Stability analysis of the numerical Method of characteristics applied to a class of energy-preserving hyperbolic systems. Part II: Nonreflecting boundary conditions

T.I. Lakoba*, Z. Deng

Department of Mathematics and Statistics, 16 Colchester Ave.,
University of Vermont, Burlington, VT 05401, USA

February 23, 2019

Abstract

We show that imposition of non-periodic, in place of periodic, boundary conditions (BC) can alter stability of modes in the Method of characteristics (MoC) employing certain ordinary-differential equation (ODE) numerical solvers. Thus, using non-periodic BC may render some of the MoC schemes stable for most practical computations, even though they are unstable for periodic BC. This fact contradicts a statement, found in some textbooks and known as part of the Babenko–Gelfand criterion, that an instability detected by the von Neumann analysis for a given numerical scheme implies an instability of that scheme with arbitrary (i.e., non-periodic) BC. We explain the mechanism behind this contradiction, which lies in a certain property of the scheme's eigenmodes that is assumed by the Babenko–Gelfand criterion but does not hold for eigenmodes of some of the MoC-based schemes. We also show that, and explain why, for the MoC employing some other ODE solvers, stability of the modes may indeed not be improved by non-periodic BC, as the Babenko–Gelfand criterion implies.

Keywords: Method of characteristics, Coupled-wave equations, Numerical instability, Non-periodic boundary conditions, Block-Toeplitz matrices.

*tlakoba@uvm.edu, 1 (802) 656-2610

1 Introduction

In Part I [1] of this study we considered the numerical stability of the Method of characteristics (MoC) applied to a class of hyperbolic partial differential equations (PDEs) with periodic boundary conditions (BC). More specifically, the ordinary differential equation (ODE) numerical solvers employed by the MoC in [1] along the characteristics were the simple Euler (SE), modified Euler (ME), and Leapfrog (LF) ones. The class of the PDEs considered in [1] has the linearized form:

$$\tilde{\mathbf{u}}_t + \mathbf{\Sigma}\, \tilde{\mathbf{u}}_x = \mathbf{P}\, \tilde{\mathbf{u}}, \tag{1}$$

where $\tilde{\mathbf{u}}$ is a small perturbation on top of the background solution $\mathbf{u}^{(0)}$ of the original nonlinear PDE, $\mathbf{\Sigma} = \mathrm{diag}(I_N, -I_N)$, with I_N being the N-dimensional identity matrix, and $\mathbf{P}$ is a constant $2N \times 2N$ matrix. Importantly, the considered class of the PDEs possesses a number of "conservation laws". In particular, the solution $\mathbf{u} \equiv \left[\underline{u}^{+}, \underline{u}^{-}\right]^T$ of the nonlinear PDE satisfies:

$$(\partial_t \pm \partial_x)\, |\underline{u}^{\pm}|^2 = 0, \tag{2a}$$

where $|\cdots|$ stands for the length of the corresponding N-dimensional vector. For periodic BC on the interval $x \in [0,\, L]$, (2a) implies

$$\partial_t \int_0^L |\underline{u}^{\pm}|^2 = 0; \tag{2b}$$

hence the name "conservation law". (For future reference, let us note that the meaning of the integral here is the energy of the solution.) Therefore, we initially believed that the LF solver, which is well known to (almost) preserve conserved quantities of energy-preserving ODEs over indefinitely long integration times, would also perform the best among the three aforementioned solvers when employed by the MoC to integrate the above energy-preserving PDE.

In [1] we showed that for periodic BC, the MoC with any of those three ODE solvers exhibits numerical instability. For the SE solver, the strongest such an instability occurs in the "ODE" and "anti-ODE" limits, i.e. for the wavenumbers $k = 0$ and $|k| = k_{\max} \equiv \pi/h$, respectively. (Here h is the grid spacing in both space and time.) This mild instability of the SE applied to conservative ODEs (e.g., the harmonic oscillator model) is well known: see, e.g., [2]. The ME solver is known to exhibit a similar, although much weaker, instability when applied to conservative ODEs. We have found, however, that when the ME is used within the MoC framework, its most unstable modes occur in the "middle"[1] of the Fourier spectrum, i.e. for $|k| \sim k_{\max}/2$, and that the instability's growth rate is much greater than that in the ODE (and anti-ODE) limit. Finally, the LF solver used along the characteristics was found to exhibit by far the strongest instability among those three solvers. In Fig. 1 we show the amplification factor, $|\lambda| \equiv \left|\tilde{\mathbf{u}}^{n+1}\right| / \left|\tilde{\mathbf{u}}^{n}\right|$, of the three numerical schemes: MoC-SE, MoC-ME, and MoC-LF, as obtained in [1]; here $\tilde{\mathbf{u}}^n$ is the numerical error at the nth time level. Note the order-of-magnitude difference of the instability growth rate of the MoC-LF compared to that rate of the MoC-Euler methods.

[1] We use quotes here and in what follows because, strictly speaking, this refers to the middle of either left or right *half* of the spectrum.

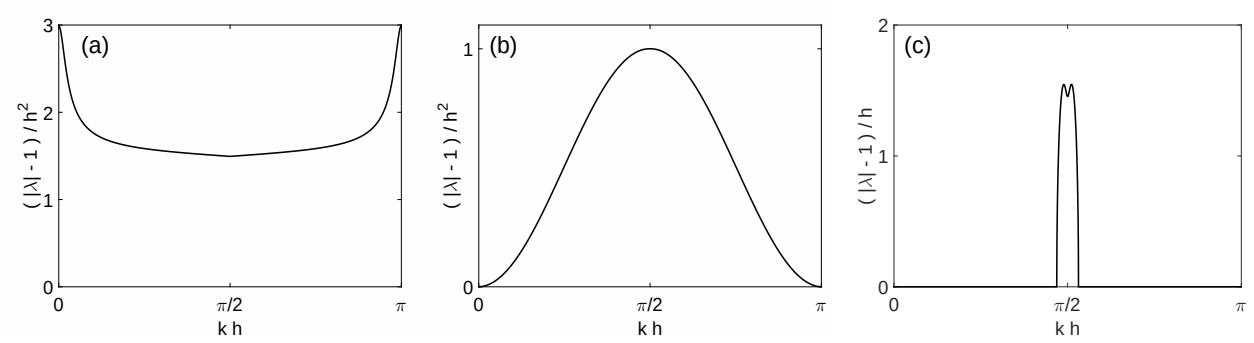

Figure 1: Representative amplification factors of the MoC-SE (a), MoC-ME (b), and MoC-LF (c) with periodic BC, as obtained in [1]. Note a drastically different vertical scale in (c).

In this paper we extend the stability analysis of these three schemes to the case where the BC are not periodic. More specifically, we consider so-called nonreflecting BC, whose exact form will be stated in Section 2 and which are often relevant for hyperbolic PDEs whose left-hand side (lhs) is of the same form as that of (1). The motivation for our study came from the following numerical observation. When we simulated the given PDE system with nonreflecting BC by each of the three schemes, we found that the MoC-SE and MoC-LF exhibited the same growth rates of the numerical error as in the case of periodic BC. This is in agreement with the common knowledge that the instability of a problem with periodic BC implies an instability of a problem with any other BC (see below). *However*, the instability of the MoC-ME (see Fig. 1(b)) was suppressed when we used nonreflecting BC.[2] This surprising observation prompted two questions, which we address in this work.

First, as we have already mentioned, it is commonly stated in textbooks on numerical analysis of PDEs that the von Neumann stability analysis of a scheme (with constant coefficients) gives a *necessary* condition for its stability (see, e.g., [3]–[5]). In other words, if numerical instability is found in a problem with periodic BC, then the same problem with some non-periodic BC will also be numerically unstable. This statement follows from the commonly made assumption that the eigenmodes of the periodic problem (i.e., Fourier harmonics) approximate a proper subset of the eigenmodes of the non-periodic problem. We will discuss this in detail in Section 7.3.

Our finding about the MoC-ME clearly violates this common knowledge: the scheme is unstable for periodic BC but becomes stable (see footnote 2 above) for nonreflecting BC. Thus, the *first question* that we will answer is:

(i) Why does this occur?

However, as we have mentioned, this suppression of numerical instability takes place for only one out of the three schemes considered. Thus, the *second question* is:

(ii) Why does it occur for the MoC-ME but not for the MoC-SE and MoC-LF?

The organization of the main part of this paper is outlined in the next three paragraphs. In

[2]More precisely, its instability in the ODE limit remained, but it was several orders of magnitude weaker and did not affect the solution over the simulation times of interest for this study.

