+2} \\ \frac{5}{48\Delta x}f_{j-2} - \frac{17}{24\Delta x}f_{j-1} + \frac{17}{24\Delta x}f_{j+1} - \frac{5}{48\Delta x}f_{j+2} \\ -\frac{1}{16\Delta x^2}f_{j-2} + \frac{3}{4\Delta x^2}f_{j-1} - \frac{11}{8\Delta x^2}f_j + \frac{3}{4\Delta x^2}f_{j+1} - \frac{1}{16\Delta x^2}f_{j+2} \\ -\frac{1}{12\Delta x^3}f_{j-2} + \frac{1}{6\Delta x^3}f_{j-1} - \frac{1}{6\Delta x^3}f_{j+1} + \frac{1}{12\Delta x^3}f_{j+2} \\ \frac{1}{24\Delta x^4}f_{j-2} - \frac{1}{6\Delta x^4}f_{j-1} + \frac{1}{4\Delta x^4}f_j - \frac{1}{6\Delta x^4}f_{j+1} + \frac{1}{24\Delta x^4}f_{j+2} \end{bmatrix}. \tag{3.2.6}$$

Evaluating $\hat{f}^5(x)$ at $x_{j+\frac{1}{2}} = \Delta x/2$ gives us,

$$\hat{f}^5_{j+\frac{1}{2}} = \frac{2}{60} f_{j-2} - \frac{13}{60} f_{j-1} + \frac{47}{60} f_j + \frac{27}{60} f_{j+1} - \frac{3}{60} f_{j+2}. \tag{3.2.7}$$

Here, we recall $f_k = f(u_k) = f(u(x_k, \cdot))$.

Since the procedure is via the average $\bar{h}_k = \bar{h}(x_k) = f(u(x_k, \cdot))$ in Eq. (3.1.2), we call this the reconstruction and $\hat{f}^5(x)$ is the reconstruction polynomial.

To justify the approximation error, we denote the polynomial $H(x)$ such that

$$H'(x) = h(x). \tag{3.2.8}$$

We then deduce from Eq. (3.1.2) that

$$f(u(x, \cdot)) = \frac{1}{\Delta x} \int_{x-\frac{\Delta x}{2}}^{x+\frac{\Delta x}{2}} H'(y) dy = \frac{H(x + \frac{\Delta x}{2}) - H(x - \frac{\Delta x}{2})}{\Delta x}. \tag{3.2.9}$$

Substituting Eq. (3.2.9) into the approximation (3.2.7), evaluating at x_k, $k = j-2, \ldots, j+2$, and applying Taylor expansions of $H(x)$ at $x = x_{j+\frac{1}{2}}$, we obtain the following truncation error,

$$\begin{aligned} \hat{f}^5_{j+\frac{1}{2}} &= \frac{1}{60\Delta x}(-2H_{j-\frac{5}{2}} + 15H_{j-\frac{3}{2}} - 60H_{j-\frac{1}{2}} + 20H_{j+\frac{1}{2}} + 30H_{j+\frac{3}{2}} - 3H_{j+\frac{5}{2}}) \\ &= H'_{j+\frac{1}{2}} - \frac{1}{60} \frac{d^6 H}{dx^6}\bigg|_{x=x_{j+\frac{1}{2}}} \Delta x^5 + \mathcal{O}(\Delta x^6) = h_{j+\frac{1}{2}} - \frac{1}{60} \frac{\partial^5 f}{\partial x^5}\bigg|_{x=x_j} \Delta x^5 + \mathcal{O}(\Delta x^6). \end{aligned} \tag{3.2.10}$$

The last equality in Eq. (3.2.10) is justified as follows. Thanks to the relation in (3.2.9), by a Taylor expansion around $x_{j+\frac{1}{2}}$ we have that

$$\frac{\partial^5 f}{\partial x^5}\bigg|_{x=x_j} = \frac{1}{\Delta x}\left[\frac{d^5 H}{dx^5}\bigg|_{x=x_{j+\frac{1}{2}}} - \frac{d^5 H}{dx^5}\bigg|_{x=x_{j+\frac{1}{2}}-\Delta x} \right] = \frac{d^6 H}{dx^6}\bigg|_{x=x_{j+\frac{1}{2}}} + \mathcal{O}(\Delta x). \tag{3.2.11}$$

Together with Eq. (3.2.8), we deduce the last equality in Eq. (3.2.10).

Similarly, we have

$$\begin{aligned} \hat{f}^5_{j-\frac{1}{2}} &= \frac{2}{60} f_{j-3} - \frac{13}{60} f_{j-2} + \frac{47}{60} f_{j-1} + \frac{27}{60} f_j - \frac{3}{60} f_{j+1} \\ &= h_{j-\frac{1}{2}} - \frac{1}{60} \frac{\partial^5 f}{\partial x^5}\bigg|_{x=x_j} \Delta x^5 + \mathcal{O}(\Delta x^6). \end{aligned} \tag{3.2.12}$$

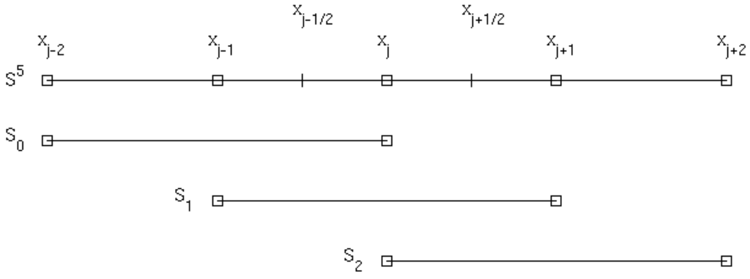


Figure 3-1: Stencils for 5th-order WENO schemes.

Hence, we have

$$\frac{du_j}{dt} \approx -\frac{\hat{f}^5_{j+\frac{1}{2}} - \hat{f}^5_{j-\frac{1}{2}}}{\Delta x} =: \hat{\mathcal{L}}^5(u). \tag{3.2.13}$$

The scheme is 5th-order of accuracy in space.

For non-smooth solutions, we employ WENO reconstruction. The idea of WENO schemes is that, instead of the 5-point stencil S^5, a convex combination of three 3-point sub-stencils are facilitated for an adaptive choice of candidates for the reconstruction. That is,

$$\hat{f}_{j+\frac{1}{2}} = \sum_{k=0}^{2} \omega_k \hat{f}^k_{j+\frac{1}{2}}, \tag{3.2.14}$$

where $\hat{f}^k_{j+\frac{1}{2}}$'s are defined below and ω_k is the non-linear weight satisfying $\omega_k \geq 0,\ \forall k$ and

$$\sum_{k=0}^{2} \omega_k = 1. \tag{3.2.15}$$

The necessity of non-negative nonlinear weights is discussed in [95] and in [131] for practical implementations. And $\hat{f}^k_{j+\frac{1}{2}}$ is the approximation of $h_{j+\frac{1}{2}}$ by the reconstruction polynomial $\hat{f}^k(x) = b_0 + b_1 x + b_2 x^2$ over the sub-stencil $S_k,\ k = 0, 1, 2$. Here, $S_0 = \{x_{j-2}, x_{j-1}, x_j\}$, $S_1 = \{x_{j-1}, x_j, x_{j+1}\}$, and $S_2 = \{x_j, x_{j+1}, x_{j+2}\}$ (see Fig. 3-1). Carrying a similar process as for the large stencil S^5, we find that around $x_j = 0$,

$$\hat{f}^0(x) = \frac{-f_{j-2} + 2f_{j-1} + 23f_j}{24} + \left(\frac{f_{j-2} - 4f_{j-1} + 3f_j}{2\Delta x}\right) x + \left(\frac{f_{j-2} - 2f_{j-1} + f_j}{2\Delta x^2}\right) x^2, \tag{3.2.16}$$

$$\hat{f}^1(x) = \frac{-f_{j-1} + 26f_j - f_{j+1}}{24} + \left(\frac{f_{j+1} - f_{j-1}}{2\Delta x}\right) x + \left(\frac{f_{j-1} - 2f_j + f_{j+1}}{2\Delta x^2}\right) x^2, \tag{3.2.17}$$

$$\hat{f}^2(x) = \frac{23f_j + 2f_{j+1} - f_{j+2}}{24} + \left(\frac{-3f_j + 4f_{j+1} - f_{j+2}}{2\Delta x}\right) x + \left(\frac{f_j - 2f_{j+1} + f_{j+2}}{2\Delta x^2}\right) x^2. \tag{3.2.18}$$

Evaluating each of these $\hat{f}^k(x)$'s at $x_{j+\frac{1}{2}}$, we obtain that

$$\hat{f}^0_{j+\frac{1}{2}} = \frac{2}{6}f_{j-2} - \frac{7}{6}f_{j-1} + \frac{11}{6}f_j, \tag{3.2.19}$$

$$\hat{f}^1_{j+\frac{1}{2}} = -\frac{1}{6}f_{j-1} + \frac{5}{6}f_j + \frac{2}{6}f_{j+1}, \tag{3.2.20}$$

$$\hat{f}^2_{j+\frac{1}{2}} = \frac{2}{6}f_j + \frac{5}{6}f_{j+1} - \frac{1}{6}f_{j+2}. \tag{3.2.21}$$

Carrying a similar process as given in Eqs. (3.2.10) - (3.2.11) with $\hat{f}^k_{j+\frac{1}{2}}$ replacing $\hat{f}^5_{j+\frac{1}{2}}$, we obtain that

$$h_{j+\frac{1}{2}} = \hat{f}^k_{j+\frac{1}{2}} + \mathcal{O}(\Delta x^3). \tag{3.2.22}$$

Comparing between $\hat{f}^5_{j+\frac{1}{2}}$ given in Eq. (3.2.7) and the ones in Eqs. (3.2.19) - (3.2.21), we deduce the following linear relation

$$\hat{f}^5_{j+\frac{1}{2}} = \sum_{k=0}^{2} \gamma_k \hat{f}^k_{j+\frac{1}{2}}, \tag{3.2.23}$$

where

$$\gamma_0 = \frac{1}{10}, \quad \gamma_1 = \frac{6}{10}, \quad \gamma_2 = \frac{3}{10}, \tag{3.2.24}$$

are called the linear (optimal) weights. We note that

$$\sum_{k=0}^{2} \gamma_k = 1. \tag{3.2.25}$$

Adding and subtracting $\sum_{k=0}^{2} \gamma_k \hat{f}^k_{j+\frac{1}{2}}$ into and from Eq. (3.2.14), thanks to the truncation errors in Eqs. (3.2.10), (3.2.22), the normalization in Eqs. (3.2.15), (3.2.25), and the linear relation (3.2.23) we obtain that

$$\begin{aligned}
\hat{f}_{j\pm\frac{1}{2}} &= \sum_{k=0}^{2}(\omega_k^{\pm} - \gamma_k^{\pm})\hat{f}^k_{j\pm\frac{1}{2}} + \sum_{k=0}^{2}\gamma_k^{\pm}\hat{f}^k_{j\pm\frac{1}{2}} \\
&= \sum_{k=0}^{2}(\omega_k^{\pm} - \gamma_k^{\pm})(h_{j\pm\frac{1}{2}} + \mathcal{O}(\Delta x^3)) + \left(h_{j\pm\frac{1}{2}} - \frac{1}{60}\frac{\partial^5 f}{\partial x^5}\bigg|_{x=x_j}\Delta x^5 + \mathcal{O}(\Delta x^6)\right) \\
&= h_{j\pm\frac{1}{2}} - \frac{1}{60}\frac{\partial^5 f}{\partial x^5}\bigg|_{x=x_j}\Delta x^5 + \sum_{k=0}^{2}(\omega_k^{\pm} - \gamma_k^{\pm})\mathcal{O}(\Delta x^3) + \mathcal{O}(\Delta x^6),
\end{aligned} \tag{3.2.26}$$

where $\gamma_k^{\pm}$, $\omega_k^{\pm}$ are the linear and non-linear weights of the sub-stencils $S_k^{j\pm\frac{1}{2}}$ corresponding to the interfaces $x_{j\pm\frac{1}{2}}$, respectively.

Hence, in order that the discretization in Eq. (3.2.13) where $\hat{f}^5_{j\pm\frac{1}{2}} = \hat{f}_{j\pm\frac{1}{2}}$ follow the nonlinear relation (3.2.14) to be 5th-order, we deduce the sufficient condition for the nonlinear weights as follows,

$$\omega_k^{\pm} - \gamma_k^{\pm} = \mathcal{O}(\Delta x^3), \quad \forall k. \tag{3.2.27}$$

Different WENO schemes depends on how these nonlinear weights and the smoothness indicators are defined. The latter ones are introduced in the below section. In the following subsections, we summarize the 5th-order upwind and 6th-order central WENO schemes discussed previously.

3.2.2 WENO-JS

In [78], Jiang and Shu defined the nonlinear weights as follows,

$$\omega_k^{JS} = \frac{\alpha_k^{JS}}{\sum_{l=0}^{2}\alpha_l^{JS}}, \quad \alpha_k^{JS} = \frac{\gamma_k}{(\varepsilon + \beta_k)^p}, \tag{3.2.28}$$

where γ_k is defined in Eq. (3.2.24), β_k is called the smoothness indicator of S_k which measures how smooth the solution is over this sub-stencil. The authors defined these β_k's through the normalized L^2-norm of high-order variations of the reconstruction polynomials given in Eqs. (3.2.19) - (3.2.21). Explicitly, for a 5th-order scheme, we have

$$\beta_k = \Delta x \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left(\frac{d}{dx}\hat{f}^k(x)\right)^2 dx + \Delta x^3 \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left(\frac{d^2}{dx^2}\hat{f}^k(x)\right)^2 dx, \tag{3.2.29}$$

where $\hat{f}^k(x)$'s are as in Eqs. (3.2.16) - (3.2.18) and the sub-stencils S_0, S_1, S_2 are centered around $x_j = 0$.

Evaluating for each k, we obtain that

$$\begin{aligned}\beta_0 &= \frac{13}{12}(f_{j-2} - 2f_{j-1} + f_j)^2 + \frac{1}{4}(f_{j-2} - 4f_{j-1} + 3f_j)^2 \\ &= f'^2\Delta x^2 + \left(\frac{13}{12}f''^2 - \frac{2}{3}f'f'''\right)\Delta x^4 + \mathcal{O}(\Delta x^5),\end{aligned} \tag{3.2.30}$$

$$\begin{aligned}\beta_1 &= \frac{13}{12}(f_{j-1} - 2f_j + f_{j+1})^2 + \frac{1}{4}(f_{j+1} - f_{j-1})^2 \\ &= f'^2\Delta x^2 + \left(\frac{13}{12}f''^2 + \frac{1}{3}f'f'''\right)\Delta x^4 + \mathcal{O}(\Delta x^6),\end{aligned} \tag{3.2.31}$$

$$\begin{aligned}\beta_2 &= \frac{13}{12}(f_j - 2f_{j+1} + f_{j+2})^2 + \frac{1}{4}(3f_j - 4f_{j+1} + f_{j+2})^2 \\ &= f'^2\Delta x^2 + \left(\frac{13}{12}f''^2 - \frac{2}{3}f'f'''\right)\Delta x^4 + \mathcal{O}(\Delta x^5),\end{aligned} \tag{3.2.32}$$

where the derivatives are evaluated at $x = x_j$.

In formula (3.2.28), ε is a small parameter to prevent zero division. In most cases, WENO-JS works well with $\varepsilon = 10^{-6}$. A thorough analysis of the role of ε can be found in [50]. The parameter p is to increase the dissipation of the scheme. For WENO-JS, $p = 2$ is chosen.

If $f'_j = f'(x_j) \neq 0$, $\forall k$, β_k can be written in the form

$$\beta_k = (f'_j\Delta x)^2(1 + \mathcal{O}(\Delta x^2)). \tag{3.2.33}$$

Substituting these into Eq. (3.2.28) with the removal of ε, since $(1+y)^{-2} = 1 + \mathcal{O}(y)$, by Eq. (3.2.25) we obtain that

$$\omega_k^{JS} = \frac{\gamma_k(f'_j\Delta x)^{-2}(1 + \mathcal{O}(\Delta x^2))}{(f'_j\Delta x)^{-2}\sum_{l=0}^{2}\gamma_l(1 + \mathcal{O}(\Delta x^2))} = \gamma_k + \mathcal{O}(\Delta x^2), \tag{3.2.34}$$

which is a relaxed form of (3.2.27). We notice that for WENO-JS, ω_k^{JS} cannot satisfy condition (3.2.27) directly. Moreover, if $f'_j = 0$, it is observed from Eqs. (3.2.30) - (3.2.32) that $\beta_k = \frac{13}{12}(f''_j)^2\Delta x^4(1 + \mathcal{O}(\Delta x))$ for $k = 0, 2$ and $\beta_1 = \frac{13}{12}(f''_j)^2\Delta x^4(1 + \mathcal{O}(\Delta x^2))$, in which the condition (3.2.33) is not satisfied for all k's. Similarly, we find that

$$\omega_k^{JS} = \gamma_k + \mathcal{O}(\Delta x), \tag{3.2.35}$$

which is a loss of accuracy near this critical point. Numerically checks show that with $\varepsilon = 10^{-40}$, WENO-JS is only 3rd-order near simple smooth critical points, i.e., $f'_j = 0 \neq f_j^{(k)}$, $k \geq 2$. This accuracy loss is improved by the mapped WENO, and later by the WENO-Z scheme presented in the next subsections.

3.2.3 Mapped WENO

Henrick et al. ([50]) proposed a method to improve the relaxed condition (3.2.34) of WENO-JS by introducing the mapping function, for $k = 0, 1, 2$,

$$g_k(\omega) = \frac{\omega(\gamma_k + \gamma_k^2 - 3\gamma_k\omega + \omega^2)}{\gamma_k^2 + \omega(1 - 2\gamma_k)}, \tag{3.2.36}$$

where γ_k is the linear weight in Eq. (3.2.24).

All of these functions are monotonically increasing with $g_k(0) = 0$, $g_k(1) = 1$, $g_k(\gamma_k) = \gamma_k$, and they are flat up to $2nd$-order, i.e.,

$$g_k'(\gamma_k) = g_k''(\gamma_k) = 0 \tag{3.2.37}$$

about the optimal value.

The new non-linear weights are modified as

$$\omega^M = \frac{\alpha_k^M}{\sum_{l=0}^{2} \alpha_l^M}, \quad \alpha_k^M = g_k(\omega_k^{JS}), \tag{3.2.38}$$

whereas the other procedures follows exactly as those of the WENO-JS scheme.

We check that by the mapping functions $g_k(x)$, condition (3.2.27) is satisfied directly. By a Taylor expansion abut γ_k of $g_k(\omega)$, we deduce that

$$\begin{aligned}\alpha_k^M &= g_k(\gamma_k) + g_k'(\gamma_k)\left(\omega_k^{JS} - \gamma_k\right) + \frac{1}{2}g''\left(\omega_k^{JS} - \gamma_k\right)^2 + \frac{1}{6}g'''\left(\omega_k^{JS} - \gamma_k\right)^3 + \mathcal{O}(\omega_k^{JS} - \gamma_k)\\ &= \gamma_k + \mathcal{O}(\left(\omega_k^{JS} - \gamma_k\right)^3)\\ &= \gamma_k + \mathcal{O}(\Delta x^3),\end{aligned} \tag{3.2.39}$$

thanks to condition (3.2.37) and the observation (3.2.35). It implies that

$$\omega_k^M = \gamma_k + \mathcal{O}(\Delta x^3), \tag{3.2.40}$$

which is condition (3.2.27).

The disadvantage of mapped WENO is that the scheme is more expensive since the non-linear weights are computed twice, one for those of WENO-JS and the other for the mapping, and it is not easy to generalize. We discuss the WENO-Z scheme below that somehow overcomes these difficulties.

3.2.4 WENO-Z

Borges et al. in [8] proposed a new WENO-Z. In their scheme, the nonlinear weights are defined differently from those of WENO-JS. They are as follows,

$$\omega_k^Z = \frac{\alpha_k^Z}{\sum_{l=0}^{2} \alpha_l^Z}, \quad \alpha_k^Z = \gamma_k \left(1 + \left(\frac{\tau^Z}{\varepsilon + \beta_k}\right)^q\right), \tag{3.2.41}$$

where the smoothness indicators β_k's are the same as those given in Eqs. (3.2.30) - (3.2.32), $\varepsilon = 10^{-40}$, and τ^Z is the smoothness indicator of S^5. Power q is used to tune the relation between the dispersive and dissipative property of the scheme. It is checked numerically in [8] that the scheme becomes more dissipative when q is increased. For WENO-Z, τ^Z is defined as follows,

$$\tau^Z = |\beta_0 - \beta_2| = \frac{13}{3}|f'' f'''|\Delta x^5 + \mathcal{O}(\Delta x^6). \tag{3.2.42}$$

We note that if $f_j' \neq 0$ and $q = 1$, $\beta_k = \mathcal{O}(\Delta x^2)$, $\forall k$. Then

$$\frac{\tau^Z}{\beta_k} = \mathcal{O}(\Delta x^3), \quad \forall k. \tag{3.2.43}$$

Similarly to Eq. (3.2.34), we obtain that

$$\omega_k^Z = \gamma_k + \mathcal{O}(\Delta x^3), \tag{3.2.44}$$

directly without using the relaxed version as the WENO-JS scheme. It was also proven in [8] that WENO-Z is 4th-order near simple smooth critical points (i.e. where $f_j' = 0$) for $q = 1$ and attains the designed 5th-order for $q = 2$. The tradeoff for the latter case is that the scheme is more dissipative. Throughout this work, we choose $q = 1$. One more advantage of WENO-Z over WENO-JS is that the former is more central in a sense that the stencil over which the solution is discontinuous plays more roles in the approximation of the numerical flux. This assessment is checked as follows. We suppose that S_2 contains a discontinuity whereas the solution is smooth over the other two sub-stencils. Hence, $\tau^Z = \mathcal{O}(1)$, $\beta_2 = \mathcal{O}(1)$, and $\beta_k = \mathcal{O}(\Delta x^2)$, $k = 0, 1$. Then,

$$\frac{\alpha_2^Z}{\alpha_k^Z} = \frac{\gamma_2\left(1 + \frac{\tau^Z}{\beta_2}\right)}{\gamma_k\left(1 + \frac{\tau^Z}{\beta_k}\right)} = \frac{\frac{\gamma_2}{\beta_2}(\beta_2 + \tau^Z)}{\frac{\gamma_k}{\beta_k}(\beta_k + \tau^Z)} = \frac{\alpha_2^{JS}(\beta_2 + \tau^Z)}{\alpha_k^{JS}(\beta_k + \tau^Z)} \geq \frac{\alpha_2^{JS}}{\alpha_k^{JS}}, \tag{3.2.45}$$

since

$$\frac{\beta_2 + \tau^Z}{\beta_k + \tau^Z} \approx \frac{\beta_2 + \tau^Z}{\tau^Z} \geq 1. \tag{3.2.46}$$

Hence, WENO-Z has a sharper capturing of discontinuities than WENO-JS.

Remark 3.2.1.

i. In case the condition $f'(u) \geq 0$ is not satisfied, which is general in real applications, we apply a flux splitting technique to decompose $f(u)$ into positive and negative components. The global Lax-Friedrichs flux splitting is used in most applications (see [78] and the references therein),

$$f^{\pm}(u) = \frac{1}{2}(f(u) \pm \alpha u), \tag{3.2.47}$$

where $\alpha = \max |f'(u)|$ over the whole computational domain. Then,

$$\hat{f}_{j+\frac{1}{2}} = \hat{f}^{+}_{j+\frac{1}{2}} + \hat{f}^{-}_{j+\frac{1}{2}}. \tag{3.2.48}$$

The negative flux $\hat{f}^{-}_{j+\frac{1}{2}}$ is symmetric to $\hat{f}^{+}_{j+\frac{1}{2}}$ with respect to $x_{j+\frac{1}{2}}$.

ii. If the flux splitting is employed, the overall number of grid points used in the reconstruction of the numerical flux is increased by one. That is, let S^{5+} and S^{5-} be the stencils over which $\hat{f}^{+}_{j+\frac{1}{2}}$ and $\hat{f}^{-}_{j+\frac{1}{2}}$ are determined, then

$$S^6 := S^{5+} \bigcup S^{5-} = \{x_{j-2}, x_{j-1}, x_j, x_{j+1}, x_{j+2}, x_{j+3}\}, \tag{3.2.49}$$

which consists of six points. We note that with this S^6, there exists a polynomial of degree five which reconstructs $h(x)$ over the stencil. Therefore, the accuracy of WENO schemes can be increased up to sixth. These schemes are discussed in the subsection below.

3.3 6th-order Central WENO Schemes

As indicated in the previous remark, for a general flux where the signs of the eigenvalues of the Jacobian $A(\mathbf{u})$ are not uniform throughout the domain, a flux splitting technique, for example, the global or local Lax-Friedrichs or the Roe with entropy fix (see [78] and the references therein) is needed. This increases the number of grid points in the numerical flux approximating procedure by one. We take the 5th-order WENO-JS scheme for example, the total number of grid points used in the reconstruction for both positive and negative fluxes will be six instead of five. We also note that with these six points, one can indeed improve the scheme up to 6th-order in smooth regions. The difficulty of this approach lies in the dispersive nature of a central scheme if six points are employed. In this case, oscillations are expected to occur near discontinuities. In [157] (see also [16] for the boundary condition treatment), Yamaleev and Carpenter for the first time introduced a 6th-order WENO scheme by adding one more sub-stencil into the numerical flux approximation. We denote this scheme WENO-NW6. For this most downwind sub-stencil, an *ad hoc* treatment on the smoothness indicator β_3 was suggested. The idea is originated from that of Martín et al. in [102]. In order that oscillations do not happen, β_3 is computed using the information of $\mathbf{f}(\mathbf{u})$ on all grid points of the large stencil, i.e, six points. Hence, the sub-stencil only plays roles in case the solution is smooth over this large stencil. In a similar manner, recently Hu, Wang, and Adams in [64] proposed an adaptive central-upwind WENO-CU6 scheme which switches between a 5th-order upwind and 6th-order central WENO scheme automatically. The difference of their work from that given in [157] is that β_3 is defined via a Lagrange interpolating polynomial of degree five over the large stencil. In [49], the authors successfully applied WENO-CU6 in the LES simulation of scale separation. Other hybrid WENO schemes can be found in, for examples, [14], or [93], [61], etc.

3.3.1 Reconstruction

Carrying a similar process with instead stencil S^6, we obtain that

$$\hat{f}^6(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5, \tag{3.3.1}$$

where the coefficients are

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_5 \\ a_5 \end{bmatrix} = \begin{bmatrix} \frac{3}{640} f_{j-2} - \frac{29}{480} f_{j-1} + \frac{1067}{960} f_j - \frac{29}{480} f_{j+1} + \frac{3}{640} f_{j+2} \\ \frac{341}{5760\Delta x} f_{j-2} - \frac{557}{1152\Delta x} f_{j-1} - \frac{259}{576\Delta x} f_j + \frac{667}{576\Delta x} f_{j+1} - \frac{379}{1152\Delta x} f_{j+2} + \frac{259}{5760\Delta x} f_{j+3} \\ -\frac{1}{16\Delta x^2} f_{j-2} + \frac{3}{4\Delta x^2} f_{j-1} - \frac{11}{8\Delta x^2} f_j + \frac{3}{4\Delta x^2} f_{j+1} - \frac{1}{16\Delta x^2} f_{j+2} \\ -\frac{5}{144\Delta x^3} f_{j-2} - \frac{11}{144\Delta x^3} f_{j-1} + \frac{35}{72\Delta x^3} f_j - \frac{47}{72\Delta x^3} f_{j+1} + \frac{47}{144\Delta x^3} f_{j+2} - \frac{7}{144\Delta x^3} f_{j+3} \\ \frac{1}{24\Delta x^4} f_{j-2} - \frac{1}{6\Delta x^4} f_{j-1} + \frac{1}{4\Delta x^4} f_j - \frac{1}{6\Delta x^4} f_{j+1} + \frac{1}{24\Delta x^4} f_{j+2} \\ -\frac{1}{120\Delta x^5} f_{j-2} + \frac{1}{24\Delta x^5} f_{j-1} - \frac{1}{12\Delta x^5} f_j + \frac{1}{12\Delta x^5} f_{j+1} - \frac{1}{24\Delta x^5} f_{j+2} + \frac{1}{120\Delta x^5} f_{j+3} \end{bmatrix}. \tag{3.3.2}$$

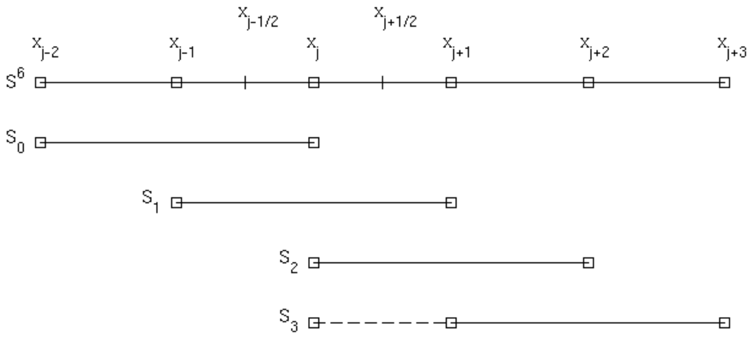


Figure 3-2: Stencils for 6th-order WENO schemes.

Evaluating $\hat{f}^6(x)$ at $x_{j+\frac{1}{2}} = \frac{\Delta x}{2}$ gives us the approximation $\hat{f}^6_{j+\frac{1}{2}}$ as follows,

$$\begin{aligned} \hat{f}^6_{j+\frac{1}{2}} &= \frac{1}{60}f_{j-2} - \frac{8}{60}f_{j-1} + \frac{37}{60}f_j + \frac{37}{60}f_{j+1} - \frac{8}{60}f_{j+2} + \frac{1}{60}f_{j+3} \\ &= h_{j+\frac{1}{2}} + \frac{1}{140}\frac{\partial^6 f}{\partial x^6}\bigg|_{x=x_j} \Delta x^6 + \mathcal{O}(\Delta x^7). \end{aligned} \tag{3.3.3}$$

Similarly,

$$\begin{aligned} \hat{f}^6_{j-\frac{1}{2}} &= \frac{1}{60}f_{j-3} - \frac{8}{60}f_{j-2} + \frac{37}{60}f_{j-1} + \frac{37}{60}f_j - \frac{8}{60}f_{j+1} + \frac{1}{60}f_{j+2} \\ &= h_{j-\frac{1}{2}} + \frac{1}{140}\frac{\partial^6 f}{\partial x^6}\bigg|_{x=x_j} \Delta x^6 + \mathcal{O}(\Delta x^7). \end{aligned} \tag{3.3.4}$$

Hence we obtain that

$$\frac{du_j}{dt} \approx -\frac{\hat{f}^6_{j+\frac{1}{2}} - \hat{f}^6_{j-\frac{1}{2}}}{\Delta x} =: \hat{\mathcal{L}}^6(u). \tag{3.3.5}$$

The scheme is increased to 6th-order of accuracy in space.

Adding one more sub-stencil $S_3 = \{x_{j+1}, x_{j+2}, x_{j+3}\}$ into the approximation of the interfaced value $\hat{f}_{j+\frac{1}{2}}$ (see Fig. 3-2), we deduce a similar linear relation with Eq. (3.2.23) as follows

$$\hat{f}^6_{j+\frac{1}{2}} = \sum_{k=0}^{3} \gamma_k \hat{f}^k_{j+\frac{1}{2}}, \tag{3.3.6}$$

where

$$\gamma_0 = \gamma_3 = \frac{1}{20}, \quad \gamma_1 = \gamma_2 = \frac{9}{20}. \tag{3.3.7}$$

Here, $\hat{f}^3_{j+\frac{1}{2}}$ is the 3rd-order approximation of the numerical flux $h(x)$ at the interface $x_{j+\frac{1}{2}}$ from the reconstruction polynomial $\hat{f}^3(x)$ over the sub-stencil S_3. Explicitly, we have that

$$\begin{aligned}\hat{f}^3(x) = &\left(\frac{71}{24}f_{j+1} - \frac{35}{12}f_{j+2} + \frac{23}{24}f_{j+3}\right) + \frac{1}{\Delta x}\left(-\frac{5}{2}f_{j+1} + 4f_{j+2} - \frac{3}{2}f_{j+3}\right)x \\ &+ \frac{1}{\Delta x^2}\left(\frac{1}{2}f_{j+1} - f_{j+2} + \frac{1}{2}f_{j+3}\right)x^2.\end{aligned} \tag{3.3.8}$$

Evaluating at $x_{j+\frac{1}{2}} = \frac{\Delta x}{2}$, we obtain that

$$\hat{f}^3_{j+\frac{1}{2}} = \frac{11}{6}f_{j+1} - \frac{7}{6}f_{j+2} + \frac{2}{6}f_{j+3}. \tag{3.3.9}$$

The other approximations $\hat{f}^k_{j+\frac{1}{2}}$'s, $k = 0, 1, 2$ follow Eqs. (3.2.19) - (3.2.21).

The nonlinear combination using the nonlinear weights is similar to that of the linear case (3.3.6), except for the linear weight γ_k replaced by its nonlinear version ω_k, $k = 0, \dots, 3$.

Remark 3.3.1.

i. Since the accuracy order is increased to sixth, the sufficient condition for the nonlinear weights given in (3.2.27) is also increased by one. That is,

$$\omega^{\pm}_k - \gamma^{\pm}_k = \mathcal{O}(\Delta x^4),\ \forall k. \tag{3.3.10}$$

ii. Observing from Eq. (3.3.6) that the approximations $\hat{f}_{j+\frac{1}{2}}$'s are now symmetric with respect to $x_{j+\frac{1}{2}}$. It means that the scheme now is central. Hence, spurious oscillations are expected to occur near discontinuities. A treatment on the most downwind sub-stencil S_3 is needed to sustain the ENO property of the scheme. We now overview the 6th-order WENO schemes for this case.

3.3.2 WENO-NW6

In [157], Yamaleev and Carpenter proposed a 6th-order energy-stable WENO scheme. They introduced an artificial dissipative term and proved that this makes the new scheme be stable in L^2 sense. In this work, we only discuss their treatment on the nonlinear weights and omit this artificial dissipative term (see [157] for a detailed discussion on the term). Hence, the scheme here is denoted by WENO-NW6, not ESWENO as in their paper.

The nonlinear weights follow those defined in the WENO-Z scheme which are given in Eq. (3.2.41), for $k = 0, \dots, 3$. The differences lie on the smoothness indicator of the most downwind sub-stencil β_3 and the one for the large stencil S^6. For the former, in order that there are no oscillations occurring near discontinuities, all grid values of the flux over S^6 are accounted for the computation of β_3. It is as follows,

$$\beta_3 = \frac{1}{4}(\beta_0^4 + \beta_1^4 + \beta_2^4 + \tilde{\beta}_3^4)^{1/4}, \tag{3.3.11}$$

where $\tilde{\beta}_3$ is computed using the formula given in Eq. (3.2.29). That is,

$$\begin{aligned}\tilde{\beta}_3 &= \frac{13}{12}(f_{j+1} - 2f_{j+2} + f_{j+3})^2 + \frac{1}{4}(-5f_{j+1} + 8f_{j+2} - 3f_{j+3})^2 \\ &= f'^2\Delta x^2 + \left(\frac{13}{12}f''^2 - \frac{11}{3}f'f'''\right)\Delta x^4 + \mathcal{O}(\Delta x^5),\end{aligned} \tag{3.3.12}$$

where the derivatives are evaluated at $x = x_j$. The other indicators β_k's, $k = 0, 1, 2$ follow Eqs. (3.2.30) - (3.2.32).

The smoothness indicator of the large stencil S^6 is defined as the highest, i.e., fifth-degree, undivided difference as follows

$$\begin{aligned}\tau^{NW} &= (f_{j-2} - 5f_{j-1} + 10f_j - 10f_{j+1} + 5f_{j+2} - f_{j+3})^2 \\ &= (f^{(5)})^2\Delta x^{10} + \mathcal{O}(\Delta x^{11}).\end{aligned} \tag{3.3.13}$$

Carrying a similar procedure as in Eqs. (3.2.43) - (3.2.44) with a change in τ, thanks to Eq. (3.3.13), we obtain that

$$\frac{\tau^{NW}}{\beta_k} = \begin{cases} \mathcal{O}(\Delta x^8), & \text{if } f_j' \neq 0, \\ \mathcal{O}(\Delta x^6), & \text{if } f_j' = 0, \end{cases} \quad \forall k, \tag{3.3.14}$$

thus condition (3.3.10) is satisfied. Hence, the scheme is 6th-order in smooth regions. The case where the derivatives of f vanish will be checked numerically in the below section.

3.3.3 WENO-CU6

In [64], Hu et al. developed the adaptive central-upwind scheme WENO-CU6 based on the principle that the most downwind sub-stencil only plays roles in smooth regions and is suppressed near discontinuities. Hence the scheme is central in smooth regions and upwind near discontinuities. The scheme is different from WENO-NW6 in defining the smoothness indicators β_3

and τ. In particular, they are as below,

$$\begin{aligned}\beta_3 = {} & \frac{1}{120960}[271779f_{j-2}^2 + f_{j-2}(-2380800f_{j-1} + 4086352f_j - 3462252f_{j+1} + 1458762f_{j+2} - 245620f_{j+3}) \\ & + f_{j-1}(5653317f_{j-1} - 20427884f_j + 17905032f_{j+1} - 7727988f_{j+2} + 1325006f_{j+3}) + f_j(19510972f_j \\ & - 35817664f_{j+1} + 15929912f_{j+2} - 2792660f_{j+3}) + f_{j+1}(17195652f_{j+1} - 15880404f_{j+2} \\ & + 2863984f_{j+3}) + f_{j+2}(3824847f_{j+2} - 1429976f_{j+3}) + 139633f_{j+3}^2] \\ & = f'^2\Delta x^2 + \frac{13}{12}f''^2\Delta x^4 + \mathcal{O}(\Delta x^6).\end{aligned} \tag{3.3.15}$$

It is noticed that there is a typo in Eq. (25) in [64], and we have corrected it in Eq. (3.3.15).

From there, the smoothness indicator of the large stencil S^6 is defined as follows,

$$\tau^{CU} = \beta_3 - \frac{1}{6}(\beta_0 + 4\beta_1 + \beta_2) = \mathcal{O}(\Delta x^6). \tag{3.3.16}$$

Hence, we have for $k = 0, \ldots, 3,$,

$$\frac{\tau^{CU}}{\beta_k} = \begin{cases} \mathcal{O}(\Delta x^4), & \text{if } f_j' \neq 0, \\ \mathcal{O}(\Delta x^2), & \text{if } f_j' = 0, \end{cases} \tag{3.3.17}$$

which satisfies the condition (3.3.10).

It is also noteworthy that α_k^{CU} in Eq. (3.2.41) has a change as below,

$$\alpha_k^{CU} = \gamma_k \left(C + \frac{\tau^{CU}}{\varepsilon + \beta_k} \right), \tag{3.3.18}$$

where $C \gg 1$ is to increase the contribution of the linear weights when the smoothness indicators have comparable magnitudes (see [142]). Following [64], we choose $C = 20$.

As indicated in example 3.4.1, the 6th-order WENO-NW6 and WENO-CU6 schemes suffer from the loss of accuracy near the smooth critical points just right behind, with respect to the characteristic direction, a critical point where the first derivative of the solution is undefined, that is, the solution is just C^0 at that point. The explanation for this defect is given in the below section. In the next section, we propose a new scheme which automatically switches between a 6th-order central and 5th-order upwind scheme and overcomes the defect occurred in the mentioned schemes.

3.4 The New $6th$-order WENO-θ Scheme

3.4.1 An illustrative example

A drawback of the presented 6th-order WENO schemes (i.e., WENO-NW6, WENO-CU6) is that they suffer from a loss of resolution near the smooth critical region which is just behind another one where

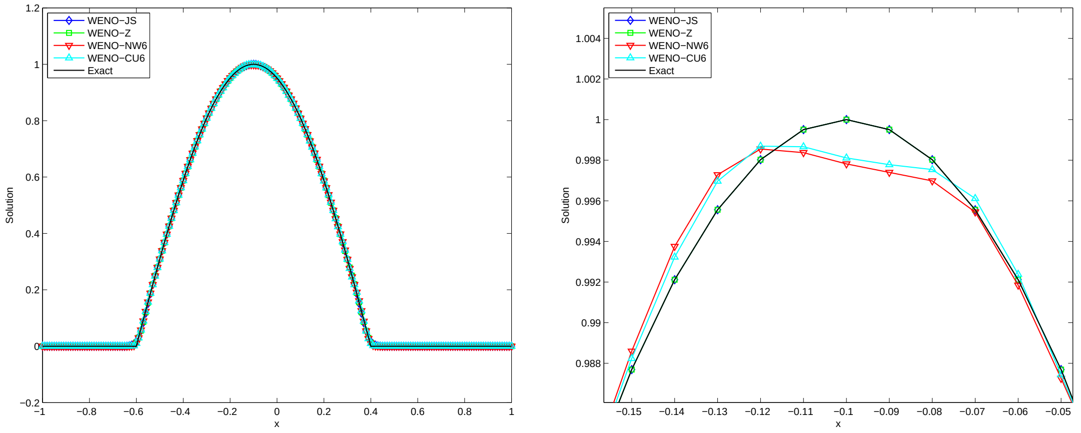


Figure 3-3: Left: Numerical solutions of Eq. (3.4.1) at time $t = 2.4$ obtained from different WENO schemes. Right: Zoom near the critical region.

the first derivative of the flux is undefined. To illustrate this, we consider Eq. (2.1.1) in a scalar case where $f(u) = u$ in the following example.

