Uncommon (in)stability properties in the numerical method of characteristic applied to hyperbolic equations

Zihao Deng

University of Vermont

Follow this and additional works at: http://scholarworks.uvm.edu/graddis

Part of the Mathematics Commons

Recommended Citation

UNCOMMON NUMERICAL INSTABILITY IN THE METHOD OF CHARACTERISTIC APPLIED TO HYPERBOLIC EQUATIONS

A Thesis Presented

by

Zihao Deng

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements for the Degree of Master of Science Specializing in Mathematics

October, 2016

Defense Date: August 9, 2016
Dissertation Examination Committee:

Taras Lakoba, Ph.D., Advisor
Jun Yu, Ph.D.
Jeffrey Marshall, Ph.D., Chairperson
Cynthia J. Forehand, Ph.D., Dean of Graduate College
Abstract

This thesis presents the stability analysis of the numerical method of characteristic (MoC) that is applied to hyperbolic partial differential equations (PDEs). Simple Euler method, Modified Euler method and Leapfrog method are used for numerical integration along the characteristics. The corresponding MoC schemes are referred to as MoC-SE, MoC-ME, and MoC-LF respectively.

We discovered and explained two unusual phenomena. First, certain non-periodic boundary condition (b.c.) could eliminate the numerical instability for some schemes such as the MoC-ME, where the instability exists for periodic b.c. However, it is commonly believed that if von Neumann analysis, i.e., assuming periodic b.c., predicts numerical instability, then there must be numerical instability if one uses non-periodic b.c..

Second, a symplectic method (LF), which is known to work well for energy-preserving ordinary differential equations, introduces a strong numerical instability when integrating energy-preserving PDEs by the MoC.

In this thesis, we worked out a new method of analyzing the numerical scheme with non-periodic boundary conditions and explained in details why the non-periodic boundary conditions could eliminate the numerical instability. We also illustrated why our result contradicts the common knowledge that von Neumann stability is necessary for numerical stability of the scheme.
3.3.2 Numerical calculation of \(|\lambda|\) ................................. 74
3.3.3 Conclusion for the von Neumann instability of LF ........................... 76

4 Stability analysis for Simple Euler scheme with non-reflecting boundary condition 78
  4.1 The impact of non-reflecting b.c. on the numerical stability ............ 78
  4.2 Set up of the method .................................................. 79
  4.3 Agreement of new method for periodic b.c. with von Neumann analysis 84
  4.4 Applying the equation to non-reflecting b.c. .................................. 89
  4.5 Computing the eigenvalue of SE scheme with non-reflecting b.c. ... 92
    4.5.1 Finding the approximate solution of (4.39) ....................... 93
    4.5.2 Finding the approximate solution of (4.42) .......................... 101
    4.5.3 Finding the approximate solution of (4.47) ....................... 108
    4.5.4 Calculations of \(\lambda\) and its numerical verification ............. 116
  4.6 Preview for the instability suppressing mechanism of MoC-ME with
      non-reflecting b.c. .................................................. 127

5 Stability analysis ME with non-reflecting b.c. 131
  5.1 Set up of the method .................................................. 131
  5.2 Finding the approximate solution \(\rho(\lambda)\) .............................. 135
  5.3 Finding the approximate solution \(\xi(\lambda)\) .............................. 138
  5.4 Finding the approximate value of \(|\lambda|\) ................................ 149
  5.5 Numerical verification of \(\lambda^{2M} = O(h^4)\) .............................. 156
  5.6 Explanation of stabilizing role of non-reflecting b.c. ..................... 158

6 Summary ........................................................................ 161

A Derivation of (4.61) .......................................................... 164

B Derivation of (4.116) .......................................................... 167

C The solutions for the hyperbolic PDE system (1.4) .......................... 171
List of Figures

1.1 Plot of $y_1$ versus $y_2$, as each numerical scheme evolves from $t = 0$ to $t = 1000$ with $h = 0.04$. According to (1.20), if the numerical scheme preserves the energy of the oscillator, the plot of $y_1$ versus $y_2$ should be a unit circle. The first panel shows that the SE scheme performs the worst for this conservative system, since the plot deviates from the unit circle very fast. Note that the scale of this panel is much larger than that of the other two. The second panel shows that the solution obtained by the ME scheme deviates from the unit circle, albeit slightly. So the ME is not a energy preserving scheme either. The last panel shows that the plot obtained by the LF barely deviates from the unit circle and, more importantly, those deviations do not accumulate with time. So the LF is the only scheme among the three that can preserve energy for a very long time.

1.2 The line $x = \xi + ct$ is the characteristic. The value of $f$ evolves as one moves from $(x_0, t_0)$ to $(x, t)$. Since $\xi = x - ct = x_0 - ct_0$, we must have $x = x_0 + c(t - t_0)$. Similarly, for each point $(x_\tau, \tau)$ on the characteristic that is between $(x_0, t_0)$ and $(x, t)$, i.e. $t_0 < \tau < t$, we have $x_\tau = x_0 + c(\tau - t_0)$. Thus, we have the following relations between the original variables ($(x, t)$, etc.) and transformed ones ($(\xi, \eta)$, etc.): $(x, t) = (x_0 + c(t - t_0), t) = (\xi, \eta)$, $(x_\tau, \tau) = (x_0 + c(\tau - t_0), \tau) = (\xi, \eta_\tau)$, $(\xi_0, t_0) = (\xi, \eta_0)$. The reason why in the pair $(\xi, t)$, $\xi$ stays constant, and only $\eta$ is changing, is that $\xi$ is a fixed number on this characteristic.
1.3 In this graph, the horizontal axis is the space grid and the vertical axis is the time grid. The slanted lines are the characteristics. The thickened vertical line denotes the boundary. The LF scheme (1.35) gives an evolution from the \((n - 1)\)th, \(n\)th time levels to the \((n + 1)\)th time level along the characteristic lines. This evolution is illustrated in the graph, where circles denote the prerequisite points and rectangles denote the desired points. When the value of \(f\) at a prerequisite point is known, we put a black circle, otherwise a white circle. Similarly, if the value of \(f\) at the square can be calculated using the information from the previous iteration, then we put a grey rectangle, otherwise a white rectangle. Since we only know the values of \((f)^n_m\), \((f)^{n-1}_m\) for \(1 \leq m \leq M\) by the previous iteration (or simply given if \(n - 1 = 0\)), the value of \((f)^{n+1}_1\) and \((f)^{n+1}_2\) can not be calculated. This is also shown in the graph: at least one of the prerequisite points for \((f)^{n+1}_1\) and \((f)^{n+1}_2\) is white. If we apply the non-reflecting b.c., the value of \((f)^{n+1}_1\) can be directly given, however, the value of \((f)^{n+1}_2\) is still unknown. We put a question mark to denote the unknown value of \((f)^{n+1}_2\).

2.1 Level curve \(F(\omega, k, b) = 0\) for values of \(b\) on opposite sides of 0. These plots are to be analysed as follows. If we draw a horizontal line \(k = const\), the intercept points between the level curve and this horizontal line are the real-valued roots \(\omega_j(k)\), where \(1 \leq j \leq 4\). If \(j < 4\), then there are complex values of \(\omega\) with nonzero imaginary part, which indicates physical instability. From the left panel we can see that when \(b > 0\), there exists “gaps” near the line \(k = \pm 1/2\), one of which is marked out. Any horizontal line \(k = const\) in such gap has less than four intercept points with the level curve, indicating that roots are not all real numbers; hence there exists physical instability. This is due to the graphical features near \((\omega, k) = (1/2, 1/2)\) and \((-1/2, -1/2)\). In the right panel where \(b < 0\), there are four intercept points for any \(k \in \mathbb{R}\), indicating that that there is no physical instability. Therefore \(b = 0\) appears to be a bifurcation point for physical instability. That is, for \(b\) that is close to 0, there is no physical instability for \(b < 0\), while there exists physical instability for \(b > 0\).
2.2 The thick line segments mark the intervals of $b$ where the physical instability definitely exists according to the results of Sec. 2.4 and Sec. 2.5. The interval with a question mark is where the physical instability cannot be explicitly determined by the analysis of Sec. 2.4 and Sec. 2.5. Here, the number $\delta$ is some small positive number such that $|G_1(\delta) - G_1(0)| \ll 1$, square bracket means the closed end of an interval, parenthesis means the open end of an interval. Note that $2 - \sqrt{3} \approx 0.2679$, $2 + \sqrt{3} \approx 3.7321$, and $1/(2 - \sqrt{3}) = 2 + \sqrt{3}$.

2.3 Level curve $F(\omega, k, b) = 0$, where $k = \frac{1}{2}$. For each $b \in (0, 1]$, there are only two real roots for $\omega$, which implies that there is physical instability. However, when $b < 0$, there will be four real roots for $\omega$, and hence no physical instability with $k \approx 1/2$. The bifurcation point is $b = 0$.

2.4 Level curve $F(\omega, k, b) = 0$ for values of $b$ changing from 0.01 to 1. We can see that the level curve $F(\omega, k, b) = 0$ changes gradually as $b$ changes from 0.01 to 1. Most importantly, it maintains one graphical features: There exist two gaps, one of which is marked in the first panel, such that any horizontal line $k = \text{const}$ in a gap has fewer than four intercepts with the level curve. This feature implies that for any $0 < b \leq 1$, there will always be values of $k$ such that there are only two real roots $\omega(k)$. Hence there exits instability for perturbations with these $k$.

2.5 Level curve $F(\omega, k, b) = 0$. The left panels are for positive values of $b$ and right panels are for their negative counter parts. In each left panel, there exists “gap” such that any horizontal line $k = \text{const}$ in between such “gap” has less than four intercept points with the level curve, while in the right panels, due to the graphical feature of symmetry described above, such “gaps” do not exist, and each horizontal line $k = \text{const}$ has four intercept points with the level curve.

3.1 The graph of $|\lambda(z)| - 1$, where $\lambda(z)$ is the “largest eigenvalue” of $N(z)$. We can see that $|\lambda|$ is always larger than 1, and $|\lambda| - 1$ is largest when $z = kh = 0$ or $\pi$. Therefore the numerically unstable modes that are observed in direct numerical simulations correspond to these values of $z$. Notice that the scale of each vertical axis is small, i.e., less than $10^{-2}$. 

vi
3.2 Fourier spectrum of the perturbation $\tilde{S}_{1,3}^{\pm}$ obtained by the SE scheme with periodic b.c. of at $t = 200$. The initial perturbations $\tilde{S}_2^{\pm}$ stay almost unchanged in time and hence are not shown. In the legend of the graph, $(+)1$, $(-)1$, $(+)3$, $(-)3$ means $\tilde{S}_1^{+}, \tilde{S}_1^{-}, \tilde{S}_3^{+}, \tilde{S}_3^{-}$ respectively. The horizontal axis is the wave number $k$, which is defined in (3.3). The relation between $k$ and $z$ is $k = z/h$, which is shown in the statements near (3.10). Here, $z$ ranges from $-\pi$ to $\pi$, so $k$ ranges from $-\pi/h = -78.5$ to $\pi/h = 78.5$. This picture shows that the sizes of $\tilde{S}_{1,3}^{\pm}$ increase most significantly when the wavenumber is $0$ or $\pm \pi/h$.

3.3 Trend of $\log(\epsilon)$ vs $t$. The error size growth from $10^{-15}$ to $10^{-3}$ for $b = -1$, $h = 0.04$. The left panel is the complete error growth. Here, the trend seems somewhat concave up, while in later iterations, the trend is very close to a straight line. The right panel is the error growth when we only take the error data that are larger than $e^{-14} \approx 10^{-6}$, i.e., $\log(\epsilon) > -14$. Here, there is very significant linear trend, and we can use this trend to calculate the numerical error growth rate.

3.4 The graph of $|\lambda(z)| - 1$, where $\lambda(z)$ is the “largest eigenvalue” of $N(z)$. We can see that $|\lambda|$ is always larger than 1, and $|\lambda| - 1$ is largest when $z = kh = \pm \pi/2$ since $\lambda(z)$ is periodic. Therefore the numerical instability that is observed in direct numerical simulation occurs for $z$ taking such values.

3.5 Fourier Spectrum for the perturbation $\tilde{S}_{1,3}^{\pm}$ obtained by ME scheme with periodic b.c. at $t = 610$. In this picture, the legend, notation, etc. are the same as those in Fig. 3.2, and $\tilde{S}_2^{\pm}$ are not shown for the same reason as in Fig. 3.2. This picture shows that the sizes of $\tilde{S}_{1,3}^{\pm}$ increase most significantly when the wavenumber is $\pm \pi/(2h)$. Notice that in the beginning of the ME scheme calculation, there is some error growth at $k = 0, \pm \pi/h$, however, this increase is eventually overpowered by the error increase at $k = \pm \pi/(2h)$.

3.6 Eigenvalue analysis for MoC-LF with periodic b.c. Here, we take $b = -1, h = 0.01$.

3.7 Fourier spectrum of the numerical error for the MoC-LF with periodic b.c. at $t = 10$. Here, we take $b = -1, h = 0.01$. The legend is the same as that of Fig. 3.5. The instability is the strongest at wave number $kh = \pi/2$.

4.1 $\lambda^2$ shifted to the left by one unit
4.2 Theoretical $|\lambda|$ for MoC-SE with non-reflecting b.c. when $L = 50$, $h = 0.01$, $b = -1$. Notice that since in the calculations, we have two values of $z$: $z_1$, $z_2$, there should be two sets of curves (see (4.123)). However, since $|z_1| = |z_2|$ for the parameters we choose, we only need to show one set of curves. 119

4.3 Super-Gaussian function $f(x) = e^{-\left(\frac{25}{L/2}\right)^8}$ for $x \in [-L/2, L/2]$ with $L = 50$. 121

4.4 Evolution of $\log_{10}(|\text{FFT}_{\text{error}}(t)|)$ vs $t$. The corresponding values of $L$ and $h$ are shown in the legend; $b = -1$ in all cases. 122

4.5 The solid curve shows the evolution of $\log_{10}(|\text{FFT}_{\text{error}}(t)|)$ vs $t$. Here, we use the parameters $b = -1, h = 0.00625, L = 25$ as an example. The data points marked with asterisks are such that every two consecutive points are separated by the same time interval (see the main text). Connecting these points gives the dashed line. 124

4.6 The values $\ln(|\lambda|^{2M})$ found from numerical simulations vs $\ln(h)$ for MoC-SE with non-reflecting b.c. The parameters are $b = -1$, $L = 25$ and 50, and $h = 0.01, 0.00625, 0.005, 0.004, 0.002, 0.001, 0.0005$. 125

4.7 Factor $\frac{|\lambda|^{2M}}{|\lambda|^{2M}}$ vs $Lh$. The parameters used are $b = -1$, $L = 25$, and 50 and $h = 0.01, 0.00625, 0.005, 0.004, 0.002, 0.001, 0.0005$. 127

4.8 See text in Sec. 4.6. 129

4.9 See text in Sec. 4.6. 130

5.1 Fourier spectra for $L = 25, h = 0.05$. Panel (a) is the spectrum when $t = 25$; panel (b) is the spectrum when $t = 50$; panel (c) is when $t = 75$. The box with solid line illustrates the half-width of 2.5%; the box with dashed line illustrates the half-width of 10%. The horizontal dash-dotted lines indicate the magnitude of the spectrum at grid point $m = M/4$ and at grid points $m = M/4 \pm M/10$. 157

5.2 Dependence of $\ln(|\lambda_{\text{num}}|^{2M})$ on $\ln(h)$, as described in the text. The slopes found by linear regression are 2.1, 2.1, 4.3, 4.2 for half-widths of 15%, 10%, 5%, 2.5% respectively (see the legend). The dashed lines with their slopes indicated are guides for the eye. 158

5.3 Schematically illustrating the spatial dependence, $(\rho, \rho^2, \ldots, \rho^M)$ (see (4.36)), of the eigenvectors of the problem (4.17) and (5.8) vs $x$. The left (right) panel is the case described in textbooks (this chapter). 160
List of Tables

3.1 Growth rates with different increment $h$. ................................. 67
3.2 Growth rates with different increment $h$. ................................. 72
Chapter 1

Introduction

1.1 Scope of the problem

In this thesis, we consider stability of the numerical method of characteristics (MoC) applied to a class of energy-preserving hyperbolic partial differential equations (PDEs). Characteristics are lines (curved in general) in the space-time domain, along which the information propagates. To accurately compute the solution, it is important to accurately describe both the characteristics and how information propagates. The main strength of the MoC is that it reduces hyperbolic a PDE to a small set of ordinary differential equations (ODEs) based on the propagation of information along the characteristics. This simplification reduces the time of computations. The MoC is, therefore, a widely used finite-difference method for computing the numerical solution of hyperbolic PDE [1]. It is often used in various fields, such as gas dynamics [2], biomechanics [3], and geomechanics [4].

While stability of many other numerical methods for hyperbolic PDEs is well studied, stability of the numerical MoC has only been studied by few researchers.
Moreover, in studies that are especially focused on stability of the numerical MoC, one commonly considers hyperbolic PDEs of the form:

\[ f_t + A f_x = b, \tag{1.1} \]

where vector \( f \) is the sought solution, \( A \) is a constant matrix, and \( b \) is a constant vector, independent of \( f \). On the other hand, a PDE of a more general form:

\[ f_t + A f_x = b(f), \tag{1.2} \]

where vector \( b(f) \) depends on solution \( f \), has been less studied. Here, \( b(f) \) can be a linear or nonlinear function of (the components of) \( f \). Furthermore, the most common method for analyzing stability is the von Neumann analysis, which assumes periodic boundary conditions (b.c.). Generally, the results given by the von Neumann analysis are perceived as sound to explain numerical instability, because the effect of b.c. is usually negligible. Methods for analyzing stability that go beyond the von Neumann analysis and consider non-periodic b.c. are much more difficult and hence less studied (\cite{5}, \cite{6}, \cite{7}). As a matter of fact, we could only find one paper, \cite{1}, that studies stability of the numerical MoC for hyperbolic equations using a method different from von Neumann analysis. Namely, \cite{1} studied stability of the numerical MoC for (1.2) with \( b(f) = B f \) and non-periodic b.c. using a set of novel matrix representations. Here, \( B \) is a constant matrix. We will use a method based on the ideas presented in Ref. \cite{1} in Chapter 4.

In this thesis, we apply the numerical MoC to a nonlinear system of hyperbolic equations that preserves energy. This system assumes the form of (1.2). In order
to study the numerical instability (NI) of the numerical MoC, we first use the von Neumann analysis. When we use periodic b.c., the observed NI agrees with the prediction of the von Neumann analysis. However, in one of the cases, the NI predicted by the von Neumann analysis disappears when we apply another type of b.c., called non-reflecting b.c.. Thus, in this case, b.c. have significant effect on the NI. Therefore, we have to use techniques that go beyond the von Neumann analysis, which can account for non-periodic b.c., to explain this phenomenon.

In the numerical MoC, we use three different numerical schemes to integrate along the characteristics: Simple Euler scheme (SE), Modified Euler scheme (ME), and Leapfrog scheme (LF). Among these three schemes, the LF is the energy conserving one. More specifically, in Sec. 12.7 of [5], it is shown that the LF has no numerical diffusion for the hyperbolic PDE $f_t + cf_x = 0$, where $c$ is constant. Therefore, we expect the LF to be the best choice for our problem. However, in our numerical experiments that used periodic b.c., we found that the LF is numerically unstable, and this NI was confirmed by the von Neumann analysis. So we changed the periodic b.c. to the non-reflecting b.c. mentioned above, but the NI still existed. Therefore, we decided to use other numerical schemes. When we applied the SE, we saw that NI appeared no matter whether we applied the periodic or the non-reflecting b.c.. Thus, the b.c. were found not to affect the existence of NI for the SE. When we applied the ME with periodic b.c., we again found NI, which agreed with the von Neumann analysis. However, when we applied the ME with the non-reflecting b.c., the strong NI was suppressed. This observation shows that for this particular method, the ME, the type of b.c. has a strong effect on NI. This phenomenon is unexpected, and thus we will focus on studying it in this thesis.
In Section 1.2 of this chapter we will introduce a nonlinear system of hyperbolic
PDEs which we will consider in this thesis. We will also demonstrate its energy-
preserving nature. Since the MoC reduces a system of PDEs to that of ODEs, then
in Section 1.4, we will review several schemes that are commonly used to numerically
solve ODEs, including those which preserve the energy (Hamiltonian) of the system.
Then, in Section 1.5, we will introduce the MoC in detail.

1.2 THE PHYSICAL PROBLEM

The model that we consider in this thesis is:

\[ S_t^+ + S_x^+ = S^+ \times JS^- \]
\[ S_t^- - S_x^- = S^- \times JS^+, \]  (1.3)

where \( t \) is the time variable, \( x \) is the space variable, vector \( S^\pm = (S^\pm_1, S^\pm_2, S^\pm_3)^T \)
with \( T \) denoting the transpose, matrix \( J = \text{diag}(1, -1, -2) \), and \( \times \) denotes the cross
product. Physically, this set of equations describes the situation where two counter-
propagating light beams nonlinearly interact inside a birefringent optical fiber via the
Kerr effect ([8], [9]). In (1.3), \( S^\pm \) are the so-called Stokes vectors that describe the
polarization state of the forward- and backward- propagating beams. We will not
further discuss the physical meaning of the variables in (1.3) because it is not related
to the subject of this thesis, which is the study of the instability of the numerical
MoC applied to (1.3).
The component form of system (1.3) is:

\[
\begin{align*}
(\partial_t + \partial_x) S_1^+ &= S_3^+ S_2^- - 2S_2^+ S_3^-,
(\partial_t + \partial_x) S_2^+ &= 2S_1^+ S_3^- + 2S_3^+ S_1^-,
(\partial_t + \partial_x) S_3^+ &= - (S_1^+ S_2^- + S_2^+ S_1^-),
(\partial_t - \partial_x) S_1^- &= S_3^- S_2^+ - 2S_2^- S_3^+,
(\partial_t - \partial_x) S_2^- &= 2S_1^- S_3^+ + S_3^- S_1^+,
(\partial_t - \partial_x) S_3^- &= (\partial_t + \partial_x) S_3^+.
\end{align*}
\]

(1.4)

For the convenience of future discussions, we define

\[
\begin{align*}
f_1^+ (S^\pm) &= S_3^+ S_2^- - 2S_2^+ S_3^-,
f_2^+ (S^\pm) &= 2S_1^+ S_3^- + 2S_3^+ S_1^-,
f_3^+ (S^\pm) &= - (S_1^+ S_2^- + S_2^+ S_1^-),
f_1^- (S^\pm) &= S_3^- S_2^+ - 2S_2^- S_3^+,
f_2^- (S^\pm) &= \tilde{S}_3^- - 2b \tilde{S}_3^+,
f_3^- (S^\pm) &= - (S_1^+ S_2^- + S_2^+ S_1^-),
\end{align*}
\]

(1.5)

which corresponds to the equations in the right hand side (r.h.s.) of (1.4).

Equations (1.3) have a so-called soliton solution, first found in [8]:

\[
\begin{align*}
S_1^\pm &= \pm \frac{P^\pm}{\sqrt{3}} \text{sech} \left( \sqrt{2} P_{ave} \theta \right) \\
S_2^\pm &= \mp P^\pm \tanh \left( \sqrt{2} P_{ave} \theta \right) \\
S_3^\pm &= \sqrt{2} S_1^\pm,
\end{align*}
\]

(1.6)
where

\[ P_{ave} = \frac{P^+ + P^-}{2}, \quad \theta = z - z_0 - Vt, \quad V = \frac{P^+ - P^-}{P^+ + P^-}. \]  

(1.7)

It describes a localized pulse (sech) in components \( S_{1,3}^\pm \) and a “step” (tanh) in components \( S_2^\pm \), all moving together with speed \( V \). However, in this thesis we will not consider such solutions, varying in space. Instead, we will focus on the evolution of small perturbations over a simpler — constant — solution. A number of such constant solutions exist; they are presented in Appendix C. We will specifically consider the solution

\[ S_{1,3}^\pm = 0, \quad S_2^+ = a, \quad S_2^- = b, \]  

(1.8a)

where \( a, b \) are constant real numbers. Here, one can rescale \( a = 1 \), but keep \( b \) as a free parameter by using the following transformation:

\[ t_{old} = t_{new}/a, \quad x_{old} = x_{new}/a. \]

Thus, the solution whose small perturbation we will consider is

\[ S_{1,3}^\pm = 0, \quad S_2^+ = 1, \quad S_2^- = b. \]  

(1.8b)

Small perturbation of this solution will be described in Section 2.2.

Equations (1.3) are hyperbolic PDEs, and there exist conservation laws for this
system. The proof is as follows: take the dot product of each equation in (1.3) with $S^\pm$ and get

\[
S^+ \cdot (S^+_t + S^+_x) = S^+ \cdot (S^+ \times JS^-)
\]
\[
S^- \cdot (S^-_t - S^-_x) = S^- \cdot (S^- \times JS^+).
\] (1.9)

Since the dot product of two perpendicular vectors is 0, one can see that the r.h.s. of each equation in (1.9) is 0. Thus,

\[
(\partial_t \pm \partial_x) |S^\pm|^2 = 0.
\] (1.10)

This means that quantities $|S^+|^2$ and $|S^-|^2$ are conserved along the characteristics $x - t = \text{const}$ and $x + t = \text{const}$, respectively. It is outside the scope of this thesis to discuss the relation between $|S^\pm|^2$ and the energy of the system. Yet, we will still say that system (1.3) has conserved energy. Therefore, numerical methods for it should, ideally, also preserve or nearly preserve energy.

In the Sec. 1.4 we describe several well-known schemes for solving ODEs, one of which has such an energy preservation property.

1.3 Periodic b.c. and Non-reflecting b.c.

Since in practice, we need to solve (1.4) on a finite domain $[0, L]$, we have to consider conditions at the boundaries $x = 0$ and $x = L$. In this section, we introduce the analytical periodic b.c. and non-reflecting b.c. for the system of hyperbolic PDEs (1.4). In Sec. 1.5 and later in Chap. 3-5, we will apply these two kinds of b.c. to the
numerical calculations.

If we are using periodic b.c., then not only the values of $S^\pm$ at the boundaries must be equal to each other, but also the values of $S_x^\pm$ at the boundaries must be equal to each other. This ensures that the function $S^\pm$ is smooth across the boundaries. Thus, the analytical periodic b.c. for (1.4) is

$$S^\pm(0,t) = S^\pm(L,t), \quad S_x^\pm(0,t) = S_x^\pm(L,t), \quad t \geq 0.$$  \hspace{1cm} (1.11)

For non-reflecting b.c., the values of $S^+$ at $x = 0$ and $S^-$ at $x = L$ must be given:

$$S^+(0,t) = G_1(t), \quad S^-(L,t) = G_2(t), \quad t \geq 0,$$  \hspace{1cm} (1.12)

where $G_{1,2}$ are given vector functions of dimension $3 \times 1$. Notice that they could either be constant function or changing over time.

1.4 THREE NUMERICAL SCHEMES FOR ODEs

The main strength of the MoC is that it reduces a set of hyperbolic PDEs to a small set of ODEs based on the propagation of information along the characteristics. Since we are applying the numerical MoC to hyperbolic PDEs, we need only to consider numerical schemes for ODEs. Note that the system of PDEs of our interest, (1.3), is energy-preserving, which is proven by (1.10). Therefore, below we consider a simple ODE that also preserves energy. Such a system is the harmonic oscillator.
The simple harmonic oscillator model is:

\[ \ddot{y} = -y, \quad y(0) = 0, \quad y'(0) = 1. \] (1.13)

Denote \( y_1 = y(t), \ y_2 = y'(t), \) and \( \mathbf{y} = (y_1, y_2)^T; \) then we rewrite (1.13) as:

\[ \begin{align*}
    y'_1 &= y_2, \\
    y'_2 &= f(y_1), \quad \text{where} \quad f(y_1) = -y_1. 
\end{align*} \] (1.14)

We can write the equations above in matrix form:

\[
\begin{pmatrix}
    y_1 \\
    y_2
\end{pmatrix}' = \begin{pmatrix}
    0 & 1 \\
    -1 & 0
\end{pmatrix} \begin{pmatrix}
    y_1 \\
    y_2
\end{pmatrix}. \] (1.15)

Since the eigenvalues of the coefficient matrix of (1.15) are \( \lambda = \pm i, \) the exact solution is

\[ y(t) = C_1 e^{it} + C_2 e^{-it}. \] (1.16)

Considering the initial condition, we have:

\[ y(t) = \frac{i}{2} \left( e^{it} - e^{-it} \right). \] (1.17)

Therefore, according to the definition of \( y_1 \) and \( y_2 \) given above, we have

\[
\begin{align*}
    y_1(t) &= y(t) = \frac{i}{2} \left( e^{it} - e^{-it} \right) = -\sin t, \\
    y_2(t) &= y'(t) = -\frac{1}{2} \left( e^{it} + e^{-it} \right) = -\cos t.
\end{align*} \] (1.18) (1.19)
Thus, we have the relation

\[ y_1^2 + y_2^2 = 1, \quad (1.20) \]

whose r.h.s. is the energy (Hamiltonian) of this system. Thus, (1.20) shows that the energy of the harmonic oscillator is preserved.

Below we present three schemes: the SE, the ME, and the LF, applied to (1.14), with general function \( f(y) \). In these schemes, we denote \( \Delta t = h \) as time step, with \( t_n = nh \) and \( (y_{1,2})^n \equiv y_{1,2}(t_n) \) etc.

The SE is:

\[
\begin{align*}
(y_1)^{n+1} &= (y_1)^n + h (y_2)^n \\
(y_2)^{n+1} &= (y_2)^n + hf((y_1)^n).
\end{align*}
\quad (1.21)
\]

The ME is:

\[
\begin{align*}
\bar{y}_1 &= (y_1)^n + h (y_2)^n \\
\bar{y}_2 &= (y_2)^n + hf((y_1)^n) \\
(y_1)^{n+1} &= (y_1)^n + \frac{h}{2} ((y_2)^n + \bar{y}_2) \\
(y_2)^{n+1} &= (y_2)^n + \frac{h}{2} (f((y_1)^n) + f(\bar{y}_1)).
\end{align*}
\quad (1.22)
\]

The LF is:

\[
\begin{align*}
(y_1)^{n+1} &= (y_1)_{n-1} + 2h (y_2)^n \\
(y_2)^{n+1} &= (y_2)_{n-1} + 2hf((y_1)^n). \quad (1.23)
\end{align*}
\]
The results of applications of these schemes to the harmonic oscillator model (1.13), where \( f(y) = -y \), are shown in Fig. 1.1. According to (1.20), if the numerical scheme preserves the energy of this oscillator, the plot of \( y_1 \) versus \( y_2 \) should be a unit circle. Since the graph for the LF is still a unit circle after many iterations and barely deviates, the LF can (approximately) preserve the energy of the harmonic oscillator, and hence it is an appropriate numerical integration method for non-dissipative system. On the other hand, the SE and, to a lesser extent, the ME schemes do not preserve the energy of the harmonic oscillator. Hence they are not considered as appropriate numerical integration methods for non-dissipative systems. These conclusions can also be arrived at by considering the stability regions of these methods; however, this is outside the scope of this thesis.

\[ \text{Figure 1.1: Plot of } y_1 \text{ versus } y_2, \text{ as each numerical scheme evolves from } t = 0 \text{ to } t = 1000 \text{ with } h = 0.04. \text{ According to (1.20), if the numerical scheme preserves the energy of the oscillator, the plot of } y_1 \text{ versus } y_2 \text{ should be a unit circle. The first panel shows that the SE scheme performs the worst for this conservative system, since the plot deviates from the unit circle very fast. Note that the scale of this panel is much larger than that of the other two. The second panel shows that the solution obtained by the ME scheme deviates from the unit circle, albeit slightly. So the ME is not a energy preserving scheme either. The last panel shows that the plot obtained by the LF barely deviates from the unit circle and, more importantly, those deviations do not accumulate with time. So the LF is the only scheme among the three that can preserve energy for a very long time.} \]
1.5 Introduction of MoC

In hyperbolic PDEs with one spatial and one temporal dimension, characteristics are curves in the space-time domain along which information propagates. This feature allows one to reduce the hyperbolic PDEs to a set of simple ODEs along the characteristics. Thus, in order to solve hyperbolic PDEs, we only need to solve these ODEs.

1.5.1 One hyperbolic PDE

The idea of the MoC for a hyperbolic PDE with one characteristic curve is as follows (see Sec. 12.2 of [5]). Consider the inhomogeneous hyperbolic equation

\[ f_t + cf_x = F(x, t, f). \] (1.24)

Consider the following variable transformation:

\[ \xi = x - ct \]
\[ \eta = t. \] (1.25)

Using the chain rule of derivatives of multi-variable function,

\[
\frac{\partial f}{\partial t} = \frac{\partial \xi}{\partial t} \frac{\partial f}{\partial \xi} + \frac{\partial \eta}{\partial t} \frac{\partial f}{\partial \eta} = -c \frac{\partial f}{\partial \xi} + \frac{\partial f}{\partial \eta},
\]

\[
\frac{\partial f}{\partial x} = \frac{\partial \xi}{\partial x} \frac{\partial f}{\partial \xi} + \frac{\partial \eta}{\partial x} \frac{\partial f}{\partial \eta} = \frac{\partial f}{\partial \xi},
\] (1.26)
in (1.24), we obtain

\[
\left( -c \frac{\partial f}{\partial \xi} + c \frac{\partial f}{\partial \eta} \right) + \frac{\partial f}{\partial \eta} = F \quad \Rightarrow \quad \frac{\partial f}{\partial \eta} = F.
\] (1.27)

Note that here, \( f = f(\xi, \eta) \) and \( F = F(\xi, \eta, f) \), so their explicit formulae are different from those of \( f(x, t) \) and \( F(x, t, f) \). From the calculation above we can see that, if we use the set of independent variables \((\xi, \eta)\) instead of \((x, t)\), then the PDE (1.24) is equivalent to (1.27). This implies that, for a fixed \( \xi = x - ct \), we only need to solve the ODE

\[
\frac{\partial f(\xi, \eta)}{\partial \eta} = F(\xi, \eta, f(\xi, \eta)), \quad \text{where} \quad \xi \equiv x - ct = \text{const.}
\] (1.28)

For each fixed value of \( \xi \), the line \( x = ct + \xi \) is a characteristic. Thus, the process of solving the PDE (1.24) reduces to solving the ODE (1.28) along the characteristic lines. This solution is given by

\[
f(\xi, \eta) = f(\xi, \eta_0) + \int_{\eta_0}^{\eta} F(\xi, \tau, f(\xi, \tau))d\tau,
\] (1.29)

where \( f(\xi, \eta_0) \) is the initial condition of the PDE (1.24) at \( \eta_0 = t_0 \) for the fixed \( \xi \). Note that in the original variables, \((x, t)\), Eq. (1.29) becomes

\[
f(x, t) = f(x_0, t_0) + \int_{t_0}^{t} F(x_\tau, \tau, f(x_\tau, \tau))d\tau.
\] (1.30)
The meaning of terms in this formula is illustrated by Fig. 1.2, where \( x_0 \), \( x_\tau \), and \( x \) are related by the following formulas:

\[
x = x_0 + c(t - t_0), \quad x_\tau = x_0 + c(\tau - t_0).
\]  (1.31)

Based on the illustration, we can see that in Eq. (1.29), \( f(\xi, \eta) = f(x, t) \), \( f(\xi, \eta_0) = f(x_0, t_0) \), and \( f(\xi, \tau) = f(x_\tau, \tau) \). Moreover, each point \((\xi, \tau)\), where \( t_0 < \tau < t \), becomes \((x_\tau, \tau)\) if we change the variables from \((\xi, \eta)\) to \((x, t)\); so \( F(\xi, \tau, f(\xi, \tau)) = F(x_\tau, \tau, f(x_\tau, \tau)) \). Thus, Eq. (1.29) gives Eq. (1.30) after we consider these relations.

Note that it may not be possible to compute \( f \) analytically using either (1.29) or (1.30) since \( f \) is unknown inside \( \int_{t_0}^t F \, d\tau \). Therefore, one needs to use numerical schemes to compute \( f \) based on (1.30). We describe this approach below.

**Figure 1.2:** The line \( x = \xi + ct \) is the characteristic. The value of \( f \) evolves as one moves from \((x_0, t_0)\) to \((x, t)\). Since \( \xi = x - ct = x_0 - c t_0 \), we must have \( x = x_0 + c(t - t_0) \). Similarly, for each point \((x_\tau, \tau)\) on the characteristic that is between \((x_0, t_0)\) and \((x, t)\), i.e. \( t_0 < \tau < t \), we have \( x_\tau = x_0 + c(\tau - t_0) \). Thus, we have the following relations between the original variables \((x, t)\) and transformed ones \((\xi, \eta)\): \( x = x_0 + c(t - t_0), t = (\xi, \eta) \), \( (x_\tau, \tau) = (x_0 + c(\tau - t_0), \tau) = (\xi, \eta_\tau) \), \( (x_0, t_0) = (\xi, \eta_0) \). The reason why in the pair \((\xi, t)\), \( \xi \) stays constant, and only \( \eta \) is changing, is that \( \xi \) is a fixed number on this characteristic.

Denote the time step as \( \Delta t = h \); then \( \Delta x = c\Delta t = ch \) is the space step and
\( f_m^n \) is the value of \( f \) at the grid point \((x_m = m\Delta x, t_n = n\Delta t)\). The idea of the numerical MoC is to integrate along characteristic line step by step. Note that we choose Eq. (1.30) instead of Eq. (1.29) because the numerical solution \( f(x, t) \) is more naturally computed in the original variables \((x, t)\). If we evolve the solution from \( t_n \) to \( t_{n+1} \) using (1.30), we have

\[
(f)_m^{n+1} = (f)_{m-1}^n + \int_{nh}^{(n+1)h} F(x_{m-1} + c(\tau - t_n), \tau, f(x_{m-1} + c(\tau - t_n), \tau)) d\tau. \quad (1.32)
\]

If we use the SE method (1.21) for the numerical approximation of the integral in (1.32), then the numerical scheme for (1.32) is

\[
(f)_{m}^{n+1} = (f)_{m-1}^n + hF(x_{m-1}, t_n, (f)_{m-1}^n). \quad (1.33)
\]

Now, if we use the ME method (1.22) for the numerical approximation of the integral in (1.32), then the numerical scheme for (1.32) is

\[
(\tilde{f})_m^n = (f)_{m-1}^n + hF(x_{m-1}, t_n, (f)_{m-1}^n)
\]

\[
(f)_{m}^{n+1} = \frac{1}{2} [(f)_{m-1}^n + (\tilde{f})_m^n + hF(x_{m-1}, t_n, (\tilde{f})_m^n)]. \quad (1.34)
\]

Similarly, if we use the LF method (1.23) for the numerical approximation of the integral in (1.32), then we should consider the evolution over two time levels per step and get the numerical scheme for (1.32) as follows:

\[
(f)_{m}^{n+1} = (f)_{m-2}^{n-1} + 2hF(x_{m-1}, t_n, (f)_{m-1}^n). \quad (1.35)
\]

Now consider applying the two kinds of b.c. that we discussed in Sec. 1.3. However,
unlike there, here we have only one hyperbolic PDE, and, strictly speaking, the b.c. is to be applied only at its left boundary. Let the index of the grid points on the space line be within these bounds: \(1 \leq m \leq M\).