Section 2 we give more details about the PDE system under study. In Section 3 we set up the framework for the stability analysis of the MoC schemes with non-periodic BC. Let us stress that this framework is different from the von Neumann stability analysis, which is used for problems with periodic BC. In fact, we have found only one paper, [6], where a similar (although less general) type of analysis for a MoC method had been used. In Section 4, we will apply our analysis to the MoC-SE with nonreflecting BC. Let us stress that this scheme is *not* at the focus of our study — because nonreflecting BC do not suppress instability for it, — and yet we will spend a substantial amount of effort on this case. The reason is that in this simplest case, we will be able to not only clearly outline the steps of the analysis, but also to *quantitatively* verify its predictions by direct numerical simulations. In Section 5 we will apply this analysis to the MoC-ME and show that for it, unlike for the MoC-SE, nonreflecting BC suppress instability. A *qualitative* reason for that is explained at the end of Section 4. This explanation will address the 'second question' listed above, but only partially: it will leave open the question as to why the same mechanism does not suppress an instability for the MoC-LF. We will address the MoC-LF case in Section 6, thereby completely answering the 'second question'.

In Section 7 we will summarize our conclusions and will present a *qualitative* explanation to the 'first question' (the (i) above), i.e., why an instability of a problem with periodic BC does not always imply an instability of the same problem with non-periodic BC. As this explanation pertains to the *main result* of our study, we compartmentalized it in a separate subsection, 7.3. The reader who is not interested in the details of the analysis may proceed directly to that subsection. In subsection 7.4 we complement the quantitative answer to the 'second question' (the (ii) above), given in Section 6, by a *qualitative* consideration. Appendices A and B contain technical derivations for Sections 3 and 4.

We would like to emphasize that the focus of this work is *not* on a specific application of MoC schemes but on the understanding of the effect of non-periodic BC on the (in)stability of these schemes, which may be in contradiction to properties stated in textbooks. *However*, in Section 7.2 we discuss one practical application of our analysis, namely: the stabilization of the MoC-ME scheme by the imposition of nonreflecting BC. To that end, we first present, in Appendix C, the background on a system different from the constant-coefficient system considered in the main part of the paper. That other system is a soliton (i.e., localized and thus non-constant) solution of the Gross–Neveu model in the relativistic field theory, which has been actively studied in the past decade. In subsection 7.2 we illustrate with direct numerical simulations that suppression of numerical instability of the MoC-ME by nonreflecting BC, predicted by our analysis of the constant-coefficient system, also occurs for the Gross–Neveu soliton.

2 Physical model

While our study will focus on a linear problem of a rather general form (see Eq. (1) or (6a) below), we will begin by stating a specific nonlinear problem which had originally motivated our study and

whose linearization leads to (6a). The vector form of the PDE system under consideration is:

$$\underline{\mathbf{S}}^{\pm}_t \pm \underline{\mathbf{S}}^{\pm}_x = \underline{\mathbf{S}}^{\pm} \times \hat{\mathbf{J}} \underline{\mathbf{S}}^{\mp}, \tag{3}$$

where $\underline{\mathbf{S}}^{\pm} \equiv [S_1^{\pm}, S_2^{\pm}, S_3^{\pm}]^T$, $\hat{\mathbf{J}} = \mathrm{diag}(1, -1, -2)$, and superscript 'T' denotes the transposition. This system is a representative of a class of models that arise in studying propagation of light in birefringent optical fibers with Kerr nonlinearity [7]–[10]. It should be noted that the specific numerical entries of $\hat{\mathbf{J}}$ arise from physical considerations and therefore should *not* be replaced with arbitrary numbers, as that would not correspond to a physical situation. The component form and a non-constant, soliton/kink solution of (3) can be found in, e.g., [1]. Here (as in [1]) we consider the numerical stability of a *constant* solution of (3):

$$S^{\pm}_{1,3} = 0, \qquad S^{\pm}_2 = \pm 1, \tag{4}$$

when it is simulated by the MoC. Considering only one representative, (3), of a broader class of models (see [1]) and its simplest solution, (4), will allow us to *focus on the analysis of the numerical scheme* without being distracted by complexities of the physical model. The relevance of the numerical (in)stability of the constant solution (4) to the numerical (in)stability of "more physically interesting", soliton/kink and related, solutions was discussed in [1] after Eq. (8).

We will linearize Eqs. (3) on the background of solution (4) using

$$S^{\pm}_j = S^{\pm}_{j\,0} + s^{\pm}_j, \qquad j = 1, 2, 3, \tag{5}$$

where $S^{\pm}_{j\,0}$ are the components of the exact solution (4) and $s^{\pm}_j$ are small perturbations. The linearized system (3) has the form:

$$\mathbf{s}_t + \mathbf{\Sigma}\, \mathbf{s}_x = \mathbf{P}\, \mathbf{s}, \tag{6a}$$

$$(s^{\pm}_2)_t \pm (s^{\pm}_2)_x = 0, \tag{6b}$$

where $\mathbf{s} = [s_1^+, s_3^+, s_1^-, s_3^-]^T$, the 4×4 matrices in (6a) are:

$$\mathbf{\Sigma} = \mathrm{diag}(I, -I), \qquad \mathbf{P} \equiv \begin{pmatrix} P^{++} & P^{+-} \\ P^{-+} & P^{--} \end{pmatrix} = \begin{pmatrix} -A & B \\ -B & A \end{pmatrix}, \tag{7a}$$

and the 2×2 matrices in (7a) are:

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad B = -\begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}. \tag{7b}$$

Given the trivial dynamics (6b) of $s^{\pm}_2$, below we will consider only the dynamics of $s^{\pm}_{1,3}$, given by (6a). As we pointed out in [1], systems of the latter form describe linear or linearized dynamics in a wide class of physical models in plasma physics, various areas of optics, and relativistic field theory.

Let us reiterate, however, that the *focus of this work is not on a particular physical application but, rather, on understanding the behavior of the MoC schemes with non-periodic BC*, as we explained in the Introduction. The reason[3] why we chose to consider the specific form of matrix $\mathbf{P}$

[3] apart from our original interest in simulating soliton-kink solutions of system (3)

is that the forthcoming analysis simply cannot be performed for an unspecified form of P. (Note that $\mathbf{P}$ in (6a) cannot be diagonalized without affecting the matrix on the lhs.) While the *results* of the analysis, of course, depend on the specific form of $\mathbf{P}$, its methodology does not. Moreover, the conclusion about suppression of the instability of the MoC-ME by nonreflecting BC holds for at least one other system of current research interest, as we demonstrate in Section 7.2 with direct numerical simulations.[4]

In Eqs. (6) and (7) and in what follows we have adopted the following notations. Boldfaced quantities with an underline, $\underline{\mathbf{S}}^{\pm}$, will continue to denote 3×1 vectors, as in (3). Boldfaced quantities *without* an underline or a hat will denote 4×4 matrices or 4×1 vectors, as in (6a); the ambiguity of the same notations for matrices and vectors here will not cause any confusion. Finally, underlined letters in regular (not boldfaced) font will denote 2×1 vectors; e.g.:

$$\underline{s}^{\pm} \equiv [s_1^{\pm}, s_3^{\pm}]^T. \tag{8}$$

Clearly then, $\mathbf{s} \equiv \big[(\underline{s}^{+})^T, (\underline{s}^{-})^T\big]^T$.

The nonreflecting BC for system (3) are:

$$\underline{\mathbf{S}}^{+}(0,t) = \underline{\mathbf{F}}^{+}(t), \qquad \underline{\mathbf{S}}^{-}(L,t) = \underline{\mathbf{F}}^{-}(t), \tag{9}$$

where $x \in [0,\, L]$ and $\underline{\mathbf{F}}^{\pm}(t)$ are given. These BC specify the right (left)-propagating components at the left (right) boundary of the spatial domain. For the small perturbation in (5), BC (9) transform into:

$$\underline{s}^{+}(0,t) = \underline{0}, \qquad \underline{s}^{-}(L,t) = \underline{0}. \tag{10}$$

3 Setup of the stability analysis of the MoC with BC (10)

We will illustrate details of this setup using the simplest "flavor" of the MoC, the MoC-SE. This will allow us to skip most details in later sections, devoted to the more complex schemes, the MoC-ME and MoC-LF.

The MoC-SE scheme for system (3) is:

$$(S_j^{\pm})_m^{n+1} = (S_j^{\pm})_{m\mp1}^{n} + h\, f_j^{\pm}\big(\,(\underline{\mathbf{S}}^{+})_{m\mp1}^{n},\, (\underline{\mathbf{S}}^{-})_{m\mp1}^{n}\,\big), \qquad j = 1,2,3; \tag{11}$$

where $f_j^{\pm}$ are the nonlinear functions on the rhs of (3). The spatial grid has $M+1$ nodes: $m = 0,1,\ldots,M$; in (11), the quantities with superscript '+' ('−') have $m = 1,\ldots,M$ ($m = 0,\ldots,M-1$). The nonreflecting BC (9) for solution (4) take on the form:

$$(\underline{\mathbf{S}}^{+})_0 = [0,1,0]^T, \qquad (\underline{\mathbf{S}}^{-})_M = [0,-1,0]^T. \tag{12}$$

[4]Since that system has non-constant coefficients, our analysis cannot be applied to it.