Example 3.4.1.

$$\begin{cases} u_t + u_x = 0, \quad x \in (-1, 1), \\ u_0(x) = \max(-\sin(\pi x), 0), \end{cases} \tag{3.4.1}$$

subject to periodic boundary conditions.

We approximate the solution of (3.4.1) by the WENO-JS, WENO-Z, WENO-NW6, and WENO-CU6 schemes. The results at time $t = 2.4$ with 201 grid points are plotted in Fig. 3-3 with the critical region zoomed in. It is clearly shown the above mentioned defect of the WENO-NW6 and WENO-CU6 schemes. Near the smooth critical region, we note that these schemes are worse than both WENO-JS and WENO-Z. Since there are many problems whose solution often exhibits the same behavior as mentioned above, we notice that this loss of accuracy is an important issue.

In [76], Jung and Nguyen constructed a new WENO scheme which overcomes the drawback of WENO-NW6 and WENO-CU6 presented in the previous example. They introduced a different switching mechanism between a 5th-order upwind and 6th-order central scheme. Unlike the WENO-NW6 or WENO-CU6 scheme in which the change depends on the smoothness indicator of the most downwind sub-stencil, in our scheme, whether the scheme is upwind or central is due to the smoothness indicator of the large stencil. Moreover, instead of using all six points for the indicator β_3, the number of points is reduced down to only four. The reason for this is explained in the below section. They also introduced a new set of smoothness indicators which are constructed in a central sense in Taylor expansions with respect to the point x_j. The feature of the new scheme is that it automatically switches between a 5th-order upwind scheme near discontinuities to prevent spurious oscillations, and a 6th-order central scheme in smooth regions which improves the loss of resolution of WENO-NW6 and WENO-CU6. We start the discussion on this scheme by the central reconstruction.

3.4.2 The New Scheme

We first observe that the $5th$- and $6th$-order linear approximations given in Eqs. (3.2.7) and (3.3.3), respectively, can be combined linearly in the following manner,

$$\hat{f}_{j+\frac{1}{2}} = \gamma_0^\theta \hat{f}^0_{j+\frac{1}{2}} + \gamma_1^\theta \hat{f}^1_{j+\frac{1}{2}} + \gamma_2^\theta \hat{f}^2_{j+\frac{1}{2}} + \gamma_3^\theta \hat{f}^3_{j+\frac{1}{2}}, \tag{3.4.2}$$

where

$$\gamma_0^\theta = \frac{1}{20}(1+\theta), \quad \gamma_1^\theta = \frac{3}{20}(3+\theta), \quad \gamma_2^\theta = \frac{3}{20}(3-\theta), \quad \gamma_3^\theta = \frac{1}{20}(1-\theta), \tag{3.4.3}$$

and $\hat{f}^k_{j+\frac{1}{2}}$, $k = 0, 1, 2, 3$ is given in Eqs. (3.2.19) - (3.2.21) and (3.3.9), respectively.

We deduce that

$$\hat{f}_{j+\frac{1}{2}} = \begin{cases} \hat{f}^5_{j+\frac{1}{2}} & \text{if } \theta = 1, \\ \hat{f}^6_{j+\frac{1}{2}} & \text{if } \theta = 0. \end{cases} \tag{3.4.4}$$

We then propose a new scheme in which $\hat{f}_{j+\frac{1}{2}}$ is chosen between $\hat{f}^5_{j+\frac{1}{2}}$ and $\hat{f}^6_{j+\frac{1}{2}}$ adaptively. Hence, the scheme is 5th-order upwind or 6th-order central depending on the smoothness of the stencils S^5 and S^6. We expect that this will get over the drawback of accuracy degeneration of the above mentioned central 6th-order schemes. To proceed, we first rewrite Eq. (3.4.2) using instead the non-linear weights ω_k^θ's as follows,

$$\hat{f}_{j+\frac{1}{2}} = \omega_0^\theta \hat{f}^0_{j+\frac{1}{2}} + \omega_1^\theta \hat{f}^1_{j+\frac{1}{2}} + \omega_2^\theta \hat{f}^2_{j+\frac{1}{2}} + \omega_3^\theta \hat{f}^3_{j+\frac{1}{2}}, \tag{3.4.5}$$

where, for $k = 0, \ldots, 3$, and

$$\omega_k^\theta = \frac{\alpha_k^\theta}{\sum_{l=0}^3 \alpha_l^\theta}, \quad \alpha_k^\theta = \gamma_k^\theta \left(1 + \frac{\tau^\theta}{\varepsilon + \tilde{\beta}_k}\right). \tag{3.4.6}$$

Here, τ^θ is the smoothness indicator of the large stencil, and $\tilde{\beta}_k$ is the smoothness indicator of the sub-stencil S_k. We note that the scheme works best with $\varepsilon = 10^{-10}$. We define these indicators in the following subsection.

3.4.3 Central Reconstruction, New Central Smoothness Indicators $\tilde{\beta}_k$

For a 6th-order central scheme over the large stencil S^6, spurious oscillations are expected to occur near discontinuities. To overcome this, WENO-NW6 and WENO-CU6 choose to construct their β_3 over all points of S^6. A more careful observation reveals that this cost can be reduced in the following way. We remind that the principle of WENO schemes is that there is at least one smoothest sub-stencil is used in the reconstruction of the numerical flux. We suppose that β_3 follows Eq. (3.3.12), that is, it measures the smoothness of the most downwind S_3 only. We further assume that the grid size Δx is so small that a discontinuity does not spread over two neighboring grid points, then for a 6th-order WENO scheme, the only case where oscillations occur is when a discontinuity is in between x_j and x_{j+1}. In this case, both $\hat{f}^0_{j+\frac{1}{2}}$ and $\hat{f}^3_{j+\frac{1}{2}}$ play main roles in the combination (3.4.5) since β_0 and β_3 are much smaller than the other two. This leads to oscillations since the downwind $\hat{f}^3_{j+\frac{1}{2}}$ is wrongly chosen. To prevent this from happening, we choose β_3 to measure the smoothness of an extended sub-stencil $\tilde{S}_3 := \{x_j, x_{j+1}, x_{j+2}, x_{j+3}\}$ (see Fig. 3-2 with S_3 extended by the dashed line). It is observed that S_2 is now a subset of $\tilde{S}_3$ and all sub-stencils share the point x_j. Hence, the case where oscillations occur is essentially eliminated.

For 5th-order schemes, all sub-stencils are symmetric with respect to x_j. As a result, the smoothness indicators β_k's are also symmetric with respect to x_j (see Eqs. (3.2.30) - (3.2.32)). That is, β_0 and β_2 are equal to each other up to order $\mathcal{O}(\Delta x^4)$ in Taylor expansions. We recall that WENO discretizations choose the sub-stencils depending on the non-linear weights ω_k's which are very sensitive to the smoothness indicators β_k's due to the latter's smallness in smooth regions (see Eq. (3.2.28) for WENO-JS, Eq. (3.2.41) for WENO-Z, WENO-NW6, and WENO-CU6). For the sensitivity, we mean that a small change in any β_k leads to a large difference among α_k's, thus ω_k's. In that sense, the symmetry in terms of Taylor expansions of β_k's reduces the effects of this sensitivity, especially in transition regions where the solution is smooth and discontinuous. We refer to Figs. 3-5 and 4-1 below for numerical evidences for this assessment in which the schemes with symmetric β_k's (i.e., WENO-Z and WENO-θ) show better results than the ones without this property. Unfortunately, the 6th-order methods lack of this (comparing β_3 in Eq. (3.3.11) for WENO-NW6, and Eq. (3.3.15) for WENO-CU6 with the other β_k's, $k = 0, 1, 2$, defined Eqs. (3.2.30) - (3.2.32)). In our new scheme, we try to recover the property. We devise our new smoothness indicators in a central sense. That is, they are constructed based on the reconstruction polynomials which are symmetric with respect to $x_{j+\frac{1}{2}}$. In addition, it is shown below that the new indicators are symmetric in terms of Taylor expansions with respect to x_j. We notice that Taylor expansions about x_j are natural since the approximation of $f(u)_x$ is at the interval center x_j (see Eq. (3.1.3)) although the reconstruction of the numerical flux function $h(x)$ is at the interface $x_{j+\frac{1}{2}}$ (see Eqs. (3.2.7) and (3.3.3) for 5th-order and 6th-order schemes, respectively).

Substituting the 4th-degree reconstruction polynomial

$$\tilde{f}^5(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4, \tag{3.4.7}$$

into Eq. (3.1.2) and evaluating at

$$x_k = \left(k - j - \frac{1}{2}\right)\Delta x, \quad k = j-2, \dots, j+2, \text{ around } x_{j+\frac{1}{2}} = 0, \tag{3.4.8}$$

we obtain the following,

$$\begin{cases} \frac{1}{\Delta x}\int_{-3\Delta x}^{-2\Delta x}(b_0+b_1y+b_2y^2+b_3y^3+b_4y^4)dy = f_{j-2}, \\ \frac{1}{\Delta x}\int_{-2\Delta x}^{-\Delta x}(b_0+b_1y+b_2y^2+b_3y^3+b_4y^4)dy = f_{j-1}, \\ \frac{1}{\Delta x}\int_{-\Delta x}^{0}(b_0+b_1y+b_2y^2+b_3y^3+b_4y^4)dy = f_{j}, \\ \frac{1}{\Delta x}\int_{0}^{\Delta x}(b_0+b_1y+b_2y^2+b_3y^3+b_4y^4)dy = f_{j+1}, \\ \frac{1}{\Delta x}\int_{\Delta x}^{2\Delta x}(b_0+b_1y+b_2y^2+b_3y^3+b_4y^4)dy = f_{j+2}. \end{cases} \tag{3.4.9}$$

The corresponding Vandermonde matrix of the coefficients b_k's is as follows,

$$N = \begin{bmatrix} 1 & -\frac{5}{2}\Delta x & \frac{19}{3}\Delta x^2 & -\frac{65}{4}\Delta x^3 & \frac{211}{5}\Delta x^4 \\ 1 & -\frac{3}{2}\Delta x & \frac{7}{3}\Delta x^2 & -\frac{15}{4}\Delta x^3 & \frac{31}{5}\Delta x^4 \\ 1 & -\frac{1}{2}\Delta x & \frac{1}{3}\Delta x^2 & -\frac{1}{4}\Delta x^3 & \frac{1}{5}\Delta x^4 \\ 1 & \frac{1}{2}\Delta x & \frac{1}{3}\Delta x^2 & \frac{1}{4}\Delta x^3 & \frac{1}{5}\Delta x^4 \\ 1 & \frac{3}{2}\Delta x & \frac{7}{3}\Delta x^2 & \frac{15}{4}\Delta x^3 & \frac{31}{5}\Delta x^4 \end{bmatrix}. \tag{3.4.10}$$

By solving the linear system

$$N\tilde{\mathbf{c}} = \mathbf{f}, \tag{3.4.11}$$

where $\tilde{\mathbf{c}}$ is the column vector of the unknown coefficients, $\mathbf{f}$ is the grid values of the flux function $f(u)$ given in the right-hand side of Eq. (3.4.9), the reconstruction polynomial $\tilde{f}^5(x)$ is computed with the coefficients as follows,

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} \frac{1}{30}f_{j-2} - \frac{13}{60}f_{j-1} + \frac{47}{60}f_j + \frac{9}{20}f_{j+1} - \frac{1}{20}f_{j+2} \\ \frac{1}{12\Delta x}f_{j-1} - \frac{5}{4\Delta x}f_j + \frac{5}{4\Delta x}f_{j+1} - \frac{1}{12\Delta x}f_{j+2} \\ -\frac{1}{8\Delta x^2}f_{j-2} + \frac{3}{4\Delta x^2}f_{j-1} - \frac{1}{\Delta x^2}f_j + \frac{1}{4\Delta x^2}f_{j+1} + \frac{1}{8\Delta x^2}f_{j+2} \\ -\frac{1}{6\Delta x^3}f_{j-1} + \frac{1}{2\Delta x^3}f_j - \frac{1}{2\Delta x^3}f_{j+1} + \frac{1}{6\Delta x^3}f_{j+2} \\ \frac{1}{24\Delta x^4}f_{j-2} - \frac{1}{6\Delta x^4}f_{j-1} + \frac{1}{4\Delta x^4}f_j - \frac{1}{6\Delta x^4}f_{j+1} + \frac{1}{24\Delta x^4}f_{j+2} \end{bmatrix}. \tag{3.4.12}$$

Evaluating $\tilde{f}^5(x)$ at $x_{j+\frac{1}{2}} = 0$ gives us exactly the approximation $\hat{f}^5_{j+\frac{1}{2}}$ in Eq. (3.2.7) above (compare with $\hat{f}^5(x)$ at $x_{j+\frac{1}{2}} = \frac{\Delta x}{2}$ in subsections 3.2.1 and 3.3.1). This $\tilde{f}^5(x)$ is also used in computing the smoothness indicator τ^θ in Eq. (3.4.26) of the new WENO-θ scheme.

Similarly, we have

$$\tilde{f}^6(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4 + b_5 x^5, \tag{3.4.13}$$

where

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_5 \\ b_5 \end{bmatrix} = \begin{bmatrix} \frac{1}{60} f_{j-2} - \frac{2}{15} f_{j-1} + \frac{37}{60} f_j + \frac{37}{60} f_{j+1} - \frac{2}{15} f_{j+2} + \frac{1}{60} f_{j+3} \\ -\frac{1}{90\Delta x} f_{j-2} + \frac{5}{36\Delta x} f_{j-1} - \frac{49}{36\Delta x} f_j + \frac{49}{36\Delta x} f_{j+1} - \frac{5}{36\Delta x} f_{j+2} + \frac{1}{90\Delta x} f_{j+3} \\ -\frac{1}{16\Delta x^2} f_{j-2} + \frac{7}{16\Delta x^2} f_{j-1} - \frac{3}{8\Delta x^2} f_j - \frac{3}{8\Delta x^2} f_{j+1} + \frac{7}{16\Delta x^2} f_{j+2} - \frac{1}{16\Delta x^2} f_{j+3} \\ \frac{1}{36\Delta x^3} f_{j-2} - \frac{11}{36\Delta x^3} f_{j-1} + \frac{7}{9\Delta x^3} f_j - \frac{7}{9\Delta x^3} f_{j+1} + \frac{11}{36\Delta x^3} f_{j+2} - \frac{1}{36\Delta x^3} f_{j+3} \\ \frac{1}{48\Delta x^4} f_{j-2} - \frac{1}{16\Delta x^4} f_{j-1} + \frac{1}{24\Delta x^4} f_j + \frac{1}{24\Delta x^4} f_{j+1} - \frac{1}{16\Delta x^4} f_{j+2} + \frac{1}{48\Delta x^4} f_{j+3} \\ -\frac{1}{120\Delta x^5} f_{j-2} + \frac{1}{24\Delta x^5} f_{j-1} - \frac{1}{12\Delta x^5} f_j + \frac{1}{12\Delta x^5} f_{j+1} - \frac{1}{24\Delta x^5} f_{j+2} + \frac{1}{120\Delta x^5} f_{j+3} \end{bmatrix}. \tag{3.4.14}$$

Evaluating $\tilde{f}^6(x)$ at $x_{j+\frac{1}{2}} = 0$ gives us exactly the approximation $\hat{f}^6_{j+\frac{1}{2}}$ in Eq. (3.3.3) above. This $\tilde{f}^6(x)$ is used in computing the smoothness indicator τ^θ in Eq. (3.4.27) of the new WENO-θ scheme.

With the same setting of grid points as in Eq. (3.4.8), applying the same process to the sub-stencil S_k, $k = 0, 1, 2, 3$, we obtain the central reconstruction polynomial $\tilde{f}^k(x)$ given as follows,

$$\tilde{f}^0(x) = \frac{2f_{j-2} - 7f_{j-1} + 11f_j}{6} + \left(\frac{f_{j-2} - 3f_{j-1} + 2f_j}{\Delta x}\right) x + \left(\frac{f_{j-2} - 2f_{j-1} + f_j}{2\Delta x^2}\right) x^2, \tag{3.4.15}$$

$$\tilde{f}^1(x) = \frac{-f_{j-1} + 5f_j + 2f_{j+1}}{6} + \left(\frac{f_{j+1} - f_j}{\Delta x}\right) x + \left(\frac{f_{j-1} - 2f_j + f_{j+1}}{2\Delta x^2}\right) x^2, \tag{3.4.16}$$

$$\tilde{f}^2(x) = \frac{2f_j + 5f_{j+1} - f_{j+2}}{6} + \left(\frac{f_{j+1} - f_j}{\Delta x}\right) x + \left(\frac{f_j - 2f_{j+1} + f_{j+2}}{2\Delta x^2}\right) x^2, \tag{3.4.17}$$

$$\begin{aligned} \tilde{f}^3(x) = {} & \frac{3f_j + 13f_{j+1} - 5f_{j+2} + f_{j+3}}{12} + \left(\frac{-11f_j + 9f_{j+1} + 3f_{j+2} - f_{j+3}}{12\Delta x}\right) x \\ & + \left(\frac{3f_j - 7f_{j+1} + 5f_{j+2} - f_{j+3}}{4\Delta x^2}\right) x^2 + \left(\frac{-f_j + 3f_{j+1} - 3f_{j+2} + f_{j+3}}{6\Delta x^3}\right) x^3. \end{aligned} \tag{3.4.18}$$

We notice that these reconstruction polynomials are different from those given in Eqs. (3.2.16) - (3.2.18) which are constructed symmetrically with respect to $x_j = 0$. Substituting these into Eq. (3.2.29) with $\tilde{f}^k(x)$ replacing $\hat{f}^k(x)$, $k = 0, 1, 2, 3$, we deduce the new central smoothness indicators as follows,

$$\begin{aligned}\tilde{\beta}_0 &= \frac{13}{12}(f_{j-2} - 2f_{j-1} + f_j)^2 + (f_{j-2} - 3f_{j-1} + 2f_j)^2 \\ &= f'^2 \Delta x^2 + f' f'' \Delta x^3 + \left(\frac{4}{3} f''^2 - \frac{5}{3} f' f'''\right) \Delta x^4 + \mathcal{O}(\Delta x^5),\end{aligned} \tag{3.4.19}$$

$$\begin{aligned}\tilde{\beta}_1 &= \frac{13}{12}(f_{j-1} - 2f_j + f_{j+1})^2 + (f_{j+1} - f_j)^2 \\ &= f'^2 \Delta x^2 + f' f'' \Delta x^3 + \left(\frac{4}{3} f''^2 + \frac{1}{3} f' f'''\right) \Delta x^4 + \mathcal{O}(\Delta x^5),\end{aligned} \tag{3.4.20}$$

$$\begin{aligned}\tilde{\beta}_2 &= \frac{13}{12}(f_j - 2f_{j+1} + f_{j+2})^2 + (f_j - f_{j+1})^2 \\ &= f'^2 \Delta x^2 + f' f'' \Delta x^3 + \left(\frac{4}{3} f''^2 + \frac{1}{3} f' f'''\right) \Delta x^4 + \mathcal{O}(\Delta x^5),\end{aligned} \tag{3.4.21}$$

and

$$\begin{aligned}\tilde{\beta}_3 &= \frac{13}{48}(3f_j - 7f_{j+1} + 5f_{j+2} - f_{j+3})^2 + (2f_{j+1} - 3f_{j+2} + f_{j+3})^2 \\ &= f'^2 \Delta x^2 + f' f'' \Delta x^3 + \left(\frac{4}{3} f''^2 - \frac{5}{3} f' f'''\right) \Delta x^4 + \mathcal{O}(\Delta x^5),\end{aligned} \tag{3.4.22}$$

where the derivatives are evaluated at $x = x_j$. We note that for the most downwind $\tilde{\beta}_3$, we treated $\tilde{f}_3(x)$ in Eq. (3.4.18) as a 2nd-degree polynomial by ignoring the 3rd-degree term when substituting it into Eq. (3.2.29) so that $\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left(\frac{d^2}{dx^2}\tilde{f}^3(x)\right)^2 dx$ is the highest-order variation. This is for the consistency with the other reconstruction polynomials $\tilde{f}^k$'s, $k = 0, 1, 2$, which are of lower-degree. We also modified the second term in the obtained indicator so that its Taylor expansion agrees with that of $\tilde{\beta}_0$ up to order $\mathcal{O}(\Delta x^4)$; thus all $\tilde{\beta}_k$'s are now symmetric with respect to x_j in Taylor expansions, which is our goal in designing these new smoothness indicators. The original $\hat{\beta}_3$ obtained from Eq. (3.2.29) was as below,

$$\hat{\beta}_3 = \frac{13}{48}(3f_j - 7f_{j+1} + 5f_{j+2} - f_{j+3})^2 + \frac{1}{144}(-11f_j + 9f_{j+1} + 3f_{j+2} - f_{j+3})^2. \tag{3.4.23}$$

In order to enhance the dispersion of the scheme, following the approach by Taylor et al. ([142]), we set a restriction on the smoothness indicators as below, for $k = 0, \ldots, 3$,

$$\tilde{\beta}_k = \begin{cases} 0, & \text{if } R(\tilde{\beta}) \leq \alpha_R, \\ \tilde{\beta}_k, & \text{otherwise;} \end{cases} \tag{3.4.24}$$

where

$$R(\tilde{\beta}) = \frac{\max_k(\tilde{\beta}_k)}{\varepsilon + \min_k(\tilde{\beta}_k)}. \tag{3.4.25}$$

Here, α_R is a threshold value depending on the configurations of flows. α_R is taken small for flows with the presence of shocks. For a detailed discussion, consult [142].

3.4.4 New τ^θ

We next devise the smoothness indicator of the large stencil S^6. Since the one proposed by Yamaleev and Carpenter in [157] (see Eq. (3.3.13)) is too dispersive and may lead to oscillations (see Fig. 4-4 below and the evidence in [157]), we introduce a new smoothness indicator τ^θ which is based on Eq. (3.2.29) but at a much higher order variations for the large stencil S^6. We consider the following τ's,

$$\begin{aligned}
\tau_5 &= \Delta x^5 \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left(\frac{d^3}{dx^3} \tilde{f}^5(x) \right)^2 dx + \Delta x^7 \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left(\frac{d^4}{dx^4} \tilde{f}^5(x) \right)^2 dx \\
&= \frac{13}{12}(f_{j-2} - 4f_{j-1} + 6f_j - 4f_{j+1} + f_{j+2})^2 + (-f_{j-1} + 3f_j - 3f_{j+1} + f_{j+2})^2 \\
&= f'''^2 \Delta x^6 + f''' f^{(4)} \Delta x^7 + \left(\frac{1}{2} f''' f^{(5)} + \frac{4}{3} (f^{(4)})^2 \right) \Delta x^8 + \mathcal{O}(\Delta x^9),
\end{aligned} \tag{3.4.26}$$

and

$$\begin{aligned}
\tau_6 &= \Delta x^7 \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left(\frac{d^4}{dx^4} \tilde{f}^6(x) \right)^2 dx + \Delta x^9 \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left(\frac{d^5}{dx^5} \tilde{f}^6(x) \right)^2 dx \\
&= \frac{13}{12}(-f_{j-2} + 5f_{j-1} - 10f_j + 10f_{j+1} - 5f_{j+2} + f_{j+3})^2 \\
&+ \frac{1}{4}(f_{j-2} - 3f_{j-1} + 2f_j + 2f_{j+1} - 3f_{j+2} + f_{j+3})^2 \\
&= (f^{(4)})^2 \Delta x^8 + f^{(4)} f^{(5)} \Delta x^9 + \left(\frac{5}{6} f^{(4)} f^{(5)} + \frac{4}{3} (f^{(5)})^2 \right) \Delta x^{10} + \mathcal{O}(\Delta x^{11}),
\end{aligned} \tag{3.4.27}$$

where $\tilde{f}^5(x)$ and $\tilde{f}^6(x)$ are the central reconstruction polynomials around $x_{j+\frac{1}{2}} = 0$ constructed in a similar way as with $\tilde{f}^k(x)$'s in Eqs. (3.4.15) - (3.4.18) but for the large stencils S^5 and S^6, respectively; and the derivatives are evaluated at x_j.

We then choose our τ^θ and set θ in Eq. (3.4.3) as follows,

$$(\tau^\theta, \theta) = \begin{cases} (\tau_6, 0) & \text{if } \tau_6 < \tau_5, \\ (\tau_5, 1) & \text{if } \tau_6 \geq \tau_5. \end{cases} \tag{3.4.28}$$

It is noted that by choosing such τ^θ and θ as in Eq. (3.4.28), the scheme achieves a 6th-order in smooth regions since $\tau_6 \ll \tau_5$; whereas it adaptively chooses the smoother large stencil between S^5 and S^6 in the WENO reconstruction near discontinuities or unresolved regions. The new scheme now chooses the smoothest not only sub-stencils but also large one in the reconstruction procedure. The non-linear weights follow Eq. (3.4.6) above. Since our new method depends on θ to switch between a 5th-order upwind and 6th-order central scheme, we name it WENO-θ. In the numerical simulations below, we use the name WENO-θ6 for the compatibility with the other 6th-order schemes.

Remark 3.4.1.

i. Although the definition of τ^{θ} has a switching mechanism by an *if* statement, it does not ruin the methodology of WENO schemes. This is because the switching applies to the smoothness indicator of the large stencil, not to the choice of smoother sub-stencils.

ii. Although the switching is discontinuous in nature, WENO-θ is robust for problems with highly unstable fluid flows. We illustrate this by conducting a numerical simulation of the Rayleigh-Taylor instability problem in the below section.

iii. The role of ε is well investigated in [50]. We note here that for cases with increasing number of vanishing derivatives, since both τ and β_k are very small at critical points, ε does play roles to sustain the designed formal accuracy order. For this reason, except for WENO-JS where $\varepsilon = 10^{-6}$, we choose $\varepsilon = 10^{-10}$ for other schemes. See the accuracy tests in the below section.

In the next step, we test the accuracy, efficiency, and resolutions of the new scheme.

3.4.5 Accuracy Tests

We note that either τ_5 or τ_6 is chosen in Eq. (3.4.28), the sufficient condition (3.3.10) is always satisfied. Hence the new scheme is 6th-order in smooth regions.

For the tests of accuracy, we choose the linear scalar conservation law,

$$\begin{cases} u_t + u_x = 0, \quad x \in (-1,1), \\ u(x,0) = u_0(x), \end{cases} \tag{3.4.29}$$

subject to periodic boundary conditions. The following initial data are considered:

- *Initial condition 1:*

$$u_0(x) = \sin(\pi x); \tag{3.4.30}$$

and

- *Initial condition 2:*

$$u_0(x) = \left(x + \frac{1}{2}\right)^k \exp\left(-100\left(x+\frac{1}{2}\right)^2\right), \tag{3.4.31}$$

where $k-1$ is the number of vanishing spatial derivatives at $x = -\frac{1}{2}$, that is, $0 = \frac{\partial f}{\partial x}\big|_{x=0} = \ldots = \frac{\partial^{(k-1)} f}{\partial x^{(k-1)}}\big|_{x=0} \neq \frac{\partial^{(k)} f}{\partial x^{(k)}}\big|_{x=0}$.

L^1 and L^{∞} errors of 6th-order schemes at time $t = 1$ are measured and listed in Table 3-1 together with the order of accuracy (in brackets), and are plotted in Fig. 3-4. We choose the time step $\Delta t = \Delta x^{6/3}$ so that the numerical errors in time do not contribute to the results. In the figures, we also show the errors of the 5th-order schemes for comparison.

Table 3-1: Convergence of $u_t + u_x = 0$ with initial conditions (3.4.30) and (3.4.31), at time $t = 1$.

		Eq. *(3.4.30)*		*Eq.* *(3.4.31),* $k = 2$		*Eq.* *(3.4.31),* $k = 3$	
	N	L^1 error	L^∞ error	L^1 error	L^∞ error	L^1 error	L^∞ error
WENO-	40	4.5E-07 (-)	3.4E-07 (-)	4.5E-04 (-)	2.2E-03 (-)	4.2E-05 (-)	1.3E-04 (-)
CU6	80	6.9E-09 (6.0)	5.4E-09 (6.0)	3.8E-05 (3.6)	2.0E-04 (3.5)	8.9E-06 (2.2)	4.5E-05 (1.5)
	160	1.1E-10 (6.0)	8.4E-11 (6.0)	6.5E-07 (5.9)	3.6E-06 (5.8)	1.3E-07 (6.1)	8.1E-07 (5.8)
	320	4.1E-13 (8.0)	3.8E-13 (7.8)	1.1E-08 (5.9)	6.0E-08 (5.9)	1.8E-09 (6.1)	1.1E-08 (6.2)
WENO-	40	4.5E-07 (-)	3.4E-07 (-)	5.0E-04 (-)	2.2E-03 (-)	4.3E-05 (-)	1.4E-04 (-)
NW6	80	6.9E-09 (6.0)	5.3E-09 (6.0)	4.1E-05 (3.6)	2.2E-04 (3.4)	8.7E-06 (2.3)	4.7E-05 (1.6)
	160	1.1E-10 (6.0)	8.4E-11 (6.0)	6.4E-07 (6.0)	3.6E-06 (5.9)	1.4E-07 (5.9)	9.6E-07 (5.6)
	320	4.1E-13 (7.8)	3.6E-13 (7.9)	1.0E-08 (5.9)	6.0E-08 (5.9)	1.8E-09 (6.3)	1.1E-08 (6.5)
WENO-	40	4.5E-07 (-)	3.4E-07 (-)	3.7E-04 (-)	1.8E-03 (-)	3.2E-05 (-)	1.2E-04 (-)
θ6	80	6.9E-09 (6.0)	5.3E-09 (6.0)	4.2E-05 (3.1)	2.3E-04 (2.9)	4.8E-06 (2.7)	2.2E-05 (2.4)
	160	1.1E-10 (6.0)	8.4E-11 (6.0)	7.6E-07 (5.8)	3.9E-06 (5.9)	1.2E-07 (5.3)	6.3E-07 (5.1)
	320	4.1E-13 (8.0)	3.7E-13 (7.8)	1.3E-08 (5.9)	8.0E-08 (5.6)	2.1E-09 (5.9)	1.1E-08 (5.9)

3.4.6 Resolution Tests

We now test if our new scheme overcomes the loss of accuracy of WENO-CU6 and WENO-NW6. We revisit the initial condition given in example 3.4.1 which is as follows,

$$u_0(x) = \max(\sin(\pi x), 0), \quad x \in (-1, 1). \tag{3.4.32}$$

The numerical solution obtained from our new scheme is added and shown in Fig. 3-5, together with those given in example 3.4.1. We also plot the pointwise errors in the same figure. It is shown that the new WENO-θ6 scheme approximates the critical region around $x = -0.1$ much better than WENO-CU6 and WENO-NW6. Indeed, the pointwise errors of the former around this region is comparable to those of WENO-JS and WENO-Z.

The nonlinear weights ω_k's of these schemes are plotted in Fig. 3-6. We observe that around the critical point, the ω_k's of WENO-Z and WENO-θ6 are stable and converge to their optimal values γ_k's. We note that for the latter scheme, the nonlinear weights keep fluctuating between the optimal weights

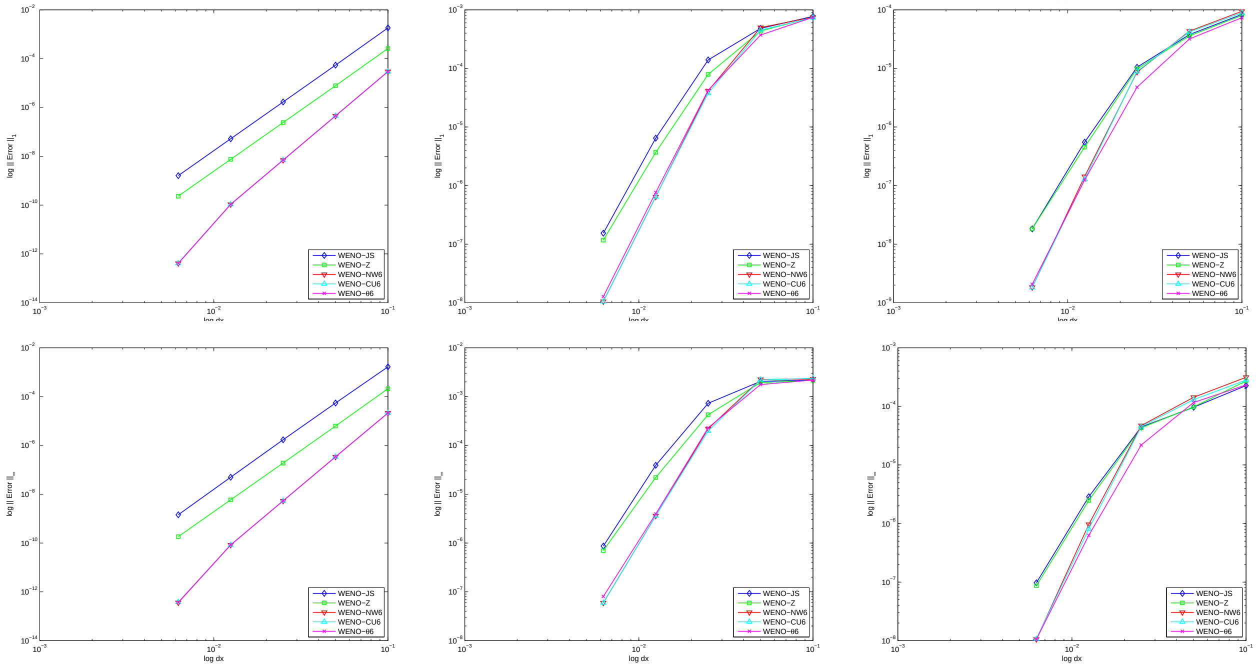


Figure 3-4: Convergence of Eq. (3.4.29) at time $t = 1$. Left: Initial condition (3.4.30); Middle: Initial condition (3.4.31) with $k = 2$. Right: Initial condition (3.4.31) with $k = 3$. Top: in L^1 norm; Bottom: in L^∞ norm.

of the 5th-order upwind and 6th-order central linear schemes. We also notify the non-convergence of ω_k's of the WENO-CU6 and WENO-NW6 schemes around the critical region. This shows the improvement of our new scheme over the other 6th-order ones.

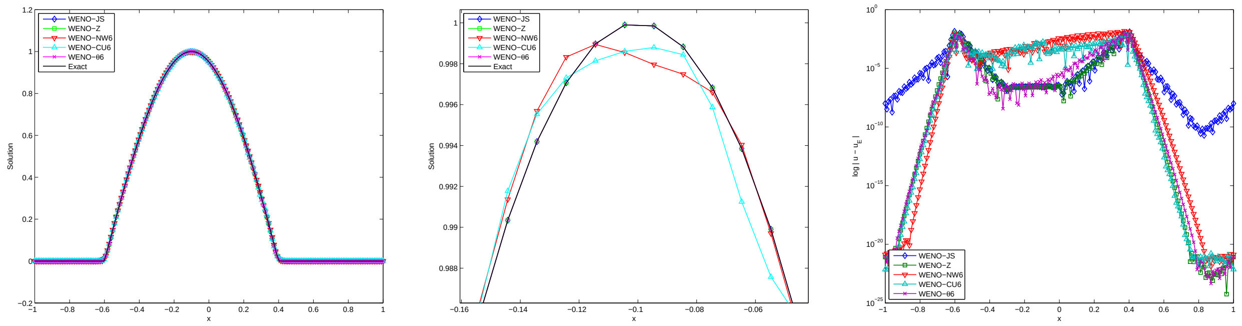


Figure 3-5: Left: Numerical solutions of Eq. (3.1.1) with initial condition (3.4.32) at time $t = 2.4$. Grid 200. Middle: Zoom near the critical point. Right: Pointwise errors in log scale.

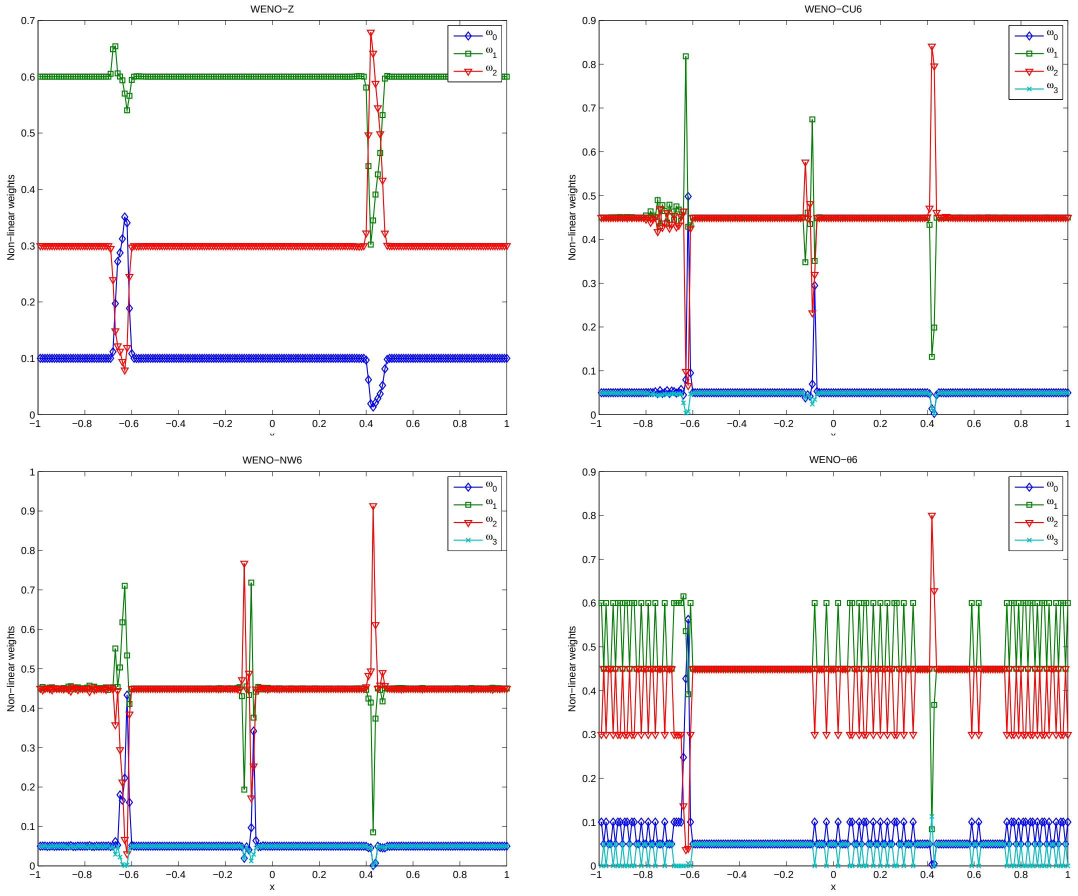


Figure 3-6: Distribution of the non-linear weights for the initial data (3.4.32). From top to bottom, left to right: WENO-Z, WENO-CU6, WENO-NW6, and WENO-θ6.

4

Numerical Results

In this chapter, we perform a number of tests to compare the results of our new scheme described in Chapter 3 with those obtained from the other WENO schemes, including the 5th-order upwind WENO-JS, WENO-Z, and the 6th-order central WENO-CU6, WENO-NW6. Since WENO-Z is a good replacement for the mapped WENO, we omit the latter one in our numerical experiments.

4.1 Scalar Conservation Laws

4.1.1 TEST 1: Linear Case

We solve the one-dimensional linear advection equation (3.4.29) with the following initial condition $u_0(x)$ which contains a C^∞ Gaussian, a square wave, a triangle, and a semi-ellipse (see [78]),

$$u_0(x) = \begin{cases} \frac{1}{6}[G(x,\beta,z-\delta)+4G(x,\beta,z)+G(x,\beta,z+\delta)], & -0.8 \le x \le -0.6, \\ 1, & -0.4 \le x \le -0.2, \\ 1-|10(x-0.1)|, & 0 \le x \le 0.2, \\ \frac{1}{6}[F(x,\alpha,a-\delta)+4F(x,\alpha,a)+F(x,\alpha,a+\delta)], & 0.4 \le x \le 0.6, \\ 0, & \text{otherwise}, \end{cases} \tag{4.1.1}$$

where

$$G(x,\beta,z) = \exp(-\beta(x-z)^2), \tag{4.1.2}$$

$$F(x,\alpha,a) = \sqrt{\max(1-\alpha^2(x-a)^2, 0)}; \tag{4.1.3}$$

the constants are $z = -0.7$, $\delta = 0.005$, $\beta = \frac{\log 2}{36\delta^2}$, $a = 0.5$, and $\alpha = 10$.

