STRUCTURE-OPTIMIZED AND ORDER-PRESERVING METHODS FOR ADVECTION AND TIME STEPPING

A Dissertation
Submitted to
the Temple University Graduate Board

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

by
Abhijit Biswas
May, 2021

Examining Committee Members:

Prof. Benjamin Seibold, Mathematics (Advisory Chair)
Prof. Gillian Queisser, Mathematics
Prof. Isaac Klapper, Mathematics
Prof. David Shirokoff, Mathematics, NJIT (External Reader)
ABSTRACT

STRUCTURE-OPTIMIZED AND ORDER-PRESERVING METHODS
FOR ADVECTION AND TIME STEPPING

Abhijit Biswas
DOCTOR OF PHILOSOPHY

Temple University, May, 2021

Professor Benjamin Seibold, Chair

This dissertation addresses some of the existing fundamental shortcomings in computational methods for differential equations by developing novel high-order accurate numerical schemes that can apply to multi-physics or multi-scale problems. Motivated by the need for numerical methods with long-time accurate solutions in complex advection-dominated problems, we first address whether there are numerical schemes for constant-coefficient advection problems that can yield convergent solutions for an infinite time horizon. After establishing a new notion of convergence in an infinite time limit of numerical methods, we first show that linear methods on a fixed-grid cannot meet this convergence criterion. Also, numerical studies demonstrate that some widely used traditional nonlinear methods fail to produce accurate long-time solutions. Then we present a new numerical methodology based on a nonlinear jet scheme framework. We show that these methods satisfy the new convergence criterion, establishing that such numerical methods exist that converge on an infinite time horizon and demonstrate the long-time accuracy gains incurred by this property. Secondly, we develop finite volume methods with optimal limiter functions based on the class of WENO type methods. Limiter functions are applied to finite volume methods to avoid oscillations near discontinuities and sharp transitions for hyperbolic equations. Many limiter functions have been introduced in the literature. The general approach is first
to devise a limiter function and then demonstrate that it performs well on some test problems. Here, we solve the inverse problem instead: given a portfolio of representative test cases, and a cost functional, determine the optimal limiter function that is then designed to do well. Lastly, we address the question of whether diagonally implicit Runge-Kutta (DIRK) schemes with weak stage order (WSO) four or higher and order at least four exist. The concept of WSO, which manifests in algebraic conditions on the Runge-Kutta Butcher tableau, can relax the stage order conditions so that it becomes compatible with the DIRK structure. Thus, DIRK schemes with high WSO can rectify order reduction. DIRK schemes up to order four with WSO up to three have already been constructed, based on a simplified theory. In this part of the dissertation, we develop a general theory of WSO and a methodology that can yield stiffly accurate, $A$-stable DIRK schemes up to order five with WSO up to five.
ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank my advisor, Prof. Benjamin Seibold, for his guidance, assistance, and patience over the last several years. I also want to express my most profound appreciation for his support outside of my research. His suggestions on my presentation skills, communicating mathematics, and interacting with other academics were invaluable. I feel very fortunate to have a chance to work with such a great mentor and mathematician. Without his assistance, this dissertation would not have been possible.

I want to acknowledge Prof. David Ketcheson (KAUST) and Prof. David Shirokoff (NJIT) for their wonderful collaboration during my dissertation research. Their expertise on time-stepping methods was essential for advancing my research. I would particularly like to single out Prof. David Shirokoff for sharing his ideas on weak stage order theory, which have helped me advance my research. Prof. Gillian Queisser, with whom I worked on a modeling project, deserves special recognition. I have learned a lot from our work and discussion together over the past years, and it is my great pleasure working with these excellent mathematicians.

I would like to express my gratitude to my committee members, Prof. Gillian Queisser, Prof. Issac Klapper, and Prof. David Shirokoff, for their comments and suggestions during this thesis’ writing.

I want to thank my fellow applied mathematics students and postdocs who provided many stimulating discussions and happy distractions for resting my mind outside of my research. We have shared many inspiring conversations and grown together as scholars and researchers. I would never forget them.

Lastly, I want to express my deepest gratitude to my family for their love, support, and sacrifices. They gave me moral support, encouragement, and motivation to accomplish my goal. This journey would not have been possible without their love and support.
To my mother and father,
TABLE OF CONTENTS

ABSTRACT iv
ACKNOWLEDGEMENT vi
DEDICATION vii
LIST OF FIGURES xi
LIST OF TABLES xvii
1 INTRODUCTION 1
2 LONG TIME NUMERICAL SOLUTIONS FOR ADVECTION PROBLEMS 8
  2.1 Notion of Convergence and Formulation of the Problem . . . . . . 10
  2.2 Long Time Solution Behavior of Linear Methods . . . . . . 12
    2.2.1 Illustrative examples . . . . . . . . . . . . . . . . . . 14
    2.2.2 Qualitative analysis . . . . . . . . . . . . . . . . . . . 18
    2.2.3 Quantitative scaling . . . . . . . . . . . . . . . . . . 25
  2.3 Long Time Solution Behavior of Traditional Nonlinear Methods 27
  2.4 Jet Scheme with Nonlinear Interpolation . . . . . . . . . . . . 33
    2.4.1 Approximation of the initial condition . . . . . . . . . 35
    2.4.2 Quantifying approximation errors and convergence . . . 36
    2.4.3 Two equivalent interpretations of jet schemes . . . . . 38
    2.4.4 Jet scheme between discrete spaces: Fixed points as vec-
          tors . . . . . . . . . . . . . . . . . . . . . . . . . . . 39
    2.4.5 Jet scheme between function spaces: Fixed points as
          functions . . . . . . . . . . . . . . . . . . . . . . . . . 45
    2.4.6 Numerical analysis for nonlinear jet scheme . . . . . . 46
    2.4.7 Other initial condition approximation: good news and
          some caveats . . . . . . . . . . . . . . . . . . . . . . . . 49
  2.5 Beyond Constant-Coefficient Advection . . . . . . . . . . . 52
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5.1</td>
<td>Variable coefficient linear advection problems</td>
<td>52</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Linear advection problems in multiple space dimensions</td>
<td>53</td>
</tr>
<tr>
<td>2.6</td>
<td>Conclusions and Outlook</td>
<td>56</td>
</tr>
<tr>
<td>3</td>
<td>OPTIMAL WENO TYPE FINITE VOLUME LIMITER FUNCTIONS</td>
<td>59</td>
</tr>
<tr>
<td>3.1</td>
<td>Limiter Functions in 1D</td>
<td>62</td>
</tr>
<tr>
<td>3.2</td>
<td>3rd Order FV Methods with Limiters in the 2D Slope Domain</td>
<td>67</td>
</tr>
<tr>
<td>3.3</td>
<td>Limiter Functions Induced by third-Order WENO Type Methods</td>
<td>69</td>
</tr>
<tr>
<td>3.3.1</td>
<td>WENO3 limiter</td>
<td>69</td>
</tr>
<tr>
<td>3.3.2</td>
<td>CWENO3 and CWENO3-Z limiters</td>
<td>71</td>
</tr>
<tr>
<td>3.4</td>
<td>Learning Optimal Parameters by an Inverse Problem</td>
<td>76</td>
</tr>
<tr>
<td>3.5</td>
<td>Numerical Results</td>
<td>81</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Linear advection equation with smooth initial condition</td>
<td>81</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Linear advection equation with discontinuous initial condition</td>
<td>82</td>
</tr>
<tr>
<td>3.5.3</td>
<td>Inviscid Burgers’ equation</td>
<td>84</td>
</tr>
<tr>
<td>3.5.4</td>
<td>Sod shock tube problem</td>
<td>85</td>
</tr>
<tr>
<td>3.5.5</td>
<td>Shu-Osher problem</td>
<td>88</td>
</tr>
<tr>
<td>3.6</td>
<td>Conclusion: Challenges and Future Work</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>DIRK SCHEMES WITH WEAK STAGE ORDER BEYOND THREE</td>
<td>92</td>
</tr>
<tr>
<td>4.1</td>
<td>Stiff Problem and Stability</td>
<td>93</td>
</tr>
<tr>
<td>4.2</td>
<td>Order Reduction and Its Remedy</td>
<td>98</td>
</tr>
<tr>
<td>4.3</td>
<td>Weak Stage Order</td>
<td>105</td>
</tr>
<tr>
<td>4.4</td>
<td>WSO Theory</td>
<td>108</td>
</tr>
<tr>
<td>4.5</td>
<td>DIRK Schemes with WSO Beyond Three</td>
<td>115</td>
</tr>
<tr>
<td>4.6</td>
<td>Numerical Results for Problems Covered by the WSO Theory</td>
<td>121</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Linear ODE test problem</td>
<td>122</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Heat equation</td>
<td>123</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Schrödinger equation</td>
<td>124</td>
</tr>
<tr>
<td>4.6.4</td>
<td>Advection-diffusion equation</td>
<td>125</td>
</tr>
<tr>
<td>4.6.5</td>
<td>Linear advection equation</td>
<td>126</td>
</tr>
<tr>
<td>4.6.6</td>
<td>Airy’s equation</td>
<td>127</td>
</tr>
<tr>
<td>4.6.7</td>
<td>Heat equation with spatially varying coefficient</td>
<td>127</td>
</tr>
<tr>
<td>4.7</td>
<td>Numerical Results for Problems NOT Covered by the WSO Theory</td>
<td>128</td>
</tr>
<tr>
<td>4.7.1</td>
<td>Stiff nonlinear ODE: Van der Pol’s equation</td>
<td>129</td>
</tr>
<tr>
<td>4.7.2</td>
<td>Heat equation with slowly and rapidly varying time dependent coefficient</td>
<td>130</td>
</tr>
<tr>
<td>4.7.3</td>
<td>Stiff nonlinear PDE: viscous Burgers’ equation</td>
<td>130</td>
</tr>
</tbody>
</table>
5 CONCLUSION AND OUTLOOK

REFERENCES
LIST OF FIGURES

2.1 Upwind method for $u_t + u_x = 0$ with $\mu = 0.8$ (CFL) and $m = \text{100 grid points}$: Amplification factor $g(\xi) = (1 - \mu) + \mu e^{-i\xi h}$, and eigenvalues (left) and numerical solution at different times (right). All the eigenvalues except 1 are clearly inside the unit circle, resulting in decay towards a constant. .......................... 15

2.2 SSP-RK3 in time and central in space method for $u_t + u_x = 0$ with $\mu = 0.8$ and $m = \text{100 grid points}$: Amplification factor $g(\xi) = \left(1 - \frac{\mu^2}{4}\right) + \frac{\mu^2}{4} \cos 2\xi h - i \left[\frac{\mu^3}{24} \sin 3\xi h + \left(\mu - \frac{\mu^3}{8}\right) \sin \xi h\right]$, and eigenvalues (left) and numerical solution at different times (right). Despite high order, all the eigenvalues except 1 are inside the unit circle, resulting in decay towards a constant. ....... 17

2.3 Crank-Nicolson in time and fourth-order in space method for $u_t + u_x = 0$ with $\mu = 0.8$ (CFL) and $m = \text{100 grid points}$: Amplification factor $g(\xi) = \frac{1 - i(\frac{2\mu}{3} \sin \xi h - \frac{2}{3} \sin 2\xi h)}{1 + i(\frac{2\mu}{3} \sin \xi h - \frac{2}{3} \sin 2\xi h)}$, and eigenvalues (left) and numerical solution at different times (right). All eigenvalues are on the unit circle, hence no modes decay; however, they move at wrong speeds. ......................... 18

2.4 Geometry for bound on the estimate. ................................. 22

2.5 Approximation error $\varepsilon_{h,t}$ as function of $h$ and $t_f$. Left: For the second order linear Lax-Wendroff scheme, and Right: For the nonlinear jet scheme. .................................................. 26

2.6 Long-time evolution of numerical solutions for different non-linear methods (with $m = 100$ grid points): WENO5, Lax-Wendroff (LW) with van Leer limiter, and Lax-Wendroff with Superbee limiter. The bottom right show the temporal evolution of the maximum for long times. ......................... 32

2.7 Genuine interpolant. .................................................. 34

2.8 Fallback case. .................................................. 34
2.9 Domain with three cells. Black dots denote the grid points; blue dots denote the foot points; red dots are the sub-cell boundaries; green denotes the kink positions that are $\delta \ll h$ distance away from the grid points. The black arrow shows the next relative position of kink.

2.10 Example of long-time evolution of nonlinear jet scheme with $(\phi, \psi)$ initial condition approximation, for $u_t + u_x = 0$ with $\mu = 0.9$ on $m = 100$ grid points, in comparison with three classical nonlinear methods (‘LW’ stands for Lax-Wendroff). While the classical methods yield deformed long-time solutions, the jet scheme assumes (after finitely many steps) a fix point (i.e., an exact solution) that is a good approximation of the true solution. The bottom right panel shows the temporal evolution of the maximum, and it confirms the jet scheme’s superiority over the existing methods in capturing long-time evolution.

2.11 Eigenvector of the one-step matrix $M$ corresponding to an eigenvalue with modulus less than 1. Left: initial condition. Right: solution at $t_f = 5$, which is exactly same as the initial profile, but multiplied by $10^{-6}$.

2.12 Long-time evolution of nonlinear jet scheme for variable coefficient advection problem. The numerical approximation is of high quality until $t = 1000$, but degrades eventually for much larger times.

2.13 Nonlinear interpolant in 2D.

2.14 Numerical approximation with nonlinear jet scheme on 2D advection problem, for the initial condition $u_0(x, y) = \sin(2\pi x) + \sin(\pi y)$.

2.15 Numerical approximation with nonlinear jet scheme on 2D advection problem, for the initial condition $u_0(x, y) = \sin(2\pi x) \sin(\pi y)$.

3.1 The light shaded regions show the TVD region. The deep shaded regions show the Sweby region of 2nd order methods. All the 2nd order linear schemes leave the TVD region and hence produce spurious oscillations. The right panel shows all the limiters in equation (3.11) lie inside the Sweby region of 2nd order methods.

3.2 The light shaded regions show the TVD region. The deep shaded regions show the Sweby region of 2nd order methods. $\phi_3$: Full 3rd order limiter function; $\phi_{AS}$: limiter function given in (3.13); and $\phi_{CT}$: limiter function given in (3.14). The figure shows that these limiter functions do not lie in the TVD regions entirely.
3.3 Combined limiter function in equation (3.23) .................................................. 68
3.4 Each surface shows the plot of \((H_{CWENO3} - H_3)\) function for different choices of reasonable parameter values, i.e., the third-order limiter function is subtracted off from the \(H_{CWENO3}\) limiter function. So the \(H_{CWENO3}\) limiter function is third-order wherever the values are zeros in the plots. ................................................................. 73
3.5 Each surface shows the plot of \((H_{CWENO3-Z} - H_3)\) function for different choices of reasonable parameter values, i.e., the third-order limiter function is subtracted off from the \(H_{CWENO3-Z}\) limiter function. So the \(H_{CWENO3-Z}\) limiter function is third-order wherever the values are zeros in the plots. ................................................................. 75
3.6 Initial function with smooth and non-smooth parts having a few sharp transitions. .................................................................................................................. 77
3.7 Left: Cost functional corresponding to \(H_{CWENO3}\) limiter. Right: Cost functional corresponding to \(H_{CWENO3-Z}\) limiter. 1st half \(L_1\) error and the 2nd half \(L_1\) error refer to the first and the second integral, respectively, in the cost functional (3.44). Different errors in the cost functional are calculated and plotted separately for different choices of parameters. Each black dot represents the optimal parameters for each of these energy surfaces. ................................................................. 79
3.8 Left: Numerical solutions of the linear advection equation with CFL number \(\nu = 0.8\), grid points \(n = 170\), and smooth initial condition in equation (3.46) by different numerical methods on \([0.2, 0.8]\). Right: Zoomed-in view near the maximum of the profile. .................................................................................................................. 82
3.9 The figure illustrates a closeup view of numerical solutions of the linear advection equation with a discontinuous initial condition in equation (3.47) by different numerical methods in two different parts of the domain. ................................................................ 83
3.10 Approximate solutions (with zoomed-in view) of Burgers’ equation with a smooth initial condition, CFL number \(\nu = 0.8\), and \(n = 100\) grid points by different methods at time \(t = 0.5\). ................................................................. 84
3.11 Left: Density profile for the Sod’s problem as computed by different methods at time \(t = 0.8\) with CFL number \(\nu = 0.95\) and \(n = 100\) grid points. Right: Zoomed-in region at the contact discontinuity. .................................................................................................................. 86
3.12 Left: Velocity profile for the Sod’s problem as computed by different methods. Right: Zoomed-in region near \(x = 1.5\). .................................................................................................................. 87
3.13 Left: Pressure profile for the Sod’s problem as computed by different methods. Right: Zoomed-in region near \(x = -1\). .................................................................................................................. 87
3.14 Solution of the Mach 3 shock-density wave interaction by different methods on $[-4.5, 4.5]$ with $n = 640$ grid points, CFL number $\nu = 0.95$ at the final time $t = 1.8$.  

3.15 Enlarged view of the density profile near a sharp transition region on a high resolution with $n = 1280$ mesh points.  

3.16 A closeup view of the oscillatory area of the density profile on a high resolution with $n = 1280$ grid points.  

4.1 Comparing the Euler methods. The problem (4.5) is solved for $\lambda = -2$ with $m = 5$ (left panels) and $m = 10$ (right panels) grid points, comparing explicit Euler (top) to implicit Euler (bottom). The exact solution is given by black curve, and the red curve shows the numerical solution.  

4.2 The test problem (4.5) is solved under the same assumptions before except for $\lambda = -20$. Black and red curves show the true and numerical solutions, respectively. The explicit method shows numerical instability (top left), as indicated by growing oscillations. This is due to the sharp transition of the neighboring solutions (as $\lambda = -20$). This is typical for stiff problems.  

4.3 Stability regions of explicit and implicit Euler methods. Left panel shows that the explicit Euler method is stable when $\lambda \Delta t$ is inside the unit circle centered at $-1$. The right panels show that the stability region of the implicit Euler method contains the stability region of the ODE problem (4.8), the entire left half plane.  

4.4 Error convergence for the Prothero-Robinson test problem (4.5) by a 2nd order DIRK scheme (left), and a 3rd order DIRK scheme (right).  

4.5 Left: Scaled BL profiles via $(x/\Delta t^{0.5}, e/\Delta t^{2.0})$ shows that BL thickness $= \mathcal{O}(\sqrt{\Delta t})$ and BL height $= \mathcal{O}(\Delta t^2)$. Middle: Global spatial error. Right: The error convergence for the heat equation (4.9) by a 3rd order DIRK scheme. The middle panel shows the global errors as functions of $x$ with three $\Delta t$ choices. The error convergence for function values and derivatives ($u$ blue circles; $u_x$ red squares) are shown in the right panel.  

4.6 Stability regions for DIRK-$(7, 4, 4)$ (left), DIRK-$(13, 5, 4)$ (middle), and DIRK-$(13, 5, 5)$ (right) schemes.  

4.7 Left: $E$-polynomials for different DIRK methods are plotted with $y$-axis in log scale. Right: $|R(\imath y)|$ for $y \in \mathbb{R}$ for different DIRK schemes, and $|R(\imath y)| \leq 1$ along the $y$-axis, hence these methods are $A$-stable.
4.8 Error convergence for the Prothero-Robinson test problem using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (blue circles) and WSO 1 (red squares) (left), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (blue circles) and WSO 1 (red squares) (right). 122

4.9 Error convergence \((u\text{ blue circles; } u_x \text{ red squares})\) for heat equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right). 123

4.10 Error convergence \((u\text{ blue circles; } u_x \text{ red squares})\) for Schrödinger equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right). 124

4.11 Error convergence \((u\text{ blue circles; } u_x \text{ red squares})\) for the advection-diffusion equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right). 125

4.12 Error convergence \((u\text{ blue circles; } u_x \text{ red squares})\) for the linear advection equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right). 126

4.13 Error convergence \((u\text{ blue; } u_x \text{ red})\) for the Airy’s equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right). 127

4.14 Error convergence \((u\text{ blue circles; } u_x \text{ red squares})\) for heat equation with spatially varying diffusion coefficient \(k(x) = \cos(x + 0.1)\) using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right). 128

4.15 Error convergence for the Van Der Pol oscillator using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (blue circles) (left), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (blue circles) (right). 130
4.16 Error convergence ($u$ blue circles; $u_x$ red squares) for heat equation with diffusion coefficient $k(x, t) = \cos(0.1t + 0.2)$ (top) and $k(x, t) = 1 + 0.5 \cos(20t)$ (bottom) using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right). . . . . . . . . . . . . 131

4.17 Error convergence for the viscous Burgers’ equation using DIRK-(7, 4, 4), DIRK-(13, 5, 4), and DIRK-(13, 5, 5). . . . . . . . . . 132
LIST OF TABLES

3.1 Different pairs of optimal parameters ($\epsilon, \sigma$) determined by our study. .......................................................... 79

4.1 Butcher’s tableaus for the Euler’s methods ................................................. 94
4.2 Butcher’s tableau for a 2nd order method with $\gamma = 1 - \frac{1}{\sqrt{2}}$. .... 98
4.3 Butcher’s tableau for a 3rd order method. ................................................. 99
4.4 Butcher’s tableau for DIRK-(7, 4, 4). ......................................................... 134
4.5 Butcher’s tableau for DIRK-(13, 5, 4). ......................................................... 134
4.6 Butcher’s tableau for DIRK-(13, 5, 5). ......................................................... 134
CHAPTER 1

INTRODUCTION

Differential equations appear everywhere in our lives. Mathematical models based on differential equations are used to analyze and understand various real-world problems around us. Oftentimes, these model equations are complex and hard to solve analytically. Numerical methods are applied to the governing equations to calculate approximate solutions. Hence, the need for accurate, robust, and efficient numerical methods for solving ODE and PDE problems continues to grow. Despite the advances made in numerical methods for differential equations, there are still many fundamental shortcomings in existing numerical methods, and approaches that overcome (some of) these shortcomings may significantly improve the performance of simulations.

One of the fundamental problems we specifically focus on is accurately capturing the long-time numerical solution of advection problems. Advection or transport phenomena appear as sub-problems in more complicated situations in multi-physics problems. The accurate numerical resolution of transport processes is a crucial ingredient in many computational mathematics problems. We focus on linear transport under a velocity field that is given or arises from other components in a multi-physics computation, such as: (i) the evolution of structures and/or interfaces defined by a level set function $\phi(\vec{x}, t)$ [62, 77, 65, 61, 60]; (ii) the passive convection of density fields or marker distributions under a flow field [88]; (iii) the free-streaming step in particle
transport computations [16, 44, 58, 10] solved via fractional steps. Furthermore, while some applications, such as particle methods [25, 49, 59, 27], allow for Lagrangian approximations, we focus on the accurate capture of advection via Eulerian methods, i.e., fixed-grid approaches, possibly with adaptive mesh refinement (AMR) [9, 56, 84, 51, 57].

Powerful computers combined with efficient numerical algorithms enable one to resolve and evolve the motion of small structures (such as droplets in a multi-phase fluid flow simulation, represented by a level set function) over millions of time steps. And with most existing Eulerian methods, even if the structure itself remains unchanged (the droplet is merely moved by a locally constant velocity field), the numerical scheme produces a spurious slow change of shape of the solution over time. In some applications, overlapping grids [34] can mitigate the problem; but these approaches are not always possible to use. Hence, We ask the question: are there numerical schemes for advection problems, which can produce accurate solutions for a long time or potentially in an infinite time horizon? To investigate this key question, we start at a fundamental, conceptual level, by considering the constant-coefficient linear advection equation in one space dimension. We first define a new notion of convergence of numerical solutions for constant-coefficient linear advection equation in an infinite time horizon. To study the infinite time limit solutions of numerical schemes, we examine the existing methods and provide a generic proof to show that traditional linear fixed-grid methods fail to meet infinite time limit convergence criteria. Given the structural limitations of linear methods, we turn our attention to nonlinear methods. Based on a systematic computational study we conclude that many existing nonlinear numerical methods lack the ability to produce accurate numerical solutions over many time steps. Then we propose a new class of numerical methods based on the jet scheme methodology [60, 76, 17], equipped with a special nonlinear Hermite interpolant, and show that numerical schemes do exist that possess convergent infinite time limit solutions.

Of course, the infinite time limit solution does not make sense for every
PDE problem. It is the specific nature of (nearly) constant-coefficient trans-
port that renders this limit of particular relevance. At the same time, given
the importance of transport in many multi-physics problems, it is justified to
consider a numerical analysis concept specific to it. For other advection type
problems or hyperbolic conservation laws, we can alternatively ask, can one
design numerical methods that can provide highly accurate solutions for a fixed
final time? Among the existing numerical approaches, finite volume methods
with limiter functions are among the most popular numerical methods to ac-
curately simulate many computational fluid dynamics problems. Many limiter
functions have been proposed in the literature, yet there are open questions,
for example, how to devise optimal limiter functions?

The problem of finding the optimal limiter functions is very general. We
propose to solve this problem by inverting the traditional paradigm. Namely,
considering a combination of well-represented test cases and a cost functional,
we learn the limiter function by optimizing the cost functional over a set of
limiters, which are generally chosen to be continuous. The sense of optimal
limiter functions depends on the user-specifications on what really matters in
the problem at hand, and these need to be incorporated in the cost functional.
The choice of a feasible set of limiter functions is also a crucial component
of this problem. One way to construct a feasible set is to characterize the
limiter functions using parameters in a finite-dimensional space. We can allow
as many degrees of freedom as we want, but more degrees of freedom in the
limiter function makes the problem harder to solve for the optimal limiters. As
a first toward this study, we consider the 3rd order WENO [38, 78, 79] type
methods, which can be reformulated as finite volume methods with limiter
functions. Our work focuses on this particular class of 3rd order WENO type
limiters that can be viewed in the 2D slope domain [73]. Recent studies show
that the interpretation of third-order limiter functions in the 2D slope domain
provides a new perspective to understand limiters’ mechanism [73]. Using this
framework, two new limiter functions are devised, which are motivated by the
3rd order central WENO and the WENO-Z methods [6, 15, 26, 93, 94]. These
new limiters contain two unknown parameters. Traditionally, these parameters’ values are chosen based on some mathematical principles, intuition, or in an ad-hoc fashion. We present a new approach to determine these parameters by considering an inverse problem. Considering a single but generic test problem and an objective function that takes different structure-preserving features of the solution, we learn the parameters by optimizing the objective function over a class of WENO type limiters that are characterized by two parameters. The limiter functions with the optimal parameters are then tested on different test problems and compared with other limiters. These limiters turn out to be the best among the WENO type limiters and comparable with other third order limiter functions in the literature.

Another problem that we consider is high-order time integration, particularly for multi-scale time-evolution problems involving differential operators. Time accuracy is an essential component to accurately simulate ODEs or PDEs that appear in many scientific problems in real life. Examples of ODE models include chemical reaction systems, electric circuits, neuron models, and many more. The governing equations in most of the examples above are stiff, i.e., the solution of these systems consists of rapidly changing parts and slowly varying parts. These systems require special attention to simulate accurately. One of the most important fields where PDE models appear is fluid dynamics. Many important fluid dynamics phenomena, for example, aerodynamics, turbulent flows, transitional flows, and other multi-phase fluid flows are intrinsically unsteady and governed by nonlinear time-dependent PDEs. The most common way to solve partial differential equations is to discretize the spatial derivatives and then apply a time integrator. This process generates a coupled system of nonlinear ODEs. These problems are required to solve at a high level of accuracy due to the complex nature of flows. It is essential to have efficient time integration methods to capture the underlying phenomena accurately, specifically time integrators that can preserve the convergence order in so-called stiff regimes.

Stiff problems need to be handled carefully as Runge-Kutta (RK) methods
applied to solve these problems often experience order reduction [91], i.e., the solution may fail to exhibit the method’s actual order of convergence. Among the family of RK methods, linear multi-step methods (LMMs) [8, 23, 36, 89] do not suffer from order reduction [32, 52], but they generally have stability issues [91], specifically at higher order to solve stiff problems. Several approaches via modified boundary conditions for explicit RK methods have been studied in [1, 3, 4, 14, 64] to overcome order reduction in different problems. However, stiff problems are generally solved by implicit Runge-Kutta methods (IRK). RK methods with high stage orders [91] can cure the order reduction phenomenon, but they are expensive when high accuracy is required. Our particular interest is the diagonally implicit Runge-Kutta (DIRK) methods [40] that are possibly the most widely used IRK time integrators for solving stiff problems due to the low cost and the ease of implementation. They can achieve high order by solving backward-type Euler methods at each stage. We mainly focus on developing time integration methods to solve stiff ODEs and PDE initial boundary value problems (IBVPs) by diagonally implicit Runge-Kutta methods. DIRK schemes with invertible matrix structure are limited to low stage order [42]. So they suffer from order reduction. The order reduction in ODEs can be explained in terms of stiff limits [67, 91]. The order reduction in PDE IBVPs depends on the geometric structures of the spatial error [70]. A geometric explanation of spatial manifestation of order reduction can be found in [70]. The order reduction in PDE IBVPs was first pointed out in [22, 24]. To solve this problem, several efforts have been made in [53, 54, 72, 90]. Error analysis for RK methods applied to linear PDEs have been rigorously studied in [5, 30, 63, 74]. In particular, a criterion that is weaker than stage order based on ROW methods applied to linear problems was derived in [74]. This criterion, if satisfied, can diminish order reduction. Similar conditions were proposed in [63], but no corresponding numerical schemes were provided satisfying those conditions. Based on the understanding from literature, the concept of weak stage order (WSO) that is a generalization of the conditions stated in [68], was proposed in [70]. WSO conditions are some algebraic con-
ditions on the Runge-Kutta Butcher tableau that can relax the stage order conditions. WSO conditions are compatible with DIRK structure, and DIRK schemes with high weak stage order (WSO) remedy the order reduction phenomenon for a certain class of problems. DIRK schemes up to order 4 with WSO up to 3 have already been constructed based on the WSO eigenvector criterion [42]. It is also shown that the WSO eigenvector criterion is limited to WSO 3. We study the question of whether diagonally implicit Runge-Kutta (DIRK) schemes of order and weak stage order (WSO) 4 and higher exist, and how to effectively construct such methods. We employ an unpublished work called the general WSO theory developed by Professor David Shirokoff and construct DIRK schemes with weak stage orders four or higher. The order reduction phenomenon in RK methods applied to stiff problems can be addressed via these newly developed DIRK schemes with high weak stage order.

This dissertation is structured as follows. In Chapter 2, the long-time numerical solution for advection problems is presented. In §2.1, we define the new notion of convergence and formulate the mathematical problem. In §2.2, we show that the fixed-grid linear methods cannot achieve convergent infinite time limit solutions. There, §2.3 ends with an observation that widely used traditional nonlinear schemes do not produce accurate solutions for a long time. Then in §2.4 we present the new nonlinear jet scheme and establish the numerical analysis to prove that the method yields convergent infinite time limit solutions. Moreover, it is demonstrated how the new method produces superior long-time accuracy relative to classical nonlinear schemes. We demonstrate numerical results by new nonlinear jet scheme for more general advection problems in §2.5, and §2.6 provides an outlook and conclusions.

Chapter 3 presents the optimal WENO type finite volume limiter functions. We start this chapter by reviewing the limiter functions in 1D in §3.1 and the interpretation of 3rd order limiters in the 2D slope domain in §3.2. The new 3rd order WENO type limiters are introduced in §3.3. We determine the optimal limiters in §3.4. The optimal limiters’ performance on various test problems is shown in §3.5, which is followed by the conclusions and future
work in §3.6.

In Chapter 4, we develop order-preserving time integrators. We introduce the stiff problems in §4.1. We demonstrate the order reduction phenomenon via examples and present its remedy in §4.2. The notion of weak stage order is defined in §4.3, and we develop the weak stage order theory in §4.4. A systematic methodology to construct high-order DIRK schemes with high WSO is provided in §4.5. The numerical results are shown by the newly developed schemes on a variety of test problems in §4.6 and §4.7. We close the thesis with a broader discussion and outlook in Chapter 5.
CHAPTER 2

LONG TIME NUMERICAL SOLUTIONS FOR ADVECTION PROBLEMS

Accurately solving advection problems is fundamentally challenging if one requires highly accurate solutions over a long time. Many existing high-order numerical methods do not produce satisfactory solutions over time, even for simple problems like the mere translation of structures in an advection-dominated multi-phase fluid flow simulation. The need for numerical schemes with long time accurate solutions inspires us to study whether there are numerical schemes for advection problems that can produce accurate solutions for an infinite time horizon. We begin to study this problem by considering the linear advection equation with constant-coefficients in 1D

\[ u_t + au_x = 0 \]  \hspace{1cm} (2.1)

on the periodic domain \( x \in [0, 1] \), with initial conditions \( u(x, 0) = u_0(x) \). This problem is to be understood as a model problem for the advection of a field quantity (or level set function), \( u_t + \vec{a} \cdot \nabla u = 0 \), under a locally constant (\( \vec{a} = \text{const.} \)) or slowly varying (\( \vec{a}(\vec{x}, t) \)) velocity field; and the periodic boundary conditions stand as a proxy for a long evolution on a large or unbounded
domain.

Despite the simplicity of constant-coefficient linear advection equation, developing numerical methods that accurately evolve the solution on a fixed grid is a non-trivial task. Of course, for such problems, there are methods that are truly exact, namely most standard linear schemes when the CFL number is chosen exactly 1, so that the solution is moved exactly from one grid point to the next in one time step. However, because such a choice is not an option in most practical problems (e.g., due to varying or unknown velocity fields), we explicitly exclude this trivial case and restrict to CFL numbers truly less than 1.

The weaker objective to design numerical schemes that exhibit little deformation of the numerical solution over long times has been pursued, for instance via Fourier continuation methods [2, 11]. Due to low dispersive errors, these methods achieve high accuracy for advection-dominated problems compared to traditional finite difference/volume methods. Yet, the methods are unable to generate convergent infinite time limits, for the simple reason that they are linear (see §2.2).

To construct numerical methods that do possess convergent infinite time limits, we employ the jet scheme methodology [76], equipped with a special nonlinear interpolant. Jet schemes are semi-Lagrangian fixed-grid approaches that achieve high order by tracking function values and derivatives along characteristic curves, defined via optimally local Hermite interpolation at the foot points of the characteristic curves. These schemes preserve the true solution’s structure over an infinite time and hence recover the full order of convergence of the methods (see §2.4.2).