In the SE scheme (1.33), we assume that the value of \(f\) at each discrete space point of the \(n\)th time level is known. When \(m = 1\), it requires an unknown value at the \(n\)th time level \((f)_0^n\), which is the only required unknown value at the \(n\)th time level. Thus, it is the treatment of the value of \((f)_0^n\) that is the major difference of the two b.c.’s. If we apply the periodic b.c., we identify

\[
(f)_0^n = (f)_M^n. \quad (1.36)
\]

Therefore, according to (1.33), we have

\[
(f)_{m+1}^n = (f)_{m-1}^n + hF (x_{m-1}, t_n, (f)^n_{m-1}), \quad 2 \leq m \leq M \\
(f)_1^{n+1} = (f)_M^n + hF (x_M, t_n, (f)_M^n). \quad (1.37)
\]

When we apply the non-reflecting b.c. to (1.33), we simply should not compute \((f)_1^n\) because it is prescribed by this type of b.c.:

\[
(f)_1^n = G(t_n). \quad (1.38)
\]

Thus, according to (1.33), we have

\[
(f)_{m+1}^n = (f)_{m-1}^n + hF (x_{m-1}, t_n, (f)^n_{m-1}), \quad 2 \leq m \leq M \\
(f)_1^{n+1} = G(t_{n+1}). \quad (1.39)
\]
In the ME scheme (1.34), we also assume that the value of \( f \) at each discrete space point of the \( n \)th time level is known. Therefore, similar to (1.37), if we apply the periodic b.c. to the ME scheme (1.34), then (1.36) yields

\[
\begin{align*}
\left( \bar{f} \right)_m &= (f)_m^n + hF(x_{m-1}, t_n, (f)_{m-1}^n) \\
(f)_{m-1}^{n+1} &= \frac{1}{2} \left[ (f)_{m-1}^n + (\bar{f})_m^1 + hF(x_{m-1}, t_n, (\bar{f})_m^1) \right]. \\
\left( \bar{f} \right)_1 &= (f)_M^n + hF(x_M, t_n, (f)_M^n) \\
(f)_{M+1}^{n+1} &= \frac{1}{2} \left[ (f)_M^n + (\bar{f})_1^1 + hF(x_M, t_n, (\bar{f})_1^1) \right].
\end{align*}
\]

(1.40)

Similar to (1.39), when we apply the non-reflecting b.c. to the ME scheme (1.34), Eq. (1.38) yields

\[
\begin{align*}
(f)_{1}^{n+1} &= G(t_{n+1}) \\
\left( \bar{f} \right)_m &= (f)_{m-1}^n + hF(x_{m-1}, t_n, (f)_{m-1}^n) \\
(f)_{m-1}^{n+1} &= \frac{1}{2} \left[ (f)_{m-1}^n + (\bar{f})_m^1 + hF(x_{m-1}, t_n, (\bar{f})_m^1) \right], \quad 2 \leq m \leq M.
\end{align*}
\]

(1.41)

In the LF scheme (1.35), we assume that on both the \( n \)th and \( (n-1) \)th time levels, the value of \( f \) at each discrete space point is known. When \( m = 2 \), the scheme requires a value of \( (f)_{0}^{n-1} \); when \( m = 1 \), the scheme requires values of \( (f)_{0}^{n-1} \) and \( (f)_{0}^{n-1} \). If we apply the periodic b.c., we have

\[
(f)_{0}^n = (f)_M^n, \quad (f)_{0}^{n-1} = (f)_{M-1}^{n-1}, \quad (f)_{-1}^{n-1} = (f)_{M-1}^{n-1}.
\]

(1.42)

Therefore, according to (1.35), we have
Figure 1.3: In this graph, the horizontal axis is the space grid and the vertical axis is the time grid. The slanted lines are the characteristics. The thickened vertical line denotes the boundary. The LF scheme (1.35) gives an evolution from the \((n-1)th\), \(n\)th time levels to the \((n+1)th\) time level along the characteristic lines. This evolution is illustrated in the graph, where circles denote the prerequisite points and rectangles denote the desired points. When the value of \(f\) at a prerequisite point is known, we put a black circle, otherwise a white circle. Similarly, if the value of \(f\) at the square can be calculated using the information from the previous iteration, then we put a grey rectangle, otherwise a white rectangle. Since we only know the values of \((f)_{m}^{n}\), \((f)_{m}^{n-1}\) for \(1 \leq m \leq M\) by the previous iteration (or simply given if \(n-1 = 0\)), the value of \((f)_{1}^{n+1}\) and \((f)_{2}^{n+1}\) can not be calculated. This is also shown in the graph: at least one of the prerequisite points for \((f)_{1}^{n+1}\) and \((f)_{2}^{n+1}\) is white. If we apply the non-reflecting b.c., the value of \((f)_{1}^{n+1}\) can be directly given, however, the value of \((f)_{2}^{n+1}\) is still unknown. We put a question mark to denote the unknown value of \((f)_{2}^{n+1}\).

\[
\begin{align*}
(f)_{m}^{n+1} &= (f)_{m-2}^{n-1} + 2hF(x_{m-1}, t_{n}, (f)_{m-1}^{n}), \quad 3 \leq m \leq M \\
(f)_{2}^{n+1} &= (f)_{M}^{n-1} + 2hF(x_{1}, t_{n}, (f)_{1}^{n}) \\
(f)_{1}^{n+1} &= (f)_{M-1}^{n-1} + 2hF(x_{M}, t_{n}, (f)_{M}^{n}).
\end{align*}
\]  

(1.43)

If we apply the non-reflecting b.c., we do not need to compute the value of \((f)_{1}^{n+1}\), because this is the value of \(f\) prescribed at the boundary. However, we do need to compute \((f)_{2}^{n+1}\) (see Fig. 1.3), and this cannot be done using the above scheme, since it requires \((f)_{0}^{n-1}\), which is undefined. Therefore, one has to approximate the value of \((f)_{2}^{n+1}\), and there are multiple methods to do so. We list two sensible methods below:
The first method is to use simple Euler method to calculate the value of \((f)_{2}^{n+1}\):

\[
\begin{align*}
(f)_{1}^{n+1} &= G(t_{n+1}) \\
(f)_{2}^{n+1} &= (f)_{1}^{n} + hF(x_1, t_n, (f)_1^n) \\
(f)_{m}^{n+1} &= (f)_{m-2}^{n-1} + 2hF(x_{m-1}, t_n, (f)_{m-1}^n), \quad 3 \leq m \leq M.
\end{align*}
\]  
\(1.44\)

The second method is to use the classical Runge–Kutta method to calculate the value of \((f)_{2}^{n+1}\):

\[
\begin{align*}
(f)_{1}^{n+1} &= G(t_{n+1}) \\
(fk)_1 &= hF(x_1, t_n, (f)_1^n) \\
(fk)_2 &= hF(x_1, t_n, (f)_1^n + \frac{1}{2}(fk)_1) \\
(fk)_3 &= hF(x_1, t_n, (f)_1^n + \frac{1}{2}(fk)_2) \\
(fk)_4 &= hF(x_1, t_n, (f)_1^n + (fk)_3) \\
(f)_{2}^{n+1} &= (f)_{1}^{n} + \frac{1}{6}((fk)_1 + 2(fk)_2 + 2(fk)_3 + (fk)_4) \\
(f)_{m}^{n+1} &= (f)_{m-2}^{n-1} + 2hF(x_{m-1}, t_n, (f)_{m-1}^n), \quad 3 \leq m \leq M.
\end{align*}
\]  
\(1.45\)

1.5.2 TWO HYPERBOLIC PDES WITH DIFFERENT CHARACTERISTICS

Now we explain how the idea of the MoC can be extended to a PDE (1.2) with two characteristics. For simplicity, let us set in (1.2) \(A = \text{diag}(c, -c)\); also, \(f = (f_1, f_2)^T\)
and \( \mathbf{b}(\mathbf{f}) = (b_1(f_1, f_2), b_2(f_1, f_2))^T \). Then the PDE (1.2) takes on the form:

\[
\begin{align*}
\frac{\partial f_1(x, t)}{\partial t} + c \frac{\partial f_1(x, t)}{\partial x} &= b_1(f_1(x, t), f_2(x, t)) \quad (1.46a) \\
\frac{\partial f_2(x, t)}{\partial t} - c \frac{\partial f_2(x, t)}{\partial x} &= b_2(f_1(x, t), f_2(x, t)). \quad (1.46b)
\end{align*}
\]

Let

\[
\begin{align*}
\xi_1 &= x - ct, \quad \eta_1 = t \\
\xi_2 &= x + ct, \quad \eta_2 = t. \quad (1.47)
\end{align*}
\]

Similar to the derivation of (1.26) and (1.27), using the chain rule for derivatives of a multi-variable function,

\[
\begin{align*}
\frac{\partial f_1}{\partial t} &= \frac{\partial \xi_1}{\partial t} \frac{\partial f_1}{\partial \xi_1} + \frac{\partial \eta_1}{\partial t} \frac{\partial f_1}{\partial \eta_1} = -c \frac{\partial f_1}{\partial \xi_1} + \frac{\partial f_1}{\partial \eta_1} \\
\frac{\partial f_1}{\partial x} &= \frac{\partial \xi_1}{\partial x} \frac{\partial f_1}{\partial \xi_1} + \frac{\partial \eta_1}{\partial x} \frac{\partial f_1}{\partial \eta_1} = \frac{\partial f_1}{\partial \xi_1} \quad (1.48)
\end{align*}
\]

and then from (1.46a) we obtain

\[
\left( -c \frac{\partial f_1}{\partial \xi_1} + \frac{\partial f_1}{\partial \eta_1} \right) + c \frac{\partial f_1}{\partial \xi_1} = b_1 \quad \Rightarrow \quad \frac{\partial f_1}{\partial \eta_1} = b_1. \quad (1.49)
\]

Similarly,

\[
\begin{align*}
\frac{\partial f_2}{\partial t} &= \frac{\partial \xi_2}{\partial t} \frac{\partial f_2}{\partial \xi_2} + \frac{\partial \eta_2}{\partial t} \frac{\partial f_2}{\partial \eta_2} = c \frac{\partial f_2}{\partial \xi_2} + \frac{\partial f_2}{\partial \eta_2} \\
\frac{\partial f_2}{\partial x} &= \frac{\partial \xi_2}{\partial x} \frac{\partial f_2}{\partial \xi_2} + \frac{\partial \eta_2}{\partial x} \frac{\partial f_2}{\partial \eta_2} = \frac{\partial f_2}{\partial \xi_2} \quad (1.50)
\end{align*}
\]
and from (1.46b) we obtain

\[ \left( c \frac{\partial f_2}{\partial \xi_2} + \frac{\partial f_2}{\partial \eta_2} \right) - c \frac{\partial f_2}{\partial \xi_2} = b_2 \quad \Rightarrow \quad \frac{\partial f_2}{\partial \eta_2} = b_2. \] (1.51)

Thus, similar to (1.28), we get

\[ \frac{\partial f_1(\xi_1, \eta_1)}{\partial t} = b_1(f_1(\xi_1, \eta_1), f_2(\xi_1, \eta_1)), \quad \xi_1 = \text{const} \] (1.52a)

\[ \frac{\partial f_2(\xi_2, \eta_2)}{\partial t} = b_2(f_1(\xi_2, \eta_2), f_2(\xi_2, \eta_2)), \quad \xi_2 = \text{const}. \] (1.52b)

Therefore, for fixed \( \xi_1 \) and \( \xi_2 \), the lines \( x = \xi_1 + ct \) and \( x = \xi_2 - ct \) are the two characteristics of the system. Again, notice that the formulae for \( f_1(\xi_1, \eta_1), f_2(\xi_2, \eta_2) \) etc. are different from those of \( f_1(x, t), f_2(x, t) \): (1.52a) is the representation of (1.46a) in variables \((\xi_1, \eta_1)\) instead of \((x, t)\), and (1.52b) is the representation of (1.46b) in variables \((\xi_2, \eta_2)\) instead of \((x, t)\).

Similar to (1.29) and (1.30), we can obtain:

\[ f_1(\xi_1, \eta_1) = f_1(\xi_1, \eta_{10}) + \int_{\eta_{10}}^{\eta_1} b_1(f_1(\xi_1, \tau), f_2(\xi_1, \tau)) \, d\tau \] (1.53a)

\[ f_2(\xi_2, \eta_2) = f_2(\xi_2, \eta_{20}) + \int_{\eta_{20}}^{\eta_2} b_2(f_1(\xi_2, \tau), f_2(\xi_2, \tau)) \, d\tau, \] (1.53b)

and

\[ f_1(x, t) = f_1(x_0, t_0) + \int_{t_0}^{t} b_1(f_1(x_0 + c(\tau - t_0), \tau), f_2(x_0 + c(\tau - t_0), \tau)) \, d\tau \] (1.54a)

\[ f_2(x, t) = f_2(x_0, t_0) + \int_{t_0}^{t} b_2(f_1(x_0 - c(\tau - t_0), \tau), f_2(x_0 - c(\tau - t_0), \tau)) \, d\tau. \] (1.54b)

Similar to (1.32), if we evolve the solution over one time level for each step of
iterations based on (1.54a) and (1.54b), we have

\[(f_1)^{n+1}_m = (f_1)^{n-1}_m + \int_{nh}^{(n+1)h} b_1 \left( (f_1(x_{m-1} + c(\tau - t_n), \tau)), f_2(x_{m-1} + c(\tau - t_n), \tau) \right) d\tau \]  \hspace{1em} (1.55a)

\[(f_2)^{n+1}_m = (f_2)^{n+1}_m + \int_{nh}^{(n+1)h} b_2 \left( (f_1(x_{m+1} - c(\tau - t_n), \tau)), f_2(x_{m+1} - c(\tau - t_n), \tau) \right) d\tau. \]  \hspace{1em} (1.55b)

Now we use a numerical approximation for the integrals in (1.55a) to get the numerical scheme. If we use the SE method (1.21) for the numerical approximation of the integrals, then the numerical scheme for (1.55a) is

\[(f_1)^{n+1}_m = (f_1)^{n-1}_m + h b_1((f_1)^{n-1}_m, (f_2)^{n-1}_m) \]

\[(f_2)^{n+1}_m = (f_2)^{n+1}_m + h b_2((f_1)^{n+1}_m, (f_2)^{n+1}_m). \]  \hspace{1em} (1.56)

Now, if we use the ME method (1.22) for the numerical approximation of the integrals, then the numerical scheme for (1.55a) is

\[\bar{f}_1 = (f_1)^{n-1}_m + h b_1((f_1)^{n-1}_m, (f_2)^{n-1}_m)\]

\[\bar{f}_2 = (f_2)^{n+1}_m + h b_2((f_1)^{n+1}_m, (f_2)^{n+1}_m)\]

\[(f_1)^{n+1}_m = \frac{1}{2} \left[ (f_1)^{n}_m + \bar{f}_1 + h b_1(\bar{f}_1, \bar{f}_2) \right] \]

\[(f_2)^{n+1}_m = \frac{1}{2} \left[ (f_2)^{n+1}_m + \bar{f}_2 + h b_2(\bar{f}_1, \bar{f}_2) \right]. \]  \hspace{1em} (1.57)

Similarly, if we use the LF method (1.23) for the numerical approximation of the integrals in (1.55a), then we should consider the evolution over two time levels per
step and get the numerical scheme for (1.55a) as follows:

\[
(f_1)_m^{n+1} = (f_1)_{m-2}^{n-1} + 2hb_1((f_1)_{m-1}^{n}, (f_2)_{m-1}^{n})
\]

\[
(f_2)_m^{n+1} = (f_2)_{m+2}^{n-1} + 2hb_2((f_1)_{m+1}^{n}, (f_2)_{m+1}^{n}).
\]  

(1.58)

Now consider the two kinds of b.c. that we discussed in Sec. 1.3. Unlike in Sec. 1.5.1, but just as in Sec. 1.3, we now are to impose the b.c. at both the left and right boundaries. Let the index of the grid points on the space line be within these bounds: \(1 \leq m \leq M\).

In the SE scheme (1.56), we assume that the values of \(f_1, f_2\) at each discrete space point of the \(n\)th time level are known. When \(m = 1\), the calculation of \((f)_1^{n+1}\) requires values of \((f_1)_0^n, (f_2)_0^n\); when \(m = M\), the calculation of \((f)_M^{n+1}\) requires values of \((f_1)_{M+1}^n, (f_2)_{M+1}^n\). For the periodic b.c., they are assigned similarly to (1.36):

\[
(f_1)_0^n = (f_1)_M^n, \quad (f_2)_0^n = (f_2)_M^n
\]

\[
(f_1)_{M+1}^n = (f_1)_1^n, \quad (f_2)_{M+1}^n = (f_2)_1^n.
\]  

(1.59)
Therefore, according to (1.56), we have

\[
\begin{aligned}
(f_1)_m^{n+1} &= (f_1)_m^n + hb_1((f_1)_m^n, (f_2)_m^n) \\
(f_2)_m^{n+1} &= (f_2)_{m+1}^n + hb_2((f_1)_{m+1}^n, (f_2)_{m+1}^n); \\
(f_1)_1^{n+1} &= (f_1)_M^n + hb_1((f_1)_M^n, (f_2)_M^n) \\
(f_2)_1^{n+1} &= (f_2)_2^n + hb_2((f_1)_2^n, (f_2)_2^n); \\
(f_1)_M^{n+1} &= (f_1)_{M-1}^n + hb_1((f_1)_{M-1}^n, (f_2)_{M-1}^n) \\
(f_2)_M^{n+1} &= (f_2)_1^n + hb_2((f_1)_1^n, (f_2)_1^n).
\end{aligned}
\]  

(1.60)

When we apply the non-reflecting b.c. to (1.56), we have

\[(f_1)_1^n = G_1(t^n), \quad (f_2)_M^n = G_2(t^n),\]  

(1.61)

where $G_1$, $G_2$ are the given boundary values of $f_1$, $f_2$, respectively. Note, again, that they are assigned at the opposite end points of the spatial grid. Thus, according to
(1.56), we have

\[
\begin{align*}
(f_1)_m^{n+1} &= (f_1)_m^n + h b_1((f_1)_m^n, (f_2)_m^n) \quad 2 \leq m \leq M - 1 \\
(f_2)_m^{n+1} &= (f_2)_{m+1}^n + h b_2((f_1)_m^{n+1}, (f_2)_{m+1}^n); \\
(f_1)_1^{n+1} &= G_1(t_{n+1}) \\
(f_2)_1^{n+1} &= (f_2)_2^n + h b_2((f_1)^n_2, (f_2)^n_2); \\
(f_1)_M^{n+1} &= (f_1)_{M-1}^n + h b_1((f_1)_{M-1}^n, (f_2)_{M-1}^n) \\
(f_2)_M^{n+1} &= G_2(t_{n+1}).
\end{align*}
\]

(1.62)

In the ME scheme (1.57), we also assume that the values of \( f_1, f_2 \) at each discrete space point of the \( n \)th time level are known, and \((f_1)_0^n, (f_2)_0^n, (f_1)_M^{n+1}, (f_2)_M^{n+1}\) are the only required unknown values at \( n \)th time level. Therefore, similar to (1.60), if we apply the periodic b.c. to the ME scheme (1.57), then (1.59) yields

\[
\begin{align*}
(\tilde{f}_1)_m &= (f_1)_m^n + h b_1((f_1)_m^n, (f_2)_m^n) \\
(\tilde{f}_2)_m &= (f_2)_{m+1}^n + h b_2((f_1)_{m+1}^n, (f_2)_{m+1}^n) \quad 2 \leq m \leq M - 1 \\
(f_1)_m^{n+1} &= \frac{1}{2} \left[ (f_1)_m^n + (\tilde{f}_1)_m + h b_1 ((\tilde{f}_1)_m, (\tilde{f}_2)_m) \right] \\
(f_2)_m^{n+1} &= \frac{1}{2} \left[ (f_2)_m^n + (\tilde{f}_2)_m + h b_2 ((\tilde{f}_1)_m, (\tilde{f}_2)_m) \right].
\end{align*}
\]

(1.63a)
Similarly to (1.62), when we apply the non-reflecting b.c. to the ME scheme (1.57), Eqs. (1.61) yield

\[
\begin{aligned}
\left\{
\begin{array}{l}
(f_1)_1 = (f_1)_M^n + \Delta t \left( (f_1)_M^n, (f_2)_M^n \right) \\
(f_2)_1 = (f_2)_M^n + \Delta t \left( (f_1)_M^n, (f_2)_M^n \right)
\end{array}
\right.
\]  \hspace{1cm} (1.63b)

\[
\begin{aligned}
\left\{
\begin{array}{l}
(f_1)_{n+1}^1 = \frac{1}{2} \left[ (f_1)_M^n + (f_2)_M^n + \Delta t \left( (f_1)_M^n, (f_2)_M^n \right) \right] \\
(f_2)_{n+1}^1 = \frac{1}{2} \left[ (f_2)_M^n + (f_2)_M^n + \Delta t \left( (f_1)_M^n, (f_2)_M^n \right) \right].
\end{array}
\right.
\]  \hspace{1cm} (1.63c)

Similarly to (1.62), when we apply the non-reflecting b.c. to the ME scheme (1.57), Eqs. (1.61) yield

\[
\begin{aligned}
\left\{
\begin{array}{l}
(f_1)_1 = (f_1)_{M-1}^n + \Delta t \left( (f_1)_{M-1}^n, (f_2)_{M-1}^n \right) \\
(f_2)_1 = (f_2)_M^n + \Delta t \left( (f_1)_M^n, (f_2)_M^n \right)
\end{array}
\right.
\]  \hspace{1cm} (1.64a)

\[
\begin{aligned}
\left\{
\begin{array}{l}
(f_1)_{n+1}^m = \frac{1}{2} \left[ (f_1)_{m-1}^n + (f_2)_{m-1}^n + \Delta t \left( (f_1)_{m-1}^n, (f_2)_{m-1}^n \right) \right] \\
(f_2)_{n+1}^m = \frac{1}{2} \left[ (f_2)_{m+1}^n + (f_2)_{m+1}^n + \Delta t \left( (f_1)_{m+1}^n, (f_2)_{m+1}^n \right) \right]
\end{array}
\right. \quad 2 \leq m \leq M - 1
\]  \hspace{1cm} (1.64b)
\[
\begin{aligned}
(f_1)_1^n &= G_1(t_{n+1}) \\
(f_2)_1^n &= (f_2)_2^n + hb_2((f_1)_1^n, (f_2)_2^n) \\
(f_1)_{n+1}^1 &= G_1(t_{n+1}) \\
(f_2)_{n+1}^1 &= \frac{1}{2}[(f_2)_2^n + (f_2)_1^n + hb_2((f_1)_1^n, (f_2)_1^n)].
\end{aligned}
\] (1.64b)

\[
\begin{aligned}
(\tilde{f}_1)_M &= (f_1)_{M-1}^n + hb_1((f_1)_{M-1}^n, (f_2)_{M-1}^n) \\
(\tilde{f}_2)_M &= G_2(t_{n+1}) \\
(f_1)_{M+1}^n &= \frac{1}{2}[(f_1)_M^n + (\tilde{f}_1)_M^n + hb_1((\tilde{f}_1)_M^n, (\tilde{f}_2)_M^n)] \\
(f_2)_{M+1}^n &= G_2(t_{n+1}).
\end{aligned}
\] (1.64c)

In the LF scheme (1.58), we assume that at both the \(n\)th and \((n-1)\)th time levels, the values of \(f_1, f_2\) at each discrete space point are known. When \(m = 1\), the scheme requires the values: \((f_1)_{-1}^{n-1}, (f_1)_0^n, (f_2)_0^n\); when \(m = 2\), the scheme requires the value of \((f_1)_0^{n-1}\); when \(m = M\), the scheme requires the values: \((f_2)_{M+2}^{n-1}, (f_1)_{M+1}^n, (f_2)_{M+1}^n\); when \(m = M - 1\), the scheme requires the value of \((f_2)_{M+1}^{n-1}\). However, none of the values required above are defined. Therefore, in order to calculate the values of \(f\) at \(x = 1, 2, M, M - 1\), we need to use periodic b.c., or non-reflecting b.c.. If we apply the periodic b.c. to these undefined values, we have

\[
\begin{aligned}
(f_1)_{-1}^{n-1} &= (f_1)_{M-1}^{n-1}, & (f_1)_0^n &= (f_1)_M^n, & (f_2)_0^n &= (f_2)_M^n, & (f_1)_0^{n-1} &= (f_1)_M^{n-1} \\
(f_2)_{M+1}^{n-1} &= (f_2)_{M+1}^{n-1}, & (f_2)_{M+2}^{n-1} &= (f_2)_2^{n-1}, & (f_1)_{M+1}^n &= (f_1)_1^n, & (f_2)_{M+1}^n &= (f_2)_1^n.
\end{aligned}
\] (1.65)
Therefore, according to (1.58), we have

\[
\begin{align*}
(f_1)_m^{n+1} &= (f_1)_{m-2}^{n-1} + 2hb_1((f_1)_{m-1}^{n}, (f_2)_{m-1}^{n}) \quad 3 \leq m \leq M - 2 \\
(f_2)_m^{n+1} &= (f_2)_{m+2}^{n-1} + 2hb_2((f_1)_{m+1}^{n}, (f_2)_{m+1}^{n});
\end{align*}
\]

\[
\begin{align*}
(f_1)_1^{n+1} &= (f_1)_{M-1}^{n-1} + 2hb_1((f_1)_M^{n}, (f_2)_M^{n}) \\
(f_2)_1^{n+1} &= (f_2)_3^{n-1} + 2hb_2((f_1)_2^{n}, (f_2)_2^{n});
\end{align*}
\]

\[
\begin{align*}
(f_1)_2^{n+1} &= (f_1)_1^{n-1} + 2hb_1((f_1)_1^{n}, (f_2)_1^{n}) \\
(f_2)_2^{n+1} &= (f_2)_3^{n-1} + 2hb_2((f_1)_2^{n}, (f_2)_3^{n});
\end{align*}
\]

\[
\begin{align*}
(f_1)_{M-1}^{n+1} &= (f_1)_{M-3}^{n-1} + 2hb_1((f_1)_{M-2}^{n}, (f_2)_{M-2}^{n}) \\
(f_2)_{M-1}^{n+1} &= (f_2)_1^{n-1} + 2hb_2((f_1)_M^{n}, (f_2)_M^{n});
\end{align*}
\]

\[
\begin{align*}
(f_1)_M^{n+1} &= (f_1)_{M-2}^{n-1} + 2hb_1((f_1)_{M-1}^{n}, (f_2)_{M-1}^{n}) \\
(f_2)_M^{n+1} &= (f_2)_1^{n-1} + 2hb_2((f_1)_1^{n}, (f_2)_1^{n});
\end{align*}
\]

If we apply the non-reflecting b.c. to the LF scheme (1.58) then, similar to what is illustrated in Fig. 1.3, we can directly prescribe the values: \((f_1)_1^{n+1}, (f_2)_M^{n+1}\), but we need to choose specific method to calculate the values \((f_1)_2^{n+1}, (f_2)_{M-1}^{n+1}\). Using the
two sensible methods (1.44) and (1.45), we have

\[
\begin{align*}
(f_1)^{n+1}_1 &= G_1(t_{n+1}), \\
(f_1)^{n+1}_2 &= (f_1)^n_1 + hb_1 ((f_1)^n_1, (f_2)^n_1) \\
(f_2)^{n+1}_m &= G_2(t_{n+1}), \\
(f_2)^{n+1}_{M-1} &= (f_2)^n_M + hb_2 ((f_1)^n_M, (f_2)^n_M) \quad (1.67a) \\
\begin{cases}
(f_1)^{n+1}_m &= (f_1)^{n-1}_{m-2} + 2hb_1 ((f_1)^n_{m-1}, (f_2)^n_{m-1}) & 3 \leq m \leq M - 2, \quad (1.67b) \\
(f_2)^{n+1}_m &= (f_2)^{n-1}_{m+2} + 2hb_2 ((f_1)^n_{m+1}, (f_2)^n_{m+1}), \\
\end{cases}
\end{align*}
\]

which is using simple Euler method to calculate the values \((f_1)^{n+1}_2\), \((f_2)^{n+1}_{M-1}\), and

\[
\begin{align*}
(f_1)^{n+1}_1 &= G_1(t_{n+1}) \\
(f_1^1)_{k1} &= h b_1 ((f_1)^n_1, (f_2)^n_1), \\
(f_1^1)_{k2} &= h b_1 \left( (f_1)^n_1 + \frac{1}{2} (f_1^1)_{k1}, (f_2)^n_1 + \frac{1}{2} (f_2^1)_{k1} \right) \\
(f_1^1)_{k3} &= h b_1 \left( (f_1)^n_1 + \frac{1}{2} (f_1^1)_{k2}, (f_2)^n_1 + \frac{1}{2} (f_2^1)_{k2} \right) \\
(f_1^1)_{k4} &= h b_1 ((f_1)^n_1 + (f_1^1)_{k3}, (f_2)^n_1 + (f_2^1)_{k3}) \\
(f_2^1)_{k1} &= h b_2 ((f_1)^n_1, (f_2)^n_1) \\
(f_2^1)_{k2} &= h b_2 \left( (f_1)^n_1 + \frac{1}{2} (f_1^1)_{k1}, (f_2)^n_1 + \frac{1}{2} (f_2^1)_{k1} \right) \\
(f_2^1)_{k3} &= h b_2 \left( (f_1)^n_1 + \frac{1}{2} (f_1^1)_{k2}, (f_2)^n_1 + \frac{1}{2} (f_2^1)_{k2} \right) \\
(f_2^1)_{k4} &= h b_2 ((f_1)^n_1 + (f_1^1)_{k3}, (f_2)^n_1 + (f_2^1)_{k3}) \\
(f_1)^{n+1}_2 &= (f_1)^n_1 + \frac{1}{6} ((f_1^1)_{k1} + 2 (f_1^1)_{k2} + 2 (f_1^1)_{k3} + (f_1^1)_{k4}); \quad (1.68a)
\end{align*}
\]
\[(f_2)^{n+1}_M = G_2(t_{n+1})\]

\[(f_{1k1})_M = hb_1 ((f_1)_M^n, (f_2)_M^n), \quad (f_{2k1})_M = hb_2 ((f_1)_M^n, (f_2)_M^n)\]

\[(f_{1k2})_M = hb_1 \left((f_1)_M^n + \frac{1}{2} (f_{1k1})_M^n, (f_2)_M^n + \frac{1}{2} (f_{2k1})_M^n\right)\]

\[(f_{2k2})_M = hb_2 \left((f_1)_M^n + \frac{1}{2} (f_{1k1})_M^n, (f_2)_M^n + \frac{1}{2} (f_{2k1})_M^n\right)\]

\[(f_{1k3})_M = hb_1 \left((f_1)_M^n + \frac{1}{2} (f_{1k2})_M^n, (f_2)_M^n + \frac{1}{2} (f_{2k2})_M^n\right)\]

\[(f_{2k3})_M = hb_2 \left((f_1)_M^n + \frac{1}{2} (f_{1k2})_M^n, (f_2)_M^n + \frac{1}{2} (f_{2k2})_M^n\right)\]

\[(f_{1k4})_M = hb_1 ((f_1)_M^n + (f_{1k3})_M^n, (f_2)_M^n + (f_{2k3})_M^n)\]

\[(f_{2k4})_M = hb_2 ((f_1)_M^n + (f_{1k3})_M^n, (f_2)_M^n + (f_{2k3})_M^n)\]

\[(f_2)^{n+1}_M = (f_2)_M^n + 1 \cdot 6 ((f_{2k1})_M^n + 2 (f_{2k2})_M^n + 2 (f_{2k3})_M^n + (f_{2k4})_M^n); \quad (1.68b)\]

\[
\begin{align*}
(f_1)^{n+1}_m = (f_1)^{n-1}_m + 2 hb_1 ((f_1)^n_{m-1}, (f_2)^n_{m-1}) & \quad 3 \leq m \leq M - 2, \quad (1.68c) \\
(f_2)^{n+1}_m = (f_2)^{n-1}_m + 2 hb_2 ((f_1)^n_{m+1}, (f_2)^n_{m+1})
\end{align*}
\]

which is using the classical Runge–Kutta method to calculate the values: \((f_1)^{n+1}_2\), \((f_2)^{n+1}_M\).

In Chap. 3, we use schemes (1.56) to (1.58) for numerical integration of Eqs. (1.4).
Chapter 2

Physical instability of system (1.4)

2.1 Significance of considering physical instability

The goal of this thesis is to find a stable numerical scheme for (1.4). However, the physical solution for the dynamical system itself can be unstable. Since a scheme can be tested for its stability only on a physical solution that is stable, we first need to determine for what parameters our solution is stable.

We illustrate the idea of the physical stability with a simple example [13]: Suppose we are given a one-dimensional differential equation

\[ S'(t) = f(S). \]  (2.1)

Let \( S_0(t) \) be the exact solution of this equation. Let \( S \) deviate away from this solution...
by a small perturbation \( \tilde{S} \):

\[
S(t) = S_0(t) + \tilde{S}.
\]  

(2.2)

To see whether this perturbation grows or decays, we need to derive a differential equation for \( \tilde{S} \). Substituting \( S \) into the given differential equation gives \((S_0 + \tilde{S})' = f(S_0 + \tilde{S})\). Now using Taylor expansion we obtain

\[
S_0' + \tilde{S}' = f(S_0) + \tilde{S} f'(S_0) + O(\tilde{S}^2),
\]  

(2.3)

where \( O(\tilde{S}^2) \) denotes quadratically small terms in \( \tilde{S} \). The cancellation in the equation above is due to the fact that \( S_0 \) is a solution of the differential equation: hence \( S_0' = f(S_0) \). Now if \( f'(S_0) \) is not near 0, the \( O(\tilde{S}^2) \) terms are negligible and we may write

\[
\tilde{S}' = \tilde{S} f'(S_0).
\]  

(2.4)

This is a linear equation in \( \tilde{S} \), and it is called the linearization about \( S_0 \). It shows that the perturbation \( \tilde{S}(t) \) grows exponentially if \( f'(S_0) > 0 \) and decays if \( f'(S_0) < 0 \). When the perturbation term grows, we say there is physical instability, and the original solution is unstable. Similarly, physical instability may also occur in the solution to the PDE (1.4). We will derive the linearized PDE for Eqs. (1.4) in Sec. 2.3.

Therefore, in order to make sure that the numerical scheme is stable, we have to ensure that a physically stable perturbation does not grow systematically after many iterations of the numerical algorithm. Otherwise, if there is a physical instability in
the system, we may still observe a significant increase of the initial perturbation, even though the numerical scheme perfectly simulates the original PDE. Thus, to study the numerical stability of the scheme, one needs to investigate the conditions where the physical instability in the solution of the linearized PDE does not exist. As we will show in the remainder of this chapter, this condition determines the range of $b$ that we will consider in later chapters.

2.2 Perturbation and linearization

According to Sec.1.2, system (1.4) has a set of constant solutions (1.8b). Here, we will denote them as

$$S_{i0}^{\pm} = 0, \quad \text{where} \quad i = 1, 3, \quad \text{and} \quad S_2^{+} = 1, \ S_2^{-} = b. \quad (2.5)$$

We consider an initial condition as the sum of the exact solution (1.8b) and some small initial perturbations $\tilde{S}_i^{\pm}$, $i = 1, 2, 3$:

$$S_i^{\pm} = S_{i0}^{\pm} + \tilde{S}_i^{\pm}, \quad |\tilde{S}_i^{\pm}| \ll 1, \quad i = 1, 2, 3. \quad (2.6)$$

In the numerical computation, we may introduce this perturbation initially as a white noise. Below we consider the evolution of $\tilde{S}_i^{\pm}$. After we substitute (2.6) into (1.4) and linearize, which is similar to the linearization process in Sec. 2.1, i.e. ignore the smaller terms of the form $O(\tilde{S}_i^{\pm} \tilde{S}_j^{\pm})$, the six equations of (1.4) take on the form:
Note that, in the linear approximation, $\tilde{S}_2^\pm$ do not change along the characteristics and they are completely decoupled from the other $\tilde{S}_i^\pm$. Therefore, we will exclude these two variables and consider only the remaining four. Hence, we do not need the second and fifth equation of (2.7). The problem is reduced to four equations with four variables.

We write the reduced equations (2.7) in matrix form:

$$
(I\partial_t + \Sigma \partial_x) \tilde{S} = P\tilde{S},
$$

(2.8)

where $\tilde{S} = (\tilde{S}_1^+, \tilde{S}_3^+, \tilde{S}_1^-, \tilde{S}_3^-)^T$, $I$ is the $4 \times 4$ identity matrix, and

$$
\Sigma = \begin{pmatrix}
I & O \\
O & -I \\
\end{pmatrix}, \quad P = \begin{pmatrix}
bA & B \\
bB & A \\
\end{pmatrix}.
$$

(2.9)
Here, $I$ is the $2 \times 2$ identity matrix, $O$ is the $2 \times 2$ zero matrix, and

\[ A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & -2 \\ 1 & 0 \end{pmatrix}. \]  \hspace{1cm} \text{(2.10)}

Note that the boldface capital letters denote $4 \times 4$ matrices, while regular capital letters denote $2 \times 2$ matrices.