We compute the solution up to time $t = 6.3$ with $N = 401$ and periodic boundary conditions. The

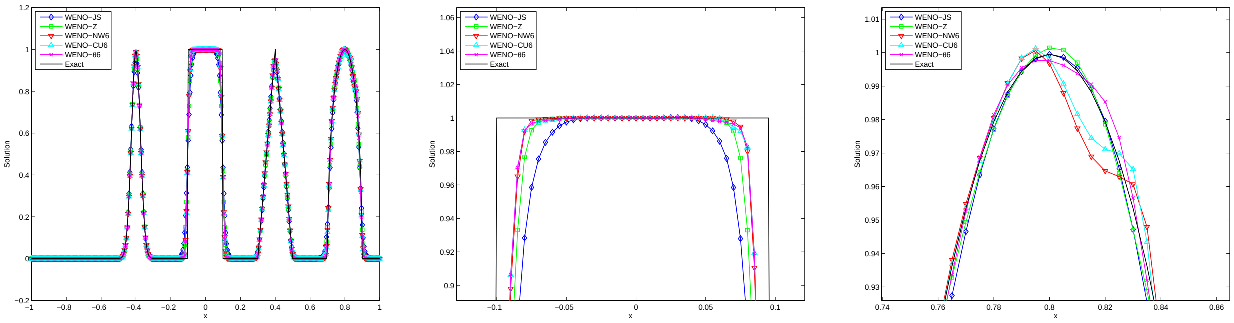


Figure 4-1: Left: Linear advection Eq. (3.4.29) with initial condition (4.1.1). Time $t = 6.3$. Grid 400. The others: zooms at critical regions.

results obtained from the WENO-CU6, WENO-NW6, and WENO-θ6 schemes are plotted in Fig. 4-1. Zooms around the shocks and top of the semi-ellipse are also shown in the same figure. It is observed that WENO-θ6 is comparable to WENO-NW6 in capturing the shocks, but the former is much better than the latter and WENO-CU6 in approximating top of the semi-ellipse.

4.1.2 TEST 2: Nonlinear Case

For this, we choose the Burgers equation

$$\begin{cases} u_t + \left(\dfrac{u^2}{2}\right)_x = 0, \quad x \in (-1,1), \\ u(x,0) = u_0(x), \end{cases} \tag{4.1.4}$$

subject to periodic boundary conditions.

In Fig. 4-2, we show the numerical results of the 6th-order WENO schemes for the initial condition

$$u_0(x) = -\sin(\pi x); \tag{4.1.5}$$

at time $t = 1.5$ and

$$u_0(x) = \frac{1}{2} + \sin(\pi x) \tag{4.1.6}$$

at $t = 0.55$. We choose a grid of $N = 200$ grid points. It is shown that the shocks are very well captured by all schemes.

4.2 Euler Equations of Gas Dynamics

In this subsection, we consider the one-dimensional Euler equations of gas dynamics given in the following,

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x = 0, \tag{4.2.1}$$

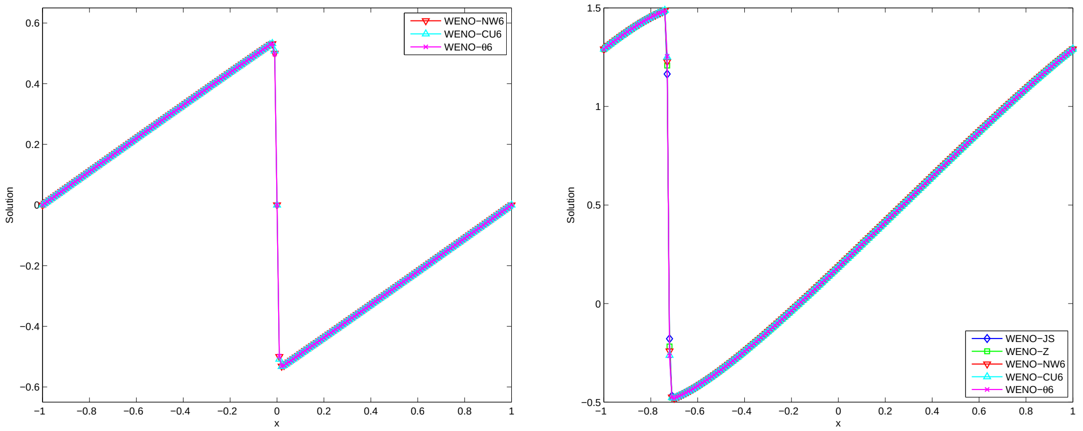


Figure 4-2: Burgers' Eq. (4.1.4). Grid 200. Left: initial condition (4.1.5) at time $t = 1.5$; Right: initial condition (4.1.6) at time $t = 0.55$.

where

$$\begin{aligned} \mathbf{u} &= (u_1, u_2, u_3) = (\rho, \rho u, E)^T, \\ \mathbf{f}(\mathbf{u}) &= (f_1, f_2, f_3) = (\rho u, p + \rho u^2, (E + p)u)^T, \end{aligned} \tag{4.2.2}$$

where ρ, u, p, E are density, velocity, pressure, and total energy, respectively. The total energy has the form

$$E = \frac{1}{2}\rho u^2 + \rho e, \tag{4.2.3}$$

where

$$e = e(\rho, p), \tag{4.2.4}$$

is the specific internal energy. For ideal gases, e has the form

$$e = \frac{p}{(\gamma - 1)\rho}. \tag{4.2.5}$$

Here, γ is the ratio of specific heats. Except indicated, for all numerical simulations in this chapter, we choose $\gamma = 1.4$.

Thus the total energy becomes

$$E = \frac{1}{2}\rho u^2 + \frac{p}{\gamma - 1}, \tag{4.2.6}$$

hence, the pressure is

$$p = (\gamma - 1)\left(E - \frac{1}{2}\rho u^2\right). \tag{4.2.7}$$

Writing Eq. (4.2.1) in a non-conservative form, we obtain that

$$\mathbf{u}_t + A(\mathbf{u})\mathbf{u}_x = 0. \tag{4.2.8}$$

Here, $A(\mathbf{u})$ is the Jacobian of the flux $\mathbf{f}(u)$, which is

$$A(\mathbf{u}) = \begin{bmatrix} 0 & 1 & 0 \\ -\frac{1}{2}(\gamma-3)\left(\frac{u_2}{u_1}\right)^2 & (3-\gamma)\frac{u_2}{u_1} & (\gamma-1) \\ -\gamma\frac{u_2u_3}{u_1^2}+(\gamma-1)\left(\frac{u_2}{u_1}\right)^3 & \gamma\frac{u_3}{u_1}-\frac{3}{2}(\gamma-1)\left(\frac{u_2}{u_1}\right)^2 & \gamma\frac{u_2}{u_1} \end{bmatrix}. \tag{4.2.9}$$

Let

$$H = \frac{E+p}{\rho} = \frac{1}{2}u^2 + h, \quad h = e + \frac{p}{\rho}, \tag{4.2.10}$$

be the total specific enthalpy and specific enthalpy, respectively. After some computation (see, e.g., [138]), we obtain the eigenvalues of right eigenvectors of $A(\mathbf{u})$ to be as follows,

$$\lambda_1 = u - c, \quad \lambda_2 = c, \quad \lambda_3 = u + c, \tag{4.2.11}$$

and

$$R = \Big[\mathbf{r}_1\Big|\mathbf{r}_2\Big|\mathbf{r}_3\Big] = \begin{bmatrix} 1 & 1 & 1 \\ u-c & u & u+c \\ H-uc & \frac{1}{2}u^2 & H+uc \end{bmatrix}. \tag{4.2.12}$$

Here,

$$c = \sqrt{\frac{\gamma p}{\rho}}, \tag{4.2.13}$$

is the speed of sound. It implies that the waves corresponding to λ_1 and λ_3 are only non-linear ones, i.e., either shocks or rarefactions; whereas the ones corresponding to λ_2 are only contact discontinuities.

For the purpose of computation, we also list the left eigenvectors, which are as follows,

$$L = \Big[\mathbf{l}_1\Big|\mathbf{l}_2\Big|\mathbf{l}_3\Big]^T = \begin{bmatrix} \frac{1}{2}\left(\frac{u}{c}+\frac{u^2}{2H-u^2}\right) & -\frac{1}{2}\left(\frac{1}{c}+\frac{2u}{2H-u^2}\right) & \frac{1}{2H-u^2} \\ 1-\frac{u^2}{2H-u^2} & \frac{2u}{2H-u^2} & -\frac{2}{2H-u^2} \\ -\frac{1}{2}\left(\frac{u}{c}-\frac{u^2}{2H-u^2}\right) & \frac{1}{2}\left(\frac{1}{c}-\frac{2u}{2H-u^2}\right) & \frac{1}{2H-u^2} \end{bmatrix}. \tag{4.2.14}$$

We notice that $LR = I$.

For all below numerical simulations, we apply the WENO schemes in characteristic fields of the flux $\mathbf{f}(\mathbf{u})$. That is, we first find an average of the Jacobian $A_{j+\frac{1}{2}}$ at the interface $x = x_{j+\frac{1}{2}}$. For this, we apply the Roe's mean matrix (see [118]). Then the eigenvalues λ_s's, $L = [\mathbf{l}_s]_{s=1}^m$, $R = [\mathbf{r}_s]_{s=1}^m$ the complete sets of the left and right eigenvectors, respectively, of $A_{j+\frac{1}{2}}$ are determined. We next project the flux $\mathbf{f}(\mathbf{u})$ into the characteristic fields by left multiplying it with L. WENO schemes with a global Lax-Friedrichs flux splitting are applied to approximate the components of the flux. After that, the approximation in each characteristic field is projected back to the component space by a right multiplying with the matrix R.

4.2.1 TEST 3: Riemann Problems

We consider the shock-tube problems which are Eq. (4.2.1) with Riemann initial data. In particular, the Sod problem, the Lax problem, and the 123 problem are given below.

- Sod's problem:

$$(\rho, u, p) = \begin{cases} (0.125, \quad 0, \quad 0.1), & -5 < x < 0, \\ (1, \quad 0, \quad 1), & 0 < x < 5; \end{cases} \tag{4.2.15}$$

and the final time $t = 1.7$.

- Lax's problem:

$$(\rho, u, p) = \begin{cases} (0.445, \quad 0.698, \quad 3.528), & -5 < x < 0, \\ (0.5, \quad 0, \quad 0.571), & 0 < x < 5; \end{cases} \tag{4.2.16}$$

and the final time $t = 1.3$.

- The 123 problem:

$$(\rho, u, p) = \begin{cases} (1, \quad -2, \quad 0.4), & -5 < x < 0, \\ (1, \quad 2, \quad 0.4), & 0 < x < 5; \end{cases} \tag{4.2.17}$$

and the final time $t = 1$.

We apply a transmissive condition at both boundaries. The exact solution of these shock-tube problems can be found in, for example, [138].

Numerical results of the density obtained from all WENO schemes with a grid of $N = 300$ are shown in Figs. 4-3 - 4-5, respectively. It is shown that WENO-NW6 and WENO-θ6 have the sharpest capturing of the discontinuities for the Sod problem. For the Lax problem, we notice an overshoot at the contact discontinuity about $x = 2$ of the WENO-NW6 and WENO-Z schemes; whereas WENO-θ6 does not experience this. It is also observe that for the 123 problem, WENO-JS, WENO-Z, and WENO-θ6 are better than WENO-NW6, WENO-CU6 in approximating the trivial contact discontinuity around $x = 0$.

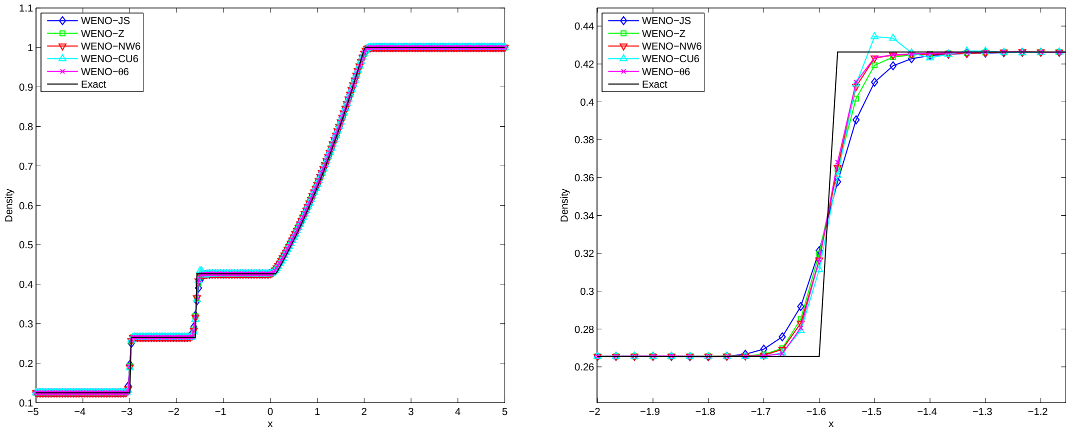


Figure 4-3: Left: Sod's problem with initial data (4.2.15). Time $t = 1.7$. Grid 300. Right: zoom at the contact discontinuity.

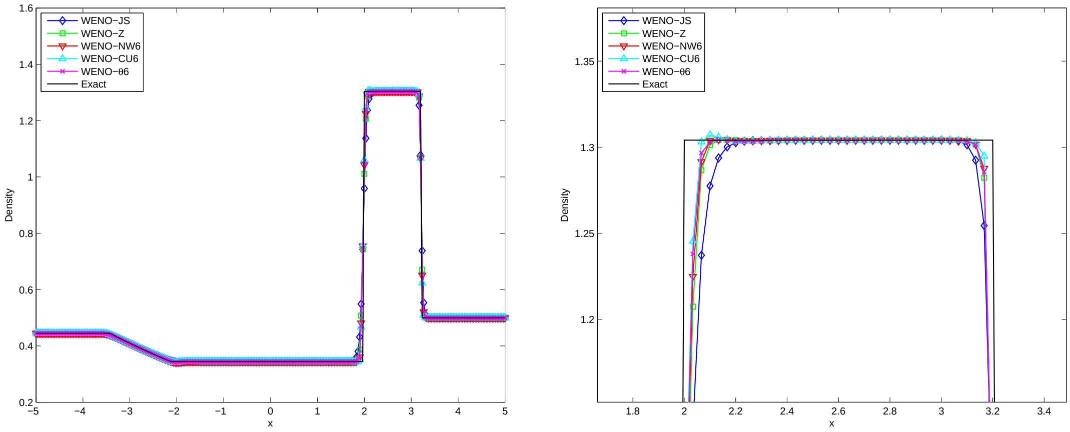


Figure 4-4: Left: Lax problem with initial data (4.2.16). Time $t = 1.3$. Grid 300. Right: zoom at the contact discontinuity.

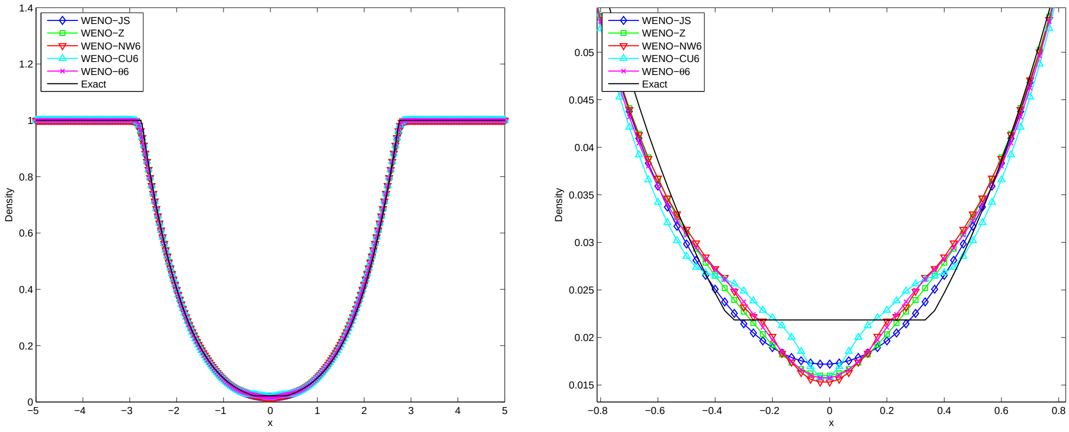


Figure 4-5: Left: The 123 problem with initial data (4.2.17). Time $t = 1.0$. Grid 300. Right: zoom at the trivial contact discontinuity.

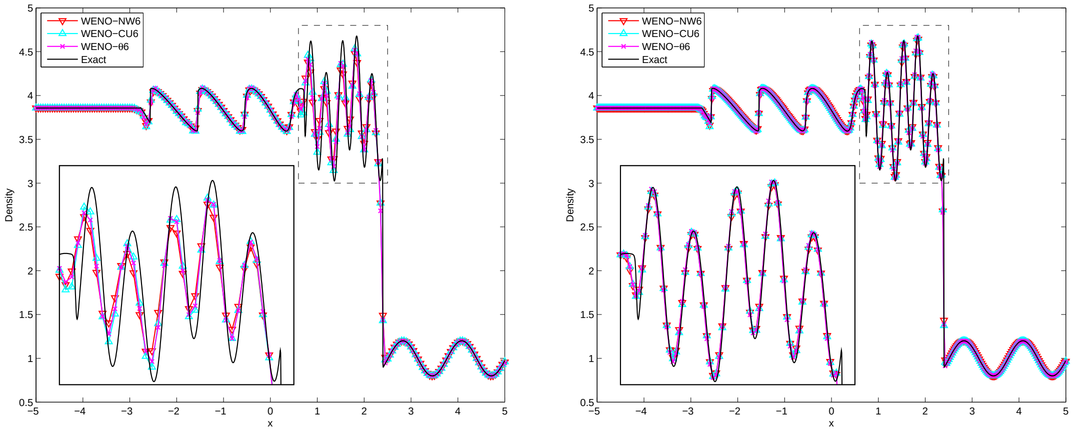


Figure 4-6: Shock density wave interaction with initial data (4.2.18). Density. Time $t = 1.8$. Left: medium grid 200; Right: fine grid 400.

4.2.2 TEST 4: Shock Density Wave Interaction

We consider the following initial data,

$$(\rho, u, p) = \begin{cases} (3.857143, \quad 2.629369, \quad 31/3), & -5 < x < -4, \\ (1 + 0.2\sin(5x), \quad 0, \quad 1), & -4 < x < 5, \end{cases} \tag{4.2.18}$$

with zero-gradient boundary conditions.

The problem simulates the interaction of a right-moving Mach 3 shock with a wavelike perturbed density whose magnitude is much smaller than the shock. As a result, a flow field of compressed and amplified wave trails is created right behind the shock. For more details, see [78]. In Fig. 4-6, we show the numerical results of the 6th-order WENO schemes at time $t = 1.8$ with grids of $N = 200$ and $N = 400$ points. The "exact" solution is computed by WENO-JS with a fine grid $N = 4000$. It is shown that all schemes give satisfactory approximations of the compressed wavelike structures behind the shock. A careful observation reveals that WENO-NW6 and WENO-θ6 are better than WENO-CU6 in case $N = 201$.

4.2.3 TEST 5: Two Interacting Blast Waves

In this test, we show that our new scheme WENO-θ6 passes the tough test of two interacting blast waves which the initial data are given as follows,

$$(\rho, u, p) = \begin{cases} (1, \quad 0, \quad 1000), & 0 < x < 0.1, \\ (1, \quad 0, \quad 0.01), & 0.1 < x < 0.9, \\ (1, \quad 0, \quad 100), & 0.9 < x < 1, \end{cases} \tag{4.2.19}$$

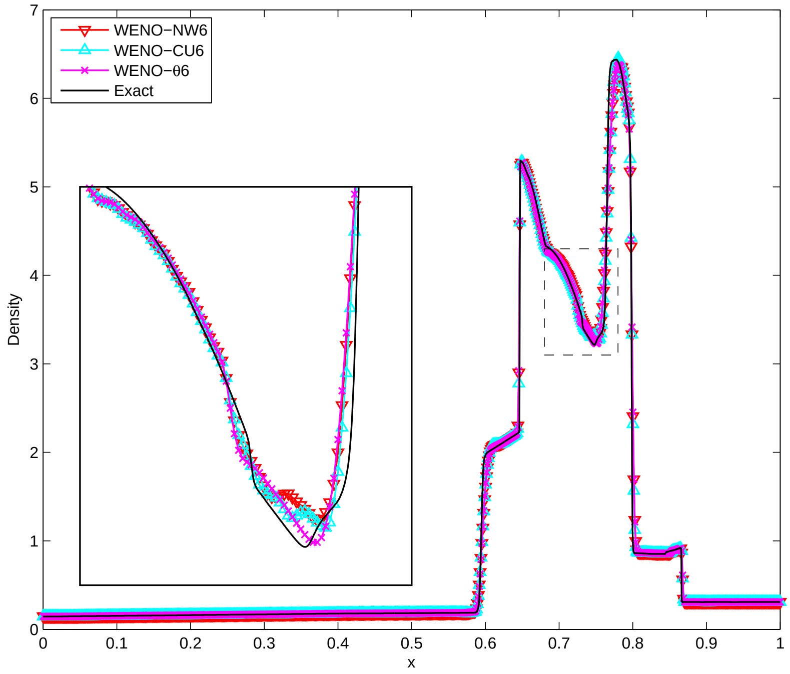


Figure 4-7: Two interacting blast waves with initial data (4.2.19). Density. Time $t = 0.038$. Grid 801.

and a reflective condition is applied at both boundaries. This problem is used to test the robustness of shock-capturing methods since many interactions are observed in a small area. A detailed discussion of this problem can be found in [151].

Numerical results of 6th-order WENO schemes are computed up to time $t = 0.038$ with a grid of $N = 801$ and plotted in Fig. 4-7 for the density. The exact solution is approximated by WENO-JS with a much fine grid $N = 4001$. It is shown that all schemes well capture the shocks as well as contact discontinuities. A zoom near $x = 0.745$ indicates that WENO-θ6 gives better resolutions than WENO-NW6 and WENO-CU6. We also emphasize that there exists a stair-casing phenomenon in the solutions of the latter methods in this region, which is similar to that at the top of the semi-ellipse in test 1 (see Fig. 4-1), and the 123 problem (see Fig. 4-5).

4.3 Two-dimensional Euler's Equations

In this subsection, we extend the problem to two-dimensional cases. We choose the 2D Euler equations which are as follows,

$$\begin{cases} \mathbf{u}_t + \mathbf{f}(\mathbf{u})_x + \mathbf{g}(\mathbf{u})_y = 0, \\ \mathbf{u}(x, y, 0) = \mathbf{u}_0(x, y), \end{cases} \tag{4.3.1}$$

where $\mathbf{u} = (\rho, \rho u, \rho v, E)^T$, $\mathbf{f}(\mathbf{u}) = (\rho u, p + \rho u^2, \rho uv, u(E+p))^T$, $\mathbf{g}(\mathbf{u}) = (\rho v, \rho uv, p + \rho v^2, v(E+p))^T$. The relation of pressure and conservative quantities is through the equation of state

$$p = (\gamma - 1)\left(E - \frac{1}{2}|\mathbf{u}|^2\right), \tag{4.3.2}$$

where $|\mathbf{u}|^2 = u^2 + v^2$. Here, we choose the ratio of specific heats $\gamma = 1.4$.

Eq. (4.3.1) is written in a non-conservative form as below,

$$\mathbf{u}_t + A(\mathbf{u})\mathbf{u}_x + B(\mathbf{u})\mathbf{u}_y = 0, \tag{4.3.3}$$

where the Jacobian matrices are (see [115], [138])

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{2}(\gamma-1)|\mathbf{u}|^2 - u^2 & (3-\gamma)u & -(\gamma-1)v & (\gamma-1) \\ -uv & v & u & 0 \\ u\left[\frac{1}{2}(\gamma-1)|\mathbf{u}|^2 - H\right] & H - (\gamma-1)u^2 & -(\gamma-1)uv & \gamma u \end{bmatrix}, \tag{4.3.4}$$

$$B = \begin{bmatrix} 0 & 0 & 1 & 0 \\ -uv & v & u & 0 \\ \frac{1}{2}(\gamma-1)|\mathbf{u}|^2 - v^2 & -(\gamma-1)u & (3-\gamma)v & (\gamma-1) \\ v\left[\frac{1}{2}(\gamma-1)|\mathbf{u}|^2 - H\right] & -(\gamma-1)uv & H - (\gamma-1)v^2 & \gamma v \end{bmatrix}. \tag{4.3.5}$$

The eigenvalues and corresponding eigenvectors are

$$\lambda_1^A = u - c, \quad \lambda_2^A = \lambda_3^A = u, \quad \lambda_4^A = u + c, \tag{4.3.6}$$

$$R^A = \begin{bmatrix} 1 & 0 & 1 & 1 \\ u - c & 0 & u & u + c \\ v & 1 & v & v \\ H - uc & v & \frac{1}{2}|\mathbf{u}|^2 & H + uc \end{bmatrix}, \tag{4.3.7}$$

$$L^A = \begin{bmatrix} \frac{1}{2}\left(\frac{u}{c} + \frac{|\mathbf{u}|^2}{2H - |\mathbf{u}|^2}\right) & -\frac{1}{2}\left(\frac{1}{c} + \frac{2u}{2H - |\mathbf{u}|^2}\right) & -\frac{v}{2H - |\mathbf{u}|^2} & \frac{1}{2H - |\mathbf{u}|^2} \\ -v & 0 & 1 & 0 \\ 1 - \frac{|\mathbf{u}|^2}{2H - |\mathbf{u}|^2} & \frac{2u}{2H - |\mathbf{u}|^2} & \frac{2v}{2H - |\mathbf{u}|^2} & -\frac{2}{2H - |\mathbf{u}|^2} \\ -\frac{1}{2}\left(\frac{u}{c} - \frac{|\mathbf{u}|^2}{2H - |\mathbf{u}|^2}\right) & \frac{1}{2}\left(\frac{1}{c} - \frac{2u}{2H - |\mathbf{u}|^2}\right) & -\frac{v}{2H - |\mathbf{u}|^2} & \frac{1}{2H - |\mathbf{u}|^2} \end{bmatrix}; \tag{4.3.8}$$

and

$$\lambda_1^B = v - c, \quad \lambda_2^B = \lambda_3^B = v, \quad \lambda_4^B = v + c, \tag{4.3.9}$$

$$R^B = \begin{bmatrix} 1 & 0 & 1 & 1 \\ u & 1 & u & u \\ v - c & 0 & v & v + c \\ H - vc & u & \frac{1}{2}|\mathbf{u}|^2 & H + vc \end{bmatrix}, \tag{4.3.10}$$

$$L^B = \begin{bmatrix} \frac{1}{2}\left(\frac{v}{c} + \frac{|\mathbf{u}|^2}{2H - |\mathbf{u}|^2}\right) & -\frac{u}{2H - |\mathbf{u}|^2} & -\frac{1}{2}\left(\frac{1}{c} + \frac{2v}{2H - |\mathbf{u}|^2}\right) & \frac{1}{2H - |\mathbf{u}|^2} \\ -u & 1 & 0 & 0 \\ 1 - \frac{|\mathbf{u}|^2}{2H - |\mathbf{u}|^2} & \frac{2u}{2H - |\mathbf{u}|^2} & \frac{2v}{2H - |\mathbf{u}|^2} & -\frac{2}{2H - |\mathbf{u}|^2} \\ -\frac{1}{2}\left(\frac{v}{c} - \frac{|\mathbf{u}|^2}{2H - |\mathbf{u}|^2}\right) & -\frac{u}{2H - |\mathbf{u}|^2} & \frac{1}{2}\left(\frac{1}{c} - \frac{2v}{2H - |\mathbf{u}|^2}\right) & \frac{1}{2H - |\mathbf{u}|^2} \end{bmatrix}. \tag{4.3.11}$$

4.3.0.1 TEST 6: Rayleigh-Taylor Instability

In the following tests, we show numerical evidence that WENO-θ6 maintains symmetry in the solutions much better than the other 6th-order schemes, and outperforms 5th-order schemes in resolving small-scaled structures occurring in flow configurations. We first simulate the Rayleigh-Taylor instability. The instability occurs where there is a heavy fluid falling into a light fluid (see [38], [2], [35]). Following [2], we set up the problem as follows. The domain is $(x, y) = (-0.25, 0.25) \times (-0.75, 0.75)$. Initial density has a discontinuity at the interface, i.e., $\rho = 2$ for $y \geq 0$ and $\rho = 1$ for $y < 0$. The pressure is set at hydrostatic equilibrium initially $p = 2.5 - \rho g y$ where $g = 0.1$ is the gravitational acceleration. The x-component velocity $u = 0$, while the y-component is perturbed with $v = \frac{0.01}{4}(1 + \cos(4\pi x))(1 + \cos(\frac{4}{3}\pi y))$ for a single mode perturbation. Boundary conditions are set periodic in x-direction, and reflective in y-direction. The ratio of specific heats $\gamma = 1.4$. We add $-g\rho$ and $-g\rho v$ in the y-momentum and energy equations of (4.3.1) as source terms.

In Fig. 4-8, we plot the density with 20 equally spaced contours obtained from 5th- and 6th-order WENO schemes at time $t = 9.5$ with a 120×360 grid. It is shown that the 6th-order schemes have much better numerical resolution comparing with the 5th-order ones. We notice that WENO-θ6 preserves the symmetry of the solution; whereas WENO-NW6 and WENO-CU6 do not. We conjecture the lack of symmetry of WENO-NW6 is due to the loss of accuracy around critical regions which is shown in previous numerical tests. The test also shows that the discontinuous switching of τ^θ in Eq. (3.4.28) does not have its effect on the robustness of the new WENO-θ scheme, even for problem with highly unstable fluid flows as the Rayleigh-Taylor instability.

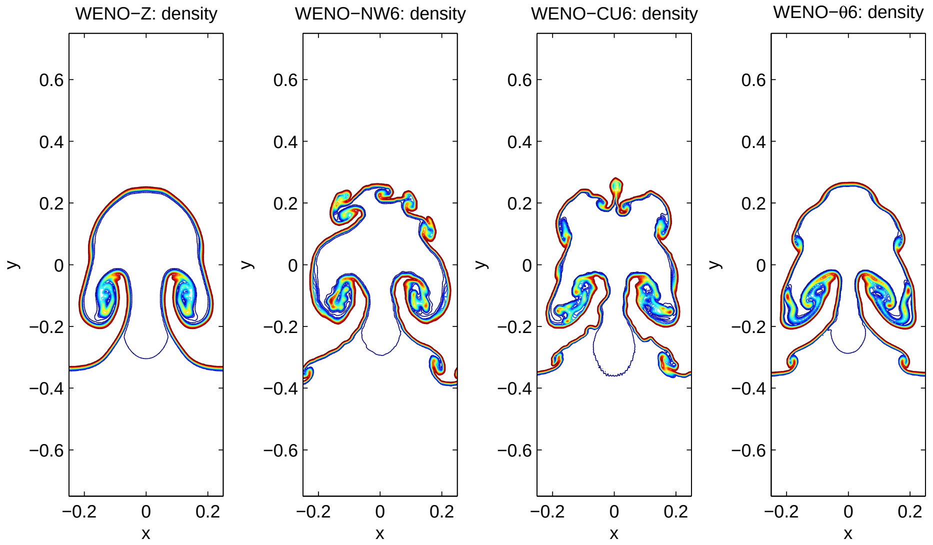


Figure 4-8: The Rayleigh-Taylor instability. Density at time $t = 9.5$. Grid 120×360. $CFL = 0.5$. From left to right: WENO-Z, WENO-NW6, WENO-CU6, and WENO-θ6.

4.3.0.2 TEST 7: Implosion problem

The next numerical test is the implosion problem (see [2], [98]) with initial data as follows,

$$(\rho, p) = \begin{cases} (1,1) & \text{for } x + y > \frac{1}{2}, \\ (0.125, 0.14) & \text{otherwise}, \end{cases} \tag{4.3.12}$$

and zero velocity everywhere initially. We choose reflecting conditions for all boundaries.

Symmetry is important for this test. For such a scheme, due to the interactions of shock waves and reflecting boundaries, jets along the diagonal are created. Longer and narrower jets are produced for less dissipative schemes.

In Fig. 4-9, we show the results obtained from different schemes on computational domain $(0,1) \times (0,1)$ at final time $t = 5$. We choose a grid of 400×400. For WENO-θ, we choose $\alpha_R = 1$. It is shown that only WENO-Z and WENO-θ well preserve the symmetry of the problem; whereas the other 6th-order schemes do not. The jets created by WENO-NW6 and WENO-CU6 tend to diverge from the main diagonal $x = y$. We also note that the jets produced by WENO-θ is much longer and narrower than those of WENO-Z, which means that the former scheme is less dissipative than the latter one.

4.3.0.3 TEST 8: 2D Riemann Initial Data

The 2D Riemann problem is set up by assigning different constant states of (ρ_k, u_k, v_k, p_k), $k = 1, 2, 3, 4$, to four quadrants of the computational domain $\Omega = (0,1) \times (0,1)$. The constant states are chosen so that there is only a single elementary wave, namely, shock-, rarefaction-, and contact-wave, connecting two neighboring quadrants (see [130]). For our test, we choose the following configuration for the initial

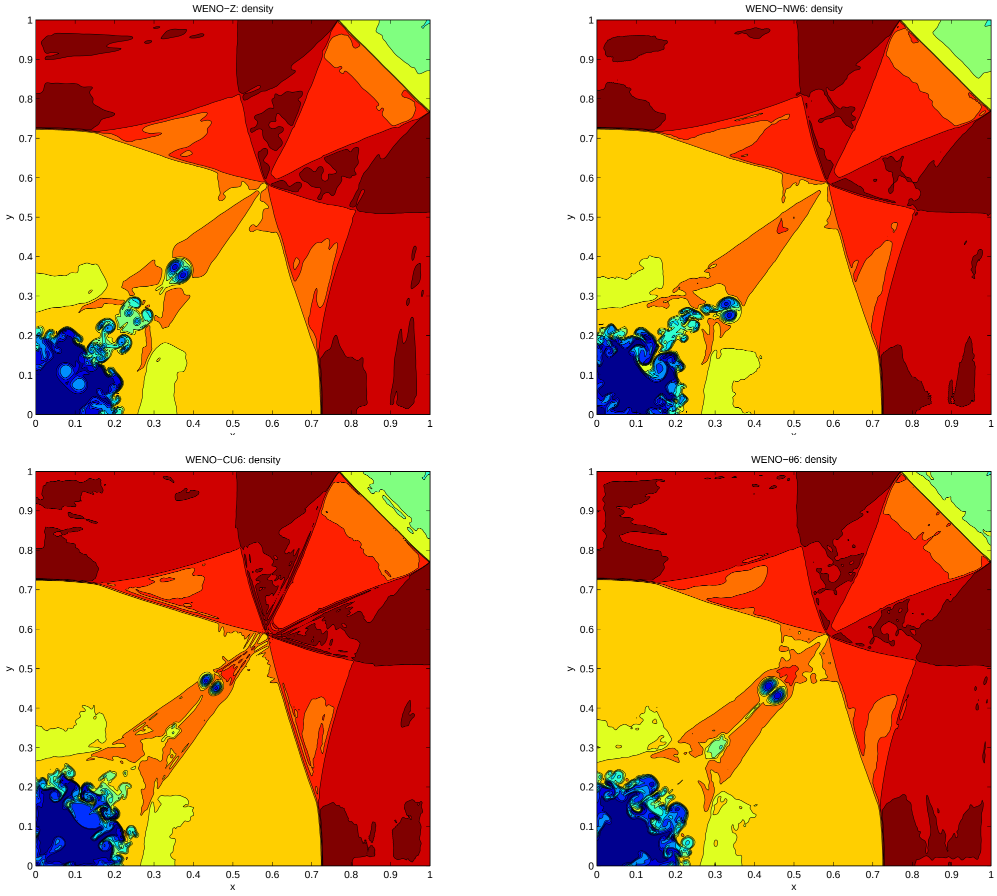


Figure 4-9: The implosion problem. Density with 20 contours uniformly distributing from 0 to 1. Grid 400 × 400. Final time $t = 5$. Left to right, top to bottom: WENO-Z, WENO-NW6, WENO-CU6, WENO-θ6.

data, respectively, for quadrants 1, 2, 3, 4,

$$(\rho, u, v, p) = \begin{cases} (0.5313,\ 0,\ 0,\ 0.4), & x > 0.5,\ y > 0.5, \\ (1,\ 0.7276,\ 0,\ 1), & x < 0.5,\ y > 0.5, \\ (0.8,\ 0,\ 0,\ 1), & x < 0.5,\ y < 0.5, \\ (1,\ 0,\ 0.7276,\ 1), & x > 0.5,\ y < 0.5, \end{cases} \tag{4.3.13}$$

which has shocks through quadrants 1 - 2 and 1 - 4, and contact discontinuities through quadrants 2 - 3 and 3 - 4. Transmissive boundary conditions are imposed on all boundaries for these two cases.

The approximations of the density with initial data (4.3.13) at time $t = 0.25$ are plotted in Fig. 4-10 with 50 contours for WENO-NW6 and WENO-θ6. Here, we use a fine grid with 1000×1000 intervals for the capturing of the small vortices along the contacts. Zooms near the spirals region are also shown on the right column in the same figure. Again, we observe a better performance of the WENO-θ6 over WENO-NW6 and WENO-CU6 schemes over these small-scaled structures without oscillations on the contours.