In the next section, we start by defining a new notion of convergence in an infinite time limit of numerical methods, which will be employed to study a method’s accuracy.
2.1 Notion of Convergence and Formulation of the Problem

To study numerical schemes for equation (2.1) in terms of their long-time, or even infinite time, accuracy, a departure from established numerical analysis notions of convergence of numerical approximations is needed, as follows. Traditionally, one chooses a fixed finite final time $t_f$, and considers a sequence of meshes, parameterized by a mesh-size $h$. On each mesh, with time step $\Delta t = \frac{\mu}{h} \, h$ (where $0 < \mu < 1$ is the fixed Courant-Friedrichs-Lewy (CFL) number [20]) one computes a numerical approximation $U^{h,t_f}$ at the fixed final time $t_f$. The numerical scheme is then called convergent if the error $\varepsilon^{h,t_f} = \|I(U^{h,t_f}) - u^t\|$ satisfies $\varepsilon^{h,t_f} \to 0$ as $h \to 0$, where $u^t$ is the true solution at time $t_f$, the norm $\|\cdot\|$ is a suitable function space norm [47], and $I(U^{h,t_f})$ extends $U^{h,t_f}$ to the full domain via a suitable interpolation (see §2.2.2 and §2.4.3).

The proposed new notion of convergence considers the error as a function of both $h$ and time $t$. For instance, given $h$ with $\frac{1}{h} \in \mathbb{N}$ and $t$, compute the numerical approximation (with $\Delta t = \frac{\mu}{a} \, h$) at time $t$, denoted $U^{h,t}$. Then define the error $\varepsilon^{h,t} = \|I(U^{h,t}) - u^t\|$, using the same norm as above. With this bivariate error, the limit $h \to 0$ with $t$ fixed (i.e., pointwise convergence in time) recovers the traditional notion of convergence. We introduce a new concept of convergence that defines a numerical scheme’s convergent infinite time limit solutions. We say a scheme has the convergent infinite time limit property if

\begin{enumerate}
\item[(i)] the method is convergent in the traditional sense, and
\item[(ii)] $\lim_{h \to 0} \limsup_{t \to \infty} \varepsilon^{h,t} = 0$.
\end{enumerate}

(2.2)

The convergence of infinite time solution by a method is defined by the repeated limit in $(ii)$. Here the true solution is to be understood in a moving frame of reference $x - at$, which is $u(x,t) = u_0(x - at)$, or $u(x,t) = u_0(\text{mod}(x - at, 1))$ on the periodic domain $[0, 1]$. Note that the $\limsup_{t \to \infty}$ is used because
a common occurrence in numerical methods is that they produce a quasi-periodic temporal evolution of the error. So an error limit may not exist in the strict sense, and the \( \lim \sup \) is the right measure of the maximum error over time in such a quasi-periodic case.

**Remark 2.1.** It is worth noting that the methods satisfy the following commuting limits property as a consequence of the property (2.2).

\[
\lim_{h \to 0} \limsup_{t \to \infty} \varepsilon_{h,t} = 0 = \limsup_{t \to \infty} \lim_{h \to 0} \varepsilon_{h,t}.
\]

If a numerical scheme is convergent in the traditional sense, then for any time \( t \) fixed, \( \lim_{h \to 0} \varepsilon_{h,t} = 0 \) by definition. So \( \limsup_{t \to \infty} \lim_{h \to 0} \varepsilon_{h,t} = 0 \).

Our objective is to have a numerical scheme that produces accurate numerical solutions at any time step. In the case of equation (2.1), we want a numerical scheme that mimics the translation property of the advection equation, i.e., the method correctly shifts a numerical approximation from some instance in time to the next time step. Given that on the periodic domain \( x \in [0,1] \) the solution returns to its initial configuration every \( \frac{1}{a} \) times, a numerical solution is said to be a **fixed point** if it agrees at time \( t + \frac{n}{a} \), for some integer \( n \), with its state at time \( t \). Moreover, given an interpolant \( \mathcal{I} \) that defines a numerical solution on the whole domain (see above), a numerical approximation is said to be a **single step fixed point** if the scheme precisely moves it from one time step to the next time step. Hence, having a single step fixed point for all steps is a special case of having a fixed point solution. Finally, we call a scheme an **exact method** if any smooth solution can be approximated by a convergent sequence of fixed points (see §2.4.6 for establishing these properties).
2.2 Long Time Solution Behavior of Linear Methods

The first question is: can linear schemes have the convergent infinite time limit property (2.2), or even more: do linear schemes exist that precisely shift the numerical solutions (to constant-coefficient linear advection problems) from one time step to next time step? Moreover, if such numerical solutions exist, then are those exact solutions rich enough (i.e., form a dense set) to allow the approximation of any arbitrary initial conditions from a class of functions?

Here we do not study the all possible linear schemes; instead, we concentrate on a crucial class of linear methods that use the same local approximation at every grid point and are based on a fixed-grid and equidistant space and time step. To show that this class of linear methods does not have the convergent infinite time limit property (2.2), we consider the problem (2.1) on the entire real line and make the following assumptions about a linear scheme.

Assumption 2.1.

(i) Consider a sequence of schemes (parameterized by $h$) with fixed CFL number $\mu = \frac{a \Delta t}{h}$. Moreover, assume $0 < \mu < 1$ to exclude trivial schemes.

(ii) The scheme is defined with constant coefficients $c^h_j$ for each grid point.

(iii) The scheme uses the same update stencil for each grid point at every time step, and (for fixed $\mu$) the stencil is independent of $h$, i.e., $c^h_j = c_j \forall h$.

(iv) The scheme is stable and consistent.

(v) The initial condition $u_0 \in C(\mathbb{R}) \cap L^\infty(\mathbb{R})$.

Given a grid with mesh size $h$, let $U^{h,t_n}$ denote the state vector of the numerical approximation at the $n$-th time step. There is a wide variety of numerical methodologies, for example: finite difference methods [47] choose the components of $U^{h,t_n}$ to be approximations to the solution at grid points; finite
volume methods [29, 46] use cell averages; discontinuous Galerkin methods [18, 19] use polynomial moments; jet schemes [76, 17] use point-wise function values and derivatives. However, what is common in the particular class of linear methods considered here, is that the effects of many steps are fully captured by understanding a single step of the method. Here we consider two frameworks to understand the long time solutions of a class of linear methods. Given $h$ and $\Delta t$ related via a CFL number, we either take a finite domain with periodic boundary conditions (to demonstrate numerical examples), or we consider an infinite domain (to carry out a Von Neumann stability analysis [47] without boundary condition artifacts). We consider the recurrence relation

$$U_{j+1}^n = g(\xi)U_j^n$$

(2.3)

to obtain numerical solutions $U_{h,t}^n = \{U_j^n\}_{j=-\infty}^{\infty}$ on an infinite grid at time $t_n$, where the *amplification factor* $g(\xi)$ is a function of the wave number $\xi$. Note that, for notational economy, we here omit the superscript $h$ in the components of $U_{h,t}^n$ and write $n$ instead of $t_n$.

For a finite grid with periodic boundary conditions, $X_m = \{x_0, x_1, \ldots, x_{m-1}\}$, where $x_j = jh$ for $j = 0, 1, \ldots, m-1$ and $h = \frac{1}{m}$, using the analogous notation, the update rule of the method can be written as

$$U_{h,t}^{n+1} = M_h U_{h,t}^n.$$

Now $g(\xi)$ assumes only discrete values, and they are exactly the eigenvalues of the matrix $M_h$. The assumptions Item (ii) and Item (iii) imply that the scheme uses the same stencil for each grid point. Then the scheme is defined by a vector $c = (c_0, c_1, \ldots, c_{m-1}) \in \mathbb{R}^m$, and $M_h$ becomes a circulant matrix. The scheme’s behavior is then determined by the spectral properties of $M_h$. 

The matrix $M_h$ and its $j$-th eigenvector are given by

$$M_h = \begin{pmatrix} c_0 & c_1 & c_2 & \cdots & c_{m-1} \\ c_{m-1} & c_0 & c_1 & \cdots & c_{m-2} \\ c_{m-2} & c_{m-1} & c_0 & \cdots & c_{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & \cdots & c_{m-1} & c_0 \end{pmatrix} \quad \text{and} \quad v_j = \begin{pmatrix} \omega_0^j \\ \omega_1^j \\ \omega_2^j \\ \vdots \\ \omega_{m-1}^j \end{pmatrix},$$

corresponding to the eigenvalue $\lambda_j = \sum_{k=0}^{m-1} c_k \omega_m^k$ for $j = 0, 1, \ldots, m - 1$. Here $\omega_m = e^{\frac{2\pi i}{m}}$ and $\omega_0^m, \omega_1^m, \ldots, \omega_{m-1}^m$ are the $m$-th roots of unity. Circulant matrices are diagonalizable; let $M_h = F_h \Lambda_h F_h^{-1}$ with $F_h = [v_0 | v_1 | \cdots | v_{m-1}]$ and $\Lambda_h = \text{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{m-1})$. $F_h$ is a symmetric matrix with $F_h^{-1} = \frac{1}{m} F_h^*$. We obtain the numerical approximation after $n$ steps

$$U^{h,t_n} = (M_h)^n U^{h,t_0} = F_h (\Lambda_h)^n F_h^{-1} U^{h,t_0} = F_h (\Lambda_h)^n \alpha_h = \sum_{j=0}^{m-1} v_j \alpha_j^h(\lambda_j)^n,$$

where $\alpha_h = F_h^{-1} U^{h,t_0} = \left[\alpha_0^h, \alpha_1^h, \ldots, \alpha_{m-1}^h\right]^T$. All the components corresponding to $\lambda_j$ with $|\lambda_j| < 1$ will decay to zero as $n \to \infty$, and all that remains are the components corresponding to $\lambda_j$ with $|\lambda_j| = 1$. Let $C = \{j : |\lambda_j| = 1\}$, and for all $j \in C$ write $\lambda_j = e^{i\theta_j}$, $\theta_j \in [0, 2\pi)$. Then the infinite time limit becomes

$$\lim_{n \to \infty} U^{h,t_n} = \sum_{j \in C} v_j \alpha_j^h e^{i\theta_j n}.$$

Before we move on to providing a generic proof on an infinite grid, we demonstrate the long time solutions via three illustrative examples on a finite domain.

### 2.2.1 Illustrative examples

We consider the equation (2.1) on $[0, 1]$ with initial conditions $u(x, 0) = \sin(2\pi x) + \sin(10\pi x)$, and periodic boundary conditions. On a finite grid $X_m = \{x_0, x_1, \ldots, x_{m-1}\}$, with $x_j = jh$ for $j = 0, 1, \ldots, m - 1$ and $h = \frac{1}{m}$, the one-step update rule always takes the form $U^{h,t_{n+1}} = M_h U^{h,t_n}$, where $U^{h,t_n} = [U_0^n, U_1^n, \ldots, U_{m-1}^n]^T$ is the discrete solution vector at time step $t_n$, and the matrix $M_h$ is determined by the numerical method that we apply.
Our first example is the very well-known upwind method. With a positive velocity \( a > 0 \), the scheme is given by

\[
\frac{U_j^{n+1} - U_j^n}{\Delta t} + a \frac{U_j^n - U_{j-1}^n}{h} = 0 \quad \Leftrightarrow \quad U_j^{n+1} = U_j^n - \frac{a \Delta t}{h} \left( U_j^n - U_{j-1}^n \right). \tag{2.4}
\]

The upwind scheme 2.4 takes the form \( U^{h,t,n+1} = M_h U^{h,t,n} \), where the matrix \( M_h \) is given by \( \mu = \frac{a \Delta t}{h} \)

\[
M_h = \begin{pmatrix}
1 - \mu & 0 & 0 & \cdots & \mu \\
\mu & 1 - \mu & 0 & \cdots & 0 \\
0 & \mu & 1 - \mu & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \mu & 1 - \mu \\
\end{pmatrix}.
\]

Figure 2.1: Upwind method for \( u_t + u_x = 0 \) with \( \mu = 0.8 \) (CFL) and \( m = 100 \) grid points: Amplification factor \( g(\xi) = (1 - \mu) + \mu e^{-i\xi h} \), and eigenvalues (left) and numerical solution at different times (right). All the eigenvalues except 1 are clearly inside the unit circle, resulting in decay towards a constant.

Figure 2.1 shows the behavior of the basic upwind method, whose behavior is typical for many other schemes as well. All eigenvalues are truly inside the unit disc, except for one at \( \lambda_1 = 1 \) whose eigenvector corresponds to a constant function (a consequence of having a consistent scheme). Hence, as \( n \to \infty \), all modes decay except for the constant mode, and the numerical solution always decays towards a constant function (the average of the initial condition), independent of \( h \), thus \( \lim_{h \to 0} \limsup_{n \to \infty} \varepsilon^{h,t,n} = \text{const.} \). Consequently, the scheme is not exact.
Our second example considers a method where the spatial derivative is approximated by standard second order central finite differences
\[ u_x \approx \frac{u(x + h) - u(x - h)}{2h} . \]

We apply the method of lines [47] first to discretize the space, and obtain a coupled system of \( m \) ODEs for the variables \( U_j(t) \), where \( U_j(t) \) is the solution along the line forward in time at the grid \( x_j \). The system of ODEs is given by
\[
\frac{dU^h(t)}{dt} = -aD_h U^h(t) ,
\]
where
\[
D_h = \frac{1}{2h} \begin{pmatrix}
0 & 1 & 0 & \cdots & \cdots & -1 \\
-1 & 0 & 1 & \cdots & \cdots & 0 \\
0 & -1 & 0 & 1 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & \cdots & \cdots & -1 & 0 & 1 \\
1 & 0 & \cdots & \cdots & -1 & 0
\end{pmatrix}
\]
and \( U^h(t) = \begin{pmatrix} U_0(t) \\ U_1(t) \\ U_2(t) \\ \vdots \\ U_{m-1}(t) \end{pmatrix} \),

thus the corresponding method of lines matrix has purely imaginary eigenvalues. To make the scheme stable, we need to time-step via an ODE solver whose stability region contains at least some portion of the imaginary axis. We choose the Shu-Osher method [80] (a SSP scheme [31]), whose intercept on the imaginary axis renders the scheme conditionally stable. Then the one-step update rule becomes \( U^{h,t_{n+1}} = M_h U^{h,t_n} \), where
\[
M_h = \left( I - a\Delta t D_h + \frac{(a\Delta t)^2}{2!}D_h^2 - \frac{(a\Delta t)^3}{3!}D_h^3 \right) .
\]
The eigenvalues of \( M_h \) (shown in Figure 2.2) again have only \( \lambda_1 = 1 \) on the unit circle; all other modes decay to zero as \( n \to \infty \). So the \( n \to \infty \) limit is the same as with upwind. Note that for this scheme (and even more so for methods of higher order), some modes decay slowly (and slower and slower as \( h \to 0 \)). However, the method does not have the convergent infinite time limit property.
Figure 2.2: SSP-RK3 in time and central in space method for $u_t + u_x = 0$ with $\mu = 0.8$ and $m = 100$ grid points: Amplification factor $g(\xi) = \left(1 - \frac{\mu^2}{4}\right) + \frac{\mu^2}{4} \cos 2\xi h - i \left[\frac{\mu^3}{24} \sin 3\xi h + \left(\mu - \frac{\mu^3}{8}\right) \sin \xi h\right]$, and eigenvalues (left) and numerical solution at different times (right). Despite high order, all the eigenvalues except 1 are inside the unit circle, resulting in decay towards a constant.

The last example is a method where the spatial derivative is approximated by the fourth-order central finite differences

$$u_x \approx -u(x+2h) + 8u(x+h) - 8u(x-h) + u(x-2h) / 12h,$$

leading to a method of lines matrix with all purely imaginary eigenvalues. Following the method of lines as previous example we get a system of ODEs as

$$\frac{dU^h(t)}{dt} = -aD_h U^h(t),$$

where

$$D_h = \frac{1}{12h} \begin{pmatrix}
0 & 8 & -1 & \cdots & \cdots & \cdots & 1 & -8 \\
-8 & 0 & 8 & -1 & \cdots & \cdots & \cdots & 1 \\
1 & -8 & 0 & 8 & -1 & \cdots & \cdots & 0 \\
0 & 1 & -8 & 0 & 8 & -1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & \cdots & \cdots & 1 & -8 & 0 & 8 & -1 \\
-1 & \cdots & \cdots & \cdots & 1 & -8 & 0 & 8 \\
8 & -1 & \cdots & \cdots & 1 & -8 & 0 & 8
\end{pmatrix}.$$
Figure 2.3: Crank-Nicolson in time and fourth-order in space method for \( u_t + u_x = 0 \) with \( \mu = 0.8 \) (CFL) and \( m = 100 \) grid points: Amplification factor 
\[
g(\xi) = \frac{1-i(\frac{\mu}{4} \sin \xi h - \frac{\mu}{2} \sin 2\xi h)}{1+i(\frac{\mu}{4} \sin \xi h - \frac{\mu}{2} \sin 2\xi h)},
\]
and eigenvalues (left) and numerical solution at different times (right). All eigenvalues are on the unit circle, hence no modes decay; however, they move at wrong speeds.

Now, we time-step via Crank-Nicolson (which we use for the sake of argument — implicit methods are non-popular for advection problems [47]). Because the imaginary axis is exactly the stability boundary for Crank-Nicolson, the resulting scheme is neutrally stable. In this case, the matrix of the one-step update rule becomes
\[
M_h = \left( I + \frac{a\Delta t}{2} D_h \right)^{-1} \left( I - \frac{a\Delta t}{2} D_h \right).
\]
In contrast to the other examples above, the third scheme shown in Figure 2.3 has all eigenvalues on the unit circle, i.e., the scheme is purely dispersive. Consequently, all modes contained in the initial \( (t = 0) \) approximation persist for all time. However, they move at different speeds, and generally do not reproduce the initial conditions after integer multiples of \( 1/a \), a fact that holds true broadly, see below.

### 2.2.2 Qualitative analysis

The numerical results above highlight some principles of how a crucial class of linear schemes tends to fail to yield the convergent infinite time limit property. This section formally proves that a fixed-grid linear method under
the Assumption 2.1 can not have the convergent infinite time limit property.

We employ Von Neumann stability analysis to characterize the stability assumption of the scheme. The analysis is particularly simple for a Cauchy problem [47], i.e., we consider equation (2.1) on the real line \(-\infty < x < \infty\) with no boundaries. We discretize the whole domain and approximate the initial condition to get the grid function \(u^{h,0} = \{u_j^0\}_{j=-\infty}^{\infty}\), where \(u_j^0\) is the evaluation of initial condition at grid points \(x_j = jh\) for \(j \in \mathbb{Z}\). As above, here we discard the superscript \(h\) in the components of \(u^{h,0}\), for notational economy. For a grid function \(u^{h,0}\) defined on \(h\mathbb{Z}\) the semidiscrete Fourier transform [85] is defined by
\[
\hat{u}(\xi) = h \sum_{j=\infty}^{j=-\infty} u_j^0 e^{-ijh\xi}.
\]
The inverse semidiscrete Fourier transform is defined by
\[
u_j^0 = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} \hat{u}(\xi) e^{ijh\xi} d\xi, \quad j \in \mathbb{Z}.
\]
For the time being, we skip the technicalities of validating these definitions and come back to them later. Note that our physical variable is discrete and unbounded, while the frequency variable is bounded and continuous in the interval \([-\pi/h, \pi/h]\). Now, by Item (ii) and Item (iii), the one-step update rule for any linear scheme using the stencil \([x_j - sh, x_j - (s-1)h, \ldots, x_j + sh]\] can be written as
\[
U_{n+1}^j = \sum_{\nu=-s}^{s} c_{\nu} U_{n}^{j+\nu}.
\]
We decompose the initial grid function into Fourier modes by using the semidiscrete Fourier transform and express the grid values as
\[
u_j^0 = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} \hat{u}(\xi) e^{ijh\xi} d\xi, \quad \forall \ j.
\]
We now look at the effect of the numerical method on each of the Fourier modes and combine all the modified modes to obtain the numerical solution at some given time. Using equation (2.3) and the numerical scheme (2.5) on the grid function \(U_j^n = e^{\xi_j h}\) of a single wave number \(\xi\), we calculate \(g(\xi) = \)
\[ \sum_{\nu=-s}^{s} c_{\nu} e^{i\nu h \xi}. \]

Using the recurrence relation equation (2.3) on each mode and combining them all we obtain the numerical solution \( U_{h,t}^{n} \) at \( t_{n} \) given by its components

\[ U_{j}^{n} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{u}(\xi)[g(\xi)]^{n} e^{i\xi j h} d\xi, \quad \forall \ j. \]

We seek to compare the numerical solution with the true solution in the continuum \( L^{\infty} \) norm, so it is natural to consider the trigonometric interpolant based on the grid function \( \{ U_{j}^{n} \}_{j=-\infty}^{\infty} \) that is defined by

\[ I(U_{h,t}^{n})(x) := \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{u}(\xi)[g(\xi)]^{n} e^{i\xi x} d\xi, \quad \forall \ x \in \mathbb{R}, \quad (2.6) \]

as our numerical solution function (defined for all \( x \)) at \( t_{n} \). A natural (yet not the only possible) way to initialize the numerical scheme is with the discrete data \( \{ u_{j}^{0} \}_{j=-\infty}^{\infty} \), directly recovered from the initial profile. The numerical schemes now evolve this discrete solution over time, and the interpolant (2.6) defines a corresponding continuum function.

Here we are only interested in studying the schemes’ evolution error, so we employ an error measure that yields zero error at time \( t = 0 \): rather than comparing the numerical approximation (2.6) to the true initial profile \( u_{0} \), we instead compare it to a proxy of the true solution, the trigonometric interpolant of the initial grid function,

\[ I(u_{h,0})(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{u}(\xi) e^{i\xi x} d\xi, \quad \forall \ x \in \mathbb{R}. \]

Moreover, like the true solution is just the shift of the initial profile, so is the proxy of the true solution at time \( t_{n} \) obtained via shifting the function \( I(u_{h,0})(x) \) by \( n \Delta t \) to yield

\[ I(u_{h,t}^{n})(x) := \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{u}(\xi) e^{i(x-n\Delta t)\xi} d\xi, \quad \forall \ x \in \mathbb{R}. \quad (2.7) \]

Having a stable scheme requires that \( |g(\xi)| \leq 1 \). All components corresponding to \( \xi \) with \( |g(\xi)| < 1 \) will decay to zero as \( n \to \infty \), and all that remains are the components corresponding to \( \xi \) with \( |g(\xi)| = 1 \). Let \( \mathcal{C} = \{ \xi : |g(\xi)| = 1 \} \), and
for all $\xi \in C$ write $g(\xi) = e^{i\theta(\xi)}$. This means that the exact numerical solutions of a linear scheme are precisely the Fourier modes that have their amplification on the unit circle and move with the correct speed as the true solution. For a linear scheme to have the convergent infinite time limit property (2.2), we would need $C$ to possess increasingly many elements that move with the correct speed, as $h \to 0$ (and thus the number of grid points goes to infinity). However, all linear schemes satisfying Assumption 2.1 have the property that as $h \to 0$, the Fourier modes approaches an algebraic curve in the complex plane. Consequently, one of two scenarios arises: either $C$ has only finitely many elements, or all Fourier modes lie on unit circle. In the latter case, the scheme is purely dispersive; however, expect for trivial cases, the wave speeds of the different Fourier modes do not match up in a way to recover an exact solution. Regarding this result, we now present the following technical lemma, which we are going to use to prove our goal.

**Lemma 2.1.** Under Assumption 2.1, one has

$$\limsup_{n \to \infty} \left| [g(\xi)]^n - e^{-i\xi nh} \right| \geq C \neq 0,$$

where $\xi$ is almost any wave number, $\mu$ is the CFL number, and $C$ is a positive constant.

**Proof.** The amplification factor is $g(\xi) = \sum_{\nu=-s}^{s} c_{\nu} e^{i\nu \xi}$. By a change of variable $\omega = \xi h$, $g$ becomes a function of $\omega$, and we denote this function by

$$\tilde{g}(\omega) = \sum_{\nu=-s}^{s} c_{\nu} e^{i\nu \omega}, \quad \omega \in [-\pi, \pi].$$

The $\limsup_{n \to \infty} \left| [g(\xi)]^n - e^{-i\xi nh} \right|$ becomes $\limsup_{n \to \infty} \left| [\tilde{g}(\omega)]^n - e^{-i\omega n \mu} \right|$. Now, if $|\tilde{g}(\omega)| < 1$ for $\omega \in [-\pi, \pi]$, then

$$\limsup_{n \to \infty} \left| [\tilde{g}(\omega)]^n - e^{-i\omega n \mu} \right| = 1.$$

For $\omega$’s with $|\tilde{g}(\omega)| = 1$ we can write $\tilde{g}(\omega) = e^{i\theta(\omega)}$, where $\theta(\omega) \in [0, 2\pi)$ and
then

\[
\limsup_{n \to \infty} \left| \tilde{g}(\omega)^n - e^{-i\omega n \mu} \right| = \limsup_{n \to \infty} \left| e^{in\theta(\omega)} - e^{-i\omega n \mu} \right|
\]

\[
= \limsup_{n \to \infty} \left| e^{i(\theta(\omega) + \omega \mu)n} - 1 \right|
\]

\[
= \limsup_{n \to \infty} \left| z^n(\omega) - 1 \right|,
\]

where \( z(\omega) = e^{i(\theta(\omega) + \omega \mu)} \). Clearly, \( |z(\omega)| = 1 \) for all \( \omega \in [-\pi, \pi] \), and \( z(0) = 1 \) as \( g(0) = 1 \) follows from the consistency (Item (iv)) of the scheme. We claim that \( z(\omega) = 1 \) only for finitely many \( \omega \)'s in \([-\pi, \pi]\). If not, then it follows that \( \tilde{g}(\omega) = e^{-i\omega \mu} \) on a subset \( D \) of \([-\pi, \pi]\) with a limit point. As \( \tilde{g}(\omega) \) and \( e^{-i\omega \mu} \) are analytic functions, we have \( \tilde{g}(\omega) = e^{-i\omega \mu} \) on the whole domain \([-\pi, \pi]\) (a property of complex analytic functions), which implies

\[
\sum_{\nu = -s}^{s} c_{\nu} e^{i\nu \omega} = e^{-i\omega \mu}, \quad \forall \omega \in [-\pi, \pi].
\]

This is a contradiction to Item (i) because \( \{e^{i\nu \omega} : \nu = -s, \ldots, s\} \) is a linearly independent set, and the only way this identity is true is when \( \mu = \nu, c_{-\nu} = \delta_{\nu \mu} \) and \( |\mu| \leq s \). But this will lead to a trivial numerical scheme. So we conclude that \( z(\omega) \neq 1 \) almost everywhere in \([-\pi, \pi]\). Then for almost all \( \omega \in [-\pi, \pi] \),

\[
\limsup_{n \to \infty} \left| \tilde{g}(\omega)^n - e^{-i\omega n \mu} \right| = \limsup_{n \to \infty} \left| z^n(\omega) - 1 \right|
\]

\[
\geq \left| e^{\frac{2\pi}{3}} - 1 \right| = \sqrt{3}.
\]

Figure 2.4: Geometry for bound on the estimate.
The last inequality follows by considering all possible values of the argument of the complex number \( z = e^{i[\theta(\omega) + \omega \mu]} \neq 1 \) that is taken to arbitrary powers \( n \), see Figure 2.4. Hence, for almost all \( \xi \in [-\frac{\pi}{h}, \frac{\pi}{h}] \),

\[
\limsup_{n \to \infty} |g(\xi)|^n - e^{-i\xi n \mu h} \geq \min\{1, \sqrt{3}\} = 1, \text{ independent of } h.
\]

We now define the Fourier transform from the space \((L^\infty(\mathbb{R}), \| \cdot \|_{L^\infty})\) to \((\mathcal{S}'(\mathbb{R}), \| \cdot \|_{L^1})\), the space of tempered distributions. Let \( \mathcal{D} = \mathcal{F}(L^\infty(\mathbb{R})) \) and \( \mathcal{S}(\mathbb{R}) \) is the Schwartz space consists of functions whose derivatives are rapidly decreasing. Then define the Fourier transformation map by

\[
\mathcal{F}: (L^\infty(\mathbb{R}), \| \cdot \|_{L^\infty}) \to (\mathcal{D}, \| \cdot \|_{L^1}) \quad f \mapsto \mathcal{F}(f) = \hat{f}, \text{ where } \hat{f}(\phi) := f(\hat{\phi}) := \int_{\mathbb{R}} f(x) \phi(x) dx, \phi \in \mathcal{S}(\mathbb{R}).
\]

Here the \( \| \cdot \|_{L^1} \) norm of a distribution \( T \) is defined as \( \| T \|_{L^1} = \sup\{|T(\phi)| : \|\phi\|_{L^\infty} \leq 1\} \). Then the map is well-defined and it can be shown that

\[
\|f\|_{L^\infty} = \|\hat{f}\|_{L^1}.
\] (2.8)

Now we prove the theorem for those initial conditions that are the linear combination of finitely many wave functions of different frequencies, i.e.,

\[
u(x) = \sum_{j=1}^{J} c_j e^{i\xi_j x},
\] (2.9)

where \( \xi_j \) are some selected frequencies with \( c_j \neq 0 \) for all \( j \).

**Theorem 2.1.** A linear scheme for equation (2.1) with initial conditions (2.9) under the Assumption 2.1 cannot have the convergent infinite time limit property, i.e.,

\[
\lim_{h \to 0} \limsup_{n \to \infty} \varepsilon_{h,t_n}^{h,t_n} \neq 0.
\]

**Proof.** For a given \( h \) and \( t_n \), we measure the error in the \( \| \cdot \|_{L^\infty} \) norm,

\[
\varepsilon_{h,t_n}^{h,t_n} = \| I(U_{h,t_n}) - I(u_{h,t_n}) \|_{L^\infty}.
\]
Using (2.6), (2.7) and the fact that $\Delta t = \mu h$, we get

$$(\mathcal{I}(\mathbf{U}^{h,t_n}) - \mathcal{I}(\mathbf{u}^{h,t_n}))(x) = \frac{1}{2\pi} \int_{\frac{-\pi}{h}}^{\frac{\pi}{h}} \hat{u}(\xi) \left([g(\xi)]^n - e^{-i\xi n \mu h}\right) e^{i\xi x} d\xi, \; \forall \; x \in \mathbb{R}.$$  

For $u(x) = \sum_{j=1}^{J} c_j e^{i\xi_j x}$ we have $\hat{u}(\xi) = \sum_{j=1}^{J} c_j \delta(\xi - \xi_j)$ and let

$$d(\xi) = \hat{u}(\xi) \left([g(\xi)]^n - e^{-i\xi n \mu h}\right), \; \text{for} \; \xi \in \left[\frac{-\pi}{h}, \frac{\pi}{h}\right].$$

Then clearly $d(\xi) \in (D, \|\cdot\|_{L^1})$, as $\hat{u}$ is a superposition of Dirac delta functions and $|g(\xi)| \leq 1$, by Item iv. Also by construction, $d(\xi)$ is the Fourier transform of the function $(\mathcal{I}(\mathbf{U}^{h,t_n}) - \mathcal{I}(\mathbf{u}^{h,t_n}))(x)$, for $x \in \mathbb{R}$. Using the identity equation (2.8) established above we can write

$$\epsilon^{h,t_n} = \| (\mathcal{I}(\mathbf{U}^{h,t_n}) - \mathcal{I}(\mathbf{u}^{h,t_n}))(x) \|_{L^\infty} = \|d(\xi)\|_{L^1}.$$  

First taking $\limsup_{n \to \infty}$ and then taking the limit as $h \to 0$ of the error $\epsilon^{h,t_n}$, we get

$$\lim_{h \to 0} \limsup_{n \to \infty} \epsilon^{h,t_n} = \lim_{h \to 0} \limsup_{n \to \infty} \|d(\xi)\|_{L^1}$$

$$= \lim_{h \to 0} \limsup_{n \to \infty} \int_{\frac{-\pi}{h}}^{\frac{\pi}{h}} \left| \sum_{j=1}^{J} c_j \delta(\xi - \xi_j) \left([g(\xi)]^n - e^{-i\xi n \mu h}\right) \right| d\xi$$

$$\geq \lim_{h \to 0} \limsup_{n \to \infty} \int_{\frac{-\pi}{h}}^{\frac{\pi}{h}} \left| c_{j_0} \delta(\xi - \xi_{j_0}) \left([g(\xi)]^n - e^{-i\xi n \mu h}\right) \right| d\xi,$$

for some $j_0 \in \{1, 2, \cdots, J\}$,

$$= |c_{j_0}| \lim_{h \to 0} \limsup_{n \to \infty} \left| \left([g(\xi_{j_0})]^n - e^{-i\xi_{j_0} n \mu h}\right) \right|,$$

Lemma 2.1

$$\geq |c_{j_0}| \lim_{h \to 0} 1 = |c_{j_0}| > 0.$$  

**Remark 2.2.** Although the proof of the theorem is based on initial conditions that are linear combinations of a finite number of wave functions, we conjecture that the statement of the theorem holds for more general initial conditions.
2.2.3 Quantitative scaling

Having shown that a class of linear schemes are not suitable to produce accurate solutions for really long times, we now provide scaling laws of the errors incurred by these schemes depending on the grid size $h$ and the final time $t_f$. In order to have a chance to obtain an accurate infinite time limit solution, we require that, for a fixed $h$, the numerical schemes produce non-constant numerical approximations as $n \to \infty$. In practice, we cannot take arbitrarily small $\Delta t$ or $h$, but we are happy to accept $h$ for which the numerical scheme will produce the desired error to exhibit its correct order of convergence. For a given final time $t_f$, there exists a critical value of $h = h^*$, for which the method will produce that desired error. Now, if we take a larger final time $t_f$, then it will be reasonable to believe that we would require a smaller value of $h^*$ to produce the desired error. So, the critical value of $h$ depends on the choice of final time $t_f$, and it decreases as $t_f$ increases. To rigorously understand this fact we consider (2.1) and solve this problem by second-order accurate standard Lax-Wendroff method (see §2.3). The modified equation [47] for this numerical method is given by

$$v_t + av_x = \frac{1}{6}ah^2 \left(1 - \frac{a\Delta t}{h}\right)^2 v_{xxx}. \tag{2.10}$$

The true solution of (2.10) captures the behavior of the numerical solutions produced by the Lax-Wendroff method well. As $\frac{a\Delta t}{h} = \mu$ is a fixed number (CFL), we can write this equation as $v_t + av_x = Ch^2v_{xxx}$, where $C = -\frac{a}{6}(1-\mu^2)$. The differential operators $\frac{\partial}{\partial t}$ and $\frac{\partial^3}{\partial x^3}$ commute, which implies that we can solve the equation by splitting into advection and dispersion problems in any order. To understand the error made by the scheme, we can ignore the advection part as it is also present in the original problem (2.1). So, it suffices to consider only $v_t = Ch^2v_{xxx}$ to understand the change in the numerical solution over time. We compare the solution of this equation, denoted by $w_h(x,t)$, with the solution of the original problem with zero velocity, i.e., $u(x,t) = u_0(x)$. By the time-scaling $t \mapsto h^2t$, we can deduce that $w_h(x,t) = w_1(x,h^2t)$. We measure
the error for a given final time $t_f$ and the corresponding critical resolution $h^*$, by $\varepsilon^{h,t_f} = ||w_1(\cdot, (h^*)^2 t_f) - u_0(\cdot)||_{L^1}$ in the continuum $L^1$ norm. We want this error to be a constant which is the desired error to get second order accuracy for the Lax-Wendroff method. The desired error depends on $h^*$ and $t_f$, and this forces $(h^*)^2 t_f = C_1$, for some constant $C_1$. In fact, for a $p$-th order linear scheme, self-similarity arguments based on the modified equation reveal that the error generally scales like $\varepsilon^{h,t_f} \propto h^p t_f$, i.e., doubling $t_f$ also doubles the error, and $h^* \propto t_f^{-1/p}$. Even though for a fixed $t_f$ the error decreases as $h \to 0$, for a fixed $h$ the error grows as $t_f \to \infty$, thus preventing the convergent infinite time limit property (2.2). The typical shape of the bivariate error function $\varepsilon^{h,t_f}$ is shown in Figure 2.5 (left panel), here for the Lax-Wendroff method. For any fixed $t_f$, the error $\varepsilon^{h,t_f}$ is L-shaped, with a transition mesh size $h^*(t_f)$ below which a clean second order manifests.