### 2.3 Physical Stability Analysis

Since Eqs. (2.8) have constant coefficients, we seek their solution in the following form:

\[ \tilde{S} = \mathbf{v} e^{(ikx - i\omega t)}, \]  \hspace{1cm} \text{(2.11)}

where $i = \sqrt{-1}$ and $\mathbf{v} = (v_1, v_2, v_3, v_4)^T$, with $v_1, \ldots, v_4$ being constants. For simplicity, we assume periodic boundary conditions (or an infinite interval as the domain of space variable $x$); then $k$ is a real number. Notice that $\omega$ determines whether the perturbation grows with time. In particular, if $\omega$ is a real number, the perturbation exhibits no physical instability. Substituting (2.11) into (2.8) gives:

\[ (i\omega I - ik\Sigma + P) \mathbf{v} = 0. \]  \hspace{1cm} \text{(2.12)}
Since we are only interested in nontrivial solutions, that is, $\mathbf{v} \neq \mathbf{0}$, we need to set
\[ \det (i\omega \mathbf{I} - ik \mathbf{\Sigma} + \mathbf{P}) = 0, \]
and thus get the following equation:

$$
\omega^4 + (-2k^2 - (b^2 - 4b + 1))\omega^2 + k(-2b^2 + 2)\omega + k^2(k^2 - (b^2 + 4b + 1)) = 0. \tag{2.13}
$$

It has four roots $\omega_j(k)$, $j = 1, \ldots, 4$. In order to ensure physical stability, below we will find the condition such that these four roots are all real for $k \in \mathbb{R}$. Since the only constant parameter in this polynomial is $b$, we need to find a range of values of $b$ such that $\omega_j(k)$ is real for any $j = 1, \ldots, 4$ and any $k \in \mathbb{R}$. We begin by rewriting (2.13) as follows:

$$
F(\omega, k, b) \equiv (\omega^2 - k^2)^2 - (\omega - k)^2 + 4b(\omega^2 - k^2) - b^2(\omega + k)^2 = 0, \tag{2.14}
$$

and subsequently analyze Eq. (2.14) instead of Eq. (2.13).

However, Eq. (2.14) cannot be solved analytically in compact form. Thus, one needs to consider special cases that are tractable. Eventually, the goal is to analyze if there exist a range of $k$ where $\omega$ is complex, which will indicate that there is physical instability.

### 2.4 Tractable case: $|b| \ll 1$

#### 2.4.1 Bifurcation at $b = 0$: Generic case

Let us note that there is a bifurcation, leading to physical instability, at $b = 0$, as explained in the caption to Fig. 2.1. Therefore, we first analyse Eq. (2.14) for $b = 0$.
Figure 2.1: Level curve $F(\omega, k, b) = 0$ for values of $b$ on opposite sides of 0. These plots are to be analysed as follows. If we draw a horizontal line $k = \text{const}$, the intercept points between the level curve and this horizontal line are the real-valued roots $\omega_j(k)$, where $1 \leq j \leq 4$. If $j < 4$, then there are complex values of $\omega$ with nonzero imaginary part, which indicates physical instability. From the left panel we can see that when $b > 0$, there exists “gaps” near the line $k = \pm 1/2$, one of which is marked out. Any horizontal line $k = \text{const}$ in such gap has less than four intercept points with the level curve, indicating that roots are not all real numbers; hence there exists physical instability. This is due to the graphical features near $(\omega, k) = (1/2, 1/2)$ and $(-1/2, -1/2)$. In the right panel where $b < 0$, there are four intercept points for any $k \in \mathbb{R}$, indicating that there is no physical instability. Therefore $b = 0$ appears to be a bifurcation point for physical instability. That is, for $b$ that is close to 0, there is no physical instability for $b < 0$, while there exists physical instability for $b > 0$.

and then extend the analysis to the case where $|b| \ll 1$. For $b = 0$, Eq. (2.14) reduces to:

$$ (\omega - k)^2((\omega + k)^2 - 1) = 0. \quad (2.15) $$

This equation has four roots for $\omega$: $\omega = k$ (double root) and $\omega = -k \pm 1$. These four curves are the asymptotes of the two hyperbolas in each panel of Fig. 2.1. Therefore, we will look for the roots $\omega(k)$ near $\omega = k$ and $\omega = -k \pm 1$ for $|b| \ll 1$.

For $\omega \approx k$, we let $\omega = k + \tilde{\omega}$ with $|\tilde{\omega}| \ll |k|$ in (2.14), assume $|b| \ll 1$, linearize,
and get the following approximation:

\[ \tilde{\omega} = -4kb \pm 2\sqrt{k^2 b^2 (4k^2 + 3)} \]
\[ \quad \frac{4k^2 - 1}{4k^2 - 1}. \]  \hspace{1cm} (2.16)

Therefore, if \( k^2 \) is not near \( 1/4 \), i.e. \( k \) is not near \( 1/2 \) or \( -1/2 \), we have

\[ \omega = k + \frac{-4kb \pm 2\sqrt{k^2 b^2 (4k^2 + 3)}}{4k^2 - 1}. \] \hspace{1cm} (2.17)

Now, for \( \omega \approx -k + 1 \), where \( k \) is not near \( 1/2 \), we use a similar procedure and get:

\[ \omega = (-k + 1) + \frac{2b}{2k - 1}. \] \hspace{1cm} (2.18)

Similarly, for \( \omega \approx -k - 1 \) and when \( k \) is not near \( -1/2 \), we can get:

\[ \omega = (-k - 1) + \frac{2b}{2k + 1}. \] \hspace{1cm} (2.19)

Thus, for \( k \) not near \( \pm 1/2 \), the perturbation analysis provides the explicit expression of the four roots when \( b \) is perturbed away from 0. These roots are all real numbers, which implies that for \( k \) not near \( 1/2 \) or \( -1/2 \), there will always be four real roots for \( F(\omega, k, b) = 0 \). However, these results all fail when \( k \) is near \( 1/2 \) or \( -1/2 \). Therefore, we need further perturbation analysis for \( k \) near \( \pm 1/2 \).
2.4.2 Bifurcation at $b = 0$:

**ANALYSIS NEAR** $(\omega, k) = \pm \left( \frac{1}{2}, \frac{1}{2} \right)$

For $k = 1/2$, the double root and the root $\omega = -k + 1$ all yield $\omega = 1/2$. For $k = -1/2$, the double root and the root $\omega = -k - 1$ yield $\omega = -1/2$. We first perform perturbation analysis of $F(\omega, k, b) = 0$ near $\omega = k = 1/2$.

$$F \left( \frac{1}{2} + \tilde{\omega}, \frac{1}{2} + \tilde{k}, b \right) = \left[ 2\tilde{\omega}^3 - 2\tilde{k}\tilde{\omega}^2 - 2\tilde{k}^2\tilde{\omega} + 2\tilde{k}^3 + \cdots \right] + 4b \left[ \tilde{\omega} - \tilde{k} \cdots \right] - b^2 \left[ 1 + \cdots \right]$$

$$= [O(\omega^3) + O(\tilde{k}\omega^2) + O(\tilde{k}^2\omega) + O(\tilde{k}^3)] + [O(b\tilde{\omega}) + O(b\tilde{k})] + O(b^2) = 0. \quad (2.20)$$

The “$\cdots$” in the first line of Eq. (2.20) stand for higher-order terms which we omit.

We will now discuss the order of magnitude of the three parts in the last line of (2.20) with the purpose to find possible relations among $\tilde{k}, \tilde{\omega}$, and $b$. By inspection, we have found only two such relations, which are listed below:

**Case 1:** If $O(\tilde{\omega}^3) \sim O(b\tilde{\omega})$, then $\tilde{\omega} \sim O(b^{1/2})$, and the last line of Eq. (2.20) yields:

$$F = [O(\omega^3) + O(\tilde{k}\omega^2) + O(\tilde{k}^2\omega) + O(\tilde{k}^3)] + [O(b\tilde{\omega}) + O(b\tilde{k})] + O(b^2)$$

$$= [O(b^{3/2}) + O(\tilde{k}b) + O(\tilde{k}^2b^{1/2}) + O(\tilde{k}^3)] + [O(b^{3/2}) + O(b\tilde{k})] + O(b^2)$$

$$= O(b^{3/2}) + O(\tilde{k}b) + O(\tilde{k}^2b^{1/2}) + O(\tilde{k}^3), \quad (2.21)$$

where the last term in the second line is omitted as being of a higher order. For Eq. (2.21) to hold, at least two of the terms must have the same order of magnitude, and the remaining terms must have a higher order. This yields six possible cases on the r.h.s of (2.21): $O(b^{3/2}) \sim O(\tilde{k}b)$, $O(b^{3/2}) \sim O(\tilde{k}^2b^{1/2})$, $O(b^{3/2}) \sim O(\tilde{k}^3)$,
\(O(\tilde{k}b) \sim O(\tilde{k}^2b^{1/2})\), \(O(\tilde{k}b) \sim O(\tilde{k}^3)\), and \(O(\tilde{k}^2b^{1/2}) \sim O(\tilde{k}^3)\). They all give the same result:

\[
\tilde{k} \sim O(b^{1/2}),
\] (2.22)

whereby all terms in Eq. (2.21) have the same order of magnitude. Keeping these terms in the first line of Eq. (2.20) yields:

\[
\dot{\omega}^3 - \tilde{k}\dot{\omega}^2 - \tilde{k}^2\dot{\omega} + \tilde{k}^3 + 2b\dot{\omega} - 2b\tilde{k}
=\dot{\omega}^2(\dot{\omega} - \tilde{k}) - \tilde{k}^2(\dot{\omega} - \tilde{k}) + 2b(\dot{\omega} - \tilde{k})
=(\dot{\omega} - \tilde{k})[(\dot{\omega}^2 - \tilde{k}^2) + 2b] = 0,
\] (2.23)

which gives two possible solutions: \(\dot{\omega} = \tilde{k}\) and \(\dot{\omega}^2 = \tilde{k}^2 - 2b\). Let us focus on the latter one.

- If \(b < 0\), then there are always two real solutions \(\dot{\omega} = \pm\sqrt{\tilde{k}^2 - 2b}\);
- If \(b > 0\), then there are two real solutions \(\dot{\omega} = \pm\sqrt{\tilde{k}^2 - 2b}\) if \(|\tilde{k}| \geq \sqrt{2b}\), and no real solutions if \(|\tilde{k}| < \sqrt{2b}\).

Case 2: If \(O(b\tilde{\omega}) \sim O(b^2)\), then \(\tilde{\omega} \sim O(b)\), and the last line of Eq. (2.20) yields:

\[
0 = [O(\tilde{\omega}^3) + O(\tilde{k}\tilde{\omega}^2) + O(\tilde{k}^2\tilde{\omega}) + O(\tilde{k}^3)] + [O(b\tilde{\omega}) + O(b\tilde{k})] + O(b^2),
\] (2.24)

where the first group of terms is omitted as being of higher order. The remaining
terms in (2.24) yields:

\[ \tilde{k} \sim O(b). \quad (2.25) \]

Substituting \( \tilde{\omega} \sim \tilde{k} \sim \tilde{b} \) in the first line of Eq. (2.20) yields:

\[ 4b\tilde{\omega} - 4b\tilde{k} + b^2 = 0, \quad \tilde{\omega} = \tilde{k} - b/4. \quad (2.26) \]

From the discussion above we can see that near the value \( k = 1/2 \) (i.e. for \( k = 1/2 + \tilde{k}, \ |\tilde{k}| \ll 1 \)), the number of real solutions \( \omega \) of \( F(\omega, k, b) = 0 \) in Eq. (2.14) depends on the sign of \( b \) when \( |b| \ll 1 \):

- \( b > 0 \): When \( |\tilde{k}| \geq \sqrt{2b} \), there are three real solutions for \( F(\omega, k, b) = 0 \) near \( \omega = \frac{1}{2} \):

\[ \omega = \frac{1}{2} + \sqrt{\tilde{k}^2 - 2b}, \quad \omega = \frac{1}{2} - \sqrt{\tilde{k}^2 - 2b}, \quad \omega = \frac{1}{2} + \tilde{k} - \frac{b}{4}. \quad (2.27) \]

Along with the solution near \( \omega = -\frac{3}{2} \), there would be \emph{four} real solutions of \( F(\omega, k, b) = 0 \) for \( k \approx 1/2 \), which means there are no complex solution \( \omega \) with nonzero imaginary part. Combining this with the results in Sec. 2.4.1, we can see that physical instability does not exist for \( |\tilde{k}| \geq \sqrt{2b} \). However, when \( |\tilde{k}| < \sqrt{2b} \), there will only be one real solution of \( \omega \):

\[ \omega = \frac{1}{2} + \tilde{k} - \frac{b}{4}. \quad (2.28) \]

Along with the solution near \( \omega = -\frac{3}{2} \), there are only two real solutions for the
fourth-degree (in $\omega$) polynomial $F(\omega, k, b) = 0$ for $k \approx 1/2$. Thus, there exists complex solutions for $\omega$ with nonzero imaginary part for certain values of $k$, indicating that there exists physical instability if $0 < b \ll 1$.

- $b \leq 0$: There will always be three real solutions for $F(\omega, k, b) = 0$ near $\omega = 1/2$:

$$\omega = \frac{1}{2} + \sqrt{\tilde{k}^2 - 2b}, \quad \omega = \frac{1}{2} - \sqrt{\tilde{k}^2 - 2b}, \quad \omega = \frac{1}{2} + \tilde{k} - \frac{b}{4}. \quad (2.29)$$

since $\tilde{k}^2 - 2b \geq 0$ for any $\tilde{k}$ if $b \leq 0$. Along with the solution near $\omega = -\frac{3}{2}$, there are always four real roots for $F(\omega, k, b) = 0$ for any $k \approx 1/2$. Combining this with the results in Sec. 2.4.1, we conclude that there is no physical instability if $b < 0$ and $|b| \ll 1$.

Similarly, we have the same situation for $k \approx -1/2$ (i.e. for $k = -1/2 + \tilde{k}$, $|k| \ll 1$ and $|b| \ll 1$). First, of all the four roots of $F(\omega, k, b) = 0$ in terms of $\omega$, there will always be one real root near $3/2$. Then there are three possible real roots near $-1/2$. This is also shown in Fig. 2.1. But two of the real roots near $-1/2$ exits only when $b < 0$, in which case there will be four real roots in total, hence no physical instability. When $b > 0$, these two real roots near $-1/2$ don’t exist, in which case there are only two real roots in total, making the solution unstable. Combining the results of perturbation analysis in Sec. 2.4.1, 2.4.2, we conclude that when $|b| \ll 1$, there exits physical instability if $b > 0$ and no physical instability if $b \leq 0$. 

42
2.5 Tractable case: $|k| \ll 1$

The other tractable case is for $|k| \ll 1$. To make our notation compact, write (2.14) as:

$$F(\omega, k, b) = \omega^4 + (-2k^2 + G_1)\omega^2 + G_2k\omega + k^2(k^2 + G_3),$$  \hspace{1cm} (2.30)

where

$$G_1 = -(b^2 - 4b + 1)$$
$$G_2 = -2b^2 + 2$$
$$G_3 = -(b^2 + 4b + 1).$$  \hspace{1cm} (2.31)

For $k = 0$, Eq. (2.30) reduces to

$$F(\omega, 0, b) = \omega^4 + G_1\omega^2 = 0,$$  \hspace{1cm} (2.32)

whose solutions are

$$\omega_{1,2}(0) = \pm \sqrt{-G_1}, \quad \omega_{3,4}(0) = 0.$$  \hspace{1cm} (2.33)

From this result, we can see that if $G_1 > 0$, then $\omega_{1,2}(0)$ are complex with non-zero imaginary part and hence the perturbation term (2.11) will grow exponentially, which means that solution (2.5) of system (1.4) is undoubtedly unstable in that case. However, it is not accurate to say that $G_1 \leq 0$ is sufficient condition for all four
solutions to be real, in which case, there is no physical instability. We need further
analysis.

For \(|k| \ll 1\), we can perform perturbation analysis by substituting \(\omega = \omega_j(0) + \tilde{\omega}\),
\(j = 1, \ldots, 4\), into (2.30):

\[
0 = F(\omega_j + \tilde{\omega}, k, b) = (\omega_j + \tilde{\omega})^4 + (-2k^2 + G_1)(\omega_j + \tilde{\omega})^2 + G_2 k(\omega_j + \tilde{\omega}) + k^2(k^2 + G_3)
\]

\[
= (\omega_j^4 + G_1 \omega_j^2) + 4\omega_j^3 \tilde{\omega} + 6\omega_j^2 \tilde{\omega}^2 + 4\omega_j \tilde{\omega}^3 + \tilde{\omega}^4
\]

\[
+ 2\omega_j G_1 \tilde{\omega} + G_1 \tilde{\omega}^2 + (-2\omega_j^2)k^2 + (-4\omega_j)k^2 \tilde{\omega} - 2k^2 \tilde{\omega}^2
\]

\[
+ G_2 k\omega_j + G_2 k\tilde{\omega} + k^4 + k^2G_3, \tag{2.34}
\]

where \(\omega_j\) is one of the solutions in (2.33), and \(\tilde{\omega}\) is the corresponding perturbation
such that \(|\tilde{\omega}| \ll 1\). The first term in the last expression vanished due to (2.32). Note
that we do not know the relative size of \(\tilde{\omega}\) in comparison with \(k\), since they are both
small. Therefore, in (2.34), we have not yet discarded any terms and hence it is not
simpler than the original Eq. (2.30).

To simplify (2.33), we need to consider the cases \(\omega_j = \omega_{1,2}(0)\) and \(\omega_j = \omega_{3,4}(0)\)
separately. Let \(\omega_{1,2}(0) = \pm \sqrt{-G_1} \neq 0\). Then we can neglect all terms \(\tilde{\omega}^2, \tilde{\omega}^3, \tilde{\omega}^4,\)
and (2.34) is simplified to :

\[
(4\omega_j^3 + 2\omega_j G_1)\tilde{\omega} + (-2\omega_j^2)k^2 + (-4\omega_j)k^2 \tilde{\omega} + G_2 k\omega_j + G_2 k\tilde{\omega} + k^4 + k^2 G_3 = 0 \tag{2.35}
\]

In (2.35), we omitted all the terms other than \((4\omega_j^3 + 2\omega_j G_1)\tilde{\omega}\) and \(G_2 k\omega_j\) because,
firstly, in comparison with \(G_2 k\omega_j\), which is \(O(k)\), any of the omitted terms are one
of undoubtedly smaller terms \(O(k^2), O(k^2 \tilde{\omega}), O(k \tilde{\omega}), O(k^4), \) and \(O(k^2)\). We cannot
omit any of \((4\omega_j^3 + 2\omega_j G_1)\hat{\omega}\) or \(G_2 k \omega_j\) since we cannot determine which one of them is smaller. Since \(\omega_j = \omega_{1,2} \neq 0\), we can cancel out \(\omega_j\) in the rest two terms by dividing \(\omega_j\) on both sides and get

\[
(4\omega_j^2 + 2G_1)\hat{\omega} + G_2 k = 0. \tag{2.36}
\]

Finally, since \(\omega_j^2 = \omega_{1,2}^2 = -G_1\), (2.36) becomes \(-2G_1 \hat{\omega} + G_2 k = 0\) and yields \(\hat{\omega} = \frac{G_2 k}{2G_1}\). This result means that the corresponding \(\hat{\omega}\) for \(\omega_j = \omega_{1,2}(0)\) are real numbers for \(|k| \ll 1\). Thus, if \(\omega_j = \omega_{1,2}\), then whether or not \(\omega = \omega_j + \hat{\omega}\) is real number depends on whether or not \(\omega_j = \omega_{1,2} = \pm \sqrt{-G_1}\) is real number.

Now, when \(\omega_j = 0\), which means either \(\omega_j = \omega_{3,4}\) or \(\omega_j = \omega_{1,2}\) with \(G_1 = 0\), we need to consider terms of order higher than \(O(\hat{\omega})\) in (2.34). Substituting \(\omega_j = 0\) into (2.34) yields.

\[
(\hat{\omega}^4 + G_1 \hat{\omega}^2) + (2k^2 \omega^2 + G_2 k \hat{\omega}) + (k^4 + k^2 G_3) = 0. \tag{2.37}
\]

In (2.37), we can omit \(\hat{\omega}^4\) since it is smaller than \(G_1 \hat{\omega}^2\), which is \(O(\hat{\omega}^2)\); we can omit \(-2k^2 \omega^2\), which is \(O(k^2 \omega^2)\), since it is smaller than \(G_2 k \hat{\omega}\), which is \(O(k \hat{\omega})\); we can omit \(k^4\) since it is smaller than \(k^2 G_3\), which is \(O(k^2)\). After the omission, we have:

\[
G_1 \hat{\omega}^2 + G_2 k \hat{\omega} + k^2 G_3 = 0. \tag{2.38}
\]

According to the equation above, in order for the corresponding \(\hat{\omega}\) to be real numbers,
we need:

\[ (G_2^2 - 4G_1G_3)k^2 \geq 0 \quad \Rightarrow \quad G_2^2 - 4G_1G_3 \geq 0. \quad (2.39) \]

Thus, collecting the conclusions stated after (2.33), (2.36), and (2.38), we conclude that in order to have \( \omega(k) \in \mathbb{R} \), for \(|k| \ll 1\), we need

\[
\begin{cases}
G_1 \leq 0 \\
G_2^2 - 4G_1G_3 \geq 0
\end{cases}
\quad \Rightarrow \quad \begin{cases}
b^2 - 4b + 1 \geq 0 \\
(-2b + 2)^2 - 4(b^2 - 4b + 1)(b^2 + 4b + 1) \geq 0
\end{cases}
\quad \Rightarrow \quad \begin{cases}
b^2 - 4b + 1 \geq 0 \\
12b^2 \geq 0
\end{cases}
\quad \Rightarrow \quad b \in (-\infty, 2 - \sqrt{3}] \cup [2 + \sqrt{3}, +\infty).
\]

From the discussion above, we can see that if \( b \in (-\infty, 2 - \sqrt{3}] \cup [2 + \sqrt{3}, +\infty) \) and \(|k| \ll 1\), there is no \( \omega(k) \) with nonzero imaginary part, and hence the solution (1.8b) will not have physically unstable perturbations with \(|k| \ll 1\). Conversely, \( \omega(k) \) for \(|k| \ll 1\) will always have a nonzero imaginary part if \( b \in (2 - \sqrt{3}, 2 + \sqrt{3}) \). Thus, we conclude that when \( b \in (2 - \sqrt{3}, 2 + \sqrt{3}) \), physical instability exists.
2.6 Remaining cases

2.6.1 Relations between regions $b \in (0, 1]$ and $b \in [1, \infty)$

Here, we will show that the existence or non-existence of physical instability for some $b$ implies the same conclusion for $1/b$. This idea comes from the inspection of the result of Sec. 2.5. There, we showed that there is physical instability for $b \in (2 - \sqrt{3}, 2 + \sqrt{3})$.

Now notice that $\frac{1}{2 - \sqrt{3}} = 2 + \sqrt{3}$, i.e., this interval of physical instability is symmetrical around 1. Below, we show that this symmetry exists for all $b$.

First, we investigate the solution of $F(\omega, k, \frac{1}{b}) = 0$. Let us rewrite Eq. (2.14) in the following equivalent form:

$$F(\omega, k, b) = (\omega + k)^2(\omega - k)^2 - (\omega - k)^2 + 4b(\omega + k)(\omega - k) - b^2(\omega + k)^2. \quad (2.43)$$

Notice that

$$F \left( \omega, k, \frac{1}{b} \right) = (\omega + k)^2(\omega - k)^2 - (\omega - k)^2 + \frac{4}{b}(\omega + k)(\omega - k) - \frac{1}{b^2}(\omega + k)^2$$

$$= (\omega + k)^2(\omega - k)^2 - \frac{1}{b^2}(\omega + k)^2 + \frac{4}{b}(\omega + k)(\omega - k) - (\omega - k)^2$$

$$= \frac{1}{b^4} \left( b^4(\omega - k)^2(\omega + k)^2 - b^2(\omega + k)^2 + 4b^3(\omega - k)(\omega + k) - b^4(\omega - k)^2 \right)$$

$$= \frac{1}{b^4} F(b\omega, -bk, b). \quad (2.44)$$
Thus, for any $b \neq 0$,

$$F\left(\omega, k, \frac{1}{b}\right) = 0$$

$$\iff \frac{1}{b^4} F(b\omega, -bk, b) = 0$$

$$\iff F(\bar{\omega}, \bar{k}, b) = 0,$$  \hspace{1cm} (2.45)

where

$$\bar{\omega} = b\omega, \quad \bar{k} = -bk.$$  \hspace{1cm} (2.46)

This result shows that solving $F(\omega, k, 1/b) = 0$ for $\omega$ is equivalent to solving $F(\bar{\omega}, \bar{k}, b) = 0$ for $\bar{\omega}$. Suppose that when we solve

$$F(x, y, z) = 0,$$  \hspace{1cm} (2.47)

we get $x = f(y, z)$. Then when we solve $F(\bar{\omega}, \bar{k}, b) = 0$, we will get $\bar{\omega} = f(\bar{k}, b)$.

Substituting $x = \bar{\omega} = b\omega$, $y = \bar{k} = -bk$, and $z = b$ in (2.47), we find that $\omega = \frac{1}{b} f(-bk, b)$ is the solution of $F(b\omega, -bk, b) = 0$. Hence, according to (2.45),

$$\omega = \frac{1}{b} f(-bk, b)$$  \hspace{1cm} (2.48)

is also a solution of $F(\omega, k, 1/b) = 0$.

Now, let us investigate the relation between the solution of $F (\omega, k, \frac{1}{b}) = 0$ and the solution of $F (\omega, k, b) = 0$. According to (2.47), the solution of $F(\omega, k, b) = 0$ is
\[ \omega = f(k, b). \] Suppose now that through solving \( F(\omega, k, b) = 0 \), we found that

\[ \omega = f(k, b) \in \mathbb{R}, \quad \forall k \in \mathbb{R}, \]  

(2.49)

where \( b \neq 0 \) is fixed. Obviously this implies that

\[ \frac{1}{b} f(k, b) \in \mathbb{R}, \quad \forall k \in \mathbb{R}. \]  

(2.50)

If we replace \( k \) with \(-bk\) in (2.50), then

\[ \frac{1}{b} f(-bk, b) \in \mathbb{R}, \quad \forall (-bk) \in \mathbb{R}. \]  

(2.51)

Since \(-bk \in (-\infty, \infty)\) is equivalent to \( k \in (-\infty, \infty) \) when \( b \neq 0 \), (2.51) implies that

\[ \frac{1}{b} f(-bk, b) \in \mathbb{R}, \quad \forall k \in \mathbb{R}. \]  

(2.52)

From the facts that: (i) by (2.48), \( \frac{1}{b} f(-bk, b) \) is the solution of \( F\left(\omega, k, \frac{1}{b}\right) = 0 \) and (ii) Eq. (2.49) (2.52), we conclude that: If the solution for \( F(\omega, k, b) = 0 \) is real for all \( k \in \mathbb{R} \), then the solution of \( F(\omega, k, 1/b) = 0 \) must also be real for all \( k \in \mathbb{R} \). Thus, for any fixed \( b \neq 0 \), if the four roots of \( F(\omega, k, b) = 0 \), i.e., \( \omega_j(k) \), with \( j = 1, 2, 3, 4 \) are real for all \( k \in \mathbb{R} \), we must have the same conclusion for \( F(\omega, k, 1/b) \). According to this argument, the intervals \( b \in (0, 1] \) and \( b \in [1, +\infty) \) are equivalent in regards to the existence of physical instability. Similarly, the intervals \( b \in [-1, 0) \) and \( b \in (+\infty, -1] \) are also equivalent in regards to the existence of physical instability.
2.6.2 Physical instability for any $b \in (0, +\infty)$

According to the conclusion of Sec. 2.4, there exists physical instability when $0 < b \ll 1$. Also, according to the result of Sec. 2.5, there exists physical instability when $b \in (2 - \sqrt{3}, 2 + \sqrt{3})$. If we could consider $b = 2 - \sqrt{3} \approx 0.26$ to be small enough to be regarded as $|b| \ll 1$, then we could have claimed that there had been physical instability for any $b \in (0, 2 - \sqrt{3})$. Thus we could have concluded that there must have existed physical instability for any $b \in (0, 1]$. Hence, because of the scaling in Sec. 2.6.1, there would have been physical instability for any $b \in [1, \infty)$. This would have implied that physical instability existed for any $b \in (0, \infty)$. However, we cannot consider $b = 2 - \sqrt{3} \approx 0.26$ as small enough. This is because we need to see how the coefficient $G_1(b)$ in (2.40) changes when $b$ changes from 0 to $2 - \sqrt{3}$. According to the expression of $G_1$ in (2.31), $G_1(0) = 1$, while $G_1(2 - \sqrt{3}) = 0$; thus the change is $O(1)$, hence making $b = 2 - \sqrt{3}$ not small enough. Therefore for some small number $0 < \delta < 2 - \sqrt{3}$, where $|G_1(\delta) - G_1(0)| \ll 1$, $(0, \delta]$ is the interval of $b$ where the physical instability exists for sure, but the interval $(\delta, 2 - \sqrt{3})$ is a “grey area” where we cannot analytically determine the existence of the physical instability. This is illustrated by Fig. 2.2. Thus, we resolve the question of whether instability exists for $b \in (\delta, 2 - \sqrt{3})$ numerically.

One numerical method to verify if there is physical instability for any $b \in (0, 1]$ consists of plotting the level curve $F(\omega, 1/2, b) = 0$ in the $(b, \omega)$ plane. The result is shown in Fig. 2.3. From this numerical result we can see that for any $b \in (0, 1]$, and $k = \frac{1}{2}$, the fourth-degree polynomial $F(\omega, k, b)$ has only two real roots $\omega$, hence implying the existence of physical instability for any $b \in (0, 1]$. 

50
Figure 2.2: The thick line segments mark the intervals of $b$ where the physical instability definitely exists according to the results of Sec. 2.4 and Sec. 2.5. The interval with a question mark is where the physical instability cannot be explicitly determined by the analysis of Sec. 2.4 and Sec. 2.5. Here, the number $\delta$ is some small positive number such that $|G_1(\delta) - G_1(0)| \ll 1$, square bracket means the closed end of an interval, parenthesis means the open end of an interval. Note that $2 - \sqrt{3} \approx 0.2679$, $2 + \sqrt{3} \approx 3.7321$, and $1/(2 - \sqrt{3}) = 2 + \sqrt{3}$.

Alternatively, the existence of unstable physical perturbation can be demonstrated as follows. We can plot the level curve $F(\omega, k, b) = 0$ in $(\omega, k)$ plane, for some representative values of $b$. The results for $b = 0.01, 0.1, 0.2, 0.3, 0.6, 1$ are shown in Fig. 2.4. From these graphs we can see that the same feature is preserved throughout all the panels of Fig. 2.4: there exists “gaps” such that any horizontal line $k = \text{const}$ in these “gaps” has only two intercept points with the level curve. As in the previous paragraph, this implies that there will always be physical instability when $b$ is gradually changing from 0 to 1.

Based on this numerical evidence, we conclude that there is physical instability for all $b \in (0, 1]$. Again, due to the scaling discussed in Sec. 2.6.1, we can conclude that there is also physical instability for all $b \in [1, +\infty)$. Therefore, there is physical instability for all $b > 0$. 

51
Figure 2.3: Level curve $F(\omega, 0.5, b) = 0$, where $k = \frac{1}{2}$. For each $b \in (0, 1]$, there are only two real roots for $\omega$, which implies that there is physical instability. However, when $b < 0$, there will be four real roots for $\omega$, and hence no physical instability with $k \approx 1/2$. The bifurcation point is $b = 0$.

2.6.3 PHYSICAL STABILITY FOR ANY $b \in (-\infty, 0)$

First, we point out a symmetry between the level curve $F(\omega, k, b) = 0$ with positive $b$ and that with negative $b$. In Eq. (2.14), if we replace $b$ with $-b$ and interchange $k$ with $\omega$, we find the following relation:

$$F(\omega, k, b) \equiv F(k, \omega, -b). \quad (2.53)$$

Therefore, the level curve $F(\omega, k, b) = 0$ with $b < 0$ is obtained from that with $b > 0$ by the reflection about the line $k = \omega$.

This symmetry can be found in Fig. 2.1 when we compare the level curves for $b = 1$ and $b = -1$. Note that because of it, the “gap” discussed at the end of Sec. 2.6.2 exists for $b = 1$, but not for $b = -1$. Thus, the situation where a line $k = \text{const}$ can have fewer than four intercepts with the level curve occurs only for $b = 1$, whereas there are
Figure 2.4: Level curve $F(\omega, k, b) = 0$ for values of $b$ changing from 0.01 to 1. We can see that the level curve $F(\omega, k, b) = 0$ changes gradually as $b$ changes from 0.01 to 1. Most importantly, it maintains one graphical feature: There exist two gaps, one of which is marked in the first panel, such that any horizontal line $k = \text{const}$ in a gap has fewer than four intercepts with the level curve. This feature implies that for any $0 < b \leq 1$, there will always be values of $k$ such that there are only two real roots $\omega(k)$. Hence, there exists instability for perturbations with these $k$.

always four intercepts when we take $b = -1$. This graphical feature is preserved when $|b|$ changes from 0 to 1, as shown in Fig. 2.5. We can see from Fig. 2.5 that for $b < 0$, the number of intercepts between the level curve $F(\omega, k, b) = 0$ and any horizontal
Figure 2.5: Level curve $F(\omega, k, b) = 0$. The left panels are for positive values of $b$ and right panels are for their negative counter parts. In each left panel, there exists “gap” such that any horizontal line $k = \text{const}$ in between such “gap” has less than four intercept points with the level curve, while in the right panels, due to the graphical feature of symmetry described above, such “gaps” do not exist, and each horizontal line $k = \text{const}$ has four intercept points with the level curve.

Line $k = \text{const}$ is always four. This implies that there are always four real roots $\omega(k)$ and hence, there is no physical instability. Note that in Fig. 2.5, the shape of the level curves changes with $|b|$ gradually. This assures us that the graphical features described above are preserved for all values of $|b| \in (0, 1)$. Hence, for $-1 \leq b \leq 0$,......
there is no physical instability. Finally, because of the scaling discussed in Sec. 2.6.1, there is no physical instability for any \( b \in (\infty, -1] \) either. Therefore, we conclude that there is no physical instability for any \( b \in (\infty, 0] \).
CHAPTER 3

VON NEUMANN ANALYSIS FOR THE NUMERICAL SCHEMES

3.1 Simple Euler Method

In this section, our goal is to analyze the instability of the SE scheme when it is applied to (1.4) based on the MoC; see Eqs. (1.56) in Chap. 2. The approach we use here is the von Neumann analysis, which considers the b.c. to be periodic, or the space line to be infinite. This approach is straightforward and gives accurate information about stability of the numerical scheme with periodic b.c..

To study a numerical instability, we need to investigate how the numerical scheme affects the initial perturbation. To that end, we will first derive the corresponding numerical scheme for (2.7) and write its effect on the initial perturbation term $\tilde{S}$ in the form of matrix multiplication (3.5). Based on this matrix multiplication, we will use the fact that the instability is dominated by the largest eigenvalue of the numerical matrix $\mathbf{N}$ (see (3.5)). Therefore, we need to analyze the magnitude of
the dominating eigenvalue. We will do so in two ways: numerically and analytically. In the numerical approach, we will obtain the dominating eigenvalue using Matlab. Then, for those wavenumbers where numerics predict the eigenvalue with the largest magnitude, we will derive an analytical formula of this eigenvalue. Finally, by direct numerical simulations of Eq. (1.4), we will show that the error growth rate is explained by this dominant eigenvalue.

3.1.1 **Matrix form of the SE**

Applying the general form (1.56) of the numerical SE based on the MoC to the system of coupled PDEs (1.4), we obtain a numerical scheme in the general form:

\[
(S_j^+)_{m+1}^n = (S_j^+)_{m-1}^n + \Delta t \cdot f_j^+ \left( (S^+)_{m-1}^n \right) \\
(S_j^-)_{m+1}^n = (S_j^-)_{m+1}^n + \Delta t \cdot f_j^- \left( (S^+)_{m+1}^n \right),
\]

where \( f_j^+, j = 1, 2, 3 \), are the nonlinear functions defined in (1.5), i.e., the r.h.s. of (1.4), and \( \Delta t \) is the time step. Since in this chapter we assume periodic b.c., The explicit form of the periodic b.c. will be specified in Sec. 3.1.4.