4.3.0.4 TEST 9: Double Mach Reflection of a Strong Shock

Finally, we investigate the double Mach reflection of a strong shock which is a typical benchmark test for shock-capturing methods. The problem simulates the reflection occurring when a simple planar shock interacts with a wedge making with the x-axis an angle α. The strength of the moving shock is characterized by the Mach number M_s. For a double Mach reflection problem, $M_s = 10$ and the wedge angle is chosen as $\alpha = 30°$. Detailed discussions on this type of problems can be found in [151] and the references therein. For numerical purpose, we choose the computational domain $\Omega = (0, 4) \times (0, 1)$. Initially the shock is located at $x_0 = 1/6$, inclined with the x-axis by the angle $90° - \alpha$. For boundary treatments, we apply Inflow and zero gradients conditions on the left and right boundaries, respectively. On the bottom one, a reflective condition is applied to the interval $[x_0, 4]$ representing the wedge, and the exact post-shock state is imposed over $[0, x_0]$. The top boundary is treated in a way that there are no interactions of the shock with this boundary. That is, the exact post- and pre-shock states are employed over the intervals $[0, x_s(t)]$ and $[x_s(t), 4]$, respectively, on the top boundary. Here, $x_s(t) = x_0 + \dfrac{1}{\tan 60°} + \dfrac{M_s a_{pre}}{\cos 30°} t$, where a_{pre} is the sound speed of the pre-shock state, is the location of the shock in time. These states can be computed exactly when one of them is pre-described (see, e.g., [138]). In particular, for our problem the initial data is given as follows,

$$(\rho, u, v, p) = \begin{cases} (8,\ 8.25\cos 30°,\ -8.25\sin 30°,\ 116.5), & x < x_0 + \frac{y}{\tan 60°}, \\ (1.4,\ 0,\ 0,\ 1), & x \geq x_0 + \frac{y}{\tan 60°}. \end{cases} \tag{4.3.14}$$

Numerical results of the density obtained from the 5th-order WENO-Z, the 6th-order WENO-NW6, WENO-CU6, and WENO-θ6 schemes at time $t = 0.2$ are plotted in Fig. 4-11 with 30 contours. For this

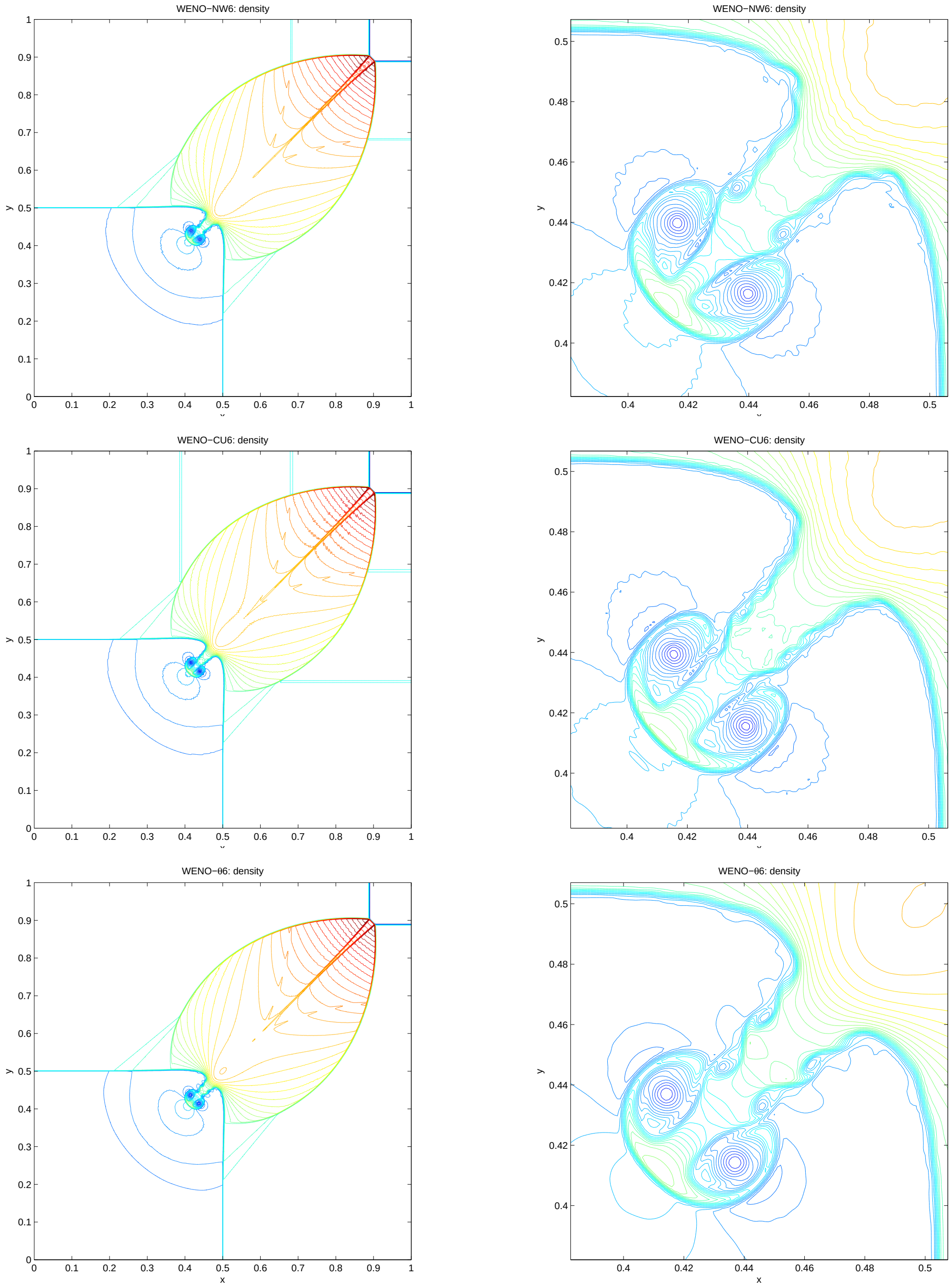
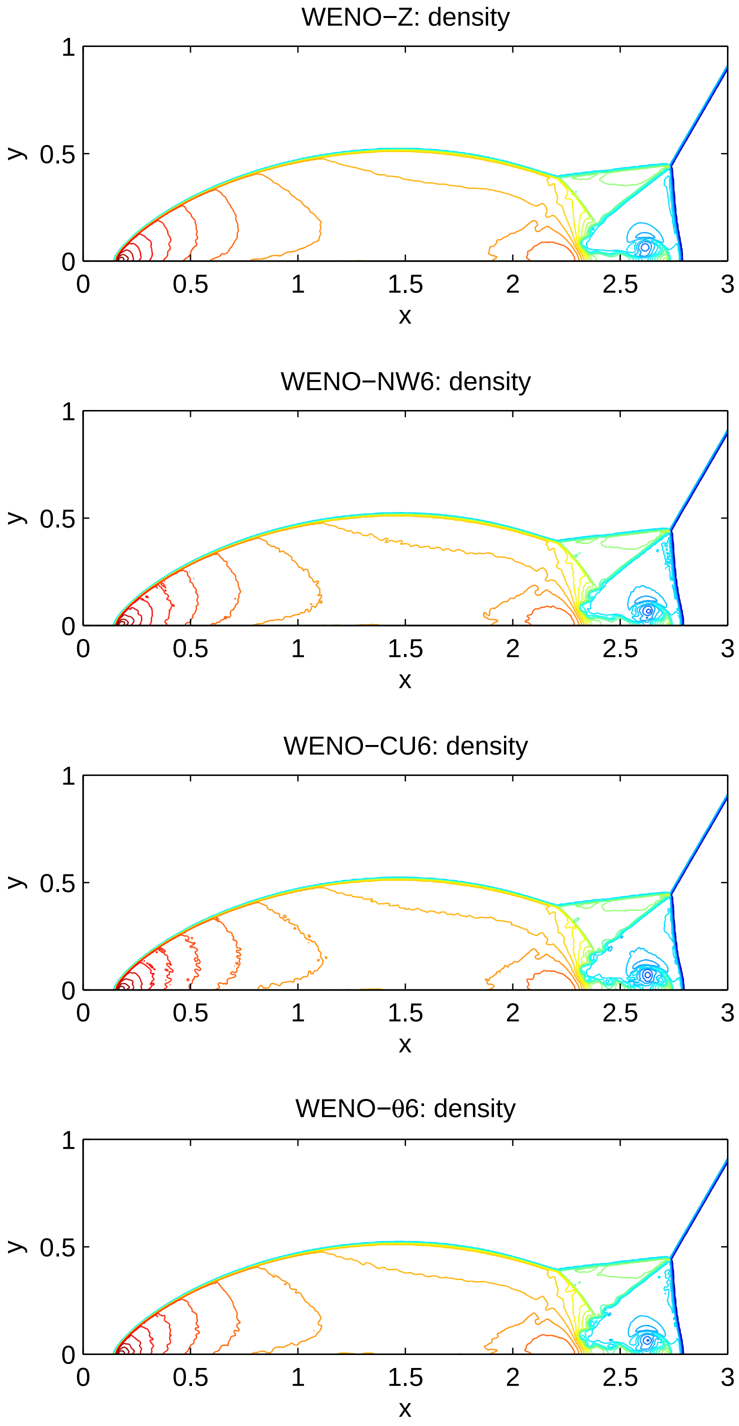


Figure 4-10: Left: The 2D Riemann problem with initial data (4.3.13). Density with 50 contours. Time $t = 0.25$. Grid 1000×1000. Right: Zoom at the spirals. From top to bottom, respectively: WENO-NW6, WENO-CU6, WENO-θ6.



WENO−Z: density

WENO−NW6: density

WENO−CU6: density

WENO−θ6: density

Figure 4-11: The double-Mach reflection problem with initial data (4.3.14). Density with 30 contours. Time $t = 0.2$. Grid 800×200. From top to bottom, respectively: WENO-Z, WENO-NW6, WENO-CU6, and WENO-θ6.

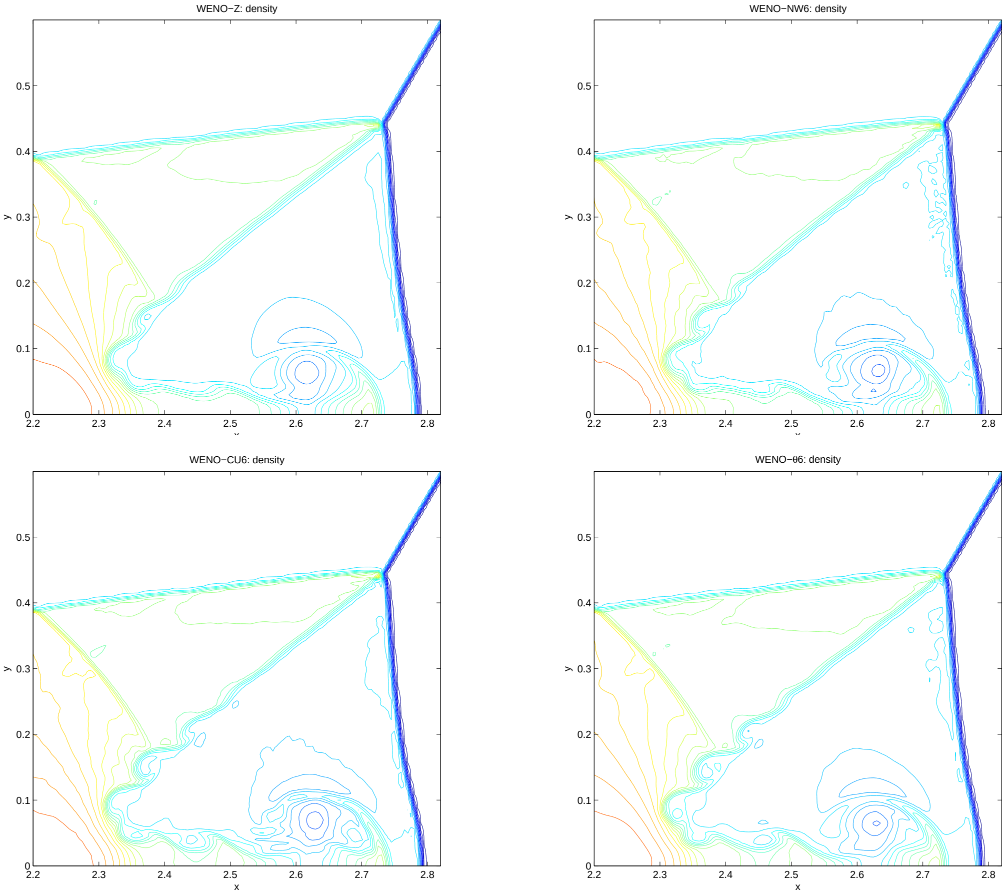


Figure 4-12: The double-Mach reflection problem with initial data (4.3.14). Zoom at the double Mach stems region. From top to bottom, respectively: WENO-Z, WENO-NW6, WENO-CU6, and WENO-θ6.

case, we choose a fine grid of 800 × 200 points for all schemes. We notice the rendering of small vortices at the end of the slip line and the wall jet, starting from WENO-Z and becoming clearer for the 6th-order schemes. The zoom-in on the Mach stems region shown in Fig. 4-12 reveals that the WENO-θ6 scheme gives more satisfactory resolution than the WENO-NW6 and WENO-CU6 ones.

Part II

Numerical Methods for Hyperbolic PDEs with Uncertainties and Dispersive PDEs

5

Maxwell Solutions in Media with Multiple Random Interfaces

5.1 Introduction

Time evolution of waves in random media has important applications in a wide range of areas such as medical imaging, wave scattering, radar detection, ionospheric plasmas and photonic devices (see e.g. [65]). Although the problem under consideration here is a forward problem, our approach reveals the effects of random inputs and provide some insights into inverse problems, e.g., the reconstruction of the interior of a human body from MRI or Ultrasound, recovery of the interior structural parameters of machines from non-destructive measurements, ionospheric dynamics and related problems.

In this chapter, we study the evolution of the cumulative distribution functions (CDF) in time of electromagnetic(EM) fields, which are governed by the following 2D transverse magnetic (TM) Maxwell equations (e.g., [19], [100], [140]): for $(x, y, t) \in \mathbb{R}^2 \times (0, \infty)$,

$$\begin{cases} \dfrac{\partial H_1}{\partial t} = -\dfrac{1}{\mu}\dfrac{\partial E_3}{\partial y}, \\ \dfrac{\partial H_2}{\partial t} = \dfrac{1}{\mu}\dfrac{\partial E_3}{\partial x}, \\ \dfrac{\partial E_3}{\partial t} = \dfrac{1}{\epsilon}\dfrac{\partial H_2}{\partial x} - \dfrac{1}{\epsilon}\dfrac{\partial H_1}{\partial y}. \end{cases} \tag{5.1.1}$$

The initial conditions are $H_1(x, y, 0) = h_1(x, y)$, $H_2(x, y, 0) = h_2(x, y)$, $E_3(x, y, 0) = e_3(x, y)$, where $H = (H_1, H_2, 0)^T$ is the magnetic field, $E = (0, 0, E_3)^T$ is the electric field, and h_1, h_2, e_3 are smooth functions. The boundary conditions will be specified below. Here the parameters (permeability, permittivity) are, e.g., $(\mu, \epsilon) = (\mu_1, \epsilon_1)$ for $z(x, y) < \xi_1$, $(\mu, \epsilon) = (\mu_i, \epsilon_i)$ for $\xi_{i-1} < z(x, y) < \xi_i$, $i = 2, 3, \cdots, n$ and $(\mu, \epsilon) = (\mu_{n+1}, \epsilon_{n+1})$ for $z(x, y) > \xi_n$ where ξ_i's are random variables and $\mu_i, \epsilon_i > 0$, μ_i's and ϵ_i's may be distinct (see Fig.5-1).

The randomness of the EM fields is inherited from the randomness of the locations of interfaces, i.e., it is uncertain where there are two or more different media interfaces (see [33], [41], [45], and [143]). In

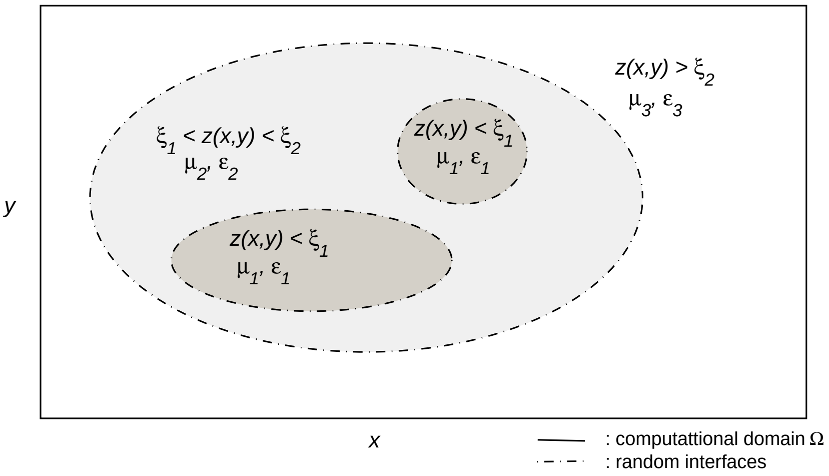


Figure 5-1: Two random interfaces for the model (5.1.1).

particular, the permeability and permittivity fluctuate randomly in space around their mean values.

For this type of problems, it has been demonstrated that the polynomial chaos expansion (PCE) methods are superior to Monte Carlo methods (e.g., [121]) in a number of applications (see e.g., [17], [18], [43], [44], [57], [86], [99], [152], [153], [155], [154]). The latter method depends on a sufficiently large number of sampling to obtain the results; whereas for the PCE methods, the solutions are projected into a random space spanned by the orthogonal polynomials whose arguments are the random variables, and are equipped with an inner product. In our problems, we choose shifted Legendre polynomials for the random variables ξ_i's following a uniform distribution.

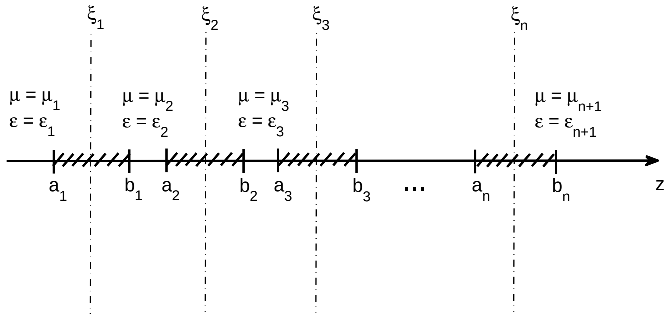


Figure 5-2: Multiple random interfaces $\{z(x, y) = \xi_i\}$ described by a level set function $z(x, y)$.

The EM fields with a single interface were studied and simulated in [68]. In this chapter, we extend our scope to discuss the case of two or multiple interfaces which are described as level sets, $\{z(x, y) = \xi_i\}$ where $z(x, y)$ is a function of x, y, and ξ_i is a random variable (see Fig.5-2). Here, we assume that ξ_i's are uniformly distributed over (a_i, b_i).

We note that the conventional PCE methods have some severe limitations. If the number of random variables increases, the computational cost will grow exponentially as indicated in the Polynomial chaos (PC) expansions (5.2.1), (5.2.9) below while the polynomials pertaining to each random variable are multiplied in a tensor product form. Thus the number of unknown coefficients (PC modes) grows exponentially and the computations are very expensive. Monte Carlo methods are then more feasible. To avoid this curse of dimensionality, along with the time explicit scheme, we will update the PC modes in each medium one by one. The computational cost then grows linearly as explained in Sec. 5.2.1.2 and 5.2.2 below.

The following discussions and results follow what presented in [69].

5.2 PCE Methods for Multiple Random Interfaces

5.2.1 2-random Interfaces

We begin with two random interfaces (see Fig. 5-1). We consider two cases. The first one is that the two random interfaces depend on a single random variable. In this case, a one-dimensional PC is used to approximate the randomness. The other one is that the two random interfaces depend on different random variables which are independent.

5.2.1.1 Case 1

In this section we consider the case $\xi_1 = \xi \in (a,b)$, $\xi_2 = \xi + \delta$, $\delta > b - a > 0$, where ξ is a uniform random variable over (a,b), for which its probability density function (PDF) is $f(\xi) = \frac{1}{b-a}\chi_{(a,b)}(\xi)$. Here χ is the corresponding characteristic function. To deal with the random interfaces $\{z(x,y) = \xi_1\}$ and $\{z(x,y) = \xi_2\}$, employing the so-called Legendre polynomial chaos (PC) we write:

$$\begin{cases} H_1 = \sum_{k=0}^{N} H_{1k}(x,y,t)P_k(\xi), \\ H_2 = \sum_{k=0}^{N} H_{2k}(x,y,t)P_k(\xi), \\ E_3 = \sum_{k=0}^{N} E_{3k}(x,y,t)P_k(\xi), \end{cases} \tag{5.2.1}$$

where $P_k = P_k^{a,b}$ are shifted Legendre polynomials. Here,

$$P_k(\xi) = P_k^{a,b}(\xi) = \widetilde{P}_k\left(\frac{2\xi - a - b}{b - a}\right) \tag{5.2.2}$$

where $\widetilde{P}_k$ are the standard Legendre polynomials with span $(-1,1)$.

Note that the number of PC modes in Eq. (5.2.1) does not grow in time and is fixed as N. This is because the fluctuations of the random variable(s) ξ or ξ_i below does not depend on t, x, y but depend only on the span of ξ. Substituting the PC expansions (5.2.1) in Eq. $(5.1.1)_1$, multiplying $P_i(\xi)$ and integrating over (a,b) in ξ we obtain the PC mode equations for H_{1i}:

$$\frac{\partial H_{1i}}{\partial t} = -\sum_{k=0}^{N} c^{\mu}_{ik} \frac{\partial E_{3k}}{\partial y}, \tag{5.2.3}$$

$$c^{\mu}_{ik} = \frac{\int_{-\infty}^{\infty} \frac{1}{\mu} P_i(\xi) P_k(\xi) \chi_{(a,b)}(\xi) d\xi}{\int_a^b P_i^2(\xi) d\xi}. \tag{5.2.4}$$

The PC mode equations for H_{2i}, E_{3i} similarly follow.

Depending on the value of z the coefficients $c^{\mu}_{ik} = c^{\mu}_{ik}(x,y)$ can be computed as follows. We note that $\mu = \mu_1$ for $z < \xi$, $\mu = \mu_2$ for $\xi < z < \xi + \delta$, and $\mu = \mu_3$ for $z > \xi + \delta$ (the parameter ε follows similarly). As in Fig. 5-2, $a_1 = a$, $b_1 = b$, $a_2 = a + \delta$ and $b_2 = b + \delta$ with $n = 2$. Using $\int_a^b P_i(\xi) P_k(\xi) d\xi = \dfrac{b-a}{2i+1} \delta_{ik}$, we first obtain that

$$c^{\mu}_{ik} = \delta_{ik} \begin{cases} \mu_1^{-1} & \text{if } z < a, \\ \mu_2^{-1} & \text{if } b < z < a + \delta, \\ \mu_3^{-1} & \text{if } z > b + \delta, \end{cases} \tag{5.2.5a}$$

and

$$c^{\mu}_{ik} = \begin{cases} \frac{1}{\mu_1}\delta_{ik} + \left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right) \frac{2i+1}{b-a} \int_a^z P_i(\xi) P_k(\xi) d\xi & \text{if } a < z < b, \\ \frac{1}{\mu_2}\delta_{ik} + \left(\frac{1}{\mu_3} - \frac{1}{\mu_2}\right) \frac{2i+1}{b-a} \int_a^{z-\delta} P_i(\xi) P_k(\xi) d\xi & \text{if } a + \delta < z < b + \delta, \end{cases} \tag{5.2.5b}$$

where the integrations $\int_a^z P_i(\xi) P_k(\xi) d\xi$ are given in explicit forms as in [68], [72] which can be evaluated with low computational cost.

We deduce the PC mode equations: for $\mathbf{H}_1(x,y,t) = (H_{10}, \cdots, H_{1N})$, $\mathbf{H}_2(x,y,t) = (H_{20}, \cdots, H_{2N})$ and $\mathbf{E}_3(x,y,t) = (E_{30}, \cdots, E_{3N})$,

$$\begin{cases} \mathbf{H}_{1t} = -\Lambda^{\mu}(z) \mathbf{E}_{3y}, \\ \mathbf{H}_{2t} = \Lambda^{\mu}(z) \mathbf{E}_{3x}, \\ \mathbf{E}_{3t} = \Lambda^{\epsilon}(z) \left(\mathbf{H}_{2x} - \mathbf{H}_{1y} \right), \end{cases} \tag{5.2.6}$$

where the initial conditions are $\mathbf{H}_1(x,y,0) = (h_1(x,y), 0, \cdots, 0)$, $\mathbf{H}_2(x,y,0) = (h_2(x,y), 0, \cdots, 0)$ and $\mathbf{E}_3(x,y,0) = (e_3(x,y), 0, \cdots, 0)$. The matrices $\Lambda^{\mu}(z) = \Lambda^{\mu}(z(x,y)) = (c^{\mu}_{ik})$ and $\Lambda^{\epsilon}(z) = \Lambda^{\epsilon}(z(x,y)) = (c^{\epsilon}_{ik})$, c^{ϵ}_{ik} is the c^{μ}_{ik} in (5.2.5) with μ_1 and μ_2 being replaced by, respectively, ϵ_1 and ϵ_2. Note that the entries c^{μ}_{ik}, c^{ϵ}_{ik} are continuous in z.

To compute the solutions of the system (5.2.6), we apply the Finite-Difference Time-Domain (FDTD) method by K. Yee (see [156], [140], and [68]). 2^{nd}-order centered finite difference is employed on a staggered grid for the space discretizations of the magnetic and electric fields. On this staggered grid, the magnetic and electric fields are located on the sides and at the center of each grid cell, respectively. The time derivatives are approximated in a same manner: the magnetic fields are first updated to the half time step; then the electric field is marched to the next time step based on the updated values of the former fields. For our Eqs. (5.2.6), we use the following notations for the magnetic fields $\mathbf{H}_{1,2}$ and electric field $\mathbf{E}_3$ at the node points on the staggered grid as in Fig. 5-10: for $I, J, n \in \mathbb{Z}^+$, on the space domain $\Omega = (-2,2) \times (0,1)$,

$$\begin{cases} \mathbf{H}^{n-1/2}_{1,i,j+1/2} = \mathbf{H}_1(x_i, y_{j+1/2}, t_{n-1/2}), & i = 0:I,\ j = 0:J-1, \\ \mathbf{H}^{n-1/2}_{2,i+1/2,j} = \mathbf{H}_2(x_{i+1/2}, y_j, t_{n-1/2}), & i = 0:I-1,\ j = 0:J, \\ \mathbf{E}^{n}_{3,i,j} = \mathbf{E}_3(x_i, y_j, t_n), & i = 0:I, j = 0:J, \end{cases} \tag{5.2.7}$$

where $x_p = -2 + p\Delta x,\ y_q = q\Delta y$ and $t_r = r\Delta t,\ p,q,r \in \mathbb{Z}^+,\ p \in [0,I],\ q \in [0,J],\ r \in [0,T]$; and $\Delta x = 4/I$, $\Delta y = 1/J$ are the space steps and Δt is the time step. We note that $\mathbf{H}_{1,2}$, $\mathbf{E}_3$ are vectors of the deterministic PC modes defined in Eqs. (5.2.6). We then discretize the system (5.2.6) as follows: for $n \in \mathbf{Z}^+$,

$$\begin{cases} \mathbf{H}^{n+1/2}_{1,i,j+1/2} = \mathbf{H}^{n-1/2}_{1,i,j+1/2} - \dfrac{\Delta t}{\Delta y}\Lambda^{\mu}_{i,j+1/2}\left[\mathbf{E}^n_{3,i,j+1} - \mathbf{E}^n_{3,i,j}\right], \\ \quad i = 1:I-1,\ j = 0:J-1, \\ \mathbf{H}^{n+1/2}_{2,i+1/2,j} = \mathbf{H}^{n-1/2}_{2,i+1/2,j} + \dfrac{\Delta t}{\Delta x}\Lambda^{\mu}_{i+1/2,j}\left[\mathbf{E}^n_{3,i+1,j} - \mathbf{E}^n_{3,i,j}\right], \\ \quad i = 0:I-1,\ j = 1:J-1, \\ \mathbf{E}^{n+1}_{3,i,j} = \mathbf{E}^n_{3,i,j} + \Lambda^{\epsilon}_{i,j}\left\{\dfrac{\Delta t}{\Delta x}\left[\mathbf{H}^{n+1/2}_{2,i+1/2,j} - \mathbf{H}^{n+1/2}_{2,i-1/2,j}\right] - \dfrac{\Delta t}{\Delta y}\left[\mathbf{H}^{n+1/2}_{1,i,j+1/2} - \mathbf{H}^{n+1/2}_{1,i,j-1/2}\right]\right\}, \\ \quad i = 1:I-1,\ j = 1:J-1, \end{cases} \tag{5.2.8}$$

where $\Lambda^{\mu}_{i,j+1/2} = \Lambda^{\mu}(z(x_i, y_{j+1/2}))$, $\Lambda^{\mu}_{i+1/2,j} = \Lambda^{\mu}(z(x_{i+1/2}, y_j))$ and $\Lambda^{\epsilon}_{i,j} = \Lambda^{\epsilon}(z(x_i, y_j))$.

5.2.1.2 Case 2

In this section we consider two random variables which determine two random interfaces, respectively. The random variables $\xi_i \in (a_i, b_i)$, $i = 1, 2$, are uniformly distributed over (a_i, b_i) where $a_1 < \xi_1 < b_1 < a_2 < \xi_2 < b_2$, and the PDFs are, respectively, $f_1(\xi_1) = \frac{1}{b_1 - a_1}\chi_{(a_1,b_1)}(\xi_1),\ f_2(\xi_2) = \frac{1}{b_2 - a_2}\chi_{(a_2,b_2)}(\xi_2)$.

To deal with the random interfaces $\{z(x,y) = \xi_1\}$ and $\{z(x,y) = \xi_2\}$, employing the PC expansions we write:

$$\begin{cases} H_1 = \displaystyle\sum_{k=0}^{N_1}\sum_{l=0}^{N_2} H_{1kl}(x,y,t) P^{a_1,b_1}_k(\xi_1) P^{a_2,b_2}_l(\xi_2), \\ H_2 = \displaystyle\sum_{k=0}^{N_1}\sum_{l=0}^{N_2} H_{2kl}(x,y,t) P^{a_1,b_1}_k(\xi_1) P^{a_2,b_2}_l(\xi_2), \\ E_3 = \displaystyle\sum_{k=0}^{N_1}\sum_{l=0}^{N_2} E_{3kl}(x,y,t) P^{a_1,b_1}_k(\xi_1) P^{a_2,b_2}_l(\xi_2). \end{cases} \tag{5.2.9}$$

Substituting the PC expansions in (5.1.1), multiplying $P_i(\xi_1)P_j(\xi_2)$ and integrating over $(a_1, b_1) \times$

(a_2, b_2) we obtain the following modal equations:

$$\frac{\partial H_{1ij}}{\partial t} = -\sum_{k=0}^{N_1}\sum_{l=0}^{N_2} c^{\mu}_{ijkl}\frac{\partial E_{3kl}}{\partial y}, \tag{5.2.10}$$

$$c^{\mu}_{ijkl} = \frac{\int_{\mathbb{R}^2}\frac{1}{\mu}P_i(\xi_1)P_k(\xi_1)\chi_{(a_1,b_1)}(\xi_1)P_j(\xi_2)P_l(\xi_2)\chi_{(a_2,b_2)}(\xi_2)d\xi_1 d\xi_2}{\int_{a_1}^{b_1}P_i^2(\xi_1)d\xi_1\int_{a_2}^{b_2}P_i^2(\xi_2)d\xi_2}. \tag{5.2.11}$$

Then the coefficients c^{μ}_{ijkl} are evaluated as follows:

$$c^{\mu}_{ijkl} = \delta_{ik}\delta_{jl}\begin{cases}\mu_1^{-1} & \text{if } z < a_1,\\ \mu_2^{-1} & \text{if } b_1 < z < a_2,\\ \mu_3^{-1} & \text{if } z > b_2,\end{cases} \tag{5.2.12a}$$

and

$$c^{\mu}_{ijkl} = \begin{cases}\delta_{jl}\left[\frac{1}{\mu_1}\delta_{ik} + \left(\frac{1}{\mu_2}-\frac{1}{\mu_1}\right)\frac{2i+1}{b_1-a_1}\int_{a_1}^{z}P_i(\xi_1)P_k(\xi_1)d\xi_1\right] & \text{if } a_1 < z < b_1,\\ \delta_{ik}\left[\frac{1}{\mu_2}\delta_{jl} + \left(\frac{1}{\mu_3}-\frac{1}{\mu_2}\right)\frac{2j+1}{b_2-a_2}\int_{a_2}^{z}P_j(\xi_2)P_l(\xi_2)d\xi_2\right] & \text{if } a_2 < z < b_2,\end{cases} \tag{5.2.12b}$$

Hence we similarly deduce the PC mode equations (5.2.6) and the Yee (FDTD) scheme (5.2.8). Here the corresponding vectors are $\mathbf{H}_1(x,y,t) = (H_{1,ij})$ (the indices i, j are in a dictionary order), $\mathbf{H}_2(x,y,t) = (H_{2,ij})$ and $\mathbf{E}_3(x,y,t) = (E_{3,ij})$, and the initial conditions are $H_{1,00} = h_1(x,y)$, $H_{2,00} = h_2(x,y)$, and $E_{3,00} = e_3(x,y)$. The matrices $\Lambda^{\mu}(z) = (c^{\mu}_{ijkl})$ and $\Lambda^{\epsilon}(z) = (c^{\epsilon}_{ijkl})$.

One disadvantage of the modal equation (5.2.10) is that it leads to a huge algebraic system when the Yee scheme (5.2.8) is applied since all the PC modes are arranged into vector forms, which in turn causes the coefficient matrices Λ^{μ} and Λ^{ε} to have very large sizes, i.e., $(1+N_1)(1+N_2)\times(1+N_1)(1+N_2)$. The case becomes worse when more interfaces are introduced into the modal problem. In the following part, we introduce a modification in deriving the PC modes which overcomes the drawback discussed above.

Substituting the PC expansions (5.2.9) in (5.1.1), multiplying by $P_i(\xi_1)P_j(\xi_2)$ and integrating over Ω, if $a_1 \le z < b_1$, since $\mu = \mu(\xi_1)$, $\varepsilon = \varepsilon(\xi_1)$ (see Fig. 5-1), ξ_2 does not affect the parameters μ, ε and thus the integration with respect to ξ_2 has no effect in the resulting equations. Hence, we obtain that

$$\frac{\partial H_{1il}}{\partial t} = -c_i\sum_{k=1}^{N_1}\lambda_{ik}\frac{\partial E_{3kl}}{\partial y}, \quad i = 0\cdots N_1, \tag{5.2.13}$$

where

$$c_i = \left[\int_{a_1}^{b_1}(P_i(\xi_1))^2 d\xi_1\right]^{-1} = \frac{2i+1}{b_1-a_1}, \tag{5.2.14}$$

and

$$\lambda_{ik} = \lambda^{\mu}_{ik} = \int_{a_1}^{b_1}\frac{1}{\mu}P_k(\xi_1)P_i(\xi_1)d\xi_1. \tag{5.2.15}$$

For other intervals, a similar procedure follows. Since $\lambda_{ik}^{\mu} = \lambda_{ki}^{\mu}$, c_i and λ_{ik}^{μ} do not change when we march the indexes, after the marching, we obtain,

$$\begin{cases} \dfrac{\partial \mathbf{H}_1}{\partial t} = -\dfrac{1}{\mu_1}\dfrac{\partial \mathbf{E}_3}{\partial y}, & z < a_1, \\ \dfrac{\partial \mathbf{H}_1}{\partial t} = -\Lambda_1^{\mu}\dfrac{\partial \mathbf{E}_3}{\partial y}, & a_1 \leq z < b_1, \\ \dfrac{\partial \mathbf{H}_1}{\partial t} = -\dfrac{1}{\mu_2}\dfrac{\partial \mathbf{E}_3}{\partial y}, & b_1 \leq z < a_2, \\ \dfrac{\partial \mathbf{H}_1}{\partial t} = -\Lambda_2^{\mu}\dfrac{\partial \mathbf{E}_3}{\partial y}, & a_2 \leq z < b_2, \\ \dfrac{\partial \mathbf{H}_1}{\partial t} = -\dfrac{1}{\mu_3}\dfrac{\partial \mathbf{E}_3}{\partial y}, & b_2 \leq z, \end{cases} \tag{5.2.16}$$

where $\mathbf{H}_1 = (H_{1kl})$, $\mathbf{E}_3 = (E_{3kl})$, $k = 1, \cdots, N_1$, $l = 1, \cdots, N_2$; and $\Lambda_{1,2}^{\mu} = (c_k\lambda_{kl})_{1,2}$ are matrices of sizes $(1+N_1)\times(1+N_1)$ and $(1+N_2)\times(1+N_2)$, respectively. The entries $(c_k\lambda_{kl})^{1,2}$ are similar to the coefficients c_{ik}^{μ} defined in (5.2.5b), respectively with the removal of δ_{ik} and a, b replaced by appropriate a_1, b_1 and a_2, b_2, respectively. We notice that the sizes of $\Lambda_{(.)}^{\mu}$ are much reduced. Moreover, the computational cost is much reduced in case more random interfaces are introduced. H_2, E_3 follow similarly.

Remark 5.2.1.

It is noted that in the system (5.2.16), we treat the matrices $\mathbf{H}_1$, $\mathbf{E}_3$ as vectors by fixing one index and marching the other. Order of the indexes is important for intervals containing the random interfaces ξ_1, ξ_2, i.e., (a_1, b_1) and (a_2, b_2), respectively. In particular, for the interval (a_1, b_1), by fixing l, we march $\mathbf{H}_1$ by columns and solve for each column vector $(H_{1il}),\ i = 0, \cdots, N_1$ as in Eq. (5.2.13); whereas for the interval (a_2, b_2), we fix k and solve for each row vector $(H_{1kj}),\ j = 0, \cdots, N_2$.

Applying the Yee scheme to (5.2.16), we obtain an algebraic system as below, for $n, i, j \in \mathbb{Z}^+$,

$$\left\{\begin{aligned}
&\bullet \text{ if } z < a_1,\\
&\mathbf{H}^{n+\frac12}_{1,i,j+\frac12} = \mathbf{H}^{n-\frac12}_{1,i,j+\frac12} - \frac{\Delta t}{\Delta y}\frac{1}{\mu_1}\left[\mathbf{E}^n_{3,i,j+1} - \mathbf{E}^n_{3,i,j}\right],\\
&\mathbf{H}^{n+\frac12}_{2,i+\frac12,j} = \mathbf{H}^{n-\frac12}_{2,i+\frac12,j} + \frac{\Delta t}{\Delta y}\frac{1}{\mu_1}\left[\mathbf{E}^n_{3,i+1,j} - \mathbf{E}^n_{3,i,j}\right],\\
&\mathbf{E}^{n+1}_{1,i,j} = \mathbf{E}^n_{1,i,j} + \frac{1}{\epsilon_1}\left\{\frac{\Delta t}{\Delta x}\left[\mathbf{H}^{n+\frac12}_{2,i+\frac12,j} - \mathbf{H}^{n+\frac12}_{2,i-\frac12,j}\right] - \frac{\Delta t}{\Delta y}\left[\mathbf{H}^{n+\frac12}_{1,i,j+\frac12} - \mathbf{H}^{n+\frac12}_{1,i,j-\frac12}\right]\right\},\\
&\bullet \text{ if } a_1 \le z < b_1,\\
&\mathbf{H}^{n+\frac12}_{1,i,j+\frac12} = \mathbf{H}^{n-\frac12}_{1,i,j+\frac12} - \frac{\Delta t}{\Delta y}\Lambda^\mu_{1,i,j+\frac12}\left[\mathbf{E}^n_{3,i,j+1} - \mathbf{E}^n_{3,i,j}\right],\\
&\mathbf{H}^{n+\frac12}_{2,i+\frac12,j} = \mathbf{H}^{n-\frac12}_{2,i+\frac12,j} + \frac{\Delta t}{\Delta y}\Lambda^\mu_{1,i+\frac12,j}\left[\mathbf{E}^n_{3,i+1,j} - \mathbf{E}^n_{3,i,j}\right],\\
&\mathbf{E}^{n+1}_{1,i,j} = \mathbf{E}^n_{1,i,j} + \Lambda^\epsilon_{1,i,j}\left\{\frac{\Delta t}{\Delta x}\left[\mathbf{H}^{n+\frac12}_{2,i+\frac12,j} - \mathbf{H}^{n+\frac12}_{2,i-\frac12,j}\right] - \frac{\Delta t}{\Delta y}\left[\mathbf{H}^{n+\frac12}_{1,i,j+\frac12} - \mathbf{H}^{n+\frac12}_{1,i,j-\frac12}\right]\right\},\\
&\bullet \text{ if } b_1 \le z < a_2,\\
&\mathbf{H}^{n+\frac12}_{1,i,j+\frac12} = \mathbf{H}^{n-\frac12}_{1,i,j+\frac12} - \frac{\Delta t}{\Delta y}\frac{1}{\mu_2}\left[\mathbf{E}^n_{3,i,j+1} - \mathbf{E}^n_{3,i,j}\right],\\
&\mathbf{H}^{n+\frac12}_{2,i+\frac12,j} = \mathbf{H}^{n-\frac12}_{2,i+\frac12,j} + \frac{\Delta t}{\Delta y}\frac{1}{\mu_2}\left[\mathbf{E}^n_{3,i+1,j} - \mathbf{E}^n_{3,i,j}\right],\\
&\mathbf{E}^{n+1}_{1,i,j} = \mathbf{E}^n_{1,i,j} + \frac{1}{\epsilon_2}\left\{\frac{\Delta t}{\Delta x}\left[\mathbf{H}^{n+\frac12}_{2,i+\frac12,j} - \mathbf{H}^{n+\frac12}_{2,i-\frac12,j}\right] - \frac{\Delta t}{\Delta y}\left[\mathbf{H}^{n+\frac12}_{1,i,j+\frac12} - \mathbf{H}^{n+\frac12}_{1,i,j-\frac12}\right]\right\},\\
&\bullet \text{ if } a_2 \le z < b_2,\\
&\mathbf{H}^{n+\frac12}_{1,i,j+\frac12} = \mathbf{H}^{n-\frac12}_{1,i,j+\frac12} - \frac{\Delta t}{\Delta y}\Lambda^\mu_{2,i,j+\frac12}\left[\mathbf{E}^n_{3,i,j+1} - \mathbf{E}^n_{3,i,j}\right],\\
&\mathbf{H}^{n+\frac12}_{2,i+\frac12,j} = \mathbf{H}^{n-\frac12}_{2,i+\frac12,j} + \frac{\Delta t}{\Delta y}\Lambda^\mu_{2,i+\frac12,j}\left[\mathbf{E}^n_{3,i+1,j} - \mathbf{E}^n_{3,i,j}\right],\\
&\mathbf{E}^{n+1}_{1,i,j} = \mathbf{E}^n_{1,i,j} + \Lambda^\epsilon_{2,i,j}\left\{\frac{\Delta t}{\Delta x}\left[\mathbf{H}^{n+\frac12}_{2,i+\frac12,j} - \mathbf{H}^{n+\frac12}_{2,i-\frac12,j}\right] - \frac{\Delta t}{\Delta y}\left[\mathbf{H}^{n+\frac12}_{1,i,j+\frac12} - \mathbf{H}^{n+\frac12}_{1,i,j-\frac12}\right]\right\},\\
&\bullet \text{ if } z \ge b_2,\\
&\mathbf{H}^{n+\frac12}_{1,i,j+\frac12} = \mathbf{H}^{n-\frac12}_{1,i,j+\frac12} - \frac{\Delta t}{\Delta y}\frac{1}{\mu_3}\left[\mathbf{E}^n_{3,i,j+1} - \mathbf{E}^n_{3,i,j}\right],\\
&\mathbf{H}^{n+\frac12}_{2,i+\frac12,j} = \mathbf{H}^{n-\frac12}_{2,i+\frac12,j} + \frac{\Delta t}{\Delta y}\frac{1}{\mu_3}\left[\mathbf{E}^n_{3,i+1,j} - \mathbf{E}^n_{3,i,j}\right],\\
&\mathbf{E}^{n+1}_{1,i,j} = \mathbf{E}^n_{1,i,j} + \frac{1}{\epsilon_3}\left\{\frac{\Delta t}{\Delta x}\left[\mathbf{H}^{n+\frac12}_{2,i+\frac12,j} - \mathbf{H}^{n+\frac12}_{2,i-\frac12,j}\right] - \frac{\Delta t}{\Delta y}\left[\mathbf{H}^{n+\frac12}_{1,i,j+\frac12} - \mathbf{H}^{n+\frac12}_{1,i,j-\frac12}\right]\right\},
\end{aligned}\right. \tag{5.2.17}$$

where $\Lambda^\mu_{(.),i,j+\frac12} = \Lambda^\mu(z(x_i, y_{j+\frac12}))$, $\Lambda^\mu_{(.),i+\frac12,j} = \Lambda^\mu(z(x_{i+\frac12}, y_j))$, and $\Lambda^\epsilon_{.,i,j} = \Lambda^\epsilon(z(x_i, y_j))$.