**Figure 2.5:** Approximation error $\varepsilon^{h,t_f}$ as function of $h$ and $t_f$. Left: For the second order linear Lax-Wendroff scheme, and Right: For the nonlinear jet scheme.

For the linear (second order) scheme, $\varepsilon^{h,t_f}$ decreases with $h$ but steadily grows in $t_f$; hence there is no $h \to 0$ convergence that is uniform in $t_f$. For the nonlinear jet scheme, the light blue/red curves show the total error at different final times for two approximations of a smooth initial function: (i) direct evaluation of $u_0(jh)$ and $u'_0(jh)$, denoted $(\phi, \psi)$; and (ii) the special strategy in (2.36), denoted $(\phi^\delta, \psi^\delta)$. The corresponding errors of only the jet scheme
evolution (without i.c. approximation errors) are shown by the dark blue/red curves for (i)/(ii), respectively. For the jet scheme, the total error $e^{h,t}$ decreases with $h$ and it is (essentially) constant w.r.t. $t$. Hence the scheme has the potential to be exact. Even though there is no traditional convergence rates due to the structure of fixed points (see §2.4.2 for a more detailed discussion of the convergence of jet schemes), the jet scheme’s evolution error (dark blue) is observed to decrease (in $h$) more rapidly than the approximation error of the i.c.

2.3 Long Time Solution Behavior of Traditional Nonlinear Methods

As we have established that an important class of linear methods are severely limited at possessing reasonable $t \rightarrow \infty$ limits (see §2.2), we now turn our attention to nonlinear methods. Clearly, using a nonlinear scheme has the potential to “break the spell” that boils linear schemes down to the single-step update rule; which is a nice analogy to nonlinear methods overcoming Godunov’s limitation theorem [46]. At the same time, analytical statements about the $t \rightarrow \infty$ limit of a nonlinear scheme are complicated. We look at some of the classical nonlinear schemes, for example, Lax-Wendroff with Van Leer limiter [87, 46], Lax-Wendroff with Superbee limiter [69, 46], and a high-order WENO [38, 50] method for our computational studies. We consider the finite volume framework for deriving the Lax-Webdroff method with different limiter functions as finite volume methods are preferable in general to solve the hyperbolic conservation laws

$$u_t + f(u)_x = 0 .$$

(2.11)

Finite volume methods [43, 46] subdivide the domain into intervals (in 1D) and compute an approximation to the integral of $u$ over each interval using the flux through the endpoints of the intervals. We denote the $i$th grid cell by
\( C_i = [x_i - h/2, x_i + h/2] \), where \( h \) is the cell size, and define the approximation of \( u \) at the cell center by the cell average as
\[
\bar{u}_i(t) \approx \frac{1}{h} \int_{x_{i - \frac{1}{2}}}^{x_{i + \frac{1}{2}}} u(x, t) \, dx .
\] (2.12)

The integral form of the conservation laws (2.11) is given by
\[
\frac{1}{h} \int_{C_i} u(x, t) \, dx = \frac{1}{h} \int_{C_i} u(x, t_n) \, dx
- \frac{1}{h} \left[ \int_{t_n}^{t_{n+1}} f(u(x_{i+1/2}, t)) \, dt - \int_{t_n}^{t_{n+1}} f(u(x_{i-1/2}, t)) \, dt \right] .
\] (2.13)

Integrating equation (2.13) in time from \( t_n \) to \( t_{n+1} \) we obtain
\[
\frac{1}{h} \int_{C_i} u(x, t_{n+1}) \, dx = \frac{1}{h} \int_{C_i} u(x, t_n) \, dx
- \frac{1}{h} \left[ \int_{t_n}^{t_{n+1}} f(u(x_{i+1/2}, t)) \, dt - \int_{t_n}^{t_{n+1}} f(u(x_{i-1/2}, t)) \, dt \right] .
\] (2.14)

This equation illustrates how the cell averages can be updated in time. But as we do not know the exact solution \( u(x, t) \) a priori, we cannot compute the time integrals on the right-hand side. However, this provides a guideline for designing numerical methods. Given the cell averages \( \bar{u}_i^n = \bar{u}_i(t_n) \) at time \( t_n \), we update the solution \( \bar{u}_i^{n+1} \) at time \( t_{n+1} \) by the following form of numerical methods
\[
\bar{u}_i^{n+1} = \bar{u}_i^n - \frac{\Delta t}{h} \left( F_{i+1/2}^n - F_{i-1/2}^n \right) ,
\] (2.16)
where \( F_{i-1/2} \) is called the numerical flux function, which is some approximation to the average flux at \( x_{i-1/2} \):
\[
F_{i-1/2}^n \approx \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} f(u(x_{i-1/2}, t)) \, dt .
\] (2.17)

For constant-coefficient linear advection problem (2.1), the original flux function is given by \( f(u) = au \). With positive velocity \( a > 0 \), the Lax-Wendroff method with different limiter functions are characterized by the following numerical flux function
\[
F_{i-1/2}^n = a\bar{u}_{i-1}^n + \frac{1}{2} \left( 1 - \frac{a\Delta t}{h} \right) \delta_{i-1/2}^n ,
\] (2.18)
where $\delta_{i-1/2}^n$ is some limited version of the jump $\Delta \bar{u}_{i-1/2}^n := \left( \bar{u}_i^n - \bar{u}_{i-1}^n \right)$. Choosing $\delta_{i-1/2}^n = \left( \bar{u}_i^n - \bar{u}_{i-1}^n \right)$ yields the Lax-Wendroff method. Other choices of $\delta_{i-1/2}^n$ in the numerical flux function (2.18) will lead to the numerical methods known as flux-limiter methods [46]. We set

$$\delta_{i-1/2}^n := \phi \left( \theta_{i-1/2}^n \right) \left( \bar{u}_i^n - \bar{u}_{i-1}^n \right), \quad (2.19)$$

where, for $a > 0$

$$\theta_{i-1/2}^n = \frac{\Delta \bar{u}_{i-3/2}^n}{\Delta \bar{u}_{i-1/2}^n}.$$  

The ratio $\theta_{i-1/2}^n$ can be interpreted as the measure of smoothness near $x_{i-1/2}$. The function $\phi$ depending on the measure of smoothness $\theta$, is defined as the limiter function. Different choices for the limiter function results in Lax-Wendroff method with different limiter functions. For example, $\phi(\theta) = 1$ results in the classical Lax-Wendroff method. Lax-Wendroff with superbee limiter method corresponds to the superbee limiter [69] function

$$\phi(\theta) = \max \left( 0, \min \left( 1, \frac{2\theta}{\min (2, \theta)} \right), \min (2, \theta) \right). \quad (2.20)$$

The Van Leer limiter [87] function is give by

$$\phi(\theta) = \frac{\theta + |\theta|}{\theta - |\theta|}. \quad (2.21)$$

As an example of a high-resolution method, we consider fifth-order WENO methods for our numerical studies. These methods are highly accurate in the smooth region and essentially non-oscillatory near the discontinuous or sharp transition region of the solution profile. WENO methods can be described by the method of lines, i.e., first, approximate the spatial derivative and then apply a time-stepping scheme to the system of ODEs. These methods also consider the integral form of conservation laws (2.13)

$$\frac{d\bar{u}_i(t)}{dt} + \frac{1}{h} \left( f \left( u_{i+1/2} \right) - f \left( u_{i-1/2} \right) \right) = 0,$$

where $u_{i+1/2} = u(x_{i+1/2}, t)$. We replace $f \left( u_{i+1/2} \right)$ by a numerical flux function $\hat{f} \left( u_{i+1/2}^-, u_{i+1/2}^+ \right)$ to be consistent with the upwinding for stability.
$u_{i+1/2}^-, u_{i+1/2}^+$ are the approximations of $u$ at $x_{i+1/2}$ defined below. Thus the ODE becomes

$$\frac{d\bar{u}_i(t)}{dt} = -\frac{1}{h} \left[ \hat{f} \left( u_{i+1/2}^-, u_{i+1/2}^+ \right) - \hat{f} \left( u_{i-1/2}^-, u_{i-1/2}^+ \right) \right]. \quad (2.22)$$

Given the cell averages at some time step, WENO reconstruction procedure [50] is used to calculate the approximates $u_{i\pm1/2}^\pm$. The numerical flux function $\hat{f} (u^-, u^+)$ satisfies the following:

1. $\hat{f} (u^-, u^+)$ is non-decreasing in $u^-$ and non-increasing in $u^+$,

2. $\hat{f} (u^-, u^+)$ is consistent with the physical flux $f(u)$, i.e. $\hat{f}(u, u) = f(u)$,

3. $\hat{f} (u^-, u^+)$ is Lipschitz continuous with respect to both arguments $u^-$ and $u^+$.

There are many examples of such flux functions in the literature. Here we use the Lax-Friedrichs flux function given by

$$\hat{f}^{LF} (u^-, u^+) = \frac{1}{2} \left( f(u^-) + f(u^+) - \alpha (u^+ - u^-) \right), \quad (2.23)$$

where $\alpha = \max_u |f'(u)|$. The approximations $u_{i+1/2}^-$ and $u_{i+1/2}^+$ are computed by using the WENO reconstructions based on stencils one point biased to the left and one point biased to the right, respectively. We use 5th order WENO reconstruction, which uses five cell combined stencil. The stencils $\{I_{i-2}, I_{i-1}, I_i, I_{i+1}, I_{i+2}\}$ and $\{I_{i-1}, I_i, I_{i+1}, I_{i+2}, I_{i+3}\}$ are used for the reconstruction of $u_{i+1/2}^-$ and $u_{i+1/2}^+$, respectively. To construct an approximation to $u_{i+1/2}^-$ based on $\{I_{i-2}, I_{i-1}, I_i, I_{i+1}, I_{i+2}\}$, we first, divide the five-point stencil into three three-point stencils $S_0 = I_{i-2} \cup I_{i-1} \cup I_i$, $S_1 = I_{i-1} \cup I_i \cup I_{i+1}$, and $S_2 = I_i \cup I_{i+1} \cup I_{i+2}$. Then define third order accurate polynomials $p_0(x)$, $p_1(x)$, and $p_2(x)$ on $S_0$, $S_1$, and $S_2$, respectively, based on the cell averages $\{\bar{u}_{i-2}, \bar{u}_{i-1}, \bar{u}_i, \bar{u}_{i+1}, \bar{u}_{i+2}\}$. For example, on $S_1$, we find the unique polynomial
$p_1(x)$ such that
\[
\frac{1}{h} \int_{x_{i-3/2}}^{x_{i-1/2}} p_1(x) \, dx = \bar{u}_{i-1}, \tag{2.24}
\]
\[
\frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} p_1(x) \, dx = \bar{u}_i, \tag{2.25}
\]
\[
\frac{1}{h} \int_{x_{i+1/2}}^{x_{i+3/2}} p_1(x) \, dx = \bar{u}_{i+1}. \tag{2.26}
\]

Following the same strategy, the other two polynomials $P_0(x)$ and $p_2(x)$ are calculated. Evaluating these polynomials at $x_{i+\frac{1}{2}}$ we obtain
\[
u^{(0)}_{i+\frac{1}{2}} := p_0(x_{i+\frac{1}{2}}) = \frac{1}{3} \bar{u}_{i-2} - \frac{7}{6} \bar{u}_{i-1} + \frac{11}{6} \bar{u}_i, \tag{2.27}
\]
\[
u^{(1)}_{i+\frac{1}{2}} := p_1(x_{i+\frac{1}{2}}) = -\frac{1}{6} \bar{u}_{i-1} + \frac{5}{6} \bar{u}_i + \frac{1}{3} \bar{u}_{i+1} \tag{2.28}
\]
\[
u^{(2)}_{i+\frac{1}{2}} := p_2(x_{i+\frac{1}{2}}) = \frac{1}{3} \bar{u}_i + \frac{5}{6} \bar{u}_{i+1} - \frac{1}{6} \bar{u}_{i+2}. \tag{2.29}
\]

If the function $u(x, t)$ is smooth then the approximations $\nu^{(0)}_{i+\frac{1}{2}}, \nu^{(1)}_{i+\frac{1}{2}}$ and $\nu^{(2)}_{i+\frac{1}{2}}$ are all third order accurate. To get the right cell interface value, a convex combination of these values is taken with nonlinear weights $w_0, w_1$ and $w_2$

\[
u^-_{i+\frac{1}{2}} = w_0 \nu^{(0)}_{i+\frac{1}{2}} + w_1 \nu^{(1)}_{i+\frac{1}{2}} + w_2 \nu^{(2)}_{i+\frac{1}{2}}, \tag{2.30}
\]

where
\[
w_k = \frac{\tilde{w}_k}{\bar{w}_0 + \bar{w}_1 + \bar{w}_2}, \quad \tilde{w}_k = \frac{\gamma_k}{(\varepsilon + \beta_k)^2}, \tag{2.31}
\]
with $\gamma_0 = \frac{1}{10}, \gamma_1 = \frac{3}{5}, \text{ and } \gamma_2 = \frac{3}{10}$ and $\varepsilon = 10^{-6}$. If the function $u(x, t)$ is smooth then $w_k = \gamma_k + O(\Delta x^2)$, $k = 0, 1, 2$, and if it has a discontinuity at $S_k$, then $w_k = O(\Delta x^4)$. With the choices $\gamma_0 = \frac{1}{10}, \gamma_1 = \frac{3}{5}, \text{ and } \gamma_2 = \frac{3}{10}$, the approximation becomes 5th order accurate in smooth region, and the reconstruction puts small weight to the approximation value coming from the stencil containing non-smooth region. The smoothness in the stencil $S_k$ is measured by a smoothness indicator $\beta_k$ defined as
\[
\beta_k = h \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (p''_k(x))^2 \, dx + h^2 \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (p''_k(x))^2 \, dx \text{ for } k = 0, 1, 2. \tag{2.32}
\]
These smoothness indicators can be worked out explicitly as

\[
\beta_0 = \frac{13}{12} (\bar{u}_{i-2} - 2\bar{u}_{i-1} + \bar{u}_i)^2 + \frac{1}{4} (\bar{u}_{i-2} - 4\bar{u}_{i-1} + 3\bar{u}_i)^2, \
\]

\( (2.33) \)

\[
\beta_1 = \frac{13}{12} (\bar{u}_{i-1} - 2\bar{u}_i + \bar{u}_{i+1})^2 + \frac{1}{4} (\bar{u}_{i-1} - \bar{u}_{i+1})^2, \
\]

\( (2.34) \)

\[
\beta_2 = \frac{13}{12} (\bar{u}_i - 2\bar{u}_{i+1} + \bar{u}_{i+2})^2 + \frac{1}{4} (3\bar{u}_i - 4\bar{u}_{i+1} + \bar{u}_{i+2})^2. \
\]

\( (2.35) \)

This 5th order WENO is used for the reconstruction of the states at cell boundaries, and the 3rd order SSP Shu-Osher scheme is used for the time-stepping in our numerical simulations.

Figure 2.6: Long-time evolution of numerical solutions for different nonlinear methods (with \(m = 100\) grid points): WENO5, Lax-Wendroff (LW) with van Leer limiter, and Lax-Wendroff with Superbee limiter. The bottom right show the temporal evolution of the maximum for long times.

Figure 2.6 shows the temporal evolution of the numerical solutions, using three popular nonlinear schemes: a fifth-order WENO scheme; Lax-Wendroff with the Superbee limiter; and Lax-Wendroff with the Van Leer limiter. The equation \(u_t + u_x = 0\) is solved on \(m = 100\) grid points with CFL number \(\mu = 0.9\). Note that the numerical approximations shown are at exponentially increasing times (the final profile has moved 100,000 times around the periodic
domain). One can observe that both WENO and the van Leer limiter yield a decay towards a constant state, like in the linear case, albeit smearing the solution in some nonlinear fashion. In contrast, the Superbee limiter appears to not yield a decay to a constant state—a remarkable property. However, the observed $t \to \infty$ limit is clearly far away from the initial condition. Hence, these schemes are unlikely to meet our requirements on exactness formulated in §2.1. Based on this (admittedly not fully exhaustive) numerical evidence, we conjecture that traditional nonlinear methods generally do not yield exact schemes (see §2.4 for attempts to actually find some).

## 2.4 Jet Scheme with Nonlinear Interpolation

We now demonstrate that there are in fact numerical methods that are exact. To that end, we use the jet scheme formalism, established in [60, 76, 17] and extended, e.g., in [71]. Jet schemes are semi-Lagrangian approaches that track and evolve derivative information in addition to function values. In the original papers, the derivative information enables one to achieve high order, while maintaining optimally local update rules, based on Hermite interpolation. This concept can also be extended to higher derivatives and higher space dimensions [76]. The jet scheme methodology is as follows.

**Method 2.1.**

1. Given function value ($\phi$) and derivative ($\psi$) approximations at all grid points at time $t$, to obtain the state at time $t + \Delta t$, first, for each grid point, solve the characteristic equation $\dot{x} = a(x, t)$ backwards from $t + \Delta t$ to $t$ to find the associated foot point $x_R$.

2. Evaluate (based on the grid data) the Hermite interpolant $P(x_R)$ and its derivative $P'(x_R)$ at this foot point.

3. Then, using these evaluations, evolve the characteristic equations $\dot{\phi} = 0$ and $\dot{\psi} = -a_x(x, t)\psi$ (if $a_x = 0$, this reduces to $\dot{\psi} = 0$) forward from $t$ to
t + \Delta t to yield the new function value and derivative information at the grid points.

Suitable approximations of the initial condition are discussed later in this section.

While the jet scheme methodology works in principle with any Hermite interpolant, it has thus far been established only with interpolations that are linear operators (including in the context of the gradient-augmented level set method [60]). In contrast, here we propose a specific nonlinear interpolant that is tailored to produce the exactness properties requested in §2.1. Thus the method is nonlinear as the update rule depends on the state vector nonlinearly. The nonlinear jet scheme interpolant is defined as follows (in 1D). Given data \( \phi(x_L), \phi(x_R) \) (function values) and \( \psi(x_L), \psi(x_R) \) (derivatives) we wish to define an interpolant \( P(x) \) on \([x_L, x_R]\). We consider the two linear functions emanating from \( x_L \) and \( x_R \):

\[
L_L(x) := \phi(x_L) + \psi(x_L)(x - x_L) \quad \text{and} \quad L_R(x) := \phi(x_R) + \psi(x_R)(x - x_R).
\]

If \( \psi(x_L) \neq \psi(x_R) \) and the lines’ intersection point \( x_k = \frac{\phi(x_L) - \phi(x_R) - \psi(x_L)x_L + \psi(x_R)x_R}{\psi(x_R) - \psi(x_L)} \) lies inside \((x_L, x_R)\), we use the piecewise linear function (see Fig. 2.7) formed by them:

\[
P(x) = \begin{cases} 
L_L(x) & \text{for } x \leq x_k \\
L_R(x) & \text{for } x > x_k. 
\end{cases}
\]

Otherwise, we use the simple linear interpolant

\[
P(x) = \phi(x_L) + \frac{\phi(x_R) - \phi(x_L)}{x_R - x_L}(x - x_L)
\]

as a fallback case (see Fig. 2.8). This last situation can be interpreted as having an extra line between the segments \( L_L \) and \( L_R \) whose length is taken to the maximum. Clearly, the fact that this new interpolant possesses jumps in its derivatives requires particular care in how evaluations of derivatives are defined. We discuss these aspects below.
2.4.1 Approximation of the initial condition

There are many ways how one can approximate a given initial condition $u_0$ via an initial function that is a jet scheme interpolant. For a given $m = 1/h$, denote the set of grid points by $X_m = \{x_1, x_2, \ldots, x_m\}$. If $u_0 \in C^1$, a natural approach, denoted $(\phi, \psi)$, is the direct evaluation of values and slopes on the grid, and the assignment $\phi_j = u_0(x_j)$ and $\psi_j = u'_0(x_j)$ to define the initial jet scheme approximation. This approach, discussed in §2.4.7, turns out to be accurate, but it does not guarantee to have the convergent infinite time limit property.

For that reason, we now introduce a different approximation strategy of the initial condition, denoted $(\phi^\delta, \psi^\delta)$, that only requires $u_0 \in C^0$ and that establishes the jet scheme’s exactness. First, we evaluate $u_0$ on the grid points to obtain preliminary values $\hat{\phi}_j = u_0(x_j)$ for $j = 1 \ldots, m$. Without loss of generality, we can assume that no three consecutive points are collinear (if there is such a triple, then we can modify the position of the middle point with minimal error to make them non-collinear). Based on these data $\{(x_1, \hat{\phi}_1), (x_2, \hat{\phi}_2), \ldots, (x_m, \hat{\phi}_m)\}$, let $I(x)$ be the standard piecewise linear interpolation (that connects the dots, forming kinks at the grid points). We now choose a parameter $\delta$ with $\epsilon \ll \delta \ll h$, where $\epsilon$ is the machine precision, and consider the function $I(x + \delta)$. This is the interpolant $I(x)$ shifted to the left by $\delta$, i.e., its kinks are at $x_j - \delta$ for $j = 1 \ldots, m$. Hence, $I(x + \delta)$ has well-defined slopes at the grid points, and we can assign

$$\phi_j = I(x_j + \delta) = \hat{\phi}_j + \frac{\hat{\phi}_{j+1} - \hat{\phi}_j}{h} \delta \quad \text{and} \quad \psi_j = I'(x_j + \delta) = \frac{\hat{\phi}_{j+1} - \hat{\phi}_j}{h}. \quad (2.36)$$

Below we show that this particular approximation of the initial condition gives rise to the jet scheme’s exactness, thus providing a rigorous way to construct methods that possess the convergent infinite time limit property. Afterwards, we discuss the pros and cons of other sampling strategies.
2.4.2 Quantifying approximation errors and convergence

Conceptualizing convergence of numerical methods requires defining a distance between the true solution (a function) and the sequence of numerical approximations (as $h \to 0$). This is commonly done via a function space norm (e.g., $L^p$) as in finite elements, or a scaled discrete norm that is induced from a continuum norm as in finite differences [47]. The connection between the $h$-scaled grid norm and the continuum norm is: (a) the former is a quadrature rule for the latter; and moreover: (b) the former can be defined by means of the latter via an interpolation.

Because jet schemes are fundamentally based on an interpolation step, this last interpretation comes naturally for them: rather than working with a discrete norm on the data ($\phi_j, \psi_j$), we directly measure errors in a continuum norm, based on the nonlinear Hermite interpolant defined above. In contrast to the Von Neumann analysis for linear methods considered in §2.2, we do not employ Fourier modes for the analysis of jet schemes. Hence we use the $L^1$ norm.

An important consequence of measuring errors in a continuum norm is that in general there is a nonzero error between the true initial condition and the initial ($t = 0$) interpolant that approximates it. Hence, there are two ways to measure the error of the numerical approximation at some time $t = t_f$: the total error, i.e., the $L^1$-distance of the interpolant at $t = t_f$ to the true solution; and the evolution error, i.e., the $L^1$-distance of the interpolant at $t = t_f$ to a proxy of the true solution, namely the interpolant from $t = 0$ translated via equation (2.1) to $t_f$. We have already encountered this important distinction in §2.2, and it is equally important for jet schemes.

The right panel of Figure 2.5 presents a convergence study of the jet scheme error $\varepsilon^{h,t_f}$ as function of $h$ and $t_f$. The light blue/red curves show the total error (at different final times), while the dark blue/red curves display the pure evolution errors (blue vs. red are the two different ways to approximate the initial conditions, see §2.4.1). The results show that the total error is
generally dominated by the approximation error of the initial condition (the 
\( (\phi, \psi) \) strategy yields a factor 2 smaller errors, but the \( (\phi^\delta, \psi^\delta) \) strategy does 
not require \( u_0 \) to be differentiable), while the scheme’s evolution error is much 
smaller.

One can show that the best one can do to approximate a general smooth 
initial condition by a piecewise linear function is second order (in the number 
of kinks), so the jet scheme cannot be better than second order if the total 
error is considered. However, as we are interested in the (loss of) accuracy as 
\( t \to \infty \) of the scheme itself, we focus on the evolution error.

Because the evolution error is the \( L^1 \)-distance between two jet scheme inter-
polants that are piecewise linear, this error can be calculated up to machine 
precision by (i) forming sub-intervals bounded by the kinks (of both inter-
polants) and the intersection points between the interpolants, and (ii) inte-
grating the difference on each sub-interval exactly. In contrast, the total error 
cannot be calculated exactly (for general \( u_0 \)), but it can be approximated very 
accurately by forming similarly splitting grid cells into sub-intervals and then 
using Gaussian quadrature on each sub-interval (we find that a 3-point Gauss 
rule suffices here).

Considering the evolution error in the right panel of Figure 2.5 one can 
observe that the special \( (\phi^\delta, \psi^\delta) \) approximation strategy (2.36) turns out to 
yield an exact method by means of creating fixed point solutions (see §2.4.6) 
right from the get go (the error (dark red) is essentially machine precision times 
the number of time steps taken). In contrast, the \( (\phi, \psi) \) approximation strategy 
yields smaller total errors, but the evolution errors (dark blue) are larger than 
with \( (\phi^\delta, \psi^\delta) \). However, they do decay rapidly as \( h \to 0 \). Moreover, the scheme 
appears to be exact nevertheless (the errors do not (up to machine precision) 
grow with \( t_f \)), albeit not via fixed point solutions right away. Comparing those 
results to the Lax-Wendroff method (left panel of Figure 2.5), we can see that 
the exactness property of the jet scheme results in substantially smaller errors, 
particularly for large final times (the black reference line is the same line in 
both panels).
2.4.3 Two equivalent interpretations of jet schemes

An important property of jet schemes is that they can be interpreted in two ways [76]: (a) an update rule in a finite-dimensional state space; or (b) an update rule of functions. Here we establish that duality for the nonlinear scheme defined above. For a given $m = 1/h$, let $U^{h,t_n} = [\phi_1^n, \psi_1^n, \phi_2^n, \psi_2^n, \ldots, \phi_m^n, \psi_m^n]^T \in \mathbb{R}^{2m}$ denote the state vector that defines the jet scheme approximation at time step $n$. Here $\phi_j^n$ and $\psi_j^n$ approximate values and derivatives, respectively, of the true solution at the grid point $x_j$. Moreover, let $\mathcal{H}$ denote the space of all continuous piecewise linear functions on $[0,1]$ with finitely many kinks. With these spaces, method 2.1 can be written as the update rule

$$U^{h,t_{n+1}} = \tilde{N}U^{h,t_n}, \quad \text{where} \quad \tilde{N} = \mathcal{E} \circ \mathcal{A} \circ \mathcal{I}.$$

Here $\mathcal{A} : \mathcal{H} \to \mathcal{H}$ is the true solution advection operator (which for problem (2.1) is simply the simple shift operator $Au(x) = u(x - a\Delta t)$). It is followed by the evaluation operator $\mathcal{E} : \mathcal{H} \to \mathbb{R}^{2m}$ that evaluates values and derivatives at the grid points with the convention that left-sided derivative values are taken at kinks (note that the particular choice of convention does change the method, but it does not compromise its exactness property); and preceded by the interpolation operator $\mathcal{I} : \mathbb{R}^{2m} \to \mathcal{H}$, defined on each grid cell $[x_j, x_{j+1}]$ by the nonlinear interpolant $P(x)$ introduced above.

At $t = 0$, an initial state vector $U^{h,0}$ must be defined that represents an approximation $\mathcal{I}(U^{h,0})$ to the true initial condition $u_0$. Methodologies to do so are discussed in §2.4.1. With this setup, conducting many steps of the jet scheme can be interpreted in two equivalent ways:

(a) $S^n = \cdots \cdots \mathcal{E} \mathcal{I} \mathcal{E} \mathcal{A} \mathcal{I} \mathcal{E} \mathcal{A} (U^{h,0})$, or (b) $S^n = \cdots \cdots \mathcal{E} \mathcal{A} \mathcal{I} \mathcal{E} \mathcal{A} \mathcal{I} \mathcal{E} \mathcal{A} (\mathcal{I}(U^{h,0}))$.

Here $\tilde{N} = \mathcal{I} \circ \mathcal{E} \circ \mathcal{A}$ is the induced jet scheme one-step operator acting between functions, i.e., the jet scheme is the true solution operator $\mathcal{A}$ followed by the operator $\mathcal{I} \mathcal{E}$. Note that for the linear jet schemes considered in [76], the operator $\mathcal{I} \mathcal{E}$ is a projection. This last property is generally not given for the
nonlinear jet scheme; however, it is not needed for the characterization of the $t \to \infty$ limits of the jet scheme.

In the next two subsections, we elaborate on two different ways to interpret the method, which will provide insights to understand the jet scheme’s exactness. First, in subsection §2.4.4, we outline the discrete interpretation of the method via a nonlinear mapping between finite-dimensional spaces. Then in subsection §2.4.5, we provide an interpretation of the method as a mapping between function spaces. To conclude, both the finite-dimensional update rule (a), which represents how the scheme is implemented, as well as the update rule of functions, (b), can be employed for the numerical analysis carried out in §2.4.6.

2.4.4 Jet scheme between discrete spaces: Fixed points as vectors

Thus far, the exactness of the jet scheme has been based on numerical evidence. Now we establish the exactness of the methods with the strategy (2.36) via rigorous arguments. Moreover, we provide a pathway towards a novel numerical analysis for the exactness of numerical methods (in one space dimension). Here we assume that the CFL number $\mu \in \mathbb{Q}$, and let $\mu = \frac{p}{q}$ with $p$ and $q$ co-prime. It is numerically observed that the restriction $\mu \in \mathbb{Q}$ can be relaxed, but for the analysis of the jet schemes done here, we stick to this mild assumption. On the periodic domain $[0, 1]$ with CFL number $0 < \mu < 1$ and $h = 1/m$, we have $\Delta t = \frac{\mu h}{a}$. Hence, after $N = \frac{1}{\mu h}$ time steps (as $\mu \in \mathbb{Q}$, we can choose $m$ such that $N \in \mathbb{N}$) the true solution has traveled a distance 1 and thus has returned to its initial configuration, i.e., $u(x, t + \frac{1}{a}) = u(x, t)$ considering periodicity. First, we consider the the $N$-step mapping from $\mathbb{R}^{2m}$ to itself defined by

$$
\tilde{\Psi}^h = \tilde{\mathcal{N}}^N : \mathbb{R}^{2m} \to \mathbb{R}^{2m}
$$

which approximates the identity map between two finite-dimensional state spaces. Of key interest are now the fix points of $\tilde{\Psi}^h$, because any such $S \in \mathbb{R}^{2m}$
for which $\tilde{\Psi}^h S = S$ constitutes an exact solution as a state vector. Define the set of all fix points of the map $\tilde{\Psi}^h$ by

$$F_\mu^h = \left\{ S \in \mathbb{R}^{2m} : \tilde{\Psi}^h S = S \right\}.$$ 

We now show that the nonlinear map has many fix points as vectors when the method is interpreted as an update rule between finite-dimensional state spaces. One way to see that the map $\tilde{\Psi}^h$ has plenty of fix points is the following.

At a given grid point $x_{j+1}$, we have the characteristic foot point $x_{ft} = x_{j+1} - a \Delta t = x_{j+1} - \mu h$, where $\mu$ is the fraction of a cell that the true solution moves per step. The update rule $\tilde{\mathcal{N}}$ is characterized by $\phi_{j+1}^{n+1} = P(x_R)$ and $\psi_{j+1}^{n+1} = P'(x_R)$, where $P(x)$ is the piecewise linear interpolant defined by the data $(\phi_j^n, \psi_j^n, \phi_{j+1}^n, \psi_{j+1}^n)$. There are exactly three possibilities: (L) $x_R$ is left of the kink position $c$; (R) $x_R > c$; and (F) the fallback case. In case (F), a new derivative value is generated, $\psi_{j+1}^{n+1} = \phi_{j+1}^n - \phi_j^n h$, which renders it hard to create a fix point. We therefore search for fix points in which only cases (L) or (R) happen. Then the nonlinear map $\tilde{\mathcal{N}}$ becomes a switched linear system, defined by

$$\phi_{j+1}^{n+1} = \phi_{j+1}^{n+1-\theta} + \psi_{j+1}^{n+1-\theta}(\theta - \mu)h \quad \text{and} \quad \psi_{j+1}^{n+1} = \psi_{j+1}^{n+1-\theta}.$$  

Here $\theta = 1$ in case (L) and $\theta = 0$ in case (R). Note that $\theta = \theta_j^n$ depends on the data at step $n$, so the update rule remains nonlinear. In finding (some) fix points $S$, we now make another restriction: in each step, all cells possess the same type of update, i.e., $\theta_1^n = \cdots = \theta_m^n$. Hence, the update rule becomes one of two matrices

$$\mathcal{N} = \begin{cases} M_L & \text{if } \theta^n = 1 \\ M_R & \text{if } \theta^n = 0 \end{cases},$$

where $M_L, M_R \in \mathbb{R}^{2m \times 2m}$ encode the rule (2.37). The matrices $M_L, M_R$ are given by
The commutator of the two matrices $M_L$ and $M_R$ is

$$[M_L, M_R] = M_L M_R - M_R M_L = 0.$$  (2.38)

**Lemma 2.2.** Let we are given $m$ number of grid points, CFL number $\mu$, velocity of the profile $a = 1$ and assume $N$ is the number of time steps it takes to reach from $t = 0$ to $t = 1$ with the time step size $\Delta t$. If a sequence of L’s at each cell and R’s at each cell happens with $m$ times L and $(N - m)$ times R for the scheme to reach an integer time from its previous integer time, then $M_L^m M_R^{N-m} = I_{2m \times 2m}$.

**Proof.** Computing $M_L^m$ and $M_R^{N-m}$ we get
where $d_1$ and $d_2$ are two block matrices of size $2 \times 2$ respectively given by

$$d_1 = \begin{bmatrix} 1 & m(1-\mu)h \\ 0 & 1 \end{bmatrix}, \quad d_2 = \begin{bmatrix} 1 & -(N-m)\mu h \\ 0 & 1 \end{bmatrix}$$

Now we claim $M_L^m M_{R}^{N-m} = I_{2m \times 2m}$, as $d_1 d_2 = I_{2 \times 2}$, because

$$-(N-m)\mu h + m(1-\mu)h = -N\mu h + mh = 0$$

as $N = \frac{1}{k} = \frac{m}{\mu}$.