In the numerical simulations, we will use (2.6) as the initial condition. Note that the evolution will occur for the perturbations \( \tilde{S}_j \). Thus, we will focus on such evolution. Applying the same linearization that led to the linearized PDE (2.7), and ignoring \( \tilde{S}_2 \) as being zero in this order of approximation, we found that scheme (3.1)
for the perturbations $\tilde{S}_{1,3}^\pm$ becomes:

$$
\begin{align*}
(S_1^+)^{n+1}_{m} &= (S_1^+)^n_{m-1} + b\Delta t(\tilde{S}_3^+)^n_{m-1} - 2\Delta t(\tilde{S}_3^-)^n_{m-1} \\
(S_3^+)^{n+1}_{m} &= (S_3^+)^n_{m-1} - b\Delta t(\tilde{S}_1^+)^n_{m-1} - \Delta t(\tilde{S}_1^-)^n_{m-1} \\
(S_1^-)^{n+1}_{m} &= (S_1^-)^n_{m+1} - 2b\Delta t(\tilde{S}_3^+)^n_{m+1} + \Delta t(\tilde{S}_3^-)^n_{m+1} \\
(S_3^-)^{n+1}_{m} &= (S_3^-)^n_{m+1} - b\Delta t(\tilde{S}_1^+)^n_{m+1} - \Delta t(\tilde{S}_1^-)^n_{m+1}.
\end{align*}
$$

According to Sec.6.3.5 in [11], the method of separation of variables can be applied to the partial differential equation. Thus, there are special solutions with wavenumber $k$ of the form $(\tilde{S}_j^\pm)^n_m = \rho_j^+(t_n)e^{ikx_m}$, where $j = 1, 3$. With this assumption, we have:

$$
\begin{align*}
(S_j^\pm)^n_{m-1} &= \rho_j^+(t_n)e^{ikx_m-\Delta x} = (S_j^\pm)^n_{m} \cdot e^{-ik\Delta x} \\
(S_j^\pm)^n_{m+1} &= \rho_j^+(t_n)e^{ikx_m+\Delta x} = (S_j^\pm)^n_{m} \cdot e^{ik\Delta x},
\end{align*}
$$

where $\Delta x$ is the increment of space variable $x$. Applying (3.3) to Eq. (3.2), we get

$$
\begin{align*}
(S_1^+)^{n+1}_{m} &= e^{-ik\Delta x}(S_1^+)^n_{m} + be^{-ik\Delta x}\Delta t(\tilde{S}_3^+)^n_{m} - 2e^{-ik\Delta x}\Delta t(\tilde{S}_3^-)^n_{m} \\
(S_3^+)^{n+1}_{m} &= e^{-ik\Delta x}(S_3^+)^n_{m} - be^{-ik\Delta x}\Delta t(\tilde{S}_1^+)^n_{m} - e^{-ik\Delta x}\Delta t(\tilde{S}_1^-)^n_{m} \\
(S_1^-)^{n+1}_{m} &= e^{ik\Delta x}(S_1^-)^n_{m} - 2be^{ik\Delta x}\Delta t(\tilde{S}_3^+)^n_{m} + e^{ik\Delta x}\Delta t(\tilde{S}_3^-)^n_{m} \\
(S_3^-)^{n+1}_{m} &= e^{ik\Delta x}(S_3^-)^n_{m} - be^{ik\Delta x}\Delta t(\tilde{S}_1^+)^n_{m} - e^{ik\Delta x}\Delta t(\tilde{S}_1^-)^n_{m}.
\end{align*}
$$

Notice that $\forall z \in \mathbb{R}$, $e^{iz}$ is periodic in terms of $z$ on the interval $z \in (-\pi, \pi]$. Thus, for $e^{ik\Delta x}$, the range of $k\Delta x$ is $(-\pi, \pi]$. Therefore, for fixed $\Delta x$, we may set a range of $k$: $\min(k) = 0$ and $\max(k) = \pi/\Delta x$. This way, we only need to study the value of $e^{\pm ik\Delta x}$ in (3.4) within this range since it repeats periodically.
According to (3.4), after each iteration, we see that the effect of SE scheme on the perturbation term $\tilde{S}$ can be written as matrix multiplication:

$$
(\tilde{S})^{n+1}_m = N (\tilde{S})^n_m ,
$$

(3.5)

where

$$
N = e^{-ik\Sigma \Delta x} (I + \Delta t P), \quad e^{-ik\Sigma \Delta x} \equiv \begin{pmatrix}
-ik\Delta x I & O \\
O & e^{ik\Delta x} I
\end{pmatrix} .
$$

(3.6)

Here, $\Sigma$, $I$, $P$ and $I$ are defined after (2.8). If we set $\Delta t = \Delta x = h$, then

$$
N = \begin{pmatrix}
e^{-ikh} I & O \\
O & e^{ikh} I
\end{pmatrix} + \begin{pmatrix}
e^{-ikh} hbA & e^{-ikh} hbB \\
e^{ikh} hbB & e^{ikh} hA
\end{pmatrix} .
$$

(3.7)

Here, matrices $I$, $A$, $B$ are defined in Sec. 2.2.

Let $\lambda_j$, $j = 1, 2, 3, 4$ be eigenvalues of the $4 \times 4$ matrix $N$ and $v_j$ be the corresponding eigenvectors. Expanding the perturbation $\tilde{S}$ over this set of eigenvectors (see Sec.6.3.6 in [11]), we have:

$$
(\tilde{S})^n_m = \sum_{j=1}^{4} c_j^{(0)} \lambda^n_j v_j \quad \Rightarrow \quad \left| (\tilde{S})^n_m \right| \leq \sum_{j=1}^{4} \left| c_j^{(0)} \right| \cdot |\lambda_j|^n \cdot |v_j| ,
$$

(3.8)

where $c_j^{(0)}$ are constant coefficients that depend on the initial perturbation $\left(\tilde{S}\right)^0_m$. Let $\lambda$ be the eigenvalue with the largest magnitude; without confusion, we will refer to it as the “largest eigenvalue". Then after many iterations $n$, $|\lambda|^n$ will eventually become significantly larger than the smaller $|\lambda_j|^n$. Thus, if the number of iterations is large
enough, the size of the perturbation $\tilde{S}$ will eventually be approximately proportional to $|\lambda|^n$. This is because in comparison with $|\lambda|^n$, the contribution of the other $|\lambda_j|^n$ can be ignored (since $|\lambda_j/\lambda|^n$ decays with $n$ exponentially when $|\lambda_j/\lambda| < 1$). Thus, the magnitude $|\lambda|$ determines the stability of the numerical scheme: if $|\lambda| \leq 1$, the magnitude of the initial perturbation does not grow, which means that the numerical scheme will be stable; if $|\lambda| > 1$, then perturbations will increase exponentially with iterations, and hence the numerical scheme will be unstable.

### 3.1.2 Numerical Calculation of $|\lambda|$

According to the discussion in Sec. 3.1.1, the numerical stability is decided by $|\lambda|$. Only when $|\lambda| \leq 1$, can we ensure that the magnitude of the initial perturbation does not grow and hence the numerical scheme will be stable. To find eigenvalues of $N$, we need to solve

$$|N - \lambda I| = 0 \quad (3.9)$$

for $\lambda$. To make such a solution feasible, we have to provide the value of $k$ inside (3.7).

If we let $z = kh$, then (3.7) becomes

$$N(z) = e^{-iz\Sigma} (I + hP)$$

$$= \begin{pmatrix} e^{-izI} & O \\ O & e^{izI} \end{pmatrix} + \begin{pmatrix} e^{-izhbA} & e^{-izhB} \\ e^{izhbB} & e^{izhA} \end{pmatrix}. \quad (3.10)$$

Notice that $e^{iz}$ is a periodic function. Therefore, if we fix $h$, then $N(z)$ is periodic in $z$. Hence, the eigenvalues of $N(z)$ must also be periodic in $z$. Since $N(z)$ is
periodic, we need only to consider $z \in (0, \pi)$ where we can solve (3.9) numerically. The result is shown in Fig. 3.1. Since $|\lambda(z)|$ is obviously continuous, then within each

![Graph of $|\lambda(z)|$](image)

**Figure 3.1:** The graph of $|\lambda(z)| - 1$, where $\lambda(z)$ is the “largest eigenvalue” of $N(z)$. We can see that $|\lambda|$ is always larger than 1, and $|\lambda| - 1$ is largest when $z = kh = 0$ or $\pi$. Therefore the numerically unstable modes that are observed in direct numerical simulations correspond to these values of $z$. Notice that the scale of each vertical axis is small, i.e., less than $10^{-2}$.

period of $z$, there must be a $z_{\text{max}}$ such that $|\lambda(z)|$ attains the maximum value. Since $|\lambda(z_{\text{max}})|$ determines the asymptotic behavior of the perturbation, the instability that we observe must be dominated by the mode with this eigenvalue. Thus, to investigate the numerical instability, if will suffice to consider the eigenvalue of $N(z)$ for $z = z_{\text{max}}$. According to Fig. 3.1, $z_{\text{max}} = 0$ and $\pi$, or we can say $k_{\text{max}} = 0$ and $\pi/h$. Therefore, in the following section, we will analytically derive $|\lambda(z_{\text{max}})|$ at these values of $z_{\text{max}}$. 
3.1.3 Analytical calculation of $|\lambda|$ 

For $z_{\text{max}} = 0$ and $\pi$, one can show form (3.7) that the characteristic equation reduces to:

$$|\mathbf{N}(z_{\text{max}}) - \lambda \mathbf{I}| = (\lambda + 1)^4 - h^2 G_1 (\lambda + 1)^2 = 0,$$  \hspace{1cm} (3.11)

where $G_1$ is defined in (2.31). Therefore, the eigenvalues are

$$\lambda = -1 \hspace{1cm} \text{(double root)}$$

$$\lambda = -1 \pm h \sqrt{G_1} = -1 \pm i h \sqrt{-G_1}.$$ \hspace{1cm} (3.12)

Recall that we set $b$ to be such that there is no physical instability, and then according to the result of Sec. 2.5, $G_1 \leq 0$. Therefore,

$$|\lambda| = \sqrt{1 - h^2 G_1} \approx 1 - \frac{1}{2} G_1 h^2 = 1 + \left(\frac{1}{2} h (b^2 - 4b + 1)\right) h.$$ \hspace{1cm} (3.13)

To estimate how repeated multiplication by $|\lambda|$ changes the size of the initial perturbation, we will write $|\lambda|$ in the following form:

$$|\lambda| = 1 + v h,$$ \hspace{1cm} (3.14)

where

$$v = \frac{(b^2 - 4b + 1) h}{2}.$$ \hspace{1cm} (3.15)
Notice that if the time is $t$, then the number of iterations $n = t/h$. Since $h \to 0$, we have

$$|\lambda|^n = (1 + vh)^n = (1 + vh)^{t/h} = \left[(1 + vh)^{\frac{1}{vh}}\right]^{vt} \to e^{vt}. \quad (3.16)$$

Therefore, in the numerical calculation, the initial perturbation will be approximately multiplied by $e^{vt}$, where $t$ is the time. We may call $v$ the error growth rate.

According to (3.14) and (3.15), when $b = -1$, $h = 0.01$, we have $|\lambda| - 1 = vh = 3 \times 10^{-4}$ and when $b = -1$, $h = 0.04$, we have $|\lambda| - 1 = vh = 4.8 \times 10^{-3}$. This analytical result agrees with the numerically obtained “largest eigenvalue” shown in Fig. 3.1.

### 3.1.4 Verification by Direct Numerical Simulations of Eq. (3.1)

According to the discussion in Sec. 3.1.3, the error growth rates obtained in Sec. 3.1.2 and Sec. 3.1.3 agree with each other, so we may refer to these two results as one. Let us recall that the error growth rate in Sec. 3.1.2 and Sec. 3.1.3 was found from the analysis of the eigenvalues of the matrix in (3.5), which is derived using von Neumann analysis. Since this result is not obtained by running direct numerical simulations using SE scheme (3.1) (with periodic b.c.), we may call the error growth rate obtained in Sec. 3.1.2 and Sec. 3.1.3 the “analytical error growth rate”, even though in Sec. 3.1.2 we used Matlab to solve (3.9). To check the validity of the analytical error growth rate that we obtained above, we ran the SE scheme (3.1) with periodic b.c. and used the initial condition (2.6). Similarly to (1.60), in order to apply periodic b.c. to the SE
scheme (3.1), we should set $2 \leq m \leq M - 1$ in (3.1) and use the following equations at the boundaries:

\[
\begin{align*}
(S^+_j)_1^{n+1} &= (S^+_j)_M^n + hf^+_j ((S^\pm)_M^n) \\
(S^-_j)_1^{n+1} &= (S^-_j)_2^n + hf^-_j ((S^\pm)_2^n) \\
(S^+_j)_M^{n+1} &= (S^+_j)_{M-1}^n + hf^+_j ((S^\pm)_{M-1}^n) \\
(S^-_j)_M^{n+1} &= (S^-_j)_1^n + hf^-_j ((S^\pm)_1^n) .
\end{align*}
\]

(3.17a) (3.17b)

We checked the validity of the numerical results in two steps. First, we examined the evolution of the numerical error by looking at the Fourier spectrum. For $b = -1, h = 0.04$, Fig. 3.2 shows this Fourier spectrum when $t = 200$. Notice that

Figure 3.2: Fourier spectrum of the perturbation $\tilde{S}_1^\pm$ obtained by the SE scheme with periodic b.c. of at $t = 200$. The initial perturbations $\tilde{S}_2^\pm$ stay almost unchanged in time and hence are not shown. In the legend of the graph, $(+)1$, $(−)1$, $(+)3$, $(−)3$ means $\tilde{S}_1^+, \tilde{S}_1^-, \tilde{S}_3^+, \tilde{S}_3^-$ respectively. The horizontal axis is the wave number $k$, which is defined in (3.3). The relation between $k$ and $z$ is $k = z/h$, which is shown in the statements near (3.10). Here, $z$ ranges from $−\pi$ to $\pi$, so $k$ ranges from $−\pi/h = −78.5$ to $\pi/h = −78.5$. This picture shows that the sizes of $\tilde{S}_1^\pm$ increase most significantly when the wavenumber is $0$ or $±\pi/h$.

Fig. 3.2 agrees with Fig. 3.1 in that they both exhibit the dominant error growth at
\( kh = 0, \pm \pi. \)

Second, we calculated the numerical error growth rate and compared it with the analytical error growth rate for \( b = -1 \) and \( h = 0.01, 0.04 \). According to the analysis of Sec. 2.2 and as confirmed by the direct numerics (see the caption to Fig. 3.2), the perturbations \( \tilde{S}_2^\pm \) remain almost unchanged as long as \( \tilde{S}_{1,3}^\pm \) remain small. Thus, we only collected the data on \( \tilde{S}_{1,3}^\pm \), and used these data over time to calculate the numerical error growth rate. Notice that if we fix \( x = x_o \) and let \( \omega = u + iv \), then (2.11) yields

\[
\tilde{S}(x_o, t) = v e^{-i\omega t + ikx_o} = v e^{-iut + ikx_o} e^{vt} \Rightarrow |\tilde{S}(x_o, t)| = |v| e^{vt}, \tag{3.18}
\]

where \( v \) is the error growth rate that we seek. Thus, for the calculation of \( v \), we filtered the perturbations near \( x = x_o = 0 \). To do this, we used a filter function:

\[
W(x) = e^{-5x^2} \tag{3.19}
\]

computed the filtered perturbations \( S_{f1,3}^\pm \) by

\[
S_{f1,3}^\pm = \tilde{S}_{1,3}^\pm W(x). \tag{3.20}
\]

We did not take only \( \tilde{S}_{1,3}^\pm(0, t) \) because such data could be too noisy. Let \( \left( S_{f1,3}^\pm \right)^n = S_{f1,3}^\pm(x, t_0) \), where \( x = (x_1, x_2, \cdots, x_M) \) is the vector of points on the spatial grid. Then the "smoothened" data that we used to calculate the error growth rate is ob-
tained by

\[(\epsilon)^n = \sqrt{\| (S_{f_1}^+)^n \|^2 + \| (S_{f_3}^-)^n \|^2 + \| (S_{f_1}^-)^n \|^2 + \| (S_{f_3}^+)^n \|^2}, \quad (3.21)\]

Here, \(\| \cdot \|\) means \(l^2\)-norm. According to (3.18), we have the approximate relation

\[(\epsilon)^n \sim e^{vt_n} \Rightarrow \ln ((\epsilon)^n) \sim vt_n. \quad (3.22)\]

Thus, the relation between \(\ln(\epsilon)\) and \(t\) is linear. To calculate \(v\) using the data \(\epsilon\) over time, we can take two time points \(T_1, T_2\) and observe the corresponding errors \(\epsilon_1, \epsilon_2\). Then the error growth rate is calculated as

\[v = \frac{\ln(\epsilon_1) - \ln(\epsilon_2)}{T_2 - T_1}. \quad (3.23)\]

An alternative method is linear fitting. However, no matter which approach we choose, we should notice that it might take some time for the error growth to become steady in time. That is, at the beginning the of the iterations, the error can be too noisy, which would make the pattern of (3.22) not very clear. That pattern becomes cleaner when the iterations have proceeded for a while. This phenomenon is illustrated by Fig. 3.3, where a steady, linear trend of the graph occurs when \(\epsilon\) reaches the size of \(e^{-14} \approx 10^{-6}\). Thus, there is a threshold of error size starting from which the error growth becomes steady. We may choose this threshold by looking at the error growth rate plot as in the first panel of Fig. 3.3. Then we compare the so obtained growth rate with the analytical growth rate. According to (3.14) and (3.16), if the numerically calculated error growth rate \(v\) agrees with analytical result, then
Figure 3.3: Trend of \( \log(\epsilon) \) vs \( t \). The error size growth from \( 10^{-15} \) to \( 10^{-3} \) for \( b = -1, \ h = 0.04 \). The left panel is the complete error growth. Here, the trend seems somewhat concave up, while in later iterations, the trend is very close to a straight line. The right panel is the error growth when we only take the error data that are larger than \( e^{-14} \approx 10^{-6} \), i.e, \( \log(\epsilon) > -14 \). Here, there is very significant linear trend, and we can use this trend to calculate the numerical error growth rate.

Table 3.1: Growth rates with different increment \( h \).

| \( h \) | \( v \) | \( vh \) | \( |\lambda(z_{\text{max}})| - 1 \) |
|---|---|---|---|
| 0.01 | 0.0302 | \( 3.0200 \times 10^{-4} \) | \( 2.9996 \times 10^{-4} \) |
| 0.04 | 0.1209 | \( 0.0048 \) | \( 4.7885 \times 10^{-3} \) |

we must have \( |\lambda(z_{\text{max}})| - 1 = vh \), Where \( |\lambda(z_{\text{max}})| \) is obtained using the result shown in Fig. 3.1. According to Tab. 3.1, \( |\lambda(z_{\text{max}})| - 1 \approx vh \) is indeed true. Thus, the error growth rate obtained in numerical simulations agrees with the analytical result.

3.2 **Modified Euler Method**

We will now follow the same process as in Sec. 3.1 for the MoC-based ME method.
3.2.1 Matrix form of the ME

According to (1.57), which is the numerical ME based on the MoC, the corresponding scheme for (1.4) is as follows:

\[
\begin{align*}
\left( S_j^+ \right)_m &= (S_j^+)^n_{m-1} + \Delta t \cdot f_j^+ \left( \left( S_j^\pm \right)^n_{m-1} \right) \\
\left( S_j^- \right)_m &= (S_j^-)^n_{m+1} + \Delta t \cdot f_j^- \left( \left( S_j^\pm \right)^n_{m+1} \right) \\
(S_j^+)^{n+1}_m &= \frac{1}{2} \left[ (S_j^+)^n_{m-1} + (S_j^-)^n_{m-1} + \Delta t \cdot f_j^+ \left( \left( S_j^\pm \right)^n_{m-1} \right) \right] \\
(S_j^-)^{n+1}_m &= \frac{1}{2} \left[ (S_j^-)^n_{m+1} + (S_j^-)^n_{m+1} + \Delta t \cdot f_j^- \left( \left( S_j^\pm \right)^n_{m+1} \right) \right].
\end{align*}
\]

(3.24a)

where \( f_j^\pm, j = 1, 2, 3 \), are the nonlinear functions defined in (1.5). Here, we also set \( \Delta x = \Delta t = h \). When deriving the matrix form of the linearization of this scheme, we will also use (3.3). Notice that (3.24a) has the structure of the SE scheme (3.1). Therefore, (3.24a) can be written in the following matrix form:

\[
\left( \mathbf{S} \right)_m = \mathbf{N} \left( \mathbf{S} \right)^n_m,
\]

(3.25)

where \( \mathbf{N} \) is defined in (3.6). Then, similarly to the derivation of (3.2), we find that (3.24b) becomes

\[
\begin{align*}
(S_1^+)^{n+1}_m &= \frac{1}{2} \left[ (S_1^+)^n_{m-1} + (S_1^-)^n_{m-1} + bh \left( S_3^+ \right)_m - 2h \left( S_3^- \right)_m \right] \\
(S_3^+)^{n+1}_m &= \frac{1}{2} \left[ (S_3^+)^n_{m-1} + (S_3^-)^n_{m-1} - bh \left( S_1^+ \right)_m - h \left( S_1^- \right)_m \right] \\
(S_1^-)^{n+1}_m &= \frac{1}{2} \left[ (S_1^-)^n_{m+1} + (S_1^-)^n_{m+1} - 2bh \left( S_3^+ \right)_m + h \left( S_3^- \right)_m \right] \\
(S_3^-)^{n+1}_m &= \frac{1}{2} \left[ (S_3^-)^n_{m+1} + (S_3^-)^n_{m+1} - bh \left( S_1^+ \right)_m - h \left( S_1^- \right)_m \right].
\end{align*}
\]

(3.26)
We can write (3.26) in the following matrix form:

\[
(\tilde{S})_{m}^{n+1} = \frac{1}{2} N_2 (\tilde{S})_m^n + \frac{1}{2} N_3 (\tilde{S})_m^n,
\]

(3.27)

where

\[
N_2 = e^{-ik\Sigma\Delta x}, \quad N_3 = I + hP,
\]

(3.28)

and \(\Sigma\), \(I\) and \(P\) are defined after (2.8). Combining (3.25) and (3.27), we have

\[
(\tilde{S})_{m}^{n+1} = \frac{1}{2} [N_2 + N_3 N] (\tilde{S})_m^n.
\]

(3.29)

Let

\[
N_4 = \frac{1}{2} [N_2 + N_3 N];
\]

(3.30)

then after each iteration, the effect of the ME scheme on the perturbation \(\tilde{S}\) can be written as the following matrix multiplication:

\[
(\tilde{S})_{m}^{n+1} = N_4 (\tilde{S})_m^n.
\]

(3.31)

Similarly to the argument at the end of Sec. 3.1.1, in order to analyze stability of the ME scheme, we need to study the "largest eigenvalue" of \(N_4\), which we will also denote by \(\lambda\).
3.2.2 Numerical calculation of $|\lambda|$

According to the discussion in Sec. 3.1.1, the numerical stability is decided by $|\lambda|$. Only when $|\lambda| \leq 1$, can we ensure that the magnitude of the initial perturbation does not grow; and hence the numerical scheme will be stable. To find eigenvalues of $N_4$, we need to solve

$$|N_4 - \lambda I| = 0$$

(3.32)

for $\lambda$. To make such a solution feasible, we have to provide the value of $k$ inside $N_4$. Similarly to Sec. 3.1.2, we will consider setting $z = kh$ and fixing $h$. Then $N_4(z)$ is periodic in $z$. For the same reason given in Sec. 3.1.2, we need to solve (3.32) for $z \in (-\pi, \pi)$, which we will do in this section numerically. As in Sec. 3.1.2, to investigate the numerical instability, it will suffice to consider $|\lambda(z_{max})|$. Performing numerical calculations similar to those reported in Sec. 3.1.2, we obtain results shown in Fig. 3.1. We found that $z_{max} = 0, \pi/2, 3\pi/2$, equivalently, $k_{max} = \pi/(2h), 3\pi/(2h)$.

3.2.3 Numerical result given by ME code

With the same motivation as in Sec. 3.1.4, to check the validity of the result we obtained from the analysis above, we ran the ME scheme (3.24) with periodic b.c., and the initial condition (2.6). Similarly to (1.63), in order to apply periodic b.c. to the ME scheme (3.24), we should set $2 \leq m \leq M - 1$ in (3.24) and use the following
Figure 3.4: The graph of $|\lambda(z)| - 1$, where $\lambda(z)$ is the “largest eigenvalue” of $N(z)_4$. We can see that $|\lambda|$ is always larger than 1, and $|\lambda| - 1$ is largest when $z = kh = \pm \pi/2$ since $\lambda(z)$ is periodic. Therefore the numerical instability that is observed in direct numerical simulation occurs for $z$ taking such values.

at the boundaries:

$$\begin{align*}
\left(\overline{S}_j^+\right)_1 &= (S_j^+)_M^n + hf_j^+ (S_j^\pm)_M^n \\
\left(\overline{S}_j^-\right)_1 &= (S_j^-)_2^n + hf_j^- (S_j^\pm)_2^n \\
(S_j^+_1)^{n+1} &= \frac{1}{2} \left[ (S_j^+_1)_1^n + \left(\overline{S}_j^+\right)_1^n + hf_j^+ \left(\overline{S}_j^\mp\right)_1^n \right] \\
(S_j^-_1)^{n+1} &= \frac{1}{2} \left[ (S_j^-_1)_2^n + \left(\overline{S}_j^-\right)_1^n + hf_j^- \left(\overline{S}_j^\mp\right)_1^n \right];
\end{align*}$$

(3.33a)

$$\begin{align*}
\left(\overline{S}_j^+\right)_M &= (S_j^+)_M^n + hf_j^+ (S_j^\pm)_M^n \\
\left(\overline{S}_j^-\right)_M &= (S_j^-)_1^n + hf_j^- (S_j^\pm)_1^n \\
(S_j^+_M)^{n+1} &= \frac{1}{2} \left[ (S_j^+_M)_1^n + \left(\overline{S}_j^+\right)_M^n + hf_j^+ \left(\overline{S}_j^\mp\right)_M^n \right] \\
(S_j^-_M)^{n+1} &= \frac{1}{2} \left[ (S_j^-_M)_1^n + \left(\overline{S}_j^-\right)_M^n + hf_j^- \left(\overline{S}_j^\mp\right)_M^n \right].
\end{align*}$$

(3.33b)

We followed the same process as in Sec. 3.1.4: First, we examined the evolution
of the numerical error by looking at the Fourier Spectrum. For $b = -1, h = 0.04$, Fig. 3.5 shows this Fourier Spectrum graph when $t = 610s$. Notice that Fig. 3.5 agrees with Fig. 3.4 in that they both exhibit the dominant error growth at $kh = \pm \pi/2$.

With the same approach as in Sec. 3.1.4, we can calculated the numerical error growth rate and compared it with the corresponding analytical result in Fig. 3.4. The result is shown in Tab. 3.2. According to Tab. 3.2, $|\lambda(z_{max})| - 1 \approx vh$ is indeed true. Thus, the error growth rate obtained in numerical simulations agrees with the analytical result.

Table 3.2: Growth rates with different increment $h$.  

| $h$  | $v$  | $vh$       | $|\lambda(z_{max})| - 1$  |
|------|------|------------|--------------------------|
| 0.01 | 0.0088 | $8.8000 \times 10^{-5}$ | $10.0008 \times 10^{-5}$  |
| 0.04 | 0.0388 | 0.0016     | $16.02 \times 10^{-4}$   |
3.3 Leapfrog Method

We will now follow the same process as in Sec. 3.1 for the MoC-based LF method.

3.3.1 Matrix Form of the MoC-based LF Method

According to (1.58), which is the numerical LF based on the MoC, the corresponding scheme for (1.4) is as follows:

\[
(S_j^+)_{m+1} = (S_j^+)_{m-2} + 2\Delta t \cdot f_j^+ \left( (S^\pm)_{m-1}^n \right),
\]

\[
(S_j^-)_{m+1} = (S_j^-)_{m+2} + 2\Delta t \cdot f_j^- \left( (S^\pm)_{m+1}^n \right),
\]

(3.34)

where \(f_j^\pm, j = 1, 2, 3\), are the nonlinear functions defined in (1.5). Here, we also set \(\Delta x = \Delta t = h\). When deriving the matrix form of the linearization of this scheme, we will also use (3.3). Therefore, (3.24a) can be written in the following matrix form:

\[
S^{n+1} = P_1 S^n + P_0 S^{n-1}
\]

(3.35)

where \(S^n = [(\tilde{S}_1)^n_m \ldots (\tilde{S}_6)^n_m]^T\), etc. and \(P_1, P_0\) are matrices given as follows:

\[
P_1 = 2hQP, \quad P_0 = Q^2, \quad Q = e^{-ik\Sigma\Delta x},
\]

(3.36)

and \(\Sigma, I\) and \(P\) are defined after (2.8). One can say that here, matrix \(Q\) serves to provide a phase distortion.
3.3.2 Numerical Calculation of $|\lambda|$

This subsection follows the outline of Sec. 3.1.4 and Sec. 3.2.3. We solve the difference equation (3.35) via the standard substitution $S^{n+1} = \lambda S^n$, which leads to the following matrix characteristic polynomial:

$$\left(\lambda^2 I - 2hQP\lambda - Q^2\right) S = 0. \quad (3.37)$$

We set $kh = z$ and use Matlab to solve for $\lambda$ numerically, similarly to how we did it for $N$ and $N_4$ in the previous sections. The result is shown in Fig. 3.6 (where the “analytical” $\lambda$ is the $\lambda$ computed as described above; we will denote its value at $z = \pi/2$ by $\lambda_{an}$).

According to Fig. 3.6, our analytical calculations show that the instability is the strongest for $kh = \pi/2$, where $|\lambda|-1 = 1.5 \times 10^{-3}$. Note that this instability has a very
narrow spectral support. This result is confirmed by direct numerical simulations of the PDEs (for simplicity, we call the $\lambda$ that is obtained by numerical simulations the “numerical” $\lambda$; we will denote its value at $z = \pi/2$ by $\lambda_{num}$). The Fourier spectrum of the numerical error is shown in Fig. 3.7.

![Figure 3.7: Fourier spectrum of the numerical error for the MoC-LF with periodic b.c. at $t = 10$. Here, we take $b = -1, h = 0.01$. The legend is the same as that of Fig. 3.5. The instability is the strongest at wave number $kh = \pi/2$](image)

Since the “analytical” $\lambda$ is shown in Fig. 3.6 and the “numerical” $\lambda$ is shown in Fig. 3.7, we can visually confirm, the qualitative similarity between the two $\lambda$’s. However, in order to quantitatively verify the “analytical” $\lambda$, we need to compare the exact value of the “analytical” $|\lambda|$ with that of the “numerical” $|\lambda|$.

In the following calculations, for both $|\lambda_{an}|$ and $|\lambda_{num}|$, we take $b = -1$ and $h = 0.01$. By solving (3.37) for $\lambda$, we can find that $|\lambda_{an}|$ is:

$$|\lambda_{an}| \approx 1 + 1.50 \times 10^{-3}. \quad (3.38)$$
Following the same process as in Sec. 3.1.4 and Sec. 3.2.3, we find that the numerical $|\lambda|$ of the MoC-LF with periodic b.c. is:

$$|\lambda_{num}| \approx 1 + 1.49 \times 10^{-3}.$$  \hfill (3.39)

Comparing (3.38) and (3.39), we see that the “numerical” $|\lambda|$ agrees with the “analytical” $|\lambda|$.

3.3.3 Conclusion for the von Neumann instability of LF

From the results of this section, we can conclude that for the MoC-LF with periodic b.c., both von Neumann analysis and numerical simulations show that (see Fig. 3.6, Fig. 3.7), the strong numerical instability only exists near wave number $kh = \pi/2$. Thus, for the MoC-LF with periodic b.c., the Fourier mode with wave number $kh = \pi/2$ is the “culprit” mode that causes the instability.

Based on this observation, we give the following explanation for the occurrence of the instability in the MoC-LF with periodic b.c. The characteristic equation, i.e., counterpart of (3.37), for the LF scheme used to solve ODEs (as opposed to being used in combination with the MoC) is:

$$\left(\lambda^2 I - 2h\lambda P - I\right) S = 0,$$  \hfill (3.40)

where $P$ is the same matrix as in (3.37), i.e., the evolution matrix of the linearized physical problem. Comparing (3.37) and (3.40), one can notice that the major dif-
ference between these two is that the matrix $Q$ in (3.37) is not the identity matrix $I$. This is because in the MoC-LF scheme, the characteristic lines are slanted in the $(x, t)$–plane. Comparing (3.37) and (3.40), one can say that in (3.37), $Q$ serves as a “phase distortion” matrix. According to (3.40), $\lambda$ of the ODE-LF scheme is determined by the “physical matrix” $P$. In particular, if the system is physically stable, then the ODE-LF scheme is also stable (for sufficiently small $h$). On the other hand, in (3.37), $\lambda$ is not determined by $P$ alone due to the presence of the “phase distortion” matrix $Q$. Thus, even if our PDE system is physically stable, the MoC-LF scheme may be unstable.

Moreover, Eq. (3.36), shows that $Q$ is most dissimilar from the identity matrix when $kh = \pi/2$, whereas for $kh = 0$ or $\pi$, $Q \propto I$, in which case the MoC-LF is reduced to the ODE-LF. In this “ODE limit”, the MoC-LF scheme is no different than the LF scheme used to solve ODEs. Since LF is stable for conservative ODEs, there is naturally no instability for modes with wave number near 0, $\pi$ in the MoC-LF scheme. However, since $Q$ deviates from the identity matrix the most when the $kh = \pi/2$, one can expect the strongest instability in modes with the corresponding wave numbers. This is precisely what Fig. 3.6 and Fig. 3.7 demonstrate.
Chapter 4

Stability analysis for Simple Euler scheme with non-reflecting boundary condition

4.1 The impact of non-reflecting b.c. on the numerical stability

When we use periodic b.c., we found that the numerical result agrees with von Neumann analysis for both SE scheme, ME scheme and LF scheme. However, when we changed the periodic b.c. to non-reflecting b.c., which we defined in Sec. 1.3, we found that the numerical result for ME scheme does not agree with the von Neumann analysis. More importantly, the non-reflecting b.c. rendered the ME scheme stable. Therefore, in the following chapter, we will consider using a different method to analysis the stability. Since the SE scheme is similar to the ME scheme but simpler, we
will first use this method on the SE in this chapter so that we can establish a good understanding of this method and its validity. Then in the next chapter we will apply it to the ME scheme.

4.2 Set up of the method

Similarly to Sec. 3.1, we use the notations that are defined in Sec. 2.2, such as \( \tilde{S} = (\tilde{S}_1^+, \tilde{S}_3^+, \tilde{S}_1^-, \tilde{S}_3^-)^T \). The notations that we employ below were introduced in [1]. Denote:

\[
y^+ = \begin{pmatrix} \tilde{S}_1^+ \\ \tilde{S}_3^+ \end{pmatrix}, \quad y^- = \begin{pmatrix} \tilde{S}_1^- \\ \tilde{S}_3^- \end{pmatrix}, \quad y = \begin{pmatrix} y^+ \\ y^- \end{pmatrix}, \quad (y)^n_m = \begin{pmatrix} y^+ \\ y^- \end{pmatrix}_m^n, \quad (4.1)
\]

where we use \( \left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_m^n \) as a simplification for \(( (y^+)^n_m, (y^-)^n_m )^T \), we will also use similar simplification in the rest of this section. Using (4.1), we can write the SE scheme for the perturbations \( \tilde{S}_{1,3}^\pm \), which is defined in (3.2), as:

\[
(y^+)^n_m = I (y^+)^n_{m-1} + bhA (y^+)^n_{m-1} + hB (y^-)^n_{m-1}, \\
(y^-)^n_m = I (y^-)^n_{m+1} + bhB (y^+)^n_{m+1} + hA (y^-)^n_{m+1}, \quad (4.2)
\]
where \( I, A, B \) are defined after (2.8). Combining the two equations above, we have

\[
(y)_{m+1}^n = 
\begin{pmatrix}
  y^+ \\
  y^-
\end{pmatrix}^m +
\begin{pmatrix}
  (I + bhA) (y^+)^n_{m-1} + hB (y^-)^n_{m-1} \\
  (I + hA) (y^-)^n_{m+1} + bhB (y^+)^n_{m+1}
\end{pmatrix} = \Gamma \begin{pmatrix}
  y^+ \\
  y^-
\end{pmatrix}_m^n + \Omega \begin{pmatrix}
  y^+ \\
  y^-
\end{pmatrix}_m^n = \Gamma (y)_{m-1}^n + \Omega (y)_{m+1}^n, \quad (4.3)
\]

where

\[
\Gamma = 
\begin{pmatrix}
  I + bhA & hB \\
  O & O
\end{pmatrix}, \quad \Omega = 
\begin{pmatrix}
  O & O \\
  bhB & I + hA
\end{pmatrix}. \quad (4.4)
\]

Consider the space line to be infinite for a moment and define an infinite vector:

\[
(Y)^n = 
\begin{pmatrix}
  \vdots \\
  (y)_m^n \\
  \vdots
\end{pmatrix}; \quad (4.5)
\]
then

\[
(Y)^{n+1} = \begin{pmatrix}
\cdots & O & \Omega & O & \cdots & \cdots & O \\
\Gamma & O & \Omega & O & \cdots & \cdots & \cdots \\
O & \Gamma & O & \Omega & \cdots & \cdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\cdots & \cdots & \cdots & \Gamma & O & \Omega & \cdots \\
O & \cdots & \cdots & O & \Gamma & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots 
\end{pmatrix}
(Y)^n. \quad (4.6)
\]

In real numerical calculations, the space line is finite, so the dimension of \( Y \) is \( 4M \times 1 \), where \( M \) is the number of grid points, and

\[
(Y)^n = \begin{pmatrix}
(y)^n_1 \\
\vdots \\
(y)^n_M
\end{pmatrix}, \quad (4.7)
\]

where each \((y)^n_m\) has dimension \( 4 \times 1 \). To explore how the finite length of \((Y)^n\) affects the matrix in the iteration scheme (4.6), let us first consider periodic b.c.. In this case, we identify \( y \) at the fictitious grid point \( m = 0 \) with \((y)_M^n\):

\[
(y)^n_0 \equiv (y)^n_M. \quad (4.8)
\]
Then from (4.3),

\[(y)^{n+1}_1 = \Gamma (y)^n_0 + \Omega (y)^n_2 \equiv \Gamma (y)^n_M + \Omega (y)^n_2. \quad (4.9)\]

Similarly,

\[(y)^{n+1}_M = \Gamma (y)^n_{M-1} + \Omega (y)^n_1. \quad (4.10)\]

Then

\[
(Y)^{n+1} = \left( \begin{array}{cccccccc}
O & \Omega & O & \cdots & \cdots & \Gamma \\
\Gamma & O & \Omega & \cdots & \cdots & \cdots \\
O & \Gamma & O & \cdots & \cdots & \cdots \\
& \vdots & \vdots & \vdots & \vdots & \vdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\Omega & \cdots & \cdots & \cdots & \cdots & \cdots
\end{array} \right) (Y)^n
\]

\[= \Psi (Y)^n. \quad (4.11)\]

If we consider the non-reflecting b.c., which was introduced in Sec. 1.3 and applied in (1.62), then we know the value of \(y^+_1\) and \(y^-_M\): We may let

\[y^+_1 = F, \quad y^-_M = G. \quad (4.12)\]

However, note that we do not have information about \(y^-_1\) and \(y^+_M\). We will state this
fact as:

\[
(y)^{n+1}_1 = \left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_1 = \left( \begin{array}{c} F \\ \ast \end{array} \right), \tag{4.13}
\]

where “\ast” is the unknown value, determined by the evolution from the time level \( n \).

Similarly,

\[
(y)^{n+1}_M = \left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_M = \left( \begin{array}{c} \ast^* \\ G \end{array} \right), \tag{4.14}
\]

where “\ast^*” is the unknown value, determined by the evolution from the time level \( n \).