Notice that the coefficient matrices $\Lambda^{(.)}_1(z)$ and $\Lambda^{(.)}_2(z)$ are independent of time. We just store those matrices only at the first step $n = 1$ and reuse them at $n \ge 2$.

5.2.2 n-random Interfaces

We now extend the PC expansions (5.2.9) to the case with n-interfaces. Substituting the expansion

$$H_1 = \sum H_{1j_1\dots j_n}(x, y, t)P_{j_1}(\xi_1)\dots P_{j_n}(\xi_n), \tag{5.2.18}$$

into Eq. (5.1.1)$_1$, we obtain

$$\sum \frac{\partial}{\partial t} H_{1j_1\dots j_n}(x,y,t)P_{j_1}(\xi_1)\dots P_{j_n}(\xi_n) = -\frac{1}{\mu}\sum \frac{\partial}{\partial t} E_{3j_1\dots j_n}(x,y,t)P_{j_1}(\xi_1)\dots P_{j_n}(\xi_n). \tag{5.2.19}$$

Here, $\sum = \sum_{j_1=1}^{N_1}\sum_{j_2=1}^{N_2}\dots\sum_{j_n=1}^{N_n}$ and $\mu = \mu_i$ if $\xi_{i-1} \le z < \xi_i$ with $\xi_0 = -\infty$, $\xi_{n+1} = \infty$ as in Fig. 5-2. Here, the random variables ξ_i are uniformly distributed over (a_i, b_i). The PDFs are respectively $f_i(\xi_i) = \frac{1}{b_i - a_i}\chi_{(a_i,b_i)}(\xi_i)$. Multiplying Eq. (5.2.19) by $P_{j_1}(\xi_i)\dots P_{j_n}(\xi_n)$ and integrating over Ω, we find that, if $a_1 \le z < b_1$,

$$\frac{\partial}{\partial t} H_{1i_1\dots i_n}(x,y,t) = -c_{i_1}\sum_{j_1=1}^{N_1}\lambda_{j_1}\frac{\partial}{\partial y}E_{3j_1,i_2\dots i_n}(x,y,t), \tag{5.2.20}$$

where

$$c_{i_1} = \left[\int_{a_1}^{b_1}(P_{i_1}(\xi_1))^2 d\xi_1\right]^{-1} = \frac{2i_1+1}{b_1-a_1}, \tag{5.2.21}$$

and

$$\begin{aligned} \lambda_{j_1} = \lambda_{i1,j1} &= \int_{a_1}^{z}\frac{1}{\mu}P_{j_1}(\xi_1)P_{i_1}(\xi_1)d\xi_1 \\ &= \frac{1}{\mu_2}\int_{a_1}^{b_1}P_{j1}(\xi_1)P_{i1}(\xi_1)d\xi_1 + \frac{1}{\mu_1}\int_{z}^{b_1}P_{j1}(\xi_1)P_{i1}(\xi_1)d\xi_1. \end{aligned} \tag{5.2.22}$$

Note here that although we consider the n-tensor product of one-dimensional PC expansions, only an integration in one dimension needs to be computed to evaluate the coefficients in Eq. (5.2.20) and this makes the whole computational cost grows linearly. Furthermore, the explicit formula of the integration in Eq. (5.2.22) is available in [68].

For other intervals, the same procedure can be applied. In particular, if $b_1 \le z < a_2$, $\mu = \mu_2$, and hence

$$\frac{\partial}{\partial t} H_{i_1\dots i_n}(x,y,t) = -\frac{1}{\mu_2}\frac{\partial}{\partial y}E_{i_1,i_2\dots i_n}(x,y,t), \tag{5.2.23}$$

and H_2, E_3 follow similarly.

We can now discretize (5.2.20), (5.2.23) using the Yee scheme as in (5.2.17). The numerical simulations with 3 and 5 random interfaces are shown in Figs. 5-9, 5-12 below.

5.3 Statistics

Thanks to the orthogonality of Legendre polynomials, we can explicitly obtain the mean and variance. For example, we can obtain the mean and variance of E_3: for Case I, as in section 1.1.1,

$$\mathbb{E}(E_3) = \int_{-\infty}^{\infty} E_3 f_1(\xi)d\xi = E_{3,0}(x,y,t), \tag{5.3.1}$$

$$Var(E_3) = \int_{-\infty}^{\infty} (E_3)^2 f_1(\xi)d\xi - (\mathbb{E}(E_3))^2 = \sum_{i=1}^{N} \frac{(E_{3,i}(x,t))^2}{2i+1}. \tag{5.3.2}$$

and, in general,

$$\mathbb{E}(E_3) = \int_{\mathbb{R}^n} E_3 \prod_{i=1}^{n} f_i(\xi_i)d\xi_1\cdots d\xi_n = E_{3,0\cdots0}(x,y,t), \tag{5.3.3}$$

$$\begin{aligned} Var(E_3) &= \int_{\mathbb{R}^n} (E_3)^2 \prod_{i=1}^{n} f_i(\xi_i)d\xi_1\cdots d\xi_n - (\mathbb{E}(E_3))^2 \\ &= \sum_{\substack{i_1=0 \\ (i_1,\cdots,i_n)\neq(0,\cdots,0)}}^{N_1} \sum_{i_2=0}^{N_2} \cdots \sum_{i_n=0}^{N_n} \frac{(E_{3,i_1 i_2\cdots i_n}(x,y,t))^2}{\Pi_{k=i_1}^{i_n}(2i_k+1)}. \end{aligned} \tag{5.3.4}$$

Here, $f_i(\xi_i) = \frac{1}{b_i - a_i}\chi_{(a_i,b_i)}(\xi_i)$ are uniform probability density functions.

The Cumulative distribution functions (CDF), and the Probability density functions (PDF) can be also evaluated via the PC expansions and the post-processing Monte Carlo method. We randomly choose a Monte Carlo sampling $\{E_3^1, \ldots, E_3^k, \ldots, E_3^M\}$ with $E_3^k \in (s_0, s_K)$, for all $1 \leq k \leq M$. Here, M is sufficiently large, and k is the multi-index of the random variables xi_i's. The CDF and PDF are evaluated as follows (see [68] for detailed discussions), respectively,

$$F(s) = P(E_3 \leq s) = \int_{-\infty}^{s} f(E_3)dE_3 \approx \frac{1}{M}\sum_{k=1}^{M} \chi_{(-\infty,s)}(E_3^k), \tag{5.3.5}$$

$$f(s_k) \approx \frac{F(s_{k+1} - F(s_k))}{h}, \tag{5.3.6}$$

where $s_k = s_0 + kh$.

5.4 Numerical Results

For the simulation purposes, we use the level set function z as below,

$$z(x,y) = \frac{3}{4}x^2 + 12\left(y - \frac{1}{2}\right)^2 - \frac{3}{2}, \tag{5.4.1}$$

for which two random interfaces $\{z(x,y) = \xi_1\}$ and $\{z(x,y) = \xi_2\}$, where ξ_1, ξ_2 are uniform random variables over (a_1, b_1) and (a_2, b_2), respectively, are plotted in Fig. 5-3. The values of a_i, b_i, $i = 1, 2$, are

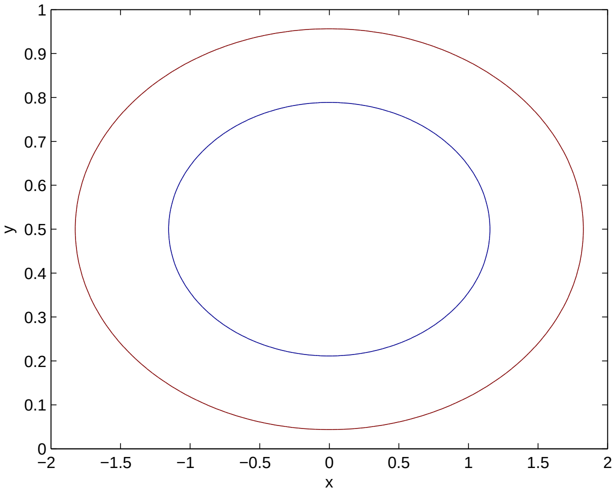


Figure 5-3: Two random interfaces $\{z(x,y) = \xi_1\}$ and $\{z(x,y) = \xi_2\}$; the case $\xi_1 = -0.5$ (inner) and $\xi_2 = 1$ (outer) are plotted.

specified in the simulations below.

Here and after, we use the computational domain $\Omega = (-2,2) \times (0,1)$. In Figs. 5-4 and 5-5, we consider the initial conditions:

$$H_1 = H_2 = 0,\ E_3 = (x+2)(x-2)y(y-1), \tag{5.4.2}$$

boundary conditions

$$E_3 = 0, \tag{5.4.3}$$

and the parameters,

$$\begin{cases} \mu_1 = \varepsilon_1 = 1 & \text{if } z < \xi_1, \\ \mu_2 = \varepsilon_2 = 2 & \text{if } \xi_1 \le z < \xi_2, \\ \mu_3 = \varepsilon_3 = 3 & \text{if } \xi_2 \le z < \xi_3. \end{cases} \tag{5.4.4}$$

We notice that the boundary conditions for H_1, H_2 are not needed since following the Yee scheme (5.2.17), H_1, H_2 are first updated using the information of the previous time step of E_3; then E_3 is updated by the newly obtained H_1, H_2.

We consider the case 1 in section 1.1.1 of two random interfaces fluctuating with the same random parameter ξ and thickness δ between the two level sets. Two cases are taken into account. One for a small fluctuation with $a = -0.5$, $b = 0$, $\delta = 0.6$; and $a = -1$, $b = 0.8$, $\delta = 2$ for a large fluctuation. We observe the CDFs of the Maxwell solutions evolving in time for both cases as in Fig. 5-4. It is shown that the fluctuation of the CDF of E_3 for the latter case is much wider than the former case. This is due

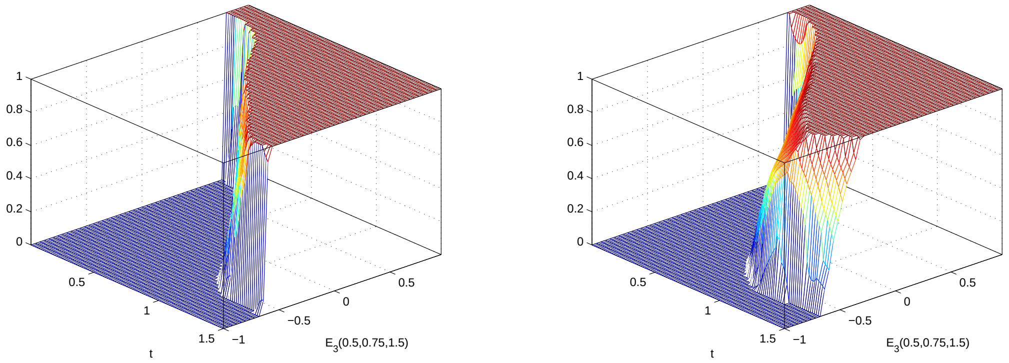


Figure 5-4: Case I: Evolution of Cumulative distribution function (CDF) of E_3 at $x = 0, y = 0.75$ with $N = 10$, (μ_i, ε_i), $\mu_1 = \varepsilon_1 = 1$, $\mu_2 = \varepsilon_2 = 2$ and $\mu_3 = \varepsilon_3 = 3$; LEFT: with a relatively small fluctuation ($a = -0.5, b = 0, \delta = 0.6$); RIGHT: with a relatively large fluctuation ($a = -1, b = 0.8, \delta = 2$).

to the wider variance of the random variable ξ in the latter case compared with that in the former one.

We now test with case 2 in which the two random variables ξ_1, ξ_2 are independent of each other. In this case, we use the intervals $a_1 = -1$, $b_1 = -0.1$, $a_2 = 0$, $b_2 = 0.9$. The CDF, mean and variance of E_3 at $t = 1.5$ are plotted in Fig. 5-5. Here, we also observe the fluctuation of the CDF of the field E_3 when it evolves in time; and the fluctuation occurs in the regions locating the random interfaces (see Fig. 5-5$_3$ showing the variance of E_3).

In Fig. 5-6, for the compatible initial/boundary conditions, we use some known solutions for (5.1.1) as follows:

$$\begin{pmatrix} H_1 \\ H_2 \\ E_3 \end{pmatrix} = \begin{pmatrix} -\beta \\ \alpha \\ \sqrt{\dfrac{\mu}{\epsilon}} \end{pmatrix} \exp\left(i\omega \left(\frac{t}{\sqrt{\epsilon\mu}} + \alpha x + \beta y + \gamma \right) \right), \tag{5.4.5}$$

where $\alpha, \beta, \gamma, \omega \in \mathbb{R}$, $\alpha^2 + \beta^2 = 1$. We can then choose the compatible boundary conditions which satisfy (5.4.5) on $\partial\Omega$ and the compatible initial conditions (5.4.5) with $t = 0$. We try the real part of solutions (5.4.5) with $\alpha = \beta = 1/\sqrt{2}$, $\epsilon = \mu = 2$ and $\omega = 2$, $\gamma = 0$. Hence the initial conditions are: for all $(x, y) \in \Omega$,

$$\left\{ \begin{array}{l} \begin{pmatrix} H_{1,00} \\ H_{2,00} \\ E_{3,00} \end{pmatrix} = \begin{pmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \\ 1 \end{pmatrix} cos\left(\sqrt{2}x + \sqrt{2}y\right), \\ H_{1kl} = H_{2kl} = E_{3kl} = 0 \quad \text{for } k, l = 1, \cdots, N. \end{array} \right. \tag{5.4.6}$$

Here we set the parameters $\mu_1 = \varepsilon_1 = 1$, $\mu_2 = \varepsilon_2 = 3$, $\mu_3 = \varepsilon_3 = 2$. Since $\partial\Omega \subset \{z(x, y) = \xi\}$ for all $\xi \in [-1, 1]$, the conditions (5.4.6) are compatible with the boundary conditions. For the boundary

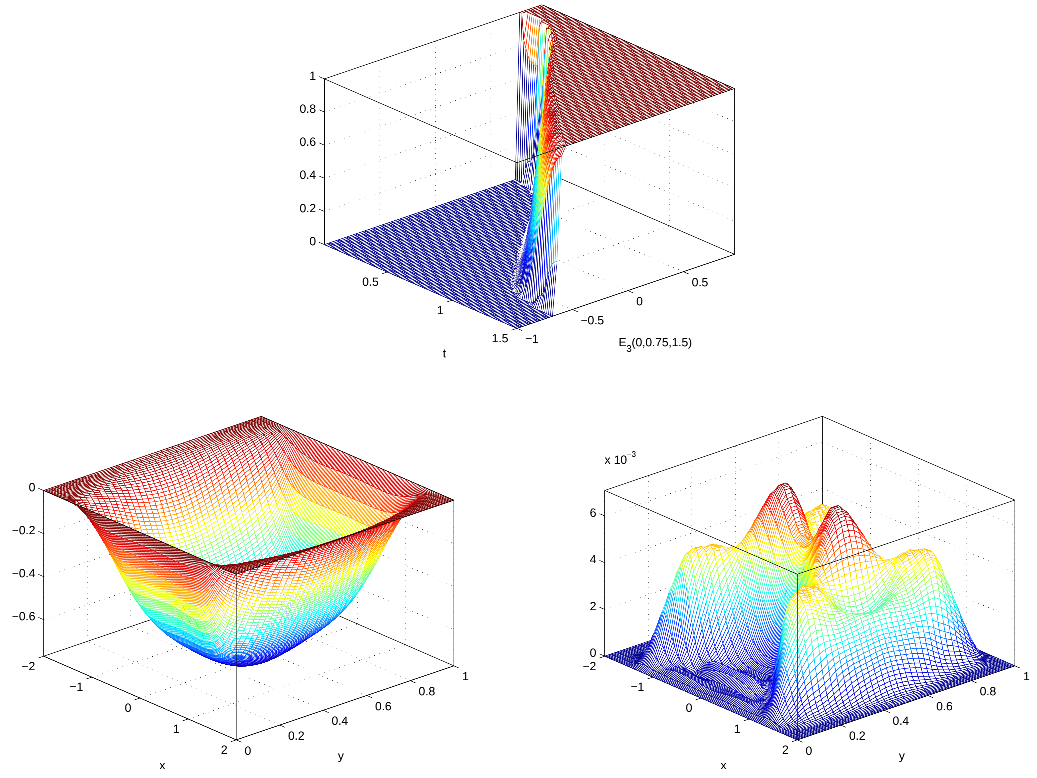


Figure 5-5: Case II: TOP: Evolution of Cumulative distribution function (CDF) of E_3 at $x = 0, y = 0.75$ with $N = 10$, initial and boundary conditions as in (5.4.2) and (5.4.3); BOTTOM: mean and variance of E_3 at $t = 1.5$, respectively.

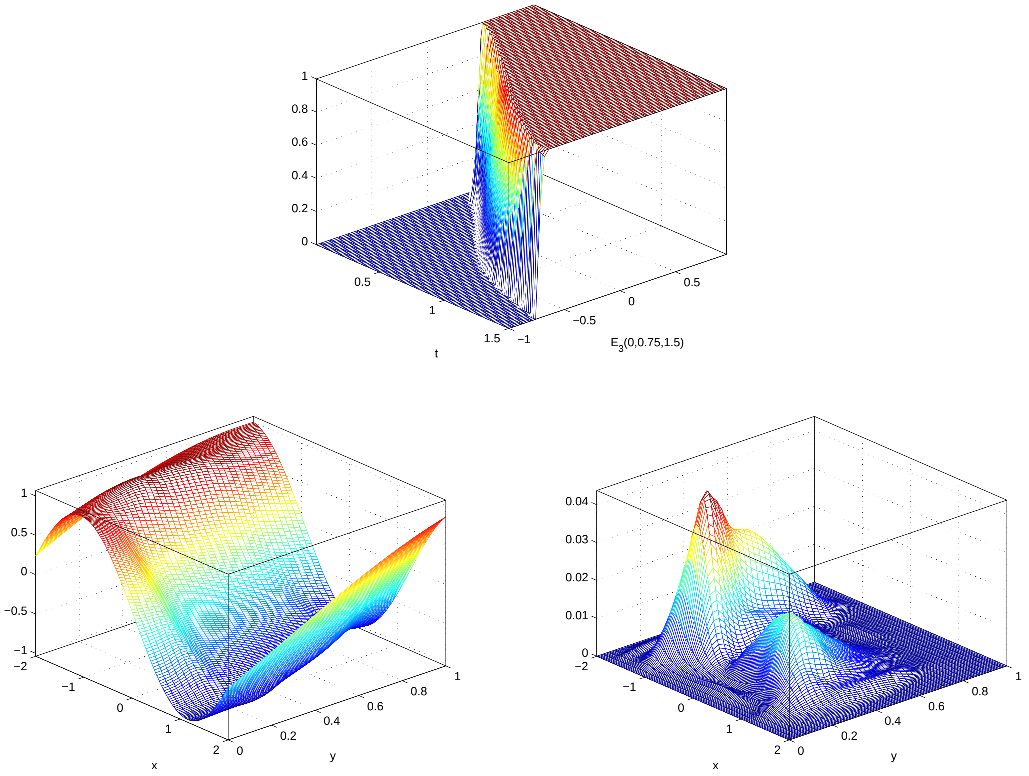


Figure 5-6: Case II: TOP: Evolution of Cumulative distribution function (CDF) of E_3 at $x = 0, y = 0.75$ with $N = 10$ using compatible initial/boundary conditions (5.4.6) and (5.4.7), μ and ε as in (5.4.4); BOTTOM: mean and variance of E_3 at $t = 1.5$, respectively.

conditions, we only have to impose on E_3 field as indicated in (5.2.8) and (5.2.17) (i.e., the H- fields are updated from the E- field): for all (x,y) $\in \partial\Omega,\ t > 0$,

$$\begin{cases} E_{3,00} = cos\left(t + \sqrt{2}x + \sqrt{2}y\right), \\ E_{3,kl} = 0, \quad \text{for } k, l = 1, \cdots, N. \end{cases} \tag{5.4.7}$$

Notice also that the initial/boundary conditions are deterministic and thus only the conditions of the zero PC mode are imposed as in (5.4.7) and the conditions of other PC modes are zero.

The evolution in time of the electric field E_3 for the case 2 of the compatible initial and boundary conditions (5.4.6)–(5.4.7) is then simulated and the results are shown in Fig. 5-7 at time $t = 0, 1, 2, 3$ respectively with two random interfaces. It is shown in the figures that even though we have deterministic initial and boundary conditions, the mean of E_3 becomes rough in the regions where the random interfaces fluctuate. This can be explained by the discontinuity of the parameters μ and ε across the interfaces.

We also check the decay of the PC modes of E_3 for both cases of initial/boundary conditions (5.4.2)–(5.4.6). The results are presented in Fig. 5-8 with the z-axis is plotted in log scale. We notice that the PC modes decay exponentially. This implies that computations with only the first few modes, e.g., $N = 10$ in our numerical simulations, can produce a good approximation.

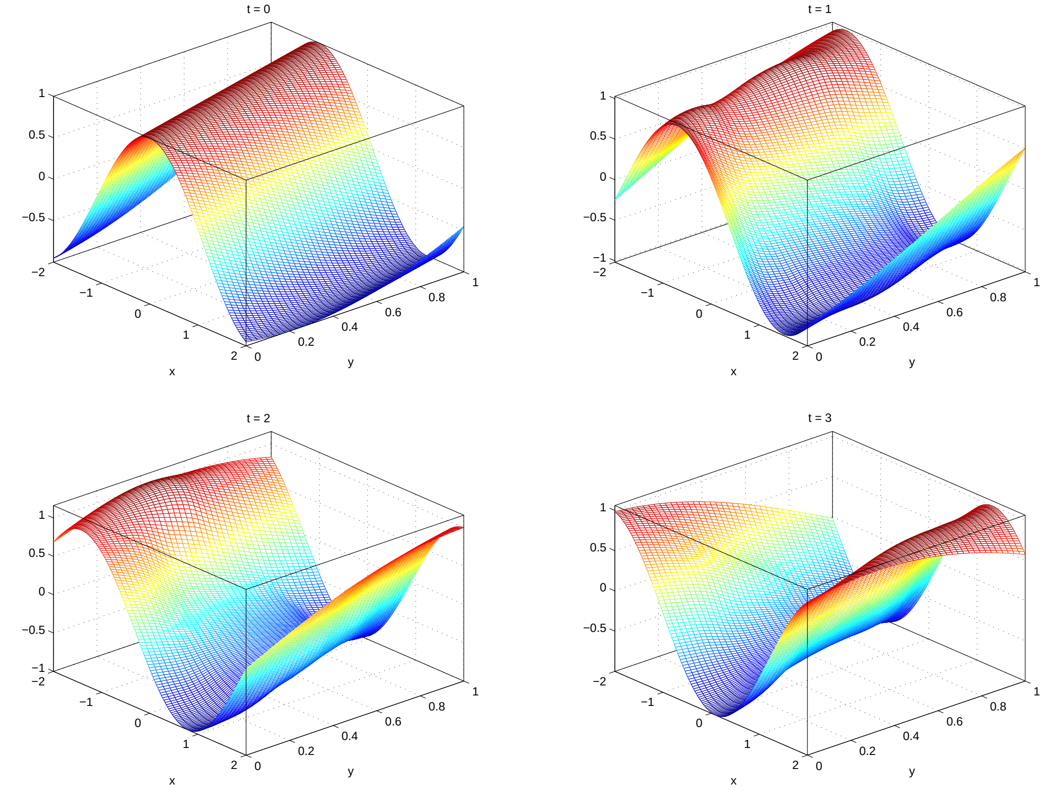


Figure 5-7: Evolution of the mean of E_3 with $N = 10$ for the case of compatible initial/boundary conditions (5.4.6)–(5.4.7) at $t = 0, 1, 2, 3$, respectively.

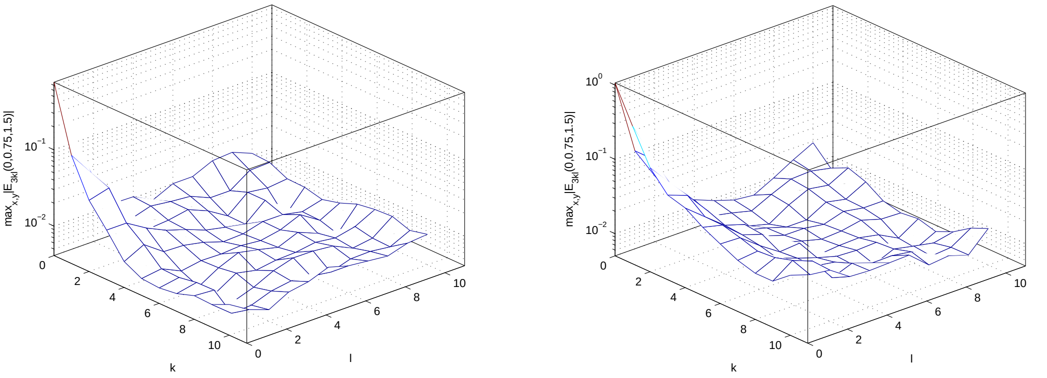


Figure 5-8: PC modes decay of E_3 with 2 random interfaces at $x = 0, y = 0.75$, $t = 1.5$. The z-axis is plotted in *log* scale. LEFT: initial/boundary conditions as in (5.4.2) and (5.4.3); RIGHT: compatible initial/boundary conditions (5.4.6) and (5.4.7).

Table 5-1: Mean and variance of E_3 with 3 random interfaces at some points (x, y), $t = 1.5$ with different number of PC modes.

		PC		
	(x, y)	5	10	15
Mean	$(-0.8, 0.2)$	-0.4535	-0.4525	-0.4517
	$(0, 0.75)$	-0.6236	-0.6227	-0.6220
	$(-1.2, 0.8)$	-0.2212	-0.2200	-0.2194
	$(1.2, 0.5)$	-0.2645	-0.2631	-0.2625
	$(0.4, 0.4)$	-0.6585	-0.6572	-0.6565
Variance	$(-0.8, 0.2)$	4.25E-03	4.35E-03	4.35E-03
	$(0, 0.75)$	8.51E-04	8.89E-04	8.98E-04
	$(-1.2, 0.8)$	1.26E-03	7.27E-03	7.29E-03
	$(1.2, 0.5)$	5.19E-03	5.18E-03	5.17E-03
	$(0.4, 0.4)$	8.15E-03	9.15E-03	9.17E-03

In Fig. 5-9, we plot the CDF, mean and variance of E_3 at $x = 0, y = 0.75$ at $t = 1.5$ with 3 random interfaces $\{z(x, y) = \xi_1\}$, $\{z(x, y) = \xi_2\}$, $\{z(x, y) = \xi_3\}$. We use the same initial/boundary conditions as in (5.4.2), μ and ε are as in (5.4.4) with additional $\mu_4 = \varepsilon_4 = 4$ if $\xi_3 \leq z < \xi_4$ and $a_1 = -1$, $b_1 = -0.5$, $a_2 = -0.4$, $b_2 = 1.4$, $a_3 = 1.5$, $b_3 = 2$. In Table 5-1, we compare the means and variances of the E_3 field at $t = 1.5$ at some specified points with different number of PC modes $PC = 5, 10, 15$. It is shown that these results are not much different from each other. This is due to the exponential decay of the PC modes (see Fig. 5-8). Hence, for a simulation of increasing number of interfaces, a relatively small number of PC modes ($N = 5$) can be used for saving computational costs.

5.5 Parallel Computing

In this section, we present the parallelization of our numerical codes using the Message Passing Interface (MPI) library (see [85], [40], and [114]) for the purpose of saving computational time.

Before proceeding, we recall that our objective in this article is to study the evolution of the CDF in time of the electromagnetic fields in case there appear some uncertainties in the governing equations (5.1.1). By projecting the solutions of Eqs. (5.1.1) into the random space using the PC projection method, we were able to separate the deterministic and random parts of the solutions. We, then, could apply the well-known Yee scheme for solving the former parts numerically. In order to obtain the CDF evolution in time, we applied the Monte Carlo sampling method in a post-processing stage. To compare our Monte Carlo sampling with the conventional one, we just note that the latter is applied as a preprocessing to the equations and then the resulting equations are deterministic for which we may use classical numerical methods. However, to get reliable statistics, large samples are required and this causes a high cost in computations. Hence, there are two independent stages in the algorithm, i.e., the Yee scheme and the Monte Carlo sampling stages. Thus, it is natural that these two stages are objectives of our parallelization.

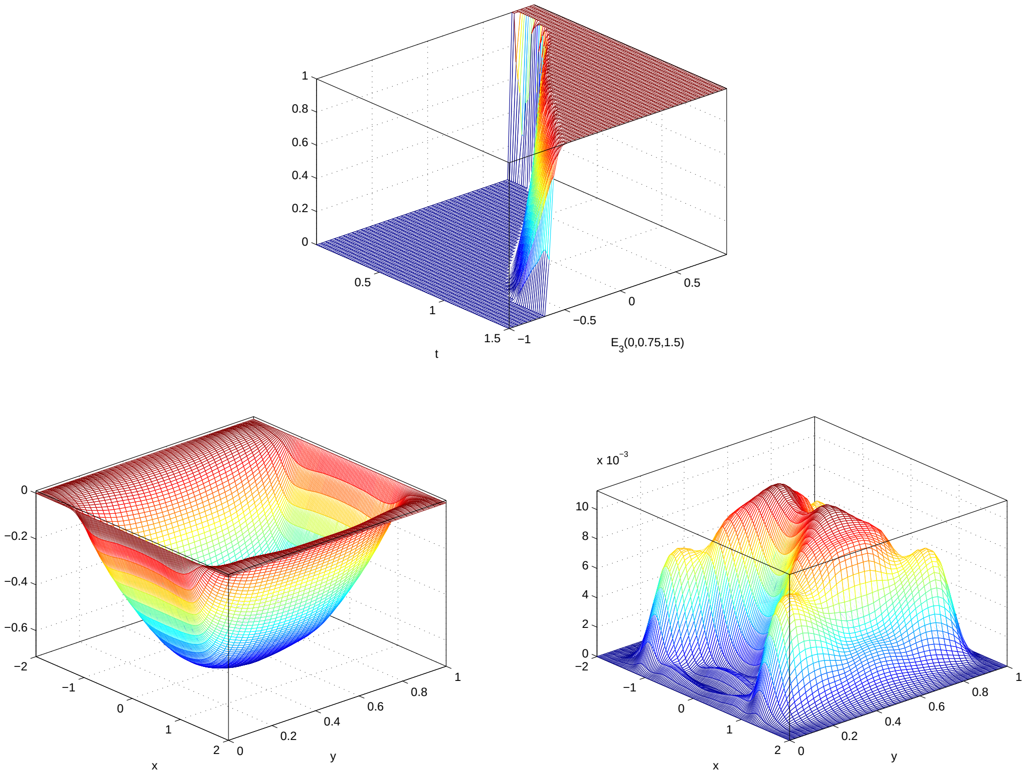


Figure 5-9: CDF, mean and variance of E_3 at $x = 0, y = 0.75$ with 3 independent random variables and 5 PC modes using initial/boundary conditions (5.4.2) and (5.4.3), μ and ε as in (5.4.4) with additional $\mu_4 = \varepsilon_4 = 4$ if $\xi_3 \le z < \xi_4$.

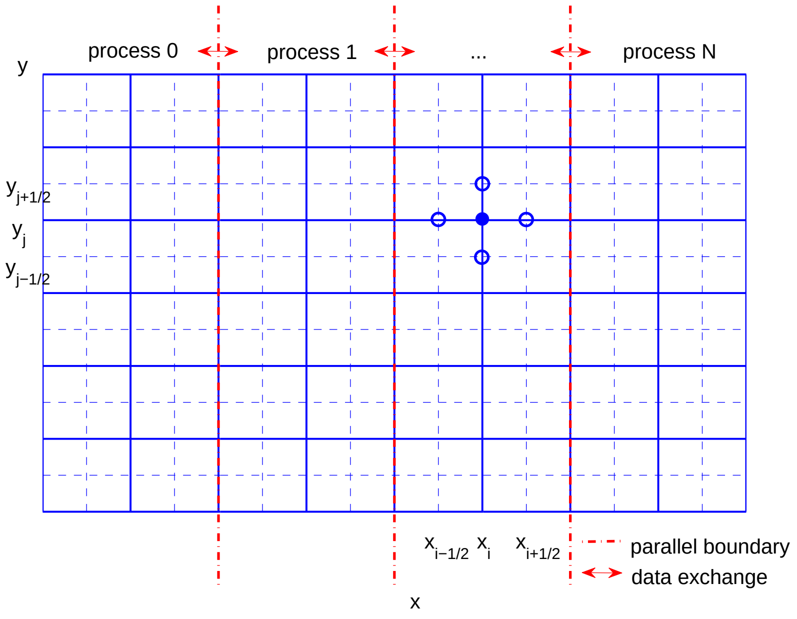


Figure 5-10: Domain decomposition for the parallel codes in a staggered grid.

In the first stage, i.e., the Yee scheme, for parallel purposes, we apply the so-called domain decomposition in the x-direction to the $2D$ grid of the computational domain. We divide the computational domain into sub-domains and assign all computational loads relating to each of them to a process. We notice that between two neighboring sub-domains, there exists an interface in which data transfer is required between the two processes responsible for these sub-domains during the parallel computing (see Fig. 5-10). We call these interfaces "parallel boundaries" to distinguish with the physical boundaries defined in the governing equations; and the data transfer among processes "parallel boundary update". Hence, for each time step, we need to update both the physical and parallel boundaries in each sub-domain. We note that since we parallel our computation based on the computational domain, not on the PC modes, the amounts of data transferring through the parallel boundaries are equal to each other and do not depend on the number of random interfaces and the fact that whether the sub-domains contain the random interfaces $\{z = \xi_i\}$ or not. We also notice that the Yee scheme is a type of multi-stage methods, i.e., at each time step, the update of $\mathbf{H}_1$, $\mathbf{H}_2$, and $\mathbf{E}_3$ in the system (5.2.17) consists of two sub-steps: firstly, the new $\mathbf{H}_1$ and $\mathbf{H}_2$ are updated based on the old $\mathbf{E}_3$ of the previous time step; then, the new $\mathbf{E}_3$ is updated following the new $\mathbf{H}_1$, $\mathbf{H}_2$ of the current time step. Hence, we need to have the parallel boundaries updated twice. Since we decompose the computational domain in the $x-$direction, only parallel boundaries for $\mathbf{H}_2$ and $\mathbf{E}_3$ need updating (see the Yee scheme in (5.2.17)).

The parallelization of the Monte Carlo sampling stage as a post-processing is simpler than that of the Yee scheme stage because we parallel our codes based on the number of samples used in the Monte Carlo method and all these samples are independent of each other. Thus we only need to divide the samples into portions and assign each portion to a process. We, then, collect the results from all processes to obtain the aiming CDFs.

For numerical simulations, we test our parallel codes with two cases. In the first case, we simulate with two random variables or interfaces with initial and boundary data as in (5.4.6) and (5.4.7) to illustrate for the necessity of updating of both physical and parallel boundaries in each sub-domain. And in the second case, we test with three random interfaces with initial and boundary conditions (5.4.2) and (5.4.3) aiming for the saving of computational time. The numerical results of these tests are similar to those of the serial cases shown in Figs. 5-6, 5-9, respectively.

In Fig. 5-11, we plot the computational time (in minutes) for different numbers of processes used in the parallel computation. The result obtained from the number of processes 1 is equivalent to that of a serial code. We conclude that using parallel computation, we can save a considerable amount of time. As shown in the figure, a parallel code with 8 processes is about 8 times faster than a serial code. The increase of computational time in case of two random variables or interfaces with 12 processes can be explained by the increase of the data transferring cost among the processes. Hence, in our problem, we conclude that parallel codes using 8 processes are more optimal for cases of 2 random interfaces; whereas for 3 and 5 interfaces, 16 processes yield faster computational time.

In Fig. 5-12, we further test with the case of 5 random interfaces with initial / boundary conditions

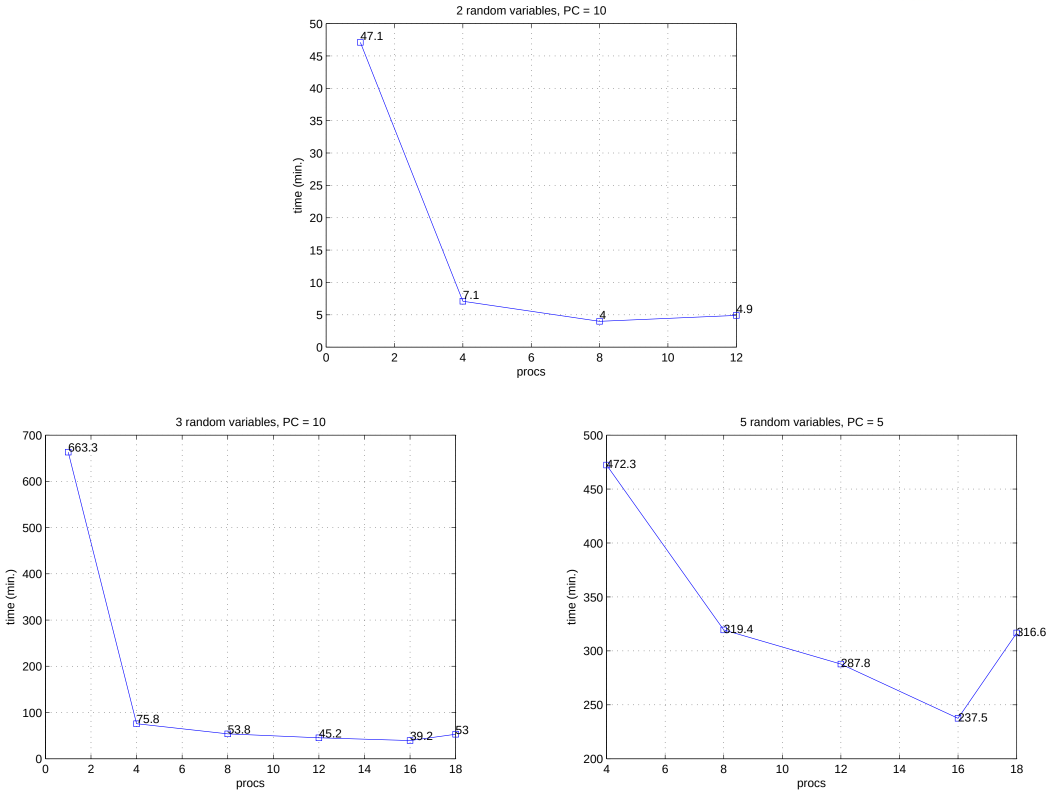


Figure 5-11: Computational time of different number of PC modes. x-axis is the number of processes used in the computation. Number of processes = 1 is equivalent to the serial codes.

as in (5.4.2)–(5.4.3) and other parameters are as below:

$$\left\{\begin{array}{l} \mu_1 = \varepsilon_1 = 1, \\ \mu_2 = \varepsilon_2 = 2, \\ \mu_3 = \varepsilon_3 = 3, \\ \mu_4 = \varepsilon_4 = 4, \\ \mu_5 = \varepsilon_5 = 1, \\ \mu_6 = \varepsilon_6 = 2, \end{array}\right. \quad \text{and} \quad \left\{\begin{array}{l} \xi_1 \in (a_1, b_1) = (-1, -0.5), \\ \xi_2 \in (a_2, b_2) = (-0.4, -0.1), \\ \xi_3 \in (a_3, b_3) = (0, 0.5), \\ \xi_4 \in (a_4, b_4) = (1, 1.4), \\ \xi_5 \in (a_5, b_5) = (1.5, 2), \end{array}\right. \tag{5.5.1}$$

It is shown in the figure that the mean and variance of E_3 behave more wildly than those of lesser numbers of random interfaces. It is because with the introduction of 5 random interfaces, their fluctuations occur in most of the computational domain Ω."