A proper combination of these two matrices $M_L$ and $M_R$ in fact maps a state $S$ onto itself after $N$ steps.

What remains to verify is that there are in fact states $S$ that generate such a sequence of matrices (note that $\theta_j^n$ depends on the state at time $t_n$, which in turn depends on the updates taken before step $n$). Let us now see sufficient conditions for a vector $S$ to be a fixed point for the map $\tilde{\Psi}^h$. Consider the cell $[x_j^h, x_{j+1}^h] = [jh, (j+1)h]$. The lines passing through the end points of the cell are

$$y = \phi_j + \psi_j (x - jh), \quad \text{and}$$

$$y = \phi_{j+1} + \psi_{j+1} (x - (j+1)h).$$

The intersection point of these lines is given by

$$c_{j,j+1} = \frac{\phi_{j+1} - \phi_j + jh \psi_j - (j+1)h \psi_{j+1}}{\psi_j - \psi_{j+1}}.$$

In case (L) for all the cells, we should have $x_R \leq c_{j,j+1} \leq x_{j+1}$, $\forall \ j = 0, 1, 2, \cdots, m - 1$; where $x_R = (j + 1 - \mu)h$, $\phi_0 = \phi_m$, and $\psi_0 = \psi_m$, as
we have periodic boundary conditions. If we map these sub-intervals \([x_j^h, x_{j+1}^h]\) to \([0, 1]\) for each \(j\), then we get those conditions as

\[
1 - \mu \leq \frac{\phi_{j+1} - \phi_j - h\psi_{j+1}}{(\psi_j - \psi_{j+1})h} \leq 1, \quad \forall \ j = 0, 1, 2, \ldots, m - 1. \tag{2.39}
\]

Similarly, in case (R) we get the conditions as

\[
0 \leq \frac{\phi_{j+1} - \phi_j - h\psi_{j+1}}{(\psi_j - \psi_{j+1})h} \leq 1 - \mu, \quad \forall \ j = 0, 1, 2, \ldots, m - 1. \tag{2.40}
\]

Adapting the notations

\[
x \in [\alpha, \beta] \implies \begin{cases} 
\alpha \leq x \leq \beta, & \text{if } \alpha \leq \beta \\
\beta \leq x \leq \alpha, & \text{if } \beta \leq \alpha
\end{cases}
\]

we can write down the those conditions as follows:

\[
\frac{\phi_{j+1} - \phi_j}{h} \in \left[ (1 - \mu)\psi_j + \mu\psi_{j+1}, \ \psi_j + (1 - \theta)\psi_{j+1} \right], \quad \forall \ j = 0, 1, 2, \ldots, m - 1.
\]

\(\theta = 1\) gives the conditions for case (L) and \(\theta = 0\) corresponds to case (R).

Consider a sequence \(\{\theta^1, \theta^2, \ldots, \theta^N\}\) for the map \(\tilde{\Psi}^h\), where \(\theta^n \in \{0, 1\} \quad \forall \ n = 1, 2, \ldots, N\) and each \(\theta^n\) determines the case (either (L) or (R)) for all the cells at the time step \(n\). Any such sequence with \(m\) numbers of \(\theta^n\)'s being 1 and \((N - m)\) number of \(\theta^n\)'s being 0 in a certain order would correspond to a fixed point. All together it boils down to a set of interval constraints that the state vector \(S\) must satisfy to be a fix point of the map.

\[
G^h_{\mu} := \left\{ S \in \mathbb{R}^{2m} \left| \frac{\phi_{j+1} - \phi_j}{h} \in \left[ (1 - \mu)\psi_j + \mu\psi_{j+1} - (\psi_{j+1} - \psi_j) \sum_{\ell=1}^{n-1} (\theta_\ell - \mu), \right. \right. \\
\left. \left. \theta_i \psi_j + (1 - \theta_i n)\psi_{j+1} - (\psi_{j+1} - \psi_j) \sum_{\ell=1}^{n-1} (\theta_\ell - \mu) \right] \right\}.
\]

\(i = 1, \ldots, N, \ \forall \ j = 0, \ldots, m-1\)

This set \(G^h_{\mu}\) characterizes some fix points of \(\tilde{\Psi}^h = \tilde{N}^N\) via many inequality constraints. A geometric interpretation of these constraints is that the piecewise linear interpolant defined by \(S\) gets exactly moved a distance \(\mu h\) in each
step. We now show the existence of fix points; namely, we show by providing particular sequences of \( \{h_n\} \) and \( \{\theta^1, \theta^2, \cdots, \theta^N\} \), that the vector defined in (2.36) satisfies all the inequality constraints for a given CFL number \( \mu = \frac{p}{q} \), where \( p \) and \( q \) are co-prime. Consider \( m = p \). Then the number of time steps it will take to travel a distance of 1 will be \( N = q \). Now we divide each of \( m \) cells into \( q \) sub-cells, and we order them from the left as 1st, 2nd up to \((m \times q)\)-th sub-cells. According to the construction of the vector in (2.36), the kinks are in numbered \( q \times (1 : 1 : m) \) sub-cells. The foot points are on the right boundary of the sub-cells indexed by \( q \times (1 : 1 : m) - p \). As the foot points are left of the kink positions in all the cells, we get a case (L). So \( \theta^1 = 1 \). The update rule of the method tells that the function and the derivative values are just shifted from foot points to the grid points in the next time step. Furthermore, if we reconstruct the jet scheme interpolation, then we get the same piecewise linear function just shifted by a distance \( \frac{p}{nq} \). It is enough to focus on one cell to determine the rest of the entries of the sequence \( \{\theta^1, \theta^2, \cdots, \theta^q\} \) for \( m = q \), as the relative positions of foot points and kinks are the same for each cell.

We consider the first cell and observe that the kink is in the \( q \)-th sub-cell. We define this as our current relative kink position by \( k_r := q \). It is easy to see that if \( q - p < k_r \leq q \), then we get a case (L) and if \( 1 \leq k_r \leq q - p \) then we get a case (R). The current relative position of kink gets updated by \( k_r = \text{mod} (k_r + p, q) \) and we determine the sequence \( \{\theta^1, \theta^2, \cdots, \theta^q\} \) accordingly till we get \( k_r = q \). After \( q \) steps of the method, we get \( k_r = q \), as in each step, the kinks are shifted by a distance \( \frac{p}{mq} \). Now we consider the sequence \( h_n = \frac{n}{q} \), then for each of these resolutions we just repeat the sequence \( \{\theta^1, \theta^2, \cdots, \theta^q\} \) \( n \) times in order to travel a distance of 1. With this sequence of \( \theta \)'s the vector in (2.36) will satisfy the inequality constraints given above, as the sequence is designed such that it provides the right combination of \( m \) numbers of (L)'s and \((N - m)\) numbers of (R)'s. For example, take \( \mu = \frac{3}{4} \), and consider \( m = 3 \). Then we get the sequence \([L L L R]\) which will correspond to a fixed point (see Figure 2.9).
2.4.5 Jet scheme between function spaces: Fixed points as functions

Now we consider the $N$-step mapping from the function space $\mathcal{H}$ defined in §2.4.3 into itself

$$\Psi^h = \mathcal{N}^N : \mathcal{H} \to \mathcal{H}$$

that is an approximation to the identity map. Because we are now concerned with the $h \to 0$ convergence of the family of methods, we now use the superscript $h$ to denote the maps’ and state vectors’ dependence on the resolution. As motivated above, our goal is to characterize the functions that are fixed points of $\Psi^h$, because any such function $\mathcal{I}(U^h)$, where $U^h \in \mathbb{R}^{2m}$, that satisfies $\Psi^h \mathcal{I}(U^h) = \mathcal{I}(U^h)$ constitutes an exact solution. Moreover, given the family of maps $\Psi^h$, parameterized by $h$, we have an exact method if for a given true solution $u_0(x)$, a sequence $\{\mathcal{I}(U^h)\}_{h \to 0}$ of fixed points (i.e., $\Psi^h \mathcal{I}(U^h) = \mathcal{I}(U^h)$) exists that converges (in $L^1$) to $u_0$ as $h \to 0$.

Now we provide a condition for a function $\mathcal{I}(U^h)$ to be a fixed point (of the map $\Psi^h$) in terms of its kinks’ positions. The function $\mathcal{I}(U^h)$ can have at most one kink in the interior of each cell, and there could be kinks on the boundary of the cell. We divide each of the $m$ cells into $q$ sub-cells of size $h/q$, and we consider these resulting $mq$ sub-cells from left to right. We can now determine the positions of all the kinks of the function $\mathcal{I}(U^h)$. For kinks that fall on the boundary between adjacent sub-cells, we use the consistent convention to associate them to the left of those two sub-cells.

As established in §2.4.3, the one-step map $\mathcal{N}$ amounts to first shifting the function $\mathcal{I}(U^h)$ to the right by $\mu h$ via the true solution operator $\mathcal{A}$, followed by the operator $\mathcal{IE}$ that replaces the shifted function by the jet scheme interpolant based on its values and slopes evaluated at the grid points. Due to the construction of the jet scheme interpolant, the application of $\mathcal{IE}$ leaves the shifted interpolant $\mathcal{I}(U^h)(x - \mu h)$ from the prior step unchanged, except on those grid cells that contain two kinks (it is impossible for the shifted interpolant to have more than two kinks per cell). On those cells, the application of
IE replaces the shifted interpolant by a new genuine piecewise linear function with a single kink in the cell, or by the fallback function.

A key mechanism to finding fixed points of the $N$-step mapping $\Psi^h = N^h$ is to characterize functions for which such discrete changes in the approximating function do not occur. We define:

Definition 2.1. A piecewise linear jet scheme interpolant $I(U^h)$, defined by a state vector $U^h$, is called **defective** if there exists at least one adjacent pair of kinks that have less than $q - 1$ sub-cells in between them. Otherwise, it is called **non-defective**.

Lemma 2.3. If $I(U^h)$ is non-defective then it is a fixed point of the mapping $\Psi^h$.

*Proof.* As $I(U^h)$ is non-defective, it has at least $q - 1$ sub-cells in between any pair of adjacent kinks. Hence, it is impossible to ever generate a situation where two kinks fall within the same grid cell, if the function $I(U^h)$ is only shifted by integer multiples of the sub-cell width $h/q$. And in fact, that is the case, because by construction each update step $N$ shifts $I(U^h)$ by $\mu h = p^h$, i.e., by exactly $p$ sub-cells. $\blacksquare$

2.4.6 Numerical analysis for nonlinear jet scheme

Based on the intuitive geometric characterization of an important class of fixed points of the jet scheme in §2.4.5, we can easily analyze the two initial condition approximation strategies described in §2.4.1. While the $(\phi, \psi)$ approach generally generates defective interpolants (e.g., two adjacent grid cells may have kinks very close to their shared grid point), the $(\phi^\delta, \psi^\delta)$ method, defined in (2.36), always generates non-defective interpolants, because adjacent kinks are precisely a distance $h$ apart from each other.

As an illustration, consider $m = 3$ grid cells with CFL number $\mu = \frac{3}{4}$, i.e., traveling the distance 1 requires $N = 4$ steps. The $(\phi^\delta, \psi^\delta)$ strategy produces a piecewise linear function whose kinks are all a distance $\delta$ left of a grid point;
Figure 2.9: Domain with three cells. Black dots denote the grid points; blue dots denote the foot points; red dots are the sub-cell boundaries; green denotes the kink positions that are $\delta \ll h$ distance away from the grid points. The black arrow shows the next relative position of kink.

hence they all fall into the rightmost sub-cell of each grid cell. In turn, the foot points fall on the boundary between the first and second sub-cell, see Figure 2.9. Hence, $N = 4$ steps of the jet scheme amount to the sequence [L L L R], where (L)/(R) means that in each cell, the left/right-sided line of the jet scheme interpolant is used.

We are now in a position to prove the main theorem, establishing the exactness of the jet scheme with the specific approximation of the initial condition given by the equation (2.36).

**Theorem 2.2 (Exactness).** Method 2.1 with the strategy (2.36) of approximating the periodic initial function $u_0(x) \in C^0([0, 1])$, applied to problem (2.1), has the convergent infinite time limit property (2.2).

**Proof.** The initial vector $U^{h,0} = [\phi_1^0, \psi_1^0, \phi_2^0, \psi_2^0, \ldots, \phi_m^0, \psi_m^0]^T$ in equation (2.36) induces a piecewise linear interpolant $I(U^{h,0})$ which is exactly equal to $I(x + \delta)$ in equation (2.36), where $I(x)$ is the standard piecewise linear function that connects the grid point values. We have already shown that $I(U^{h,0})$ is a fixed point of the jet scheme. So for a given $h$, the numerical solution at any time step $t_n$ satisfies $I(U^{h,t_n})(x) = I(U^{h,0})(\text{mod}(x - n\mu h, 1))$. As established in §2.4.2, we consider the jet scheme’s total error as the $L^1$-distance of the
approximate function from the true solution:

\[ \varepsilon^{h,t_n} = \| I(U^{h,t_n}) - u^{t_n} \|_{L^1} \]

\[ = \| I(U^{h,0})(\text{mod}(x - n\mu h, 1)) - u_0(\text{mod}(x - n\mu h, 1)) \|_{L^1} \]

\[ = \| I(U^{h,0})(x) - u_0(x) \|_{L^1} = \| I(x + \delta) - u_0(x) \|_{L^1} \]

\[ \leq \| I(x + \delta) - u_0(x + \delta) \|_{L^1} + \| u_0(x + \delta) - u_0(x) \|_{L^1} \]

\[ \leq \omega(u_0, h) + \omega(u_0, \delta) , \]

where \( \omega(u_0, r) = \sup\{|u_0(x) - u_0(y)| : x, y \in [0, 1], |x - y| < r\} \) is the modulus of continuity of \( u_0 \in C^0([0, 1]) \). The last inequality arises as follows. On each sub-interval \([x_j, x_{j+1}]\), one has \( \int_{x_j}^{x_{j+1}} |I(x) - u_0(x)| \, dx \leq h \omega(u_0, h) \), and thus \( \| I(x + \delta) - u_0(x + \delta) \|_{L^1} \leq \omega(u_0, h) \); and by similar arguments \( \| u_0(x + \delta) - u_0(x) \|_{L^1} \leq \omega(u_0, \delta) \). Consequently, with \( \delta \ll h \) (assuming exact arithmetic here), we obtain that

\[ \lim_{h \to 0} \lim_{n \to \infty} \varepsilon^{h,t_n} \leq \lim_{h \to 0} [\omega(u_0, h) + \omega(u_0, \delta)] \leq \lim_{h \to 0} [\omega(u_0, h) + \omega(u_0, h)] = 0 , \]

where \( \omega(u_0, \delta) \leq \omega(u_0, h) \) follows from the \( r \)-monotonicity of \( \omega(u_0, r) \), and \( u_0 \) uniformly continuous on \([0, 1]\) implies \( \lim_{h \to 0} \omega(u_0, h) = 0 \). Moreover, it is easy to see that the scheme is convergent in the traditional sense. Hence, we have established the convergent infinite time limit property.

Remark 2.3. Obviously, in floating point arithmetic, there are practical limitations to the just established convergence notions. As in §2.4.1, one should choose \( \epsilon \ll \delta \ll h \) (where \( \epsilon \) is the machine precision); and the results in Figure 2.5 indicate that a linear accumulation of round-off errors could occur. Both aspects imply technical limitations on the \( h \to 0 \) as well as \( t_f \to \infty \) limits; however, those are not much different from what is incurred in other numerical methods. Moreover, the results in Figure 2.5 indicate that for practically relevant choices of \( h \) and \( t_f \), there are no detriments to the exactness of the jet schemes.
2.4.7 Other initial condition approximation: good news and some caveats

With Theorem 2.2 we have established the exactness of jet schemes, if the \((\phi^\delta, \psi^\delta)\) initial function approximation (2.36) is employed. The underlying reason is that the initial function \(I(U_{h,0})\) is a fixed point of the jet scheme update step. However, other approximations of the initial conditions may yield smaller error constants. For instance, Figure 2.5 shows that the \((\phi, \psi)\) strategy, which is natural if \(u_0 \in C^1\) as it is based on direct evaluations of the initial condition and its derivatives, turns out to generate second order convergent approximations with smaller errors than the \((\phi^\delta, \psi^\delta)\) strategy. A key drawback is that \((\phi, \psi)\) strategy generally yields an initial vector \(U_{h,0}^p \in \mathbb{R}^{2n}\), or function \(I(U_{h,0}^p)\), that is not a fixed point of the jet scheme. However, numerical evidence (Figure 2.5 and below) suggests that, for a wide range of sufficiently smooth initial functions, the jet scheme does always assume a fixed point after a finite number of update steps, and thus nevertheless gives rise to the convergent infinite time limit property (2.2).

To demonstrate this property, and its consequences, more distinctly, we present a numerical test of the jet scheme with the \((\phi, \psi)\) strategy up to a really large final time \(t_f = 100000\). Figure 2.10 shows the numerical results on \(m = 100\) grid points with CFL number \(\mu = 0.9\) at \(t = 100\), \(t = 1000\), and the final time. One can see that the jet scheme solution persists up to infinite time. Moreover, we observe (not shown here) that if the computational grid is refined, one always obtains exact solutions after finitely many steps, and they converge (as \(h \to 0\)) to the true solution. In other words, this scheme is exact for this problem.

Moreover, we compare the long-time evolution of jet schemes with other nonlinear (we already know from §2.2 that a crucial class of linear methods have no chance of converging at \(t \to \infty\)) classical high resolution methods: fifth-order WENO [38], Lax-Wendroff with van Leer limiter [46], and Lax Wendroff with Superbee limiter [46]. In Figure 2.10, the three snapshots in
time are complemented (bottom right panel) by the temporal evolution of the maximum of the numerical solutions. Clearly, all classical schemes keep on deforming in time, with the WENO and van Leer solutions being essentially constant at $t_f$, and the Superbee solution flattening on a very slow time scale. These results demonstrate the jet scheme’s superiority over the classical methods in terms of its long-time numerical behavior for constant-coefficient advection problems.

Figure 2.10: Example of long-time evolution of nonlinear jet scheme with $(\phi, \psi)$ initial condition approximation, for $u_t + u_x = 0$ with $\mu = 0.9$ on $m = 100$ grid points, in comparison with three classical nonlinear methods (‘LW’ stands for Lax-Wendroff). While the classical methods yield deformed long-time solutions, the jet scheme assumes (after finitely many steps) a fix point (i.e., an exact solution) that is a good approximation of the true solution. The bottom right panel shows the temporal evolution of the maximum, and it confirms the jet scheme’s superiority over the existing methods in capturing long-time evolution.

While these observations are very promising, it is actually not true that any possible initial vector $U^{h,0}$ will be transformed (by the jet scheme) into a fixed point. Here we establish a procedure to construct examples where the jet scheme does not assume a fixed point after a finite number of time steps.
To find those, we seek an eigenvector of the mapping $\Psi^h$ with associated
eigenvalue of modulus strictly less than 1. We do so by restricting the map
$\Psi^h$ to be of the form $M^N$, where $M$ is a specific one-step update rule that is
linear. Because in this case the jet scheme applies the same update rule in
every time step, it suffices to analyze just one step of the scheme.

Specifically, consider $m = 6$ and $\mu = \frac{3}{4}$, so that after $N = 8$ steps the
solution has traveled a distance 1. Now assume that the updates on the 6
cells are (from left to right, where 'L' denotes the case that the foot point
is left to the kink and 'F' denotes the fallback case): $[L\ L\ F\ L\ L\ F]$. It
is straightforward to check that the resulting matrix $M$ has a real eigenvalue
with modulus less than 1. Hence, if we start with the associated eigenvector
as initial condition, the numerical solution will exponentially decay to zero,
and it will never turn into a fixed point in finite time, as shown in Figure 2.11.

![Figure 2.11: Eigenvector of the one-step matrix $M$ corresponding to an eigen-
value with modulus less than 1. Left: initial condition. Right: solution at
t_{\ell} = 5, which is exactly same as the initial profile, but multiplied by $10^{-6}$.

Based on this example, we can also construct an infinite family of cor-
responding examples for larger grid sizes $m$ (multiples of 6), by simply con-
catenating the sequence $[L\ L\ F\ L\ L\ F]$. However, such functions are by
construction highly oscillatory (on the scale $h$). In contrast, $h \to 0$ sequences
of approximations to a smooth initial function are the opposite of such grid-
scale oscillatory functions; and in fact, decaying eigenvectors like the ones
constructed here never arose (for $h$ sufficiently small) in any test cases we
conducted. An intuitive explanation for this fact is as follows. As $h$ becomes
small, any smooth function locally is $O(h^3)$ close to a parabola on a grid cell
of size $h$. If one had an exact parabola (with nonzero quadratic part) on that
cell, then the associated piecewise linear jet scheme interpolant would have its
kink precisely in the center of the interval. Consequently, for the true func-
tion, kinks will be $O(h^3)$ away from the center of each cell, except for those
few cells in which the function has inflection points. As a consequence, for
$h$ sufficiently small, the approximation of a smooth initial condition will be
away from those oscillatory eigenvectors. That being said, even for tiny $h$, one
generally cannot expect to start with a fixed point of the jet scheme because
near inflection points, fallback cases may arise. However, as discussed above,
we do observe that for $h$ sufficiently small, a fixed point is generally assumed
after $O(1)$ steps (independent of $h$).

2.5 Beyond Constant-Coefficient Advection

The jet scheme with nonlinear interpolant has been constructed for 1D
constant-coefficient linear advection problems. Of course, in many practical
problems, transport or advection is just one component and often time it is
not constant-coefficient. While the jet scheme and the analysis for the method
established above rely on the fact that the problem is constant-coefficient linear
advection in 1D, the methodology can also be carried over to other problems,
such as variable-coefficient problems in 1D or problems in higher dimensions.
In this section, we apply the methodology to more general advection problems
and investigate to what extent the concepts have the potential to be generalized
to more complex situations.

2.5.1 Variable coefficient linear advection problems

As a first step, we consider variable coefficient advection (in 1D):

$$
\begin{cases}
  u_t + a(x, t)u_x = 0 \quad \text{on } [0, 1] \times (0, T], \\
  u_0(x) = \sin(2\pi x) + \sin(10\pi x),
\end{cases}
$$
with periodic boundary conditions and periodic velocity field \( a(x, t) = \frac{\sqrt{5}}{2}(0.9 + 0.1 \sin(2\pi x)) \), which is scaled so that the solution returns to its starting configuration exactly at integer times. This can be interpreted as a small droplet moving while undergoing small oscillatory deformations.

Figure 2.12: Long-time evolution of nonlinear jet scheme for variable coefficient advection problem. The numerical approximation is of high quality until \( t = 1000 \), but degrades eventually for much larger times.

The results with the nonlinear jet scheme on \( m = 100 \) grid points and CFL number \( \mu = 0.9 \) are shown in Figure 2.12. While the scheme produces very reasonable approximations up to \( t = 1000 \), the numerical solutions eventually degrade and become unstable. This means that the jet scheme methodology developed above does not carry over directly to variable coefficient problems. A plausible underlying reason is that piecewise linear functions are not exact solutions of the PDE.

### 2.5.2 Linear advection problems in multiple space dimensions

we consider constant coefficient advection (in 2D):
\[
\begin{aligned}
\left\{ u_t + a(\vec{x}, t) u_x = 0 \quad \text{on} \quad [0, 1]^2 \times [0, T], \\
u(x, y, 0) = u_0(x, y).
\end{aligned}
\] (2.41)

While linear Hermite interpolants generalize naturally to higher space dimensions [76], there is no canonical generalization of the piecewise linear jet scheme interpolant introduced in §2.4 to 2D and 3D. Hence, it is an open research question how to devise exact methods in higher dimensions, even in the constant coefficient case.

Here we test one particular way to extend the piecewise linear interpolant to 2D, as follows. Consider a rectangular grid cell with vertices \((a_1, b_1), (a_2, b_1), (a_1, b_2)\) and \((a_2, b_2)\), and let function values and gradients be given at those vertices. First, evaluate the 1D interpolants (§2.4) at the top and bottom edges, defined by the data \((\phi^{1,2}_x, \phi^{1,2}_x, \phi^{2,2}_x, \phi^{2,2}_x)\) and \((\phi^{1,1}_x, \phi^{1,1}_x, \phi^{2,1}_x, \phi^{2,1}_x)\), respectively, at \(x_t\) to yield \(\phi^{ft,\text{top}}\) and \(\phi^{ft,\text{bottom}}\). Second, at the same edges, use linear interpolants based on \((\phi^{1,2}_y, \phi^{2,2}_y)\) and \((\phi^{1,1}_y, \phi^{1,1}_y)\) to construct \(\phi^{ft,\text{top}}\) and \(\phi^{ft,\text{bottom}}\). Third, using the just constructed data, define another 1D interpolant along the line \(x = x_t\) and evaluate it at \(y_t\) to yield the value \(\phi^{x-y}\). Analogously, compute the value \(\phi^{y-x}\) via the same construction, but with the roles of \(x\) and

Figure 2.13: Nonlinear interpolant in 2D.
$y$ reversed. Then define the actual interpolation value as $\phi = \frac{1}{2}(\phi^{x \rightarrow y} + \phi^{y \rightarrow x})$.

This yields a nonlinear Hermite interpolant that is continuous. Because it is defined everywhere, it also defines an interpolant for $\nabla \phi$ (that is piecewise constant).

$$u_0(x,y) = \sin(2\pi x) + \sin(\pi y)$$

Figure 2.14: Numerical approximation with nonlinear jet scheme on 2D advection problem, for the initial condition $u_0(x,y) = \sin(2\pi x) + \sin(\pi y)$.

We test the resulting nonlinear jet scheme on the 2D advection problem (2.41) with velocity field $\vec{a}(\vec{x},t) = (1,2)^T$, on the periodic domain $[0,1]^2$. Using $m = 50$ grid points in each dimension, and with the initial condition $u(x,y,0) = \sin(2\pi x) + \sin(\pi y)$ we get the results shown in Figure 2.14.

The results look very promising. In fact, a numerical inspection reveals that the differences between the numerical states at $t = 1000$ and $t = 10000$ is on the order of machine precision. In other words, we have found an exact solution.

The problem is that the scheme itself is, unfortunately, not exact, as the results in Fig. 2.15 show, corresponding to the different initial condition
$u_0(x, y) = \sin(2\pi x) \sin(\pi y)$. While one still obtains an exact solution, it is not a reasonable approximation to the true solution. In other words, it remains an open research problem to find Hermite interpolants that generate exact jet schemes in higher space dimensions.

### 2.6 Conclusions and Outlook

The key takeaways from this study are that: (a) there is merit in conceptualizing infinite time limits of numerical methods for advection problems; and (b) schemes that possess convergent infinite time limits do exist. Conceptually, the focus on an infinite (or asymptotically large) time horizon is a departure from classical numerical analysis paradigms, and this work establishes new notions of numerical analysis and formalizes the infinite time convergence via the convergent infinite time limit property (2.2) and the concept of exact meth-
ods that possess convergent sequences of fixed point solutions. Moreover, in establishing these new concepts, traditional numerical analysis perspectives have been extended, such as considering error convergence in both mesh size $h$ and final time $t_f$, and the focus on long-time behavior of existing numerical methods.

The specific results and insights found in this work are as follows. First, it is formally shown that a widely used important class of linear methods for advection problems cannot (except for trivial cases) yield exact solutions, and thus they are unable to produce approximations that do not deteriorate in time. Second, the formal proof is complemented by asymptotic arguments, based on the modified equation, that establish the scaling of the truncation error $\mathcal{E}^{h,t_f}$ interpreted as a bivariate function of $h$ and $t_f$. Third, the jet scheme methodology with a specific nonlinear interpolant is introduced, including a duality between being an update rule in a finite-dimensional state space vs. in a function space. Fourth, exploiting that duality, we prove that the jet scheme, with a specific initial condition approximation, does indeed satisfy the stringent requirement of producing convergent infinite time solutions. And finally, the benefit of the jet schemes’ infinite time convergence, relative to existing widely used nonlinear methods, is illustrated in a comparative long-term computation example.

There is a broad variety of questions that are triggered by the concepts and results presented here. One important (open) question is whether there are traditional fixed-grid advection schemes that are exact. While we have proven that exactness is impossible to achieve with a class of linear methods, and have demonstrated numerically that various popular nonlinear methods all fail to be exact, this does not mean that there are no such methods. Certainly, the results in Figure 2.6 demonstrate that different finite volume limiters may result in vastly different long-time behavior of the numerical solutions. Specifically, while Lax-Wendroff with van Leer limiter clearly produces an essentially constant solution after some large but finite time ($t = 10^5$), Lax-Wendroff with Superbee limiter preserves a (deformed but) clearly non-constant profile even
after such a long time. This property plausibly has to do with the fact that
the Superbee limiter follows the uppermost boundary of the TVD region [46],
and it might hint at the existence of limiters that generate exact methods.

Another important question is to what extent the concepts established
herein generalize to more complicated transport problems. Clearly, it is a
long route from problem (2.1) to real application problems that may involve
(a) variable coefficients, (b) higher dimensions, and/or (c) coupling of trans-
port to other equations. Regarding (a): while the jet scheme methodology
can be applied to variable coefficient advection, the exactness properties fail
to carry over. At the same time, simple numerical tests indicate that for non-
rapidly varying coefficients, it does take a long time for the numerical solution
to degrade. Regarding (b): while linear Hermite interpolants generalize natu-
rally to higher space dimensions [76], there is no such canonical generalization
of the piecewise linear interpolant.

Finally, regarding (c), an important question for future research is to what
extent exact numerical methods can give rise to more accurate long-time so-
lutions for problems where the accurate resolution of linear advection is a
critical sub-problem, such as level set methods in fluid flow simulations or
contact discontinuities in compressible gas dynamics. Likewise, one may ask
whether some of the presented ideas and concepts can also be extended to
other problems with nonlinear traveling waves (solitons, diffusion-reaction,
advection-reaction).
CHAPTER 3

OPTIMAL WENO TYPE
FINITE VOLUME LIMITER FUNCTIONS

This chapter focuses on developing numerical methods to accurately capture the solutions’ structure for advection problems or hyperbolic conservation laws. Solving hyperbolic conservation laws accurately is an important step in many applied mathematics problems. High-order finite volume methods with limiter functions are applied to resolve many important structures of the solution, such as discontinuities or sharp gradients, encountered by the hyperbolic conservation laws. Many limiter functions have been introduced in the literature, and the general approach is to design a limiter function and demonstrate that it performs well on some test problems. We wish to investigate the inverse problem instead: given a representative test case portfolio and a cost functional, determine the optimal limiter function. Designing an optimal limiter function is a very generic and complicated problem. Hence, as a first step towards this study, we focus on finite volume methods with limiter functions that involve some artificial parameters. Specifically, we consider the work by Schmidtmann et al. [73] who analyzed 3rd order finite volume methods and the WENO3 [38] methods and presented a unifying view of 3rd order FV limiter
functions. They also improved the finite volume limiter by Čada and Torrilhon [13] by considering the limiter function in the 2D slope domain and presented a parameter-free limiter function, denoted as $H_{11}^{(c)}$. The concept of studying the limiter function in the 2D slope domain is possible when the reconstruction is based on only three cell averages. But this allows us to understand and overcome some of the fundamental shortcomings of 3rd order limiter functions. Considering the semi-discrete finite volume framework, we construct two FV methods designed in an ad-hoc fashion based on the third-order central WENO [38], and the WENO-Z [15] methods. We interpret these methods with limiter functions in the 2D slope domain as Schmidtmann et al. interpreted WENO3 method in [73]. The limiter functions in these proposed methods are characterized by two parameters that are chosen based on some trial and error method or in an ad-hoc fashion. Our goal is to find the optimal values of these parameters so that the proposed methods produce the best results for a variety of test problems. We determine these parameters by studying an inverse problem. Namely, considering a cost functional and some well-represented test cases, we determine the optimal parameters by minimizing the cost functional over the set of WENO type limiters that are characterized by those artificial parameters. Several test problems are solved by these limiter functions with optimal parameters and compared their numerical results with some other existing finite volume (FV) methods.

Consider a Cauchy problem of hyperbolic conservation laws in 1D

$$u_t + f(u)_x = 0 ,$$
$$u(x,0) = u_0(x), \ x \in \mathbb{R} .$$

(3.1)

Finite volume methods update the cell averages of the true solution at each time step to advance the numerical solution in time. Here we consider regular grid in space with space interval size $\Delta x$, and denote the $i$th cell by $C_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, having the cell center $x_i, i \in \mathbb{Z}$. The cell average at $(x,t)$ is
given by
\[ \bar{U}(x, t) = \int_{x - \frac{1}{2}}^{x + \frac{1}{2}} u(s, t) \, ds. \] (3.2)

We integrate equation (3.1) over a grid cell \( C_i \) and then divide by \( \Delta x \) to get
\[ - \frac{1}{\Delta x} \left( f(u(x_{i+\frac{1}{2}}, t)) - f(u(x_{i-\frac{1}{2}}, t)) \right), \] (3.3)

where \( \bar{U}_i(t) = \bar{U}(x_i, t) \). This is the integral form of the hyperbolic conservation laws, which is still exact. The goal is to find an approximate solution \( \bar{u}_i^n \) to the average of the true solution \( \bar{U}_i^n = \bar{U}(x_i, t^n) \) at \((x_i, t^n)\). To get an approximate solution, we require the fluxes at the cell boundaries. Since the exact values of the fluxes are not known, we construct numerical flux function \( \hat{f}(u, v) \) to approximate the true flux function, which is Lipschitz continuous and consistent with the true flux function. The fluxes at the cell boundaries are calculated by a numerical flux function depending on the approximate solutions at the cell boundaries. Then the fluxes at the cell boundaries are given by
\[ \hat{f}_{i+\frac{1}{2}} = \hat{f}(\hat{u}_{i+\frac{1}{2}}, \hat{u}_{i+\frac{1}{2}}) \approx f(u(x_{i+\frac{1}{2}}, t)), \] (3.4)
\[ \hat{f}_{i-\frac{1}{2}} = \hat{f}(\hat{u}_{i-\frac{1}{2}}, \hat{u}_{i-\frac{1}{2}}) \approx f(u(x_{i-\frac{1}{2}}, t)), \] (3.5)

where \( \hat{u}_{i+\frac{1}{2}} \) and \( \hat{u}_{i-\frac{1}{2}} \) are the approximations to the true solution at the cell boundaries. We drop the \( \hat{\cdot} \) for notational economy and define the left and right interface values based on three-point stencil calculated by the following functions:
\[ u^{(-)}_{i+\frac{1}{2}} := L(\bar{u}_{i-1}, \bar{u}_i, \bar{u}_{i+1}) , \] (3.6)
\[ u^{(+)}_{i-\frac{1}{2}} := R(\bar{u}_{i-1}, \bar{u}_i, \bar{u}_{i+1}) . \] (3.7)

Now the time evolution for the approximate cell averages becomes
\[ \frac{d\bar{u}_i}{dt} = - \frac{1}{\Delta x} \left( \hat{f}_{i+1/2} - \hat{f}_{i-1/2} \right) . \] (3.8)
The approximations at the cell boundaries are calculated by evaluating a reconstructed polynomial based on cell averages. The construction of the approximate solutions at the cell boundaries depending on the cell averages is responsible for the accuracy of the finite volume methods. The Piecewise constant reconstruction leads to a first-order accurate method, and high-order accurate methods are obtained by building higher order polynomial reconstruction functions. Our study only considers three-point stencil, and the best accuracy we can get is 3rd order linear methods via constructing quadratic polynomials based on three cell averages. This is true as long as the underlying solution function is smooth. Godunov’s theorem [28] says that a linear scheme on fixed grids creates spurious oscillations. A lower order reconstruction is used to avoid oscillations near the non-smooth region of the solution, characterized by limiter functions, thus making the schemes nonlinear. Many limiter functions [7, 33, 50, 38, 87, 69] have been introduced in the literature, and we discuss them briefly in the following section (see §3.1).