The numerical error satisfies the linearized form of (11), (12):

$$\begin{pmatrix} \underline{s}^+ \\ \underline{s}^- \end{pmatrix}_m^{n+1} = \begin{pmatrix} \underline{s}^+ \\ \underline{0} \end{pmatrix}_{m-1}^{n} + \begin{pmatrix} \underline{0} \\ \underline{s}^- \end{pmatrix}_{m+1}^{n} + h \begin{pmatrix} P^{++} & P^{+-} \\ \mathcal{O} & \mathcal{O} \end{pmatrix} \begin{pmatrix} \underline{s}^+ \\ \underline{s}^- \end{pmatrix}_{m-1}^{n} + h \begin{pmatrix} \mathcal{O} & \mathcal{O} \\ P^{-+} & P^{--} \end{pmatrix} \begin{pmatrix} \underline{s}^+ \\ \underline{s}^- \end{pmatrix}_{m+1}^{n}, \quad (13)$$

$$(\underline{s}^+)_0 = \underline{0}, \qquad (\underline{s}^-)_M = \underline{0}. \quad (14)$$

In (13), $\mathcal{O}$ is the 2×2 zero matrix and P^{++} etc. are defined in (7).

Following the idea of [6], we rewrite (13) as:

$$(\mathbf{s})_m^{n+1} = \mathbf{\Gamma}(\mathbf{s})_{m-1}^n + \mathbf{\Omega}(\mathbf{s})_{m+1}^n, \quad m = 1, \ldots, M-1; \quad (15a)$$

$$\begin{aligned} (\underline{s}^+)_M^{n+1} &= (I + hP^{++})(\underline{s}^+)_{M-1}^n + hP^{+-}(\underline{s}^-)_{M-1}^n, \\ (\underline{s}^-)_0^{n+1} &= hP^{-+}(\underline{s}^+)_1^n + (I + hP^{--})(\underline{s}^-)_1^n, \end{aligned} \quad (15b)$$

where

$$\mathbf{\Gamma} = \mathbf{\Gamma}_0 + h\mathbf{\Gamma}_1, \qquad \mathbf{\Omega} = \mathbf{\Omega}_0 + h\mathbf{\Omega}_1, \quad (16a)$$

$$\mathbf{\Gamma}_0 = \begin{pmatrix} I & \mathcal{O} \\ \mathcal{O} & \mathcal{O} \end{pmatrix}, \quad \mathbf{\Gamma}_1 = \begin{pmatrix} P^{++} & P^{+-} \\ \mathcal{O} & \mathcal{O} \end{pmatrix}; \qquad \mathbf{\Omega}_0 = \begin{pmatrix} \mathcal{O} & \mathcal{O} \\ \mathcal{O} & I \end{pmatrix}, \quad \mathbf{\Omega}_1 = \begin{pmatrix} \mathcal{O} & \mathcal{O} \\ P^{-+} & P^{--} \end{pmatrix}. \quad (16b)$$

In writing (15b), we have used (14) and (16).

Scheme (15) with nonreflecting BC can be written in the form

$$\begin{pmatrix} (\mathbf{s})_0 \\ \vdots \\ (\mathbf{s})_M \end{pmatrix}^{n+1} = \begin{pmatrix} \mathcal{O} & \mathbf{\Omega} & \mathcal{O} & \cdots & & & \\ \mathbf{\Gamma} & \mathcal{O} & \mathbf{\Omega} & \mathcal{O} & \cdots & & \vdots \\ \mathcal{O} & \mathbf{\Gamma} & \mathcal{O} & \mathbf{\Omega} & \cdots & & \\ & & & \ddots & & & \\ \vdots & & & \cdots & \mathbf{\Gamma} & \mathcal{O} & \mathbf{\Omega} \\ & & & \cdots & \mathcal{O} & \mathbf{\Gamma} & \mathcal{O} \end{pmatrix} \begin{pmatrix} (\mathbf{s})_0 \\ \vdots \\ (\mathbf{s})_M \end{pmatrix}^n \quad (17a)$$

or, equivalently,

$$\mathbb{s}^{n+1} = \mathbb{N}\,\mathbb{s}^n, \quad (17b)$$

where $\mathbb{s}$ is the $4(M+1) \times 1$ vector and $\mathbb{N}$ is the $4(M+1) \times 4(M+1)$ matrix in (17a). Note that in (17a), $\mathcal{O}$ denotes the 4×4 zero matrix.

To analyze stability of the MoC-SE, we need to determine whether the largest-in-magnitude eigenvalue of $\mathbb{N}$ exceeds 1. Since $\mathbb{N}$ is block-tridiagonal and Toeplitz, we will find its eigenvalues using the method for non-block tridiagonal Toeplitz matrices (see, e.g., [11]); note that it is here where our approach differs from that in [6]. Namely, we first seek the 4×1 "components" of $\mathbb{s}$ in the form:

$$(\mathbf{s})_{m+1}^n = \rho\,(\mathbf{s})_m^n. \quad (18)$$

In Appendix A we explain why ρ should be considered a scalar rather than a 4×4 matrix. Substituting (18) into (15a) and using

$$\mathbb{s}^{n+1} = \lambda\, \mathbb{s}^n, \tag{19}$$

where λ is the eigenvalue that we want to find, we obtain for the eigenvector $\boldsymbol{\xi} \equiv (\mathbf{s})_m$:

$$(\rho\, \boldsymbol{\Omega} + \rho^{-1}\boldsymbol{\Gamma} - \lambda \mathbf{I})\boldsymbol{\xi} = \mathbf{0}. \tag{20}$$

We will now outline three steps of our analysis, which will be carried out in Sections 4 and 5 for the MoC-SE and MoC-ME, respectively.

Step 1: We will show that the characteristic equation for (20) yields four solutions $\rho_j(\lambda)$, $j = 1, \ldots, 4$ for a given λ. Note that according to (18), values of $\rho_j(\lambda)$ determine the spatial behavior of the modes of the problem. (For example, for periodic BC, one would have $\rho^M = 1$, which is satisfied by $\rho = \exp[ikh]$ with $k = 2\pi m/L$, $m \in \mathbb{Z}$. With that ρ, Eqs. (15a) and (18) recover the results of the von Neumann analysis in [1].)

Step 2: For the ρ_j found in Step 1, we will find their respective eigenvectors $\boldsymbol{\xi}_j$ from (20). These eigenvectors do *not* affect the spatial structure of the modes but simply reflect the distribution of "energy" among the components of $\mathbf{s} \equiv [s_1^+,\, s_3^+,\, s_1^-,\, s_3^-]^T$ at any given point in space.

Step 3: Using the form

$$(\mathbf{s})_m^n = \lambda^n \sum_{j=1}^{4} C_j \rho_j^m \boldsymbol{\xi}_j, \tag{21}$$

we will determine constants C_j by requiring (21) to satisfy the BC (14). Namely, denoting

$$\boldsymbol{\xi}_j \equiv \begin{pmatrix} \underline{\xi}_j^+ \\ \underline{\xi}_j^- \end{pmatrix} \tag{22}$$

and substituting (21) into (14), we have:

$$\sum_{j=1}^{4} C_j\, \underline{\xi}_j^+(\lambda) = \underline{0}, \qquad \sum_{j=1}^{4} C_j \rho_j^M(\lambda)\, \underline{\xi}_j^-(\lambda) = \underline{0}. \tag{23a}$$

These conditions can be rewritten as a homogeneous linear system:

$$\boldsymbol{\Phi}(\lambda) \begin{pmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \end{pmatrix} = \mathbf{0}, \qquad \boldsymbol{\Phi}(\lambda) \equiv \begin{pmatrix} \underline{\xi}_1^+ & \underline{\xi}_2^+ & \underline{\xi}_3^+ & \underline{\xi}_4^+ \\ \rho_1^M \underline{\xi}_1^- & \rho_2^M \underline{\xi}_2^- & \rho_3^M \underline{\xi}_3^- & \rho_4^M \underline{\xi}_4^- \end{pmatrix}. \tag{23b}$$

Note that the dependence of Φ on λ comes from such a dependence of $\boldsymbol{\xi}_j$ and ρ_j. Thus, the eigenvalue λ of the amplification matrix in (17) will be found from the characteristic equation

$$\det \boldsymbol{\Phi}(\lambda) = 0. \tag{24}$$

Let us note that the above steps are standard in analyzing any intial-boundary value problem (see, e.g., [12]) with separable spatial and temporal variables. Importantly, they differ from the

steps of the stability analysis of the initial-boundary value problems found in textbooks [3]–[5] in the following aspect. Following a step in the Babenko–Gelfand procedure, the textbooks *postulate* that all spatial modes of the problem fall into two groups: those localized near either of the boundaries and those resembling Fourier harmonics inside the spatial domain, sufficiently far away from the boundaries. In contrast, our Step 1 *finds* the spatial modes. Surprisingly, those modes turn out to violate the aforementioned categorization into the two groups. It is this circumstance that will be shown (see Section 7.3) to be the reason behind the "disappearance" of the numerical instability of the MoC-ME, announced in the Introduction.