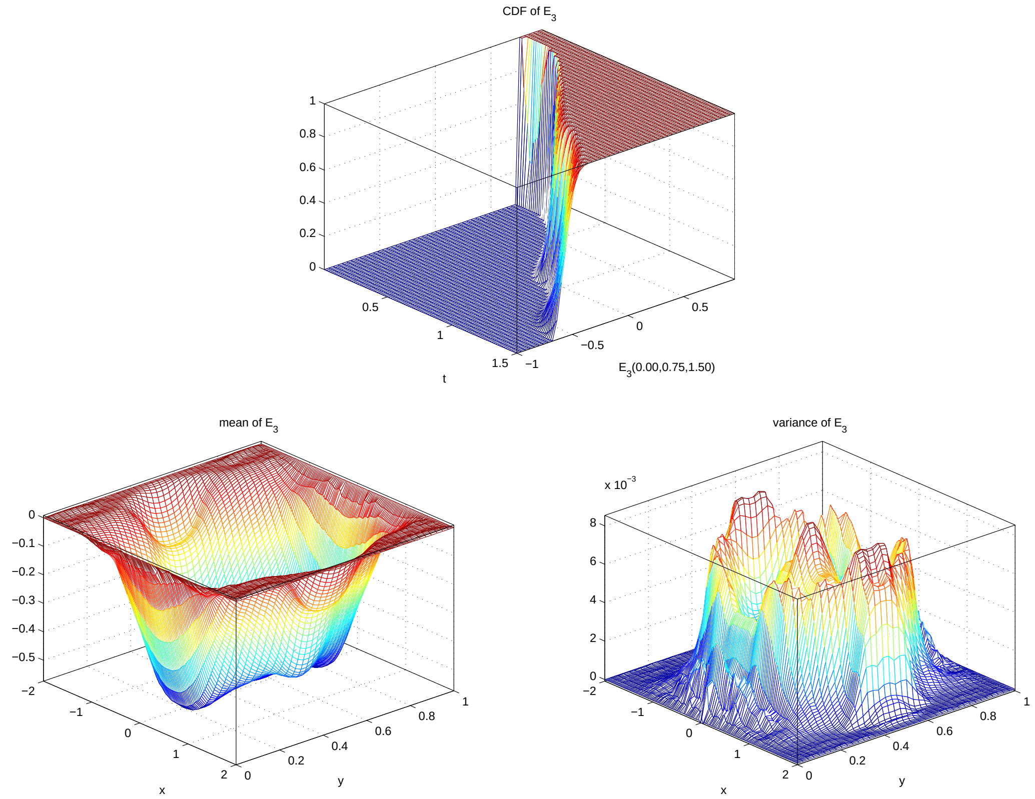


Figure 5-12: CDF, mean and variance of E_3 with 5 random interfaces with initial/boundary conditions (5.4.2)–(5.4.3) and other parameters as in (5.5.1). $PC = 5$.

6

Semi-analytical Time Differencing Methods for Dispersive PDEs

6.1 Introduction

In this chapter, we investigate the following Korteweg de Vries (KdV) equation, which belongs to the class of hyperbolic PDEs,

$$\begin{cases} u_t + uu_x + u_{xxx} = f, & \text{in } \Omega = (0, 2\pi), \\ u(x,0) = u_0(x), \end{cases} \tag{6.1.1}$$

where $u = u(x,t)$, $f = f(x,t)$, the initial and boundary conditions are 2π-periodic.

By a spectral method (see [139], [32], [20]), Eq. (6.1.1) is transformed into a finite system of ODEs whose unknowns are the modes of a truncated Fourier series, and this system is complex due to the dispersive term u_{xxx}. Hence, the waves do not decay in time but rapidly oscillate as the wavenumber k is large. Moreover, all Fourier modes interact with each other in a convolution form by the nonlinear quadratic term uu_x.

To illustrate, we first discretize the spatial domain using an equidistant grid $0 = x_0 < x_1 < \ldots < x_j < \ldots < x_{N-1} = 2\pi - \Delta x$, where the grid size $\Delta x = x_{j+1} - x_j = 2\pi/N$, $j = 0, \ldots, N-2$. That is, $x_j = 2\pi j/N$. Let $u_j = u_j^n$ be an approximation of the exact solution $u(x_j, t_n)$. By a discrete Fourier transform, we write

$$u_j = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} \widehat{u}(k) e^{i\frac{2\pi}{N}jk}, \quad j = 0, \ldots, N-1, \tag{6.1.2}$$

where

$$\widehat{u}(k) = \sum_{j=0}^{N-1} u_j e^{-i\frac{2\pi}{N}jk}, \quad k = -N/2, \ldots, -1, 0, 1, \ldots, N/2-1, \tag{6.1.3}$$

are the Fourier modes. Here, $i^2 = -1$ is the imaginary number. For simplicity, we omit the time dependence of $\widehat{u}(k)$ and we assume that N is even.

Taking the discrete Fourier transform of Eq. (6.1.1) with a note that $\widehat{u_{xxx}} = -ik^3\widehat{u}$, and $\widehat{uu_x} = \frac{1}{2}\widehat{(u^2)_x} = \frac{ik}{2}\widehat{u^2} = \frac{ik}{2}\widehat{u} * \widehat{u}$. Invoking transforms (6.1.2) and (6.1.3), the discrete convolution $*$ is derived as

$$\begin{aligned}\widehat{uv}(k) &= \sum_{j=0}^{N-1}(uv)_j e^{-i\frac{2\pi}{N}kj} = \sum_{j=0}^{N-1}\left[\frac{1}{N}\sum_{l=-N/2}^{N/2-1}\widehat{u}(l)e^{i\frac{2\pi}{N}lj}\right]v_j e^{-i\frac{2\pi}{N}lj} \\ &= \frac{1}{N}\sum_{l=-N/2}^{N/2-1}\widehat{u}(l)\sum_{j=0}^{N-1}v_j e^{-i\frac{2\pi}{N}(k-l)j} = \frac{1}{N}\sum_{l=-N/2}^{N/2-1}\widehat{u}(l)\widehat{v}(k-l) \\ &=: (\widehat{u} * \widehat{v})(k).\end{aligned} \tag{6.1.4}$$

For $k \neq 0$, let

$$\varepsilon(k) = \frac{i}{k^3}, \tag{6.1.5}$$

by Fourier transform, we obtain Eq. (6.1.1) as follows,

$$\begin{cases}\widehat{u}_t(k) + \dfrac{1}{\varepsilon(k)}\widehat{u}(k) = \widehat{F}(\widehat{u}, t), \\ \widehat{u}(k) = \widehat{u}_0(k) \quad \text{at } t = 0,\end{cases} \tag{6.1.6}$$

where

$$\widehat{F}(\widehat{u}, t) = -\frac{ik}{2}(\widehat{u} * \widehat{u})(k) + \widehat{f}(k). \tag{6.1.7}$$

We note that the system is complex since $\varepsilon(k)$ is pure imaginary as indicated above. This results from the dispersive term. In case of a dissipative PDE, e.g. the Kuramoto-Sivashinsky (KS) equation ([90], [75]), the system is real. For the former, the solutions experience oscillations; whereas for the latter, they decay in time. The wave frequencies and rates of decay are proportional to the wavenumbers k's.

When k is large, $0 < |\varepsilon(k)| \ll 1$, where $|\cdot|$ is the modulus. Due to this smallness, the solutions of Eq. (6.1.6) exhibit stiff layers (see the thin layers (6.2.32), (6.2.35) below). This causes difficulties in the implementation of numerical methods to approximate the solutions. A very small time step should be chosen for the treatment of the sharp transition in the stiff layers. Here, we use the term "stiff" loosely to indicate both rapidly decaying layers if system (6.1.6) is real (for dissipative PDEs) or rapidly oscillatory waves if it is complex for dispersive PDEs as in our case.

In order to overcome this difficulty, a good numerical time-stepping method for Eq. (6.1.6) has to handle the stiffness without requiring a fine time step. A conventional approach is the integrating factor (IF) (see Sec. 6.2.1.1 below and [20], [6], [32], etc.) Here, one multiplies both sides of Eq. (6.1.6) by the integrating factor $e^{t/\varepsilon}$, where we have omitted k; and applies a time integrating method, e.g. the Runge-Kutta type, to the new function $y = ue^{t/\varepsilon}$ with $y_t = e^{t/\varepsilon}\widehat{F}$ where the stiff term $e^{t/\varepsilon}$ is absorbed into the y. Implementations and applications of the IF methods can be found in, for instances, [22], [90], [4], or [5]. An alternative approach is the so-called Exponential time differencing (ETD) methods, which are constructed by Cox and Matthews ([22]) (see Sec. 6.2.1.2 below). Their idea is that, after multiplying with the integrating factor, unlike the IF methods, the authors approximate the nonlinear term under the integral sign; then apply some standard method for the time advancing. Numerical applications show that the ETD methods outperform the other comparing ones (see, e.g. [22], [1]). Unfortunately, the ETD methods suffer from the instability caused by the presence of exponential terms in their formulas. This drawback is tackled by Kassam and Trefethen with the application of an integral over a closed contour in a complex plane (see [90]). Since their introduction, ETD methods have been successfully applied in a variety of PDEs, including the nonlinear Schrödinger equation ([62]), the KdV and the KS equations ([22], [90]), or recently, to model the particle dynamics of low Reynolds number incompressible flows (see [104]), etc. Besides the IFs and ETDs, we would also like to mention the Implicit-Explicit (IMEX) methods, which are developed based on the idea to use an implicit scheme to approximate the stiff linear term; while an explicit one for the mild nonlinear one. Interested readers can refer to [4] or [3] and the references therein.

In general, there are various existing methods for stiff problems. For example, one can consider relaxation methods if the spectrum of the dominating operator in the problems is negative, or multi-scale methods if the spectrum is imaginary. References can be found for Jin-Xin relaxation schemes [71], [81], heterogeneous multi-scale methods [30], [27], equation free methods [88], flow averaging integrators [141], and WKB methods [84]. However, for our research we restrict to the case where the linear stiff term, u/ε, is dominating as in Eq. (6.1.6).

In this chapter, we present a new semi-analytical time differencing methods for the spectral methods (6.1.6) applying to the KdV equation (6.1.1) (see [74], [75]). For the time integrator, we restrict our scope to the Runge-Kutta (RK) methods, both the 2^{nd}- and 4^{th}-order ones. The idea is to invoke the singular perturbation analysis to analytically approximate the stiff part of the solution, whereas the non-stiff ones are obtained numerically. Notice that if we fix the wavenumber k, the system (6.1.6) is merely an ODE. Hence, it is natural to investigate the new scheme in case of ODEs first. The development to a PDE case, i.e., the KdV equation will then follow. For the latter stage, it is important to how one treats the coupling of all Fourier modes in (6.1.6) due to the nonlinear convective term. We organize the chapter following this structure.

6.2 New Semi-analytical Time Differencing Methods for Stiff ODEs

For simplicity, we rewrite the system (6.1.6) in terms of a single equation as follows,

$$\begin{cases} u_t + \dfrac{1}{\varepsilon}u = F(u,t), \\ u(0) = u_0, \end{cases} \tag{6.2.1}$$

where $u = u(t)$, $0 < |\varepsilon| \ll 1$, and $F(u(t),t)$ and u_0 are given data. Here, ε is real and positive or pure imaginary, and $|\varepsilon|$ is its modulus. However, in the analysis below, ε may be considered a complex number.

6.2.1 Classical IF and ETD Methods

In order to resolve the stiff linear term u/ε in Eq. (6.2.1), multiplying both sides by an integrating factor $e^{t/\varepsilon}$, we obtain that

$$\frac{d}{dt}\left(ue^{t/\varepsilon}\right) = e^{t/\varepsilon}F(u,t). \tag{6.2.2}$$

Notice that the stiff term u/ε has been absorbed by the integrating factor, leaving the right-hand side of Eq. (6.2.2) with only the nonlinear term F, which is considered milder than the u/ε. How one treats this right-hand side determines the numerical method to use. Below, we present the two classical methods which are mostly used in dealing with this type of problems. For deeper discussions and stability issues, readers can refer to [22], [90].

First of all, let us define a discrete time grid. We denote 0 and T the initial and final time, respectively; and discretize the time into M equal intervals $0 = t_0 < t_1 < \ldots < t_n < \ldots < t_{M-1} < t_M = T$. Let $h = t_{n+1} - t_n$, $n = 0, \ldots, M-1$, be the time step. We denote $u(t_n)$ the exact solution of Eq. (6.2.1) at time t_n, and u_n the corresponding approximation obtained from some numerical method. We now proceed to the IF methods.

6.2.1.1 Integrating Factor Methods

Let

$$y(t) = u(t)e^{t/\varepsilon}, \tag{6.2.3}$$

Eq. (6.2.2) is rewritten in term of $y(t)$ as follows,

$$y_t = e^{t/\varepsilon}F(ye^{-t/\varepsilon}, t). \tag{6.2.4}$$

Applying a typical RK2 or RK4 to Eq. (6.2.4) to approximate $y(t)$ instead of $u(t)$, we deduce the following schemes which are called IFRK2 or IFRK4, depending on which RK method to be applied.

Writing $F_n = F(u_n, t_n)$ we have:

- IFRK2 method:

$$\begin{cases} k_1 = e^{-h/\varepsilon} F_n, \\ k_2 = F(u_n e^{-h/\varepsilon} + hk_1, t_n + h), \\ u_{n+1} = u_n e^{-h/\varepsilon} + \dfrac{h}{2}\left(k_1 + k_2\right). \end{cases} \tag{6.2.5}$$

- IFRK4 method (see [1]):

$$\begin{cases} k_1 = F_n, \\ k_2 = F((u_n + hk_1/2)e^{-h/(2\varepsilon)}, t_n + h/2), \\ k_3 = F(u_n e^{-h/(2\varepsilon)} + hk_2/2, t_n + h/2), \\ k_4 = F(u_n e^{-h/\varepsilon} + hk_3 e^{-h/(2\varepsilon)}, t_n + h), \\ u_{n+1} = u_n e^{-h/\varepsilon} + \dfrac{h}{6}\left(k_1 e^{-h/\varepsilon} + 2(k_2 + k_3)e^{-h/(2\varepsilon)} + k_4\right). \end{cases} \tag{6.2.6}$$

6.2.1.2 Exponential Time Differencing Methods

Instead of directly applying the RK time integration as in the IF methods, in their paper, Cox and Matthews ([22]) suggest an alternative by interpolating the nonlinear term F. Integrating both sides of Eq. (6.2.2) with respect to time, we obtain

$$u(t_{n+1}) = u(t_n)e^{-h/\varepsilon} + e^{-h/\varepsilon}\int_0^h e^{s/\varepsilon} F(u(t_n + s), t_n + s)ds. \tag{6.2.7}$$

As a simple approximation to the integration, we take $F = F_n$ and we then obtain the scheme ETD1, $u_{n+1} = u_n e^{-h/\varepsilon} + \frac{(e^{-h/\varepsilon}-1)F_n}{-1/\varepsilon}$. We denote this by a_n. Interpolating the nonlinear term F linearly as

$$F = F_n + \frac{F(a_n, t_n + h) - F_n}{h}(t - t_n) + \mathcal{O}(h^2), \tag{6.2.8}$$

and substituting into Eq. (6.2.7) gives the ETDRK2 method. The ETDRK4 method is derived in a similar way, but using an RK4 method. The schemes are as follows:

- ETDRK2 method:

$$\begin{cases} a_n = u_n e^{-h/\varepsilon} + \dfrac{(e^{-h/\varepsilon} - 1)F_n}{-1/\varepsilon}, \\ u_{n+1} = a_n + (F(a_n, t_n + h) - F_n)\dfrac{e^{-h/\varepsilon} - 1 + h/\varepsilon}{h/\varepsilon^2}, \end{cases} \tag{6.2.9}$$

- ETDRK4 method:

$$\begin{cases} a_n = u_n e^{-h/(2\varepsilon)} + \dfrac{(e^{-h/(2\varepsilon)} - 1)F_n}{-1/\varepsilon}, \\ b_n = u_n e^{-h/(2\varepsilon)} + \dfrac{(e^{-h/(2\varepsilon)} - 1)F(a_n, t_n + h/2)}{-1/\varepsilon}, \\ c_n = a_n e^{-h/(2\varepsilon)} + \dfrac{(e^{-h/(2\varepsilon)} - 1)(2F(b_n, t_n + h/2) - F_n)}{-1/\varepsilon}, \\ t_1 = \dfrac{-4 + h/\varepsilon + e^{-h/\varepsilon}(4 + 3h/\varepsilon + h^2/\varepsilon^2)}{-h^2/\varepsilon^3}, \\ t_2 = \dfrac{2 - h/\varepsilon + e^{-h/\varepsilon}(-2 - h/\varepsilon)}{-h^2/\varepsilon^3}, \\ t_3 = \dfrac{-4 + 3h/\varepsilon - h^2/\varepsilon^2 + e^{-h/\varepsilon}(4 + h/\varepsilon)}{-h^2/\varepsilon^3}, \\ u_{n+1} = u_n e^{-h/\varepsilon} + F_n t_1 \\ \qquad + 2(F(a_n, t_n + h/2) + F(b_n, t_n + h/2))t_2 + F(c_n, t_n + h)t_3. \end{cases} \tag{6.2.10}$$

In [90], the authors point out that the ETDRK methods are not numerically stable, due to the presence of the term

$$g(z) = \frac{e^z - 1}{z}, \tag{6.2.11}$$

or its higher-order variants. When z is small, these terms induce large cancellation errors.

To remedy this, they propose a stabilizing mechanism using a so-called contour integral in a complex plane. That is, instead of a direct calculation of $g(z)$ in Eq. (6.2.11), one can evaluate the function via an integral over a contour Γ as follows

$$g(z) = \frac{1}{2\pi i} \int_\Gamma \frac{g(s)}{s - z} ds, \tag{6.2.12}$$

where $z \in \mathbb{C}$, $i = \sqrt{-1}$, and Γ encloses z. For simplicity, Γ is chosen to be a circle centered at z. Then, a typical quadrature method, e.g., the trapezoidal rule, can be applied to evaluate the integral numerically. In case z is real, by choosing Γ as a circle centered at z on the real axis, the computational cost can be reduced by approximating Eq. (6.2.12) using m equally-spaced points on the upper half of the contour Γ.

Applying the complex integral (6.2.12) to recompute the unstable terms in (6.2.9) and (6.2.10) with $z = -h/\varepsilon$, we obtain a stabilized version of the ETDRK methods. From herein, whenever we refer to an ETDRK method, we mean the one using this stability mechanism.

6.2.2 The Schemes with Correctors

In this section, we first investigate the behavior of the solution of Eq. (6.2.1) by invoking the singular perturbation analysis (see e.g., [48], [150], [132], [66]). By this, we can decompose the solution into a fast and slow part, in which both are analytically approximated by an asymptotic expansion. For our numerical treatments, we call the asymptotic expansion for the fast part u_f a corrector θ. The slow part u_s is then numerically approximated by the IFRK and ETDRK methods discussed in the previous section. Since the conventional numerical schemes are now applied for the non-stiff part $v = u - \theta \approx u_s$, it is expected that the results are much improved compared with the same schemes applied directly to the solution u. Hence, our approach is semi-analytical. Semi-analytical schemes are widely developed for problems exhibiting boundary layers or interior layers. Here, we name some relevant works, both analytically (see [52], [87], [108]) and numerically (see [79], [67], [80], [73], [134], etc.)

6.2.2.1 Asymptotic Expansions

We decompose the solution of (6.2.1) as

$$u = u_s + u_f, \tag{6.2.13}$$

where $u_s(t)$ is the slow part of u for which $u_s = \mathcal{O}(1)$ and $du_s/dt = \mathcal{O}(1)$; and the fast part $u_f(t)$ exhibits a sharp transition in the solution with $u_f = \mathcal{O}(1)$ and $du_f/dt = \mathcal{O}(1/\varepsilon)$. In [22] (see also [101]), it is pointed out that for a dissipative equation (i.e. ε is real and positive), the solution u is attracted to the slow part u_s, and this can be expressed as an asymptotic expansion of ε as follows

$$u \sim \varepsilon F - \varepsilon^2 \frac{dF}{dt} + \varepsilon^3 \frac{d^2 F}{dt^2} + \mathcal{O}(\varepsilon^4). \tag{6.2.14}$$

To justify this, we formally approximate the slow part u_s. Using singular perturbation technique, we approximate it by an asymptotic expansion, which is a power series of the small parameter ε,

$$u_s(t) \sim \sum_{j=0}^{\infty} \varepsilon^j u^j. \tag{6.2.15}$$

Substituting into Eq. (6.2.1), we obtain that

$$\sum_{j=0}^{\infty} \varepsilon^j u_t^j + \sum_{j=0}^{\infty} \varepsilon^{j-1} u^j = F\left(\sum_{j=0}^{\infty} \varepsilon^j u^j, t\right). \tag{6.2.16}$$

Assuming that $F = \mathcal{O}(1)$, at order $\mathcal{O}(\varepsilon^{-1})$, $u^0 = 0$, and at order $\mathcal{O}(\varepsilon^0)$, $u^1 = -u_t^0 + F = F$. At order $\mathcal{O}(\varepsilon^{j-1})$, $j \geq 2$, $u^j = -u_t^{j-1}$. Hence, $u_s \sim \varepsilon u^1 + \varepsilon^2 u^2 + \varepsilon^3 u^3 + \mathcal{O}(\varepsilon^4)$ which implies (6.2.14).

Expanding the nonlinear term F we obtain more precise asymptotic expansion for the u_s. We first formally write

$$F(u,t) = \sum_{j=0}^{\infty} \varepsilon^j F^j(u,t). \tag{6.2.17}$$

We then find that

$$\begin{aligned} F\left(\sum_{j=0}^{\infty}\varepsilon^j u^j, t\right) &= F^0(u^0+\varepsilon u^1+\mathcal{O}(\varepsilon^2),t)+\varepsilon F^1(u^0+\mathcal{O}(\varepsilon),t)+\mathcal{O}(\varepsilon^2) \\ &= \text{(using the formal Taylor expansion for } F^j(u,t) \text{ w.r.t. } u \text{ at } u^0) \\ &= F^0(u^0,t)+\varepsilon\frac{\partial F^0}{\partial u}(u^0,t)u^1+\varepsilon F^1(u^0,t)+\mathcal{O}(\varepsilon^2). \end{aligned} \tag{6.2.18}$$

Hence, balancing terms at each order of ε in (6.2.16), at $\mathcal{O}(\varepsilon^{-1})$, we have:

$$u^0 = 0, \tag{6.2.19}$$

at order $\mathcal{O}(\varepsilon^0)$,

$$u^1 = -u^0_t + F^0(u^0,t) = F^0(0,t), \tag{6.2.20}$$

and at order $\mathcal{O}(\varepsilon)$,

$$\begin{aligned} u^2 &= -u^1_t + \frac{\partial F^0}{\partial u}(u^0,t)u^1 + F^1(u^0,t) \\ &= -\frac{\partial F^0}{\partial t}(0,t) + \frac{\partial F^0}{\partial u}(0,t)F^0(0,t) + F^1(0,t). \end{aligned} \tag{6.2.21}$$

In this way, at each order $\mathcal{O}(\varepsilon^{j-1})$, we can recursively define u^j, $j = 3, 4, \cdots$.

For the stiff or fast part u_f with $u = u_s + u_f$, we formally write

$$u_f(t) \sim \theta = \sum_{j=0}^{\infty}\varepsilon^j\theta^j, \tag{6.2.22}$$

$$u(t) \sim \sum_{j=0}^{\infty}\varepsilon^j(u^j+\theta^j). \tag{6.2.23}$$

Substituting (6.2.23) into Eq. (6.2.1) we formally obtain that

$$\sum_{j=0}^{\infty}\varepsilon^j(u^j+\theta^j)_t + \sum_{j=0}^{\infty}\varepsilon^{j-1}(u^j+\theta^j) = F\left(\sum_{j=0}^{\infty}\varepsilon^j(u^j+\theta^j), t\right). \tag{6.2.24}$$

In order to capture the sharp transitions, using a stretched time variable,

$$\bar{t} = \frac{t}{\varepsilon}, \tag{6.2.25}$$

thanks to (6.2.16), from (6.2.24) we formally obtain that

$$\sum_{j=0}^{\infty}\varepsilon^{j-1}\frac{d}{d\bar{t}}\theta^j + \sum_{j=0}^{\infty}\varepsilon^{j-1}\theta^j = F\left(\sum_{j=0}^{\infty}\varepsilon^j(u^j+\theta^j), t\right) - F\left(\sum_{j=0}^{\infty}\varepsilon^j u^j, t\right). \tag{6.2.26}$$

As before, using the formal Taylor expansion with respect to u at $u^0 + \theta^0$, we find that

$$\begin{aligned} &F\left(\sum_{j=0}^{\infty} \varepsilon^j (u^j + \theta^j), t\right) \\ &\quad = F^0(u^0 + \theta^0, t) + \varepsilon \frac{\partial F^0}{\partial u}(u^0 + \theta^0, t)(u^1 + \theta^1) + \varepsilon F^1(u^0 + \theta^0, t) + \mathcal{O}(\varepsilon^2). \end{aligned} \tag{6.2.27}$$

We here supplement with the initial condition $(6.2.1)_2$ so that

$$\sum_{j=0}^{\infty} \varepsilon^j (u^j + \theta^j) = u_0 \quad \text{at } t = 0. \tag{6.2.28}$$

In particular, from (6.2.19) and (6.2.20), we have

$$\begin{cases} \theta^0 = -u^0 + u_0 = u_0 \quad \text{at } t = 0, \\ \theta^1 = -u^1 = -F^0(0,0) \quad \text{at } t = 0. \end{cases} \tag{6.2.29}$$

Balancing terms at each order of ε in (6.2.26), we obtain the corrector equations as below for θ^j. At $\mathcal{O}(\varepsilon^{-1})$, we thus have:

$$\begin{cases} \theta^0_{\bar{t}} + \theta^0 = 0, \\ \theta^0 = u_0 \text{ at } t = 0; \end{cases} \tag{6.2.30}$$

changing back to the variable t gives

$$\begin{cases} \theta^0_t + \dfrac{1}{\varepsilon}\theta^0 = 0, \\ \theta^0 = u_0 \text{ at } t = 0, \end{cases} \tag{6.2.31}$$

whose solution is

$$\theta^0(t) = u_0 \exp\left(-\frac{t}{\varepsilon}\right). \tag{6.2.32}$$

We note that if $\varepsilon > 0$, θ^0 rapidly changes in the thickness of $\mathcal{O}(\varepsilon)$, then decays to zero.

At order $\mathcal{O}(\varepsilon^0)$, we have

$$\begin{cases} \theta^1_{\bar{t}} + \theta^1 = F^0(u^0 + \theta^0, t) - F^0(u^0, t), \\ \theta^1 = -F^0(0,0) \text{ at } t = 0. \end{cases} \tag{6.2.33}$$

Since $u^0 = 0$, we may write that

$$\begin{cases} \theta^1_t + \dfrac{1}{\varepsilon}\theta^1 = \dfrac{1}{\varepsilon}(F^0(\theta^0, t) - F^0(0,t)), \\ \theta^1 = -F^0(0,0) \text{ at } t = 0. \end{cases} \tag{6.2.34}$$

The exact solution is available,

$$\theta^1(t) = \frac{1}{\varepsilon}\int_0^t \exp\left(-\frac{t-s}{\varepsilon}\right) G^0(\theta^0(s), s)ds - F^0(0,0)\exp\left(-\frac{t}{\varepsilon}\right). \tag{6.2.35}$$

where

$$G^0(\theta^0(s), s) = F^0(\theta^0(s), s) - F^0(0, s). \tag{6.2.36}$$

At $\mathcal{O}(\varepsilon^{j-1})$, $j \geq 2$, we have that

$$\begin{cases} \theta^j_t + \theta^j = \tilde{F}(\theta^0, \ldots, \theta^{j-1}, t), \\ \theta^j = -u^j + u^j_0 \text{ at } t = 0, \end{cases} \tag{6.2.37}$$

in which $\tilde{F}$ keeps only the $\mathcal{O}(\varepsilon^{j-1})$ terms from $F\left(\sum_{j=0}^{\infty} \varepsilon^j(u^j + \theta^j), t\right) - F\left(\sum_{j=0}^{\infty} \varepsilon^j u^j, t\right)$. It turns out that only θ^0 and θ^1 are sufficient for our numerical treatments.

6.2.2.2 The IFRK and ETDRK Schemes with Correctors

Taking θ^0 and θ^1 into account, the solution of Eq. (6.2.1) is decomposed as

$$u = u_s + u_f = v + \theta, \tag{6.2.38}$$

where θ is called the corrector, which approximates the fast part u_f, and v is then close to the slow part u_s. Here, we use numerically the corrector θ in two cases, with only $\theta = \theta^0$ and with both $\theta = \theta^0 + \varepsilon\theta^1$, and v is defined accordingly as below. We denote the first case by Cor1 and the second by Cor2.

For the numerical implementation purpose, since in the combination $\theta^0 + \varepsilon\theta^1$ the last term in (6.2.35) which is multiplied by ε is milder than θ^0, dropping the mild term we define the correctors, θ^0, θ^1,

$$\theta^0(t) = u_0 \exp\left(-\frac{t}{\varepsilon}\right), \quad \theta^1(t) = \frac{1}{\varepsilon}\int_0^t \exp\left(-\frac{t-s}{\varepsilon}\right) G^0(\theta^0(s), s)ds. \tag{6.2.39}$$

Here, G^0 is defined in (6.2.36). We now numerically solve Eq. (6.2.1) incorporating the correctors (6.2.39).

- New scheme with Cor1: $\theta = \theta^0$: Substituting

$$v = u - \theta^0, \tag{6.2.40}$$

into Eq. (6.2.1), thanks to Eq. (6.2.31), we obtain

$$\begin{cases} v_t + \dfrac{1}{\varepsilon} v = F(v + \theta^0, t), \\ v(0) = 0. \end{cases} \tag{6.2.41}$$

- New scheme with Cor2: $\theta = \theta^0 + \varepsilon\theta^1$: Substituting

$$v = u - \theta^0 - \varepsilon\theta^1, \tag{6.2.42}$$

into Eq. (6.2.1) with (6.2.31), (6.2.34) and (6.2.39) gives us

$$\begin{cases} v_t + \dfrac{1}{\varepsilon} v = F(v + \theta^0 + \varepsilon\theta^1, t) - G^0(\theta^0, t), \\ v(0) = 0. \end{cases} \tag{6.2.43}$$

We now apply the conventional IFRK and ETDRK schemes to v, which is considered slower than u because the fast part is subtracted by the correctors. Then u is recovered from $u = v+\theta^0$, $u = v+\theta^0+\varepsilon\theta^1$, respectively.

To justify the scheme (6.2.43), we compare the schemes (6.2.41) and (6.2.43). Expanding $F^0(u,t)$ w.r.t. u at θ^0, the right-hand sides are estimated respectively,

$$F(v + \theta^0, t) = F^0(\theta^0, t) + \frac{\partial F^0}{\partial u}(\theta^0, t)v + \mathcal{O}(v^2) + \mathcal{O}(\varepsilon), \tag{6.2.44}$$

$$F(v + \theta^0 + \varepsilon\theta^1, t) - G^0(\theta^0, t) = F^0(0, t) + \frac{\partial F^0}{\partial u}(\theta^0, t)v + \mathcal{O}(v^2) + \mathcal{O}(\varepsilon). \tag{6.2.45}$$

Obviously, the term $F^0(\theta^0, t)$ is faster than $F^0(0, t)$ and we expect that the solution v in Eq. (6.2.43) is milder and better approximated than in (6.2.41).

6.3 Application to Spectral Methods for the Dispersive KdV Equation

We now apply the analysis given above to the system (6.1.6). We rewrite it as follows, for $k \neq 0$,

$$\begin{cases} L_\varepsilon \widehat{u} := \widehat{u}_t(k) + \dfrac{1}{\varepsilon(k)}\widehat{u}(k) = -\dfrac{ik}{2}(\widehat{u} * \widehat{u})(k) + \widehat{f}(k), \\ \widehat{u}(k) = \widehat{u}_0(k) \quad \text{at } t = 0, \end{cases} \tag{6.3.1}$$

where

$$\varepsilon(k) = \frac{i}{k^3}, \tag{6.3.2}$$

We remind that the discrete convolution is given in Eq. (6.1.4).

For $k = 0$, since $\widehat{u}_t(0) = \widehat{f}(0)$, we can set

$$\widehat{v}(0) = \widehat{u}(0) = \widehat{u}_0(0) + \int_0^t \widehat{f}(k)(s)ds, \quad \widehat{\theta}(0) = 0. \tag{6.3.3}$$

The main issue in the section is how to handle effectively the nonlinear convective term uu_x represented by the convolution $\widehat{u} * \widehat{u}$ in the Fourier space.

6.3.1 Derivation of the Correctors

Using the correctors $\widehat{\theta} = \widehat{\theta}(k)$, we decompose the solution $\widehat{u}(k)$ as

$$\widehat{u}(k) = \widehat{v}(k) + \widehat{\theta}(k). \tag{6.3.4}$$

Here, $\widehat{\theta}$ is derived as below and it turns out that $\widehat{\theta}$ captures the stiffness of $\widehat{u}$ and thus $\widehat{v}$ becomes mild and slow.

Substituting the decomposition (6.3.4) into Eq. (6.3.1) and utilizing the convolution (6.1.4), we obtain that, for $k = -N/2, \dots, -1, 1, \dots, N/2 - 1$,

$$\begin{aligned} L_\varepsilon(\widehat{u}) &= L_\varepsilon(\widehat{v} + \widehat{\theta}) = (\widehat{v}_t + \widehat{\theta}_t)(k) + \frac{1}{\varepsilon(k)}(\widehat{v} + \widehat{\theta})(k) \\ &= -\frac{ik}{2}(\widehat{v} * \widehat{v} + 2\widehat{v} * \widehat{\theta} + \widehat{\theta} * \widehat{\theta})(k) + \widehat{f}(k) \\ &= -\frac{ik}{2N} \sum_{l=-N/2}^{N/2-1} \Big[\widehat{v}(l)\widehat{v}(k-l) + 2\widehat{v}(l)\widehat{\theta}(k-l) + \widehat{\theta}(l)\widehat{\theta}(k-l)\Big] + \widehat{f}(k). \end{aligned} \tag{6.3.5}$$

Since $\widehat{v}(k)$ is considered slow, we first formally seek for the correctors $\widehat{\theta}(k)$ satisfying the equation

$$L_\varepsilon(\widehat{\theta}) = \widehat{\theta}_t(k) + \frac{1}{\varepsilon(k)}\widehat{\theta}(k) = -\frac{ik}{2N} \sum_{l=-N/2}^{N/2-1} \widehat{\theta}(l)\widehat{\theta}(k-l); \tag{6.3.6}$$

then Eq. (6.3.5) is formally reduced to

$$\begin{aligned} L_\varepsilon(\widehat{v}) &= L_\varepsilon(\widehat{u}) - L_\varepsilon(\widehat{\theta}) \\ &= -\frac{ik}{2N} \sum_{l=-N/2}^{N/2-1} \Big[\widehat{v}(l)\widehat{v}(k-l) + 2\widehat{v}(l)\widehat{\theta}(k-l)\Big] + \widehat{f}(k). \end{aligned} \tag{6.3.7}$$

Since the corrector $\widehat{\theta}(k)$ analytically absorbs the stiff part of the solution of each mode, we expect that $\widehat{v}(k)$ is slow. Comparing the convolution terms on the right-hand side of Eqs. (6.3.5) and (6.3.7), it is observed that the one of the latter is much milder than that of the former. Hence, applying the IFRK and ETDRK to $\widehat{v}$ is expected to produce better accuracies for $\widehat{v}$ than for $\widehat{u}$, and by adding $\widehat{\theta}$ to $\widehat{v}$, we then recover $\widehat{u}$.

From (6.3.6), using the singular perturbation analysis, we approximate $\widehat{\theta}$ to capture the stiffness.

For this, for each mode k, we use the stretched time variable,

$$\bar{t} = \frac{t}{\varepsilon} = \frac{t}{\varepsilon(k)}. \tag{6.3.8}$$

Using the following asymptotic expansion,

$$\widehat{\theta}(k) = \widehat{\theta}^0(k) + \varepsilon(k)\widehat{\theta}^1(k) + \mathcal{O}(\varepsilon(k)^2), \tag{6.3.9}$$

and substituting in (6.3.6) with the initial condition $(6.3.1)_2$, we write

$$\begin{cases} \dfrac{1}{\varepsilon}(\widehat{\theta}^0 + \varepsilon\widehat{\theta}^1)_{\bar{t}} + \dfrac{1}{\varepsilon}(\widehat{\theta}^0 + \varepsilon\widehat{\theta}^1) = -\dfrac{ik}{2N}\sum_{l=-N/2}^{N/2-1}(\widehat{\theta}^0 + \varepsilon\widehat{\theta}^1)(l)(\widehat{\theta}^0 + \varepsilon\widehat{\theta}^1)(k-l), \\ \widehat{\theta}^0 = \widehat{u}_0, \quad \widehat{\theta}^1 = 0 \quad \text{at } t = 0. \end{cases} \tag{6.3.10}$$

At the leading order of ε, i.e. $\mathcal{O}(\varepsilon^{-1})$, we obtain that

$$\widehat{\theta}^0_{\bar{t}} + \widehat{\theta}^0 = 0, \tag{6.3.11}$$

or

$$\begin{cases} \widehat{\theta}^0_t + \widehat{\theta}^0 = 0, \\ \widehat{\theta}^0 = u_0, \quad \text{at } t = 0; \end{cases} \tag{6.3.12}$$

and thus, for $k \neq 0$,

$$\widehat{\theta}^0(k) = \widehat{u}_0(k)e^{-\bar{t}} = \widehat{u}_0(k)e^{-t/\varepsilon(k)}. \tag{6.3.13}$$

We now seek for $\widehat{\theta}^1(k)$ at the next order, i.e. $\mathcal{O}(\varepsilon^0)$. Changing the time back to t, and dropping the small nonlinear terms of order $\mathcal{O}(\varepsilon(l)\varepsilon(k-l))$ in the right-hand side of Eq. (6.3.10), it is simplified as

$$\begin{cases} \varepsilon(k)\widehat{\theta}^1_t(k) + \widehat{\theta}^1(k) = -\dfrac{ik}{2N}\sum_{l=-N/2}^{N/2-1}\left(\widehat{\theta}^0(l) + 2\varepsilon(l)\widehat{\theta}^1(l)\right)\widehat{\theta}^0(k-l) \\ \qquad\qquad\qquad\qquad = -\dfrac{ik}{2}\left((\widehat{\theta}^0 + 2\varepsilon\widehat{\theta}^1) * \widehat{\theta}^0\right)(k), \\ \widehat{\theta}^1(k) = 0 \quad \text{at } t = 0. \end{cases} \tag{6.3.14}$$

Due to the convolution on the right-hand side of Eq. (6.3.14), observing that $\widehat{\theta}^0(0) = 0$, Eq. (6.3.14) for mode k is coupled with Eqs. (6.3.14) for all Fourier modes $l \neq k$.

For numerical purpose, it is sufficient to set a cut-off tolerance $H > 0$ so that

$$\widehat{\theta}^0(k) = \widehat{\theta}^1(k) = 0, \forall\ |k| > H. \tag{6.3.15}$$

The $\sum_{l=-N/2}^{N/2-1}$ in Eq. (6.3.14), thanks to the tolerance H, can be reduced to

$$\sum_{l} := \sum_{l=-H+\max(0,k)}^{H+\min(0,k)} \tag{6.3.16}$$

by excluding all zero combinations of $\widehat{\theta}^0(l)\widehat{\theta}^0(k-l)$ and $\widehat{\theta}^1(l)\widehat{\theta}^0(k-l)$. Then, Eq. (6.3.14) is rewritten as

$$\varepsilon(k)\widehat{\theta}^1_t(k) + \widehat{\theta}^1(k) = -\frac{ik}{2N}\sum_{l}\left(\widehat{\theta}^0(l) + 2\varepsilon(l)\widehat{\theta}^1(l)\right)\widehat{\theta}^0(k-l). \tag{6.3.17}$$

These equations are not closed and difficult to get explicit solutions. However, the correctors can be recursively approximated as presented below.