3.1 Limiter Functions in 1D

Methods with high-order polynomial reconstructions produce much better accuracy near the smooth regions than the first-order upwind method [47]. However, they produce oscillations near discontinuities. Limiter functions are applied with high-order accurate finite volume methods to limit the slopes near discontinuities or sharp transitions so that in the smooth regions, we gain high accuracy and avoid oscillations near discontinuities. Hence applying limiter functions produces overall accurate results. Limiter functions depend on a variable that is the ratio of two consecutive slopes at a grid, which can be considered as a measure of smoothness. The reconstructions $u_{i+\frac{1}{2}}^{(+)}$ and $u_{i-\frac{1}{2}}^{(-)}$ using three-point stencil with a limiter function $\phi$ in 1D are given by

\begin{align}
    u_{i+\frac{1}{2}}^{(-)} &= \bar{u}_i + \frac{1}{2} \phi(\theta_i) \delta_{i+\frac{1}{2}}, \\
    u_{i-\frac{1}{2}}^{(+)} &= \bar{u}_i - \frac{1}{2} \phi(\theta_i^{-1}) \delta_{i-\frac{1}{2}},
\end{align}

(3.9)
where the limiter function $\phi$ is a nonlinear function of $\theta_i = \frac{\delta_i}{\delta_{i+1/2}}$, with $\delta_{i+1/2} = \bar{u}_{i+1} - \bar{u}_i$, $\delta_{i-1/2} = \bar{u}_i - \bar{u}_{i-1}$. Linear and nonlinear methods can be written with the approximations at the cell boundaries formulated as in equation (3.9). Some of the examples of limiter function for linear methods read as

\[
\begin{align*}
\text{upwind} : & \quad \phi(\theta) = 0, \\
\text{Lax-Wendroff} : & \quad \phi(\theta) = 1, \\
\text{Beam-Warming} : & \quad \phi(\theta) = \theta, \\
\text{Fromm} : & \quad \phi(\theta) = \frac{1}{2}(1 + \theta).
\end{align*}
\]

Among these methods, upwind exhibits the first-order accuracy, and all other methods are second-order accurate. These second-order linear methods produce spurious oscillations because they violate the TVD [46] property (see Figure 3.1). The limiter functions are designed so that the reconstruction produces a higher order accurate approximation of the true solution in the smooth region and lower order approximation at the cell boundaries near discontinuity or sharp gradient. So the order of accuracy depends on the choice of limiter function $\phi$. There are many nonlinear finite volume schemes based on three-point stencil, such as Lax-Wendroff with the Superbee limiter, the Van Leer limiter (see §2.3 of chapter 2) or the MC limiter [88], can be rewritten in the formulation as in equation (3.9), where

\[
\begin{align*}
\text{superbee} : & \quad \phi(\theta) = \max(0, \min(1, 2\theta), \min(2, \theta)), \\
\text{MC} : & \quad \phi(\theta) = \max(0, \min(\frac{(1 + \theta)}{2}, 2, 2\theta)), \\
\text{Van Leer} : & \quad \phi(\theta) = \frac{\theta + |\theta|}{1 + |\theta|}.
\end{align*}
\]

These are second-order accurate methods by the following properties for limiters:

**Proposition 3.1.**

(i) If $\phi$ passes continuously through $\theta = 1$ with $\phi(1) = 1$, then the resulting scheme is at least 2nd order accurate in sufficiently smooth, monotonous regions of the solution.
(ii) If $\phi$ satisfies the conditions

$$0 \leq \phi(\theta) \leq \max(0, \min(2, 2\theta)),$$

the numerical scheme is total variation diminishing (TVD) [46], and thus does not create spurious oscillations.

The proof of this proposition can be found in [46]. The limiter functions in equation (3.11) lie in the Sweby region [45] of 2nd order methods. But they assume the value zero (different than the value taken by the 2nd order method) near the extrema where $\theta \approx -1$. As a result the accuracy drops to 1st order near extrema [45].

So far, we have seen 2nd order accurate methods based on the three-point stencil. It is possible to achieve 3rd order accurate methods using the three-point stencil by constructing a quadratic polynomial at each cell. In that case, the interface values are

$$u_{i+\frac{1}{2}}^{(-)} = \frac{1}{3} \bar{u}_{i+1} + \frac{5}{6} \bar{u}_i - \frac{1}{6} \bar{u}_{i-1}, \quad u_{i-\frac{1}{2}}^{(+)} = \frac{1}{3} \bar{u}_{i-1} + \frac{5}{6} \bar{u}_i - \frac{1}{6} \bar{u}_{i+1}. \quad (3.12)$$
These can also be rewritten as in equation (3.9) with

\[
\phi(\theta_i) = \frac{2 + \theta_i}{3}.
\]

It is well-known that this method gives 3rd order accurate solution in smooth regions but produces oscillations near discontinuities or sharp gradients. In these classical schemes, the limiter functions have been designed such a way that they produce low order approximation near smooth minima or maxima, which is not desirable. Artebrant and Schroll [7] introduced a limiter function \( \phi_{AS}(\theta_i, q) \) with \( q = 1.4 \), that assumes a value same as the third-order limiter function at \( \theta = -1 \). Thus this limiter function produces 3rd order approximation near smooth extremums. This limiter function’s only drawback is its complexity, which makes evaluating this function expensive for each cell. The

Figure 3.2: The light shaded regions show the TVD region. The deep shaded regions show the Sweby region of 2nd order methods. \( \phi_3 \): Full 3rd order limiter function; \( \phi_{AS} \): limiter function given in (3.13); and \( \phi_{CT} \): limiter function given in (3.14). The figure shows that these limiter functions do not lie in the TVD regions entirely.
formula for this limiter reads as
\[
\phi_{AS}(\theta, q) = \frac{2p [(p^2 - 2p\theta_i + 1) \log(p) - (1 - \theta_i) (p^2 - 1)]}{(p^2 - 1) (p - 1)^2},
\]
\[p = p(\theta_i, q) = \frac{|\theta_i|^q}{1 + |\theta_i|^{2q}}. \tag{3.13}\]

Čada and Torrilhon [13] later developed a limiter function \(\phi_{CT}(\theta_i)\) based on \(\phi_{AS}(\theta_i, q)\)
\[
\phi_{CT}(\theta_i) = \max \left(0, \min \left(\phi_3(\theta_i), \max \left(-\frac{1}{2} \theta_i, \min (2\theta_i, \phi_3(\theta_i), 1.6)\right)\right)\right)\) \tag{3.14}
that reduces the computational cost. Note that these limiter functions \(\phi_{AS}(\theta_i, q)\) or \(\phi_{CT}(\theta_i)\) do not lie in the TVD region (see Figure 3.2) around the neighborhood of \(\theta_i = -1\), and the reason for this was to achieve higher order accuracy near the extremums. The TVD version of \(\phi_{CT}\) limiter function is given by (see Figure 3.2)
\[
\phi_{CT,\text{TVD}}(\theta) = \max (0, \min (2\theta, \phi_3(\theta), 1.6)). \tag{3.15}\]

However, it is not necessary that discretizing the domain always gives \(\theta_i = -1\) near an extremum. One of the consecutive slopes might be approximately zero or small, depending on the domain’s discretization. Then the ratio of these slopes will be close to either 0 or \(\pm \infty\), and hence the limiter functions will be close to zero, resulting in the low order approximation to the true solution. To avoid this phenomenon, one needs a criterion that can separate smooth extremum from discontinuity or steep gradient. In order to do that, a parameter depending on two consecutive slopes \((\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}})\) is introduced as a measure of smoothness. Čada and Torrilhon proposed the function
\[
\eta_{CT} \left(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}\right) = \frac{\delta_{i-\frac{1}{2}}^2 + \delta_{i+\frac{1}{2}}^2}{(r \Delta x)^2}, \tag{3.16}\]
that defines an asymptotic region of radius \(r\) around the origin in \((\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}})\)-plane. The value of \(r\) is chosen in an ad-hoc fashion in [13], and then the limiter
function $\phi_{CT}$ is modified to a limiter function defined by

$$\phi_{CT}^{(c)}(\theta_i) := \begin{cases} 
\phi_3(\theta_i) & \text{if } \eta_{CT} < 1 \\
\phi_{CT}(\theta_i) & \text{if } \eta_{CT} \geq 1,
\end{cases}$$

(3.17)

to include the asymptotic region around the origin, where the third order reconstructions are assumed by the limiter function. This limiter function has been widely used in the literature, for example in [39, 41, 55].

Later, Schmidtmann et al. [73] considered the limiter function and the smoothness indicator function as functions of two consecutive slopes in the 2D slope domain, and proposed a parameter free smoothness indicator function and a limiter function based on $\phi_{CT}$. The reason for considering the limiter function in a 2D slope domain is that it is too restrictive to consider the limiter function as a function of the ratios of successive slopes, as we discard the magnitude of the slopes. In the next section, we discuss the 3rd order finite volume methods with limiter functions interpreted in the 2D slope domain.

### 3.2 3rd Order FV Methods with Limiters in the 2D Slope Domain

Considering the limiter $\phi$ as a function of two parameters, we reformulate the equation (3.9) as

$$u_{i+\frac{1}{2}}^{(-)} = \bar{u}_i + \frac{1}{2}H(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}),$$  
$$u_{i-\frac{1}{2}}^{(+)} = \bar{u}_i - \frac{1}{2}H(\delta_{i+\frac{1}{2}}, \delta_{i-\frac{1}{2}}).$$

(3.18)

In this framework, the third-order limiter function $\phi_3(\theta_i) = \frac{2+\theta_i}{3}$ becomes,

$$H_3(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}) = \frac{2\delta_{i+\frac{1}{2}} + \delta_{i-\frac{1}{2}}}{3}.$$  

(3.19)

The limiter function $\phi_{CT}$ can also be reformulated in the 2D slope domain and let us denote it by $H_{CT}$. One downside of this limiter function $H_{CT}$ is that it does not treat similar smoothness or non-smoothness equally. Namely,
\[ H_{CT}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}) = H_3(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}), \text{ but } H_{CT}(-\delta_{i-\frac{1}{2}}, -\delta_{i+\frac{1}{2}}) \neq H_3(-\delta_{i-\frac{1}{2}}, -\delta_{i+\frac{1}{2}}), \]

which is not natural. The authors in [73] corrected this feature and defined a new limiter function \( H_{3L}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}) \) given by

\[
H_{3L} = \text{sgn} \left( \delta_{i+\frac{1}{2}} \right) \max \left( 0, \min \left( \text{sgn} \left( \delta_{i+\frac{1}{2}} \right) H_3, \max \left( -\text{sgn} \left( \delta_{i+\frac{1}{2}} \right) \delta_{i-\frac{1}{2}}, \min \left( 2 \text{sgn} \left( \delta_{i+\frac{1}{2}} \right) \delta_{i-\frac{1}{2}}, \text{sgn} \left( \delta_{i+\frac{1}{2}} \right) H_3, 1.5 \left| \delta_{i+\frac{1}{2}} \right| \right) \right) \right).
\] (3.20)

They also defined a parameter free smoothness indicator

\[
\eta = \eta \left( \delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}} \right) = \frac{\sqrt{\delta_{i-1/2}^2 + \delta_{i+1/2}^2}}{\sqrt{\frac{5}{2} \alpha \Delta x^2}} \] (3.21)

with

\[
\alpha = \max_{x \in \Omega \setminus \Omega_d} |u_0''(x)|, \] (3.22)

where \( \Omega \) is the computational domain, and \( \Omega_d \) is the set of points where the initial condition is discontinuous. Following the equation (3.17), the combined
limiter function reads as

\[ H^{(c)}_{3L} \left( \delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}} \right) := \begin{cases} H_3 \left( \delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}} \right) & \text{if } \eta < 1 \\ H_{3L} \left( \delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}} \right) & \text{if } \eta \geq 1. \end{cases} \] (3.23)

This limiter function produces a 3rd order reconstruction when the consecutive slopes are in a neighborhood of the origin, i.e., close to an extremum, otherwise assumes a lower order reconstruction. The proposed two-dimensional limiter function in the equation (3.23) is shown in Figure 3.3.

### 3.3 Limiter Functions Induced by third-Order WENO Type Methods

Third-order WENO type methods can be reformulated as finite volume methods with limiters in the 2D slope domain. This new perspective allows us to understand the role of parameters presented in these methods better. First, we will discuss the traditional WENO3 limiters in this new framework [73]. Then we will present and analyze two new limiter functions based on 3rd order central WENO and WENO-Z methods.

#### 3.3.1 WENO3 limiter

Third-order weighted essentially non-oscillatory (WENO3) is a family of numerical schemes that use the three-point stencil for the reconstruction of the cell interface values. The three-point stencil allows us to interpret WENO3 methods with limiters in the two parameter framework. Schmidtmann et al. [73] rigorously showed that WENO3 methods can be interpreted as finite volume methods with limiter functions in the 2D slope domain. Here we briefly review the interpretation of WENO3 methods with limiters in the 2D framework. We divide the three-point stencil \( C_{i-1} \cup C_i \cup C_{i+1} \) into two two-point stencils \( S_0 = C_{i-1} \cup C_i \) and \( S_1 = C_i \cup C_{i+1} \). Then define second-order accurate polynomials \( p_0(x) \) and \( p_1(x) \) on \( S_0 \) and \( S_1 \), respectively, based on the...
cell averages $\bar{u}_{i-1}, \bar{u}_{i}, \bar{u}_{i+1}$. To get the right cell interface value by WENO3 reconstruction, we evaluate $p_0(x)$ and $p_1(x)$ at $x_{i+\frac{1}{2}}$ and take the convex combination

$$u_{i+\frac{1}{2}} = w_{i-\frac{1}{2}} p_0(x_{i+\frac{1}{2}}) + w_{i+\frac{1}{2}} p_1(x_{i+\frac{1}{2}}),$$

(3.24)

where $w_{i\pm\frac{1}{2}} \geq 0$, $w_{i-\frac{1}{2}} + w_{i+\frac{1}{2}} = 1$, $p_0(x_{i+\frac{1}{2}}) = -\frac{3}{2} \bar{u}_{i-1} + \frac{3}{2} \bar{u}_{i}$, and $p_1(x_{i+\frac{1}{2}}) = \frac{1}{2} \bar{u}_{i} + \frac{1}{2} \bar{u}_{i+1}$. If we choose $w_{i-\frac{1}{2}} = \gamma_{i-\frac{1}{2}} = \frac{1}{3}$ and $w_{i+\frac{1}{2}} = \gamma_{i+\frac{1}{2}} = \frac{2}{3}$ then $u_{i+\frac{1}{2}}$ produces 3rd order accurate approximation to $u(x_{i+\frac{1}{2}})$. WENO3 chooses $w_{i\pm\frac{1}{2}} \approx \gamma_{i\pm\frac{1}{2}}$ in the smooth regions, and near discontinuity or steep gradient, it puts weights to the polynomials such that the scheme produces less oscillations. The weights are defined by

$$w_{i\pm\frac{1}{2}} = \frac{\alpha_{i\pm\frac{1}{2}}}{\alpha_{i-\frac{1}{2}} + \alpha_{i+\frac{1}{2}}},$$

(3.25)

where

$$\alpha_{i\pm\frac{1}{2}} = \frac{\gamma_{i\pm\frac{1}{2}}}{(\epsilon + \beta_{i\pm\frac{1}{2}})^{p}};$$

(3.26)

and $\beta_{i\pm\frac{1}{2}}$ are some smoothness indicators defined by

$$\beta_{i\pm\frac{1}{2}} = (\delta_{i\pm\frac{1}{2}})^2.$$  

(3.27)

Substituting $p_0(x_{i+\frac{1}{2}})$ and $p_1(x_{i+\frac{1}{2}})$ into the equation (3.24), we get

$$u_{i+\frac{1}{2}} = \bar{u}_{i} + \frac{1}{2} \left( w_{i-\frac{1}{2}} \delta_{i-\frac{1}{2}} + w_{i+\frac{1}{2}} \delta_{i+\frac{1}{2}} \right).$$

(3.28)

Comparing equation (3.18) with equation (3.28) and denoting the WENO3 limiter function $H_{\text{WENO3}}$ in the two parameter framework, we obtain

$$H_{\text{WENO3}}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}) = w_{i-\frac{1}{2}} \delta_{i-\frac{1}{2}} + w_{i+\frac{1}{2}} \delta_{i+\frac{1}{2}}.$$  

Using equations (3.25)-(3.27), we obtain the formula for $H_{\text{WENO3}}$ that reads

$$H_{\text{WENO3}}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}) = \frac{1}{3} \left( \epsilon + (\delta_{i-\frac{1}{2}})^2 \right)^{-p} \delta_{i-\frac{1}{2}} + \frac{2}{3} \left( \epsilon + (\delta_{i+\frac{1}{2}})^2 \right)^{-p} \delta_{i+\frac{1}{2}}.$$  

(3.29)

Jiang and Shu’s original work [38] recommended that $p$ is 2 and the other parameter $\epsilon$ is taken to be a tiny positive number $10^{-6}$. Near the origin $(0, 0)$,
i.e. for $|\delta_{ij}| \ll \epsilon$, the function behaves like the 3rd order limiter function equation (3.19), and away from the origin, it is a homogeneous function, i.e., along any line through origin, the limiter function is linear. Casting WENO3 method in the 2D slope domain gives us much more insight on understanding how this method performs. Schmidtmann et al. [73] compared their proposed limiter function $H_{3L}^{(c)}$ with $H_{WENO3}$ by conducting many numerical examples. They also demonstrated that this approach helps to understand how to choose the parameter $\epsilon$ depending on the discretization of the domain of the problem.

### 3.3.2 CWENO3 and CWENO3-Z limiters

In our work, we construct two different limiters using the formulation developed in [73]; one based on the classical 3rd order central WENO method [48], and another based on combining 3rd order central WENO and the weights according to the WENO-Z method [15, 26]. We denote these new methods by CWENO3 and CWENO3-Z. The reconstruction of these methods involves another parameter (denoted by $\sigma$) other than the parameter $\epsilon$. In practice, this constant $\sigma$ is chosen to be a number from the interval $(0, 1)$. Once we reformulate the reconstruction of CWENO3 and CWENO3-Z methods in the 2D framework, we see the corresponding limiter functions $H_{CWENO3}$ and $H_{CWENO3-Z}$ both have two parameters $\epsilon$ and $\sigma$. Our main goal is to study how these methods perform on different problems with a particular set of interests, depending on the parameters $\epsilon$ and $\sigma$. Specifically, we want to determine the optimal parameters by minimizing a cost functional and understand the fundamental role of those parameters in these methods.

First, we start by briefly reviewing the 3rd order CWENO reconstruction. The reconstruction in the $i$th cell is based on two linear polynomials $P_L(x)$ and $P_R(x)$ on the stencils $S_1 = C_{i-1} \cup C_i$ and $S_2 = C_i \cup C_{i+1}$, respectively, and a central parabola $P_0(x)$ that are given by

$$P_L(x) = \bar{u}_i + \frac{\bar{u}_i - \bar{u}_{i-1}}{\Delta x} (x - x_i), \quad (3.30)$$
\[ P_R(x) = \bar{u}_i + \frac{\bar{u}_{i+1} - \bar{u}_i}{\Delta x} (x - x_i), \] (3.31)

and

\[ P_0(x) = \frac{1}{d_0} \left( P_{OPT}(x) - d_L P_L(x) - d_R P_R(x) \right), \] (3.32)

where

\[ P_{OPT} = \bar{u}_i - \frac{1}{24} (\bar{u}_{i+1} - 2\bar{u}_i + \bar{u}_{i-1}) + \frac{\bar{u}_{i+1} - \bar{u}_{i-1}}{2\Delta x} (x - x_i) + \frac{\bar{u}_{i+1} - 2\bar{u}_i + \bar{u}_{i-1}}{2\Delta x^2} (x - x_i)^2, \] (3.33)

\[ \bar{u}_{i+1} - 2\bar{u}_i + \bar{u}_{i-1} \]

is the optimal 3rd order accurate central parabola based on the cell averages on the stencil \( C_{i-1} \cup C_i \cup C_{i+1} \). The constants \( d_0, d_L, d_R \) are chosen such that they are strictly positive and \( d_0 + d_L + d_R = 1 \). Originally, it was proposed that users can choose any value for \( d_0 \) from \((0, 1)\), and the other two values are chosen such that \( d_L = d_R = (1 - d_0)/2 \). The popular choices of \( d_0 \) are \( \frac{1}{2}, \frac{3}{4} \) or 0.99, which makes \( d_L = d_R = \frac{1}{4}, \frac{1}{8} \), or 0.005, respectively. We take \( d_L = d_R := \sigma \) and leave it as a free parameter from \((0, 1)\) and, hence \( d_0 = 1 - 2\sigma \). Here, we will determine the optimal value of \( \sigma \) depending on the particular criteria according to the users’ specifications.

The reconstruction at the right cell interface of the \( i \)th cell is given by

\[ u_{i+\frac{1}{2}} = w_{L,i} P_L(x_{i+\frac{1}{2}}) + w_{0,i} P_0(x_{i+\frac{1}{2}}) + w_{R,i} P_R(x_{i+\frac{1}{2}}), \] (3.35)

where

\[ w_{k,i} = \frac{\alpha_{k,i}}{\alpha_{L,i} + \alpha_{0,i} + \alpha_{R,i}} \quad \text{and} \quad \alpha_{k,i} = \frac{d_k}{(\epsilon + \beta_{k,i})^2}, \quad \text{for} \quad k = L, 0, R. \] (3.36)

The smoothness indicators are defined by

\[ \beta_{L,i} = (\bar{u}_i - \bar{u}_{i-1})^2, \]
\[ \beta_{0,i} = \frac{13}{12} (\bar{u}_{i+1} - 2\bar{u}_i + \bar{u}_{i-1})^2 + \frac{1}{4} (\bar{u}_{i+1} - \bar{u}_{i-1})^2, \] (3.37)
\[ \beta_{R,i} = (\bar{u}_{i+1} - \bar{u}_i)^2. \]

The parameter \( \epsilon \) was initially introduced to avoid division by zero, but in our discussion, we keep \( \epsilon \) as free parameter (like \( \sigma \)) to be chosen later.
Figure 3.4: Each surface shows the plot of \((H_{CWENO3} - H_3)\) function for different choices of reasonable parameter values, i.e., the third-order limiter function is subtracted off from the \(H_{CWENO3}\) limiter function. So the \(H_{CWENO3}\) limiter function is third-order wherever the values are zeros in the plots.

With the notation \(\delta_{i+\frac{1}{2}} = \bar{u}_{i+1} - \bar{u}_i\) and \(\delta_{i-\frac{1}{2}} = \bar{u}_i - \bar{u}_{i-1}\) introduced before, we can write the equation (3.35) as follows

\[
u_{i+\frac{1}{2}} = \bar{u}_i + \frac{1}{2} H_{CWENO3}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}),
\]  
(3.38)
where the CWENO3 limiter function $H_{\text{CWENO3}}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}})$ reads as

$$
\frac{\left[\sigma \frac{1}{(\epsilon + \beta_{L,i}^2)^2} + \left(\frac{1}{2} - \sigma\right)\frac{1}{(\epsilon + \beta_{0,i}^2)^2}\right] \delta_{i-\frac{1}{2}} + \left[\sigma \frac{1}{(\epsilon + \beta_{R,i}^2)^2} + \left(\frac{1}{2} - \sigma\right)\frac{1}{(\epsilon + \beta_{0,i}^2)^2}\right] \delta_{i+\frac{1}{2}}}{\left(1 - 2\sigma\right)\left(\epsilon + \beta_{0,i}^2\right)^2 + \sigma \frac{1}{\left(\epsilon + \beta_{L,i}^2\right)^2} + \sigma \frac{1}{\left(\epsilon + \beta_{R,i}^2\right)^2}},
$$

and the smoothness indicators become

$$
\beta_{L,i} = \left(\delta_{i-\frac{1}{2}}\right)^2,
\beta_{0,i} = \frac{13}{12}(\delta_{i+\frac{1}{2}} - \delta_{i-\frac{1}{2}})^2 + \frac{1}{4}(\delta_{i+\frac{1}{2}} + \delta_{i-\frac{1}{2}})^2,
\beta_{R,i} = \left(\delta_{i+\frac{1}{2}}\right)^2.
$$

We view the CWENO3 limiter function as a two-parameter ($\epsilon$ and $\sigma$) family of functions defined in the 2D slope domain. The Figure 3.4 shows the shape of different $H_{\text{CWENO3}}$ limiter functions for different choices of the pairs ($\epsilon$ and $\sigma$). We can derive that the limiter function acts like 3rd order limiter function near the origin $(0, 0)$, i.e., for $|\delta_{i\pm\frac{1}{2}}| \ll \epsilon$

$$
\frac{1}{3}\delta_{i-\frac{1}{2}} + \frac{2}{3}\delta_{i+\frac{1}{2}},
$$

and away from the origin, i.e., for $|\delta_{i\pm\frac{1}{2}}| \gg \epsilon$

$$
\frac{1}{3}\beta_{L,i}^2 + \left(\frac{1}{2} - \sigma\right)\beta_{0,i}^2, \quad \beta_{R,i}^2,
$$

is a homogeneous function of degree one.

Similar derivation with the choices of $\alpha_{k,i}$’s

$$
\alpha_{k,i} = d_k \left(1 + \frac{\tau}{\epsilon + \sigma}\right), \quad \text{for } k = L, 0, R, \quad \text{with } \tau = \left(\delta_{i+\frac{1}{2}} - \delta_{i-\frac{1}{2}}\right)^2
$$

according to WENO-Z method will give rise to CWENO3-Z limiter function $H_{\text{CWENO3-Z}}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}})$, which is given by
\[ H_{\text{CWENO3-Z}}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}) = \left[ \sigma \left( 1 + \frac{\tau}{\epsilon + \beta_{L,i}} \right) + \left( \frac{1}{3} - \sigma \right) \left( 1 + \frac{\tau}{\epsilon + \beta_{0,i}} \right) \right] \delta_{i-\frac{1}{2}} + \left[ \sigma \left( 1 + \frac{\tau}{\epsilon + \beta_{R,i}} \right) + \left( \frac{2}{3} - \sigma \right) \left( 1 + \frac{\tau}{\epsilon + \beta_{0,i}} \right) \right] \delta_{i+\frac{1}{2}} \right] \frac{1}{\sigma \left( 1 + \frac{\tau}{\epsilon + \beta_{L,i}} \right) + (1 - 2\sigma) \left( 1 + \frac{\tau}{\epsilon + \beta_{0,i}} \right) + \sigma \left( 1 + \frac{\tau}{\epsilon + \beta_{R,i}} \right)}. \]

Figure 3.5: Each surface shows the plot of \((H_{\text{CWENO3-Z}} - H_3)\) function for different choices of reasonable parameter values, i.e., the third-order limiter function is subtracted off from the \(H_{\text{CWENO3-Z}}\) limiter function. So the \(H_{\text{CWENO3-Z}}\) limiter function is third-order wherever the values are zeros in the plots.
For $|\delta_{i+\frac{1}{2}}| \ll \epsilon$, the limiter function $H_{CWENO3-Z}$ has the third-order limiter function as the leading order term, and for $|\delta_{i+\frac{1}{2}}| \gg \epsilon$, one can derive that it has the leading order term $H_{\gg}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}})$ given by

$$H_{\gg}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}) = \left[ \sigma \left(1 + \tau \beta_{L,i}^{-1}\right) + \left(\frac{1}{3} - \sigma\right) \left(1 + \tau \beta_{0,i}^{-1}\right) \right] \delta_{i-\frac{1}{2}}$$

$$+ \left[ \sigma \left(1 + \tau \beta_{R,i}^{-1}\right) + \left(\frac{2}{3} - \sigma\right) \left(1 + \tau \beta_{0,i}^{-1}\right) \right] \delta_{i+\frac{1}{2}}$$

$$\frac{\sigma \left(1 + \tau \beta_{L,i}^{-1}\right) + (1 - 2\sigma) \left(1 + \tau \beta_{R,i}^{-1}\right) + \sigma \left(1 + \tau \beta_{0,i}^{-1}\right)}{\sigma \left(1 + \tau \beta_{L,i}^{-1}\right) + (1 - 2\sigma) \left(1 + \tau \beta_{R,i}^{-1}\right) + \sigma \left(1 + \tau \beta_{0,i}^{-1}\right)} \Delta.$$  

$H_{\gg}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}})$ is a homogeneous function of degree one, i.e.,

$$H_{\gg}(k\delta_{i-\frac{1}{2}}, k\delta_{i+\frac{1}{2}}) = kH_{\gg}(\delta_{i-\frac{1}{2}}, \delta_{i+\frac{1}{2}}).$$

Thus both the limiter functions $H_{CWENO3}$ and $H_{CWENO3-Z}$ follow the behavior of traditional finite volume limiter functions. They have two parameters, namely, $\epsilon$ and $\sigma$. It is not clear from the literature how to exactly choose any of these two parameters. Our goal is to find the right pair of these parameters so that the scheme performs better in some sense. In general, the goal is to design numerical schemes that can give better accuracy in the smooth region, especially near extremums, and no overshoot or undershoot near discontinuity or sharp gradient. So we want our numerical scheme that optimizes the numerical error and undershoot or overshoot. It is not straightforward how to design a cost functional that incorporates these errors for any initial condition. In the next section, we discuss a particular way of designing a cost functional depending on one specific test case.

### 3.4 Learning Optimal Parameters by an Inverse Problem

Up till now, we have considered the existing methods and combined them in an ad-hoc fashion to design new methods with parameters that were given to be fixed numbers in the literature. In past research, the parameter values
were chosen by a combination of intuition, trial and error, or in an ad-hoc fashion [50, 48]. Then it was demonstrated that those choices for parameters yield satisfactory numerical results. Here as the first step, we provide insights into how the new limiter functions behave depending on those parameters’ different choices. We then invert the traditional approach and learn the optimal choices for the parameters by solving an inverse problem: Determine the optimal limiters by optimizing an objective function that quantifies a suitable combination of application error and accurately capturing the solution’s structure for a set of well-chosen features. We choose one particular test case to demonstrate how one can design an objective function or a cost functional. Consider the linear advection problem, \( u_t + u_x = 0 \) with the initial condition
\[
u_0(x) = \max\left(\min\left(\frac{x}{0.1} - 2, -\left(\frac{x}{0.1} - 2\right) \right) + 1, 0\right) + \exp\left(-\left(\frac{x - 0.7}{0.15}\right)^4\right) \sin(30\pi x).\]

This initial condition combines smooth and non-smooth features having sharp transitions. Therefore the linear advection with this initial condition is a well-
represented test problem to test the potentials of limiter functions. As we
know the true solution, we can measure the undershoot or overshoot (defined
in (3.44)) for this particular test case. We want to design a cost functional
that can take numerical error, overshoot, or undershoot error into account.
For the linear advection equation with the particular initial condition in equation (3.43), we define a cost functional as

$$E(\epsilon, \sigma) = \int_0^1 |U - u_0| \, dx + c_1 \int_{\frac{1}{2}}^1 |U - u_0| \, dx + c_2 \int_0^{\frac{1}{2}} \min\{U, 0\} \, dx.$$  (3.44)

Here $U$ is the numerical solution at the final time $t_f \in \mathbb{N}$. The first integral in
the cost functional represents the $L_1$ error on the non-smooth part (referred
to 1st half $L_1$ error in Figure 3.7), the second part represents the $L_1$ error on
the smooth part (referred to 2nd half $L_1$ error in Figure 3.7), and the last part
quantifies the undershoot error. $c_1$, $c_2$ are some constants to be chosen based
on how important it is to penalize undershoot vs. error on the smooth part
and the non-smooth part. For example, consider gas dynamics problems where
it is important to resolve the undershoot error to avoid nonphysical negative
density or pressure.

In determining optimal parameters by solving a set of test problems, the
choices of grid resolution, CFL number, and the final time play a crucial role.
Different sets of choices may yield different results. So it is important to con-
sider a well-chosen portfolio that includes multiple grid resolutions, different
CFL numbers, and different final times to prevent selection bias. However, it
is observed that for WENO type limiters considered here, the different optimal
parameters produced by the optimization routine are not too far from each
other. So we consider a particular set of grid resolution, CFL number, and a
final time and determine the optimal parameters to convey the main message
of our study.

Using $n = 640$ grid cells, CFL number $\nu = 0.8$, we solve the advection equa-
tion by finite volume methods with limiter functions $H_{\text{CWENO3}}$ and $H_{\text{CWENO3-Z}}$ for
different pairs of $\sigma$’s and $\epsilon$’s till $t_f = 10$. This allows us to calculate energy
surfaces for the error depending on the limiter functions’ parameters \((\epsilon, \sigma)\) and determine the optimal parameters according to our needs.

![Figure 3.7: Left: Cost functional corresponding to \(H_{CWENO3}\) limiter. Right: Cost functional corresponding to \(H_{CWENO3-Z}\) limiter. 1st half \(L_1\) error and the 2nd half \(L_1\) error refer to the first and the second integral, respectively, in the cost functional (3.44). Different errors in the cost functional are calculated and plotted separately for different choices of parameters. Each black dot represents the optimal parameters for each of these energy surfaces.](image_url)

<table>
<thead>
<tr>
<th>Optimal Parameters ((\epsilon, \sigma))</th>
<th>(H_{CWENO3})</th>
<th>(H_{CWENO3-Z})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-smooth region</td>
<td>((1.0 \times 10^{-12}, 1.5 \times 10^{-4}))</td>
<td>((1.0 \times 10^{-14}, 4.9 \times 10^{-2}))</td>
</tr>
<tr>
<td>Smooth region</td>
<td>((1.0 \times 10^{-1}, 4.9 \times 10^{-14}))</td>
<td>((1.0 \times 10^{-1}, 1.5 \times 10^{-12}))</td>
</tr>
<tr>
<td>Undershoot</td>
<td>((1.0 \times 10^{-14}, 4.9 \times 10^{-4}))</td>
<td>((1.0 \times 10^{-12}, 1.5 \times 10^{-1}))</td>
</tr>
</tbody>
</table>

Table 3.1: Different pairs of optimal parameters \((\epsilon, \sigma)\) determined by our study.