According to Eqs.(4.3) and Eqs.(4.13), we have

\[
(y)^{n+1}_1 = \left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_1^{n+1} = \left( \begin{array}{c} F \\ bhB (y^+)_2^n + (I + hA) (y^-)_2^n \end{array} \right) = \Omega (y)_2^n + \left( \begin{array}{c} F \\ \vec{0} \end{array} \right), \tag{4.15}
\]

where \( \vec{0} = (0, 0)^T \). Similarly,

\[
(y)^{n+1}_M = \left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_M^{n+1} = \left( \begin{array}{c} (I + bhA) (y^+)_M^{n-1} + hB (y^-)_M^{n-1} \\ G \end{array} \right) = \Gamma (y)_M^{n-1} + \left( \begin{array}{c} \vec{0} \\ G \end{array} \right). \tag{4.16}
\]
Combining Eqs. (4.6), (4.15), (4.16), we obtain

\[
(Y)^{n+1} = \begin{pmatrix}
O & \Omega & O & O & \cdots & \cdots & O \\
\Gamma & O & \Omega & O & \cdots & \cdots & \cdots \\
O & \Gamma & O & \Omega & \cdots & \cdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \\
\cdots & \cdots & \cdots & \Gamma & O & \Omega & \cdots & \cdots \\
O & \cdots & \cdots & \cdots & O & \Gamma & O \\
\end{pmatrix}
\begin{pmatrix}
F \\
\tilde{\epsilon} \\
\ddots \\
\ddots \\
\ddots \\
\ddots \\
\end{pmatrix}
\] 

\[
= \Phi (Y)^n + V
\] 

(4.17)

### 4.3 Agreement of New Method for Periodic B.C. with Von Neumann Analysis

In this section, we will describe how to use (4.11) to study the stability of the SE scheme with periodic b.c.. Then we will show that the results obtained by this method agree with those of the von Neumann analysis in Sec. 3.1, where we also consider periodic b.c..

Similarly to the discussion following (3.8) in Sec. 3.1.1, we know that the magnitude of \((Y)^n\) is controlled by the eigenvalues of the matrix \(\Psi\) in (4.11). That is,

\[
|(Y)^n| \leq \sum_{j=1}^{4M} |c_j^{(0)}| |\lambda_j|^n |v_j|,
\] 

(4.18)

where \(c_j^{(0)}\) are constant coefficients that depend on \((Y)^0\); \(\lambda_j, \quad j = 1, \ldots, 4M\) are the \(4M\) eigenvalues of the matrix \(\Psi\) and \(v_j\) are the corresponding eigenvectors. The
largest $|\lambda_j|$ determines the stability. Thus, we should first compute the eigenvalues of $\Psi$. Suppose $\lambda$ is an eigenvalue of $\Psi$ and $Y$ is the corresponding eigenvector, then we have

$$\Psi Y = \lambda Y,$$

(4.19)

where $Y = [(y)_1, \ldots, (y)_M]^T$, which is similar to the definition (4.7). Note that here we do not indicate the time level, $n$, because (4.19) along with the analysis that lead to (3.8) implies

$$(y)_m^n = \lambda^n (y)_m^0,$$

(4.20)

where $(y)_m^0$ is the initial value of $(y)$. Substituting the explicit form of matrix $\Psi$ into (4.19), we have

$$\Gamma (y)_M + \Omega (y)_2 = \lambda (y)_1$$

(4.21a)

$$\Gamma (y)_{m-1} + \Omega (y)_{m+1} = \lambda (y)_m, \quad m = 2, \ldots, M - 1$$

(4.21b)

$$\Gamma (y)_{M-1} + \Omega (y)_1 = \lambda (y)_M.$$  

(4.21c)

Since the difference equations (4.21) are linear with constant coefficients, we seek the solution of (4.21) in the form:

$$(y)_{m+1} = \rho (y)_m \quad \Rightarrow \quad (y)_m = \rho^{m-1} (y)_1,$$

(4.22)

where $\rho$ is the spatial increment in the difference equation (4.21). A similar case
was discussed in Sec. 3.1. of [14]. There, the matrix is not a block matrix and the
eigenvector is not a block vector, hence, \( \rho \) is simply a scalar. However, since \( \Psi \) is a
block matrix and \( Y \) is a block vector, \( \rho \) should be initially considered to be a \( 4 \times 4 \) matrix block.

Substituting (4.22) into (4.21b), we have

\[
\Gamma \rho^{m-2}(y)_1 + \Omega \rho^m(y)_1 = \lambda \rho^{m-1}(y)_1, \quad m = 2, \ldots, M - 1
\]

\[\Rightarrow \quad D \left( \rho^{m-2}(y)_1 \right) = 0, \quad \text{where} \quad D = \Gamma - \lambda \rho + \Omega \rho^2. \quad (4.23)\]

Let us now show that we can replace the multiplication with a matrix \( \rho \) in (4.22) by
the scalar multiplication:

\[
\rho (y)_m = \rho (y)_m \quad \Rightarrow \quad (y)_m = \rho^{m-1}(y)_1 \quad (4.24)
\]

where \( \rho \) is a scalar. Indeed, if (4.24) does not hold true, then (4.23) implies that

\[
D (y)_{m-1} = 0, \quad m = 1, \ldots, M - 1, \quad (4.25)
\]

which means \( (y)_1, (y)_2, \ldots, (y)_{M-1} \) are all the zero eigenvalues of the matrix \( D \).
However, this cannot be true since \( D \) is \( 4 \times 4 \) matrix and in practice, \( M - 1 \) can be
significantly larger than 4. It is not possible for a \( 4 \times 4 \) matrix to have \( M - 1 \) zero
eigenvalues. This contradiction would not happen if (4.24) holds true since we would
be able to factor out the scalar number $\rho^{m-2}$:

\[
D(y)_{m-1} = 0, \quad m = 1, \ldots, M - 1,
\]

\[
\Rightarrow D(\rho^{m-2}(y)_1) = D(\rho^{m-2}(y)_{1}) = \rho^{m-2}D(y)_{1} = 0
\]

\[
\Rightarrow D(y)_{1} = 0,
\]

which means not all $(y)_1, (y)_2, \ldots, (y)_{M-1}$ are zero eigenvectors. Here, $(y)_1$ is the only one that must be a zero eigenvector of $D$. Since we cannot factor out the matrix $\rho^{m-2}$ in $D(\rho^{m-2}(y)_1)$, (4.24) is the only situation such that $D$ does not have $M - 1$ zero eigenvalues. Thus, (4.24) must be true.

When we apply (4.24) to the b.c. (4.8), we have

\[
\rho^{-1}(y)_1 = \rho^{M-1}(y)_1 \quad \Rightarrow \quad \rho^{M} = 1.
\]

Thus, $\rho^{M} = 1$, which means if we apply periodic b.c. then $\rho$ is one of the distinct complex $M$th roots of unity. That is

\[
\rho^{M} = 1 = e^{i2\pi s}, \quad \Rightarrow \quad \rho_s = e^{2\pi i \left(\frac{s}{M}\right)}, \quad s = 0, 1, \ldots, M - 1.
\]

According to the discussion above, we must $D(y)_1 = 0$, which implies that

\[
D(y)_1 = \Gamma(y)_1 - \lambda \rho(y)_1 + \Omega \rho^{2}(y)_1
\]

\[
= \Gamma(y)_1 - \lambda \rho(y)_1 + \Omega \rho^{2}(y)_1
\]

\[
= \left(\Gamma - \lambda \rho I + \Omega \rho^{2}\right)(y)_1 = 0.
\]
This is equivalent to

\[ (\Gamma \rho^{-1} + \Omega \rho - \lambda I) \mathbf{y} = \mathbf{0}. \]  

(4.30)

Therefore, in order to find \( \lambda \), we need to solve the following equation

\[ \begin{vmatrix} \Gamma \rho^{-1} + \Omega \rho - \lambda I \end{vmatrix} = 0 \]  

(4.31)

for \( \lambda \). Here, we are actually looking for the eigenvalue of the matrix \( \Gamma \rho^{-1} + \Omega \rho \). When we substitute the value of \( \rho \) in (4.28), which is the value \( \rho \) of periodic b.c., into this matrix, we have

\[ \Gamma \rho^{-1} + \Omega \rho = \Gamma e^{-2\pi i \left( \frac{s}{M} \right)} + \Omega e^{2\pi i \left( \frac{s}{M} \right)} \]  

(4.32)

If we let \( z = 2\pi \left( \frac{s}{M} \right) \), then \( e^{2\pi i \left( \frac{s}{M} \right)} = e^{iz} \), \( e^{-2\pi i \left( \frac{s}{M} \right)} = e^{-iz} \), and hence

\[ \Gamma \rho^{-1} + \Omega \rho = \begin{pmatrix} e^{-iz} I & O \\ O & e^{iz} I \end{pmatrix} + \begin{pmatrix} e^{-iz} hB \hbar A & e^{-iz} hB \\ e^{iz} hB \hbar A & e^{iz} hA \end{pmatrix}, \]  

(4.33)

which is exactly the matrix \( \mathbf{N} \) in (3.10). Then solving (4.31) is equivalent to solving (3.9) in Sec. 3.1.2 and thus the results of our new approach will agree with the the von Neumann analysis in Sec. 3.1.
4.4 Applying the equation to non-reflecting b.c.

In this section, we will use (4.17) to investigate the stability of the SE scheme with non-reflecting b.c.. Notice that the stability only depends on the matrix $\Phi$ in (4.17). The vector $V$ in (4.17) does not affect the stability. Thus we set $V$ to be zero vector. According to the same reasoning as in Sec. 4.3, in order to investigate the stability of this scheme, we need to compute the eigenvalues of $\Phi$. Suppose $\lambda$ is the eigenvalue of $\Phi$ and $Y$ is the corresponding eigenvector, then we have (Here I used Note TMA4205 of The eigenvalues of tridiagonal matrices Autumn 2009 from Norwegian University of Science and Technology, which is on your website, how do I reference it?)

$$\Phi Y = \lambda Y,$$  \hspace{1cm} (4.34)

where $Y$ is defined as the same one in Sec. 4.3. Similarly to (4.21), we have:

$$\Gamma (y)_{m-1} + \Omega (y)_{m+1} = \lambda (y)_m, \ m = 1, \ldots, M \hspace{1cm} (4.35a)$$

$$(y)_0 = (y)_{M+1} = 0, \hspace{1cm} (4.35b)$$

where $(y)_0$, $(y)_{M+1}$ are the artificial entry of $Y$ that are use for the purpose of computation. According to the same reasoning in in Sec. 4.3, we have the following relation that is similar to (4.24):

$$(y)_m = \rho^m (y)_0, \hspace{1cm} (4.36)$$
and the following equation:

\[ \left( \Gamma - \lambda \rho I + \Omega \rho^2 \right) (y)_0 = 0. \]  

(4.37)

Thus, in order to get \( \lambda \), we need to solve

\[ \left| \Omega \rho^2 - \lambda \rho I + \Gamma \right| = 0. \]  

(4.38)

For convenience, we will denote \( \left| \Omega \rho^2 - \lambda \rho I + \Gamma \right| \) as \( E(\rho, \lambda) \). Notice that here, we do not have the condition \( \rho^M = 1 \) as in Sec. 4.3 since we are not considering the periodic b.c.. The value of \( \rho \) in (4.38) is unknown. Thus, in order to compute \( \lambda \), we need to first solve (4.38) to get the relation between \( \rho \) and \( \lambda \), then combine this relation with the non-reflecting b.c. (4.13),(4.14). The complete process is discussed as below.

First we need to get the relation between \( \rho \) and \( \lambda \) (i.e. \( \rho = \rho(\lambda) \)) by solving following polynomial equation:

\[ E(\rho, \lambda) = 0. \]  

(4.39)

Since there are four nontrivial solutions, or nonzero solutions, of \( E(\rho, \lambda) = 0 \) in terms of \( \rho \) (see Sec. 4.5.1): \( \rho_k(\lambda), \ k = 1, 2, 3, 4 \), the solution of the difference equation (4.35) should be

\[ (y)_m = \sum_{k=1}^{4} C_k \rho_k^n(\lambda) \xi_k, \]  

(4.40)

where \( C_k \) are constant coefficients; \( \xi_k \) are the four possible values of \( (y)_0 \) in (4.36).
The reason we do not use component such as $C_k m \rho^m_k(\lambda) \xi_k$ in (4.40) is that in Sec. 4.5.1 we found that $\rho_k(\lambda)$ are distinct. Similarly to the notations of (4.1), we also have

$$\xi = \begin{pmatrix} \xi^+ \\ \xi^- \end{pmatrix},$$  \hspace{1cm} (4.41)$$

where $\xi^+$, $\xi^-$ are $2 \times 1$ vectors. Substituting $\rho = \rho_k(\lambda)$ and $(y)_0 = \xi_k$ into (4.37), we can get the following equation:

$$\left( \Gamma - \lambda \rho_k(\lambda) I + \Omega \rho^2_k(\lambda) \right) \xi_k = 0.$$  \hspace{1cm} (4.42)

By solving (4.42), we can get the values of $\xi_k$ (see Sec. 4.5.2). Thus, the values of $\xi_k$ actually depend on $\lambda$ and (4.40) should be written as

$$(y)_m = \sum_{k=1}^{4} C_k \rho^m_k(\lambda) \xi_k(\lambda).$$  \hspace{1cm} (4.43)$$

Finally, we can compute $\lambda$ by combining (4.43) and (4.13),(4.14), which gives

$$\begin{align*}
(y^+)_0 &= \sum_{k=1}^{4} C_k \xi^+_k(\lambda) = F = \vec{0} \\
(y^-)_M &= \sum_{k=1}^{4} C_k \rho^M_k \xi^-_k(\lambda) = G = \vec{0}.
\end{align*}$$  \hspace{1cm} (4.44)$$

We can write (4.44) in the matrix form:

$$\Upsilon(\lambda) \begin{bmatrix} C_1, \ C_2, \ C_3, \ C_4 \end{bmatrix}^T = \begin{bmatrix} \vec{0}, \vec{0} \end{bmatrix}^T,$$  \hspace{1cm} (4.45)$$
where

\[
\Upsilon(\lambda) = \begin{pmatrix}
\xi_1^+(\lambda) & \xi_2^+(\lambda) & \xi_3^+(\lambda) & \xi_4^+(\lambda) \\
\rho_1^M(\lambda)\xi_1^-(\lambda) & \rho_2^M(\lambda)\xi_2^-(\lambda) & \rho_3^M(\lambda)\xi_3^-(\lambda) & \rho_4^M(\lambda)\xi_4^-(\lambda)
\end{pmatrix}
\]  

(4.46)

According to (4.45), we must have

\[|\Upsilon(\lambda)| = 0\]  

(4.47)

Solving (4.47), we have the value of \(\lambda\) (see Sec. 4.5.3). In conclusion, the process discussed above consists of three main steps: finding the four \(\rho_4(\lambda)\) by solving (4.39); obtaining the \(\xi_4(\lambda)\), which we will refer to as “eigenvector of \(\rho_4(\lambda)\)”, by using \(\rho_4(\lambda)\) to solve (4.42); substituting \(\rho_4(\lambda)\) and \(\xi_4(\lambda)\) into (4.47) and solve (4.46) for \(\lambda\). These three steps are discussed in details and implemented in Sec. 4.5.1, 4.5.2, 4.5.3.

### 4.5 Computing the eigenvalue of SE scheme with non-reflecting B.C.

This section is dedicated to explicitly computing the eigenvalue of matrix \(\Phi\) in (4.17) by using following the three steps discussed in Sec. 4.4.
4.5.1 Finding the approximate solution of (4.39)

In this section we will perform the perturbation analysis to find the asymptotic solution of (4.39) so that we can get the four \( \rho_k(\lambda) \).

In order to go through the process discussed in the last part of Sec. 4.4, we need to first get the relation \( \rho(\lambda) \) by solving (4.39). However, similarly to the situation of solving (2.14), the exact solution of (4.39) can be quite difficult to obtain since \( E(\rho, \lambda) \) is proportional to a fourth-order polynomial. As a matter of fact, even if we obtain the exact solution of equation (4.39), the explicit form of \( \rho(\lambda) \) would be so complicated that the subsequent process in Sec. 4.5.2, 4.5.3 would be nearly impossible to accomplish. However, notice that

\[
E(\rho, \lambda) = \left| \begin{pmatrix} (1 - \lambda \rho) I & O \\ O & (\rho^2 - \lambda \rho) I \end{pmatrix} + h \begin{pmatrix} bA & B \\ \rho^2 b B & \rho^2 A \end{pmatrix} \right|, \quad (4.48)
\]

which yields a simple diagonal matrix structure when \( h = 0 \). Thus, the solution of \( E(\rho, \lambda)|_{h=0} = 0 \) is very easy to compute, and the structure of the solution would be simple. Since \( 0 < h \ll 1 \), we can circumvent the difficulty of solving (4.39) exactly by using perturbation analysis to obtain the asymptotic solution of (4.39). With the appropriate truncation of small terms, the asymptotic (in the limit \( h \to 0 \)) solution of (4.38) will be accurate enough for the method discussed in Sec. 4.5.2, 4.5.3 to evaluate \( \lambda \).
We now perform the first step of the perturbation analysis. When $h = 0$, we have
\[ E(\rho, \lambda)|_{h=0} = \begin{vmatrix} (1 - \lambda \rho)I & O \\ O & (\rho^2 - \lambda \rho)I \end{vmatrix} \]
\[ = (1 - \lambda \rho)^2(\rho^2 - \lambda \rho)^2 = \rho^2(\lambda \rho - 1)^2(\rho - \lambda)^2 \]
\[ = \rho^2 \lambda^2(\rho - \lambda)^2 \left( \rho - \frac{1}{\lambda} \right)^2. \tag{4.49} \]

Thus there are two nonzero roots for $E(\rho, \lambda)|_{h=0} = 0$:
\[ \rho_1 = \frac{1}{\lambda} \text{ and } \rho_2 = \lambda. \tag{4.50} \]

In order to find the asymptotic solution of (4.39), we should perform perturbation analysis near these two roots. However, it is possible that $\rho_1 \approx \rho_2$. In this case, we should not perform perturbation analysis near two distinct roots (4.50). Instead, we need to resort to a different approach to be described later. To understand when we will have to use it, note that:
\[ \rho_1 \approx \rho_2 \implies \frac{1}{\lambda} \approx \lambda \implies \lambda^2 - 1 \approx 0. \tag{4.51} \]

Thus, we can use the last condition in (4.51) to separate the two cases. That is, when $\lambda^2 - 1 \not\approx 0$, which will be referred to as “Case 1”, we will perform perturbation analysis near the two distinct roots (4.50) to calculate an asymptotic solution of (4.39). On the other hand, when $\lambda^2 - 1 \approx 0$, which will be referred to as “Case 2”, we will use another perturbation approach to calculate the asymptotic solution.
From (4.48), $E(\rho, \lambda)$ has the following form (which we verified with Mathematica):

$$E(\rho, \lambda) = \rho^2 \left[ E_0(\rho, \lambda) + h^2 E_1(\rho, \lambda) \right], \quad (4.52)$$

where

$$E_0(\rho, \lambda) = \lambda^2 (\rho - \lambda)^2 (\rho - \frac{1}{\lambda})^2 \quad (4.53)$$

and

$$E_1(\rho, \lambda) = \lambda^2 \rho^4 + (2b-1)2\lambda \rho^3$$

$$+ \left( (1 - 4b + b^2) - 4b\lambda^2 \right) \rho^2 + (2 - b)2b\lambda \rho + b^2 \lambda^2. \quad (4.54)$$

Since $E = \rho^2 (E_0 + h^2 E_1) = 0$ is equivalent to

$$E_0 + h^2 E_1 = 0 \quad (4.55)$$

when we do not consider the solution $\rho = 0$, we will refer to Eq. (4.55) instead of (4.52) for convenience.

**Case 1:** $\lambda^2 - 1 \approx 0$.

If we let $\rho = \rho_1 + \tilde{\rho}_1 = \frac{1}{\lambda} + \tilde{\rho}_1$, where $|\tilde{\rho}_1| \ll 1$ and substitute it into (4.55), then we

95
have

\[ E_0 \left( \frac{1}{\lambda} + \tilde{\rho}_1, \lambda \right) + h^2 E_1 \left( \frac{1}{\lambda} + \tilde{\rho}_1, \lambda \right) \\
= (1 - \lambda^2)^2 \tilde{\rho}_1^2 + + O(\tilde{\rho}_1^3) \\
+ b^2 \left( \frac{1}{\lambda^2} + \lambda^2 - 2 \right) h^2 + O(h^2 \tilde{\rho}_1) + O(h^3) = 0, \tag{4.56} \]

where the obvious smaller terms, \( O(\tilde{\rho}_1^3), O(h^2 \tilde{\rho}_1) \), are ignored. Now, comparing the size of \( \tilde{\rho}_1 \) and \( h \) in (4.56), we find that \( O(\tilde{\rho}_1^2) \sim O(h^2) \), and

\[ \tilde{\rho}_1 = \pm \frac{ibh}{\lambda} \quad \Rightarrow \quad \rho = \frac{1}{\lambda} (1 \pm ibh), \tag{4.57} \]

where we have used the identity \((1/\lambda^2 + \lambda^2 - 2) = (1/\lambda - \lambda)^2\). For future use, let us denote

\[ \rho_{1(+)} = \frac{1}{\lambda} \left( 1 + ibh + O(h^2) \right), \quad \rho_{1(-)} = \frac{1}{\lambda} \left( 1 - ibh + O(h^2) \right), \tag{4.58} \]

where the \( O(h^2) \) terms will be presented soon. Similarly, if we let let \( \rho = \rho_2 + \tilde{\rho}_2 = \lambda + \tilde{\rho}_2, \ |\tilde{\rho}_2| \ll 1 \), and substitute it into (4.55), we find that

\[ \tilde{\rho}_2 = \pm i\lambda h \quad \Rightarrow \quad \rho = \lambda(1 \pm ih). \tag{4.59} \]

Similarly to (4.58), we denote

\[ \rho_{2(+)} = \lambda \left( 1 + ih + O(h^2) \right), \quad \rho_{2(-)} = \lambda \left( 1 - ih + O(h^2) \right). \tag{4.60} \]

In Sec. 4.5.4 it will become apparent that we will need the explicit expression of
the $O(h^2)$-terms in (4.58) and (4.60). Their calculation in Appendix A, and the result is:

$$\rho_1(\pm) = \frac{1}{\lambda} \left( 1 \pm ibh + \frac{2b}{\lambda^2 - 1} h^2 \right)$$
$$\rho_2(\pm) = \lambda \left( 1 \pm ih - \frac{\lambda^2 + 2b - 1}{\lambda^2 - 1} h^2 \right).$$  \hspace{1cm} (4.61)

For convenience, we will use the following notations:

$$\rho_1(\pm) = \frac{1}{\lambda} \hat{\rho}_1(\pm), \quad \rho_2(\pm) = \lambda \hat{\rho}_2(\pm).$$  \hspace{1cm} (4.62)

The four $\rho(\lambda)$ in (4.61) are the ones that we need in order to proceed with the calculation in Secs. 4.5.2, 4.5.3.

**Case 2:** $\lambda^2 - 1 \approx 0$.

In this case, we have $\lambda \approx 1$ or $\lambda \approx -1$. The results of these two subcases are very similar, so here we only present the discussion of the subcase $\lambda \approx 1$. The same instability for $\lambda \approx -1$ can be demonstrated in a similar fashion. Assume that $\lambda = 1 + \tilde{\alpha}$, $|\tilde{\alpha}| \ll 1$. Hence from (4.50), $\rho_1 = 1/\lambda \approx 1 - \tilde{\alpha}$ and $\rho_2 = \lambda \approx 1 + \tilde{\alpha}$. Let us denote

$$\Delta \rho = \rho_1 - \rho_2 = \frac{1}{\lambda} - \lambda = (1 - \tilde{\alpha} + \cdots) - (1 + \tilde{\alpha}) = -2\tilde{\alpha} + \cdots.$$  \hspace{1cm} (4.63)

Let us seek solutions of (4.55) in the form

$$\rho = \left( \frac{\rho_1 + \rho_2}{2} \right) + \tilde{\beta} = \frac{1}{2} (1 - \tilde{\alpha} + \cdots + (1 + \tilde{\alpha})) + \tilde{\beta} = 1 + \tilde{\beta},$$  \hspace{1cm} (4.64)
where $|\tilde{\beta}| \ll 1$ is to be later related to $\tilde{\alpha}$. Then, omitting higher-order terms in $\tilde{\alpha}$:

$$
(\rho - \rho_1) = ((1 + \tilde{\beta}) - (1 - \tilde{\alpha})) = \tilde{\beta} + \tilde{\alpha}
$$

$$
(\rho - \rho_2) = ((1 + \tilde{\beta}) - (1 + \tilde{\alpha})) = \tilde{\beta} - \tilde{\alpha}.
$$

(4.65)

Substituting (4.65) into $E_0(\rho, \lambda)$ and $h^2 E_1(\rho, \lambda)$, we have:

$$
E_0(1 + \tilde{\beta}, 1 + \tilde{\alpha}) = (1 + \tilde{\alpha})^2(\tilde{\beta} + \tilde{\alpha})^2(\tilde{\beta} - \tilde{\alpha})^2
$$

$$
= (1 + 2\tilde{\alpha} + \cdots)(\tilde{\beta}^2 - \tilde{\alpha}^2)^2 + \cdots
$$

$$
= \tilde{\beta}^4 + \tilde{\alpha}^4 - 2\tilde{\alpha}^2\tilde{\beta}^2 + \cdots
$$

(4.66)

and

$$
h^2 E_1(\rho)(1 + \tilde{\beta}, 1 + \tilde{\alpha}) = (1 - 4b + b^2)\tilde{\alpha}^2h^2 + (1 + 4b + b^2)\tilde{\beta}^2h^2
$$

$$
+ (2 - 2b^2)h^2\tilde{\alpha}\tilde{\beta} + \cdots
$$

(4.67)

Therefore, substituting (4.66) and (4.67) into (4.55), we have

$$
E_0 + h^2 E_1 = \left(\tilde{\beta}^4 + \tilde{\alpha}^4 - 2\tilde{\alpha}^2\tilde{\beta}^2\right) + (1 - 4b + b^2)\tilde{\alpha}^2h^2 + (1 + 4b + b^2)\tilde{\beta}^2h^2
$$

$$
+ (2 - 2b^2)h^2\tilde{\alpha}\tilde{\beta} + \cdots = 0.
$$

(4.68)

Comparing the magnitude of $\tilde{\alpha}$, $\tilde{\beta}$ with the that of $h$ from the condition that at least two of the terms in (4.68) must have commensurate size, we find that at least one of
the following must hold:

\[ O(\tilde{\alpha}^4) \sim O(\tilde{\alpha}^2 h^2), \quad O(\tilde{\beta}^4) \sim O(\tilde{\beta}^2 h^2), \quad O(\tilde{\alpha}^2 \tilde{\beta}^2) \sim O(\tilde{\alpha} \tilde{\beta} h^2). \]  

(4.69)

In fact, all of these relations will hold if we let

\[ \tilde{\alpha} = \alpha h, \quad \tilde{\beta} = \beta h. \]  

(4.70)

Substituting (4.70) into (4.68) yields

\[ (\alpha^2 - \beta^2)^2 + (\alpha + \beta)^2 + b^2(\alpha - \beta)^2 - 4b(\alpha^2 - \beta^2) = 0. \]  

(4.71)

From this equation one can, in principle, determine a relation between \( \alpha \) and \( \beta \) and hence between \( \lambda \) and \( \rho \). Then, following the approach outlined for case 1 in Sec. 4.5.2 and Sec. 4.5.3, one could eventually find \( \lambda \). However, the corresponding calculations are very complicated. Therefore, we use another approach, which gives a good approximation of the instability growth rate \(|\lambda| - 1\) in this case (case 2).

According to the discussion above, we have

\[ \lambda = 1 + \alpha h, \quad \rho = 1 + \beta h; \]  

(4.72)

hence \( \rho \approx 1 \). Recall that we obtain the relation \( \lambda(\rho) \) by solving the following equation

\[ E(\rho, \lambda) = \rho^2 \left[ E_0(\rho, \lambda) + h^2 E_1(\rho, \lambda) \right] = 0. \]  

(4.73)
By continuity, we have

$$\lim_{\rho \to 1} E(\rho, \lambda) = E(1, \lambda), \quad (4.74)$$

which implies that \(\lambda(\rho \approx 1)\) must be close to \(\lambda(\rho = 1)\). Thus, we can approximate the instability growth rate by computing \(\lambda(\rho = 1)\) from \(E(1, \lambda) = 0\). Using (4.52)-(4.54), one can show that the last equation yields:

$$E(1, \lambda) = (\lambda - 1)^2 \left( (\lambda - 1)^2 + h^2(1 - 4b + b^2) \right) = 0. \quad (4.75)$$

Since we consider a physically stable PDE system, we must have \(b < 0\) (see Sec. 2.6.3), hence \((1 - 4b + b^2) > 0\). Thus, the solutions for (4.75) are

$$\lambda_{1,2} = 1, \quad \lambda_{3,4} = 1 \pm ih\sqrt{1 - 4b + b^2}. \quad (4.76)$$

The instability is associated with \(\lambda_{3,4}\), because:

$$|\lambda_{3,4}| = \sqrt{1 + h^2(1 - 4b + b^2)} \approx 1 + \frac{1}{2} h^2(1 - 4b + b^2) > 1. \quad (4.77)$$

Notice that (4.77) agrees with the von Neumann result (3.13), which, in turn, is very close to the numerical instability growth rate that we observed by applying the MoC-SE with non-reflecting BC to the PDE system.

Let us summarize the results of this subsection. We have solved the equation \(E(\rho, \lambda) = 0\), where \(E\) is given by (4.52). Our solution so far did not account for the information of the non-reflecting BC. For \(\lambda^2 - 1 \approx 0\) (case 1), solving \(E(\rho, \lambda) = 0\) gave us a relation \(\rho(\lambda)\). In order to get the value of \(\lambda\), we need to proceed with
the calculations in Secs. 4.5.2, 4.5.3, which will incorporate the information of the non-reflecting BC. As a matter of fact, in Sec. 4.5.3, we will show that modes with \( \lambda^2 - 1 \approx 0 \) are all stable, thus do not explain the observed numerical instability. For \( \lambda^2 - 1 \approx 0 \) (case 2), since \( \rho \approx 1 \), we can approximate the value of \( \lambda \) by solving for \( E(1, \lambda) = 0 \). Although we did not verify that this mode satisfy the non-reflecting BC, we still got a very good approximation for the observed numerical instability growth rate. Thus, the numerical instability for MoC-SE with non-reflecting BC is already well-explained by modes with \( \lambda^2 - 1 \approx 0 \). Therefore, in Secs. 4.5.2, 4.5.3, we will not proceed with the calculation for \( \lambda^2 - 1 \approx 0 \), because they are very complicated but will not add anything new to our understanding of the instability’s mechanism. On the other hand, even though we will not find any instability for modes with \( \lambda^2 - 1 \approx 0 \), we will proceed with calculations for them. We will do so because it is only by the detailed calculations for these modes in Secs. 4.5.2, 4.5.3 that we can show that indeed there is no instability for for modes with \( \lambda^2 - 1 \approx 0 \). What’s more, following the outline of these calculations will show in Chap. 5 why instability for the MoC-ME with non-reflecting BC is suppressed for modes with \( \lambda^2 - 1 \approx 0 \), hence explaining why MoC-ME with non-reflecting BC is stable.

4.5.2 Finding the approximate solution of (4.42)

According to the discussion in Sec. 4.4, the next step after finding \( \rho_k(\lambda) \), done in Sec. 4.5.1, is to use (4.42) to find the four eigenvectors \( \xi_k(\lambda) \). To that end, we substitute the obtained values of \( \rho_k(\lambda) \) into (4.42) and solve it for the corresponding
ξ_k(λ). As per the discussion in Sec. 4.5.1, we will only consider the case \( \lambda^2 - 1 \gtrsim 0 \).

In this subsection, we will only need to use the expressions

\[
\rho_{1(\pm)} = \frac{1}{\lambda} (1 \pm i b h), \quad \rho_{2(\pm)} = \lambda (1 \pm i h)
\]  

(4.78)

from (4.58), (4.60), i.e., the approximations of \( \rho(\lambda) \) up to \( O(h) \). This will be clear from our calculations. The \( O(h^2) \)-terms derived in Appendix A will be needed in Sec. 4.5.3. For the approximations (4.78), we will compute the \( \xi(\lambda) \) corresponding to \( \rho_{1(\pm)} \) and \( \rho_{2(\pm)} \) separately. For convenience of notation, we will denote \( \rho_0 = \rho_1 \) or \( \rho_2 \), depending on which root of \( E(\rho, \lambda)|_{h=0} \) is the solution of (4.49). We also denote

\[
s = \pm 1, \quad i b s = \tilde{R}
\]

(4.79)

when \( \rho_0 = \rho_1 = 1/\lambda \), and

\[
s = \pm 1, \quad i s = \tilde{R}
\]

(4.80)

when \( \rho_0 = \rho_2 = \lambda \) to provide a uniform notation for both roots in (4.58) and (4.60).

In order to avoid any possible confusion between the cases of these two pairs of roots, on one hand, and Cases 1 and 2 considered in Sec. 4.5.1, on the other hand, we will label the cases of \( \rho_0 = \rho_1 \) and \( \rho_0 = \rho_2 \) with Roman (as opposed to arabic) numbers.

**Case I**: Eigenvectors for \( \rho_{1(\pm)}(\lambda) = \frac{1}{\lambda} (1 \pm i b h). \)

Using the above notations, we have \( \rho_0 \equiv \rho_1 = \frac{1}{\lambda} \) and

\[
\rho = \rho_0(1 + i b s h) = \rho_0 + \rho_0 h \tilde{R}, \quad \frac{1}{\rho} \approx \frac{1}{\rho_0} - \frac{1}{\rho_0} h \tilde{R}.
\]

(4.81)
Now let us notice that (4.42) is equivalent to

\[
\left( \Omega \rho + \Gamma \rho^{-1} - \lambda I \right) \xi = 0, \tag{4.82}
\]

which we will use in the following calculations. We seek \( \xi \) to have the following structure:

\[
\xi = \begin{pmatrix} \vec{u} \\ h^\gamma \vec{v} \end{pmatrix}, \tag{4.83}
\]

where \( \vec{u} = [u_1, u_2]^T \), \( \vec{v} = [v_1, v_2]^T \) are \( 2 \times 1 \) vectors with components of order \( O(1) \).

Also, a subsequent calculation will reveal that parameter \( \gamma > 0 \), i.e., \( h^\gamma \vec{v} \) is small compared to \( \vec{u} \). The reason we assume this structure will become clear as we obtain \( \vec{u} \) and \( \vec{v} \) via a perturbation analysis for \( h \ll 1 \). To that end, we write matrices \( \Omega \) and \( \Gamma \) in (4.82) as perturbations of matrices \( \Omega_0 \) and \( \Gamma_0 \) for \( h = 0 \):

\[
\Omega = \Omega_0 + h \Omega_1, \quad \Gamma = \Gamma_0 + h \Gamma_1, \tag{4.84}
\]

where

\[
\Omega_0 = \begin{pmatrix} O & O \\ O & I \end{pmatrix}, \quad \Omega_1 = \begin{pmatrix} O & O \\ bB & A \end{pmatrix},
\]

\[
\Gamma_0 = \begin{pmatrix} I & O \\ O & O \end{pmatrix}, \quad \Gamma_1 = \begin{pmatrix} bA & B \\ O & O \end{pmatrix}, \tag{4.85}
\]

and \( 2 \times 2 \) matrices \( A \) and \( B \) were defined after (2.8). Substituting (4.81), (4.83), and
We will now perturbatively solve (4.86) to get the explicit form of $\xi(\lambda)$. To that end, we first note that by setting $\gamma = 1$, one can obtain nontrivial solutions of (4.86). The validity of having this value, $\gamma = 1$, will be verified later when we find that the numerical results, presented in Sec. 4.5.4, agree with the analytical results found here. Thus, using $\gamma = 1$ and ignoring $O(h^2)$ terms in (4.86), we have

$$\left[\left(\Omega_0 + h\Omega_1 \right)(\rho_0 + \rho_0 h \tilde{R}) + (\Gamma_0 + h\Gamma_1) \left(\frac{1}{\rho_0} - \frac{1}{\rho_0} h \tilde{R}\right) - \lambda I\right] \begin{pmatrix} \bar{u} \\ h\gamma \bar{v} \end{pmatrix} = 0. \quad (4.86)$$

Collecting the terms of (4.87) according to different orders of $h$ and again ignoring terms $O(h^2)$, we have:

$O(1)$:

$$\left(\Omega_0 \rho_0 + \Gamma_0 \frac{1}{\rho_0} - \lambda I\right) \begin{pmatrix} \bar{u} \\ 0 \end{pmatrix} = 0 \quad (4.88)$$
\[ O(h): \]
\[
\left( \Omega_0 \rho_0 + \Gamma_0 \frac{1}{\rho_0} - \lambda I \right) \begin{pmatrix} \vec{v} \\ \vec{0} \end{pmatrix} + \left( \Omega_0 \rho_0 \tilde{R} + \Omega_1 \rho_0 - \Gamma_0 \frac{1}{\rho_0} \tilde{R} + \Gamma_1 \frac{1}{\rho_0} \right) \begin{pmatrix} \vec{u} \\ \vec{0} \end{pmatrix} = 0 \tag{4.89}
\]

Notice that
\[
\left( \Omega_0 \rho_0 + \Gamma_0 \frac{1}{\rho_0} - \lambda I \right) = \begin{pmatrix} (\frac{1}{\rho_0} - \lambda) I \end{pmatrix} \begin{pmatrix} O \\ O \end{pmatrix} (\rho_0 - \lambda) I \tag{4.90}
\]
since \( \rho_0 = \frac{1}{\lambda} \). Thus, (4.88) yields
\[
\left( \Omega_0 \rho_0 + \Gamma_0 \frac{1}{\rho_0} - \lambda I \right) \begin{pmatrix} \vec{u} \\ \vec{0} \end{pmatrix} = \begin{pmatrix} O \\ O \end{pmatrix} \begin{pmatrix} \vec{u} \\ \vec{0} \end{pmatrix} = 0, \tag{4.91}
\]
which holds true no matter what \( \vec{u} \) we take, and so (4.88) gives no information about the eigenvector that we are seeking. Thus, we need to resort to (4.89) to calculate the eigenvector. Eq. (4.89) yields:
\[
\begin{pmatrix} O & O \\ O & (\rho_0 - \lambda) I \end{pmatrix} \begin{pmatrix} \vec{u} \\ \vec{0} \end{pmatrix} + \begin{pmatrix} \frac{b}{\rho_0} A - \tilde{R} B \\ \frac{1}{\rho_0} B \end{pmatrix} \begin{pmatrix} \vec{u} \\ \vec{0} \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{0} \end{pmatrix} \tag{4.92}
\]
The first row of $2 \times 2$ matrix blocks in (4.92) yields:

$$\begin{pmatrix} b \rho_0 A - \tilde{R} I \rho_0 \end{pmatrix} \vec{u} = \vec{0} \quad \Rightarrow \quad \lambda (bA - \tilde{R}I)\vec{u} = \vec{0}$$

$$\Rightarrow \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix} - ibs \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} -is & 1 \\ -1 & -is \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \quad (4.93)$$

recall that $s$ and $\tilde{R}$ are defined in (4.80). Notice that the determinant $\left|\begin{array}{cc} -is & 1 \\ -1 & -is \end{array}\right| = -s^2 + 1 = -1 + 1 = 0$, so (4.93) has a nontrivial solution:

$$-isu_1 + u_2 = 0 \quad \Rightarrow \quad u_2 = isu_1 \quad \Rightarrow \quad \vec{u} = \begin{pmatrix} 1 \\ is \end{pmatrix}. \quad (4.94)$$

Similarly, the second row of $2 \times 2$ matrix blocks in (4.92) yields:

$$\left(\frac{1}{\lambda} - \lambda\right) \vec{v} + b \frac{1}{\lambda} B\vec{u} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad (4.95)$$

which gives the solution for $\vec{v}$:

$$\vec{v} = \frac{b}{\lambda^2 - 1} B\vec{u}. \quad (4.96)$$
Recall that, in (4.94), the value of $s$ depends on which $\rho(\lambda)$ we choose:

$$\rho_{1(+)} = \frac{1}{\lambda}(1 + ibh) \quad \Rightarrow \quad s = 1,$$

$$\rho_{1(-)} = \frac{1}{\lambda}(1 - ibh) \quad \Rightarrow \quad s = -1,$$

(4.97)

where $\rho_{1(\pm)}$ are defined in (4.58).