6.3.2 Approximating $\widehat{\theta}^1$

To approximate $\widehat{\theta}^1(k)$, i.e. solutions of Eq. (6.3.17), we let $\widehat{\theta}^{1,0}$ be the solution of Eq.

$$\begin{cases} \varepsilon(k)\widehat{\theta}^{1,0}_t(k) + \widehat{\theta}^{1,0}(k) = -\dfrac{ik}{2N}\sum_l \widehat{\theta}^0(l)\widehat{\theta}^0(k-l), \\ \widehat{\theta}^{1,0}(k) = 0 \quad \text{at } t = 0. \end{cases} \tag{6.3.18}$$

For $n \geq 1$, we define:

$$\begin{cases} \varepsilon(k)\widehat{\theta}^{1,n}_t(k) + \widehat{\theta}^{1,n}(k) = -\dfrac{ik}{2N}\sum_l \left(\widehat{\theta}^0(l) + 2\varepsilon(l)\widehat{\theta}^{1,n-1}(l)\right)\widehat{\theta}^0(k-l), \\ \widehat{\theta}^{1,n}(k) = 0 \quad \text{at } t = 0. \end{cases} \tag{6.3.19}$$

We then find recursively $\widehat{\theta}^{1,n}$ by e.g. Maple. Iterating a few times gives very accurate approximations for $\widehat{\theta}^1$ as indicated in the convergence analysis below.

To understand the convergence errors, we define the norm:

$$\|\widehat{u}\| = \max_k |\widehat{u}(k)|. \tag{6.3.20}$$

Lemma 6.3.1. *Let $t > 0$, $|k| \leq H$, and $n \geq 0$. Suppose that $\widehat{\theta}^0(l_0) \neq 0$ or $\widehat{\theta}^0(-l_0) \neq 0$, and $\widehat{\theta}^0(l) = 0$ for all $|l| < l_0$ with $l_0 \geq 1$. Then, we have*

$$\left\|\varepsilon(k)\left(\widehat{\theta}^{1,n} - \widehat{\theta}^1\right)(k)(t)\right\| \leq C(n), \tag{6.3.21}$$

where

$$C(n) = \left[\frac{2H^2\varepsilon(l_0)\|\widehat{u}_0\|}{N}\left(1 - \exp\left(-\frac{t}{\varepsilon(l_0)}\right)\right)\right]^n \left\|\varepsilon(k)\left(\widehat{\theta}^{1,0} - \widehat{\theta}^1\right)(k)(t)\right\|. \tag{6.3.22}$$

Proof. Let $w^n(k) = \widehat{\theta}^{1,n} - \widehat{\theta}^1$. We first prove (6.3.21) by induction on n. For $n = 0$, it obviously follows.

Assume that it holds at order $n-1$ with $n \geq 1$. That is, we have

$$\left\|\varepsilon(k)w^{n-1}(k)(t)\right\| \leq C(n-1). \tag{6.3.23}$$

Subtracting Eqs. (6.3.17) from Eq. (6.3.19) we find that

$$\varepsilon(k)w^n_t(k) + w^n(k) = -\frac{ik}{N}\sum_l \varepsilon(l)w^{n-1}(l)\widehat{\theta}^0(k-l). \tag{6.3.24}$$

At order n, using an integrating factor we obtain

$$\varepsilon(k)w^n(k) = -ike^{-\frac{t}{\varepsilon(k)}}\int_0^t e^{\frac{s}{\varepsilon(k)}}\frac{1}{N}\sum_l \varepsilon(l)w^{n-1}(l)\widehat{\theta}^0(k-l)ds. \tag{6.3.25}$$

Hence, since $e^{\frac{s-t}{\varepsilon(k)}} \leq 1$ and $\|\widehat{\theta}^0(k-l)\| \leq \|\widehat{u}_0\|e^{-\frac{s}{\varepsilon(l_0)}}$, using (6.3.23) applied to $\varepsilon(l)w^{n-1}(l)$ and noting that $\widehat{\theta}^0(0) = 0$ with (6.3.15) we find that

$$|\varepsilon(k)w^n(k)| \leq C(n-1)\|\widehat{u}_0\|\frac{2H^2}{N}\left|\int_0^t e^{-\frac{s}{\varepsilon(l_0)}}ds\right|. \tag{6.3.26}$$

This proves (6.3.21).

□

6.3.3 New Schemes with Correctors

Utilizing the correctors $\widehat{\theta}^0(k)$, $\widehat{\theta}^1(k)$, we now consider equations for $\widehat{v}(k)$. Eq. (6.3.1) is modified as follows. Since $L_\varepsilon(\widehat{v}) = L_\varepsilon(\widehat{u}) - L_\varepsilon(\widehat{\theta})$ and $L_\varepsilon(\widehat{\theta}^0) = 0$,

- With *Cor* 1: $\widehat{\theta}(k) = \widehat{\theta}^0(k)$.

  Let

$$\widehat{u}(k) = \widehat{v}(k) + \widehat{\theta}^0(k). \tag{6.3.27}$$

$$\begin{cases} \widehat{v}_t(k) + \dfrac{1}{\varepsilon(k)}\widehat{v}(k) = L_\varepsilon(\widehat{u}) = -\dfrac{ik}{2}(\widehat{u} * \widehat{u})(k) + \widehat{f}(k), \\ \widehat{v}(k) = 0, \quad \text{at } t = 0. \end{cases} \tag{6.3.28}$$

- With *Cor* 2: $\widehat{\theta}(k) = \widehat{\theta}^0(k) + \varepsilon(k)\widehat{\theta}^1(k)$.

  Let

$$\widehat{u}(k) = \widehat{v}(k) + \widehat{\theta}^0(k) + \varepsilon(k)\widehat{\theta}^1(k). \tag{6.3.29}$$

From (6.3.14) we have

$$\begin{cases} \widehat{v}_t(k) + \dfrac{1}{\varepsilon(k)}\widehat{v}(k) = -\varepsilon(k)L_\varepsilon(\widehat{\theta}^1(k)) + L_\varepsilon(\widehat{u}) \\ \quad = \dfrac{ik}{2}\left((\widehat{\theta}^0 + 2\varepsilon\widehat{\theta}^1) * \widehat{\theta}^0\right)(k) - \dfrac{ik}{2}(\widehat{u} * \widehat{u})(k) + \widehat{f}(k), \\ \widehat{v}(k) = 0, \quad \text{at } t = 0. \end{cases} \tag{6.3.30}$$

Applying the IFRK and ETDRK methods to Eqs. (6.3.28) and (6.3.30) to approximate $\widehat{v}(k)$, and adding to $\widehat{\theta}^0$, $\widehat{u}(k)$ is obtained from Eqs. (6.3.27) and (6.3.29), respectively.

Remark 6.3.1.

Since the stiff linear terms in Eqs. (6.3.28) and (6.3.30) will be absorbed by the integrating factor used in the IFRK and ETDRK methods, as indicated in [22], [74], the accuracy of the schemes depends on the nonlinear terms, i.e. the right-hand sides of (6.3.28) and (6.3.30), respectively,

$$F(\widehat{u}, t) = -\frac{ik}{2}(\widehat{u} * \widehat{u})(k) + \widehat{f}(k), \tag{6.3.31}$$

and, after some elementary calculations,

$$F(\widehat{u}, t) = (\text{R.H.S. of } (6.3.30)) = -\frac{ik}{2}\left((2\widehat{u} - \widehat{v}) * \widehat{v} + (\varepsilon\widehat{\theta}^1) * (\varepsilon\widehat{\theta}^1)\right)(k) + \widehat{f}(k). \tag{6.3.32}$$

We observe that the former one is similar to that in Eq. (6.3.1), which shows no benefits of the correctors; while in the latter case, with the presence of the correctors, the nonlinear term $F(t, \widehat{u})$ is much milder, yielding much better numerical errors compared with those of the classical schemes (see the numerical evidences in Sec. 6.4).

6.4 Numerical Results

In this section, we show the numerical simulations for cases discussed in the previous sections. We test our new schemes for some ODEs with a rapid decay and a rapid oscillation, the dispersive KdV equation for the PDE case.

For notations, hereafter, we use IFRK2, IFRK4 for the classical 2^{nd}- and 4^{th}-order IF schemes without using any correctors; IFRK2 Cor1, IFRK4 Cor2 indicate the schemes incorporating the correctors θ^0 and $\theta^0 + \varepsilon\theta^1$, respectively. We follow similar notations for the ETDRK and the classical Runge-Kutta (RK) schemes. We remind that for the ETDRK schemes, we will use the complex contour integral for stability purposes (see Section 6.2.1.2).

For all below numerical tests, we measure the relative errors. For ODEs, we use

$$E_{rel} = \left|\frac{u_N(T) - u_{EX}(T)}{u_{EX}(T)}\right|. \tag{6.4.1}$$

Here, u_N is the numerical solution obtained from some schemes, and u_{EX} is the exact one. For PDEs, using a discrete L^2 norm in space, we use

$$E_{rel} = \left(\frac{\sum_{j=0}^{N-1} (u_N(x_j,T) - u_{EX}(x_j,T))^2}{\sum_{j=0}^{N-1} (u_{EX}(x_j,T))^2} \right)^{1/2}. \tag{6.4.2}$$

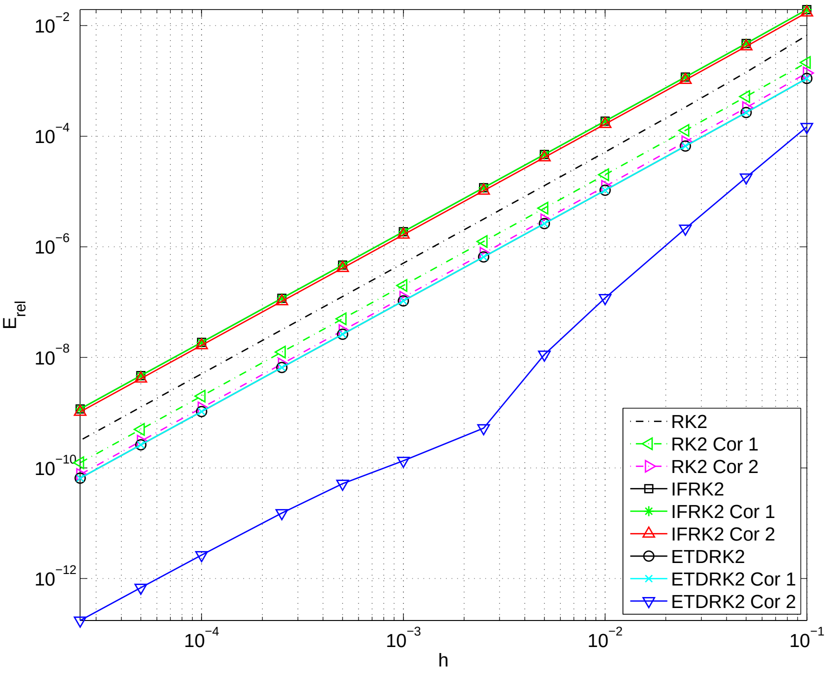


Figure 6-1: Relative numerical errors of 2^{nd}-order schemes for Eq. (6.4.3) at time $t = 1$. Initial condition $u_0 = 1 + \varepsilon$, $\varepsilon = 0.25$.

6.4.1 Stiff ODEs

We consider: for $\varepsilon > 0$,

$$\begin{cases} u_t + \dfrac{1}{\varepsilon} u = F(u,t) := u^2 + e^t + \varepsilon(1-2u)e^t + \varepsilon^2 e^{2t}, \\ u(0) = u_0 = a + \varepsilon. \end{cases} \tag{6.4.3}$$

The exact solution of Eq. (6.4.3) is then available

$$u(t) = \frac{a e^{-t/\varepsilon}}{1 - \varepsilon a(1 - e^{-t/\varepsilon})} + \varepsilon e^t. \tag{6.4.4}$$

From Eqs. (6.2.39) with $F^0(u,t) = u^2 + e^t$, thus $G^0(\theta^0(s), s) = (\theta^0(s))^2$, and from (6.2.39) we obtain

$$\theta^0(t) = u_0 e^{-t/\varepsilon}, \tag{6.4.5}$$

$$\theta^1(t) = (u_0)^2 (e^{-t/\varepsilon} - e^{-2t/\varepsilon}). \tag{6.4.6}$$

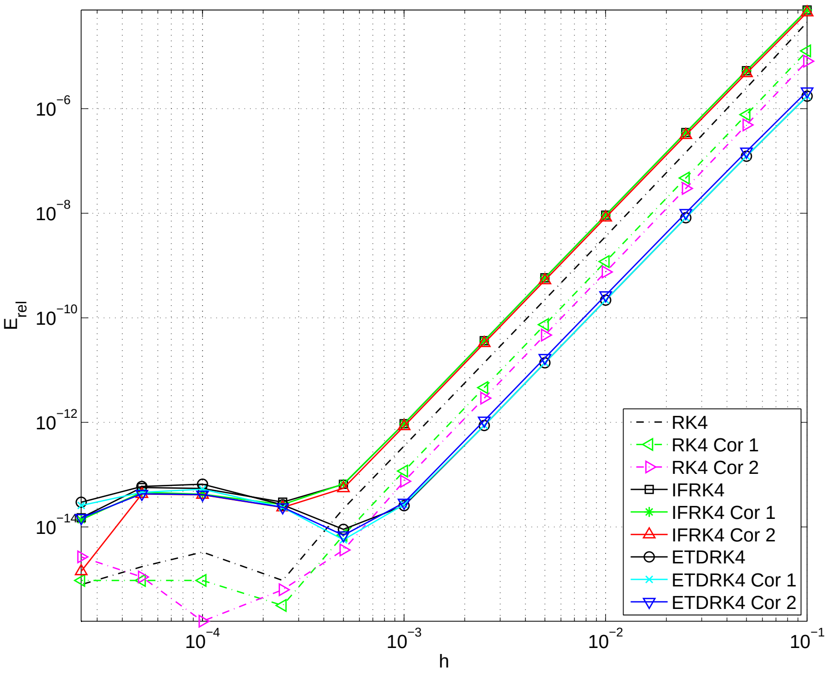


Figure 6-2: Relative numerical errors of 4^{th}-order schemes for Eq. (6.4.3) at time $t = 1$. Initial condition $u_0 = 1 + \varepsilon$, $\varepsilon = 0.25$.

The schemes with Cor1 (6.2.41) and Cor2 (6.2.43) are then written as follows, respectively,

$$\begin{cases} v_t + \dfrac{1}{\varepsilon} v = F(v + \theta^0, t), \\ v(0) = 0, \end{cases} \tag{6.4.7}$$

and

$$\begin{cases} v_t + \dfrac{1}{\varepsilon} v = F(v + \theta^0 + \varepsilon\theta^1, t) - (\theta^0)^2, \\ v(0) = 0. \end{cases} \tag{6.4.8}$$

Applying the IFRK and ETDRK methods discussed in the previous section, we numerically approximate v. The solution u is retrieved from Eq. (6.2.38). That is, $u = v + \theta^0$ or $u = v + \theta^0 + \varepsilon\theta^1$.

The relative numerical errors of this case are plotted in log-log scales in Fig. 6-1 for 2^{nd}-order methods, and in Fig. 6-2 for 4^{th}-order schemes. Here, we have chosen $u_0 = 1 + \varepsilon$, $\varepsilon = 0.25, 0.05$, and run the simulations up to time $t = 1$. In the numerical simulations, if we only use θ^0 as the corrector, the new schemes are similar to the classical IFRK and ETDRK, i.e., the ones without correctors. This implies that classical IFRK and ETDRK absorb the singularities corresponding to θ^0. On the contrary, when $\varepsilon\theta^1$ is added, the new schemes show better results than all the others. The difference in errors of the schemes are listed in Table 6-1 for the 2^{nd}- and 4^{th}-order schemes, respectively. Among all schemes, the ETDRK2 and ETDRK4 give the best results.

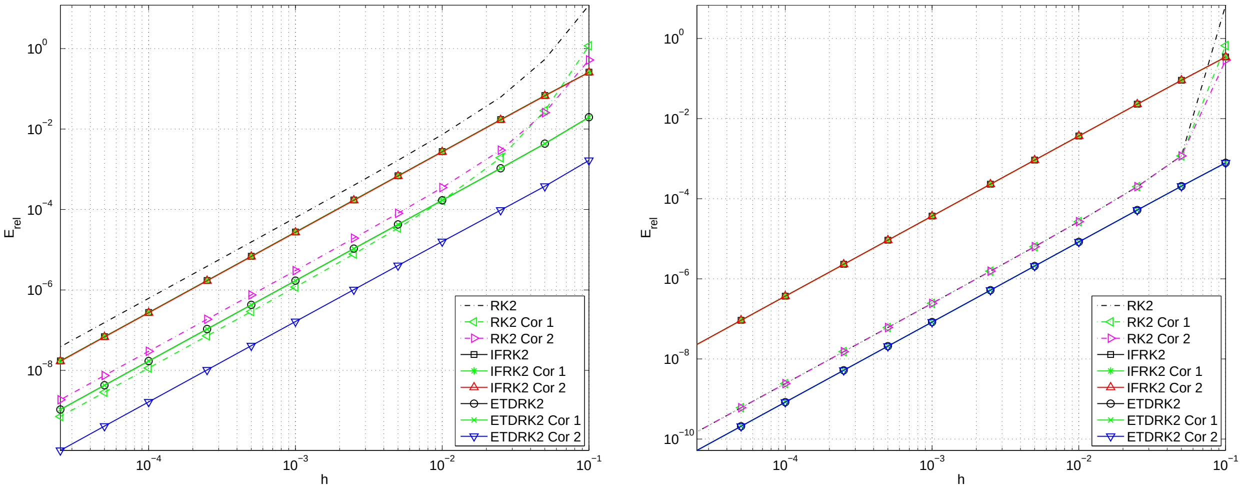


Figure 6-3: Relative numerical errors of 2^{nd}-order schemes for Eq. (6.4.3) at time $t = 0.2$(left), $t = 1$(right). Initial condition $u_0 = 1 + \varepsilon$, $\varepsilon = 0.05$.

Table 6-1: Relative numerical errors of 2^{nd} and 4^{th}-order schemes for Eq. (6.4.3) at time $t = 1$. Initial condition $u_0 = 1 + \varepsilon$, $\varepsilon = 0.25$.

h	RK2	RK2 Cor1	RK2 Cor2	IFRK2	IFRK2 Cor2	ETDRK2	ETDRK2 Cor2
10^{-1}	6.68E-03	2.16E-03	1.27E-03	1.94E-02	1.91E-02	1.11E-03	1.48E-04
10^{-2}	5.17E-05	2.01E-05	1.13E-05	1.87E-04	1.84E-04	1.05E-05	1.19E-07
10^{-3}	5.06E-07	1.99E-07	1.11E-07	1.87E-06	1.84E-06	1.05E-07	1.35E-10
10^{-4}	5.05E-09	1.99E-09	1.11E-09	1.87E-08	1.84E-08	1.05E-09	2.66E-12
h	RK4	RK4 Cor1	RK4 Cor2	IFRK4	IFRK4 Cor2	ETDRK4	ETDRK4 Cor2
10^{-1}	4.51E-05	1.27E-05	8.10E-06	7.74E-05	7.71E-05	1.74E-06	2.15E-06
10^{-2}	3.61E-09	1.19E-09	7.52E-10	9.22E-09	9.19E-09	2.18E-10	2.71E-10
10^{-3}	3.54E-13	1.17E-13	7.46E-14	9.42E-13	9.38E-13	2.56E-14	2.92E-14
10^{-4}	3.31E-15	9.46E-16	1.58E-16	5.46E-14	4.57E-14	6.58E-14	4.15E-14

As indicated in the correctors (6.4.5), (6.4.6), if ε is very small, with $Cor2$, i.e. $\theta = \theta^0 + \varepsilon\theta^1$, the term $\varepsilon\theta^1$ is not effective. Hence, we demonstrate our correctors for relatively small ε's. For $\varepsilon = 0.25$ with time $t = 1$, in the 2^{nd}-order schemes ETDRK2 with $Cor2$ outperforms the others as in Fig. 6-1. In the 4^{th}- order schemes as in Fig. 6-2, the schemes with $Cor1$ and $Cor2$ similarly behave in this example. This indicates that the 4^{th}- order schemes capture the stiffness corresponding to $\varepsilon\theta^1$. However, when the step size is close to 10^{-4}, the error plots for all schemes get flattened. This is because the relative errors are close to the machine precision which is also observed in e.g. [22].

Figs. 6-3 and 6-4 demonstrate that the corrector methods are effective near the thickness of the transition layers $\mathcal{O}(\varepsilon) = \mathcal{O}(0.05)$ which are captured by the correctors. See the left figures there for $t = 0.2$. However, for large time (away from the layers), e.g. $t = 1$ in this example, the schemes without and with correctors all similarly behave. As indicated in (6.2.14), for large time, we also note that the solutions are close to the slow part which can be well approximated by all the schemes.

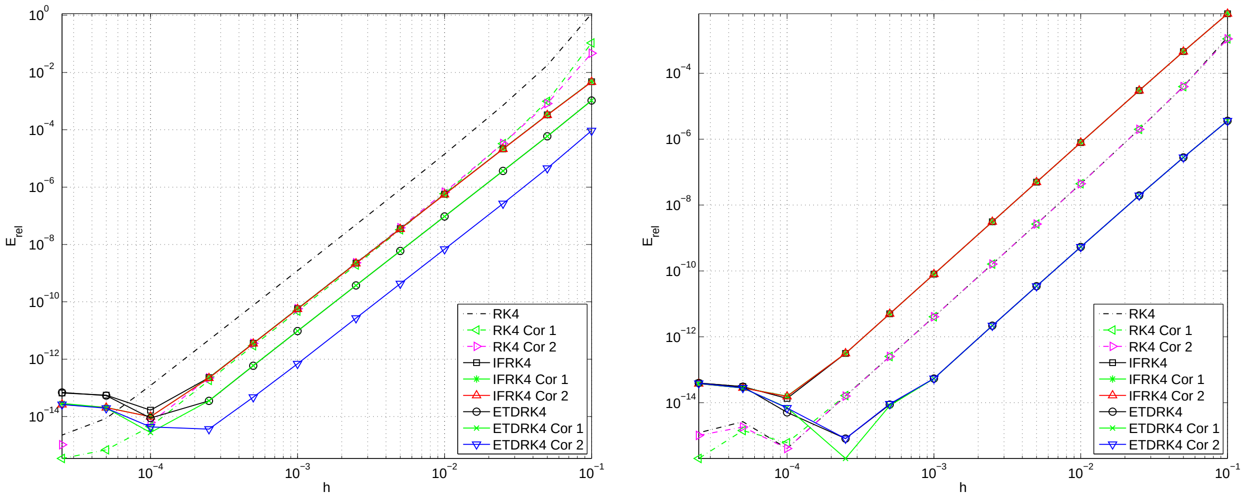


Figure 6-4: Relative numerical errors of 4^{th}-order schemes for Eq. (6.4.3) at time $t = 0.2$(left), $t = 1$(right). Initial condition $u_0 = 1 + \varepsilon$, $\varepsilon = 0.05$.

Table 6-2: Relative numerical errors of 2^{nd} and 4^{th}-order schemes for Eq. (6.4.9) at time $t = 1$. Initial condition $u_0 = 1$, $\varepsilon = 0.01$.

h	RK2	RK2 Cor1	RK2 Cor2	IFRK2	IFRK2 Cor2	ETDRK2	ETDRK2 Cor2
10^{-1}	NaN	NaN	NaN	1.29E-02	1.96E-03	1.09E-02	7.25E-05
10^{-2}	NaN	3.94E+01	1.23E+01	4.49E-04	2.83E-05	1.68E-03	4.31E-05
10^{-3}	1.66E-01	1.68E-03	1.71E-05	4.38E-06	2.57E-07	1.76E-05	4.35E-07
10^{-4}	1.65E-03	1.67E-05	1.69E-07	4.38E-08	2.57E-09	1.76E-07	4.34E-09
h	RK4	RK4 Cor1	RK4 Cor2	IFRK4	IFRK4 Cor2	ETDRK4	ETDRK4 Cor2
10^{-1}	NaN	NaN	NaN	2.59E-02	4.65E-04	8.96E-03	7.30E-05
10^{-2}	6.10E-01	6.13E-03	6.03E-05	1.92E-06	6.30E-07	6.10E-05	3.11E-07
10^{-3}	8.26E-05	8.41E-07	8.33E-09	1.84E-10	5.47E-11	2.78E-09	3.85E-11
10^{-4}	8.26E-09	8.41E-11	8.33E-13	1.47E-13	5.45E-15	3.42E-13	3.72E-15

If the thickness of the layers is longer, e.g. $\mathcal{O}(\varepsilon) = \mathcal{O}(0.25)$ in Fig. 6-1, the correctors are effective for larger times, e.g. $t = 1$.

For models with a rapid oscillation, our corrector methods turn out to be outstandingly effective. In this case, the oscillatory sharp transitions appear periodically, and so do the correctors in this case (see (6.4.11), (6.4.12) below). For large time, still the correctors are much effective unlike the dissipative case, i.e. (6.4.3).

For this, we now test the following oscillatory problem: for $\varepsilon > 0$,

$$\begin{cases} u_t - \dfrac{i}{\varepsilon} u = F(u) := -iu^2, \\ u(0) = u_0 = a. \end{cases} \tag{6.4.9}$$

Then the exact solution is known,

$$u(t) = \frac{ae^{it/\varepsilon}}{1 - \varepsilon a(1 - e^{it/\varepsilon})}. \tag{6.4.10}$$

Replacing ε by $i\varepsilon$ in Section 6.2.2.2 and (6.2.39) with $F^0(u,t) = -iu^2$, we obtain the correctors,

$$\theta^0(t) = u_0 e^{it/\varepsilon}, \tag{6.4.11}$$

$$\theta^1(t) = -i(u_0)^2(e^{it/\varepsilon} - e^{2it/\varepsilon}), \tag{6.4.12}$$

and with conventional schemes, we numerically solve the following problems with Cor1 (6.2.41) and Cor2 (6.2.43),

$$\begin{cases} v_t - \dfrac{i}{\varepsilon} v = F(v + \theta^0), \\ v(0) = 0, \end{cases} \tag{6.4.13}$$

$$\begin{cases} v_t - \dfrac{i}{\varepsilon} v = F(v + \theta^0 + i\varepsilon\theta^1) + i(\theta^0)^2, \\ v(0) = 0. \end{cases} \tag{6.4.14}$$

and then obtain $u = v + \theta^0$, $u = v + \theta^0 + i\varepsilon\theta^1$, respectively.

Relative numerical errors of all schemes are plotted in Figs. 6-5 - 6-6, and the errors for this case are listed in Table 6-2. The schemes with Cor2 outperform the ones without correctors or with only Cor1, and the IFRK Cor2 and ETDRK Cor2 show the best results.

6.4.2 The KdV Equation

For the KdV equation, due to the dispersive property, the Fourier modes of the solution do not decay with time, but oscillate rapidly as the wave number k increases. In the below results, we demonstrate numerically that the corrector (6.3.14) is able to absorb these oscillations analytically. For the first test, we try the following initial conditions.

- Initial condition 1:

$$u_0(x) = \sin(x), \tag{6.4.15}$$

whose discrete Fourier transform gives

$$\widehat{u}_0(k) = \begin{cases} -\dfrac{N}{2}i, & \text{for } k = 1, \\ \dfrac{N}{2}i, & \text{for } k = -1, \\ 0, & \text{otherwise}; \end{cases} \tag{6.4.16}$$

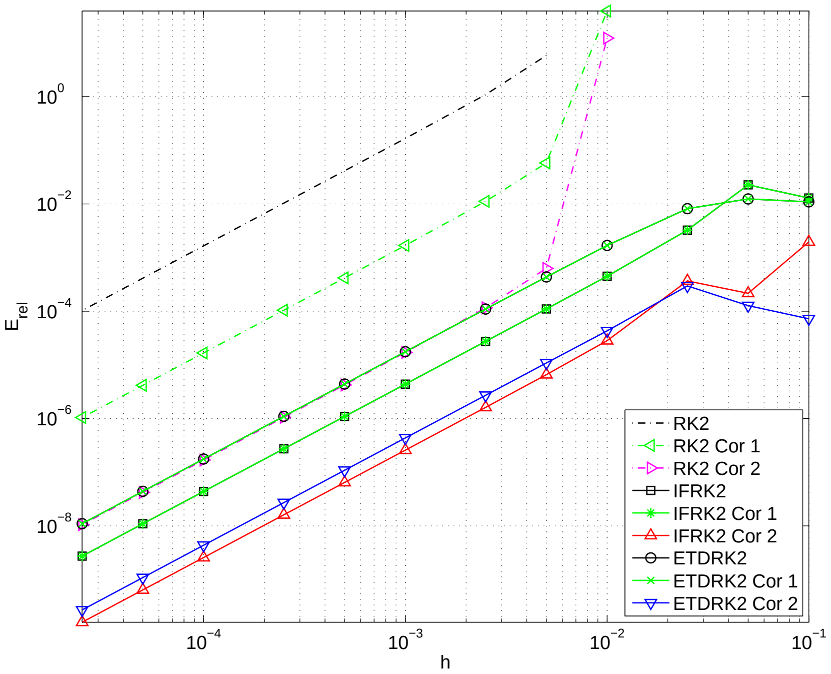


Figure 6-5: Relative numerical errors of 2^{nd}-order schemes for Eq. (6.4.9) at time $t = 1$; Initial condition $u_0 = 1$, $\varepsilon = 0.01$.

thus the correctors $\widehat{\theta}^0(k)$'s are

$$\begin{cases} \widehat{\theta}^0(k) = \widehat{u}_0(k)e^{ikt}, \quad k = \pm 1, \\ \widehat{\theta}^0(k) = 0, \qquad \text{otherwise k's;} \end{cases} \tag{6.4.17}$$

where $\widehat{u}_0(k)$'s follow those given in Eq. (6.4.16), and $\widehat{\theta}^1(k)$, $k = 1, \dots, 4$ are computed with $n = 4$ iterations as follows

$$\widehat{\theta}^1(1) = \frac{283}{432}ie^{it}t - \frac{53}{486}e^{7it} + \frac{1}{36}e^{it}t^2 - \frac{1}{31104}e^{25it} + \frac{377}{3456}e^{it}; \tag{6.4.18}$$

$$\begin{aligned} \widehat{\theta}^1(2) = \frac{2831}{270}ie^{2it} - \frac{206138}{19683}ie^{8it} + \frac{47}{54}te^{2it} - \frac{2}{729}e^{8it}t \\ - \frac{1}{27}it^2e^{2it} - \frac{215}{17496}ie^{26it} - \frac{1}{787320}ie^{62it}; \end{aligned} \tag{6.4.19}$$

$$\widehat{\theta}^1(3) = \frac{1413}{640}e^{3it} - \frac{319}{108}e^{9it} - \frac{3}{16}ite^{3it} + \frac{1}{4320}e^{63it} + \frac{859}{1152}e^{27it}; \tag{6.4.20}$$

$$\begin{aligned} \widehat{\theta}^1(4) = -\frac{353}{2025}ie^{4it} + \frac{10784}{41553}ie^{10it} - \frac{2}{135}te^{4it} - \frac{1147}{11664}ie^{28it} \\ + \frac{1}{615600}ie^{124it} + \frac{718}{54675}ie^{64it}; \end{aligned} \tag{6.4.21}$$

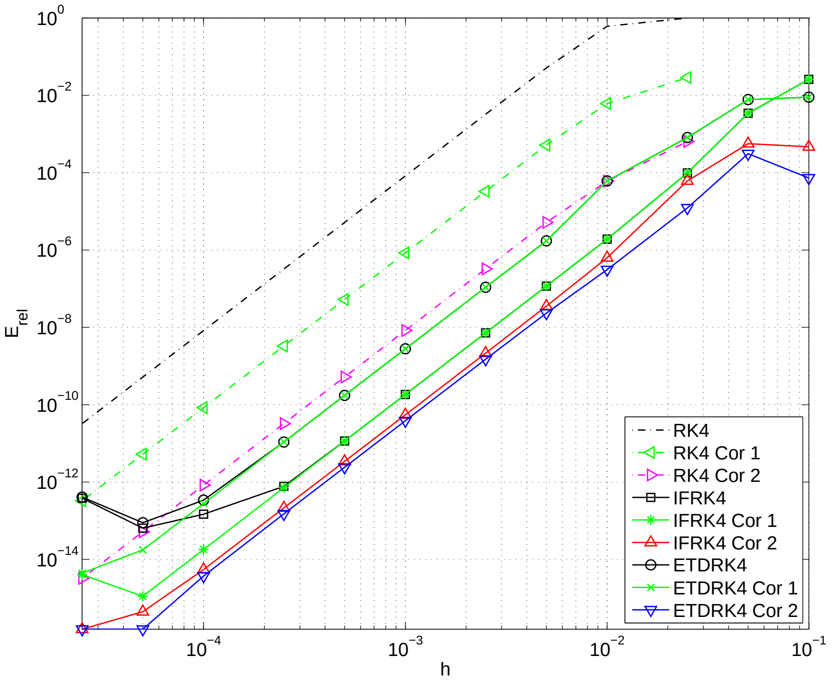


Figure 6-6: Relative numerical errors of 4^{th}-order schemes for Eq. (6.4.9) at time $t = 1$; Initial condition $u_0 = 1$, $\varepsilon = 0.01$.

$\widehat{\theta}^1(k)$, $k = -2, -4$ are the corresponding complex conjugates, and $\widehat{\theta}^1(k) = 0$ for other k's.

The numerical relative errors of this case are shown in Fig. 6-7 for both 2^{nd}- and 4^{th}-order schemes, and the correctors are plotted together with the reference solution in Fig. 6-8. We observe that the schemes with Cor 2 are better than the conventional ones, especially for the 2^{nd}-order schemes. This results from the well approximation of the corrector 2 to the oscillations. In fact, up to $k = 3$, the reference solution and the corrector 2 at each mode are almost indistinguishable.

- Initial condition 2:

$$u_0(x) = \sin(x) + \sin(2x), \tag{6.4.22}$$

whose discrete Fourier transform gives

$$\widehat{u}_0(k) = \begin{cases} -\dfrac{N}{2}i, & \text{for } k = 2, \\ -\dfrac{N}{2}i, & \text{for } k = 1, \\ \dfrac{N}{2}i, & \text{for } k = -1, \\ \dfrac{N}{2}i, & \text{for } k = -2, \\ 0, & \text{otherwise.} \end{cases} \tag{6.4.23}$$

The correctors $\widehat{\theta}^0(k)$ and $\widehat{\theta}^1(k)$ can be obtained following a similar process. We omit presenting them here due to the complexity.

The numerical results at time $t = 2$ with $n = 32$ are plotted in Fig. 6-9, and the correctors are shown in Fig. 6-10. We note a better performance of the new schemes. In Fig. 6-10, for $k = 4$, we emphasize the well approximation of the corrector 2 to the exact solution, even though the oscillations are quite rapid.

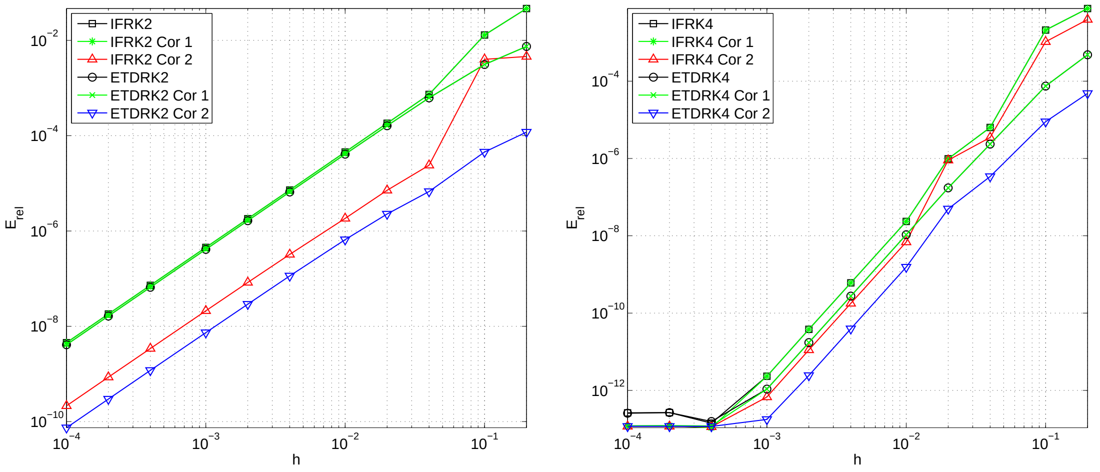


Figure 6-7: Relative numerical errors for Eq. (6.1.1) with the initial condition (6.4.15) at time $t = 1$. Left: 2^{nd}-order; Right: 4^{th}-order schemes.

Finally, we test the inhomogeneous cases where the forcing term f in Eq. (6.1.1) is non-zero. We choose the following initial condition $u_0(x)$ as follows,

$$u_0(x) = \sin(x), \tag{6.4.24}$$

whose DFT modes follow Eq. (6.4.16) above.

Proceeding as the homogeneous case, we obtain the correctors $\widehat{\theta}^0(k)$'s and $\widehat{\theta}^1(k)$'s as in Eqs. (6.4.17), (6.4.18) - (6.4.21), respectively.

The relative numerical errors of these cases at time $t = 1$ with a grid of $N = 32$ are plotted in Fig. 6-11. Here, the reference solutions are obtained from an RK4 method with a fine time step $h = 10^{-7}$. We observe that the schemes with $Cor2$ give better results than the other comparing methods. We also notice that the ETDRK, both 2^{nd}- and 4^{th}-order, give the best results.

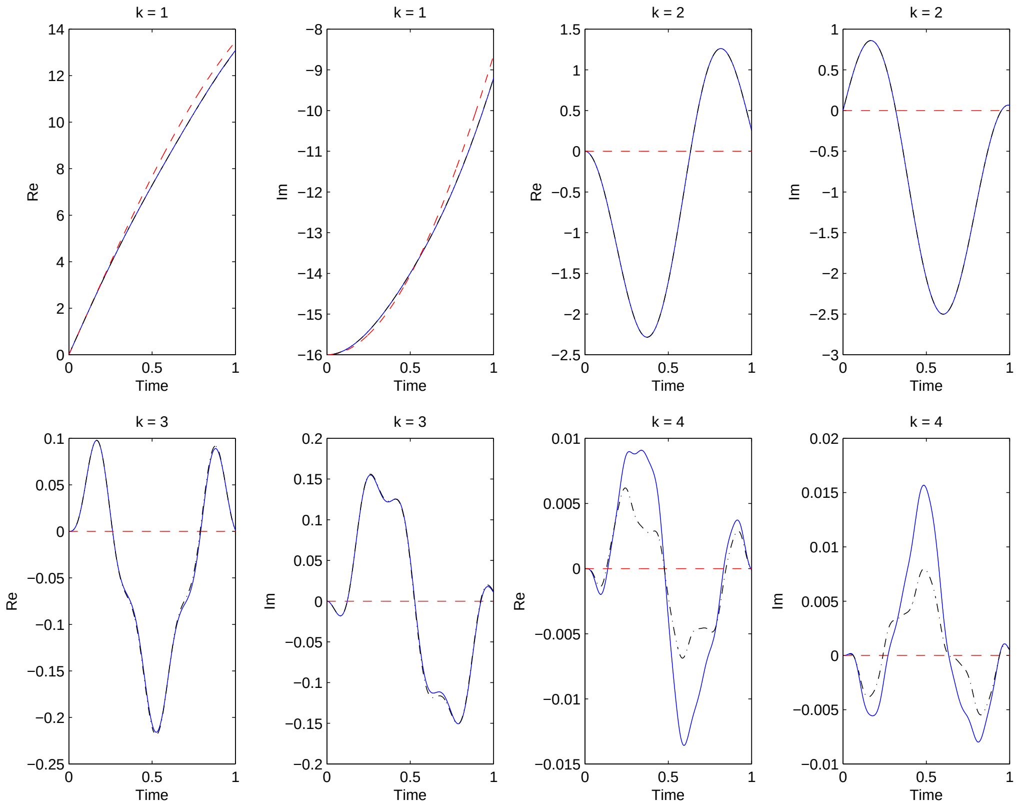


Figure 6-8: Correctors at each mode of Eq. (6.1.1) with the initial condition (6.4.15) at time $t = 1$. Solid: reference solution, dashed: corrector 1: $\widehat{\theta}(k) = \widehat{\theta}^0$, dash-dot: corrector 2: $\widehat{\theta}(k) = \widehat{\theta}^0 + i\varepsilon(k)\widehat{\theta}^1$.