The \(L_1\) error and the undershoot error by the \(H_{CWENO3}\) limiter on the non-smooth part are minimal for a small value of \(\epsilon\), and a bigger \(\sigma\) value (see Table 3.1 or Figure 3.7). The \(L_1\) error on the smooth part is almost constant along the line \(\sigma = 4.9 \times 10^{-14}\) (see left panel of Figure 3.7). So we can conclude that the error is minimal along that line. The right panel of
the Figure 3.7 shows the different errors by the \( H_{\text{CWENO3-Z}} \) limiter function. Similar observations indicate that a small value of \( \epsilon \) is always selected by the optimization process. A value of \( \sigma = 1.5 \times 10^{-1} \) is selected by our study for the minimum undershoot error, and the minimum \( L_1 \) error in the smooth region is obtained at a small value of \( \sigma \) (see Table 3.1 or Figure 3.7). To understand the intuitive reasons for different optimal pairs of \( \epsilon \) and \( \sigma \), we analyze different limiter functions for different choices of the parameters. We can see from Figure 3.4 and Figure 3.5 that for a fixed \( \epsilon \), as \( \sigma \) gets smaller and smaller, the limiter functions approach to the 3rd order limiter except for a band along the both coordinate axes in the 2D slope domain. On the other hand, for a fixed \( \sigma \), as \( \epsilon \) gets smaller and smaller, the limiter functions approach some smooth functions. These are also 3rd order limiters, except for a conic region with vertex at origin along the axes. Intuitively, bigger \( \sigma \) leads to more limiting and hence it causes less undershoot error. If the application requires to avoid overshoot or undershoot, we should choose a big value \( \sigma \). On the other hand, for a smooth solution, we take a small \( \sigma \), which produces a third-order limiter function almost everywhere except for a band around the both coordinate axes in an unusual way. As the error is almost constant along the line \( \sigma = C \), where \( C < 10^{-10} \) (see Figure 3.7), we choose a small value of \( \epsilon \) for an optimal choice. Our numerical studies show that the optimal parameters are independent of the choices of the weights \( c_1 \) and \( c_2 \) considered in the cost functional equation (3.44). So altogether, our study determines the optimal parameters in \( H_{\text{CWENO3}} \) is \((\epsilon, \sigma) = (10^{-14}, 4.9 \times 10^{-14})\) for smooth initial condition, and \((\epsilon, \sigma) = (10^{-14}, 4.9 \times 10^{-4})\) for non-smooth initial condition. Similarly, for \( H_{\text{CWENO3-Z}} \) limiter function, we choose \((\epsilon, \sigma) = (10^{-14}, 1.5 \times 10^{-12})\) for smooth initial condition, and \((\epsilon, \sigma) = (10^{-14}, 1.5 \times 10^{-1})\) for non-smooth initial condition.
3.5 Numerical Results

In this section, we demonstrate how the proposed methods with the choices of optimal parameters perform on a variety of test cases. The numerical results by the FV methods with two different limiter functions $H_{\text{CWENO3}}$ and $H_{\text{CWENO3-Z}}$ with their optimal parameters are compared with the traditional third-order scheme, WENO3, and the FV method with $H_{3L}^{c}$ limiter function. For time stepping, we use the third-order Shu-Osher RK scheme [81] for all the test problems considered below.

3.5.1 Linear advection equation with smooth initial condition

Our first example considers the linear advection problem with a smooth initial condition. Consider

$$u_t + u_x = 0 \text{ for } (x, t) \in [0, 1] \times (0, 10],$$

with the smooth initial condition

$$u_0(x) = \begin{cases} \left(0.5 + 0.5 \cos(5\pi(x - 0.5))\right)^4 & \text{if } 0.3 \leq x \leq 0.7 \\ 0 & \text{else} \end{cases}$$

and periodic boundary conditions. The numerical simulations are performed with CFL number $\nu = 0.8$, $n = 170$ grid points till the final time $t = 10$. The left panel of the Figure 3.8 shows the numerical results by different methods on $[0.2, 0.8]$, and the right panel shows an enlarged view of the maximum of the profile. The maximal second order derivative in the limiter function $H_{3L}^{c}$ is $\alpha = \max_{x \in \Omega} \|u''_0(x)\| = 493.48$ for the initial condition in equation (3.46). We fix the parameter $\epsilon$ in the WENO3 method to be $10^{-6}$ following the original work [38]. As the initial condition is smooth, we choose the optimal pairs $(\epsilon, \sigma) = (10^{-14}, 4.9 \times 10^{-14})$ for CWENO3 method and $(\epsilon, \sigma) = (10^{-14}, 1.5 \times 10^{-12})$ for CWENO3-Z method, following our analysis in the previous section. The results show that the 3rd order scheme, FV methods with limiters $H_{3L}^{c}$,
$H_{\text{CWENO3}}$, and $H_{\text{CWENO3-Z}}$ perform much better than the conventional WENO3 method.

Figure 3.8: Left: Numerical solutions of the linear advection equation with CFL number $\nu = 0.8$, grid points $n = 170$, and smooth initial condition in equation (3.46) by different numerical methods on $[0.2, 0.8]$. Right: Zoomed-in view near the maximum of the profile.

CWENO3, and CWENO3-Z methods with their optimal parameters produce equally accurate results as the proposed FV method with $H^{(c)}_{3L}$ limiter in [73]. One advantage of using $H_{\text{CWENO3}}$, and $H_{\text{CWENO3-Z}}$ limiters over the $H^{(c)}_{3L}$ is that we do not require to calculate maximal derivative of the initial profile.

3.5.2 Linear advection equation with discontinuous initial condition

Our second example considers the linear advection equation (3.45) on the domain $[-1, 1]$ with a discontinuous initial condition

$$u_0(x) = \begin{cases} 
1 & \text{if } -0.5 < x < 0.5 \\
0 & \text{else}
\end{cases} \quad (3.47)$$

that can be thought of as a model for contact discontinuities in hyperbolic systems. This initial condition has discontinuities inside the domain $[-1, 1]$, so we expect oscillations near the discontinuities in the solutions by the full
3rd order scheme. The other methods with different limiter functions $H_{3L}^{(c)}$, $H_{WENO3}$, $H_{CWENO3}$, and $H_{CWENO3-Z}$ apply limiting depending on the magnitude of the consecutive slopes near the sharp transitions and avoid spurious oscillations. Different methods have different limiting strategies. The parameter $\alpha$ in the limiter function $H_{3L}^{(c)}$ becomes 0 for this initial condition equation (3.47), and hence $H_{3L}^{(c)}$ applies the limiting via the function $H_{3L}$ (see equation (3.23)). The other limiter functions $H_{WENO3}$, $H_{CWENO3}$, and $H_{CWENO3-Z}$ apply the limiting through their respective homogeneous functions that are obtained from the asymptotic expansion of the limiter functions for $|\delta_{i+\frac{1}{2}}| \gg \epsilon$. The optimal parameter for CWENO3 is the pair $(\epsilon, \sigma) = (10^{-14}, 4.9 \times 10^{-4})$, and $(\epsilon, \sigma) = (10^{-14}, 1.5 \times 10^{-1})$ for CWENO3-Z method, which we obtained by our study in the previous section.

Figure 3.9: The figure illustrates a closeup view of numerical solutions of the linear advection equation with a discontinuous initial condition in equation (3.47) by different numerical methods in two different parts of the domain.

The problem is solved on the domain $[-1, 1]$ with the CFL number $\nu = 0.8$, $n = 320$ grid points till $t = 10$. Each plot in Figure 3.9 shows a zoomed-in view of discontinuities of the solution by different schemes. As expected, the 3rd order scheme produces a spurious oscillation. The other methods avoid oscillations by applying the limiting, but some methods perform better than others. Again we observe that the performance by CWENO3, and CWENO3-Z methods with their optimal parameters is as good as the FV method with
\( H_{3L}^{(c)} \) limiter function.

### 3.5.3 Inviscid Burgers’ equation

As a nonlinear PDE test problem, our first example is the inviscid Burgers’s equation

\[
    u_t + \left( \frac{u^2}{2} \right)_x = 0, \quad -1 \leq x \leq 1, \quad t > 0 ,
\]

subject to periodic boundary conditions and the smooth initial condition \( u_0(x) = \sin(4\pi x) \). The solution of Burgers’ equation starting with smooth sine wave data compresses in some regions and expands in some other areas. So it produces a combination of shocks and rarefactions. Eventually, it takes the form of a so-called \( N \)-wave shape between shocks. For this reason, we choose the optimal parameters of \( H_{CWENO3} \), and \( H_{CWENO3-Z} \) according to non-smooth initial data and conduct the numerical simulations. The constant in the \( H_{3L}^{(c)} \) limiter function is calculated as \( \alpha = 16\pi^2 \) for the initial data \( u_0(x) = \sin(4\pi x) \).

![Figure 3.10: Approximate solutions (with zoomed-in view) of Burgers’ equation with a smooth initial condition, CFL number \( \nu = 0.8 \), and \( n = 100 \) grid points by different methods at time \( t = 0.5 \).](image)

The approximate solutions are computed with CFL number \( \nu = 0.8 \), and \( n = 100 \) grid points till the final time \( t = 0.5 \) by different methods. A reference solution \( (u_{ex}) \) is calculated on a very fine mesh (5000 grid points) by the WENO3 method for the sake of comparison. As before, all the methods except
the full 3rd method use limiting to prevent oscillations, but some methods function better than others. The full 3rd order method lacks a mechanism to limit the slopes near the non-smooth regions, which results in oscillations near the sharp transitions.

3.5.4 Sod shock tube problem

We consider the Sod shock tube problem [82] to test and compare the proposed schemes’ accuracy with other existing methods. This problem describes the variables density \( \rho \), velocity \( v \), and pressure \( p \) of an ideal gas when it is put in different states at the left and right of a diaphragm, placed at \( x = 0 \) in a tube and the diaphragm is removed at \( t = 0 \). The time evolution of these three quantities is modeled by the one-dimensional Euler equations that come from the conservation of mass, momentum, and energy of a compressible fluid. The primitive variables density, velocity, and pressure are not conserved, but the mass, momentum, and energy are conserved quantities. So the system of Euler equations considers the conserved quantities as unknown variables, and the equations are given by

\[
\begin{bmatrix}
\rho \\
\rho u \\
E
\end{bmatrix}_t + \begin{bmatrix}
\rho u \\
\rho u^2 + p \\
(E + p)u
\end{bmatrix}_x = 0, \tag{3.49}
\]

where

\[
E = \frac{p}{\gamma - 1} + \frac{1}{2} \rho u^2 \tag{3.50}
\]

is the total energy and the ratio of specific heats is taken to be \( \gamma = 1.4 \). In line with [50], we consider the Sod shock problem on the domain \([-2, 2]\) with the states in the left and right of \( x = 0 \) are given by

\[
\begin{bmatrix}
\rho_L \\
v_L \\
p_L
\end{bmatrix} = \begin{bmatrix}
1.0 \\
0.0 \\
1.0
\end{bmatrix}, \quad \text{and} \quad \begin{bmatrix}
\rho_R \\
v_R \\
p_R
\end{bmatrix} = \begin{bmatrix}
0.125 \\
0.0 \\
0.1
\end{bmatrix}. \tag{3.51}
\]
The solutions are computed by different methods on $n = 100$ grid points with 
CFL number $\nu = 0.95$ up to the final time $t = 0.8$. As we have a discontinuous 
initial profile (3.51), we obtain $\alpha = 0$ in the limiter function $H_{3L}^{(c)}$, which implies 
that the $H_{3L}^{(c)}$ limiter function applies limiting via the function $H_{3L}$. The 
reconstruction procedure in WENO3, CWENO3, and CWENO3-Z methods is 
performed on the primitive variables $\rho$, $v$, and $p$ in line with [92]. As before, 
the optimal parameters for CWENO3 is chosen $(\epsilon, \sigma) = (10^{-14}, 4.9 \times 10^{-4})$, 
and $(\epsilon, \sigma) = (10^{-14}, 1.5 \times 10^{-1})$ for CWENO3-Z as we have discontinuities in 
the initial condition of the problem.

Figure 3.11: Left: Density profile for the Sod’s problem as computed by dif-
ferent methods at time $t = 0.8$ with CFL number $\nu = 0.95$ and $n = 100$ grid 
points. Right: Zoomed-in region at the contact discontinuity.
We see that the proposed schemes with optimal parameters perform as accurately as the FV method with $H_{3L}^{(c)}$ limiter function in all cases. The full 3rd order scheme produces spurious oscillations near the discontinuities. The WENO3 method avoids oscillations, but it is not as accurate as compared to the FV methods with $H_{CWENO3}$, $H_{CWENO3-Z}$, or $H_{3L}^{(c)}$ limiter functions.
3.5.5 Shu-Osher problem

Our last example is the Shu-Osher problem [75, 81] that is designed to test shock-capturing schemes. The governing equation is the one-dimensional Euler equations (3.49) on the domain \([-4.5, 4.5]\) with the shock is at \(x = -4\) at \(t = 0\). The states of primitive variables to the left and the right of \(x = -4\) are given by

\[
\begin{pmatrix}
\rho_L \\
v_L \\
p_L
\end{pmatrix} = \begin{pmatrix} 3.857143 \\ 2.629369 \\ 10.3333 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix}
\rho_R \\
v_R \\
p_R
\end{pmatrix} = \begin{pmatrix} 1 + 0.2 \sin(5x) \\ 0.0 \\ 1.0 \end{pmatrix}. \tag{3.52}
\]

A right-moving Mach 3 shock wave propagates into a sinusoidal density field with amplitude 0.2 and frequency 5, and this generates a density flow field with two different characteristics behind the main shock.

Figure 3.14: Solution of the Mach 3 shock-density wave interaction by different methods on \([-4.5, 4.5]\) with \(n = 640\) grid points, CFL number \(\nu = 0.95\) at the final time \(t = 1.8\).
Figure 3.15: Enlarged view of the density profile near a sharp transition region on a high resolution with $n = 1280$ mesh points.

Figure 3.16: A closeup view of the oscillatory area of the density profile on a high resolution with $n = 1280$ grid points.

The solution is computed with CFL number $\nu = 0.95$ at the final time $t = 1.8$. In accordance with [92], the reconstruction techniques have been applied to the primitive variables density, velocity, and pressure. The limiter function $H_{3L}^{(c)}$ has a constant $\alpha = 5$ for the initial condition equation (3.52). The
optimal parameters in $H_{CWENO3}$ and $H_{CWENO3-Z}$ limiter functions are chosen according to discontinuous initial conditions. We compute a reference solution by WENO3 method on a fine mesh with 10,000 grid points. Figure 3.14 provides a comparison for all schemes at $t = 1.8$ on $n = 640$ mesh points. Figure 3.15 shows a closeup view of a sharp transition, and a zoomed-in view of the oscillatory region is shown in Figure 3.16, where the solution is calculated at a high resolution with $n = 1280$ grid points. The same observation can be made in the numerical results as in the previous test cases, i.e., the full 3rd order scheme produces spurious oscillations near sharp transitions, and other schemes are able to avoid oscillations. WENO3 conducts more limiting compared to other methods leading to less accurate solutions compared to FV methods with $H_{CWENO3}$, $H_{CWENO3-Z}$, or $H_{3L}^{(c)}$ limiter functions. In terms of accuracy, the last three methods produce similar solutions over the whole computational domain.

### 3.6 Conclusion: Challenges and Future Work

We designed two different limiter functions, each involving two artificial parameters based on 3rd order central WENO and WENO-Z methods, and interpreted them in the 2D slope domain. We provided a pathway to determine optimal parameters in these limiter functions. Based on a particular but generic test case, we designed a cost functional that takes numerical error, overshoot, or undershoot into account and determined optimal parameters by optimizing that cost functional over the two-parameter family of limiter functions. These limiters with their optimal parameters were then tested on several test problems, and we found that these are the best limiter functions in the class of 3rd order WENO type limiter functions, and they perform with similar accuracy as the $H_{3L}^{(c)}$ limiter proposed in [73].

One of the main challenges here is to prevent overfitting, i.e., how to get the optimal limiter function, which not only works well for a particular set of problems but also performs well generically. Challenges related to overfitting...
are to design a well-represented portfolio of test cases and a generic cost functional. If the portfolio of test cases is not designed sufficiently generic, then a highly oscillatory optimal limiter could be found that works well for the very particular subset of problems at hand but are not generically applicable. Here we considered only one specific test case, and the also the cost functional designed above is not a generic cost functional, is very particular to that test problem. So the question remains how one can design a cost functional generically which penalizes the numerical error (diffusion and dispersion), overshoots or undershoots, etc. So one key task would be to answer the above questions and obtain a limiter function that can be used for many applications.

To overcome some of these challenges, we explored the Approximation dispersion relation (ADR) method developed by Pirozzoli et al. [66]. ADR is a tool that quantifies the effect on the solution in wave-number space when we apply some nonlinear method to solve a problem. Diffusion and dispersion of WENO methods are studied extensively in [37] by using the ADR method. Further, Cravero et al. [21] developed a measure of distortion error by a nonlinear scheme based on the ADR method, which accounts for the error introduced by a scheme leading to spurious modes. We tried to use distortion measure to design a generic cost functional, because this works on different Fourier modes and computes the numerical error of a scheme. Hence we could avoid the hard work of accumulating the generic set of test cases. But we did not get a straightforward way to do it, and the question remains open for future research.
CHAPTER 4

DIRK SCHEMES WITH WEAK STAGE ORDER BEYOND THREE

So far, our study of numerical methods was devoted to hyperbolic or, more precisely, advection type problems. We focused on numerical methods that can accurately capture the structure of the numerical solutions for a long time and for a fixed final time via a new class of jet schemes and finite volume methods with optimal WENO type limiter functions, respectively. This chapter focuses on order-preserving time stepping methods both for ODEs and PDEs. More specifically, we develop high-order accurate time integrators to avoid order reduction when used to solve time-dependent problems, in particular stiff systems [83, 91]. This is a particular class of Runge-Kutta (RK) methods, diagonally implicit Runge-Kutta methods (DIRK) that are constructed satisfying the weak stage order (WSO) criterion [42, 70]. The Runge-Kutta (RK) family consists of one-step methods that combine multi-stages in each step to solve time-dependent differential equations and achieve high-order accuracy. However, when applied to solve some stiff problems, they experience order reduction [70], i.e., the observed orders of the accuracy of the time stepping methods are less than the schemes’ formal orders. Among the many order re-
duction remedies, high stage order is the simplest one, which is not compatible with DIRK schemes. DIRK schemes are of vital interest as they can achieve high-order by sequentially solving backward-Euler-type stages. The concept of WSO, which are some algebraic conditions on the Runge-Kutta Butcher tableau, can relax the stage order conditions. WSO conditions are compatible with DIRK structure, and it has been shown that the DIRK schemes with high weak stage order rectify the order reduction phenomenon for a certain class of problems. DIRK schemes with orders up to four and WSO up to three were constructed based on a simplified version of WSO criterion, which was introduced as the WSO eigenvector criterion in [42]. The prior work [42] has also shown that the WSO eigenvector criterion is fundamentally limited to WSO 3. Hence, the main objective of this chapter is to establish that the concept of general WSO can be pushed beyond the limitation of WSO eigenvector criterion. Then we construct new DIRK schemes with high WSO and demonstrate that those schemes can resolve order reduction for a variety of test problems. We start this chapter by describing stiffness in a problem via an illustrative example in the next section.

4.1 Stiff Problem and Stability

The parameters \((A, \bar{b}, \bar{c})\) of a general \(s\)-stage RK method are displayed via the Butcher tableau:

\[
\begin{array}{c|cc}
\bar{c} & A \\
\hline
\bar{b}^T & \end{array}
\]

where the matrix \(A \in \mathbb{R}^{s \times s}\), \(\bar{b} \in \mathbb{R}^s\). We assume \(\bar{c} = A\bar{c}\), where \(\bar{c}\) is the vector of ones in \(\mathbb{R}^s\). Consider the initial value problem

\[
u'(t) = f(t, u(t)), \quad u(0) = u_0; \quad u, f \in \mathbb{R}^m,
\]

and let \(u_n\) and \(u_{n+1}\) be the numerical approximations to the true solution at time \(t_n\) and \(t_{n+1} = t_n + \Delta t\) with a time step \(\Delta t\). The RK methods compute
the numerical solution as

\[ u_{n+1} = u_n + \Delta t \sum_{j=1}^{s} b_j f(t_n + c_j \Delta t, g_j), \quad (4.3) \]

where

\[ g_i = u_n + \Delta t \sum_{j=1}^{s} a_{ij} f(t_n + c_j \Delta t, g_j), \quad i = 1, 2, \ldots, s. \quad (4.4) \]

Now consider a particular family of problems of the form (Prothero-Robinson test problem [91])

\[ u'(t) = \lambda (u(t) - \phi(t)) + \phi'(t); \quad u(0) = u_0, \text{ with Re}(\lambda) \leq 0, \quad (4.5) \]

where the true solution \( u(t) = \phi(t) \) is a smooth function that varies at a moderate rate. We solve this problem numerically by two RK methods: explicit and implicit Euler methods with the choice of \( m = 5 \) and \( m = 10 \) grid points, respectively, on the domain \([0, 1]\) for both \( \lambda = -2, -20, u_0 = 0 \) and \( \phi(t) = \sin(\pi t) \). The Butcher’s tableaus for the explicit and implicit Euler methods are given in the table 4.1. Hence the update rules become

(a) Explicit Euler

\[
\begin{array}{c|c}
0 & 0 \\
1 & 1 \\
\end{array}
\]

(b) Implicit Euler

\[
\begin{array}{c|c}
1 & 1 \\
1 & 1 \\
\end{array}
\]

Table 4.1: Butcher’s tableaus for the Euler’s methods

Explicit : \( u_{n+1} = u_n + \Delta t \cdot f(t_n, u_n) \), \quad (4.6)

Implicit : \( u_{n+1} = u_n + \Delta t \cdot f(t_{n+1}, u_{n+1}) \). \quad (4.7)

When \( \lambda = -2 \), both methods approximate the true solution with similar accuracy (at least pictorially) for \( m = 5 \) and \( m = 10 \) (see Figure 4.1). For \( \lambda = -20 \), the explicit method goes unstable in the case of \( m = 5 \), and the numerical
Figure 4.1: Comparing the Euler methods. The problem (4.5) is solved for $\lambda = -2$ with $m = 5$ (left panels) and $m = 10$ (right panels) grid points, comparing explicit Euler (top) to implicit Euler (bottom). The exact solution is given by black curve, and the red curve shows the numerical solution.

solution shows growing oscillations diverging from the exact solution (see Figure 4.2). But with $m = 10$, it regains the stability, and this is due to the stability region of the explicit Euler method. In contrast, the implicit Euler method has stable solutions for both $m = 5$ and 10, showing that it can take large time steps without being unstable. To understand the reason behind these facts, we need to understand the concept of the stability region of these methods.

The notion of absolute stability is based on the simplest linear test problem

$$u' = \lambda u; \quad u(0) = 1, \text{ with } \lambda \in \mathbb{C}, \text{ and } \Re(\lambda) \leq 0. \quad (4.8)$$

The explicit Euler method applied to this problem (4.8) produces

$$u^{n+1} = (1 + \lambda \Delta t) u^n,$$

and we say the method is absolutely stable when $|1 + \lambda \Delta t| \leq 1$; otherwise unstable. We define $z := \lambda \Delta t$ and write the condition as $|1 + z| \leq 1$. A similar
Figure 4.2: The test problem (4.5) is solved under the same assumptions before except for $\lambda = -20$. Black and red curves show the true and numerical solutions, respectively. The explicit method shows numerical instability (top left), as indicated by growing oscillations. This is due to the sharp transition of the neighboring solutions (as $\lambda = -20$). This is typical for stiff problems.

analysis for *implicit Euler method* would give us the condition as $\left| \frac{1}{1-z} \right| \leq 1$. The *stability region* of the test problem is the entire left half plane as $\text{Re}(\lambda) \leq 0$ would give us stable solutions. So the stability of the differential equation depends on the value of $\lambda$. Unlike the stability of the ODE, the stability of the numerical methods depends on both $\lambda$ and the step size $\Delta t$, specifically on the product $\lambda \Delta t$. Now in Figure 4.1, $\lambda \Delta t = -0.4$ and $-0.2$ for $m = 5$ and $m = 10$, respectively ($\Delta t = \frac{1}{m}$), which are in the region of absolute stability for both *explicit* and *implicit Euler methods* (see Figure 4.3). So the numerical solutions are stable. But in Figure 4.2, $\lambda \Delta t = -4$ and $-2$. These are in the *absolute region* of stability of *implicit Euler method* (see Figure 4.3); hence the *implicit Euler method* produces stable numerical solutions for both $m = 5$ and $m = 10$. But $\lambda \Delta t = -4$ is not in the stability region of *explicit Euler method*, producing unstable solution for $m = 5$. This is typical behavior of stiff problems, i.e., an explicit method requires a very small time step to get
Figure 4.3: Stability regions of explicit and implicit Euler methods. Left panel shows that the explicit Euler method is stable when $\lambda \Delta t$ is inside the unit circle centered at $-1$. The right panel shows that the stability region of the implicit Euler method contains the stability region of the ODE problem (4.8), the entire left half plane.

stable numerical solutions. There is no precise definition, but we are going to stick to the following definition of stiffness as follows:

**Definition 4.1** (Stiff problem). An ODE system is called stiff if explicit schemes require much smaller time steps to satisfy stability than to satisfy the accuracy.

Roughly speaking, a stiff ODE is a system in which the solution consists of rapidly changing parts and slowly changing parts. So we can also define that a stiff system is a problem where we choose a larger time step than the fastest time scale of the problem dynamics. The most general way to define a stiff problem is that when the eigenvalues of the Jacobian of $f$ differ greatly in magnitude. An example of such a system would be the chemical reaction system [91] that contains several fast and slow reactions. Stiffness also occurs in many PDE initial-boundary-value problems (IBVPs), for example, the heat equation. Discretization of the second-order spatial derivative of the heat
equation on a finite grid with some mesh space size $\Delta x$ yields a system of ODEs that gains a huge stiffness ratio (ratio between the maximum absolute value of eigenvalues over the minimum absolute value of eigenvalues) as $\Delta x \to 0$ and makes the problem very stiff. On the other hand, time discretization of the heat equation makes the problem infinitely stiff. In the next section, we will consider the heat equation with Dirichlet b.c. and demonstrate the order reduction phenomenon caused by the time stepping schemes.

4.2 Order Reduction and Its Remedy

Runge Kutta schemes, applied to some problems, often do not recover their conventional orders of convergence, and this phenomenon is known as order reduction. It occurs for linear as well as for nonlinear problems. To study the order reduction phenomenon, Prothero and Robinson introduced a particular family of problems (4.5), whose family of solutions shows a sharp transition in the particular solution’s neighborhood for the large negative values of the stiffness parameter $\lambda$. We consider a 2nd and a 3rd order DIRK schemes given in the following Butcher’s tableaus

\[
\begin{array}{c|cc}
\bar{c} & A & \bar{b}^T \\
\hline
1 & \gamma & \gamma \\
1 - \gamma & 1 - \gamma & \gamma \\
1 - \gamma & \gamma & \\
\end{array}
\]

Table 4.2: Butcher’s tableau for a 2nd order method with $\gamma = 1 - \frac{1}{\sqrt{2}}$.

and solve the problem (4.5) up to the final time $T = 10$ with the true solution $\phi(t) = \sin \left( t + \frac{\pi}{4} \right)$ and $\lambda = -10^{-4}$. The convergence results are shown in Figure 4.4. The figure clearly shows that both the schemes do not exhibit their conventional convergence orders for larger values of time step size $\Delta t$. Hence the order reduction is clearly manifested in the case of this linear ODE
\[ 0.4358665215 \quad 0.4358665215 \\
\bar{c} \quad A \quad = \\
0.7179332607 \quad 0.2820667392 \quad 0.4358665215 \\
1 \quad 1.2084966490 \quad -0.644363171 \quad 0.4358665215 \\
1.2084966490 \quad -0.644363171 \quad 0.4358665215 \\
\]

Table 4.3: Butcher’s tableau for a 3rd order method.

test problem. Now we turn our attention to a PDE IBVP and illustrate the

\begin{align*}
\text{Error Convergence by 2nd order scheme} \\
\text{2nd order scheme} & \quad \text{Slope = 1} & \quad \text{Slope = 2} \\
10^{-15} & \quad 10^{-13} & \quad 10^{-11} & \quad 10^{-9} & \quad 10^{-7} \\
\text{Error} & \quad \text{Error} & \quad \text{Error} & \quad \text{Error} & \quad \text{Error} \\
10^{-11} & \quad 10^{-9} & \quad 10^{-7} & \quad 10^{-5} & \quad 10^{-3} & \quad 10^{-1} \\
\end{align*}

Figure 4.4: Error convergence for the Prothero-Robinson test problem (4.5) by a 2nd order DIRK scheme (left), and a 3rd order DIRK scheme (right).

order reduction phenomenon incurred via boundary layers [35]. One of the vital components for the exhibition of order reduction for PDE IBVPs is the boundary condition. For simplicity, the exact b.c. is used for each stage in the numerical simulation. This causes zero error at the boundary and non-zero error away from the boundary that leads to boundary layers and hence order reduction. To demonstrate the order reduction, we consider the 1D heat equation with Dirichlet b.c.

\[ u_t = u_{xx} + f \quad \text{for} \quad (x, t) \in (0, 1) \times (0, 1), \quad u = g \quad \text{on} \quad \{0, 1\} \times (0, 1), \quad (4.9) \]
with the forcing function $f$, the boundary conditions $g$ and the initial condition such that the true solution is $u(x, t) = \cos(t)$. For the numerical simulation, a 2nd order centered finite difference on a fine grid with 10000 cells is used to discretize the spatial derivative. As a result, the spatial errors are negligible, and the actual behavior of temporal error is obtained. The 3rd order DIRK scheme in the table 4.3 is used for time stepping. All errors are measured in the maximum norm. The right panel of the Figure 4.5 shows the error convergence of function values and derivatives by a 3rd order DIRK scheme (DIRK3). The scheme suffers from order reduction ($u$ converges with $O(\Delta t^2)$)

Figure 4.5: Left: Scaled BL profiles via $(x/\Delta t^{0.5}, e/\Delta t^{2.0})$ shows that BL thickness $= O(\sqrt{\Delta t})$ and BL height $= O(\Delta t^2)$. Middle: Global spatial error. Right: The error convergence for the heat equation (4.9) by a 3rd order DIRK scheme. The middle panel shows the global errors as functions of $x$ with three $\Delta t$ choices. The error convergence for function values and derivatives ($u$ blue circles; $u_x$ red squares) are shown in the right panel.

and $u_x$ converges with $O(\Delta t^{1.5})$ for both function values and derivatives. This example shows that order reduction is rather common and can occur in very simple problems. Numerically, we can determine the thickness and the amplitude of the boundary layer (BL). We postulate a scaling of BL thickness $\Delta t^\alpha$ ($\alpha > 0$), and a scaling of BL height $\Delta t^\beta$ ($\beta > 0$), and then plot the scaled boundary layer profiles via $(x/\Delta t^\alpha, e/\Delta t^\beta)$ ($e$ denotes the error) as in the left panel of Figure 4.5. We observe that the profile converges (as $\Delta t \to 0$) to a
fixed function with asymptotic scaling for $\alpha = 0.5$ and $\beta = 2$. This implies that the BL thickness $= \mathcal{O}(\sqrt{\Delta t})$ and BL height $= \mathcal{O}(\Delta t^2)$. A geometric explanation of order reduction via boundary layers can be found in [70]. The main reason can be understood by observing the shape of global spatial errors for different choices of $\Delta t$ in the middle panel of Figure 4.5. Notice that the scheme develops boundary layers for each choice of $\Delta t$. On the other hand, if we first discretize time, leaving the space continuous, then each stage solves a 2nd order boundary value problem with the conventional Dirichlet b.c. at the boundary. The equations for stages become singularly perturbation BVPs with a small parameter $\Delta t$, and the standard perturbation analysis [35] reveals that each stage exhibits a boundary layer of thickness $\mathcal{O}(\sqrt{\Delta t})$. The amplitude of the error for function values in the boundary layer scales like $\mathcal{O}(\Delta t^2)$, and the error in the interior scales same as the order of the method. All together this leads to an order reduction of 1 for function values. On top of that, the boundary layer thickness $\mathcal{O}(\sqrt{\Delta t})$ causes an extra half order reduction for the derivatives.