4 Stability analysis of the MoC-SE

In the first three subsections of this Section we will implement the three respective steps listed at the end of Section 3. The main result of our analysis will be obtained in subsection 4.3. In the fourth subsection, we will present a verification of this analysis by direct numerical simulations of scheme (11), (12). In the fifth subsection, we will discuss what bearing the results of this Section will have on those in Section 5.

4.1 Step 1: Finding $\rho(\lambda)$ in (20)

To obtain a perturbative (for $h \ll 1$) solution of the characteristic polynomial of (20), note that it can be written as

$$\det\left(\begin{pmatrix} (\rho^{-1}-\lambda)I & \mathcal{O} \\ \mathcal{O} & (\rho-\lambda)I \end{pmatrix} + h\begin{pmatrix} -\rho^{-1}A & \rho^{-1}B \\ -\rho B & \rho A \end{pmatrix}\right) = 0, \tag{25}$$

where A, B are defined in (7b). Consequently, in the main order, one has

$$\rho_1^{(0)} = \lambda^{-1}, \qquad \rho_2^{(0)} = \lambda. \tag{26}$$

Since each of these roots is double, then for $0 < h \ll 1$ one will have four roots of (25). To find them, we substitute[5]

$$\rho = \rho_j^{(0)} + h\rho_j^{(1)} + h^2\rho_j^{(2)}, \qquad j = 1, 2 \tag{27}$$

into (25). The orders $O(h^2)$ and $O(h^3)$ of the resulting expression, found with software *Mathematica*, yield expressions for $\rho_j^{(1)}$ and $\rho_j^{(2)}$, respectively:

$$\rho_{1(\pm)} = \frac{1}{\lambda}\left(1 \mp ih - \frac{2h^2}{\lambda^2-1}\right) \equiv \frac{1}{\lambda}\,\widehat{\rho}_{1(\pm)}, \qquad \rho_{2(\pm)} = \lambda\left(1 \pm ih - h^2\frac{\lambda^2-3}{\lambda^2-1}\right) \equiv \lambda\,\widehat{\rho}_{2(\pm)}. \tag{28}$$

Let us note that the *subscript* notations '$(\pm)$' refer to distinct roots in (28). Thus, they are *in no way related* to the *superscript* notations '$\pm$', which were used in, e.g., (22) and will be used in similar context below. Those superscript notations denote the "forward"- and "backward"-propagating components of the numerical error, in analogy with that notation in (13). To emphasize this

[5] see a discussion in the paragraph that starts with Eqs. (29)

difference between the sub- and superscript $\pm$ notations, we will always use parentheses in subscript $(\pm)$.

Formulae (28) complete Step 1 of the analysis of the eigenvalue problem for the amplification matrix $\mathbb{N}$ in (17). A discussion about one of its assumptions will now be in order.

It should be noted that seeking the solution of (25) in the form (27) is valid only as long as the two roots (26) are "sufficiently distinct", i.e.:

$$\left|\rho_1^{(0)} - \rho_2^{(0)}\right| \gg h, \tag{29a}$$

or, equivalently,

$$|\lambda^2 - 1| \gg h. \tag{29b}$$

In other words, the analysis presented in this section will break down for

$$\lambda \approx \pm 1, \qquad \text{i.e.} \quad \rho \approx \pm 1, \tag{30a}$$

where the implication is based on (26), (27). (Note that this does not preclude the possibility $|\lambda|^2 \approx 1$; e.g., for $\lambda^2 \approx -1$ our analysis will be valid.) The proper approach in the case (30a) is to seek

$$\lambda = \pm 1 + \alpha h, \qquad \rho = \pm 1 + \beta, \tag{30b}$$

for which a fourth-degree polynomial in α and β will result. The corresponding calculations are considerably more complex than those in Sections 4.1–4.3 and will not be carried out.[6] Therefore, it is now important to clarify the following issues:

(i) What we will miss by not considering case (30);

(ii) Why this missed information will not be crucial for our analysis; and

(iii) How the curtailed analysis based on (27) will benefit us.

Before we address these issues, we need to establish a correspondence between modes of the problem with nonreflecting BC, characterized by parameter ρ, and Fourier harmonics of the problem with periodic BC. According to (18), ρ is the counterpart of e^{ikh} in the von Neumann analysis (see the line before Eq. (18) in [1]). Therefore, the modes in (30a) correspond to the Fourier harmonics with the lowest and highest wavenumbers, $kh = 0$ and $kh = \pi$:

$$\rho \approx 1 = e^{i0}, \qquad \rho \approx -1 = e^{i\pi}. \tag{31}$$

We will now present answers to questions (i)–(iii) above.

(i) We will be unable to rigorously describe the instability of the MoC-SE with nonreflecting BC. Indeed, as Fig. 1(a) shows, the most unstable harmonics of the periodic problem are those with $kh = 0$ and $kh = \pi$, and our numerical simulations, described in Section 4.4, indicate that the

[6] A related analysis, however, will be required for the MoC-LF and therefore will be presented in Section 6.

corresponding modes (31) are the most unstable also in the problem with nonreflecting BC. Thus, it may sound as if we will intentionally neglect the main goal of our study. However, in the next two paragraphs, we will explain why this is not so.

(ii) Let us focus on the mode with $\rho \approx 1$; the case of the mode with $\rho \approx -1$ can be considered similarly. While a nonzero mode with $\rho = 1$ will not satisfy the BC (10), one can still consider the *limit* $\rho \to 1$, where the BC can be upheld. This will correspond to the case $|\beta| \ll |\alpha|$ in (30b). Then, by taking the limit $\beta \to 0$ in the characteristic polynomial obtained from the substitution of (30b) into (20), one finds:

$$\alpha^2(\alpha^2 + 6) = 0 \qquad \Rightarrow \qquad \alpha = 0,\ \pm i\sqrt{6}. \tag{32}$$

The last two roots correspond exactly to the values $(|\lambda| - 1)/h^2$ found by the von Neumann analysis for the periodic BC [1]. Moreover, in our simulations of the MoC-SE, we found that the growth rates of the numerically unstable modes are the same for the periodic and nonreflecting BC. Thus, the simplified analysis in this paragraph has been able to quantitatively predict the instability growth rate of the MoC-SE with periodic BC.

(iii) Then, what is the point of considering the less unstable (or, rather, as we will show, even stable) modes satisfying (27), (29)? The point is that we will thereby uncover a *mechanism* which suppresses the numerical instability of those modes, and it will turn out to be the same mechanism which suppresses the *most unstable* modes of the MoC-ME! We will preview this in Section 4.5. Here we will only reiterate our earlier statement: Presenting details for the simplest scheme, the MoC-SE, will keep the analysis more transparent and will also allow us to proceed faster when considering the MoC-ME in Section 5.

4.2 Step 2: Finding $\boldsymbol{\xi}$ in (20)

We will now find the eigenvectors corresponding to the four roots (28). We will present details of the calculation for $\boldsymbol{\xi}_{1(+)}$, corresponding to the root $\rho_{1(+)}$; calculations for the other three roots are similar. Note that for $\rho = \rho_{1(+)}$, Eq. (20) in the order $O(1)$ is satisfied by the solution $[\underline{u}^T, \underline{0}^T]^T$ for an arbitrary $\underline{u}$ (recall that we consider the case $\lambda^2 - 1 \not\approx 0$). Therefore, in the order $O(h)$ we seek the solution of (20) in the form

$$\boldsymbol{\xi}_{1(+)} = \begin{pmatrix} \underline{u} \\ h\underline{v} \end{pmatrix}, \tag{33}$$

where $\underline{u}$, $\underline{v} = O(1)$ are to be determined. (It will become clear later that to explain the suppression of instability of the MoC-SE, we will not need higher-order terms in (33).) Using (16a), (28) with terms up to $O(h)$, and (33), we obtain in the $O(h)$ order of (20):

$$\left(i\boldsymbol{\Gamma}_0\lambda - i\boldsymbol{\Omega}_0\lambda^{-1} + \boldsymbol{\Gamma}_1\lambda + \boldsymbol{\Omega}_1\lambda^{-1}\right)\begin{pmatrix} \underline{u} \\ \underline{0} \end{pmatrix} + \left(\boldsymbol{\Omega}_0\lambda^{-1} - \lambda\mathbf{I}\right)\begin{pmatrix} \underline{0} \\ \underline{v} \end{pmatrix} = \mathbf{0}. \tag{34}$$

The top 2×1 block of this equation determines $\underline{u}$:

$$(A - iI)\underline{u} = \underline{0} \qquad \Rightarrow \qquad \underline{u} = \begin{pmatrix} 1 \\ i \end{pmatrix}; \tag{35a}$$

here A (and B below) is defined in (7b). The bottom 2×1 block of (34) yields:

$$\underline{v} = \frac{1}{1-\lambda^2} B \underline{u}. \tag{35b}$$

Performing similar calculations for the roots $\rho_{1(-)}$ and $\rho_{2(\pm)}$ and using the explicit form of matrix B, one finds:

$$\boldsymbol{\xi}_{j(\pm)} \equiv \begin{pmatrix} \underline{\xi}^{+}_{j(\pm)} \\ \underline{\xi}^{-}_{j(\pm)} \end{pmatrix}, \qquad j = 1, 2, \tag{36a}$$

$$\underline{\xi}^{+}_{1(\pm)} = \underline{\xi}^{-}_{2(\mp)} = \begin{pmatrix} 1 \\ \pm i \end{pmatrix}, \qquad \underline{\xi}^{-}_{1(\pm)} = -\underline{\xi}^{+}_{2(\mp)} = \frac{h}{\lambda^2 - 1} \begin{pmatrix} \pm 2i \\ 1 \end{pmatrix}. \tag{36b}$$

Recall that, as stated after (28), the super- and subscript $\pm$ notations are unrelated to each other. The superscripts '$\pm$' are used by analogy with that notation in (13) and thus denote the "forward"- and "backward"-propagating components of the numerical error. On the other hand, the subscripts $(\pm)$ simply refer to distinct roots in (28) and in (36b).