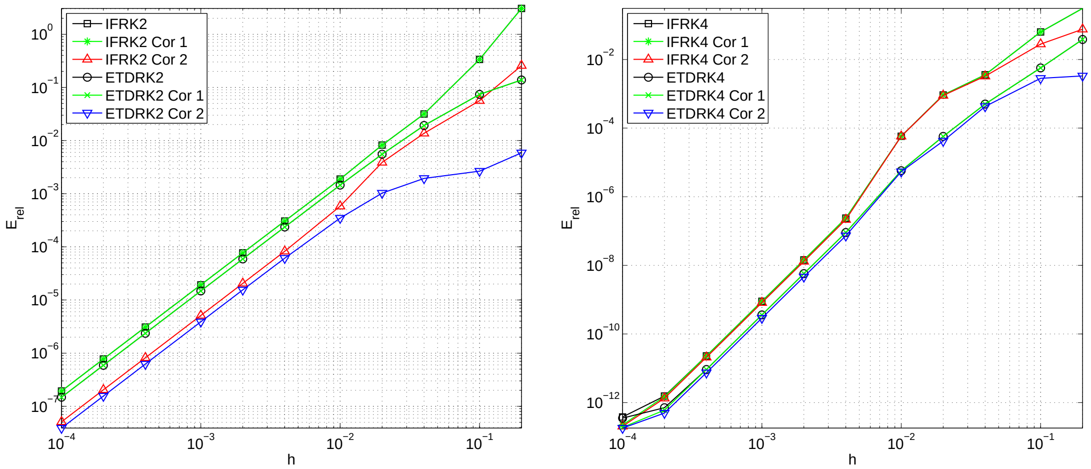


Figure 6-9: Relative numerical errors for Eq. (6.1.1) with the initial condition (6.4.22) at time $t = 2$. Left: 2^{nd}-order; Right: 4^{th}-order schemes.

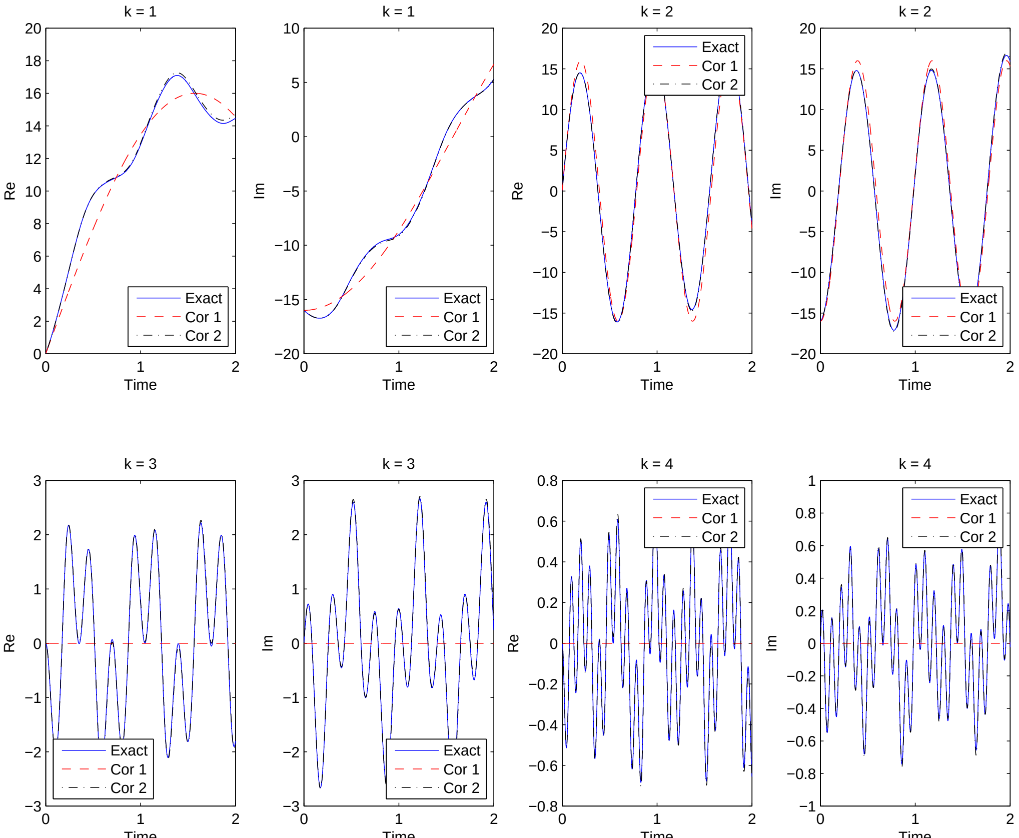


Figure 6-10: Correctors at each mode of Eq. (6.1.1) with the initial condition (6.4.22) at time $t = 2$. Solid: reference solution, dashed: corrector 1: $\widehat{\theta}(k) = \widehat{\theta}^0$, dash-dot: corrector 2: $\widehat{\theta}(k) = \widehat{\theta}^0 + i\varepsilon(k)\widehat{\theta}^1$.

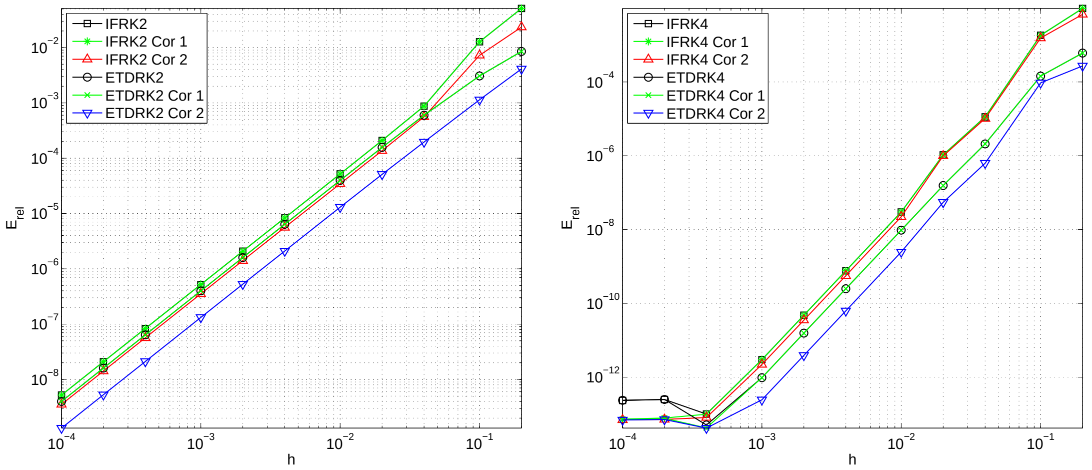


Figure 6-11: Relative numerical errors for KdV Eq. (6.1.1) at time $t = 1$ for the inhomogeneous case with initial condition (6.4.24), forcing term $f(x,t) = t\cos(x)$. Left: 2^{nd}-order; Right: 4^{th}-order schemes.

7

Conclusion

In this dissertation, we have presented our new numerical methods for different important partial differential equations which belong to the class of hyperbolic problems. These include conservation laws in gas dynamics, the 2D transverse Maxwell equations with uncertainties, and the dispersive KdV equation for waves. For all cases, our new methods improved the comparing conventional ones in terms of computational cost, accuracy, resolution, as well as robustness.

For the first part of the dissertation, we have constructed a new WENO-θ scheme for conservation laws. The new scheme adaptively switches between a 5th-order upwind and 6th-order central scheme, depending on the smoothness of not only the sub-stencils but also the large one. Unlike the other 6th-order WENO schemes in which this switch depends solely on the smoothness of the most downwind sub-stencils, it is that of the large stencil which decides this mechanism in our new scheme. The main features of the new scheme are that a sixth order is achieved in smooth regions and it overcomes the loss of accuracy which is detected for the WENO-NW6 and WENO-CU6 ones. We have also developed the new smoothness indicators of the sub-stencils $\tilde{\beta}_k$'s which are symmetric in terms of Taylor expansions around point x_j and a new τ^θ for the large stencil. The latter is chosen as the smoother one among two candidates which are computed based on the possible highest-order variations of the reconstruction polynomials in L^2-sense. From then, value of the parameter θ is determined to decide if the scheme is 5th-order upwind or 6th-order central.

A number of numerical tests for both scalar cases, linear and nonlinear, and system cases with the Euler equations of gas dynamics were carried out to check the accuracy, resolution, efficiency, and robustness of our new scheme. It has been shown that our new method is more accurate than WENO-JS, WENO-Z, and WENO-CU6; more efficient that WENO-NW6; more robust than WENO-CU6 and WENO-NW6; and outperforms comparing schemes in capturing small-scaled structures, and around critical regions.

In the second part, we then extended our scope by studying the uncertainty propagation from parameters pertaining to certain random media for which multiple media randomly interface, and proposed new semi-analytical IF and ETD methods for spectral methods applying to the KdV equation. For the former, multiple random parameters are present in the system considered here, and then the computational cost via the PC expansions grows exponentially with the dimensionality, i.e., the number of random parameters. To overcome such dimensionality issues, rather than the tensor products of one-dimensional nodal sets, the stochastic collocation on a sparse grid by, e.g. the Smolyak algorithm can be considered (see [153], [143]). However, this collocation method is for approximating the multi-dimensional integrations involved in the coefficients. To avoid the dimensionality problem, we computed the PC modes in each interval of a level set function z which makes the computational cost grow linearly with the dimensionality. Furthermore, evaluating the integrations via the explicit formula given in [68], we also avoided the integration errors with low computational cost. For numerical evidence, we considered up to five random parameters or interfaces, which showed promising results.

As mentioned above, in the last chapter we improved the well-known IF and ETD time integrators by incorporating the analytical approximations of the stiff parts in the solutions, which we named as correctors, and applied these to spectral methods for the KdV equation for waves propagation. The correctors are derived by invoking the singular perturbation analysis. By this, the stiffness of the problem is then resolved or captured by the correctors. Embedding these into the original IF and ETD methods, we further resolved the stiffness in the nonlinear term, which resulted in much better schemes comparing with the ones without correctors. We applied the correctors to improve the IFRK and ETDRK methods, both for 2nd- and 4th-order schemes. The application of the new methods to spectral methods for PDEs. Incorporating the correctors into the classical IF and ETD methods, we resolved the stiffness due to the stiffness linear term, i.e. $\frac{1}{\varepsilon(k)}\widehat{u}(k)$, in each Fourier mode k and this resulted in better accuracy of the schemes presented in the text. Here, a careful treatment of the correctors is required due to the interactions of all Fourier modes resulting from the quadratic nonlinear term uu_x. A variety on numerical evidence showed that our new semi-analytical methods a great deal outperformed the conventional ones in terms of accuracy, for ODEs of either rapid oscillations or rapid decay in time, as well as the KdV equation for the PDE case.

Finally, we discuss our relevant future works on these topics. For the WENO schemes, we will investigate problems of higher dimensions, say in 3D. Moreover, since the new smoothness indicators are constructed in a systematic manner, it is expected that the development of the scheme to higher orders would be feasible. On the other hand, thanks to the capability to capture small-scaled structures, we expect to apply WENO-θ in, e.g., the Navier-Stokes equations ([29]), or in direct numerical simulations ([102]), etc. For the PCE methods, with our new approach, increasing the number of random parameters only increases the computational time linearly, with a linear growth cost, more random parameters can be taken into account with the aid of parallel computations. We believe that our approach can be easily and efficiently applied to wave-typed equations or conservation laws with multiple random parameters. For the new semi-analytical methods, we aim to handle more complicated problems, for example, the one supplemented with a pulse-like initial condition. Here, due to the stiffness of the initial data, they need *many* Fourier modes in the spectral representations, i.e. H is large in Eq. (6.3.15). This causes considerable computational cost to obtain the explicit forms of the correctors $\widehat{\theta}^1(k)$ due to the modal interactions via the nonlinear term. This issue will be addressed elsewhere. If H is not large, as in the text, the solutions of the KdV equation involving a quadratic nonlinear term are well approximated by out methods. Other than quadratic nonlinear terms, we will approximate some polynomials or other types of nonlinear terms by mimicking the correctors $\widehat{\theta}^0(k)$ and $\widehat{\theta}^1(k)$. See also [75] for the our first attempt to the dissipative Kuramoto-Sivashinsky (KS) equation (see [89], [126], [133], [15], [105], or [58]) which have applications in flame-front propagation in laminar flames, phase dynamics in reaction-diffusion systems, etc. The interesting issue of this equation is the balance between the dissipative and anti-dissipative terms. In [75], we have tried with the case where the former term is dominant, whereas the other cases need some assumptions.

References

[1] H. Ashi, *Numerical Methods for Stiff Systems,* Ph.D. dissertation, the University of Nottingham, 2008. 107, 109

[2] Athena3D in Fortran 95, *http://www.astro.virginia.edu/VITA/athena.php.* 76, 77

[3] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri, *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations,* Appl. Num. Math., 25(2-3), pp. 151-167, 1997. 107

[4] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton, *Implicit-explicit methods for time-dependent partial differential equations,* SIAM J. Numer. Anal., Vol. 32, No. 3, pp. 797-823, 1995. 107

[5] Z. A., Aziz et. al., *Fourth-order time stepping for stiff PDEs via integrating factor,* Adv. Sci. Lett. 19(1) pp. 170-173. 107

[6] J. P. Boyd, *Chebyshev and Fourier Spectral Methods,* Dover, Mineola, NY, 2001. 107

[7] J. P. Boris, and D. L. Book, *Flux corrected transport, I, SHASTA, A fluid transport algorithm that works*, J. Comp. Phys., 11 (1973), 39-69. 29

[8] R. Borges, M. Carmona, B. Costa, and W. S. Don, *An improved weighted essentially non-oscillatory scheme for hyperbolic conservation laws*, J. Comput. Phys. 227 (2008), 3191-3211. 40, 49

[9] T. Barth, and M. Ohlberger, *Finite volume methods: foundation and analysis*, in *Encyclopedia of Computational Mechanics*, John Wiley & Sons, Ltd., 2004. 9, 11

[10] F. Bianco, G. Puppo, and G. Russo, *High order central schemes for hyperbolic systems of conservation laws*, SIAM J. Sci. Comput. 21, 294 (1999). 23

[11] A. Bressan, *Hyperbolic conservation laws: an illustrative tutorial*, Lecture Notes, Dept. Math., Penn State Univ., 2009. 5

[12] D. S. Balsara, and C.-W. Shu, *Monotonicity preserving weighted essentially non-oscillatory schemes with increasingly high order of accuracy*, J. Comput. Phys. 160, 405-452 (2000). 39

[13] M. Castro, B. Costa, and W. S. Don, *High order weighted essentially non-oscillatory WENO-Z schemes for hyperbolic conservation laws*, J. Comput. Phys. 230 (2011), 1766-1792. 40

[14] B. Costa, and W. S. Don, *High order hydrid central - WENO finite difference scheme for conservation laws*, IJCAM, 204 (2007), 209-218. 51

[15] P. Collet, J.-P. Eckmann, H. Epstein, and J. Stubbe, *Analyticity for the Kuramoto-Sivashinsky equation,* Physica D 67 (1993) 321-326, North-Holland, Amsterdam. 133

[16] M. H. Carpenter, T. C. Fisher, and N. K. Yamaleev, *Boundary closures for sixth-order energy-stable weighted essentially non-oscillatory finite-difference schemes*, Advances in Applied Mathematics, Modeling, and Computational Science, Fields Institute Communications Vol. 66, 2013, 117-160. 51

[17] Q.-Y. Chen, D. Gottlieb and J. S. Hesthaven, *Uncertainty analysis for the steady-state flows in a dual throat nozzle.* J. Comput. Phys. 204(2005), pp. 378–398. 86

[18] A. J. Chorin and O. H. Hald, *Stochastic Tools in Mathematics and Science.* Springer, New York, 2006. 86

[19] C. Chauvière, J. S. Hesthaven and L. Lurati, *Computational modeling of uncertainity in time-domain electromagnetics.* SIAM J. Sci. Comput. 28(2006), pp. 751–775. 85

[20] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics,* Springer series in computational physics, Springer-Verlag, Berlin, 1988. 105, 107

[21] M. G. Crandall, and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comp., Vol. 34, No. 149, 1980, pp. 1-21. 28

[22] S. M. Cox, and P. C. Matthews, *Exponential time differencing for stiff systems,* J. Comput. Phys. **176**, 430-455 (2002). 107, 108, 109, 111, 120, 123

[23] S. Charavarthy, and S. Osher, *High resolution applications of the Osher upwind scheme for the Euler equations*, AIAA paper presented at 6th CFD conference, 1983. 29, 31

[24] M. G. Crandall, and L. Tartar, *Some relations between non expansive and order preserving mapping*, Proc. Amer. Math. Soc. 78 (1980), pp. 385-390. 10

[25] P. Colella, and P. Woodward, *The piecewise-parabolic method (PPM) for gas dynamical simulations*, J. Comp. Phys., 54 (1984), 174-201. 32

[26] L. C. Evans, *Partial Differential Equations*, 2nd ed., American Mathematical Society, 2010. 7, 8

[27] Weinan E, Bjorn Engquist, Xiantao Li, Weiqing Ren, and Eric Vanden-Eijnden, *The Heterogeneous Multiscale Method: A Review,* Commun. Comput. Phys., Vol. 2, no. 3, pp. 367-450 (2007). 107

[28] B. Engquist, and S. Osher, *Stable and entropy satisfying approximations for transonic flow calculations*, Math. Comp., Vol. 34, No. 149, 1980, pp. 45-75. 27

[29] B. Epstein, and S. Peigin, *Application of WENO (weighted essentially non-oscillatory) approach to Navier-Stokes computations*, Int. J. Comput. Fluid D. 2004, Vol. 18 (3), 289-293. 132

[30] B. Engquist and Y.-H. Tsai, *Heterogeneous multiscale methods for stiff ordinary differential equations,* Math. Comp, Vol. 74, No.252, pp. 1707-1742, 2005. 107

[31] P. Fan, *High order weighted essentially nonoscillatory WENO-η schemes for hyperbolic conservation laws*, J. Comput. Phys. 269 (2014), 355-285. 40

[32] B. Fornberg, *A Practical Guide to Pseudospectral Methods,* Cambridge University Press, Cambridge, UK, 1996. 105, 107

[33] J.-P. Fouque, J. Garnier, G. Papanicolaou, and K. Sølna, *Wave Propagation and Time Reversal in Randomly Layered Media.* Springer, New York, 2007. 85

[34] H. Feng, F. Hu, and R. Wang, *A new mapped weighted essentially non-oscillatory scheme*, J. Sci. Comput. (2012), 51:449-473. 40

[35] P. Fan, Y. Shen, B. Tian, and C. Yang, *A new smoothness indicator for improving the weighted essentially non-oscillatory scheme*, J. Comput. Phys., 269 (2014), 329-354. 40, 76

[36] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., XVIII, 697-715 (1965). 19

[37] S. K. Godonov, *Finite-difference method for numerical computation of discontinuous solutions of the equations of fluid dynamics*, Mat. Sbornik, 47 (1959), pp. 271-306 (Russian). 23, 28

[38] J. Glimm, J. Grove, X. Li, W. Oh, and D. C. Tan, *The dynamics of bubble growth for Rayleigh-Taylor unstable interfaces*, Physics of Fluids 31 (1988), 447-465. 76

[39] J. B. Goodman, and R. J. LeVeque, *On the accuracy of stable schemes for 2D scalar conservation laws*, Math. Comp., Vol. 45, No. 171, 1985, 15-21. 22

[40] W. Gropp, E. Lusk, and A. Skjellum, *Using MPI: Portable Parallel Programming with the Message-Passing Interface.* 2nd ed., The MIT Press, Cambridge, Massachusetts, London, England, 1999. 100

[41] J. Garnier and G. Papanicolaou, *Pulse propagation and time reversal in random waveguides.* SIAM J. Appl. Math. 67(2007), pp. 1718-1739. 85

[42] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: a Spectral Approach*, Springer-Verlag, New York, 1991. 41, 86

[43] S. Gottlieb, and C.-W. Shu, *Total variation dimishing Runge-Kutta schemes*, Math. of Comp., Vol. 67, No. 221, (1998) 73-85. 41, 86

[44] D. Gottlieb and Dongbin Xiu, *Galerkin method for wave equations with uncertain coefficients.* Commun. Comput. Phys. 3(2008), pp. 505-518. 86

[45] A. Guadagnini, D. M. Tartakovsky and C. L. Winter, *Random domain decomposition for flow in heterogeneous stratified aquifers*, Stochastic Environ. Res. and Risk Assessment (SERRA), vol. 17, no. 6 (2003), pp. 394-407. 85

[46] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys. 49, 357-393 (1983). 10, 21, 22, 28

[47] A. Harten, *On a class of high resolution total-variation-stable finite-difference schemes*, SIAM J. Numer. Anal., Vol. 21, No. 1, 1984.

[48] M. H. Holmes, *Introduction to Perturbation Methods*, Springer, New York, 1995. 110

[49] X. Y. Hu, and N. A. Adams, *Scale separation for implicit large eddy simulation*, J. Comput. Phys., 230 (2011), 7240-7249. 51

[50] A. K. Henrick, T. D. Aslam, and J. M. Powers, *Mapped weighted essentially non-oscillatory schemes: Achieving optimal order near critical points*, J. Comput. Phys. 207 (2005), 542-567. 39, 47, 48, 63

[51] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy, *Uniformly high order accurate non-oscillatory schemes, III*, J. Comput. Phys. 131, 3-47 (1997). 21, 33, 35, 38

[52] H. Han and R. B. Kellogg, *A method of enriched subspaces for the numerical solution of a parabolic singular perturbation problem.* In: Computational and Asymptotic Methods for Boundary and Interior Layers, Dublin, pp.46-52 (1982). 111

[53] Y. Ha, C. H. Kim, Y. J. Lee, and J. Yoon, *An improved weighted essentially non-oscillatory scheme with a new smoothness indicator*, J. Comput. Phys., 232 (2013), 68-86. 40

[54] A. Harten, and P. Lax, *A random choice finite difference scheme for hyperbolic conservation laws*, SIAM J. Numer. Anal., 18(2), 1981. 19

[55] A. Harten, P. Lax, and B. Van Leer, *On upstream differencing and Godonov-type schemes for hyperbolic conservation laws*, SIAM Rev., 25 (1983), 35-61. 9, 15, 20, 23, 27

[56] A. Harten, J. M. Hyman, and P. D. Lax, *On finite-difference approximations and entropy conditions for shocks*, Comm. Pure Appl. Math., Vol XXIX, 297-322 (1976). 27

[57] Thomas Y. Hou, Wuan Luo, Boris Rozovskii and Hao-Min Zhou, *Wiener chaos expansions and numerical solutions of randomly forced equations of fluid mechanics.* J. of Comput. Phys. 216 (2006), 687–706. 86

[58] J. M. Hyman, and B. Nicolaenko, *The Kuramoto-Sivashinsky equation: a bridge between PDE's and dynamical systems,* Physica D, Vol. 18, pp 113-126 (1986). 133

[59] A. Harten, and S. Osher, *Uniformly high-order accurate nonoscillatory schemes, I,* SIAM J. Numer. Anal., Vol. 24, No. 2, 1987, 279-309. 33

[60] A. Harten, S. Osher, B. Engquist, and S. R. Chakravarthy, *Some results on uniformly high-order accurate essentially nonoscillatory schemes*, Applied Numerical Mathematics 2 (1986), 347-377. 33, 36, 37

[61] D. J. Hill, and D. I. Pullin, *Hydrid tuned center-difference WENO method for large-eddy simulations in the presence of strong shocks*, J. Comput. Phys. 194 (2004), 435-450. 51

[62] F. D. L. Hoz, and F. Vadillo, *An exponential time differencing method for the nonlinear Schrödinger equation,* Comp. Phys. Comm., **179** (2008) 449-456. 107

[63] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*, Springer, 1996.

[64] X. Y. Hu, Q. Wang, and N. A. Adams, *An adaptive central-upwind weighted essentially non-oscillatory scheme*, J. Comput. Phys., 229 (2010), 8952-8965. 51, 54, 55

[65] V. Isakov, *Inverse Problems for Partial Differential Equations*, Springer Science+Business Media, Inc, New York, 2006. 85

[66] R. S. Johnson, *Singular Perturbation Theory*, Springer Science+Business Media, Inc., New York, 2005. 110

[67] C. Jung, *Finite elements scheme in enriched subspaces for singularly perturbed reaction-diffusion problems on a square domain*, Asymptot. Anal. **57**, 41-69 (2008). 111

[68] C. Jung, *Evolution of probability distribution in time for solutions of hyperbolic equations.* Journal of Scientific Computing. 41 (2009), No. 1, 13–48. 86, 88, 93, 94, 132

[69] C. Jung, B Kwon, A. Mahalov, and T. B. Nguyen, *Maxwell solutions in media with multiple random interfaces,* Int. J. Numer. Anal. Mod., Vol. 11, No. 1, pp. 193-212 (2014). 87

[70] A. Jameson, and P. D. Lax, *Conditions for the construction of multi-point total variation diminishing difference schemes*, Princeton Univ., MAE Report 1650, 1984. 107

[71] S. Jin and C. D. Levermore, *Numerical Schemes for Hyperbolic Conservation Laws with Stiff Relaxation Terms,* J. Comp. phys. Vol. 126, pp. 449-467, 1996. 107

[72] C. Jung and A. Mahalov, *Wave propagation in one-dimensional random waveguides*, Discrete and continuous dynamical systems, 28, No. 1, (2010), pp. 147–159. 88

[73] C. Jung, and T. B. Nguyen, *Semi-analytical numerical methods for convection-dominated problems with turning points,* Int. J. Numer. Anal. Mod., Vol. 10, No. 2, pp. 314-332 (2013). 111

[74] C. Jung, and T. B. Nguyen, *Semi-analytical time differencing methods for stiff problems,* J. Sci. Comput., Vol. 63, No. 2, 355-373 (2015). 107, 120

[75] C. Jung, and T. B. Nguyen, *New time differencing methods for spectral methods,* J. Sci. Comput., (2015) DOI 10.1007/s10915-015-0037-0 106, 107, 133

[76] C. Jung, and T. B. Nguyen, *A new adaptive weighted essentially non-oscillatory WENO-θ scheme for hyperbolic conservation laws,* submitted. 56

[77] C. Jung, and T. B. Nguyen, *New WENO-θ scheme and its application to the 2D Riemann problem,* The ninth IMACS Conference on *Nonlinear Evolution Equations and Wave Phenomenon: Computation and Theory*, Univ. of Geogia, 2015.

[78] G.-S. Jiang, and C.-W. Shu, *Efficient implementation of Weighted ENO schemes*, J. Comput. Phys. 126, 202-228 (1996). 39, 46, 50, 67, 73

[79] C. Jung, and R. Temam, *Asymptotic analysis for singularly perturbed convection-diffusion equations with a turning point*, J. Mathematical Physics, **48**, 065301 (2007). 111

[80] C. Jung, and R. Temam, *Finite volume approximation of one-dimensional stiff convection-diffusion equations.* J. Sci. Comput. **41**, no. 3, pp 384-410. (2009). 111

[81] S. Jin and Z. Xin, *The relaxation Schemes for Systems of Conservation Laws in Arbitrary Space Dimensions,* Comm. Pure Appl. Math. Vol. 48, no. 3, pp. 235-276, 1995. 107

[82] S. Krogstad, *Generalized integrating factors methods for stiff PDEs,* J. Comput. Phys. 203 (2005) 72-88. 10

[83] S. N. Kružkov, *First order quasilinear equations in several independent variables*, Math. USSR Sbornik 10 (1970), pp. 217-243. 10

[84] J. Kevorkian and J. D. Cole, *Multiple Scale and Singular Perturbation Methods,* Springer, 1996. 107

[85] G. E. Karniadakis and R. M. Kirby II, *Parallel Scientific Computing in C++ and MPI: a Seamless Approach to Parallel Algorithms and their Implementation*, Cambridge University Press, 2003. 100

[86] O.M.Knio and O.P.Le Maître, *Uncertainty propagation in CFD using polynomial chaos decomposition.* Fluid dynamics research, 38 (2006), 616–640. 86

[87] R. B. Kellogg and M. Stynes, *Layers and corner singularities in singularly perturbed elliptic problems*, BIT **48**(2), 309-314 (2008). 111

[88] I. G. Kevrekidis and G. Samaey, *Equation-Free Multiscale Computation: Algorithms and Applications,* Annu. Rev. Phys. Chem, Vol. 60, pp. 321-344, 2009. 107

[89] Y. Kuramoto, and T. Tsuzuki, *Persistent propagation of concentration waves in dissipative media far from thermal equilibrium,* Prog. Theor. Phys., Vol. 55, No. 2, Feb. 1976. 133

[90] A.-K. Kassam, and L. N. Trefethen, *Fourth-order time-stepping for stiff PDEs,* SIAM J. Sci. Comput. Vol. 26, No. 4, pp. 1214-1233. 106, 107, 108, 110

[91] P. Lax, *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, in *Regional Conference Series Lectures in Applied Math. Vol. 11 (SIAM, Philadelphia, 1972).* 5, 11

[92] X.-D. Liu, S. Osher, and T. Chan, *Weighted essentially non-oscillatory schemes*, J. Comput. Phys. 115, 200-212 (1994). 27, 39, 41

[93] G. Li, and J. Qiu, *Hydrid weighted essentially non-oscillatory schemes with different indicators*, J. Comput. Phys., 229 (2010), 8105-8129. 51

[94] D. Levy, G. Puppo, and G. Russo, *Central WENO schemes for hyperbolic systems of conservation laws*, Math. Model. Numer. Anal. 33, 547 (1999). 23

[95] Y.-Y. Liu, C.-W. Shu, and M.-P. Zhang, *On the positivity of linear weights in WENO approximations*, Acta Mathematicae Applicatae Sinica, English series, Vol. 25, No. 3 (2009), 503-538. 44

[96] R. J. LeVeque, *Numerical Methods for Conservation Laws*, 2nd ed., Lectures in Mathematics, ETH Zürich. 5, 18, 29, 30, 33

[97] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, 2nd ed., Cambridge texts in applied mathematics, Cambridge Univerisity Press, 2012. 29, 30, 33

[98] P. Lax, and B. Wendroff, *Systems of conservation laws*, Comm. Pure Appl. Math., XIII, 217-237 (1960). 15, 77

[99] G. Lin, X. Wan, C.-H. Su and G.E. Karniadakis, *Stochastic computational fluid mechanics.* Computing in Science and Engineering, 9(2007), pp. 21-29 86

[100] A.S. Mahalov, *Mathematical investigation of periodic acoustical waveguides of an arbitrary shape.* Journal of mathematical analysis and applications. 127 (1986), no. 2, 569–576. 85

[101] M. Marion, and R. Temam, *Nonlinear Galerkin methods,* SIAM J. Numer. Anal. Vol. 26, No. 5, pp. 1139-1157, Oct. 1989. 111

[102] M. P. Martín, E. M. Taylor, M. Wu, and V.G. Weirs, *A bandwidth-optimized WENO scheme for the effective direct numerical simulation of compressible turbulence*, J. Comput. Phys. 220 (2006), 270-289. 51, 132

[103] Natanson, *Theory of Functions of a Real Variable*, Frederick Ungan, New York, 1965. 18

[104] N. Mai-Duy, D. Pan, N. Phan-Thien, and B. C. Khoo, *Dissipative particle dynamics modeling of low Reynolds number incompressible flows,* J. Rheol. **57**, 585 (2013). 107

[105] B. Nicolaenko, B. Scheurer, and R. Temam, *Some global dynamical properties of the Kuramoto-Sivashisky equations: nonlinear stability and attractors,* Physica D, Vol. 16, pp 155-183 (1985). 133

[106] H. Nessyahu, and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys. 87, 408 (1990). 23

[107] B. Øksendal, *Stochastic Differential Equations.* Springer, New York, 2007.

[108] R. E. O'Malley, *Singularly perturbed linear two-point boundary value problems.* SIAM Rev. **50** (2008), no. 3, pp 459-482. 111

[109] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM J. Numer. Anal., Vol. 21, No. 2, 1984. 24, 27, 28

[110] S. Osher, *Convergence of generalized MUSCL schemes*, SIAM J. Numer. Anal., Vol. 22, No. 5 (1985), 947-961. 32

[111] S. Osher, and S. Charavarthy, *High resolution schemes and the entrypy condition*, SIAM J. Numer. Anal., Vol. 21, No. 5 (1984), 955-984. 22, 27

[112] S. Osher, and S. Charavarthy, *Very high order accurate TVD schemes*, ICASE Report 84-44, 1984. 23

[113] S. Osher, and E. Tadmor, *On the convergence of difference approximations to scalar conservation laws*, Mathematics of Computation, Vol. 50, No. 181, (1988) 19-51.

[114] P. S. Pacheco, *Parallel Programming with MPI.* Morgan Kaufmann Publisher, Inc., San Francisco, California. An imprint of Elsevier, 1997. 100

[115] T. H. Pulliam, *The Euler equations* Personal Notes, NASA Ames Research Center, 1994. 75

[116] W. Purkert and J. Vom Scheidt, *Stochastic eigenvalue problems for differential equations.* Reports on mathematical physics. 15 (1979), no. 2, 205–227.

[117] J. Qiu, and C.-W. Shu, *High order central schemes for hyperbolic systems of conservation laws*, J. Comput. Phys. 183, 187-209 (2002). 23

[118] P. L. Roe, *Approximate Riemann solvers, parameter vectors, and difference schemes*, J. Comput. Phys., 43., 357-372 (1981). 11, 27, 38, 71

[119] P. L. Roe, *Numerical algorithms for the linear wave equation*, Royal Aircraft Establishment Technical Report 81047, 1981. 29

[120] P. L. Roe, *Some contributions to the modelling of discontinuous flows*, Proc. AMS/SIAM Seminar, 1983. 31

[121] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods.* Springer Science+Business Media, New York, 2004. 86

[122] C.-W. Shu, *High-order finite difference and finite volume WENO schemes and discontinuous Galerkin methods for CFD*, Int. J. Comput. Fluid D., 2003, Vol. 17 (2), 107-118. 39

[123] C.-W. Shu, *TVB uniformly high-order schemes for conservation laws*, Math. Comp., Vol. 49, No. 179 (1987), 105-121. 33

[124] C.-W. Shu, *Essentially non-oscillatory and Weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, B. Cockburn, C. Johnson, C.-W. Shu, E. Tadmor, and A. Quarteroni, eds., Lecture Notes in Math. 1697, Springer-Verlag, Berlin, 1998, 325-432. 35, 36, 37, 38

[125] C.-W. Shu, *High order weighted essentially nonoscillatory schemes for convection dominated problems*, SIAM Review, Vol. 51, No. 1, 82-126. 39, 41

[126] S. I. Sivashinsky, *Nonlinear analysis of hydrodynamic instability in laminar flames, Part I. Derivation of basis equations,* Acta Astronautica **4** (1977), pp. 1177-1206. 133

[127] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, 1994. 5, 11, 19

[128] P. K. Sweby, *High resolution schemes using flux limiters for hyperbolic conservation laws*, SIAM J. Numer. Anal., Vol. 21, No. 5, 1984. 27, 29, 30, 33

[129] P. K. Sweby, and M. J. Baines, *Convergence of Roe's scheme for the general non-linear scalar wave equation*, Reading University Numerical Analysis Report, 1981. 31

[130] C. W. Schulz-Rinne, J. P. Collins, and H. M. Glaz, *Numerical solution of the Riemann problem for two-dimensional gas dynamics*, SIAM J. Sci. Comput., Vol. 14, No. 6, pp. 1394-1414, 1993. 77

[131] J. Shi, C. Hu, and C.-W. Shu, *A technique of treating negative weights in WENO schemes*, J. Comput. Phys. 175, 108-127 (2002). 44

[132] S. Shih and R. B. Kellogg, *Asymptotic analysis of a singular perturbation problem.* SIAM J. Math. Anal. **18** (1987), pp . 1467-1511. 110

[133] S. I. Sivashinsky, and D. M. Michelson, *On irregular wavy flow of a liquid film dowm a vertical plane,* Prog. Theor. Phys. **63** (1980), 2112-2114. 133

[134] M. Stynes, *Steady-state convection-diffusion problems*, Acta Numer., **14** (2005), 445-508. 111

[135] Y. Shen, and G. Zha, *A robust seventh-order WENO scheme and its applications*, AIAA paper 2008-0757. 39

[136] E. Tadmor, *Convenient total variation diminishing conditions for nonlinear difference schemes*, SIAM J. Numer. Anal., Vol. 25, No. 5, 1988.

[137] E. Tadmor, *Approximate solutions of nonlinear conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, B. Cockburn, C. Johnson, C.-W. Shu, E. Tadmor, and A. Quarteroni, eds., Lecture Notes in Math. 1697, Springer-Verlag, Berlin, 1998, 1-149. 5, 10, 11

[138] E. F. Toro, *Riemann solvers and numerical methods for fluid dynamics, A practical introduction*, 3rd ed., Springer 2009. 26, 29, 33, 38, 70, 71, 75, 79

[139] L. N. Trefethen, *Spectral methods in MATLAB,* Soc. for Industr. & Appl. Math., Philadelphia, 2000. 105

[140] A. Taflove and S. Hagness, *Computational Electromagnetics: the Finite-Difference Time-Domain Method.* 3rd ed., Norwood, MA: Artech House, 2005. 85, 88

[141] M. Tao, H. Owhadi and J. E. Marsden, *Nonintrusive and structure preserving multiscale integration of stiff ODEs, SDEs, and Hamiltonian systems with hidden slow dynamics via flow averaging,* Multiscale Model. Simul., Vol. 8, No.4, pp. 1269-1324, 2010. 107

[142] E. M. Taylor, M. Wu, and M. P. Martín, *Optimization of nonlinear error for weighted essentially non-oscillatory methods in direct numerical simulations of compressible turbulence*, J. Comput. Phys., 223 (2007), 384-397. 55, 61, 62

[143] D. M. Tartakovsky and D. Xiu, *Numerical methods for differential equations in random domains*, SIAM J. Sci. Comput., vol. 28, no. 3 (2006), pp. 1167-1185. 85, 132

[144] B. Van Leer, *Towards the ultimate consevative difference scheme, II. Monotonicity and conservation combined in a second order scheme*, J. Comput. Phys. 14 (1974), 361-370. 30, 31, 32

[145] B. Van Leer, *Towards the ultimate consevative difference scheme, IV. A new approach to numerical convection*, J. Comput. Phys. 23 (1977), 276-299. 32

[146] B. Van Leer, *Towards the ultimate consevative difference scheme, V. A second order sequel to Godonov's method*, J. Comput. Phys. 32 (1979), 101-136. 32

[147] B. Van Leer, *On the relation between the upwind-difference schemes of Godunov, Engquist-Osher and Roe*, SIAM J. Sci. Statist. Comput. 5 (1984), 1-20. 23, 27, 32

[148] C.V. Verhoosel, M.A. Gutierrez and S.J. Hulshoff, *Iterative solution of the random eigenvalue problem with application to spectral stochastic finite element systems.* Int. J. for numer. Meth. Engng. 68 (2006), 401–424.

[149] M. I. Vishik and L. A. Lyusternik, *Regular degeneration and boundary layer for linear differential equations with small parameter*, Usp. Mat. Nauk **12**, 3.122 (1957).

[150] W. Wasow, *Linear Turning Point Theory*, Spinger, New York, 1985. 110

[151] P. Woodward, and P. Colella, *The numerical simulation of two-dimensional fluid flow with strong shocks*, J. Comput. Phys. 54, 115-173 (1984). 37, 74, 79

[152] D. Xiu, *Fast numerical methods for stochastic computations: A review.* Commun. Comput. Phys. 5 (2008), 242–272. 86

[153] D. Xiu and J. S. Hesthaven, *High-order collocation methods for differential equations with random inputs.* SIAM J. Sci. Comput. 27(2005), pp 1118-1139. 86, 132

[154] D. Xiu and G. E. Karniadakis, *The Wiener-Askey polynomail chaos for stochastic differential equations.* SIAM J. Sci. Comput. 24(2002), pp 619-644. 86

[155] D. Xiu and G. E. Karniadakis, *Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos.* Comput. Methods Appl. Mech. Engrg. 199(2002), pp 4927-4948. 86

[156] K. Yee, *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media.* IEEE Transactions on Antennas and Propagation, 14, No. 5 (1966), pp 302-307. 88

[157] N. K. Yamaleev, and M. H. Carpenter, *A systematic methodology for constructing high-order energy stable WENO schemes*, J. Comput. Phys. 228 (2009), 4248-4272.

51, 53, 62

ACKNOWLEDGEMENTS

First of all, I would like to deeply express my gratitude to my advisor, Professor Chang-Yeol Jung. Professor Jung is not only an advisor but also a mentor who introduced me to the very interesting research topics of WENO schemes, singular perturbation techniques, and uncertainty quantification, and instructed me how to struggle and overcome difficulties occurring during the research process. I have learned and benefited greatly from his insightful knowledge, working ethic and scientific methods in conducting research. I would not have been able to achieve this far without his constant guidance and support.

I would like to show my sincere appreciation to Professor Pilwon Kim, Professor Bongsuk Kwon, Professor Bongsoo Jang, and Professor Junseok Kim for their valuable advice and suggestions during the preparation of this dissertation.

Hereby I would also like to give my thanks to my labmates and all of my friends, especially to Soyeong Jeong, Hyojung Lee, Junyoung Lee, Insu Cho, Kyunghoon Kim, Kyung Duk Park, Junho Choi, Gnidakouong Ngouanom Joel Renaud, Mayzonee Virtudazo Ligaray, Ho Duc Tam, Luong Minh Bau, and all Byeongyeoung and Beomseo church members for their friendship and encouragement during my PhD program.

Last but not least, I would like to thank my parents, as well as my brothers and sisters. Their constant love, care, and advice give me strength to keep on moving forward.