Our goal now is to understand the order reduction formally and come up with some remedies to this phenomenon. To study this formally, we look at the accuracy of an RK scheme applied to some stiff ODE problems. The accuracy of the Runge-Kutta schemes applied to stiff problems can be analyzed by the explicit formulas for the local and global errors of the methods applied to our particular model stiff problem equation (4.5). Applying the RK method (4.3) and (4.4) to the problem (4.5) we get

$$g_i = u_n + \Delta t \sum_{j=1}^{s} a_{ij} \left( \lambda [g_j - \phi(t_n + c_j \Delta t)] + \phi'(t_n + c_j \Delta t) \right), \quad (4.10)$$

and

$$u_{n+1} = u_n + \Delta t \sum_{j=1}^{s} b_j \left( \lambda [g_j - \phi(t_n + c_j \Delta t)] + \phi'(t_n + c_j \Delta t) \right). \quad (4.11)$$

Starting from the true solution at time $t_n$ with $g_i = \phi(t_n + c_i \Delta t)$, $u_n = \phi(t_n)$, and $u_{n+1} = \phi(t_n + \Delta t)$, we get the stage error $E_{i,\Delta t}(t_n)$ and the local error
\( E_{\Delta t}(t_n) \) at \( i \)th stage and in one-step, respectively, given by

\[
E_{i,\Delta t}(t_n) = \phi(t_n) + \Delta t \sum_{j=1}^{s} a_{ij} \phi'(t_n + c_j \Delta t) - \phi(t_n + c_i \Delta t) ,
\]

(4.12)

and

\[
E_{\Delta t}(t_n) = \phi(t_n) + \Delta t \sum_{j=1}^{s} b_j \phi'(t_n + c_j \Delta t) - \phi(t_n + \Delta t) .
\]

(4.13)

Taylor series expansion of the functions in (4.12) and (4.13) produces the following:

\[
E_{i,\Delta t}(t_n) = \phi(t_n) + \Delta t \sum_{j=1}^{s} a_{ij} \left[ \phi'(t_n) + (c_j \Delta t) \phi''(t_n) + \frac{(c_j \Delta t)^2}{2!} \phi'''(t_n) + \cdots \right]
\]

\[
+ \frac{(c_i \Delta t)^q}{(q-1)!} \phi^{(q)}(t_n) + O(\Delta t^q) \right] - \left[ \phi(t_n) + (c_i \Delta t) \phi'(t_n) + \frac{(c_i \Delta t)^2}{2!} \phi''(t_n) + \cdots \right]
\]

\[
+ \frac{(c_i \Delta t)^q}{q!} \phi^{(q)}(t_n) + O(\Delta t^{q+1}) \right] ,
\]

(4.14)

and

\[
E_{\Delta t}(t_n) = \Delta t \phi'(t_n) \left[ \sum_{j=1}^{s} a_{ij} c_j^{1-1} - c_i^{1} \right] + \frac{(\Delta t)^2}{1!} \phi''(t_n) \left[ \sum_{j=1}^{s} a_{ij} c_j^{2-1} - \frac{c_i^{2}}{2} \right] + \cdots
\]

\[
+ \frac{(\Delta t)^q}{(q-1)!} \phi^{(q)}(t_n) \left[ \sum_{j=1}^{s} a_{ij} c_j^{q-1} - \frac{c_i^{q}}{q} \right] + O(\Delta t^{q+1}) ,
\]

(4.15)

In general we can also write these as

\[
E_{i,\Delta t}(t_n) = \sum_{k \geq 1} \frac{(\Delta t)^k}{(k-1)!} \left[ \sum_{j=1}^{s} a_{ij} c_j^{k-1} - \frac{c_i^k}{k} \right] \phi^{(k)}(t_n) , \quad \text{and}
\]

(4.16)

\[
E_{\Delta t}(t_n) = \sum_{k \geq 1} \frac{(\Delta t)^k}{(k-1)!} \left[ \sum_{j=1}^{s} b_j c_j^{k-1} - \frac{1}{k} \right] \phi^{(k)}(t_n) .
\]

(4.17)
The approximation error at time $t_{n+1}$ is defined and calculated using (4.11) and (4.13) as

$$
\epsilon^{n+1} := u^{n+1} - \phi(t_n + \Delta t) = \epsilon^n + \lambda \Delta t (\vec{b}^T \cdot \vec{v}) + E_{\Delta t}(t_n),
$$

(4.18)

where $\vec{v} = [g_1 - \phi(t_n + c_1 \Delta t), g_2 - \phi(t_n + c_2 \Delta t), \ldots, g_s - \phi(t_n + c_s \Delta t)]^T$. We solve for $\vec{v}$ using (4.10) and (4.12) as

$$
\vec{v} = (I - \lambda \Delta t A)^{-1} \left[ \vec{e}^{\epsilon n} + \vec{E}(t_n) \right],
$$

(4.19)

where $\vec{E}(t_n) = [E_{1,\Delta t}(t_n), E_{2,\Delta t}(t_n), \ldots, E_{s,\Delta t}(t_n)]^T$. Letting $\zeta = \lambda \Delta t$ and using (4.19) in (4.18) we obtain

$$
\epsilon^{n+1} = R(\zeta) \epsilon^n + \zeta \vec{b}^T (I - \zeta A)^{-1} \vec{E}(t_n) + E_{\Delta t}(t_n),
$$

(4.20)

where $R(\zeta) := 1 + \zeta \vec{b}^T (I - \zeta A)^{-1} \vec{e}$ is the growth factor per time step $\Delta t$. Using the equation (4.16), we can write the vector $\vec{E}(t_n)$ as

$$
\vec{E}(t_n) = \sum_{k \geq 1} \frac{(\Delta t)^k}{(k - 1)!} \vec{\tau}^{(k)} \phi^{(k)}(t_n),
$$

(4.21)

where the vector $\vec{\tau}^{(k)} = A \vec{\tau}^{k-1} - \frac{1}{k} \vec{\tau}^k$ is call the $k$th stage order residual and $\phi^{(k)}(t_n)$ is the $k$th derivative of $\phi$ at $t_n$. The collection of the conditions $\vec{\tau}^{(k)} = 0$ for $k = 1, 2, \ldots, \hat{q}$ is denoted by $C(\hat{q})$ in the literature [12], [91]. The necessary conditions for a $\hat{p}$th order Runge-Kutta scheme is $\sum_{j=1}^{s} b_j \vec{c}_j^{k-1} - \frac{1}{k} \vec{\tau} = 0$ for $k = 1, 2, \ldots, \hat{p}$, and these conditions together is denoted by $B(\hat{p})$. The condition $B(\hat{p})$ holds implies $E_{\Delta t}(t_n) = O(\Delta^{\hat{p}+1})$. So the order of accuracy of a scheme is affected by the term $\zeta \vec{b}^T (I - \zeta A)^{-1} \vec{E}(t_n)$, more specifically it depends on the asymptotic behavior of the expression

$$
g^{(k)}(\zeta) = \zeta \vec{b}^T (I - \zeta A)^{-1} \vec{\tau}^{(k)}. \quad (4.22)
$$

In the classical theory, i.e., in the non-stiff case, the scheme’s convergence is treated as $\Delta t \to 0$ with $\zeta = O(\Delta t)$. With $\zeta \ll 1$ and using a Neumann expansion, we get $\zeta (I - \zeta A)^{-1} = \zeta I + \zeta^2 A + \zeta^3 A^2 + \cdots$, which leads to the terms like $\vec{b}^T A^\ell \vec{\tau}^{(k)}$ with $\ell \geq 0$. The order conditions imply that these
terms vanish, ensuring the formal order of the scheme. In the case of stiff problems, we are interested in the time steps that are much larger than the fastest time scale of the problem dynamics, which is $\frac{1}{|\lambda|}$ for the Prothero-Robinson problem, i.e., we want $|\lambda|\Delta t \gg 1$. So we study the convergence of errors under the limits $\Delta t \to 0$ and $\zeta \to -\infty$ simultaneously, i.e., $\lambda \to -\infty$ faster than $\Delta t \to 0$. In this case $\zeta^{-1}$ is small and a Neumann expansion yields $$\zeta (I - \zeta A)^{-1} = -A^{-1} (I - \zeta^{-1} A^{-1})^{-1} = \zeta I + \zeta^2 A + \zeta^3 A^2 + \cdots,$$ leading to the terms like $\vec{b}^T A^{\ell} \vec{\tau}(k)$ but with $\ell < 0$. These quantities are not guaranteed to vanish by the order conditions, and this in general leads to order reduction in the convergence.

One way to avoid order reduction is to use schemes with high stage order which is defined as follows:

**Definition 4.2 (Stage order).** Let the conditions $B(\hat{p})$ and $C(\hat{q})$ hold, i.e., $\hat{p}$ is the largest number such that $\sum_{j=1}^{s} b_j c_j^{k-1} - \frac{1}{k} = 0$ for $k = 1, 2, \ldots, \hat{p}$, and $\hat{q}$ is the largest number with $\vec{\tau}(k) = 0$ for $k = 1, 2, \ldots, \hat{q}$. The stage order of an RK scheme is $\tilde{q} = \min\{\hat{p}, \hat{q}\}$.

If we have a scheme with stage order $\tilde{q}$, then we observe from the error formulas that the error decreases at least at an order $\tilde{q}$ in the stiff regime, avoiding the order reduction. We are particularly interested in DIRK schemes to solve stiff problems because of their implementation-friendly structure and cost efficiency. However, it is well-known that DIRK schemes are restricted to low stage order [42].

**Theorem 4.1.** The stage order of an irreducible DIRK scheme is at most 2. The stage order of a DIRK scheme with non-singular $A$ is at most 1.

In the next section, we describe a criterion weak stage order (WSO) that is weaker than stage order conditions but compatible with the DIRK structure. We will show later that the DIRK schemes with high WSO orders avoid order reduction for a certain class of problems.
4.3 Weak Stage Order

As DIRK schemes are not compatible with stage order, a weaker condition called *weak stage order* \([42, 70]\) was introduced to avoid order reduction in the context of DIRK schemes. Weak stage order puts some algebraic conditions on the Butcher tableau \((A, \vec{b}, \vec{c})\), and it is compatible with DIRK structure. Consider the following Krylov subspace generated by the stage order residuals \(\{\vec{r}(1), \vec{r}(2), \ldots, \vec{r}(q)\}\):

\[
K_q := \text{span} \{ \vec{r}(1), A\vec{r}(1), \ldots, A^{s-1}\vec{r}(1), \vec{r}(2), A\vec{r}(2), \ldots, A^{s-1}\vec{r}(q) \}.
\]

We have now two following equivalent definitions of *weak stage order*.

**Definition 4.3.** (Weak stage order, version 1) An \(s\)-stage RK scheme \((A, \vec{b}, \vec{c})\) is said to have WSO \(q\) if \(\vec{b} \perp K_q\). Namely,

\[
\vec{b}^T A^j \vec{r}(k) = 0, \quad \text{for} \quad 0 \leq j \leq s - 1, \quad 2 \leq k \leq q. \tag{4.23}
\]

**Definition 4.4.** (Weak stage order, version 2) An RK scheme has WSO \(q\), if there exists an \(A\)-invariant vector space \(V\) such that: \(\vec{r}(k) \in V\), for \(1 \leq k \leq q\) and \(\vec{b}^T y = 0\) for all \(y \in V\).

These two definitions of weak stage order are equivalent, and weak stage order is the most general criterion to guarantee the avoidance of order reduction (Theorem 2 in [42]) in linear problems, specifically the stiff problems. Our goal is to construct high order DIRK schemes that have high weak stage order.

Finding such schemes boils down to a root-finding problem when all the order conditions and the weak stage order conditions are formulated as polynomial equations in the variables as coefficients in the butcher tableau \((A, \vec{b}, \vec{c})\). To find an \(s\)-stage, \(p\)th order DIRK scheme with WSO \(q\), one needs to find a common solution to the following two systems of polynomial equations:

\[
\begin{align*}
\beta_j&(A, \vec{b}, \vec{c}) = 0, \quad \text{for} \quad j = 1, 2, \ldots, n_p; \quad (p\text{th order conditions}), \quad \text{and} \\
\gamma_{j,k}&(A, \vec{b}, \vec{c}) := \vec{b}^T A^j \vec{r}(k) = 0, \quad \text{for} \quad 0 \leq j \leq s - 1, \quad 2 \leq k \leq q, \tag{4.24}
\end{align*}
\]
where \( n_p \) (\( n_4 = 8 \) and \( n_5 = 17 \)) is the number of polynomial equations required to be satisfied for order \( p \). With the assumption \( \vec{c} = A\vec{c} \), the order conditions up to order 5 are given by the following equations:

Consistency : \( \vec{b}^T e = 1 \),

2nd order : \( \vec{b}^T \vec{c} = \frac{1}{2} \),

3rd order : \( \vec{b}^T A\vec{c} = \frac{1}{6}, \quad \vec{b}^T \vec{c}^2 = \frac{1}{3} \),

4th order :

\[
\begin{align*}
\vec{b}^T A^2 \vec{c} &= \frac{1}{24}, & \vec{b}^T A\vec{c}^2 &= \frac{1}{12} , \\
\vec{b}^T C A\vec{c} &= \frac{1}{5}, & \vec{b}^T \vec{c}^3 &= \frac{1}{4} ,
\end{align*}
\]

5th order :

\[
\begin{align*}
\vec{b}^T C^3 \vec{c} &= \frac{1}{5}, & \vec{b}^T C^2 A\vec{c} &= \frac{1}{10}, & \vec{b}^T CAC\vec{c} &= \frac{1}{15} , \\
\vec{b}^T C^2 A^2 \vec{c} &= \frac{1}{20}, & \vec{b}^T DA\vec{c} &= \frac{1}{20}, & \vec{b}^T AC^2 \vec{c} &= \frac{1}{20} , \\
\vec{b}^T ACA\vec{c} &= \frac{1}{40}, & \vec{b}^T A^2 C\vec{c} &= \frac{1}{60}, & \vec{b}^T A^3 \vec{c} &= \frac{1}{120} ,
\end{align*}
\]

where \( D = \text{diag}(A\vec{c}) \) and \( C = \text{diag}(\vec{c}) \). We call a matrix \( A \) to be a feasible scheme if it satisfies all the constraints in the equation (4.24). So a feasible scheme only satisfies order conditions and the weak stage order conditions. However, a useful time stepping scheme possesses other desirable properties. Our goal is to construct \( p \)th order schemes with WSO \( q \), which minimize the \( \| \|_\ell^2 \) norm of the residual error of \((p + 1)\)th order conditions, having (i) stiff accuracy, (ii) \( A \)-stability, (iii) \( \vec{c} > 0 \), and (iv) \( a_{ii} > 0 \). All together, finding a scheme satisfying all the constraints can be formulated as a minimization problem:

(P) Minimize: \( \| (p + 1)\)th order conditions\( \|_\ell^2 \) \hspace{1cm} (4.25)

Subject to: \( \beta_j(A, \vec{b}, \vec{c}) = 0, \quad \gamma_{j,k}(A, \vec{b}, \vec{c}) = 0, \) and properties ((i) – (iv)) . \hspace{1cm} (4.26)

Due to the high dimension of the search space and the ill-conditioning constraints, finding feasible schemes does not work by brute force search methods. So a systematic approach was needed to solving this problem. As a first step
towards finding high-order DIRK schemes with high WSO, a simpler criterion of WSO was considered known as WSO eigenvector criterion [42].

**Definition 4.5.** An RK scheme satisfies the WSO eigenvector criterion of order $q_e$ if for each $1 \leq j \leq q_e$, there exists $\mu_j$ such that $A\vec{\tau}^{(j)} = \mu_j\vec{\tau}^{(j)}$ and moreover, $\vec{b}^T \vec{\tau}^{(j)} = 0$.

The WSO eigenvector criterion of order $q_e$ guarantees of WSO at least $q_e$. This criterion was used to solve the systems of polynomials in equation (4.24) and the authors in [42] were able to construct DIRK schemes with order up to 4 and WSO up to 3. They also showed a limitation of WSO eigenvector criterion that DIRK schemes with WSO beyond 3 can not be constructed by using WSO eigenvector criterion.

**Theorem 4.2.** DIRK schemes with invertible $A$ have $q_e \leq 3$.

The main idea of this theorem’s proof is to study the WSO eigenvector relation as in Def. 4.5 row by row. This is possible because $\vec{\tau}^{(j)}$ only depends on $A$, the equation $A\vec{\tau}^{(j)} = \mu_j\vec{\tau}^{(j)}$ is independent of $\vec{b}$, and $A$ is lower triangular. $A$ being lower triangular matrix allows us to study the equation $A\vec{\tau}^{(j)} = \mu_j\vec{\tau}^{(j)}$ row by row. Each row of this equation imposes necessary conditions on the solvability of the equations for $j = 2, 3, \cdots, q$. From the first row, we get $\mu_j = a_{11}$, and the second row yields a homogeneous equation in the variables $(a_{11}, a_{21}, a_{22})$ for each $j$. It turns out that the equations for $j \in \{2, 3\}$ have common solutions, and with those solutions, non-equal-time [42] feasible schemes can be constructed. However, the equations for $j \in \{2, 3, 4\}$ have common solutions that produce only equal-time ($\vec{c} = \nu\vec{e}$ where $\nu \in \mathbb{R}$) [42] schemes, which are not useful due to their limitation to high-order. Hence, we can not construct WSO 4 or higher schemes with the WSO eigenvector criterion.

To construct WSO 4 or higher DIRK schemes, it is now necessary to use the generalized WSO criterion. Generalized WSO criterion makes the systems in the equation (4.24) more ill-conditioned, making the problem even harder.
to solve by brute search methods. To overcome the limitation of the WSO eigenvector criterion, we propose two steps; (i) based on the generalized WSO criterion, construct a feasible solution, which might not be optimal, and (ii) use an optimization routine to get other desirable properties. In the next section, we present the theory of the WSO that enables us to find feasible solutions with WSO four or higher and sufficient regular orders.

4.4 WSO Theory

The polynomial equations system (4.24) becomes more complicated for weak stage orders beyond 3, and brute force search methods fail to find a feasible solution for the system. To devise DIRK schemes with WSO order beyond 3, we consider the general WSO criterion and exploit the problem’s linear algebra structure to find out some of the necessary conditions that need to be satisfied for the solvability of the system (4.24). The main idea is to use the necessary conditions imposed by the order conditions and the WSO conditions to give it a better chance for the solvability. First, we consider the set of polynomial equations given by WSO conditions and derive a few necessary conditions that have to be satisfied by a DIRK scheme with high weak stage order. Let

\[ Y := \text{Span}\left\{ \vec{b}, A^T \vec{b}, \ldots, (A^T)^{s-1} \vec{b} \right\} \]

be the left-Krylov subspace generated by \( A^T \) and \( \vec{b} \). An \( s \)-stage scheme satisfying WSO equations of order \( q \) implies that \( Y \perp K_q \) (by WSO equations (4.23) and the Cayley-Hamilton theorem) and hence

\[ \dim(Y) + \dim(K_q) \leq s. \]  

(4.27)

Furthermore, we can prove the following:

**Lemma 4.1.** Let \( d = \dim Y, m = \dim K_q, \) and \( \ell = s - (d + m) \). Then there exist polynomials \( O, P, \) and \( Q \) of degrees \( \ell, m, \) and \( d \), respectively, such that \( \vec{b}^T Q(A) = 0, \) \( P(A) \vec{r}^{(k)} = 0, \) for \( 2 \leq k \leq q, \) and \( \text{char}(A) = P(A)Q(A)O(A). \)
Proof. \( d = \dim Y \) implies that the first \( d \) vectors \( \{\vec{b}, A^T\vec{b}, \ldots, (A^T)^{d-1}\vec{b}\} \) are linearly independent. Then there exist \( a_0, a_1, \ldots, a_{d-1} \) such that

\[
(A^T)^d\vec{b} = a_0\vec{b} + a_1A^T\vec{b} + \cdots + a_{d-1}(A^T)^{d-1}\vec{b},
\]

\[
((A^T)^d + \alpha_{d-1}(A^T)^{d-1} + \cdots + \alpha_1A^T + \alpha_0I)\vec{b} = 0,
\]

\[
\vec{b}^T(\vec{b} + \alpha_{d-1}\vec{b} + \cdots + \alpha_1A + \alpha_0I) = 0,
\]

where \( \alpha_i = -a_i \) for \( i = 0, 1, \ldots, d-1 \). Let \( Q(A) = A^d + \alpha_{d-1}A^{d-1} + \cdots + \alpha_1A + \alpha_0I \), then \( Q \) is a degree \( d \) polynomial such that \( \vec{b}^TQ(A) = 0 \). Note that, there is no lower degree polynomial \( Q \) in \( A \) that satisfies \( \vec{b}^TQ(A) = 0 \) as otherwise that would violate \( \dim Y = d \). Now we factorize the polynomial \( Q \) into linear factors as \( Q(A) = (A - \mu_1I) \cdots (A - \mu_dI) \). Then \( Q(A^T)\vec{b} = 0 \) implies

\[
(A^T - \mu_1I)(A^T - \mu_2I) \cdots (A^T - \mu_dI)\vec{b} = 0.
\]

If \( \det(A - \mu_jI) \neq 0 \) for some \( j \), then multiplying through by \( (A^T - \mu_jI)^{-1} \) we obtain

\[
(A^T - \mu_1I) \cdots (A^T - \mu_{j-1}I)(A^T - \mu_{j+1}I) \cdots (A^T - \mu_dI)\vec{b} = 0,
\]

which is a contradiction to the fact that \( Q(A) \) is the lowest degree polynomial such that \( Q(A^T)\vec{b} = 0 \). This proves \( \det(A - \mu_jI) = 0 \) for all \( j = 1, 2, \ldots, d \), and hence \( Q(A) \) is a factor of the \( \text{char}(A) \).

The subspace \( K_q \) can be written as \( K_q = \text{span}\{K^1, K^2, \ldots, K^q\} \), where \( K^j = \text{span}\{\tau_j, A\tau_j, \ldots, A^{s-1}\tau_j\} \), for \( j = 1, 2, \ldots, q \). By similar arguments above there exists a minimal polynomial \( P_j \) of degree less than equal to \( m \) (as \( m = \dim K_q \)) such that \( P_j(A)\tau_j = 0 \) for all \( j = 1, 2, \ldots, q \). Now let \( P \) be the minimal polynomial such that each \( P_j \) divides \( P \) for all \( j = 1, 2, \ldots, q \). Then by construction of \( P \), we get \( P(A)\tau_j = 0 \) for all \( j = 1, 2, \ldots, q \). The polynomials \( P_j \), \( P \) are being minimal and \( \dim K_q = m \) imply degree of \( P \) is \( m \). Define \( O(A) := \frac{\text{char}(A)}{P(A)Q(A)} \), then degree of the polynomial \( O \) is \( l = s - (d + m) \) and we get \( \text{char}(A) = P(A)Q(A)O(A) \). \( \blacksquare \)
Before we present another necessary condition for DIRK schemes with high weak stage order, we provide an observation about the lower triangular matrices and introduce some notations to be used in our next lemma. For a lower triangular matrix $A$, the first $n$ (≤ $s$) components of the vectors $\vec{\tau}_j$ and $\vec{c}$ are functions of the upper $n \times n$ sub-block of the matrix $A \in \mathbb{R}^{s \times s}$ only. We define two maps $[A]_n : \mathbb{R}^{s \times s} \to \mathbb{R}^{n \times n}$ and $[\vec{c}]_n : \mathbb{R}^s \to \mathbb{R}^n$ by

$$[A]_n := \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \ddots & \vdots \\ \vdots & & \ddots & a_{n1} \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \quad \text{and} \quad [\vec{c}]_n := \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}.$$ 

For a lower triangular matrix $A \in \mathbb{R}^{s \times s}$, it is easy to see now that $[A^k]_n = ([A]_n)^k$, $[\vec{c}^k]_n = ([\vec{c}]_n)^k$, and hence the first $n$ components of the vectors $\vec{\tau}_j$ are functions of the upper $n \times n$ sub-matrix of the matrix $A$ only.

Now we conjecture that for each $m$ there exists a $q(m)$ such that we can devise DIRK schemes only up to WSO $q(m)$, and in that case the roots the polynomial $P$ must contain the upper diagonal entries $\{a_{11}, a_{22}, \ldots, a_{mm}\}$ of $A$. It follows from Theorem 4.2 that $q(1) \leq 3$, and the only root of $P$ is $a_{11}$. Now we provide a proof of the second part of the conjecture above corresponding to $m = 2$ in terms of the following lemma.

**Lemma 4.2.** A necessary condition to develop DIRK schemes with WSO beyond 3 with $m = 2$ is that the roots of the polynomial $P$ must contain the upper diagonal entries $\{a_{11}, a_{22}\}$ of $A$, where $\text{degree}(P) = 2$.

**Proof.** From Lemma 4.1, we have $P(A)\vec{\tau}^{(k)} = 0$ for all $k = 2, \ldots, q$ ($q > 3$). Factoring the degree 2 polynomial $P$, we obtain

$$(A - \mu_1 I)(A - \mu_2 I)\vec{\tau}^{(k)} = 0, \text{ for } k = 2, 3, \ldots, q. \quad (4.28)$$

As the matrix $(A - \mu_1 I)(A - \mu_2 I)$ is upper triangular and the first $n$ components of $\vec{\tau}^{(k)}$ depend only on the $n$ upper rows of $A$, we therefore study the above
equation row by row applying the operator $\lfloor \cdot \rfloor_n$ for $n = 1, 2, \ldots, s$. Applying the operator $\lfloor A \rfloor_1$ to the equation (4.28), we get

\[
\lfloor (A - \mu_1 I)(A - \mu_2 I)\tilde{x}^{(k)} \rfloor_1 = \lfloor 0 \rfloor_1,
\]

\[
(a_{11} - \mu_1)(a_{11} - \mu_2) \left(1 - \frac{1}{k}\right) a_{11}^k = 0.
\]

We assume $k \neq 1$ and $a_{11} \neq 0$ and obtain either $\mu_1 = a_{11}$ or $\mu_2 = a_{11}$. Without loss of generality we let $\mu_1 = a_{11}$. Now applying the operator $\lfloor A \rfloor_2$ to the equation (4.28) yields

\[
\lfloor (A - \mu_1 I)(A - \mu_2 I)\tilde{x}^{(k)} \rfloor_2 = \lfloor 0 \rfloor_2,
\]

\[
\lfloor (A - a_{11} I)\lfloor (A - \mu_2 I)\rfloor_2 \tilde{x}^{(k)} \rfloor_2 = \lfloor 0 \rfloor_2.
\]

If $\mu_2 \notin \{a_{11}, a_{22}\}$, then $\lfloor (A - \mu_2 I) \rfloor_2$ is invertible. Multiplying by $\lfloor (A - \mu_2 I) \rfloor_2^{-1}$ to above equation we get

\[
\lfloor (A - a_{11} I) \rfloor_2 \lfloor \tilde{x}^{(k)} \rfloor_2 = \lfloor 0 \rfloor_2.
\]

But this is exactly the WSO eigenvector criterion and we know this criterion is limited to WSO 3 by Theorem 4.2. So, to obtain DIRK scheme with WSO beyond 3 we must have $\mu_2 \in \{a_{11}, a_{22}\}$. Hence we have $\{\mu_1, \mu_2\} \subseteq \{a_{11}, a_{22}\}$. 

So far we provided necessary conditions imposed by the weak stage order equations. Now we turn our attention on the regular order conditions and observe that satisfying the $p$th order conditions puts restrictions on the dimension of the space $Y$. This leads to the following lemma on the bound of the dimension of $Y$ in terms of the order $p$:

**Lemma 4.3.** For an $s$-stage, $p$th order RK scheme, \( \dim Y \geq \frac{p}{2} \) and hence \( \dim K_q \leq s - \frac{p}{2} \).

**Proof.** From Lemma 4.1 it follows that there exists a polynomial $Q$ of degree $d$ with coefficients $\bar{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{d-1})^T \in \mathbb{R}^d$ such that

\[
\bar{b}^T \left( A^d + \alpha_{d-1} A^{d-1} + \cdots + \alpha_1 A + \alpha_0 I \right) = \bar{0},
\]

(4.29)
where \( d = \text{dim} \ Y \). Since the scheme \((A, \bar{b}, \bar{c})\) satisfies the \( p \)th order conditions, we have the linear order conditions:

\[
\bar{b}^T A^j \bar{c} = \frac{1}{(j+1)!}, \quad 0 \leq j \leq p - 1. \tag{4.30}
\]

Let us assume for a moment that \( p = 2d \). Then, multiplying equation (4.29) by \( A^j \bar{c} \) for powers of \( 0 \leq \ell \leq d - 1 \) and using the linear order conditions from (4.30) yield \( d \) linear equations for \( \alpha \):

\[
\frac{1}{(d+1)!} \alpha_{d-1} + \cdots + \frac{1}{2!} \alpha_1 + \alpha_0 = 0,
\]

by (4.30) for \( 0 \leq j \leq d \),

\[
\frac{1}{(d+2)!} \alpha_{d-1} + \cdots + \frac{1}{3!} \alpha_1 + \frac{1}{2!} \alpha_0 = 0,
\]

by (4.30) for \( 1 \leq j \leq d + 1 \),

\[
\vdots
\]

\[
\frac{1}{(2d)!} \alpha_{d-1} + \cdots + \frac{1}{(d+1)!} \alpha_1 + \frac{1}{d!} \alpha_0 = 0,
\]

by (4.30) for \( d - 1 \leq j \leq 2d - 1 \).

This yields a linear system for \( \bar{\alpha} \):

\[
H_d \bar{\alpha} = -\bar{h}_d, \tag{4.31}
\]

where

\[
H_d = \begin{pmatrix}
\frac{1}{d!} & \frac{1}{d!} & \frac{1}{d!} & \cdots & \frac{1}{d!} \\
\frac{1}{d!} & \frac{1}{d!} & \frac{1}{d!} & \cdots & \frac{1}{d!} \\
\frac{1}{d!} & \frac{1}{d!} & \frac{1}{d!} & \cdots & \frac{1}{d!} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{d!} & \frac{1}{(d+1)!} & \frac{1}{(d+2)!} & \cdots & \frac{1}{(2d-1)!}
\end{pmatrix}
\quad \text{and} \quad
\bar{h}_d = \begin{pmatrix}
\frac{1}{(d+1)!} \\
\frac{1}{(d+2)!} \\
\frac{1}{(d+3)!} \\
\vdots \\
\frac{1}{(2d)!}
\end{pmatrix}. \tag{4.32}
\]

Now, \( H_d \in \mathbb{R}^{d \times d} \) is a Hankel matrix, and it is well-known that the determinant of a Hankel matrix is nonzero. Hence the vector \( \bar{\alpha} \) is uniquely solvable when \( p = 2d \). It is also easy to see that the corresponding system for \( p < 2d \) has solutions for the vector \( \bar{\alpha} \) with a \((2d - p)\) dimensional null space.
Now consider the case when $p > 2d$ and assume $p = 2d + 1$. In such a situation, the corresponding system for $\vec{\alpha}$ will have an added row to the equation (4.31):

$$
\begin{pmatrix}
\frac{1}{1!} & \frac{1}{2!} & \frac{1}{3!} & \cdots & \frac{1}{d!} \\
\frac{1}{2!} & \frac{1}{3!} & \frac{1}{4!} & \cdots & \frac{1}{(d+1)!} \\
\frac{1}{3!} & \frac{1}{4!} & \frac{1}{5!} & \cdots & \frac{1}{(d+2)!} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{d!} & \frac{1}{(d+1)!} & \frac{1}{(d+2)!} & \cdots & \frac{1}{(2d)!}
\end{pmatrix}
\begin{pmatrix}
\alpha_0 \\
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_{d-1}
\end{pmatrix}
= -
\begin{pmatrix}
\frac{1}{(d+1)!} \\
\frac{1}{(d+2)!} \\
\frac{1}{(d+3)!} \\
\vdots \\
\frac{1}{(2d)!} \\
0
\end{pmatrix}.
\quad (4.33)
$$

If the system (4.33) were solvable, then $\vec{\alpha}$ also solves the following equation obtained by adding an extra column to the matrix in (4.33) and extra variable 0 in the column of $\vec{\alpha}$:

$$
\begin{pmatrix}
\frac{1}{1!} & \frac{1}{2!} & \frac{1}{3!} & \cdots & \frac{1}{d!} & \frac{1}{(d+1)!} \\
\frac{1}{2!} & \frac{1}{3!} & \frac{1}{4!} & \cdots & \frac{1}{(d+2)!} & \frac{1}{(d+1)!} \\
\frac{1}{3!} & \frac{1}{4!} & \frac{1}{5!} & \cdots & \frac{1}{(d+3)!} & \frac{1}{(d+1)!} \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
\frac{1}{d!} & \frac{1}{(d+1)!} & \frac{1}{(d+2)!} & \cdots & \frac{1}{(2d)!} & \frac{1}{(d+1)!} \\
\frac{1}{(d+1)!} & \frac{1}{(d+2)!} & \frac{1}{(d+3)!} & \cdots & \frac{1}{(2d)!} & \frac{1}{(d+1)!}
\end{pmatrix}
\begin{pmatrix}
\alpha_0 \\
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_{d-1} \\
0
\end{pmatrix}
= -
\begin{pmatrix}
\frac{1}{(d+1)!} \\
\frac{1}{(d+2)!} \\
\frac{1}{(d+3)!} \\
\vdots \\
\frac{1}{(2d)!} \\
0
\end{pmatrix}.
\quad .
$$

But this equation is exactly:

$$
H_{d+1} \begin{pmatrix} \vec{\alpha} \\ 0 \end{pmatrix} = - \begin{pmatrix} \vec{h}_d \\ \frac{1}{(2d+1)!} \end{pmatrix}
$$

which has the unique (since $H_{d+1}$ is invertible) solution $(\vec{0}_d, -1)^T$ ($\vec{0}_d \in \mathbb{R}^d$ is the zero vector), and this cannot be ever equal to $(\vec{\alpha}^T, 0)^T$. Hence the system (4.33) is not solvable when $p = 2d + 1$. We can make similar argument and prove that the system is not solvable for any arbitrary $p > 2d$, forcing $p \leq 2d$. ■
Remark 4.1. Note that for an $s$-stage, $p$th order RK scheme, the bound of $\dim Y$ in terms of the order $p$ becomes $p+1 \leq 2d$ when $p$ is odd, because $p \neq 2d$ in this case.

Lemma 4.4. For DIRK schemes we can further tighten the bounds to $p+2 \leq 2 \dim Y$ when $p$ is even and $4 \leq p \leq 10$.

Proof. The eigenvalues of a DIRK scheme $A$ are the diagonal entries of the matrix $A$ that are all real numbers. From Lemma 4.1, we have a polynomial $Q(A)$ of degree $d = \dim Y$ such that $Q$ divides the characteristic polynomial of $A$. So all the roots of the polynomial $Q$ must be real numbers for a DIRK scheme. Now suppose $p = 2d$, then $d = 2, 3, 4, 5$, for $p = 4, 6, 8, 10$ (by the assumption), respectively. For each value of $d$, we can now determine the polynomial $Q$ by solving the equation (4.31) for the coefficients $\alpha$ of $Q$ from Lemma 4.3. The calculation gives

\begin{align*}
\dim Y = 2 : \quad & Q(\lambda) = \lambda^2 - \frac{1}{2} \lambda + \frac{1}{12}, \\
\dim Y = 3 : \quad & Q(\lambda) = \lambda^3 - \frac{1}{2} \lambda^2 + \frac{1}{10} \lambda - \frac{1}{120}, \\
\dim Y = 4 : \quad & Q(\lambda) = \lambda^4 - \frac{1}{2} \lambda^3 + \frac{3}{28} \lambda^2 - \frac{1}{84} \lambda + \frac{1}{1680}, \\
\dim Y = 5 : \quad & Q(\lambda) = \lambda^5 - \frac{1}{2} \lambda^4 + \frac{1}{9} \lambda^3 - \frac{1}{72} \lambda^2 + \frac{1}{1008} \lambda - \frac{1}{30240}.
\end{align*}

It is straightforward to check using MATLAB that all these polynomials have at least one complex root. This contradicts our assumption and hence we must have $p + 1 \leq 2d$ by Lemma 4.3. Moreover, because $p + 1$ is odd for $p = 4, 6, 8, 10$, we get $p + 2 \leq 2d$.

Remark 4.2. This lemma might be true when $p$ is even and $p \geq 12$. We only checked the roots of the polynomial $Q$ up to degree 5 that enables us to conclude the lemma for $p$ even and $4 \leq p \leq 10$, which is enough for our purpose.
4.5 DIRK Schemes with WSO Beyond Three

This section describes how the general theory of WSO guides us to find DIRK schemes with high WSO. To find a \( p \)th \( (p \geq 4) \) order DIRK scheme with WSO \( q \) \( (q \geq 4) \) for a sufficiently large \( s \), denoted by DIRK-\( (s, p, q) \), we must have \( \dim K_q \leq s - \frac{p+1}{2} \) from Remark 4.1 and Lemma 4.4. A priori, we have lots of choices for \( \dim K_q \), and we decided to consider the cases depending on \( \dim K_q \). Furthermore, we assume \( Y \bigoplus K_q = \mathbb{R}^s \) and all the diagonal entries of the matrix \( A \) are restricted to be distinct and nonzero for our study.