4.3 Step 3: Finding eigenvalues of $\mathbb{N}$ in (17)

Here we will obtain the main result of Section 4; it is given in qualitative form by relation (45) and in quantitative form by Eq. (43).

As we explained in Section 3, the eigenvalues of the amplification matrix $\mathbb{N}$ are found from (24). Substituting the eigenvectors (36) (with $\boldsymbol{\xi}_1 \equiv \boldsymbol{\xi}_{1(+)}$, $\boldsymbol{\xi}_2 \equiv \boldsymbol{\xi}_{1(-)}$, etc.) and roots (28) into (23b), one obtains:

$$\boldsymbol{\Phi}(\lambda) \equiv \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix}, \tag{37a}$$

$$\Phi_{11} = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}, \qquad \Phi_{12} = \frac{h}{1-\lambda^2} \begin{pmatrix} -2i & 2i \\ 1 & 1 \end{pmatrix}, \tag{37b}$$

$$\Phi_{21} = \frac{-h\lambda^{-M}}{1-\lambda^2} \begin{pmatrix} 2i\widehat{\rho}^{M}_{1(+)} & -2i\widehat{\rho}^{M}_{1(-)} \\ \widehat{\rho}^{M}_{1(+)} & \widehat{\rho}^{M}_{1(-)} \end{pmatrix}, \qquad \Phi_{22} = \lambda^{M} \begin{pmatrix} \widehat{\rho}^{M}_{2(+)} & \widehat{\rho}^{M}_{2(-)} \\ -i\widehat{\rho}^{M}_{2(+)} & i\widehat{\rho}^{M}_{2(-)} \end{pmatrix}.$$

Recall that $(M+1)$ is the number of points in the spatial grid.

Before we proceed with finding λ from (24), let us demonstrate that a naive perturbation expansion of $\det \boldsymbol{\Phi}$ in powers of h will fail. Understanding the reason for that will help us guess the forthcoming main result of this section. Setting $h \to 0$ in (37b) and using the asymptotic expression $\widehat{\rho}^{M}_{1(\pm)} = \widehat{\rho}^{M}_{2(\mp)} \to \exp[\mp iL]$ (since $M = L/h$), one sees that Φ_{12}, $\Phi_{21} \to \mathcal{O}$ and Φ_{11}, Φ_{22} are non-singular matrices. Hence in the limit $h \to 0$, (24) formally yields $\lim_{h\to 0} \lambda^{2M} = 0$. This, however, would invalidate the preceding "result" that $\Phi_{21} \to \mathcal{O}$, because $\Phi_{21} \propto \lambda^{-M}$. This inconsistency suggests that a more subtle calculation is required and, moreover, that one should expect a relation $\lambda^M = O(h^a)$ for some $a > 0$. In what follows we will obtain a quantitative form of this result.

Using the identity

$$\det \boldsymbol{\Phi} = \det \Phi_{11} \, \det \Phi_{22} \, \det \left(I - \Phi_{22}^{-1} \Phi_{21} \Phi_{11}^{-1} \Phi_{12} \right) \tag{38}$$

and (37), one transforms (24) to:

$$\det\left[I-\left(\frac{h}{2(\lambda^2-1)\lambda^M}\right)^2\begin{pmatrix}-3(\widehat{\rho}_{1(+)}/\widehat{\rho}_{2(+)})^M & (\widehat{\rho}_{1(-)}/\widehat{\rho}_{2(+)})^M\\ -(\widehat{\rho}_{1(+)}/\widehat{\rho}_{2(-)})^M & 3(\widehat{\rho}_{1(-)}/\widehat{\rho}_{2(-)})^M\end{pmatrix}\begin{pmatrix}3 & -1\\ 1 & -3\end{pmatrix}\right]=0. \tag{39}$$

Using the approximations

$$\left(\frac{\widehat{\rho}_{1(\pm)}}{\widehat{\rho}_{2(\pm)}}\right)^M=e^{\mp 2iL+3Lh},\qquad \left(\frac{\widehat{\rho}_{1(\pm)}}{\widehat{\rho}_{2(\mp)}}\right)^M=e^{3Lh}, \tag{40}$$

derived in Appendix B,[7] one obtains from (39):

$$\lambda^{2M}(\lambda^2-1)^2=h^2e^{3Lh}z, \tag{41}$$

where z is either root of a quadratic equation

$$2z^2+(9\cos(2L)-1)z+8=0; \tag{42a}$$

whence

$$z_{1,2}=\Big((1-9\cos(2L))\pm\sqrt{(1-9\cos(2L))^2-64}\Big)\ /4. \tag{42b}$$

A simple analysis (e.g., plotting $|z(L)|$) shows that $|z|\in[1,\,4]$; i.e., $z=O(1)$.

Let us rewrite (41) as

$$\lambda^M=h\,\frac{e^{3Lh/2}\,z^{1/2}}{\lambda^2-1}, \tag{43}$$

where $(\cdots)^{1/2}$ can denote either of the two values of the square root; thus, given (42b), $z^{1/2}$ can take on four distinct values. In our numerical simulations, we had $Lh=O(1)$, and hence

$$e^{Lh}=O(1). \tag{44}$$

Remembering from (29b) that $\lambda^2-1=O(1)$, one has from (43) that

$$|\lambda|^M=O(h). \tag{45}$$

Since $M=L/h\gg 1$, this has two consequences. First,

$$|\lambda|=1+O\left(\frac{h}{L}\ln h\right)\qquad\Rightarrow\qquad|\lambda|\approx 1. \tag{46}$$

Second, after (t/h) time steps, the magnitude of the considered modes is proportional to

$$|\lambda|^{t/h}=\left(|\lambda|^M\right)^{t/(Mh)}=O\left(h^{t/L}\right)\ll 1, \tag{47}$$

which means that these modes decay in time, i.e., are stable.

To conclude this subsection and prepare the ground for the next one, let us show how (43) can be solved approximately. It follows from (46) that λ^2 lies inside, and very close to, the unit circle

[7] and, to be precise, valid only for those modes for which we will perform verification of analytical results by direct numerics in Section 4.4

in the complex plane. Recall from (29b) that our analysis is valid away from the $O(h)$-vicinity of the point $\lambda^2 = 1$. One can rewrite the factor in the denominator of (43) as:

$$\lambda^2 - 1 = |\lambda^2 - 1| \cdot \exp[i \arg(\lambda^2 - 1)],$$

where $|\lambda^2 - 1| \gg h$ and $\arg(\lambda^2 - 1) \in [\pi/2,\, 3\pi/2]$ (since $|\lambda|^2 < 1$). Then, the M distinct roots of (43) are:

$$\lambda_l = \left| \frac{e^{3Lh/2}\, h\, |z|^{1/2}}{|\lambda_l^2 - 1|} \right|^{1/M} \cdot \exp\left[i\, \frac{\arg(z)/2 - \arg(\lambda^2 - 1)}{M} + i\, \frac{2\pi l}{M} \right], \qquad l = 0, 1, \ldots, M-1. \quad (48)$$

Note that the first term in the exponent is $O(1/M)$ and hence can be neglected. Since $z^{1/2}$ can take on four distinct values (see text after (43)), then (48) yields $4M$ eigenvalues λ_l of the matrix $\mathbb{N}$ in (17). (This is consistent with the dimension, $4(M+1) \times 4(M+1)$, of that matrix, given that four entries of $\mathbb{s}$ are fixed by the BC (14).) In the next subsection we will use direct simulations to verify expression (48) for $l \approx M/4$, where $\arg(\lambda^2) \approx \pi$ and hence $|\lambda^2 - 1| \approx 2$. These eigenvalues correspond to the "middle" of the spectral domain, i.e., $kh \approx \pi/2$ (see the text before (31)), in the periodic problem, shown in Fig. 1(a). This particular choice of $\arg(\lambda^2)$ is just the matter of convenience and does not limit our analysis, as long as condition (29b) holds.