**Case II:** Eigenvectors for $\rho(\lambda) = \rho_{2(\pm)}(\lambda) = \lambda (1 \pm ih)$.

Note that the counterpart of relations (4.81) in this case is:

$$\rho = \rho_0(1 + ish) = \rho_0 + \rho_0 h \tilde{R}, \quad \frac{1}{\rho} \approx \frac{1}{\rho_0} - \frac{1}{\rho_0} h \tilde{R},$$

(4.98)

where $\rho_0 = \lambda$. Similarly to (4.83), we seek $\xi$ in the following form:

$$\xi = \begin{pmatrix} h\bar{v} \\ \bar{u} \end{pmatrix}. \quad (4.99)$$

Following the steps made to obtain (4.94) and (4.96), we have

$$\bar{u} = \begin{pmatrix} 1 \\ -is \end{pmatrix} \quad (4.100)$$

and

$$\bar{v} = \frac{1}{\lambda^2 - 1} B\bar{u}. \quad (4.101)$$
Similar to (5.57), in (4.101) the value of \(s\) depends on which \(\rho(\lambda)\) we choose:

\[
\rho_{2(+)} = \lambda(1 + i\hbar) \quad \Rightarrow \quad s = 1,
\]
\[
\rho_{2(-)} = \lambda(1 - i\hbar) \quad \Rightarrow \quad s = -1,
\]

(4.102)

where \(\rho_{2(\pm)}\) are defined in (4.60).

### 4.5.3 Finding the approximate solution of (4.47)

We are now at the last step of our program, set out in Sec. 4.4. Namely, we will use (4.46) to compute the eigenvalue \(\lambda\). As we explained at the end of Sec. 4.5.1, we will consider only the case \(\lambda^2 - 1 \approx 0\).

Using (4.40) and the results of Secs. 4.5.1, 4.5.2, we have

\[
(y)_m = C_1 \rho_{1(+)}^m \xi_{1(+)} + C_2 \rho_{1(-)}^m \xi_{1(-)} + C_3 \rho_{2(+)}^m \xi_{2(+)} + C_4 \rho_{2(-)}^m \xi_{2(-)},
\]

(4.103)

where \(\rho_{1(\pm)}, \rho_{2(\pm)}\) are the four \(\rho_k\) in (4.40) and \(\xi_{1(\pm)}, \xi_{1(-)}, \xi_{2(+)}, \xi_{2(-)}\) are the four \(\xi_k\) in (4.40), and the constants \(C_1, C_2, C_3, C_4\) are to be determined below the BC. Here, the formulas of \(\rho_{1(\pm)}\) and \(\rho_{2(\pm)}\) are given in (4.58) and (4.60); the formulas for \(\xi_{1(\pm)}\) are given by (4.83), (4.94) and (4.96); the formulas for \(\xi_{2(\pm)}\) are given by (4.99), (4.100) and (4.101). For the reader’s convenience, we write their explicit expressions
In the $2 \times 1$-block form, these eigenvectors can be written as

\[
\xi_{1(+)} = \begin{pmatrix}
1 \\
i
\end{pmatrix}, \quad \xi_{1(-)} = \begin{pmatrix}
i \\
-1
\end{pmatrix}, \quad \xi_{2(+)} = \begin{pmatrix}
\frac{b}{h} \\
\frac{1}{\lambda - 1}
\end{pmatrix}, \quad \xi_{2(-)} = \begin{pmatrix}
\frac{1}{\lambda - 1} \\
i
\end{pmatrix}, \quad (4.104a)
\]

In the $2 \times 1$-block form, these eigenvectors can be written as

\[
\xi_{i(+)} = \begin{pmatrix}
\bar{\xi}_{i(+)}^+ \\
\bar{\xi}_{i(+)}^-
\end{pmatrix}, \quad \xi_{i(-)} = \begin{pmatrix}
\bar{\xi}_{i(-)}^+ \\
\bar{\xi}_{i(-)}^-
\end{pmatrix}, \quad i = 1, 2 \quad (4.104b)
\]

Note, in particular, that the superscripts $\pm$ are in no way related to the subscripts $(\pm)$. The boundary conditions (4.44) are as follows:

\[
C_1\bar{\xi}_{1(+)}^+ + C_2\bar{\xi}_{1(-)}^+ + C_3\bar{\xi}_{2(+)}^+ + C_4\bar{\xi}_{2(-)}^+ = 0; \quad (4.105a)
\]

\[
C_1\bar{\xi}_{1(+)}^- + C_2\bar{\xi}_{1(-)}^- + C_3\bar{\xi}_{2(+)}^- + C_4\bar{\xi}_{2(-)}^- = 0. \quad (4.105b)
\]

Together, (4.105a) and (4.105b) give the following linear system:

\[
\Upsilon(\lambda) [C_1, C_2, C_3, C_4]^T = 0, \quad (4.106)
\]
where

\[
\Upsilon(\lambda) = \begin{pmatrix}
\begin{pmatrix}
1 & 1 \\
\lambda & -\lambda
\end{pmatrix},
\begin{pmatrix}
\frac{h}{\lambda^2 - 1} & 2i \\
-1 & -1
\end{pmatrix}
\end{pmatrix},
\begin{pmatrix}
-2i\hat{\rho}^M_{1(+)} & 2i\hat{\rho}^M_{1(-)} \\
-\hat{\rho}^M_{1(+)} & -\hat{\rho}^M_{1(-)}
\end{pmatrix},
\begin{pmatrix}
\hat{\rho}^M_{2(+)} & \hat{\rho}^M_{2(-)} \\
-i\hat{\rho}^M_{2(+)} & i\hat{\rho}^M_{2(-)}
\end{pmatrix}
\end{pmatrix},
\]

(4.107)

where, according to (4.61) and (4.62), we have

\[
\hat{\rho}_{1(\pm)} = 1 \pm ibh + \frac{2b}{\lambda^2 - 1} h^2, \quad \hat{\rho}_{2(\pm)} = 1 \pm ih - \frac{(\lambda^2 + 2b - 1)}{\lambda^2 - 1} h^2.
\]

(4.108)

Here, we use the approximations of \(\rho_{1,2(\pm)}\) accurate up to \(O(h^2)\), i.e., (4.61), in order to achieve quantitative agreement with numerical simulations, which will be described in Sec. 4.5.4. Notice that \(\Upsilon(\lambda)\) can be written as a block matrix

\[
\Upsilon(\lambda) = \begin{pmatrix}
\Upsilon_1 & \Upsilon_2(
\lambda) \\
\Upsilon_3(\lambda) & \Upsilon_4(\lambda)
\end{pmatrix},
\]

(4.109)

where \(\Upsilon_1\) is a constant matrix that does not depend on \(\lambda\). It is important to stress that in order to solve the characteristic equation Eq. (4.46), we should not be tempted to follow the steps of the perturbation analysis used for solving (4.48). Indeed, if we follow these steps, we should first find the solution of \(|\Upsilon(\lambda)||_{h=0} = 0\). (This is simple to do: when we set \(h = 0\), the blocks \(\Upsilon_2(\lambda), \Upsilon_3(\lambda)\) appear to be zero matrices, which makes \(|\Upsilon(\lambda)||_{h=0} = 0\) to have a very simple structure.) The solution of \(|\Upsilon(\lambda)| = 0\), i.e., Eq. (4.46), is then supposedly to be found as a perturbation of the solution of
|\( \mathbf{Y}(\lambda) \)| = 0. However, suppose \( \lambda^M \sim O(h) \), which will be confirmed later. Then the blocks \( \mathbf{Y}_2(\lambda), \mathbf{Y}_3(\lambda) \) are of order \( O(1) \) and hence cannot be ignored similarly to what we did for (4.48). Thus, we will have to directly compute \( |\mathbf{Y}(\lambda)| \), without using a perturbation analysis.

Notice that

\[
\mathbf{Y}_1 = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} \quad \Rightarrow \quad |\mathbf{Y}_1| = -2i \neq 0, \quad \text{and} \quad \mathbf{T}_1^{-1} = \begin{pmatrix} 1/2 & -i/2 \\ 1/2 & i/2 \end{pmatrix},
\]

(4.110)

which implies that \( \mathbf{Y}_1 \) is not a singular matrix. Therefore, to calculate an analytical expression of \( |\mathbf{Y}(\lambda)| \), we can use the formula of the determinant of a block matrix:

\[
\begin{vmatrix} \mathbf{Y}_1 & \mathbf{Y}_2 \\ \mathbf{Y}_3 & \mathbf{Y}_4 \end{vmatrix} = |\mathbf{Y}_1| \begin{vmatrix} \mathbf{Y}_4 - \mathbf{Y}_3 \mathbf{T}_1^{-1} \mathbf{Y}_2 \end{vmatrix}.
\]

(4.111)

Since \( |\mathbf{Y}_1| \neq 0 \), the formula (4.111) shows that Eq. (4.46) is equivalent to

\[
|\mathbf{Y}_4 - \mathbf{Y}_3 \mathbf{Y}_1^{-1} \mathbf{Y}_2| = 0.
\]

(4.112)

Substituting the explicit forms of \( \mathbf{Y}_{1,2,3,4} \) from (4.107) into (4.112) yields

\[
\begin{vmatrix} \lambda^M \left( \hat{\rho}^M_{2+} & \hat{\rho}^M_{2-} \\ -i\hat{\rho}^M_{2+} & i\hat{\rho}^M_{2-} \right) - \frac{ibh^2}{2(\lambda^2 - 1)^2 \lambda^M} \left( -2i\hat{\rho}^M_{2+} & 2i\hat{\rho}^M_{2-} \\ \hat{\rho}^M_{2+} & -\hat{\rho}^M_{2-} \right) \begin{pmatrix} 3 & -1 \\ 1 & 3 \end{pmatrix} \end{vmatrix} = 0.
\]

(4.113)

Notice that in order for the entire matrix above to be singular, it is necessary that
the magnitudes of its two terms be of the same order in $h$. This gives the condition

$$|\lambda^M| \sim \frac{O(h^2)}{\lambda^M} \Rightarrow |\lambda|^M \sim O(h). \quad (4.114)$$

This estimate is the main nontrivial result of our analysis. It shows that the magnitude of eigenvalue, $|\lambda(h)|$, is less than 1 for the modes with $\lambda^2 \not\approx 0$.

We will now obtain an accurate value of $\lambda^M$ which will confirm the validity of estimate of (4.114). Upon a simplification, we can write the determinant in (4.113) as follows

$$\left| \lambda^M I + \frac{bh^2}{4(\lambda^2 - 1)^2} \lambda^M \begin{pmatrix} 3 & -1 \\ 1 & -3 \end{pmatrix} \right|. \quad (4.115)$$

Subsequent analysis can produce formulas that would provide insight beyond (4.114) into the dependence $\lambda(h)$ only for $b = -1$ (see (1.8)). Therefore, below we consider only $b = -1$, which is also the case that was simulated numerically (see Chap. 3). Using $\theta$ to denote the argument of $1 + ih$, we have

$$\begin{pmatrix} \hat{p}_{1(+)}(+) \hat{p}_{2(+)}(+) \\ \hat{p}_{1(-)}(-) \hat{p}_{2(-)}(-) \end{pmatrix}^M = e^{-2iM\theta} \cdot e^{3Lh}, \quad \begin{pmatrix} \hat{p}_{1(+)}(-) \hat{p}_{2(+)}(-) \\ \hat{p}_{1(-)}(+) \hat{p}_{2(-)}(+) \end{pmatrix}^M = e^{3Lh}, \quad (4.116)$$

The detailed calculation for (4.116) is in Appendix B. After substituting (4.116) and
Thus, we need to solve $b = -1$ into (4.115), that equation, and hence (4.112), becomes:

$$
\begin{align*}
\lambda^M & \quad - \frac{\hbar^2 e^{3L \hbar}}{4(\lambda^2 - 1)^2 \lambda^M} \begin{pmatrix}
-3e^{i(-2M\theta)} & 1 \\
-1 & 3e^{i(2M\theta)}
\end{pmatrix} \begin{pmatrix}
3 \\
-1
\end{pmatrix} \\
& = \lambda^M + \frac{\hbar^2 e^{3L \hbar}(9e^{i(-2M\theta)} - 1)}{4(\lambda^2 - 1)^2 \lambda^M} - \frac{3\hbar^2 e^{3L \hbar}(e^{i(-2M\theta)} - 1)}{4(\lambda^2 - 1)^2 \lambda^M} \\
& = \left( \lambda^M + \frac{\hbar^2 e^{3L \hbar}(9e^{i(-2M\theta)} - 1)}{4(\lambda^2 - 1)^2 \lambda^M} \right) \left( \lambda^M + \frac{\hbar^2 e^{3L \hbar}(9e^{i(2M\theta)} - 1)}{4(\lambda^2 - 1)^2 \lambda^M} \right) \\
& - \frac{9\hbar^4 e^{6L \hbar}(e^{i(2M\theta)} - 1)(e^{i(-2M\theta)} - 1)}{16(\lambda^2 - 1)^4 \lambda^M} = 0.
\end{align*}
$$

Thus, we need to solve

$$
\begin{align*}
& \left( (\lambda^2 - 1)^2 \lambda^2 e^{3L \hbar}(9e^{i(-2M\theta)} - 1) \right) \left( (\lambda^2 - 1)^2 \lambda^2 e^{3L \hbar}(9e^{i(2M\theta)} - 1) \right) \\
& - 9\hbar^4 e^{6L \hbar}(e^{i(2M\theta)} - 1)(e^{i(-2M\theta)} - 1) \\
& = 16(\lambda^2 - 1)^2 \lambda^4 + 4(\lambda^2 - 1)^2 \lambda^2 e^{3L \hbar} \left( 9(e^{i(-2M\theta)} + e^{i(2M\theta)}) - 2 \right) \\
& + \hbar^4 e^{6L \hbar} \left( 82 - 9(e^{i(-2M\theta)} + e^{i(2M\theta)}) \right) - 9\hbar^4 e^{6L \hbar} \left( 2 - (e^{i(-2M\theta)} + e^{i(2M\theta)}) \right) \\
& = 16(\lambda^2 - 1)^2 \lambda^4 + 4(\lambda^2 - 1)^2 \lambda^2 e^{3L \hbar} \left( 18 \cos(2M\theta) - 2 \right) \\
& + \hbar^4 e^{6L \hbar} \left( 82 - 18 \cos(2M\theta) \right) - 9\hbar^4 e^{6L \hbar} \left( 2 - 2 \cos(2M\theta) \right) = 0,
\end{align*}
$$

(4.117)
which can be simplified to

\[ 2(\lambda^2 - 1)^4 \lambda^{4M} + \hbar^2 e^{3Lh}(\lambda^2 - 1)^2 \lambda^{2M}(9 \cos(2M\theta) - 1) + 8\hbar^4 e^{6Lh} = 0. \]  

(4.119)

Now let

\[ \hbar^2 e^{3Lh}z = (\lambda^2 - 1)^2 \lambda^{2M}. \]  

(4.120)

Substituting (4.120) into (4.119) and cancelling out the common factor \( \hbar^4 e^{6Lh} \), one obtains:

\[ 2z^2 + (9 \cos(2M\theta) - 1)z + 8 = 0, \]  

(4.121)

which has two solutions for \( z \):

\[ z_{1,2} = \frac{(1 - 9 \cos(2M\theta)) \pm \sqrt{(9 \cos(2M\theta) - 1)^2 - 64}}{4}. \]  

(4.122)

Notice that \( M = L/h \) and according to the discussion in Appendix B, \( \theta \approx h \), hence \( M\theta \approx \frac{L}{h} \cdot h = L \) (Here, \( L \) is the length of the space line), and

\[ z_{1,2} \approx \frac{(1 - 9 \cos(2L)) \pm \sqrt{(9 \cos(2L) - 1)^2 - 64}}{4}. \]  

(4.123)

Let us note that \( |z| = O(1) \). Indeed, the only other possibility, according to (4.123), could have been \( |z| \ll 1 \) (or \( |z| \approx 0 \)). This, however, is not possible, because for any value of \( M\theta \), \( |9 \cos(2M\theta)| - 1 \leq 10 \), and hence one can show by a tedious but straightforward calculation that \( |z| \in [1, 4] \).
To simplify subsequent notations, denote

\[ \pm z^{1/2} = \zeta \equiv |\zeta|e^{i\phi}; \quad (4.124) \]

Then (4.120) reduces to

\[ \lambda^M (\lambda^2 - 1) = h e^{\frac{3Lh}{2}\zeta}. \quad (4.125) \]

From (4.125), one can deduce that

\[ \lambda^M = h \frac{e^{\frac{3Lh}{2}\zeta}}{\lambda^2 - 1}. \quad (4.126) \]

In what follows we will consider only the situation where

\[ e^{Lh} = O(1) \quad (4.127a) \]

for two reasons. First, this condition corresponded to our numerics. Second, considering the other limit, i.e.,

\[ e^{Lh} \gg 1, \quad (4.127b) \]

would not lead to the phenomenon of suppression of numerical instability of the MoC-ME, which we will aim to explain in Chap. 5.

Since \( \zeta = O(1) \), \( (\lambda^2 - 1) = O(1) \), and \( e^{Lh} = O(1) \), we must have \( |\lambda^M| = O(h) \), which verifies our estimate (4.114). This result along with \( h \ll 1 \) implies that for
\( \lambda^2 \approx 1 \), we have \(|\lambda| < 1\). Note, however, that \( M \gg 1 \), and \( h \sim 1/M \), we have

\[
\ln(h^{1/M}) \sim \ln((1/M)^{1/M}) = \frac{\ln(1/M)}{M} = -\frac{\ln M}{M} \approx 0, \tag{4.128}
\]

and

\[
\ln \left( \frac{|e^{3Lh/2} z|^{1/M}}{\lambda^2 - 1} \right) = \frac{\ln(|e^{3Lh/2} z/(\lambda^2 - 1)|)}{M} \approx 0. \tag{4.129}
\]

Therefore

\[
\ln(|\lambda|) = \ln \left( \frac{|e^{3Lh/2} z|^{1/M}}{\lambda^2 - 1} \right) + \ln(h^{1/M}) \approx 0 + 0 = 0, \tag{4.130}
\]

which implies that \(|\lambda| \to 1\). In the following subsection, we will discuss how to quantitatively compute \( \lambda \) using (4.125), and then verify this calculation using numerical simulations.

### 4.5.4 Calculations of \( \lambda \) and its Numerical Verification

In the proceeding subsection, we found a key equation needed to compute the value of \( \lambda \):

\[
\lambda^M (\lambda^2 - 1) = he^{3Lh/2} z, \tag{4.131}
\]
which is valid when $\lambda^2 - 1 \approx 0$. In order to solve this equation in the limit $M \gg 1$, we will make use of the fact that $\hat{z}$ and $\lambda^2 - 1$ are both $O(1)$. Correspondingly, we let

$$\lambda^2 - 1 = re^{i\psi}, \quad \hat{z} = |\hat{z}|e^{i\hat{\phi}}, \quad r, |\hat{z}|, \psi, \hat{\phi} \in \mathbb{R}. \quad (4.132)$$

Here, $r, |\hat{z}|$ are both $O(1)$. $\psi, \hat{\phi}$ are arguments in the interval $[-\pi, \pi)$. To calculate $\hat{z}$, we resort to (4.123) and (4.124), so as long as the space length $L$ is given, we can compute $|\hat{z}|$. Substituting (4.132) into (4.131) yields

$$\lambda^M re^{i\psi} = e^{3Lh/2|\hat{z}|h} |\hat{z}| e^{i\hat{\phi}}$$

$$\Rightarrow \lambda^M = \left(\frac{e^{3Lh/2|\hat{z}|h}}{r}\right) e^{i\hat{\phi} - i\psi}$$

$$\Rightarrow \lambda_l = \left(\frac{e^{3Lh/2|\hat{z}|h}}{r}\right)^{1/M} e^{i\left(\frac{2\pi \psi}{M}\right) + i\left(\frac{2\pi l}{M}\right)} \quad (4.133)$$

where $l = 0, 1, \ldots, M - 1$. Thus

$$|\lambda| = \left(\frac{e^{3Lh/2|\hat{z}|h}}{r}\right)^{1/M}. \quad (4.134)$$

In (4.134), the quantities $h, |\hat{z}|, M, \text{and } L$ are all known. The only quantity we do not know is $r$. Therefore, in order to calculate $|\lambda|$, we need to calculate $\lambda^2 - 1 = re^{i\psi}$ first. We will show that $r$ only depends on the argument of $\lambda$.

Recall, from the end of Sec. 4.5.3, that $|\lambda|$ is less than 1 but very close to 1. Thus, $\lambda$ is within the unit circle in the complex plane, but very close to the unit circle. We may let $|\lambda| = 1 - \epsilon$, where $0 < \epsilon \ll 1$. Then $\lambda^2 - 1$ is a complex number obtained by shifting $\lambda^2$ to the left by 1 unit. The small difference caused by $\epsilon$ can be ignored in comparison with the magnitude of $\lambda^2 - 1$. Thus, we can assume $|\lambda^2| = 1$ in subsequent
calculations. We depict this in Fig. 4.1. One can see that the only thing that matters

\[ r = |\lambda^2 - 1| \]

is the argument of \( \lambda^2 \). For instance, if \( \arg(\lambda) = \pi/2 \), then

\[ \arg(\lambda^2) = \pi \quad \Rightarrow \quad \lambda^2 - 1 \approx -2 \quad \Rightarrow \quad r = 2. \quad (4.135) \]

Following the same logic, for each argument of \( \lambda \equiv \lambda_l \) in (4.133), we can calculate a corresponding \( r \equiv r_l \).

Now we will clarify the relation between \( l \) and the argument of \( \lambda \) based on (4.133). Notice that since \( M \gg 1 \), and \( \hat{\phi}, \psi = O(1) \), we have

\[ \frac{\hat{\phi} - \psi}{M} \approx 0 \quad \Rightarrow \quad \arg(\lambda_l) \approx \frac{2\pi l}{M} \quad \Rightarrow \quad \arg(\lambda_l^2) \approx \frac{4\pi l}{M}. \quad (4.136) \]

Notice that \( l/M \) can be \( O(1) \) because \( l = 0, 1, \ldots, M - 1 \). Also notice that since we are assuming \( \lambda^2 - 1 \approx 0 \), the argument of \( \lambda \) should not be near 0, \( \pi \), or \( 2\pi \) (since otherwise we have \( \lambda \approx 1, -1 \) and hence \( \lambda^2 - 1 \approx 0 \)). Therefore, in (4.136), \( l \) cannot be
near 0, \(M/2\) or \(M - 1\). For example, when \(l \approx M/4\), or, equivalently, \(\arg(\lambda_l) \approx \pi/2\), and hence \(r \approx 2\) (see (4.135)), we will obtain an accurate approximation for \(|\lambda|\) from (4.134).

To calculate \(|\lambda|\) in a specific case, we let \(L = 50, h = 0.01\) (hence \(M = L/h = 5000\)). Then we take \(l = 100, \ldots, M/2 - 100, M/2 + 100, \ldots, (M - 1) - 100\), where the significance of using the number 100 \(\gg 1\) has been explained a few sentences above. Namely, the corresponding argument of \(\lambda\) ranges from 0 to \(2\pi\), but stays away from 0, \(\pi\) and \(2\pi\). Following the same steps that lead to (4.135), we can calculate \(r_l\) that correspond to each argument \(2\pi l/M\). Substituting these parameters, including \(r_l\), into (4.134), we can calculate \(|\lambda_l|\) corresponding each \(l\). The result is shown in Fig. 4.2. We can see that, as expected from (4.134) and (4.132), the largest \(|\lambda|\) occurs

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure42.png}
\caption{Theoretical \(|\lambda|\) for MoC-SE with non-reflecting b.c. when \(L = 50, h = 0.01, b = -1\). Notice that since in the calculations, we have two values of \(z\): \(z_1, z_2\), there should be two sets of curves (see (4.123)). However, since \(|z_1| = |z_2|\) for the parameters we choose, we only need to show one set of curves.}
\end{figure}
where \(|\lambda^2 - 1|\) is the smallest, or \(\lambda \approx \pm 1\). However, this is where we "approach" Case 2 of Sec. 4.5.2, so we will only note that the increase of \(|\lambda|\) as \(\arg(\lambda_l) \to 0, \pi, \text{ or } 2\pi\) is consistent with our simplified treatment of Case 2 in Sec. 4.5.2, where we showed that in that case, \(|\lambda|\) is slightly greater than 1.

For simplicity, we will carry out further verification for those \(\lambda_l\) where \(l \approx M/4\) because in this case we have an accurate estimate for \(r \approx 2\), and hence for \(|\lambda|\). For brevity, we simply write \(\lambda\) to denote \(\lambda\) with \(\arg(\lambda) \approx \pi/2\). We will now demonstrate that the scaling \(|\lambda|^M = O(h)\), predicted by (4.134), indeed holds, by directly simulating the PDE.

To obtain the numerical \(|\lambda|\) where the argument is near \(\pi/2\), we need to compute the Fourier spectrum of the numerical error (denoted below as FFT error) with wave numbers near \(\pi/2\). However, in order to calculate the Fourier spectrum of the error, we have to make sure that the error is periodic. In order to do that, we multiply the error by a function whose values are close to 0 at the two boundaries, but close or equal to 1 when it is not near the boundaries. The function that we use here is a super-Gaussian:

\[
f(x) = e^{-\left(\frac{x}{L/2}\right)^8},
\]  

where the space variable \(x\) ranges from \(-L/2\) to \(L/2\). This function is shown in Fig. 4.3.
The super-Gaussian function $f(x) = e^{-\left(\frac{1.5x}{L/2}\right)^8}$ for $x \in [-L/2, L/2]$ with $L = 50$.

Also notice that we cannot simply use the data at the wave number of exactly $\pi/2$. If we had done so, we would have gotten a poor approximation of $|\lambda|$ because of the randomness of the error, which can cause the data to fluctuate irregularly. In other words, we need to smoothen the data. Thus, we need to collect the data that are in some finite vicinity of the wave number $\pi/2$.

Before we discuss this process of smoothing data, let us discuss the structure of the FFT error in our numerical simulations and how to extract the data we need. Suppose the FFT errors of the four components $S_1^\pm, S_3^\pm$ at time $t$ are $\hat{s}_{1,3,4}^{(t)}$ respectively, then each of $\hat{s}_{1,2,3,4}^{(t)}$ is a vector. Each element of the vector corresponds to one wave number. For example, if the (normalized) wave number ranges from 0 to $2\pi$, then the first element of $\hat{s}_{1,2,3,4}^{(t)}$ is the FFT error of $S_1^+$ with wave number 0; we can denote this element as $\hat{s}_{1,2,3,4}^{(t)}(1)$. The second element of $\hat{s}_{1,2,3,4}^{(t)}$ is the FFT error of $S_1^+$ with wave number $2\pi/M$; we can denote this element as $\hat{s}_{1,2,3,4}^{(t)}(2)$. The $n$th element
of \( \hat{s}_1^{(t)} \) is the FFT error of \( S_1^+ \) with wave number \( 2(n - 1)\pi/M \); we can denote this element of \( \hat{s}_1^{(t)}(n) \); and so on. Following this pattern, we know that the FFT error with wave number \( \pi/2 \) would have the index \( M/4 \). Therefore, to collect the FFT errors with wave numbers near \( \pi/2 \), we need to collect the elements of vectors \( \hat{s}_{1,2,3,4}^{(t)} \) whose indices are near \( M/4 \). We chose to take the indices ranging from \( M/4 - 20 \) to \( M/4 + 20 \), where, in comparison with \( M \), 20 is a very small number. Thus in the calculation of \( |\lambda_{num}| \) described below, we used the quantity

\[
|FFT_{\text{error}}(t)| = \sqrt{\sum_{i=1}^{M/4+20} \sum_{l=M/4-20}^{M/4+20} |\hat{s}_1^{(t)}(l)|^2},
\]

(4.138)

whose evolution with time is shown in Fig. 4.4.

---

**Figure 4.4: Evolution of \( \log_{10}(|FFT_{\text{error}}(t)|) \) vs \( t \). The corresponding values of \( L \) and \( h \) are shown in the legend; \( b = -1 \) in all cases.**

For future convenience, we may refer to this number 20 as half width of the “box”
since this index range serves as the “box” that collects the FFT error near wave number $\pi/2$.

**Verification of $|\lambda|^M = O(h)$**

We need to first calculate the value of numerical $|\lambda|$, i.e., $|\lambda_{num}|$, that corresponds to each chosen value of $h$. The numerical error evolves as

$$|FFT_{\text{error}}(t)| = |FFT_{\text{error}}(0)||\lambda_{num}|(t/h),$$

(4.139)

where $FFT_{\text{error}}(0)$ is the initial perturbation of the exact solution. Taking the logarithm to base 10 of both sides of (4.139), we have:

$$\log_{10}(|FFT_{\text{error}}(t)|) = \log_{10}(|FFT_{\text{error}}(0)|) + (t/h) \log_{10}(|\lambda_{num}|).$$

(4.140)

Denoting the slope with respect to $t$ of the r.h.s. of (4.140) as $r$, we have

$$r = \frac{\log_{10}(|\lambda_{num}|)}{h} \Rightarrow |\lambda_{num}| = 10^{(hr)}.$$ 

(4.141)

Therefore, in order to obtain $|\lambda_{num}|$, we need to numerically calculate the slope $r$. To do that, we will perform linear regression on the data of $\log_{10}(|FFT_{\text{error}}(t)|)$, as described below. After obtaining $r$, we will use formula (4.141) to calculate $|\lambda_{num}|$. Fig. 4.5 helps us illustrate how we do the linear regression of $\log_{10}(|FFT_{\text{error}}(t)|)$ with respect to $t$. 

---

1 We have determined that the plateau-like shape of this curve is due to the finite time that it takes a perturbation to propagate through the computational domain. We choose the data points in Fig. 4.5 such that in between every two consecutive points there is a constant time interval which equals to $L/$ speed $= L$. After that, we perform the standard linear regression using these chosen...
Figure 4.5: The solid curve shows the evolution of $\log_{10}(\|FFT_{\text{error}}(t)\|)$. Here, we use the parameters $b = -1, h = 0.00625, L = 25$ as an example. The data points marked with asterisks are such that every two consecutive points are separated by the same time interval (see the main text). Connecting these points gives the dashed line.

Notice that for MoC-SE with non-reflecting b.c., our theoretical result in Sec. 4.5.3 gives $|\lambda|^M \sim h$. This implies that $|\lambda|^{2M} \sim h^2$. (The reason we use $|\lambda|^{2M}$ for the verification instead of $|\lambda|^M$ is that $|\lambda|^{2M}$ gives more stable data.) Thus, according to (4.134),

$$|\lambda|^{2M} = C_0 h^2 e^{3Lh},$$

where $C_0 = |\hat{z}|^2/r^2$ is a quantity that does not dependent on $h$. This implies that

$$\ln(|\lambda|^{2M}) = 2 \ln(h) + 3Lh + \ln(C_0).$$

Data points to approximate the slope $r$; then we substitute $r$ into (4.141) to calculate $|\lambda_{\text{num}}|$.
Viewing the quantity $\ln(|\lambda|^{2M})$ as a function of $\ln(h)$, we find its slope as follows:

\[
\frac{d \ln(|\lambda|^{2M})}{d \ln(h)} = \frac{d \ln(|\lambda|^{2M})}{dh} \cdot \frac{dh}{d \ln(h)} = \frac{d (2 \ln(h) + 3Lh + \ln(C_0))}{dh} \cdot h = \left(\frac{2}{h} + 3L\right) h = 2 + 3Lh = 2 + O(h).
\] (4.144)

Therefore, the plot of $\ln(|\lambda_{num}|^{2M})$ vs $\ln(h)$ should be close to a straight line with a slope being approximately $2$ (as $h \to 0$), if our theoretical result is correct.

The plot of $\ln(|\lambda_{num}|^{2M})$ vs $\ln(h)$ is shown in Fig. 4.6 and indeed is almost a straight line.

![Figure 4.6: The values $\ln(|\lambda|^{2M})$ found from numerical simulations vs $\ln(h)$ for MoC-SE with non-reflecting b.c. The parameters are $b = -1$, $L = 25$ and 50, and $h = 0.01, 0.00625, 0.005, 0.004, 0.002, 0.001, 0.0005$.](image)

Using these data for the two smallest $h$ values, $\ln(h_1) = -6.91$, which corresponds to $\ln(|\lambda_1|^{2M}) = -14.42$, and $\ln(h_2) = -7.60$, which corresponds to $\ln(|\lambda_2|^{2M}) = \ldots$
−15.84, we get an estimate of the slope:

\[
\frac{\ln \left( |\lambda_2^{2M}| \right) - \ln \left( |\lambda_1^{2M}| \right)}{\ln (h_2) - \ln (h_1)} = \frac{(-15.84) - (-14.42)}{(-7.60) - (-6.91)} = 2.04 \approx 2.00.
\] (4.145)

Thus, the theoretical result \(|\lambda|^M \sim h\) is verified.

**Factor of \(\frac{\text{numerical } |\lambda|^{2M}}{\text{theoretical } |\lambda|^{2M}}\) vs \(Lh\)**

From the discussion above, we can see that \(|\lambda|^M\) is indeed \(O(h)\). However, this is only a qualitative verification. In order to quantitatively verify the theoretical analysis of \(|\lambda|\), i.e., the accuracy of (4.134), we resort to the same process that we used in Sec. 3.1.4, 3.2.3, 3.3.2. Namely, we compare the theoretical value (4.134), which we will refer to as “theoretical result”, and the numerical \(|\lambda|^{2M}\) computed as described above, which we will refer to as “numerical result”. To do that, we calculate the factor

\[
\frac{\text{numerical } |\lambda|^{2M}}{\text{theoretical } |\lambda|^{2M}}.
\] (4.146)

If the theoretical result agrees with the numerical result, then this factor should be approximately 1. Notice that different sets of parameters \(b, L, h\) will give different values of \(|\lambda|\). However, this will not significantly affect the value of the factor (4.146) if the theoretical \(|\lambda|\) and the numerical \(|\lambda|\) agree with each other; i.e., factor (4.146) will always be close to 1. We know from (4.134) that

\[
|\lambda|^{2M} = |\hat{z}|^2/r^2 h^2 e^{3Lh},
\] (4.147)
which depends on $h$ and $Lh$. Indeed, from Fig. 4.6, we see that for the same $h$, different values of $Lh$ give different curves. Thus, we will display the factor (4.146) for two different values of $L$. We plot factor (4.146) vs $Lh$ in Fig. 4.7. It is seen to be almost 1, which verifies our theoretical result.

![Figure 4.7: Factor $\frac{\text{numerical } |\lambda|^{2M}}{\text{theoretical } |\lambda|^{2M}}$ vs $Lh$. The parameters used are $b = -1$, $L = 25$, and 50 and $h = 0.01, 0.00625, 0.005, 0.004, 0.002, 0.001, 0.0005$.]

4.6 PREVIEW FOR THE INSTABILITY SUPPRESSING MECHANISM OF MoC-ME WITH NON-REFLECTING B.C.

In PDE simulations we found that with the non-reflecting BC given above, the numerical instability of MoC-SE and MoC-LF remains the same for periodic b.c. However,
the numerical instability of MoC-ME *disappears* when we change periodic b.c. to non-reflecting b.c. This chapter studied the numerical instability of MoC-SE with non-reflecting b.c. Although these results are not for MoC-ME with non-reflecting b.c., they still provided some insight of the instability suppressing mechanism of MoC-ME with non-reflecting b.c..

According to a complicated analysis of a large circulant block matrix in Sec. 4.5, we found that for MoC-SE:

\[
|\lambda|^M = O(h) \ll 1, \quad (4.148)
\]

when the wave number of the mode is not near 0, π; where \( M \) is the number of spatial grid points. This indicates that for the modes with wave number not near 0, π, we must have

\[
|\lambda| < 1. \quad (4.149)
\]

Therefore, these modes will decay in time, i.e., they are numerically stable. In particular, instability of modes with wave number near \( kh = \pi/2 \) is strongly suppressed. This instability-suppressing mechanism of MoC-SE with non-reflecting b.c. is schematically illustrated in Fig. 4.8.
In the left panel of Fig. 4.8, we see that in MoC-SE scheme with periodic b.c., all modes are unstable: $|\lambda| > 1$. In particular, the numerical instability is the strongest for modes with wave numbers near $0, \pi$. When we use the non-reflecting b.c. for MoC-SE scheme, then for all modes whose wave numbers are not near $0, \pi$, $|\lambda|$ is “pushed” below 1; hence the numerical instability of modes with wave numbers not near $0, \pi$ is suppressed. However, this suppression does not affect the most unstable modes, i.e., modes with wave numbers near $0, \pi$. Therefore, we still observe the same instability for the MoC-SE with non-reflecting b.c. as for the MoC-SE with periodic b.c..