**Case 1: \( \dim K_q = 1 \)**

In this case we get \( \dim K_q = 1 \), and \( \dim Y = s - 1 \). From Lemma 4.1, it follows that

\[
(A - \mu_s I)\vec{\tau}(k) = 0, \quad \text{for} \quad k = 2, 3, \ldots, q, \quad (4.34)
\]

where \( \mu_s \) is one of the diagonal entries of \( A \), and Lemma 4.2 implies that \( \mu_s = a_{11} \). This equation (4.34) is the same as the WSO eigenvector criterion of order \( q \), and we know that DIRK schemes with invertible \( A \) are limited to WSO 3. So it is natural to explore \( \dim K_q = 2 \) as our next case.

**Case 2: \( \dim K_q = 2 \)**

Now we have \( \dim K_q = 2 \), and \( \dim Y = s - 2 \). From Lemma 4.1, we obtain

\[
(A - \mu_{s-1} I)(A - \mu_s I)\vec{\tau}(k) = 0, \quad \text{for} \quad k = 2, 3, \ldots, q, \quad (4.35)
\]

where \( \mu_{s-1} \) and \( \mu_s \) are from the diagonal entries of \( A \), and Lemma 4.2 implies that \( \{\mu_{s-1}, \mu_s\} = \{a_{11}, a_{22}\} \). As \( (A - a_{11} I)(A - a_{22} I) = (A - a_{22} I)(A - a_{11} I) \) and it is easy to see \( \ker (A - a_{11} I) \cap \ker (A - a_{22} I) = \{0\} \) as \( a_{11} \neq a_{22} \) by our assumption, it follows that

\[
\vec{\tau}(k) = \beta_1^{(k)} \vec{w}^{(1)} + \beta_2^{(k)} \vec{w}^{(2)}, \quad \text{for} \quad k = 2, 3, \ldots, q,
\]

where \( \beta_i^{(k)} \)'s are constants and \( \vec{w}^{(1)}, \vec{w}^{(2)} \) are the eigenvectors of the matrix \( A \) corresponding to the eigenvalues \( a_{11} \) and \( a_{22} \), respectively. In that case, the space \( K_q \) becomes \( \text{span}\{\vec{w}^{(1)}, \vec{w}^{(2)}\} \), which is an \( A \)-invariant subspace. If we can construct \( A \) such that \( \vec{\tau}(k) \in \text{span}\{\vec{w}^{(1)}, \vec{w}^{(2)}\} \), then WSO \( q \) requires only
two conditions to be satisfied: $\tilde{b}^T \tilde{w}^{(1)} = 0$ and $\tilde{b}^T \tilde{w}^{(2)} = 0$, as

$$\tilde{b}^T A^j \tilde{r}^{(k)} = \tilde{b}^T A^j (\beta_1^{(k)} \tilde{w}^{(1)} + \beta_2^{(k)} \tilde{w}^{(2)}) = \tilde{b}^T (\beta_1^{(k)} A^j \tilde{w}^{(1)} + \beta_2^{(k)} A^j \tilde{w}^{(2)})$$

$$= \beta_1^{(k)} a_{11} \tilde{b}^T \tilde{w}^{(1)} + \beta_2^{(k)} a_{22} \tilde{b}^T \tilde{w}^{(2)} = 0, \text{ for } k = 2, \ldots, q, \text{ and } \forall \ j \geq 0.$$

The WSO conditions and the classical order conditions are not independent of each other. If $A$ is constructed such that all the WSO equations are satisfied, then having the classical orders 4 and 5 only requires the following equations to be satisfied, respectively:

$$\tilde{b}^T [\tilde{w}^{(1)}, \tilde{w}^{(2)}, \tilde{c}, \tilde{c}, \tilde{c}, CA\tilde{c}] = \begin{bmatrix} 0, 0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{8} \end{bmatrix}, \text{ and } \quad (4.36)$$

$$\tilde{b}^T [\tilde{w}^{(1)}, \tilde{w}^{(2)}, \tilde{c}, \tilde{c}, \tilde{c}, CA\tilde{c}, \tilde{c}, CA^2\tilde{c}, CAC\tilde{c}, \text{diag}(Ac)Ac, CA^2c, ACA\tilde{c}, CA\tilde{c}, ACA\tilde{c}, \text{diag}(Ac)Ac, CA^2c, ACA\tilde{c}] =$$

$$\begin{bmatrix} 0, 0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{15}, \frac{1}{20}, \frac{1}{30}, \frac{1}{40} \end{bmatrix}. \quad (4.37)$$

With these insights, we now determine $p$th order DIRK schemes with weak stage order $q$ and other desired properties by solving the minimization problem $(P)$ in (4.26):

$$(P) \quad \text{Minimize: } \| (p + 1) \text{th order conditions} \|_{l^2}$$

Subject to:

$$(i) \quad \beta_j(A, \tilde{b}, \tilde{c}) = 0, \text{ for } j = 1, 2, \ldots, n_q$$

$$(ii) \quad \gamma_{j,k}(A, \tilde{b}, \tilde{c}) = 0, \text{ for } 0 \leq j \leq s - 1, \ 2 \leq k \leq q$$

$$(iii) \quad \text{stiff accuracy}, \ (iv) \quad A - \text{stability}$$

$$(v) \quad \tilde{c} > 0, \ (vi) \ a_{ii} > 0.$$  

We solve this minimization problem in two major steps:

1. **Construction of a feasible scheme**: Find a feasible scheme, i.e., a scheme that satisfies the constraints (i), (ii), and (iii).

2. **Optimization**: Using the previous step’s output as a starting point of an optimization routine, minimize the objective function subject to the conditions ((i) – (vi)) and get an optimal scheme.
1. **Construction of a feasible scheme:** In the first step of our searching algorithm, our goal is to find a stiffly accurate scheme that satisfies all the order conditions and the weak stage order conditions. We construct a matrix $A$ such that $\vec{\tau}(k) = \beta_1^{(k)} \vec{w}^{(1)} + \beta_2^{(k)} \vec{w}^{(2)}$, for $k = 2, \ldots, q$, ($p \geq q = 4$ or $5$) satisfying (4.36) or (4.37) depending on the classical order $4$ and $5$, respectively. We complete this important task in two sub-steps as follows:

(1A) First, we do a random search to find schemes that satisfy the conditions 

$$((i) - (iii)).$$

We accept those schemes whose entries are bounded by $20$ (ad hoc choice) and satisfy the conditions by a reasonable accuracy. Here we state the key steps to construct a feasible scheme.

1. Determine the coefficients $\beta_1^{(k)}, \beta_2^{(k)}$ from the first two components of the equation $\vec{\tau}(k) = \beta_1^{(k)} \vec{w}^{(1)} + \beta_2^{(k)} \vec{w}^{(2)}$ for $k = 2, 3, 4$.

2. Substitute $\beta_1^{(k)}, \beta_2^{(k)}$ to rest of the components of the equations $\vec{\tau}(k) = \beta_1^{(k)} \vec{w}^{(1)} + \beta_2^{(k)} \vec{w}^{(2)}$ for each $k$, to get $(q - 1)(s - 2)$ equations.

3. Now consider the third component of $\vec{\tau}(k) = \beta_1^{(k)} \vec{w}^{(1)} + \beta_2^{(k)} \vec{w}^{(2)}$ for $k = 2, 3, 4$, which yields $(q - 1)$ equations in six variables $(a_{11}, a_{21}, a_{22}, a_{31}, a_{32}, a_{33})$.

4. We can parameterize these $(q - 1)$ equations for $q = 4$, and express $a_{21}, a_{31}$ and $a_{32}$ in terms of $a_{11}, a_{22},$ and $a_{33}$.

5. Using the parameterization and choosing $a_{11}, a_{22},$ and $a_{33}$ randomly, we construct the upper $3 \times 3$ block of the matrix $A$ for $q = 4$. For $q = 5$, we first select $a_{11}, a_{22},$ and $a_{33}$ satisfying the third component equation corresponding to $q = 5$, and then complete the upper $3 \times 3$ block using the parameterized equations for $a_{21}, a_{31},$ and $a_{32}$.

6. Then we consider the $(q - 1)$ equations from the fourth component of $\vec{\tau}(k) = \beta_1^{(k)} \vec{w}^{(1)} + \beta_2^{(k)} \vec{w}^{(2)}$ for $k = 2, \ldots, q$. Using the values of the top $3 \times 3$ block, we get three equations in four variables $(a_{41}, a_{42}, a_{43}, a_{44})$, and obtain the fourth row of $A$ by solving the three equations numerically.
(7) Similarly, we construct the other rows one by one except the last row.

(8) To construct a stiffly accurate scheme, we assume \( b^T = (a_{s1}, a_{s2}, ..., a_{ss}) \), and solve the last row of \( A \) by simultaneously solving \((q - 1)\) equations and equations from (4.36) or (4.37) for classical order 4 and 5, respectively.

**Remark 4.3.** Note that for \( q = 4 \), there are several sets of parameterization of \( a_{21}, a_{31}, \text{and} \ a_{32} \) in terms of \( a_{11}, a_{22}, \text{and} \ a_{33} \). Some of these parameterizations produce equal-time schemes [42]. So we need to carefully choose a nontrivial parameterization to obtain a useful, feasible matrix \( A \).

**Remark 4.4.** In MATLAB, we use the ‘sqp’ algorithm in ‘fmincon’ for faster and accurate results to solve each row. The numerical algorithm ‘fmincon’ is an optimization routine that minimizes an objective function subject to linear and nonlinear equality and inequality constraints. In our case, we consider the nonlinear equations satisfied by each row as nonlinear equality constraints and use a constant objective function in ‘fmincon’ to solve the parameters of the matrix \( A \).

(1B) Post-processing: The feasible schemes constructed by the above procedure might not have satisfactory (up to machine precision) error residuals for order conditions and WSO equations. So we perform a post-processing step: using the Gauss-Newton iteration with the constructed solution in step (1A) as a starting point, we obtain a feasible scheme by driving the residual errors up to machine precision.

**2. Optimization:** In this step, we solve the optimization problem \((P)\). We have observed that the optimization routine works better with a preprocessing step, where we achieve the inequality constraints \(((iv) - (vi))\) satisfied by the schemes on the top of the order and WSO conditions. So we perform the optimization in two steps:
Pre-processing: We provide the conditions \(((i) - (iii))\) as nonlinear equality constraints and \(((iv) - (vi))\) as inequality constraints in the optimization routine 'fmincon'. With these constraints and a constant objective function we run the optimization starting with the feasible scheme from step (1B). Hence, at the end of this step, we obtain a scheme that satisfies \(((i) - (vi))\). The conditions (v) \(\bar{c} > 0\), and (vi) \(a_{ii} > 0\) are incorporated as linear inequality constraints in the algorithm as \(\bar{c} \geq \epsilon\), and \(a_{ii} \geq \epsilon\), where \(\epsilon\) is a small positive number. We implement the linear inequality constraints in this way to define a proper feasible set for the optimization algorithm. However, the schemes that we found are safely away from the boundary of the domain of the feasible set. For 

\(A\)-stability, we consider the stability function \(R(z) := \frac{P(z)}{Q(z)} = \frac{\det(I-zA+zeb^T)}{\det(I-zA)}\) that is analytic on the left half plane due to the designing criterion of \(a_{ii} > 0\). In that case, \(A\)-stability requires either \(|R(iy)| \leq 1\) for all \(y \in \mathbb{R}\), or the E-polynomial \(E(y) := |Q(iy)|^2 - |P(iy)|^2\), \(y \in \mathbb{R}\) satisfies \(E(y) \geq 0\) for all \(y \in \mathbb{R}\). We take the conditions \(|R(z)| \leq 1\) for a finite number of points along the imaginary axis and incorporate them as nonlinear inequality constraints in the optimization routine.

In the last step, we take the output of the step (2A) as an initial guess and minimize a proxy of the truncation error, the \(\| \cdot \|_\ell^2\) norm of the \((p + 1)\)th order conditions subject to the constraints \(((i) - (vi))\). For the optimization routine, we use the 'fmincon' with the gradient-based 'sqp' algorithm. Our observation is that starting from an initial guess satisfying all the constraints, the optimizer either finds a local minimum satisfying all the constraints, or the optimization routine stops because the objective function is non-decreasing in feasible directions. Thus the algorithm sometimes improves the proxy of the truncation error producing a local minimum, and sometimes it does not improve much.

We find multiple schemes by rerunning the above procedure, and we choose the schemes that have the smallest error constants.
We provide three schemes (see §4.9), each of which has the smallest truncation error among all the schemes that we found for each triple \((s, p, q)\): DIRK-(7, 4, 4), DIRK-(13, 5, 4), and DIRK-(13, 5, 5) that are (a) stiffly accurate, (b) having \(\vec{c} > 0\), (c) with positive diagonal entries, and (d) A-stable. We also demonstrate the A-stability of these schemes by (a) plotting the stability region, or (b) plotting the stability function along the imaginary axis or plotting the \(E\)-polynomial. Figure 4.6 shows the stability region of these three schemes. The left panel of the Figure 4.7 shows that \(E\)-polynomial for each of these schemes is positive. The right panel shows the stability function for each of these schemes is bounded above by one on the imaginary axis, confirming these schemes’ A-stability.
Figure 4.7: Left: E-polynomials for different DIRK methods are plotted with $y$-axis in log scale. Right: $|R(iy)|$ for $y \in \mathbb{R}$ for different DIRK schemes, and $|R(iy)| \leq 1$ along the $y$-axis, hence these methods are $A$-stable.

4.6 Numerical Results for Problems Covered by the WSO Theory

In this section, we test our schemes on some linear test problems and demonstrate the order reduction phenomenon’s remedies. The weak stage order theory only guarantees to remove order reduction for linear problems with time-independent differential operators [70]. Here we confirm the order of accuracy of our newly developed high WSO schemes by testing them on the problems involving time-independent differential operators. The schemes are parameterized by $(s, p, q)$, where $s$ is the number of stages, $p$ is the classical order of accuracy, and $q$ is the scheme’s weak stage order. For ODE test problems, we consistently apply schemes where $p = q$, and for PDE IBVs, we test all our schemes. Although WSO $(p - 1)$ schemes are enough to remedy the order reduction for PDE IBVs, they do not recover the full order of convergence for derivatives. Schemes with $p = q$ assure the full order of convergence for both function values and derivatives, shown for several problems in this section.
4.6.1 Linear ODE test problem

Consider the Prothero-Robinson test problem

\[ u' = \lambda (u - \phi(t)) + \phi'(t); \quad u(0) = u_0, \text{ with } \text{Re}(\lambda) \leq 0, \]  

with the true solution \( \phi(t) = e^{-t} \sin(10t) + \cos(20t) \), the stiffness parameter \( \lambda = -10^4 \), and the initial condition \( u(0) = \phi(0) \). The problem is solved using two 4th order DIRK schemes with WSO 4 and 1, and two 5th order schemes with WSO 5 and 1 up to the final time \( T = 10 \). We observe that the schemes with WSO 1 exhibit first order of convergence in the stiff region as expected. The 4th order DIRK scheme with WSO 4 and the 5th order DIRK scheme with WSO 5 recover their full order of convergence. Notice that even in the non-stiff region, schemes with higher WSO generate significantly lower error constants.

Figure 4.8: Error convergence for the Prothero-Robinson test problem using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (blue circles) and WSO 1 (red squares) (left), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (blue circles) and WSO 1 (red squares) (right).
4.6.2 Heat equation

We study the diffusive 1D heat equation as our first linear PDE test problem. Consider the 1D heat equation

\[ u_t = u_{xx} + f \quad \text{for} \quad (x, t) \in (0, 1) \times (0, 1], \quad u = g \quad \text{on} \quad \{0, 1\} \times (0, 1), \quad (4.39) \]

with the forcing function, the boundary conditions (b.c.) and the initial condition (i.c.) such that the true solution is \( u(x, t) = \cos(30t) \sin(10x + 10) \). We use a 4th order centered finite difference on a fine grid with 10000 cells to isolate the temporal errors due to the DIRK schemes from the spatial approximation error. Thus we obtain the actual behavior of the error convergence of the time-stepping schemes. All the errors are measured in the maximum norm in space. The problem is solved up to final time \( T = 1 \) by different DIRK schemes with high WSO, and Figure 4.9 shows the error convergence of function values and derivatives for heat equation by DIRK-(7, 4, 4) (left), DIRK-(13, 5, 4) (middle), and DIRK-(13, 5, 5) (right). We ensure that the order of convergence for the function values is \( \min\{p, q + 1\} \) \([70]\) for all the methods applied here. The time stepping schemes produce boundary layers (BLs) of width \( O(\Delta t^{0.5}) \) due to second order term \( u_{xx} \) in the equation. As a result, the order of convergence is reduced by half for derivatives when \( q < p \). Thus, we obtain full
order of convergence for function values and derivatives by DIRK-(7, 4, 4) and DIRK-(13, 5, 5), and full order for function and half order loss in derivative by DIRK-(13, 5, 4).

### 4.6.3 Schrödinger equation

We consider the Schrödinger equation as an example of dispersive problem that is given by

$$ u_t = \frac{i \omega}{k^2} u_{xx} \text{ for } (x, t) \in (0, 1) \times (0, 1.2), \quad u = g \text{ on } \{0, 1\} \times (0, 1.2), \quad (4.40) $$

with the true solution $u(x, t) = e^{i(kx - \omega t)}$, $\omega = 2\pi$ and $k = 20$. The spatial derivative is approximated by the 4th order centered finite difference on a fine grid with 10000 cells. This problem is solved up to final time $T = 1.2$ by different DIRK schemes with high weak stage order, and the error convergence of function values and derivatives for Schrödinger equation by DIRK-(7, 4, 4) (left), DIRK-(13, 5, 4) (middle), and DIRK-(13, 5, 5) (right) are displayed in Figure 4.10. Similar error convergence results are observed as for the heat equation. As we can see, the full orders of convergence are recovered for function values and derivatives by DIRK-(7, 4, 4) and DIRK-(13, 5, 5) (schemes

![Figure 4.10: Error convergence (u blue circles; u_x red squares) for Schrödinger equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right).](attachment:figure_4.10.png)
with $p = q$), and full order for function and half order loss for derivative by the DIRK-(13, 5, 4) (scheme with $q < p$).

### 4.6.4 Advection-diffusion equation

This example demonstrates that DIRK schemes with high weak stage orders remove order reduction for problems with physical boundary layers. We consider the 1D linear advection-diffusion equation

$$u_t + u_x = \nu u_{xx} + f \quad \text{for} \quad (x, t) \in (0, 1) \times (0, 1] \, , \tag{4.41}$$

with the true solution $u(x, t) = \cos(5t) \sin(10x + 10)$, and the viscosity constant $\nu = 10^{-3}$. The advection term overwhelms the diffusion term for such a small value of $\nu$, and the boundary condition at the outflow boundary at $x = 1$ leads to a physical boundary layer of width $\mathcal{O}(\nu)$. Both the spatial derivatives are approximated by the 4th order centered finite difference on a fine grid with 10000 cells. Here we compute the numerical solutions by our schemes at $T = 1$, and show the convergence plots for function values and derivatives by DIRK-(7, 4, 4) (left), DIRK-(13, 5, 4) (middle), and DIRK-(13, 5, 5) (right) in Figure 4.11. These plots confirm that order reduction is not exhibited, and each of our schemes recovers its expected convergence orders.

![Figure 4.11: Error convergence (u blue circles; $u_x$ red squares) for the advection-diffusion equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right).](image)
We consider the linear advection problem with constant-coefficient to illustrate that the order reduction phenomenon happens more generally, i.e., in the problems with 1st order spatial derivative, and our schemes remedy them as well. The 1D linear advection equation with constant velocity 1 is given by

$$u_t + u_x = 0 \text{ for } (x,t) \in (0,1) \times (0,1],$$

with the true solution $u(x,t) = \sin(2\pi(x - t))$, and the Dirichlet’s b.c. at $x = 0$. The spatial derivative is approximated by the 4th order centered finite difference on a fine grid with 10000 cells. We solve the problem up to time $T = 1$, and report the convergence results in Figure 4.12. The boundary layers’ thickness now scales like $O(\Delta t)$ due to 1st order term $u_x$ in the equation. So we expect an order loss by one for derivatives in the convergence plot when $q < p$. The figure exactly shows that DIRK-(13, 5, 4) recovers 5th order for function values and 4th order for derivatives. Moreover, DIRK-(7, 4, 4) and DIRK-(13, 5, 5) recover their full orders of convergence for both $u$ and $u_x$. 

Figure 4.12: Error convergence ($u$ blue circles; $u_x$ red squares) for the linear advection equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right).
4.6.6  Airy’s equation

We consider the Airy’s equation

\[ u_t = u_{xxx} + f \text{ for } (x, t) \in (0, 1) \times (0, 1), \]

(4.43)

with b.c. at \( u(0) = g(t), u_x(0) = h_0(t), u_x(1) = h_1(t), \) and the manufactured solution \( u(x, t) = \cos(15t). \) Here we have 3rd order spatial derivative in the equation which is approximated by the 2nd order centered finite difference on a fine grid with 10000 cells. Solving the problem by different schemes up to final time \( T = 1, \) the error convergence results are shown in Figure 4.13. The time-stepping schemes produce numerical boundary layer whose width scales like \( O(\Delta t^{1/3}), \) leading to \( \frac{1}{3} \) order loss per derivatives when \( q < p. \) For \( p = q \) we recover the schemes’ full order of convergence for both \( u \) and \( u_x. \)

Figure 4.13: Error convergence (\( u \) blue; \( u_x \) red) for the Airy’s equation using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right).

4.6.7  Heat equation with spatially varying coefficient

So far, we have seen examples where the differential operators have all constant coefficients. To demonstrate that our schemes resolve the order reduction for more general problems, we consider the following heat equation
with spatially varying diffusion coefficient:

\[ u_t = (k(x)u_x)_x + f \text{ for } (x, t) \in (0, 1) \times (0, 1], \quad u = g \text{ on } \{0, 1\} \times (0, 1], \quad (4.44) \]

with the diffusion coefficient \( k(x) = \cos(x + 0.1) \), the forcing function \( f(x, t) \), the boundary conditions (b.c.) and the initial condition (i.c.) such that the true solution is \( u(x, t) = \cos(30t)\sin(10x + 10) \). Figure 4.14 shows that our

high weak stage schemes recover the expected convergence orders similar to the constant-coefficient heat equation.

### 4.7 Numerical Results for Problems NOT Covered by the WSO Theory

Weak stage order theory was established based on linear differential operators that do not explicitly depend on time. The newly developed schemes based on WSO theory are guaranteed to remedy the order reduction on problems whose differential operators are linear and do not explicitly depend on time (as shown in sec 4.6). However, high weak stage order schemes could

Figure 4.14: Error convergence (\( u \) blue circles; \( u_x \) red squares) for heat equation with spatially varying diffusion coefficient \( k(x) = \cos(x + 0.1) \) using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right).
sometimes remove order reduction or improve the error convergence for the problems with differential operators that are not restricted to a constant coefficient or spatially dependent. Here we demonstrate the performance of high weak stage order on a few problems on which the high weak stage order schemes are not guaranteed to resolve the issue of order reduction.

4.7.1 Stiff nonlinear ODE: Van der Pol’s equation

In this example, we choose a stiff nonlinear ODE, Van Der Pol’s equation, to demonstrate that DIRK schemes with high weak stage order do not remedy order reduction for all types of problems. The standard form of the Van Der Pol oscillator is given by

$$\frac{d^2 x}{dt^2} + \mu \left( x^2 - 1 \right) \frac{dx}{dt} + x = 0, \quad \mu > 0 ,$$  \hspace{1cm} (4.45)

where $\mu$ is the stiffness parameter. We solve this equation by writing as a system of first order differential equations

$$\begin{cases}  
\frac{dx}{dt} = y, \\
\frac{dy}{dt} = \mu (1 - x^2)y - x .  
\end{cases} \hspace{1cm} (4.46)$$

We make the problem very stiff by setting the stiffness parameter to $\mu = 500$, and solve it with initial condition $(x(0), y(0)) = (2, 0)$, and the final time $t = 10$. The numerical solutions are computed by different DIRK schemes with high WSO using different time step sizes between $\Delta t = 0.5$ and $\Delta t \approx 2.44e-04$. The implicit DIRK schemes applied to nonlinear system of the Van Der Pol’s equation produces a system of nonlinear equations for each stage at each time step. We solve that system of nonlinear equations for the stages by the standard Newton iteration method. The reference solution is calculated by the standard explicit RK4 method with a time step $\Delta t = 1e - 6$. Figure 4.15 shows the error convergence plots for two different methods DIRK-(7, 4, 4) and DIRK-(13, 5, 5). The results clearly indicate that DIRK schemes with high WSO do not remove order reduction in the stiff regime.
4.7.2 Heat equation with slowly and rapidly varying time dependent coefficient

We again consider the variable coefficient heat equation

$$u_t = (k(x,t)u_x)_x + f \text{ for } (x,t) \in (0,1) \times (0,1), \quad u = g \text{ on } \{0,1\} \times (0,1),$$

(4.47)

where $k(x,t)$ depends on time. We choose two different diffusion coefficients, one which varies slowly in time, $k(x,t) = \cos(0.1t + 0.2)$ and another oscillates rapidly in time, $k(x,t) = 1 + 0.5 \cos(100t + 0.1)$. The Figure 4.16 shows that all the schemes remove the order reduction for low oscillatory diffusion coefficient. On the other hand, for a rapidly varying time-dependent coefficient, the schemes suffer from order reduction and do not produce clear convergence results.

4.7.3 Stiff nonlinear PDE: viscous Burgers’ equation

As a stiff nonlinear PDE problem, we study the nonlinear viscous Burgers’ equation to show that DIRK schemes with high WSO do not improve the convergence rates in the stiff regime. We consider the nonlinear viscous Burgers’
Figure 4.16: Error convergence ($u$ blue circles; $u_x$ red squares) for heat equation with diffusion coefficient $k(x,t) = \cos(0.1t + 0.2)$ (top) and $k(x,t) = 1 + 0.5\cos(20t)$ (bottom) using DIRK-(7, 4, 4): 4th order DIRK scheme with WSO 4 (left), DIRK-(13, 5, 4): 5th order DIRK scheme with WSO 4 (middle), and DIRK-(13, 5, 5): 5th order DIRK scheme with WSO 5 (right).

equation

$$u_t + uu_x = \nu u_{xx} + f$$

for $(x, t) \in (0, 1) \times (0, 1)$, $u = g$ on $\{0, 1\} \times (0, 1)$, \hspace{1cm} (4.48)

with the true solution $u(x, t) = \cos(t)$, and the viscosity constant $\nu = 0.1$. It has been shown in a previous work by Ketcheson et al. [42] that DIRK schemes with orders up to 3 and WSO up to 3 exhibit clean and full order of convergence for this nonlinear problem. Our observation is that DIRK schemes with orders beyond 3 or WSO beyond 3 fail to recover their full convergence order for the viscous Burgers’ equation.
Figure 4.17 shows the convergence results by DIRK schemes with high WSO for viscous Burgers’ equation with a manufactured solution $u^*(x, t) = \cos(t)$. Clearly, these schemes do not remove order reduction for the nonlinear viscous order reduction.

4.8 Conclusions and Outlook

We provided a pathway to construct DIRK schemes with high weak stage order based on the WSO theory. Using the general WSO criterion, we developed high-order DIRK schemes with WSO 4 or higher, hence overcame the limitation of the eigenvector WSO criterion. A clever construction technique is employed to find feasible schemes, which are then used as starting points of an optimization routine to discover the new schemes. We provided three DIRK schemes: (i) 7-stages 4th order DIRK with WSO 4, (ii) 13-stages 5th order DIRK with WSO 4, and (iii) 13-stages 5th order DIRK with WSO 5, that are stiffly accurate and $L$-stable. We attempted to optimize a proxy of the truncation of these methods and came up with schemes with decent error constants. Hence they can not be claimed to be the optimal methods. These schemes are tested on different test problems starting from constant-coefficient differential operators of different orders to spatially varying differential operators. All the new schemes gained their respective full orders of accuracy in
the stiff region for all the test problems.

We observed cliffs systematically in the spatial error of solution when tested these schemes on linear advection problems with both constant coefficients and variable coefficients. These cliffs possibly arise from the mismatch between the error $t = 0$ and $t > 0$. It is observed that the cliffs emerge and get transported in the interior of the domain, causing the order reduction specifically in the derivative of the solution. They are the effects of boundary layers originated at $x = 0$ and $t = 0$. They get transported along the characteristic curve starting from $x = 0$ at $t = 0$. So if we run the problem for long enough final time or choose the velocity such that the cliffs move out of the domain, we get no order reduction. We do not know whether this is a feature of the advection problem or a fundamental drawback of the high WSO schemes. One way to check this would be to test the advection problem with high-stage order schemes and examine if we get those same cliffs. Then we could decide on this issue, and this is a topic of future investigation.

The new schemes are observed to exhibit order reduction when tested on problems with time-dependent differential operators or nonlinear problems. This is not a contradiction because the general WSO criterion was derived to remedying order reduction based on constant coefficients or spatially dependent differential operators. A natural question for future research is: Can a criterion be derived for DIRK schemes that can mitigate the boundary order reduction for problems with time-dependent differential operators or nonlinear problems?
4.9 Butcher Tableau For DIRK Schemes

Table 4.4: Butcher’s tableau for DIRK-(7, 4, 4).

Table 4.5: Butcher’s tableau for DIRK-(13, 5, 4).

Table 4.6: Butcher’s tableau for DIRK-(13, 5, 5).
CHAPTER 5

CONCLUSION AND OUTLOOK

Many numerical methods for differential equations possess structural shortcomings. A lot of research needs to be done that is specifically focused on remediying these fundamental challenges. This thesis has served as an example of such an investigation. The special methods presented herein successfully remedy some of these shortcomings on a fundamental level, at least for the situations and test problems presented in this work.

For instance, we studied the long-time accurate solution of numerical methods for advection problems. It is often important to accurately capture the structure of solutions for a large final time, specifically for advection-dominated problems. So it is of particular interest to devise numerical methods that can produce accurate solutions for a large final time or potentially in an infinite time horizon because many traditional numerical methods are not specifically designed to yield accurate solutions for a long time. We presented a class of numerical methods based on jet scheme methodology with a particular nonlinear interpolant that possesses long-time accurate solutions for advection problems. To study the numerical solutions’ long-time accuracy, we introduced a new notion of convergence in an infinite time horizon. We established new numerical analysis concepts, first to show that an important class of linear methods on a
fixed-grid or widely used nonlinear methods fail to produce accurate solutions for a long time. We then rigorously proved that the new nonlinear jet schemes have convergent infinite time limit solutions. Although this research focused on transport on a conceptual (1D constant-coefficient) level, the principles and methods presented here may be used to build new accurate schemes for more general advection-dominated problems in the future.

The jet scheme methodology established above does not carry over directly to variable coefficient problems, or the nonlinear jet scheme interpolant does not have an immediate generalization into 2D and 3D. Therefore, an open research question is how to develop numerical methods to produce long-time accurate solutions for variable-coefficient advection problems in 1D or even constant-coefficient advection problems in higher dimensions.

Determining the optimal finite volume limiter functions is worth investigating in this context because FV methods with limiter functions are applied to accurately capture the essential features of solutions for the advection type problems or hyperbolic conservation laws. We mainly focused on a class of third-order WENO type limiter functions, which have parameters. The values of these parameters were chosen based on some mathematical principles in the literature. However, many fixed suggested choices tend to not produce a desirable balance between the accuracy and the structural properties, such as overshoots or undershoots. So an important question is how we can determine optimal limiter functions to capture the essential structures of the solution that are of particular interest to the users’ applications. We solved the problem by reversing the standard paradigm: we learned the optimal limiters from a class of limiters based on how well they performed on a collection of well-represented test problems. The degree to which certain aspects of the solution should be avoided depends on the application. As a consequence, there is no one-size-fits-all parameter collection, and the inverse problem structure studied here is more suitable. As a first step, we focused on a very special class of third-order WENO type finite volume limiter functions. We presented two new classes of third-order WENO type limiters based on the third-order cen-
tral WENO and the WENO-Z methods with two artificial parameters. After studying the limiter function’s mechanism depending on the pair of parameters, we learned the optimal parameters by optimizing an objective function over a two-parameter family of limiters. On various test problems, the performance of the optimal limiters obtained from this study was found to be the best among the third-order WENO type limiters.

The new paradigm considered in this thesis opens up many possibilities along with many research questions. This approach is not limited to WENO type limiter functions. It can be applied to the more generic problem to determine the optimal limiter function where the limiter function may involve more parameters. Overfitting and designing a set of well-represented test cases in a similar spirit of machine learning are the most critical issues in this approach. So one important question for future research is how one can design a well-represented portfolio of test cases and a generic cost functional to learn the best limiter function without falling into the overfitting trap.

Another problem that we studied is the accuracy and performance of simple time stepping methods that are crucial components in any time-dependent multi-physics or multi-rate problems. We concentrated on the diagonally implicit Runge-Kutta (DIRK) family of methods because of their ease of implementation and cost effectiveness. These methods often do not preserve their formal order of convergence when applied to solve some multi-rate ODE systems, i.e., the exhibited order drops significantly in the stiff regime. So new higher order time integrators need to be constructed to avoid these phenomena.

We studied DIRK schemes with high weak stage order (WSO) that overcome the order reduction phenomenon for linear problems. The WSO eigenvector criterion, a special case of WSO, allowed the construction of DIRK schemes up to order four with WSO up to three in prior work. The WSO eigenvector criterion had been found to be limited to WSO three. An important key question that has been answered in this thesis was whether this limitation is fundamental, or whether the WSO concept can actually be pushed beyond that boundary. Specifically, we showed that DIRK schemes up to order five
with WSO up to five exist. In this work, we employed the generalized WSO criterion and presented a systematic methodology based on a rigorous WSO theory that enabled to find and construct DIRK schemes up to order five with WSO up to five. Thus we overcame the limitation WSO eigenvector criterion. The WSO theory and the current methods outlined in this study have also opened up plenty of new opportunities for developing even higher weak stage order schemes. For a number of linear ODE and PDE IBVP test problems, we demonstrated that the new time stepping methods prevent order reduction. So these schemes can be used as order-preserving time integrators for linear multi-rate systems.

In the optimization routine, we attempted to reduce the norm of the local truncation error. Because of the complexity of the algorithm, we were able to find schemes with acceptable truncation errors. But they can not be claimed to optimal schemes. So an important question is how one can design the construction procedure of these schemes so that they are optimal in terms of truncation errors or in terms of other objective functions. While these new schemes are devoid of order reduction for problems where the differential operators involve only space variables, they exhibit order reduction for problems with time-dependent differential operators and/or nonlinear problems. So another future research problem is whether there is a weaker condition for nonlinear problems that prevents order reduction and is also consistent with DIRK schemes.
REFERENCES