4.4 Verification of relation (45)

More specifically, we will verify that in the "middle" of the spectrum of the eigenvalue problem (17), (14) one has

$$|\lambda|^{2M} \approx h^2\, \frac{e^{3Lh}\, |z|}{4}, \quad (49)$$

where the denominator is the value of $|\lambda^2 - 1|^2$ for

$$\lambda^2 \approx -1. \quad (50)$$

Verification of (49) requires several steps.

Step 1 We begin by extracting modes with $\lambda^2 \approx -1$ from the numerical solution of Eqs. (3) by the MoC-SE. By (26), these modes have $\rho \approx \pm i$, which according to the text before (31) means that they correspond to harmonics with $kh \approx \pi/2$ in the periodic problem. We first compute the numerical error by subtracting the exact solution (4) from the numerical solution obtained by the MoC-SE (11), (12). To force periodic BC on that error, so as to enable the application of Fourier transform, we multiply the error by a super-Gaussian "window" $f(x) = \exp\left[-(x/(L/3))^8\right]$. In what follows we will refer to the so modified error as just "error" (without the quotation marks). The Fourier spectrum of a representative error is shown in Fig. 2. It confirms that while the modes with $kh \approx 0$ and π (i.e., $\rho \approx \pm 1$) grow (see the end of Section 4.1), the rest of the modes decay, in qualitative agreement with (47). As mentioned above, we will focus on the harmonics with $kh \approx \pi/2$.

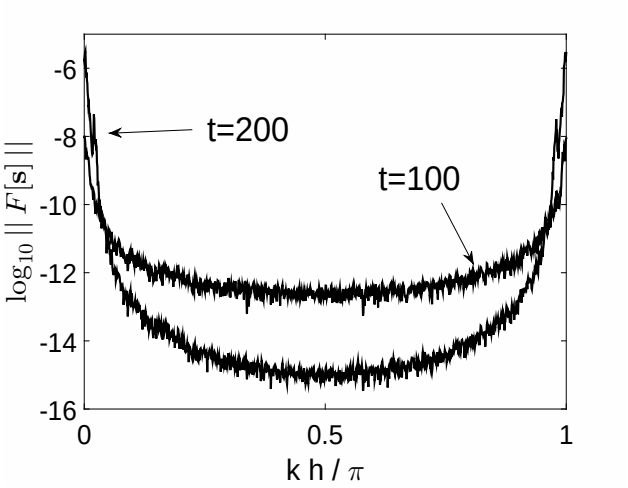

Figure 2: Spectrum of the modified (as per Step 1 in Section 4.4) numerical error for the MoC-SE with nonreflecting BC at two different times. Simulation parameters: $L = 50$, $h = 0.02$.

Step 2 In order to smooth out a noisy profile of the error, we used its value averaged over $(2m_{\rm ave} + 1)$ wavenumbers:

$$\|F[\mathbf{s}]\|_{\rm meas} \equiv \sqrt{\frac{1}{2m_{\rm ave}+1} \sum_{m=-m_{\rm ave}}^{m=m_{\rm ave}} \left\| F[\mathbf{s}] \left(\frac{\pi}{2h} + m\Delta k \right) \right\|^2}, \tag{51}$$

where $F[\cdots]$ denotes the Fourier transform, $\|\cdots\|$ is the Euclidean norm of a vector, and $\Delta k = 2\pi/L$ is the spectral grid spacing. The value $m_{\rm ave}$ has little effect on the numerical coefficient in formula (49) as long as $m_{\rm ave} \ll M/4$; we used $m_{\rm ave} = 20$. The time evolution of $\|F[\mathbf{s}]\|_{\rm meas}$ is shown in Fig. 3(a) for a few values of L and h.

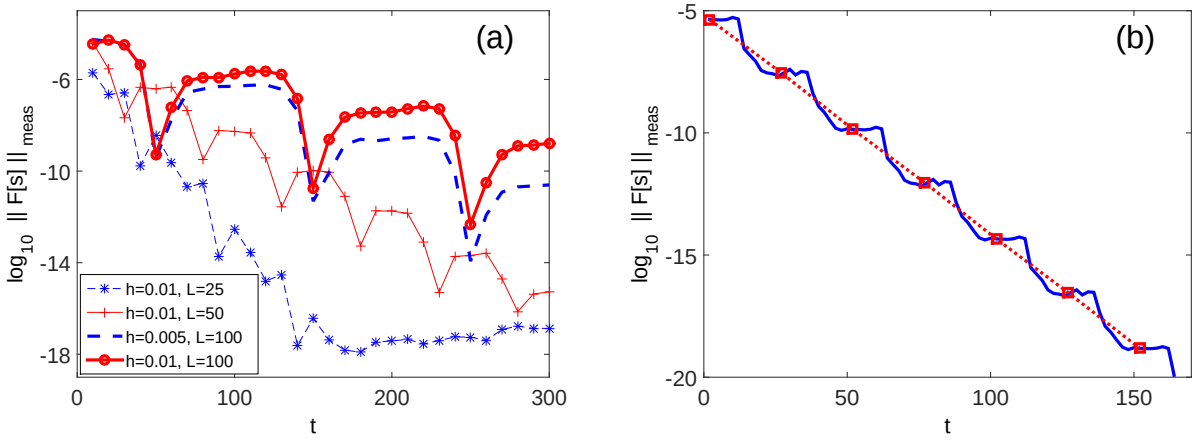

Figure 3: (a) Evolution of error (51) for the values of L and h listed in the legend. (b) Illustration of how λ is found from such an evolution; see text for details. Simulation parameters are: $L = 25$, $h = 0.00625$.

Step 3 The "staircase" shape of the curves in this Figure is explained as follows. At point $x_m = mh$ of the spatial grid, the size of any any given mode of the error is proportional to $|\rho^m \lambda^n|$:

see (18) and (19). Recalling from (26) that $\rho \approx \lambda^{-1}$ or λ, this gives $|\lambda|^{n-m}$ or $|\lambda|^{n+m}$. These expressions correspond to the error's *propagation*, rather than mere decay, to the right and to the left, respectively. Due to the finite size of the domain, the error eventually exits it. (It is not reflected back into the domain because of the special kind — nonreflecting — BC that we consider here.) As a given "batch" of error leaves the domain, the magnitude of the error inside the domain drops (as in a jump); then some "leftover" error from the middle of the domain moves towards its boundaries, exits it, and so on.

Note that the temporal period of the error's evolution in Fig. 3(a) equals the length of the domain; this is consistent with the above scenario. Therefore, to find λ from the error's evolution plots, we "straighten" each "staircase" curve by taking the data points every $t = L$ time units and perform a linear regression of these data. This is illustrated in Fig. 3(b). Then $|\lambda|$ was computed from the slope r of the resulting line (the dotted line in Fig. 3(b)) as $|\lambda| = 10^{hr}$, which follows from (47).

Step 4 As an initial confirmation of (49), we first note that the slope of $\ln|\lambda|^{2M}$ vs. $\ln h$ should be $(2+3Lh)$, which for $Lh \ll 1$ is approximately 2. This is verified in Fig. 4(a). Then in Fig. 4(b) we present a more detailed comparison between $|\lambda|^{2M}$ predicted by (49) and that obtained from direct numerical simulations. The agreement between the two is seen to be good.

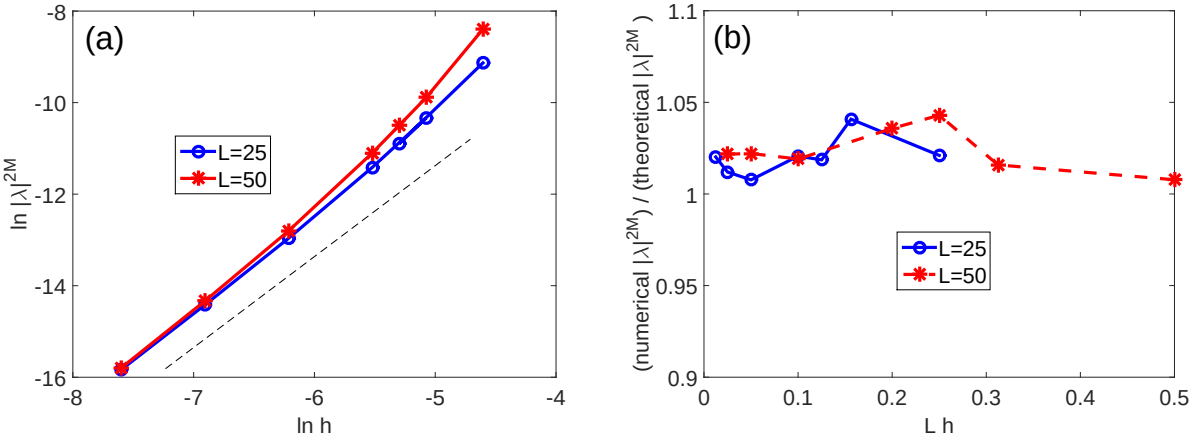

Figure 4: (a) Qualitative confirmation of (49) by direct numerical simulations; see text. For reference, $\ln 0.01 \approx -4.6$ and $\ln 0.0005 \approx -7.6$. The thin dashed line shows a line with slope 2. (b) Quantitative comparison (ratio) between $|\lambda|^{2M}$ computed by direct numerics and by the theory, i.e., Eq. (49). In both panels the symbols represent measured data, while the lines connecting them are guides for the eye.