Since the MoC-ME and MoC-SE schemes have similar structure, we may assume that the disappearance of numerical instability of the MoC-ME scheme, when periodic b.c. are replaced with non-reflecting ones, is also caused by a similar mechanism. This is schematically illustrated in Fig. 4.9.
In the left panel of Fig. 4.9, we see that in the MoC-ME with periodic b.c., the most unstable modes are those with wave numbers near $\pi/2$. (There is also instability in the modes with wave numbers near 0, $\pi$, i.e., in the ODE limit, but it is much weaker.) When we impose non-reflecting b.c. for the MoC-ME, the instability suppressing mechanism “pushes” the $|\lambda|$ of the modes whose wave number are not near 0, $\pi$ below 1. This suppresses the strongest instability of the MoC-ME scheme, which occurs for $kh \approx \pi/2$. Even though this mechanism does not affect the modes with wave numbers near 0, $\pi$, the instability for these modes is originally quite weak. Therefore, in the MoC-ME with non-reflecting b.c., we do not observe numerical instability unless we run simulations for a very long time. With this motivation, when we study the numerical stability of MoC-ME with non-reflecting b.c. in the next chapter, we will focus on the modes with wave numbers near $\pi/2$. 

*Figure 4.9: See text in Sec. 4.6.*
Chapter 5

Stability analysis ME with non-reflecting b.c.

In the numerical experiment, we found that method of characteristics using the modified Euler scheme (MoC-ME) with periodic b.c. is unstable and the instability agrees with the von Neumann analysis. However, MoC-ME with non-reflecting b.c. was found to be stable. In this chapter, we study analytically the instability of MoC-ME with non-reflecting b.c. and quantitatively explain the mechanism by which the instability is suppressed.

5.1 Set up of the method

This process is similar to the derivation of the difference operator for the MoC-SE with non-reflecting BC in Sec. 4.2. Similarly to Sec. 4.2, we use the notations that
are defined in Sec. 2.2, such as \( \tilde{S} = (\tilde{S}^+_1, \tilde{S}^+_3, \tilde{S}^-_1, \tilde{S}^-_3)^T \). Denote

\[
\begin{align*}
y^+ &= \begin{pmatrix} \tilde{S}^+_1 \\ \tilde{S}^+_3 \end{pmatrix}, \\
y^- &= \begin{pmatrix} \tilde{S}^-_1 \\ \tilde{S}^-_3 \end{pmatrix}, \\
y &= \begin{pmatrix} y^+ \\ y^- \end{pmatrix}, \\
(y)^n_m &= \begin{pmatrix} y^+_n \\ y^-_n \end{pmatrix}_m. \quad (5.1)
\end{align*}
\]

Based on (3.24), (4.3), and (4.4), we can denote the MoC-ME scheme as the following difference equation:

\[
\begin{align*}
\left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_m &= \Gamma \left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_{m-1} + \Omega \left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_{m+1} \\
\left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_{m+1} &= \frac{1}{2} \left( \begin{array}{c} y^+_{m-1} \\ y^-_{m+1} \end{array} \right) + (I + hP) \left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_m. \quad (5.2)
\end{align*}
\]

\[
\begin{align*}
\left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_{m+1} &= \frac{1}{2} \left( \begin{array}{c} y^+_{m-1} \\ y^-_{m+1} \end{array} \right) + (I + hP) \left( \begin{array}{c} y^+ \\ y^- \end{array} \right)_m. \quad (5.3)
\end{align*}
\]

where \( \Gamma, \Omega \) are, respectively, the \( \Gamma, \Omega \) in (4.4) and \( I, P \) are defined after (2.8). Here, (5.2) corresponds to (3.24a), and (5.3) corresponds to (3.24b). Note also that

\[
\Gamma + \Omega = I + hP. \quad (5.4)
\]

After detailed calculation, we can see that (5.2) and (5.3) combined can be written as a single difference equation:

\[
(y)^n_{m+1} = \Gamma ME (y)^n_{m-1} + \Omega ME (y)^n_{m+1}, \quad (5.5)
\]
where, upon using the expression (2.9) for $P$:

\[
\begin{align*}
\Gamma_{ME} & = \frac{1}{2} \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{2} \begin{pmatrix} I + bhA & 0 \\ bhB & 0 \end{pmatrix} \Gamma_{SE} \\
\Omega_{ME} & = \frac{1}{2} \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} + \frac{1}{2} \begin{pmatrix} 0 & hB \\ O & I + hA \end{pmatrix} \Omega_{SE},
\end{align*}
\]  

(5.6)

(5.7)

and $2 \times 2$ matrices $A$ and $B$ are defined in (2.10). Notice that (5.5) is a counterpart of (4.3). When we consider the non-reflecting b.c. described by (4.13), (4.14), (4.15), (4.16), the difference equation of the MoC-ME with non-reflecting b.c. can be cast in the form:

\[
(Y)^{n+1} = \begin{pmatrix} O & \Omega_{ME} & O & O & \cdots & \cdots & O \\ \Gamma_{ME} & O & \Omega_{ME} & O & \cdots & \cdots & \cdots \\ O & \Gamma_{ME} & O & \Omega_{ME} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \Gamma_{ME} & O & \Omega_{ME} \\ O & \cdots & \cdots & \cdots & O & \Gamma_{ME} & O \end{pmatrix} (Y)^n + \begin{pmatrix} F \\ 0 \\ \vdots \\ 0 \\ G \end{pmatrix} \\
\equiv \Phi_{ME} (Y)^n + V_{ME},
\]  

(5.8)

which is a counterpart of (4.17). Here, the notation $(Y)^n$ is defined in (4.7).

According to an analysis similar to the one in Sec. 4.4, in order to study the numerical instability of the MoC-ME with non-reflecting b.c., one needs to calculate the eigenvalue $\lambda$ of the matrix $\Phi_{ME}$. To do that, we will go through the same three main steps that we discussed at the end of Sec. 4.4. That is, we will need to calculate
\( \rho_k(\lambda) \) by solving the characteristic equation

\[
E(\rho, \lambda) \equiv \left| \Omega_{ME} \rho^2 - \lambda \rho I + \Gamma_{ME} \right| = 0, \quad (5.9)
\]

then calculate the corresponding eigenvector \( \xi_k(\lambda) \) by solving

\[
\left( \Gamma_{ME} - \lambda \rho_k(\lambda) I + \Omega_{ME} \rho_k^2(\lambda) \right) \xi_k = 0, \quad (5.10)
\]

and finally, use \( \xi_k(\lambda) \) and \( \rho_k(\lambda) \) to solve

\[
|\Upsilon(\lambda)| = 0 \quad (5.11)
\]

for \( \lambda \), where the \( 4 \times 4 \) matrix

\[
\Upsilon(\lambda) = \begin{pmatrix}
\xi_1^+(\lambda) & \xi_2^+(\lambda) & \xi_3^+(\lambda) & \xi_4^+(\lambda) \\
\rho_1^{M+1}(\lambda)\xi_1^{-}(\lambda) & \rho_2^{M+1}(\lambda)\xi_2^{-}(\lambda) & \rho_3^{M+1}(\lambda)\xi_3^{-}(\lambda) & \rho_4^{M+1}(\lambda)\xi_4^{-}(\lambda)
\end{pmatrix}. \quad (5.12)
\]

The only difference is that the matrices \( \Gamma, \Omega \) are now \( \Gamma_{ME}, \Omega_{ME} \); this leads to a certain quantitative difference for modes with wave numbers in the middle of the spectrum. The three steps above are implemented in Sec. 5.2, 5.3, 5.4.
5.2 Finding the approximate solution $\rho(\lambda)$

This process is the same as in Sec. 4.5.1. We will perform the perturbation analysis to find an approximate solution of (5.9) (for $h \ll 1$) so that we can get the four $\rho_k(\lambda)$. The only difference is that the polynomial $E(\rho, \lambda)$ is different from the one from Sec. 4.5.1: Compare (5.9) with (4.38) and recall that in the latter equation, $\Gamma \equiv \Gamma_{SE}$ and $\Omega \equiv \Omega_{SE}$.

Similarly to (4.48), we first solve $E(\rho, \lambda)|_{h=0} = 0$:

\[
E(\rho, \lambda)|_{h=0} = \begin{vmatrix} (1 - \lambda \rho)I & O \\ O & (\rho^2 - \lambda \rho)I \end{vmatrix} = (1 - \lambda \rho)^2(\rho^2 - \lambda \rho)^2 = \rho^2(\lambda \rho - 1)^2(\rho - \lambda)^2 = \rho^2 \lambda^2(\rho - \lambda)^2 \left( \frac{\rho - 1}{\lambda} \right)^2. \tag{5.13}
\]

Thus there are two nonzero roots for $E(\rho, \lambda)|_{h=0} = 0$:

\[
\rho_1 = \frac{1}{\lambda} \quad \text{and} \quad \rho_2 = \lambda. \tag{5.14}
\]

In order to find the asymptotic solution of (5.9), we should perform perturbation analysis near these two roots. Here, we only discuss the case when $\lambda^2 - 1 \approx 0$, i.e., when $\rho_1 \approx \rho_2$. Our justification for not considering the other case, i.e., $\rho_1 \approx \rho_2$ or, equivalently, $\lambda^2 - 1 \approx 0$, is as follows.

- First, the numerical instability for $\lambda^2 - 1 \approx 0$ (i.e. $\lambda \approx 1$ or $\lambda \approx -1$) will still
occur, but it is similar to the numerical instability of the ME for ODEs (because
the limit \( \lambda \approx 1 \) corresponds to \( kh \to 0 \) in the von Neumann analysis). In our
study of the MoC-ME with periodic b.c. in Sec. 3.2.2, we showed that it is much
weaker than the numerical instability of the MoC-ME for modes with \( \lambda^2 \approx -1 \).
Thus, we are more concerned about the instability for \( \lambda^2 - 1 \approx 0 \) than about
that for \( \lambda^2 - 1 \approx 0 \).

• Second, as we have previewed in Sec. 4.6 and will demonstrate below, suppres-
sion of numerical instability occurs for \( \lambda^2 - 1 \approx 0 \).

• Third, as we saw in Sec. 4.5.1, the analysis is different for the cases when \( \rho_1 \not\approx \rho_2 \)
and \( \lambda^2 - 1 \approx 0 \) (\( \rho_1 \approx \rho_2 \) (\( \lambda^2 - 1 \approx 0 \))). The latter is much more difficult even
for the MoC-SE; for MoC-ME, it is even more difficult.

Thus, below we perform a perturbation analysis when \( \rho_1 \not\approx \rho_2 \). As in Sec. 4.5.1,
there are two cases.

**Case I:** \( \rho \approx \rho_1 \).

Let \( \rho = \rho_1 + \tilde{\rho}_1 \), where \( |\tilde{\rho}_1| \ll 1 \). Substituting \( \rho = \rho_1 + \tilde{\rho}_1 = \frac{1}{\lambda} + \tilde{\rho}_1 \) into \( E(\rho, \lambda) = 0 \) and using (5.9), (5.6), (5.7) and (4.4), one obtains:

\[
E(\rho, \lambda) = E\left(\frac{1}{\lambda} + \tilde{\rho}_1, \lambda\right) \\
= \tilde{\rho}_1^2 \left( \lambda^2 + \frac{1}{\lambda^2} - 2 \right) + O(\tilde{\rho}_1^3) \\
+ h^2 \left( b^2 \left( 1 + \frac{1}{\lambda^4} - \frac{2}{\lambda^2} \right) + O(\tilde{\rho}_1) \right) \\
+ O(h^4) = 0.
\]

\[(5.15)\]
After omitting the smaller terms $O(\tilde{\rho}_1^3), O(h^2\tilde{\rho}_1)$ and $O(h^4)$, this equation gives

$$
\left(\tilde{\rho}_1^2\lambda^2 + b^2h^2\right) \frac{(\lambda^2 - 1)^2}{\lambda^4} = 0.
$$

(5.16)

Since we consider $\lambda^2 - 1 \not\approx 0$, the equation above implies that

$$
\tilde{\rho}_1^2\lambda^2 + b^2h^2 = 0 \quad \Rightarrow \quad \tilde{\rho}_1 = \pm i\frac{1}{\lambda}bh.
$$

(5.17)

Therefore, if $\rho$ is near $\rho_1$, we have

$$
\rho = \rho_1 + \tilde{\rho}_1 = \frac{1}{\lambda} + \tilde{\rho}_1 = (1 \pm i bh) \frac{1}{\lambda}.
$$

(5.18)

**Case II: $\rho \approx \rho_2$.**

Let $\rho = \rho_2 + \tilde{\rho}_2$, where $|\tilde{\rho}_2| \ll 1$. Substituting $\rho = \rho_2 + \tilde{\rho}_2 = \lambda + \tilde{\rho}_2$ into $E(\rho, \lambda) = 0$, one obtains:

$$
E(\rho, \lambda) = E(\lambda + \tilde{\rho}_2, \lambda)
= \tilde{\rho}_2^2 \left(\lambda^2 - 2\lambda^4 + \lambda^6\right) + O(\tilde{\rho}_2^3)
+ h^2 \left(\lambda^2 \left(\lambda^2 - 2\lambda^4 + \lambda^6\right) + O(\tilde{\rho}_2)\right)
+ O(h^4) = 0.
$$

(5.19)

After omitting the smaller terms $O(\tilde{\rho}_2^3), O(h^2\tilde{\rho}_2)$ and $O(h^4)$, this equation gives

$$
(\tilde{\rho}_2^2 + h^2\lambda^2)\lambda^2(\lambda^2 - 1)^2 = 0.
$$

(5.20)

To simplify this, recall that we consider the case $\lambda^2 - 1 \not\approx 0$. Moreover, since we
seek any possible “culprit” modes that create numerical instability, then we also have $|\lambda| > 1$. Therefore, in (5.20), $\lambda^2 \approx 0$. Thus, that equation implies that

$$\tilde{\rho}_2^2 + h^2 \lambda^2 = 0 \quad \Rightarrow \quad \tilde{\rho}_2 = \pm i h \lambda. \quad (5.21)$$

Therefore, if $\rho$ is near $\rho_2$, we have

$$\rho = \rho_2 + \tilde{\rho}_2 = \lambda + \tilde{\rho}_2 = (1 \pm i h) \lambda. \quad (5.22)$$

The expressions in (5.17) and (5.22) are the same as those in Sec. 4.5.1. Therefore, we will use the same notations (4.58) and (4.60). Here, $\rho_1(\pm), \rho_2(\pm)$ are the four $\rho_k(\lambda)$ that we need in order to proceed with the calculation in Secs. 5.3, 5.4. We will denote their corresponding eigenvectors $\xi_k$ as $\xi_{1,2(\pm)}$.

### 5.3 Finding the approximate solution $\xi(\lambda)$

This process is similar to that in Sec. 4.5.2. According to the discussion in Sec. 5.1, the next step after finding $\rho_k(\lambda)$, is to use (5.10) to find the four eigenvectors $\xi_k(\lambda)$. To that end, we substitute the obtained values of $\rho_k(\lambda)$ into (5.10) and solve it for the corresponding $\xi_k(\lambda)$. For the reason explained after (5.14), we will only consider the case $\lambda^2 - 1 \approx 0$.

According to Sec. 5.2, the four roots of (5.9) are: $\rho_{1(\pm)} = \frac{1}{\lambda} (1 \pm i h)$ and $\rho_{2(\pm)} = \ldots$
\(\lambda (1 \pm ih)\). As in Chap. 4, we will seek \(\rho_{1,2}(\pm)\) in the form:

\[
\rho_{1(\pm)} = (1 \pm ibh + w_{1,2}h^2)\frac{1}{\lambda},
\]

(5.23)

\[
\rho_{2(\pm)} = (1 \pm ih + w_{3,4}h^2)\lambda,
\]

(5.24)

where \(w_{1,2}, w_{3,4}\) are coefficients that one can calculate by considering higher order terms in, (5.15) and (5.19). However, unlike in Chap. 4, here the presence of these coefficients is essential in the following calculations, while the actual values of these coefficients are not as important, since they do not help us understand the instability suppressing mechanism in MoC-ME with non-reflecting b.c.. Thus, we do not calculate the explicit values of \(w_{1,2,3,4}\).

We will now compute the corresponding \(\xi_{1,2(\pm)}(\lambda)\) of \(\rho_{1,2(\pm)}(\lambda)\) in Case I and Case II of Sec. 5.2 separately:

**Case I:** \(\rho \approx \rho_1 = \frac{1}{\lambda}\).

According to the discussion above, we let

\[
\rho = \rho_{1(\pm)} = (1 \pm ibh + w_{1,2}h^2)\frac{1}{\lambda}.
\]

(5.25)

We substitute this \(\rho\) into (5.10), i.e.,

\[
\left(\Gamma_{ME} - \lambda \rho(\lambda)I + \Omega_{ME}\rho^2(\lambda)\right)\xi = 0,
\]

(5.26)

and solve for the eigenvector \(\xi\). The eigenvector \(\xi\) that we are seeking has the following
structure:

$$\xi = \begin{pmatrix} u_0^* + hu_1^* + h^2 u_2^* \\ h v_1^* + h^2 v_2^* \end{pmatrix}, \quad (5.27)$$

where $u_{0,1,2}, v_{1,2}$ are $2 \times 1$ vectors. To simplify notations, we write:

$$\rho = (1 + isbh + w_{1,2}h^2) \frac{1}{\lambda} \equiv \rho_0 + \rho_0 h \tilde{R}_1 + \rho_0 h^2 \tilde{R}_2 \quad (5.28)$$

where $s = \pm 1$, $\rho_0 = \frac{1}{\lambda}$, $\tilde{R}_1 = isb$, $\tilde{R}_2 = w_{1,2}$. Then with the same accuracy:

$$\rho^{-1} = (1 - isbh - (w_{1,2} + b^2)h^2)\lambda$$

$$= (1 - isbh + \hat{w}_{1,2}h^2)\lambda$$

$$= \frac{1}{\rho_0} - \frac{1}{\rho_0} h \tilde{R}_1 + \frac{1}{\rho_0} h^2 \tilde{R}_2,$$  \quad (5.29)

where we used the notation $\tilde{R}_2 = \hat{w}_{1,2} \equiv -(w_{1,2} + b^2)$.

Substituting (5.28) and (5.29) into (5.26), we obtain:

$$M\xi = M\begin{pmatrix} u_0^* + hu_1^* + h^2 u_2^* \\ h v_1^* + h^2 v_2^* \end{pmatrix} = 0, \quad (5.30)$$

where

$$M = \begin{pmatrix} \Omega_{0ME} + h\Omega_{1ME} + h^2\Omega_{2ME} \\ h\Gamma_{0ME} + h\Gamma_{1ME} + h^2\Gamma_{2ME} \end{pmatrix} \begin{pmatrix} \rho_0 + \rho_0 h \tilde{R}_1 + \rho_0 h^2 \tilde{R}_2 \\ \frac{1}{\rho_0} - \frac{1}{\rho_0} h \tilde{R}_1 + \frac{1}{\rho_0} h^2 \tilde{R}_2 \end{pmatrix} - \lambda I \quad (5.31)$$
\( \Omega_{0ME} = \begin{pmatrix} O & O \\ O & I \end{pmatrix} \quad \Gamma_{0ME} = \begin{pmatrix} I & O \\ O & O \end{pmatrix} \)

\( \Omega_{1ME} = \frac{1}{2} (P_S + P_E) \quad \Gamma_{1ME} = \frac{1}{2} (P_N + P_W) \)

\( \Omega_{2ME} = \frac{1}{2} PP_S \quad \Gamma_{2ME} = \frac{1}{2} PP_N, \tag{5.32} \)

and \( P \) was defined in (2.9). For simplicity, in (5.32) we also used the notations defined below (named after North, East, South, and West):

\[
P_N = \begin{pmatrix} P_{11} & P_{12} \\ O & O \end{pmatrix} \quad P_E = \begin{pmatrix} O & P_{12} \\ O & P_{22} \end{pmatrix} \quad P_S = \begin{pmatrix} O & O \\ P_{21} & P_{22} \end{pmatrix} \quad P_W = \begin{pmatrix} P_{11} & O \\ P_{12} & O \end{pmatrix}, \tag{5.33} \]

\[
P_{11} = bA, \quad P_{12} = B, \quad P_{21} = bB, \quad P_{22} = A. \tag{5.34} \]

Collecting the terms of (5.31) according to consecutive orders of \( h \), we can write \( M \) as

\[
M = M_0 + hM_1 + h^2M_2 + O(h^3) \tag{5.35} \]

where

\[
M_0 = \rho_0 \Omega_{0ME} + \frac{1}{\rho_0} \Gamma_{0ME} - \lambda I, \tag{5.36a} \]

141
\[ \mathbb{M}_1 = \rho_0 \tilde{R}_1 \Omega_{0ME} + \rho_0 \Omega_{1ME} - \frac{1}{\rho_0} \tilde{R}_1 \Gamma_{0ME} + \frac{1}{\rho_0} \Gamma_{1ME}, \quad (5.36b) \]

\[ \mathbb{M}_2 = \rho_0 \tilde{R}_2 \Omega_{0ME} + \rho_0 \tilde{R}_1 \Omega_{1ME} + \rho_0 \Omega_{2ME} + \frac{1}{\rho_0} \tilde{R}_2 \Gamma_{0ME} - \frac{1}{\rho_0} \tilde{R}_1 \Gamma_{1ME} + \frac{1}{\rho_0} \Gamma_{2ME}, \quad (5.36c) \]

Substituting (5.35) into (5.30), we obtain:

\[
\left( \mathbb{M}_0 + h\mathbb{M}_1 + h^2\mathbb{M}_2 + O(h^3) \right) \begin{pmatrix} \vec{u}_0 + h\vec{u}_1 + h^2\vec{u}_2 \\ h\vec{v}_1 + h^2\vec{v}_2 \end{pmatrix} = 0 \quad (5.37)
\]

After substituting (5.32), (5.33), (5.34), (5.36) into (5.37), collecting the terms of (5.37) according to consecutive orders of \( h \), and ignoring higher-order terms, we obtain after tedious but straightforward calculations:

\[ O(1): \]

\[
\begin{pmatrix} O & O \\ O & (\frac{1}{\lambda} - \lambda) I \end{pmatrix} \begin{pmatrix} \vec{u}_0 \\ \vec{v}_0 \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{0} \end{pmatrix}, \quad (5.38)
\]

\[ O(h): \]

\[
\begin{pmatrix} \lambda(P_{11} - \tilde{R}_1 I)\vec{u}_0 \\ \frac{1}{2} \left( \frac{1}{\lambda} + \lambda \right) P_{21} \vec{u}_0 + \left( \frac{1}{\lambda} - \lambda \right) \vec{v}_1 \end{pmatrix} = \begin{pmatrix} \vec{0} \\ \vec{0} \end{pmatrix}, \quad (5.39)
\]
\( O(h^2) \):

\[
\begin{align*}
\lambda(P_{11} - \tilde{R}_1 I)\vec{u}_1 + \frac{1}{2}\left( \frac{1}{\lambda} + \lambda \right) P_{12} \vec{v}_1 \\
+ (\lambda\tilde{R}_1 P_{11} + \lambda\tilde{R}_2 I + \frac{1}{2} \lambda P_{11}^2 + \frac{1}{2 \lambda} P_{12} P_{21}) \vec{u}_0 &= \vec{0} \quad (5.40a) \\
\left( \frac{1}{\lambda} - \lambda \right) \vec{v}_2 + \frac{1}{2} \left( \frac{1}{\lambda} + \lambda \right) P_{21} \vec{u}_1 + \frac{1}{\lambda} (P_{22} + \tilde{R}_1 I) \vec{v}_1 \\
+ \left[ \frac{1}{2} \left( \frac{1}{\lambda} - \lambda \right) \tilde{R}_1 P_{21} + \frac{1}{2} \left( \lambda P_{21} P_{11} + \frac{1}{\lambda} P_{22} P_{21} \right) \right] \vec{u}_0 &= \vec{0}. \quad (5.40b)
\end{align*}
\]

The left-hand side of the (5.38) holds identically, regardless of \( \vec{u}_0 \); this justifies choosing the form (5.27). We then determine \( \vec{u}_0, \vec{u}_1, \vec{v}_1 \) and \( \vec{v}_2 \) from (5.39) and (5.40).

Using (5.34) and the fact that \( A^2 = -I, \ B^2 = 2I \), we simplify (5.39) to:

\[
\lambda(bA - \tilde{R}_1 I) \vec{u}_0 = \vec{0}, \quad (5.41a)
\]

\[
- \left( \frac{\lambda^2 - 1}{\lambda^2} \right) \vec{v}_1 + \frac{1}{2} \left( \frac{\lambda^2 + 1}{\lambda^2} \right) bB \vec{u}_0 = \vec{0}. \quad (5.41b)
\]

Similarly, (5.40) becomes:

\[
\begin{align*}
\frac{1}{2} \left( \lambda + \frac{1}{\lambda} \right) B\vec{v}_1 + \lambda(bA - \tilde{R}_1 I) \vec{u}_1 \\
+ \lambda \left( -\tilde{R}_1 bA + \left( \tilde{R}_2 - \frac{b^2}{2} + \frac{b}{\lambda^2} \right) I \right) \vec{u}_0 &= \vec{0}, \quad (5.42a)
\end{align*}
\]
\[
\left(\frac{1}{\lambda} - \lambda\right)\vec{v}_2 + \frac{1}{2}\left(\lambda + \frac{1}{\lambda}\right)bB\vec{u}_1 + \frac{1}{\lambda}(A + \tilde{R}_1I)\vec{v}_1 \\
+ \frac{1}{2}\left(\frac{1}{\lambda}(b\tilde{R}_1B + bAB) + \lambda(-\tilde{R}_1bB + b^2BA)\right)\vec{u}_0 = \vec{0}.
\]

(5.42b)

Following the process of obtaining the solution (4.94), (4.96) in Sec. 4.5.2, we obtain the solution of (5.41):

\[
\vec{u}_0 = \begin{pmatrix} 1 \\ is \end{pmatrix}, \quad \vec{v}_1 = \frac{b(\lambda^2 + 1)}{2(\lambda^2 - 1)}B\vec{u}_0 = \frac{b(\lambda^2 + 1)}{2(\lambda^2 - 1)}\begin{pmatrix} -2is \\ 1 \end{pmatrix}.
\]

(5.43)

Recall that in (5.43), the value of \(s\) depends on which \(\rho(\lambda)\) in (5.23) we choose:

\[
\rho_{1(+)} = \frac{1}{\lambda}(1 + ibh + w_{1,2}h^2) \quad \Rightarrow \quad s = 1, \\
\rho_{1(-)} = \frac{1}{\lambda}(1 - ibh + w_{1,2}h^2) \quad \Rightarrow \quad s = -1.
\]

(5.44)

Notice that the only difference between the solution \(\vec{v}\) in (4.96) (for the MoC-SE) and the solution \(\vec{v}_1\) in (5.43) (for the MoC-ME) is the factor \((\lambda^2 + 1)/2\). Later we will see that it leads to a noticeable quantitative difference for eigenvalues \(\lambda\) satisfying \(\lambda^2 + 1 \approx 0\).

After obtaining \(\vec{u}_0\) and \(\vec{v}_1\), we can obtain the expression for \(\vec{u}_1\) by substituting the expressions for \(\vec{u}_0\) and \(\vec{v}_1\) into (5.42a). Then by substituting the expressions of \(\vec{u}_1, \vec{v}_1, \vec{u}_0\) into (5.42b), we can obtain the expression for \(\vec{v}_2\). In this manner, we can get expressions for all the unknown quantities \(\vec{u}_{0,1,2}, \vec{v}_{1,2}\) in \(\xi\). Thus, using (5.43), the
solution of (5.30) is
\[ \xi = \begin{pmatrix} 1 + h\bar{u}_1 + h^2\bar{u}_2 \\ is \\ \frac{bh}{2} (\frac{\lambda^2 + 1}{(\lambda^2 - 1)}) \begin{pmatrix} -2is \\ 1 \end{pmatrix} + h^2\bar{v}_2 \end{pmatrix} \right) . \] (5.45)

Calculation of the explicit expressions for \( \bar{u}_{1,2}, \bar{v}_2 \) is quite complicated. Fortunately, while the presence of \( \bar{u}_{1,2}, \bar{v}_2 \) is essential in the following calculations, the explicit expression of these vectors are not important for understanding the instability-suppressing mechanism in the MoC-ME with non-reflecting b.c.. As we have said earlier, one can use (5.42) to obtain those explicit expressions if necessary.

**Case II**: \( \rho \approx \rho_2 = \lambda \).

According to (5.24), we let
\[ \rho = \rho_{2(\pm)} = (1 \pm ih + w_{3,4}h^2)\lambda. \] (5.46)

We substitute this \( \rho \) into (5.10) and seek the eigenvector \( \xi \) in the form:
\[ \xi = \begin{pmatrix} hv_1 + h^2\bar{v}_2 \\ u_0 + h\bar{u}_1 + h^2\bar{u}_2 \end{pmatrix}, \] (5.47)

where \( u_{0,1,2}, \bar{v}_{1,2} \) are \( 2 \times 1 \) vectors. To simplify notations, we write:
\[ \rho = (1 + ish + w_{3,4}h^2 + \cdots)\lambda \equiv \rho_0 + \rho_0 h\bar{R}_1 + \rho_0 h^2\bar{R}_2 \] (5.48)
where \( s = \pm 1, \rho_0 = \lambda, \tilde{R}_1 = is, \tilde{R}_2 = w_{3,4} \). Then with the same accuracy:

\[
\rho^{-1} = (1 - ish - (w_{3,4} + 1)h^2) \frac{1}{\lambda}
= (1 - isbh + \tilde{w}_{3,4}h^2) \frac{1}{\lambda}
= \frac{1}{\rho_0} - \frac{1}{\rho_0} h\tilde{R}_1 + \frac{1}{\rho_0} h^2 \tilde{R}_2,
\]

(5.49)

where we used the notation \( \tilde{R}_2 = \tilde{w}_{3,4} \equiv -(w_{3,4} + 1) \). Following the similar steps that leads to (5.30), we obtain:

\[
\mathbf{M} \xi = \mathbf{M} \left( \begin{array}{c}
h\tilde{v}_1 + h^2 \tilde{v}_2 \\
\tilde{u}_0 + hu_1 + h^2 \tilde{u}_2 
\end{array} \right) = \mathbf{0},
\]

(5.50)

where \( \mathbf{M} \) is the has the same form as in (5.31), but with different \( \rho_0, \tilde{R}_1, \tilde{R}_2, \tilde{R}_2 \).

Using the same process by which we obtained (5.38), (5.39), (5.40), we have the following equations with respect to consecutive orders of \( h \):

\( O(1) \):

\[
\left( \begin{array}{cc}
\frac{1}{\lambda} - \lambda & I \\
O & O 
\end{array} \right) \left( \begin{array}{c}
\tilde{0} \\
\tilde{0}
\end{array} \right) = \left( \begin{array}{c}
\tilde{0} \\
\tilde{0}
\end{array} \right),
\]

(5.51)

\( O(h) \):

\[
\left( \begin{array}{c}
\frac{1}{2} \left( \frac{1}{\lambda} + \lambda \right) P_{12} \tilde{u}_0 + \left( \frac{1}{\lambda} - \lambda \right) \tilde{v}_1 \\
\lambda (P_{22} + \tilde{R}_1 I) \tilde{u}_0
\end{array} \right) = \left( \begin{array}{c}
\tilde{0} \\
\tilde{0}
\end{array} \right)
\]

(5.52)
\( \mathcal{O}(h^2) \):

\[
\begin{align*}
\left( \frac{1}{\lambda} - \lambda \right) \vec{v}_2 & + \frac{1}{2} \left( \frac{1}{\lambda} + \lambda \right) P_{12} \vec{u}_1 + \frac{1}{\lambda} (P_{11} - \tilde{R}_1 I) \vec{v}_1 \\
+ \left[ \frac{1}{2} \left( -\frac{1}{\lambda} + \lambda \right) \tilde{R}_1 P_{12} + \frac{1}{2} \left( \lambda P_{12} P_{22} + \frac{1}{\lambda} P_{11} P_{12} \right) \right] \vec{u}_0 &= \vec{0}, \quad (5.53a)
\end{align*}
\]

\[
\begin{align*}
\lambda (P_{22} + \tilde{R}_1 I) \vec{u}_1 & + \frac{1}{2} \left( \frac{1}{\lambda} + \lambda \right) P_{12} \vec{v}_1 \\
+ \left( \lambda \tilde{R}_1 P_{22} + \lambda \tilde{R}_2 I + \frac{1}{2} \lambda P_{22}^2 + \frac{1}{2} \frac{1}{\lambda} P_{21} P_{12} \right) \vec{u}_0 &= \vec{0}. \quad (5.53b)
\end{align*}
\]

The left hand side of the (5.51) holds identically, regardless of \( \vec{u}_0 \); this justifies choosing the form (5.47). We then determine \( \vec{u}_0, \vec{u}_{1,2} \) and \( \vec{v}_{1,2} \) from (5.52) and (5.53).

Using (5.34) and the fact that \( A^2 = -I, \ B^2 = 2I \), we simplify (5.52) to:

\[
- \left( \frac{\lambda^2 - 1}{\lambda^2} \right) \vec{v}_1 + \frac{1}{2} \left( \frac{\lambda^2 + 1}{\lambda^2} \right) B \vec{u}_0 = \vec{0}, \quad (5.54a)
\]

\[
\lambda (A + \tilde{R}_1 I) \vec{u}_0 = \vec{0}. \quad (5.54b)
\]

Similarly, (5.53) becomes:

\[
\begin{align*}
\left( \frac{1}{\lambda} - \lambda \right) \vec{v}_2 & + \frac{1}{2} \left( \lambda + \frac{1}{\lambda} \right) B \vec{u}_1 + \frac{1}{\lambda} (bA - \tilde{R}_1 I) \vec{v}_1 \\
+ \frac{1}{2} \left( \frac{1}{\lambda} (-\tilde{R}_1 B + bAB) + \lambda (\tilde{R}_1 B + BA) \right) \vec{u}_0 &= \vec{0}, \quad (5.55a)
\end{align*}
\]
\[
\frac{1}{2} (\lambda + \frac{1}{\lambda}) bB \bar{v}_1 + \lambda (A + \tilde{R}_1 I) \bar{u}_1 \\
+ \lambda \left( \tilde{R}_1 A + \left( \tilde{R}_2 - \frac{1}{2} + \frac{b}{\lambda^2} \right) I \right) \bar{u}_0 = \bar{0}.
\] (5.55b)

Following the process of obtaining the solution (4.100), (4.101) in Sec. 4.5.2, we obtain the solution of (5.54):

\[
\bar{u}_0 = \begin{pmatrix} 1 \\ -is \end{pmatrix}, \quad \bar{v}_1 = \frac{(\lambda^2 + 1)}{2(\lambda^2 - 1)} B \bar{u}_0 = \frac{(\lambda^2 + 1)}{2(\lambda^2 - 1)} \begin{pmatrix} 2is \\ -1 \end{pmatrix},
\] (5.56)

Recall that in (5.56), the value of \( s \) depends on which \( \rho(\lambda) \) in (5.24) we choose:

\[
\rho_{2(+)} = \lambda(1 + ibh + w_{3,4}h^2) \quad \Rightarrow \quad s = 1, \\
\rho_{2(-)} = \lambda(1 - ibh + w_{3,4}h^2) \quad \Rightarrow \quad s = -1.
\] (5.57)

Notice that the only difference between the solution \( \bar{v} \) in (4.101) (for the MoC-SE) and the solution \( \bar{v}_1 \) in (5.56) (for the MoC-ME) is the factor \((\lambda^2 + 1)/2\). Later we will see that it leads to a noticeable, quantitative difference for eigenvalues \( \lambda \) satisfying \( \lambda^2 + 1 \approx 0 \).

After obtaining \( \bar{u}_0 \) and \( \bar{v}_1 \), we can obtain the expression for \( \bar{u}_1 \) by substituting the expressions for \( \bar{u}_0 \) and \( \bar{v}_1 \) into (5.55b). Then by substituting the expressions of \( \bar{u}_1, \bar{v}_1, \bar{u}_0 \) into (5.55a) and solve it, we can obtain the expression for \( \bar{v}_2 \). In this manner, we can get expressions for all the unknown quantities \( \bar{u}_{0,1,2}, \bar{v}_{1,2} \) in \( \xi \). Thus, using (5.56),
the solution of (5.50) is
\[ \xi = \left( \begin{array}{c} \frac{h}{2} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \begin{pmatrix} 2i \lambda \\ -1 \end{pmatrix} + h^2 \vec{v}_2 \\ 1 \\ -i \lambda \end{array} \right). \]  
(5.58)

Calculation of the explicit expressions for \( \vec{u}_{1,2}, \vec{v}_2 \) is quite complicated. Fortunately, while the presence of \( \vec{u}_{1,2}, \vec{v}_2 \) is essential in the following calculations, the explicit expression of these vectors are not important for understanding the instability-suppressing mechanism in the MoC-ME with non-reflecting b.c.. As we have said earlier, one can use (5.55) to obtain those explicit expressions if necessary.

### 5.4 Finding the approximate value of \( |\lambda| \)

This process is similar to that in Sec. 4.5.3. Namely, we will use (5.11) and (5.12) to compute the eigenvalue \( \lambda \). As we explained at the end of Sec. 5.1, we need to use the \( \xi(\lambda) \) that we calculated in Sec. 5.3 to construct the matrix \( \Upsilon(\lambda) \) in (5.11). Then we can obtain \( \lambda \) by solving (5.11) for \( \lambda \); the numerical stability of the MoC-ME with non-reflecting b.c. will then be inferred from whether the magnitude \( |\lambda| \) is greater or less than 1.
As in Sec. 4.5.3, we have

\[(y)_m = C_1 \rho_{11(+)}^m \xi_{1(+)} + C_2 \rho_{11(-)}^m \xi_{1(-)} + C_3 \rho_{22(+)}^m \xi_{2(+)} + C_4 \rho_{22(-)}^m \xi_{2(-)}, \tag{5.59}\]

where \(\rho_{1,2(\pm)}\) are the four \(\rho_k\) that we obtained in Sec. 5.2; \(\xi_{1,2(\pm)}\) are the four \(\xi_k\) that we obtained in Sec. 5.3, and the constants \(C_{1,2,3,4}\) are to be determined by imposing the non-reflecting b.c.