4.5 Insight into instability suppression in MoC-ME with nonreflecting BC

The following is a heuristic description of how the imposition of nonreflecting BC in the MoC-SE changes the stability of some of the modes compared to the case of periodic BC. After presenting that description, we will apply it to explain why we expect a suppression of the instability for the MoC-ME, even though the instability of the MoC-SE is not suppressed by nonreflecting BC.

The solid line in Fig. 5(a) shows the amplification factor of the MoC-SE with periodic BC (same as in Fig. 1(a)). While all the modes are unstable, the strongest instability occurs for $kh = 0$ and π. The dashed line shows schematically the locus of the modes for which the imposition of nonreflecting BC suppresses the instability via (43), (47). Recall here that the analysis that led to (43) if valid only for the modes sufficiently "away" from having $\rho = \pm 1$, or, equivalently, $kh = 0$ or π in Fig. 5(a); this is why the "box" there does *not* occupy the entire interval $[0, \pi]$. Figure 5(b) shows schematically the effect of the imposition of nonreflecting BC on the amplification factor: That factor becomes less than unity for the modes inside the "box" in panel (a). Thus, the instability of those modes is suppressed. However, this does not affect the most unstable modes with $kh = 0$ and π. Therefore, the observed instability of the MoC-SE is the same for the periodic and nonreflecting BC.

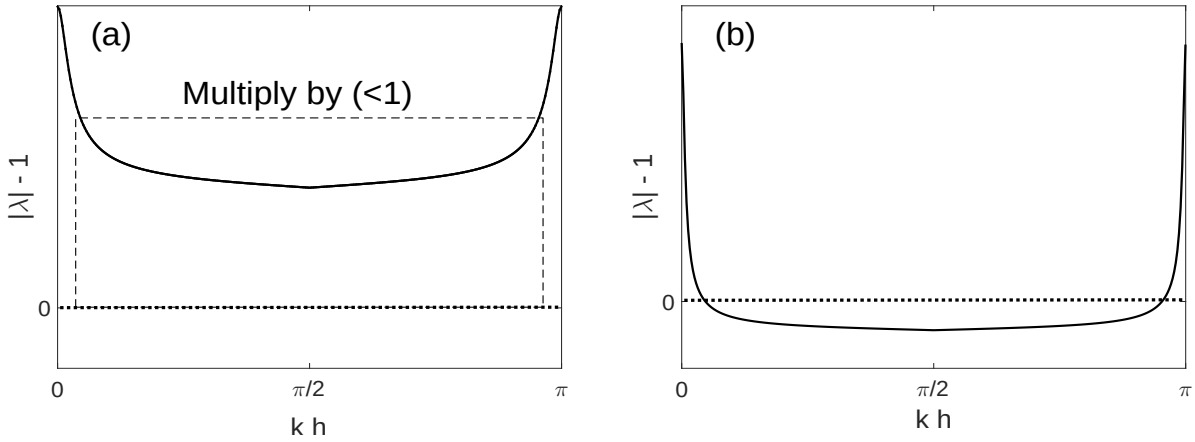

Figure 5: Schematics of instability suppression for MoC-SE; see Section 4.5 for details.

Applying the same concept to the MoC-ME and using Fig. 1(b), we obtain the result shown in Fig. 6(a). Assuming that the nonreflecting BC suppress instability in the MoC-SE and MoC-ME in a similar way, we see in Fig. 6(b) that the instability of the *most* unstable modes of the MoC-ME is to be suppressed. The modes with $\arg(\rho) \approx 0$ and π may remain unstable with nonreflecting BC, but their instability is much weaker than that of modes with $\arg(\rho) \approx \pi/2$ (see Section 4 in [1]). Such a weak instability can be ignored unless the simulation time is very long.

5 Stability analysis of the MoC-ME

Here we will follow the steps of Sections 3 and 4. We will skip most of the details which are similar for the MoC-ME and MoC-SE and will emphasize only the essential differences. To keep the numeration of subsections in Sections 5 and 4 the same, we will present the setup for the MoC-ME in the forthcoming preamble. The main result of our analysis, indicating suppression of the most strongly unstable modes, will be obtained in Section 5.3. It will then be verified numerically in Section 5.4.

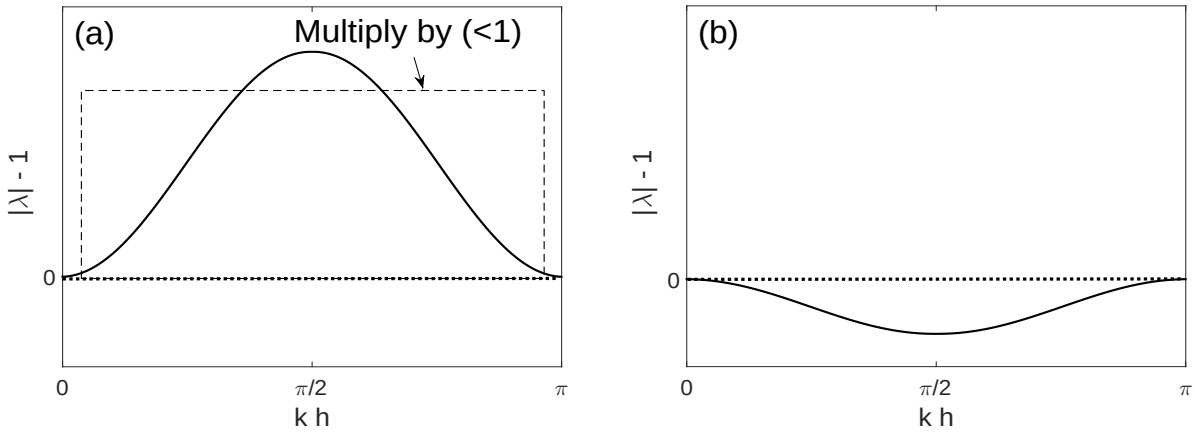

Figure 6: Schematics of instability suppression for MoC-ME; see Section 4.5 for details.

The MoC-ME scheme for Eqs. (3) is:

$$\overline{S_j^{\pm}}_{\,m} = (S_j^{\pm})_{m\mp1}^{n} + h\, f_j^{\pm}\big(\,(\underline{\mathbf{S}}^{+})_{m\mp1}^{n},\,(\underline{\mathbf{S}}^{-})_{m\mp1}^{n}\,\big), \qquad j=1,2,3; \tag{52a}$$

$$(S_j^{\pm})_{m}^{n+1} = \frac{1}{2}\left[\,(S_j^{\pm})_{m\mp1}^{n} + \overline{S_j^{\pm}}_{\,m} + h\, f_j^{\pm}\Big(\overline{\underline{\mathbf{S}}^{+}}_{\,m},\,\overline{\underline{\mathbf{S}}^{-}}_{\,m}\Big)\,\right], \tag{52b}$$

where $f_j^{\pm}$ are the nonlinear functions on the rhs of (3). The nonreflecting BC have the form (12). The numerical error satisfies the linearized form of (52):

$$\overline{\mathbf{s}}_m = \begin{pmatrix} I+hP^{++} & hP^{+-} \\ \mathcal{O} & \mathcal{O} \end{pmatrix} (\mathbf{s})_{m-1}^{n} + \begin{pmatrix} \mathcal{O} & \mathcal{O} \\ hP^{-+} & I+hP^{--} \end{pmatrix} (\mathbf{s})_{m+1}^{n} \tag{53a}$$

$$(\mathbf{s})_{m}^{n+1} = \frac{1}{2}\left(\begin{pmatrix} I & \mathcal{O} \\ \mathcal{O} & \mathcal{O} \end{pmatrix} (\mathbf{s})_{m-1}^{n} + \begin{pmatrix} \mathcal{O} & \mathcal{O} \\ \mathcal{O} & I \end{pmatrix} (\mathbf{s})_{m+1}^{n} + (\mathbf{I}+h\mathbf{P})\overline{\mathbf{s}}_m \right), \tag{53b}$$

with the BC for it being given by (14). Equations (53a) can be cast in the form (15a), where now

$$\mathbf{\Gamma} = \mathbf{\Gamma}_0 + h\mathbf{\Gamma}_1 + \frac{1}{2}h^2\mathbf{\Gamma}_2, \qquad \mathbf{\Omega} = \mathbf{\Omega}_0 + h\mathbf{\Omega}_1 + \frac{1}{2}h^2\mathbf{\Omega}_2; \tag{54a}$$