For the reader’s convenience, we write the explicit expressions of the \(\xi_{1,2(\pm)}\) below:

\[
\xi_{1(+)} = \begin{pmatrix} 1 \\ i \end{pmatrix} + h \bar{u}_{1,1(+)} \quad \hat{\xi}_{1(+)} = \begin{pmatrix} 1 \\ -i \end{pmatrix} + h \bar{u}_{1,1(-)}
\]

\[
\xi_{1(-)} = \begin{pmatrix} \frac{bh}{2} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \\ -i \end{pmatrix} + h^2 \hat{\xi}_{1(-)}
\]

\[
\xi_{2(+)} = \begin{pmatrix} \frac{b}{2} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \\ 1 \end{pmatrix} + h^2 \hat{\xi}_{2(+)}
\]

\[
\xi_{2(-)} = \begin{pmatrix} \frac{b}{2} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \\ -i \end{pmatrix} + h^2 \hat{\xi}_{2(-)}
\]

\[
\xi_{2(\pm)} = \begin{pmatrix} \bar{u}_{2,1(\pm)} \\ \bar{v}_{2,1(\pm)} \end{pmatrix}, \quad \hat{\xi}_{2(\pm)} = \begin{pmatrix} \bar{v}_{2,2(\pm)} \\ \bar{u}_{2,2(\pm)} \end{pmatrix}. \tag{5.60}
\]

As we discussed in the previous section, we will not give the explicit expressions of \(\bar{u}_{1,2}, \bar{v}_2\). Recall that (+) denotes that the quantity corresponds to \(s = 1\), and
\((-\) denotes that the quantity corresponds to \(s = -1\). For example, \(\xi_{1(+)}\) and \(\vec{u}_{2,1(+)}\) correspond to \(\rho_{1(+)}\). Following notations in (4.41), these \(\xi\) eigenvectors can be written as the following \(2 \times 1\)-block form, which is important in our next step of constructing the \(\Upsilon\) matrix:

\[
\hat{\xi}_{i(\pm)} = \begin{pmatrix} \xi_{i(\pm)}^+ \\ \xi_{i(\pm)}^- \end{pmatrix}, \quad i = 1, 2. \tag{5.62}
\]

Note that here, the superscripts \(\pm\) are in no way related to the subscripts \((\pm)\).

We now show how to construct the matrix \(\Upsilon(\lambda)\) using notations (5.62). The non-reflecting b.c. initially stated in (4.44) are as follows:

\[
C_1 \xi_{1(+)1}^++ C_2 \xi_{1(-)}^- + C_3 \xi_{2(+)2}^+ + C_4 \xi_{2(-)}^- = 0, \tag{5.63a}
\]

\[
C_1 \xi_{1(+)1}^- \rho_{1(+)}^M + C_2 \xi_{1(-)}^- \rho_{1(-)}^M + C_3 \xi_{2(+)2}^- \rho_{2(+)}^M + C_4 \xi_{2(-)}^- \rho_{2(-)}^M = 0. \tag{5.63b}
\]

Together, (5.63a) and (5.63b) give the following linear system:

\[
\Upsilon(\lambda) [C_1, C_2, C_3, C_4]^T = 0. \tag{5.64}
\]

Substituting (5.60) and (5.62) into (5.63), we can see that the matrix \(\Upsilon\) has the following block form:

\[
\Upsilon = \begin{pmatrix} \Upsilon_A & \Upsilon_B \\ \Upsilon_C & \Upsilon_D \end{pmatrix}, \tag{5.65}
\]

151
where

\[ \Upsilon_A = \left( \begin{array}{cc} 1 & 1 \\ i & -i \end{array} \right) + \hbar \left[ \bar{u}_{1,1(+)} \bar{u}_{1,1(-)} \right] + \hbar^2 \left[ \hat{\xi}_{1(+)}^+, \hat{\xi}_{1(-)}^+ \right] \]

\[ \equiv \Upsilon_{A0} + \hbar \Upsilon_{A1} + \hbar^2 \Upsilon_{A2} \equiv \Upsilon(\Upsilon_{A1} + \hbar \Upsilon_{A2}) \]

\[ \equiv \Upsilon_{A0} + \hbar \hat{\Upsilon}_A, \quad (5.66a) \]

\[ \Upsilon_C = \frac{\hbar}{2\lambda M} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \left( \begin{array}{cc} -2i\hat{\rho}_{1(+)}^M & 2i\hat{\rho}_{1(-)}^M \\ -\hat{\rho}_{1(+)}^M & -\hat{\rho}_{1(-)}^M \end{array} \right) + \frac{\hbar^2}{\lambda M} \left[ \hat{\xi}_{1(+)}^-, \hat{\xi}_{1(-)}^- \right] \]

\[ \equiv \frac{\hbar}{2\lambda M} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \Upsilon_{C0} + \frac{\hbar^2}{\lambda M} \hat{\Upsilon}_C, \quad (5.66b) \]

\[ \Upsilon_B = \frac{\hbar}{2} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \left( \begin{array}{cc} 2i & -2i \\ -1 & -1 \end{array} \right) + \hbar^2 \left[ \hat{\xi}_{2(+)}^+, \hat{\xi}_{2(-)}^+ \right] \]

\[ \equiv \frac{\hbar}{2} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \Upsilon_{B0} + \hbar^2 \hat{\Upsilon}_B \quad (5.66c) \]

\[ \Upsilon_D = \lambda^M \left( \begin{array}{cc} \hat{\rho}_{2(+)}^M & \hat{\rho}_{2(-)}^M \\ -i\hat{\rho}_{2(+)}^M & i\hat{\rho}_{2(-)}^M \end{array} \right) + \lambda^M \left[ \hat{\rho}_{2(+)}^M \bar{u}_{1,2(+)} \bar{u}_{1,2(-)} \right] + \hbar^2 \lambda^M \left[ \hat{\xi}_{2(+)}^-, \hat{\xi}_{2(-)}^- \right]. \]

\[ \equiv \lambda^M \Upsilon_{D0} + \hbar \lambda^M \Upsilon_{D1} + \hbar^2 \lambda^M \Upsilon_{D2} \]

\[ = \lambda^M \Upsilon_{D0} + \hbar \lambda^M (\Upsilon_{D1} + \hbar \Upsilon_{D2}) \]

\[ \equiv \lambda^M \Upsilon_{D0} + \hbar \lambda^M \hat{\Upsilon}_D \quad (5.66d) \]

152
In the equations above, we have denoted

\[ \hat{\rho}_1(\pm) = 1 \pm ibh + w_{1,2}h^2, \quad \hat{\rho}_2(\pm) = 1 \pm ih + w_{3,4}h^2; \]  

(5.67)

see (5.23), (5.24) and the text after those equations.

The block form of \( \Upsilon \), i.e., (5.65), shows that \( \Upsilon \) has the structure:

\[
\Upsilon = \begin{pmatrix}
\Upsilon_{A0} + h\hat{\Upsilon}_A & \frac{h}{2} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \Upsilon_{B0} + h^2\hat{\Upsilon}_B \\
\frac{bh}{\lambda^{M'}} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \Upsilon_{C0} + \frac{h^2}{\lambda^{M'}} \hat{\Upsilon}_C & \lambda^M \Upsilon_{D0} + h\lambda^{M'}\hat{\Upsilon}_D
\end{pmatrix}
\]  

(5.68)

where \( \Upsilon_{A0}, \Upsilon_{B0}, \Upsilon_{C0}, \Upsilon_{D0}, \) and \( \hat{\Upsilon}_A, \hat{\Upsilon}_B, \hat{\Upsilon}_C, \hat{\Upsilon}_D \) are matrix blocks that are defined above in (5.66). \( \Upsilon_{A0}, \Upsilon_{B0}, \Upsilon_{C0}, \Upsilon_{D0} \) are constant blocks whose explicit expressions are known. \( \hat{\Upsilon}_A, \hat{\Upsilon}_B, \hat{\Upsilon}_C, \hat{\Upsilon}_D \) are blocks that may contain \( \lambda \), and they are determined by \( \vec{u}_{1,i(\pm)}, \vec{u}_{2,i(\pm)}, \vec{v}_{2,i(\pm)}, i = 1, 2 \), which can be found as explained in Sec. 5.3.

From (5.68), one can, in principle, calculate the determinant of \( \Upsilon(\lambda) \). However, before we proceed to do so, let us consider the magnitude of the number \( \lambda^2 + 1 \). According to the previous discussion, there are two cases depending on the argument of \( \lambda \). In the first case, \( \arg(\lambda) \not\approx \pm \pi/2 \), which implies that \( \lambda^2 + 1 = O(1) \). In this case, in the main approximation, one will find that \( |\lambda|^M = O(h) \), as in Sec. 4.5. We will not repeat these calculations here. In the second case, \( \arg(\lambda) \approx \pm \pi/2 \) (i.e., \( \lambda^2 \approx -1 \)). In this case we can consider a small interval of eigenvectors for which

\[ \lambda^2 + 1 = O(h) \ll 1. \]

(5.69)

Then we have a different estimate for \( |\lambda|^M \), as shown later. We will now focus on obtaining this new estimate.
Given (5.69), in (5.66b) we have
\[ \frac{bh}{2\lambda^M} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \gamma_{C0} = O \left( \frac{h^2}{\lambda^M} \right), \]  
(5.70)
and in (5.66c), we have
\[ \frac{h}{2} \left( \frac{\lambda^2 + 1}{\lambda^2 - 1} \right) \gamma_{B0} = O \left( h^2 \right). \]  
(5.71)

Therefore, by factoring out the common power of \( h \) in (5.66b) and (5.66c), we may let
\[ \tilde{\gamma}_C = \frac{h^2}{\lambda^M} \tilde{\gamma}_C, \quad \tilde{\gamma}_B = h^2 \tilde{\gamma}_B, \quad \tilde{\gamma}_C, \tilde{\gamma}_B = O(1). \]  
(5.72)

Note that the calculation of explicit expressions of \( \tilde{\gamma}_c, \tilde{\gamma}_B \) is not feasible, not only because of the technical difficulties of computing \( \hat{\gamma}_C, \hat{\gamma}_B \) in (5.66), but also because those expressions will substantially change as \( \lambda^2 + 1 \) varies near 0. Therefore, below we can only determine the order of magnitude of \( |\lambda|^M \), whereas in Chap. 4, we were able to also determine proportionality coefficients.

We now calculate the determinant \( |\gamma(\lambda)| \) by using formulas (4.111) and (5.65):
\[ |\gamma(\lambda)| = |\gamma_A| \left| \gamma_D - \gamma_C \gamma_A^{-1} \gamma_B \right|. \]  
(5.73)

Substituting (5.66) and (5.72) into (5.73), we can write \( |\gamma(\lambda)| = 0 \) as
\[ \left| \gamma_{A0} + h\hat{\gamma}_A \right| \left| \lambda^M (\gamma_{D0} + h\hat{\gamma}_D) - \frac{h^2}{\lambda^M} \tilde{\gamma}_C (\gamma_{A0} + h\hat{\gamma}_A)^{-1} \cdot h^2 \tilde{\gamma}_B \right| = 0, \]  
(5.74)

154
which is equivalent to

\[
|\lambda^M (\Upsilon D_0 + h\hat{\Upsilon}_D) - \frac{h^2}{\lambda^M} \hat{\Upsilon}_C (\Upsilon A_0 + h\hat{\Upsilon}_A)^{-1} \cdot h^2 \hat{\Upsilon}_B| = 0. \tag{5.75}
\]

Notice that

\[
(\Upsilon A_0 + h\hat{\Upsilon}_A)^{-1} = \Upsilon A_0^{-1} + h\hat{\Upsilon}_A, \quad \Upsilon A_0^{-1} = \begin{pmatrix} 1/2 & -i/2 \\ i/2 & 1/2 \end{pmatrix}, \quad \hat{\Upsilon}_A = O(1). \tag{5.76}
\]

After substituting (5.76) into (5.75) and simplification, (5.75) can be written as

\[
|\lambda^M \Upsilon D_0 - \frac{h^4}{\lambda^M} \hat{\Upsilon}_C \Upsilon A_0^{-1} \hat{\Upsilon}_B + \lambda^M h\hat{\Upsilon}_D - \frac{h^5}{\lambda^M} \hat{\Upsilon}_C \hat{\Upsilon}_A \hat{\Upsilon}_B| = 0
\]

\[
\Rightarrow |\lambda^M \Upsilon D_0 - \frac{h^4}{\lambda^M} \hat{\Upsilon}_C \Upsilon A_0^{-1} \hat{\Upsilon}_B + O(\lambda^M h) + O \left( \frac{h^5}{\lambda^M} \right)| = 0
\]

\[
\Rightarrow |\lambda^2 \Upsilon D_0 - h^4 \hat{\Upsilon}_C \Upsilon A_0^{-1} \hat{\Upsilon}_B + O(h^2) + O \left( \frac{h^5}{h^4} \right)| = 0. \tag{5.77}
\]

If we now assume that

\[
|\lambda|^{2M} = O(h^4), \tag{5.78}
\]

then the first two terms in (5.77) can balance each other in the sense that the matrix

\[
\left( \lambda^{2M} \Upsilon D_0 - h^4 \hat{\Upsilon}_C \Upsilon A_0^{-1} \hat{\Upsilon}_B \right)
\]

can be singular. Moreover, the last two terms in (5.77) are then of higher order in \(h\) and hence can be ignored. Thus, (5.78) is the main conclusion of this section, and in the next section, we will verify this conclusion via numerical simulations.
5.5 Numerical verification of $\lambda^{2M} = O(h^4)$

Here we follow the outline of the presentation in Sec. 4.5.4. From Sec. 5.4, we have the theoretical result $|\lambda^{2M}| = O(h^4)$ for those eigenvectors where $\arg(\lambda) \approx \pi/2$. This implies that

$$|\lambda|^{2M} = C h^4 \Rightarrow \ln(|\lambda|^{2M}) = 4 \ln(h) + \ln(C), \quad (5.79)$$

where $C$ is either a constant or a slowly changing function of $h$ (see the text after (4.144)). Therefore, the slope of the graph of $\ln(|\lambda|^{2M})$ vs $\ln(h)$ should be approximately 4.

To verify that, we need to first calculate the value of the numerical $|\lambda|$, i.e., $|\lambda_{num}|$, for each value of $h$. For that, we use the same approach as in Sec. 4.5.4. Notice that in this approach, the half width of the “box” in (4.138) can affect the result. This is illustrated in Fig. 5.1. To calculate the evolution of the FFT error, defined in Sec. 4.5.4, we need to use the data for wave numbers in some vicinity of the (normalized) wave number $z = \pi/2$ (see the first paragraph of this subsection). Notice that the FFT error decreases faster when we choose a smaller vicinity (illustrated by the solid box in Fig. 5.1) around $\pi/2$, while it does so slower for a bigger vicinity (dashed box in Fig. 5.1).

Therefore, unlike in Sec. 4.5.4, we take the half width of the “box”, defined in (4.138), to be 15%, 10%, 5%, 2.5% of the number of grid points (i.e., $M$) and calculate the corresponding $|\lambda_{num}|$. Notice that the initial perturbation in the numerical simulation is generated by using random function in MATLAB, and hence is noisy.
Figure 5.1: Fourier spectra for $L = 25, h = 0.05$. Panel (a) is the spectrum when $t = 25$; panel (b) is the spectrum when $t = 50$; panel (c) is when $t = 75$. The box with solid line illustrates the half-width of 2.5%; the box with dashed line illustrates the half-width of 10%. The horizontal dash-dotted lines indicate the magnitude of the spectrum at grid point $m = M/4$ and at grid points $m = M/4 \pm M/10$.

To smoothen the data, we ran the numerical simulation 10 times using different seeds of the random number generator in MATLAB and then took the average of these 10 simulations. After calculating the values of $|\lambda_{num}|$ corresponding to the same values of $h$ as in Sec. 4.5.4, we find the dependence of $\ln \left( |\lambda|^2M \right)$ on $\ln(h)$ to be as shown in Fig. 5.2. The slopes we obtained in these numerical simulations depend on the half width of the "box". For larger half-box-sizes, i.e., 15%, 10%, the slopes are close to 2, while for smaller half-box-sizes, i.e., 5%, 2.5%, the slopes are close to 4. Since our theoretical result pertains to the (normalized) wave numbers that are very close $\pi/2$, then the smaller vicinity around $\pi/2$ we take, the more accurate the result should be. Indeed, the results for the "box"'s half-widths of 2.5% and 5% of M, reported in Fig. 5.2, confirms that the slope is indeed approximately 4. This verifies our main theoretical result of this chapter.
5.6 Explanation of stabilizing role of non-reflecting b.c.

Comparing the (in)stability features of MoC-ME with periodic and non-reflecting b.c., we find an unexpected result. Namely, according to common knowledge in textbooks [15] [16], the von Neumann stability, i.e. the stability of the scheme with periodic b.c., is necessary for the “general” stability of a scheme. That is, an instability for periodic b.c. should imply instability for any other b.c., including the non-reflecting b.c. considered in this chapter. In other words, non-periodic b.c. can only “make things worse” according to the common knowledge.

In Chap. 3, we found that when we apply periodic b.c., we observe strong nu-
numerical instability at wave number $kh \approx \pi/2$. However, in numerical simulations with non-reflecting b.c., considered in this Chapter, we did not detect any numerical instability. To support this observation, our analytical results in Chap. 5 show that non-reflecting b.c. suppress the numerical instability. So the MoC-ME with non-reflecting b.c. is stable even though its corresponding von Neumann analysis predicts instability. This disagree with the common knowledge; that is, non-reflecting boundary condition “can make things better”.

According to (3.1.2) and (4.18), we know that

$$ (Y)^n_j = \lambda^n v_j $$

(5.80)

where $(Y)^n_j$ is a perturbation mode at time $t_n = n\Delta t$, and $v_j$ is an eigenvector of the difference equation operator matrix $\Phi_{ME}$ (see (4.17), (5.8)). This implies that the shape of eigenvector $v_j$ decides the shape of its corresponding growing (i.e., unstable) perturbation $(Y)^n_j$.

In textbooks, it is assumed that there are three shapes of eigenvector: one that vanishes on the right (left b.c. problem); one that vanishes on the left (right b.c. problem), and one that is oscillating (periodic b.c. problem). We plot these three shapes in the left panel of Fig. 5.3.

Suppose that von Neumann analysis predicts instability, then it means that the perturbation proportional to the eigenvector of the periodic problem will grow, and an instability will develop. So as long as the von Neumann instability analysis predicts an instability, the scheme is unstable, regardless of whether the perturbations proportional to the eigenvectors of the right(left) b.c. problem will grow or not.

However, in our problem, the eigenvectors have only two kinds of profile that are
shown in the right panel of Fig. 5.3. Importantly, there is no eigenvector corresponding to the periodic problem. Thus, even if the von Neumann analysis for the periodic problem predicts instability, it does not imply that the perturbations proportional to these eigenvectors will grow. That is why our result contradicts that stated in textbooks.

Figure 5.3: Schematically illustrating the spatial dependence, \((\rho, \rho^2, \ldots, \rho^M)\) (see (4.36)), of the eigenvectors of the problem (4.17) and (5.8) vs \(x\). The left (right) panel is the case described in textbooks (this chapter).
Chapter 6

Summary

We considered the numerical (in)stability of the numerical MoC with three ODE solvers, SE, ME, LF, when it simulates an energy-preserving hyperbolic PDE system (1.3). The corresponding numerical schemes are (3.1), (3.24), (3.34). Using these schemes with periodic and non-reflecting b.c., which are introduced in Sec. 1.3, we discovered two unexpected results.

Firstly, we found difference in the MoC-ME with periodic b.c. (see Chap. 3) and non-reflecting b.c. (see Chap. 5). Textbooks state that the stability of the scheme with periodic b.c., is necessary for the “general” stability of a scheme. That is, an instability for periodic b.c. should imply instability for any other b.c., including the non-reflecting b.c. considered in this chapter. In other words, non-periodic b.c. can only “make things worse” according to the common knowledge. However, in Chap. 3, we found that when we apply periodic b.c., we observe strong numerical instability, while in numerical simulations with non-reflecting b.c., considered in this chapter, we did not detect any numerical instability. So, contrary to the textbook statement, we found a situation when non-periodic b.c. can “make things better”.

161
The first unexpected result is studied and explained in two steps. The first step is to analyze the stability of the MoC-ME with periodic b.c., for which we used the standard von Neumann method in Sec. 3.2. (Other parts of Chap. 3 were devoted to the von Neumann analysis of the MoC-SE and MoC-LF). We also verified these theoretical results by direct numerical simulations. Important for the following will be the fact that for the MoC-ME, the most unstable modes occur at (normalized) wavenumbers close to $\pm \pi/2$. (see Figs. 3.4 and 3.5).

For the next step of the explanation, we had to perform a stability analysis for non-periodic (in our case). It should be pointed out that we have founded only one instance of a related analysis for the MoC in the published literature [1]. Therefore, we first needed to develop the corresponding analysis that could be applied to our situation (that is, the type of the PDE and the type of b.c.). We did this in Chap. 4 for the MoC-SE. This is not the method whose stability we intended to analyze, but it was simpler and thus allowed us to fully understand the concepts of the method. Then, in Chap. 5, we applied this method to the MoC-ME. This allowed us to explain, in Sec. 5.4, why numerical instability does not occur when non-reflecting b.c. are used. In Sec. 5.5 we verified some quantitative predictions of this analysis by direct numerical simulations. In Sec. 5.6 we explained why our result contradicts that common statement in textbooks, as described at the beginning of this Summary. In Sec. 4.6 we explained why the non-reflecting b.c. do not suppress numerical instability of the MoC-SE, in contrast to what occurs for the MoC-ME.

Our second unexpected result came from the von Neumann stability analysis of the MoC-LF. Thus, here we refer to periodic b.c.. Notice that we consider an energy-preserving PDE system (see (1.10)). The LF scheme is known to work well for energy
preserving ODEs (see Sec. 1.4), However, the MoC-LF has turned out to have stronger instability than the MoC-based non-energy-preserving Euler methods, i.e. the MoC-SE and MoC-ME (Figs. 3.1, 3.4, 3.6). More specifically, we found that the MoC-LF has a strong instability only at the normalized wavenumbers \( \approx \pi/2 \), while there is no instability at other wavenumbers. We qualitatively explained this instability in Sec. 3.3.3. The reason is a certain “phase distortion matrix” \( Q \), which deviates from the identity matrix the most when the normalized wavenumber is \( \pi/2 \).

Our numerical results also show that no matter whether we use the periodic or non-reflecting b.c., the strong instability of MoC-LF is still present. We have not yet found an analytical explanation for this unexpected result. An interesting topic for a future work will be to study the instability of MoC-LF with non-reflecting boundary condition.
Appendix A

Derivation of (4.61)

As we discussed in Sec. 4.5.1, in order to obtain the expression of $\rho(\lambda)$, we need to solve

$$E_0(\rho, \lambda) + h^2 E_1(\rho, \lambda) = 0,$$

(A.1)

where $E_0(\rho, \lambda)$, $E_1(\rho, \lambda)$ are given in (4.53) and (4.54).

In Case 1 of Sec. 4.5.1, we obtained two sets of approximate solutions for $\rho(\lambda)$, one close to $\rho_1 = \frac{1}{\lambda}$ and the other close to $\rho_2 = \lambda$. We found them by adding perturbation terms $\tilde{\rho}_1$, $\tilde{\rho}_2$ ($|\tilde{\rho}_1| \ll 1$, $|\tilde{\rho}_2| \ll 1$) and substituting $\rho = \rho_1 + \tilde{\rho}_1$, or $\rho = \rho_2 + \tilde{\rho}_2$ into (A.1) to determine $\tilde{\rho}_1$ and $\tilde{\rho}_2$. According to the calculations in Sec. 4.5.1, the results are:

$$\rho_{1(\pm)} = \rho_1 + \tilde{\rho}_1 = \frac{1}{\lambda}(1 \pm ibh),$$

$$\rho_{2(\pm)} = \rho_2 + \tilde{\rho}_2 = \lambda(1 \pm ih).$$

(A.2)
To obtain more accurate approximations, we need to add even smaller perturbation terms to (A.2):

\[
\rho_1(\pm) = \rho_1 + \tilde{\rho}_1 = \frac{1}{\lambda} \pm \frac{ibh}{\lambda} + \tilde{\rho}_1,
\]

\[
\rho_2(\pm) = \rho_2 + \tilde{\rho}_2 = \lambda \pm ih\lambda + \tilde{\rho}_2,
\]

(A.3)

where \(\tilde{\rho}_{1,2}\) are terms with a higher order in \(h\) than \(\tilde{\rho}_{1,2}\). To determine \(\tilde{\rho}_{1,2}\), we substitute (A.3) into (A.1) and obtain:

\[
E_0 \left( \rho_1 + \tilde{\rho}_1 + \tilde{\rho}_1 \right) + h^2 E_1 \left( \rho_1 + \tilde{\rho}_1 + \tilde{\rho}_1 \right) = E_0 \left( \frac{1}{\lambda} \pm \frac{ibh}{\lambda} + \tilde{\rho}_1, \lambda \right) + h^2 E_1 \left( \frac{1}{\lambda} \pm \frac{ibh}{\lambda} + \tilde{\rho}_1, \lambda \right)
\]

\[
= \pm \frac{2ibh}{\lambda} (\lambda^2 - 1)^2 \tilde{\rho}_1 + O(\tilde{\rho}_1 h^2) + O(\tilde{\rho}_1^2) + h^3 \frac{4ib^2(\lambda^2 - 1)}{\lambda^2} + O(h^4) = 0
\]

(A.4)

and

\[
E_0 \left( \rho_2 + \tilde{\rho}_2 + \tilde{\rho}_2 \right) + h^2 E_1 \left( \rho_2 + \tilde{\rho}_2 + \tilde{\rho}_2 \right) = E_0 \left( \lambda \pm ih\lambda + \tilde{\rho}_2, \lambda \right) + h^2 E_1 \left( \lambda \pm ih\lambda + \tilde{\rho}_2, \lambda \right)
\]

\[
= \pm 2ih\lambda(\lambda^2 - 1)^2 \tilde{\rho}_2 + O(\tilde{\rho}_2 h^2) + O(\tilde{\rho}_2^2) + h^3 \cdot 2\lambda^2(\lambda^2 - 1)(\lambda^2 + 2b - 1) + O(h^4) = 0.
\]

(A.5)

We will proceed by ignoring terms in (A.4) and (A.5) that we have denoted by the \(O\)-symbol. This will be justified by the magnitude of \(\tilde{\rho}_{1,2}\) that we will obtain. Thus,
we find:

\[ \pm \frac{2ibh}{\lambda} (\lambda^2 - 1)^2 \tilde{\rho}_1 \mp h^3 \frac{4ib^2(\lambda^2 - 1)}{\lambda^2} = 0, \]  

(A.6)

\[ \pm 2ih\lambda(\lambda^2 - 1)^2 \tilde{\rho}_2 \pm h^3 \cdot 2\lambda^2(\lambda^2 - 1)(\lambda^2 + 2b - 1) = 0, \]  

(A.7)

which give the solutions

\[ \tilde{\rho}_1 = \frac{2b}{\lambda(\lambda^2 - 1)} h^2, \quad \tilde{\rho}_2 = -\frac{(\lambda^2 + 2b - 1)\lambda}{(\lambda^2 - 1)} h^2. \]  

(A.8)

Therefore, we obtain the more accurate approximations (4.61):

\[ \rho_{1(\pm)} = \frac{1}{\lambda} \left( 1 \pm i\lambda h + \frac{2b}{\lambda^2 - 1} h^2 \right) \]

\[ \rho_{2(\pm)} = \lambda \left( 1 \pm i\lambda - \frac{(\lambda^2 + 2b - 1)\lambda}{\lambda^2 - 1} h^2 \right). \]  

(A.9)
Appendix B

Derivation of (4.116)

To simplify the derivation, we break it into two steps. First, we will use the approximations of $\rho(\lambda)$ up to $O(h)$, i.e., (4.58), (4.60), to calculate $\hat{\rho}_1(\pm), \hat{\rho}_2(\pm)$. Then we use the $O(h^2)$ terms in (4.61), found in Appendix A, to sharpen these results, thereby obtaining the expressions in (4.116).

According to (4.58), (4.60), and (4.62), we have

$$\hat{\rho}_1(\pm) = 1 \pm ibh, \quad \hat{\rho}_2(\pm) = 1 \pm ih. \tag{B.1}$$

In what follows, we choose $b = -1$, because explicit analytical results can be found
only in this case. Then, according to (B.1), we have \( \hat{\rho}_{1(\pm)} = \hat{\rho}_{2(\pm)} \). Therefore,

\[
\begin{align*}
\hat{\rho}_{1(+)} &= \frac{\hat{\rho}_{1(+)}}{\hat{\rho}_{1(-)}} = \frac{1 - ih}{1 + ih} = \frac{\sqrt{1 + h^2} e^{-i\theta}}{\sqrt{1 + h^2} e^{i\theta}} = e^{-2i\theta} \\
\hat{\rho}_{2(+)} &= \frac{\hat{\rho}_{2(+)}}{\hat{\rho}_{1(-)}} = 1 \\
\hat{\rho}_{1(-)} &= \frac{\hat{\rho}_{1(-)}}{\hat{\rho}_{1(+)}} = 1 \\
\hat{\rho}_{2(-)} &= \frac{\hat{\rho}_{2(-)}}{\hat{\rho}_{1(+)}} = \frac{1 + ih}{1 - ih} = \frac{\sqrt{1 + h^2} e^{i\theta}}{\sqrt{1 + h^2} e^{-i\theta}} = e^{2i\theta},
\end{align*}
\]

where \( \theta \) is the argument of \( 1 + ih \). Notice that \( \theta \approx h \) since \( h \ll 0 \).

Now we account for the \( O(h^2) \) terms in the approximations of \( \hat{\rho}_{1(\pm)} \) and \( \hat{\rho}_{2(\pm)} \), i.e., use the more accurate approximations from (4.61):

\[
\begin{align*}
\hat{\rho}_{1(\pm)} &= 1 \pm i b h + (2b/(\lambda^2 - 1)) h^2, \\
\hat{\rho}_{2(\pm)} &= 1 \pm i h + (-1 - 2b/(\lambda^2 - 1)) h^2. \tag{B.3}
\end{align*}
\]

Since we are considering the case where \( b = -1 \) and \( \lambda^2 \approx -1 \) (see text after (4.136)), we have

\[
\begin{align*}
\hat{\rho}_{1(\pm)} &= 1 \mp ih + h^2, \\
\hat{\rho}_{2(\pm)} &= 1 \pm ih - 2h^2. \tag{B.4}
\end{align*}
\]
Therefore, as \( h \to 0 \), we have

\[
\left( \frac{\hat{p}_1(+) \hat{p}_2(\pm)}{\hat{p}_2(\pm)} \right)^M = \left( \frac{1 - ih + h^2}{1 + ih - 2h^2} \right)^M = \frac{(1 + h^2)^M (1 - \frac{ih}{1 + h^2})^M}{(1 - 2h^2)^M (1 + \frac{ih}{1 - 2h^2})^M}
\]

\[
= \frac{(1 + h^2)^{L/h} \left( 1 - \frac{ih}{1 + h^2} \right)^{L/h} \approx \exp \{ Lh \} \exp \left\{ -i \left( \frac{L}{1 + h^2} \right) \right\}}{(1 - 2h^2)^{L/h} \left( 1 + \frac{ih}{1 - 2h^2} \right)^{L/h} \approx \exp \{ -2Lh \} \exp \left\{ i \left( \frac{L}{1 - 2h^2} \right) \right\}}
\]

\[
= e^{3lh} \frac{\exp \{ -iL + O(Lh^2) \}}{\exp \{ iL + O(Lh^2) \}} \approx e^{3lh} e^{-2iL} \approx e^{3lh} e^{-2iM\theta}, \quad (B.5)
\]

where we ignored the \( O(Lh^2) \)-terms which has high order in \( h \), and used the fact that

\[
\theta \approx h \quad \Rightarrow \quad M\theta \approx Mh = L. \quad (B.6)
\]

Similarly, we have

\[
\left( \frac{\hat{p}_1(-) \hat{p}_2(\pm)}{\hat{p}_2(\pm)} \right)^M = \left( \frac{1 + ih + h^2}{1 + ih - 2h^2} \right)^M = \frac{(1 + h^2)^M (1 + \frac{ih}{1 + h^2})^M}{(1 - 2h^2)^M (1 + \frac{ih}{1 - 2h^2})^M}
\]

\[
\approx \frac{\exp \{ Lh \} \exp \left\{ i \left( \frac{L}{1 + h^2} \right) \right\}}{\exp \{ -2Lh \} \exp \left\{ i \left( \frac{L}{1 - 2h^2} \right) \right\}}
\]

\[
= e^{3lh} \frac{\exp \{ iL + O(Lh^2) \}}{\exp \{ iL + O(Lh^2) \}} \approx e^{3lh}, \quad (B.7)
\]

\[
\left( \frac{\hat{p}_1(+) \hat{p}_2(-)}{\hat{p}_2(-)} \right)^M = \left( \frac{1 - ih + h^2}{1 - ih - 2h^2} \right)^M = \frac{(1 + h^2)^M (1 - \frac{ih}{1 + h^2})^M}{(1 - 2h^2)^M (1 - \frac{ih}{1 - 2h^2})^M}
\]

\[
\approx \frac{\exp \{ Lh \} \exp \left\{ -i \left( \frac{L}{1 + h^2} \right) \right\}}{\exp \{ -2Lh \} \exp \left\{ -i \left( \frac{L}{1 - 2h^2} \right) \right\}}
\]

\[
= e^{3lh} \frac{\exp \{ -iL + O(Lh^2) \}}{\exp \{ -iL + O(Lh^2) \}} \approx e^{3lh}, \quad (B.8)
\]

169
\[
\begin{align*}
\left( \frac{\hat{\rho}_1}{\hat{\rho}_2} \right)^M_{(-)} &= \left( \frac{1 + \text{i}h + h^2}{1 - \text{i}h - 2h^2} \right)^M_{(-)} = \frac{(1 + h^2)^M (1 + \frac{\text{i}h}{1 + h^2})^M}{(1 - 2h^2)^M (1 - \frac{\text{i}h}{1 - 2h^2})^M} \\
&\approx \frac{\exp \{ Lh \} \exp \{ i \left( \frac{L}{1 + h^2} \right) \}}{\exp \{ -2Lh \} \exp \{ -i \left( \frac{L}{1 - 2h^2} \right) \}} \\
&= e^{3Lh} \exp \{ iL + O(Lh^2) \} \approx e^{3Lh} e^{2i\text{M} \theta},
\end{align*}
\]

Thus, we have found (4.116):

\[
\begin{align*}
\left( \frac{\hat{\rho}_1}{\hat{\rho}_2} \right)^M_{(+)} &= e^{-2i\text{M} \theta}, e^{3Lh}, & \left( \frac{\hat{\rho}_1}{\hat{\rho}_2} \right)^M_{(-)} &= e^{3Lh}, \\
\left( \frac{\hat{\rho}_1}{\hat{\rho}_2} \right)^M_{(+)} &= e^{3Lh}, & \left( \frac{\hat{\rho}_1}{\hat{\rho}_2} \right)^M_{(-)} &= e^{2i\text{M} \theta}, e^{3Lh}. \tag{B.10}
\end{align*}
\]
Appendix C

The solutions for the hyperbolic PDE system (1.4)

In (1.8), we already showed one of the constant solutions to the hyperbolic PDE system (1.4). Now we will discuss all the possible constant solutions to this PDE system. It’s obvious the first solution is the trivial solution, i.e., $S_1^\pm, S_2^\pm, S_3^\pm$ are all 0. By inspecting on the structure of (1.4), we have found that the non-trivial solutions all fall into the following cases:

**Case(1):** One of the six values $S_1^\pm, S_2^\pm, S_3^\pm$ is not 0, while the rest of them are all 0. This nonzero value can be any real number.

**Case(2):** Two of the six values $S_1^\pm, S_2^\pm, S_3^\pm$ are not 0, while the rest of them are all 0. These two nonzero values can be any real numbers. However, only the following
combinations are possible:

\[
S_3^+ \neq 0 \text{ and } S_3^- \neq 0, \text{ while the rest are all 0}
\]
\[
S_3^+ \neq 0 \text{ and } S_1^+ \neq 0, \text{ while the rest are all 0}
\]
\[
S_3^+ \neq 0 \text{ and } S_2^+ \neq 0, \text{ while the rest are all 0}
\]
\[
S_3^- \neq 0 \text{ and } S_1^- \neq 0, \text{ while the rest are all 0}
\]
\[
S_3^- \neq 0 \text{ and } S_2^- \neq 0, \text{ while the rest are all 0}
\]
\[
S_1^+ \neq 0 \text{ and } S_2^+ \neq 0, \text{ while the rest are all 0}
\]
\[
S_1^+ \neq 0 \text{ and } S_1^- \neq 0, \text{ while the rest are all 0}
\]
\[
S_2^+ \neq 0 \text{ and } S_2^- \neq 0, \text{ while the rest are all 0}
\]
\[
S_2^- \neq 0 \text{ and } S_1^- \neq 0, \text{ while the rest are all 0}
\]

(C.1)

**Case (3):** Three of \(S_1^\pm, S_2^\pm, S_3^\pm\) are not 0, while the rest of them are all 0. These three nonzero values can be any real numbers. However, only the following combinations are possible:

\[
S_{1,2,3}^+ \neq 0, \text{ while } S_{1,2,3}^- = 0
\]
\[
S_{1,2,3}^- \neq 0, \text{ while } S_{1,2,3}^+ = 0
\]

(C.2)

**Case (4):** Four of \(S_1^\pm, S_2^\pm, S_3^\pm\) are not 0, while the rest of them are all 0. These four nonzero values can be any real numbers. However, only the following combination is
possible:

\[ S_1^\pm, S_2^\pm \neq 0, \text{ while } S_3^\pm = 0 \] (C.3)
BIBLIOGRAPHY