$$\begin{aligned} \mathbf{\Gamma}_1 &= \begin{pmatrix} P^{++} & \frac{1}{2}P^{+-} \\ \frac{1}{2}P^{-+} & \mathcal{O} \end{pmatrix}, \qquad \mathbf{\Gamma}_2 = \mathbf{P}\begin{pmatrix} P^{++} & P^{+-} \\ \mathcal{O} & \mathcal{O} \end{pmatrix}; \\ \mathbf{\Omega}_1 &= \begin{pmatrix} \mathcal{O} & \frac{1}{2}P^{+-} \\ \frac{1}{2}P^{-+} & P^{--} \end{pmatrix}, \qquad \mathbf{\Omega}_2 = \mathbf{P}\begin{pmatrix} \mathcal{O} & \mathcal{O} \\ P^{-+} & P^{--} \end{pmatrix}; \end{aligned} \tag{54b}$$

and $\mathbf{\Gamma}_0$ and $\mathbf{\Omega}_0$ are as in (16b). (It should be noted that ($\mathbf{\Gamma}_1$, $\mathbf{\Omega}_1$) being different for the MoC-SE and MoC-ME does not contradict the MoC-ME being a higher-order extension of the MoC-SE. Indeed, one can easily check that up to the order $O(h)$, Eqs. (15a) with (16a) are the same as (15a) with (54).) Equations (17)–(20) have the same form for the MoC-ME, where now $\mathbf{\Gamma}$ and $\mathbf{\Omega}$ are defined by (54). Then, the paragraph of Section 3 that contains Eqs. (21)–(24) applies to the MoC-ME verbatim. We will now implement the three Steps listed in that Section.

5.1 Step 1: Finding $\rho(\lambda)$ in (20)

The characteristic equation (20) for the MoC-ME coincides with that for the MoC-SE, given by (25), in the orders $O(1)$ and $O(h)$. Therefore, one seeks its solution in the form (26), (27). Unlike in Section 4, here we will *not* obtain explicit expressions for $\rho_j^{(2)}$ as they will not be able to facilitate a quantitative comparison of the analytical and numerical results. However, in our analysis we will need to refer to the *presence* of such terms, and therefore below we will include them in some of the formulas. In the order $O(h)$ in (27), the expressions for $\rho_j^{(1)}$ are the same as for the MoC-SE. Thus, the counterpart of (28) for the MoC-ME is:

$$\rho_{1(\pm)} = \frac{1}{\lambda}\left(1 \mp ih + h^2\rho_{1(\pm)}^{(2)}\right) \equiv \frac{1}{\lambda}\,\widehat{\rho}_{1(\pm)}, \qquad \rho_{2(\pm)} = \lambda\left(1 \pm ih + h^2\rho_{2(\pm)}^{(2)}\right) \equiv \lambda\,\widehat{\rho}_{2(\pm)}. \tag{55}$$

As in Section 4, we will *not* analyze the case where $\rho_{1(\pm)} \approx \rho_{2(\pm)}$, or, equivalently, $\lambda^2 \approx \rho^2 \approx 1$. The reasons for this are as follows.

(i) The numerical instability for modes with $\rho \approx \pm 1$ will still occur. However, it corresponds to the instability in the ODE- and anti-ODE limits of the von Neumann analysis of the periodic problem and thus will be much weaker than the instability (in the periodic problem) for modes with $\rho = e^{ikh} \approx e^{i\pi/2}$: see Section 4.5 above and also Section 4 in [1]. As noted there, that weak instability will be inconsequential for moderately long simulation times considered here.

(ii) As we previewed in Section 4.5 and will demonstrate below, our forthcoming analysis, based on (55), will succeed in predicting the instability suppression for the modes with $\rho \approx e^{i\pi/2}$, which are the most unstable in the periodic problem. In other words, that analysis will describe the most important phenomenon that motivated this entire study.

(iii) The analysis for the case $\lambda^2 \approx 1$ is much more technical than the one given below. Since its results will not affect the main conclusions of the analysis based on (55), then it seems appropriate not to carry it out here in order not to distract the reader's attention. Thus, we will proceed assuming the condition (29b).

5.2 Step 2: Finding $\boldsymbol{\xi}$ in (20)

As in Section 4.2, we will outline the derivation of only $\boldsymbol{\xi}_{1(+)}$ and will simply state the results for the other three eigenvectors. Similarly to (33), we seek

$$\boldsymbol{\xi}_{1(+)} = \begin{pmatrix} \underline{u}^{(0)} + h\underline{u}^{(1)} + h^2\underline{u}^{(2)} \\ h\underline{v}^{(1)} + h^2\underline{v}^{(2)} \end{pmatrix}. \tag{56}$$

We will compute only $\underline{u}^{(0)}$ and $\underline{v}^{(1)}$. The other terms in (56) will enter some of the expressions in Section 5.3. We will *not* need their precise values, but their *form* will require $\underline{u}^{(1),(2)}$, $\underline{v}^{(2)}$ to be defined; this is why we listed them in (56).

Similarly to Section 4.2, the order $O(1)$ in the equation obtained by substitution of (56) into (20) leaves $\underline{u}^{(0)}$ undefined. The order $O(h)$ yields precisely (34), where $\boldsymbol{\Gamma}_1$, $\boldsymbol{\Omega}_1$ are defined by (54b) instead of (16b). It is the difference in the *structure* of these expressions that will cause a *substantial*

difference between the results for the MoC-ME and MoC-SE. The vector $\underline{u}^{(0)}$ in (56) is found to still be given by (35a), but the counterpart of (35b) is now:

$$\underline{v}^{(1)} = \frac{1}{2}\,\frac{1+\lambda^2}{1-\lambda^2}\,B\,\underline{u}^{(0)}. \tag{57}$$

The numerator in the fraction above, which is absent in (35b), will turn out to be "responsible" for the aforementioned substantial difference.

If one continues expanding Eqs. (20), (54)–(56) in powers of h, in the order $O(h^2)$ one can determine $\underline{u}^{(1)}$, $\underline{v}^{(2)}$, and in the order $O(h^3)$, $\underline{u}^{(2)}$. Putting the above information together and performing similar calculations for $\boldsymbol{\xi}_{1(-)}$ and $\boldsymbol{\xi}_{2(\pm)}$, one obtains:

$$\begin{aligned}
\boldsymbol{\xi}_{1(\pm)} &= \begin{pmatrix} \begin{pmatrix} 1 \\ \pm i \end{pmatrix} + h\underline{u}^{(1)}_{1(\pm)} \\ \dfrac{h}{2}\dfrac{\lambda^2+1}{\lambda^2-1}\begin{pmatrix} \pm 2i \\ 1 \end{pmatrix} \end{pmatrix} + h^2\boldsymbol{\xi}^{(2)}_{1(\pm)}\,, \\
\boldsymbol{\xi}_{2(\pm)} &= \begin{pmatrix} -\dfrac{h}{2}\dfrac{\lambda^2+1}{\lambda^2-1}\begin{pmatrix} \mp 2i \\ 1 \end{pmatrix} \\ \begin{pmatrix} 1 \\ \mp i \end{pmatrix} + h\underline{u}^{(1)}_{2(\pm)} \end{pmatrix} + h^2\boldsymbol{\xi}^{(2)}_{2(\pm)}\,,
\end{aligned} \tag{58}$$

where

$$\boldsymbol{\xi}^{(2)}_{1(\pm)} \equiv \begin{pmatrix} \underline{u}^{(2)}_{1(\pm)} \\ \underline{v}^{(2)}_{1(\pm)} \end{pmatrix}, \qquad \boldsymbol{\xi}^{(2)}_{2(\pm)} \equiv \begin{pmatrix} \underline{v}^{(2)}_{2(\pm)} \\ \underline{u}^{(2)}_{2(\pm)} \end{pmatrix}.$$

5.3 Step 3: Finding eigenvalues of $\mathbb{N}$ in (17)

Here we will obtain the main result of Section 5, which will be given by relation (63).

The equation from which the eigenvalues of the amplification matrix $\mathbb{N}$ are to be found is (24), where matrix $\boldsymbol{\Phi}$ is defined by (23b). Substituting there expressions (55) and (58), we obtain a counterpart of (37). For reasons to be explained soon, in writing it, we will specify the first two h-orders of the entries (and will also slightly change the notations):

$$\boldsymbol{\Phi}(\lambda) \equiv \begin{pmatrix} \Phi^{(0)}_{11} + h\Phi^{(1)}_{11} & h\Phi^{(1)}_{12} + h^2\Phi^{(2)}_{12} \\ h\Phi^{(1)}_{21} + h^2\Phi^{(2)}_{21} & \Phi^{(0)}_{22} + h\Phi^{(1)}_{22} \end{pmatrix}, \tag{59a}$$

$$\Phi^{(0)}_{11} = \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix}, \qquad \Phi^{(1)}_{11} = \left[\underline{u}^{(1)}_{1(+)},\, \underline{u}^{(1)}_{1(-)}\right], \tag{59b}$$

$$\Phi^{(1)}_{12} = -\frac{1}{2}\,\frac{\lambda^2+1}{\lambda^2-1}\begin{pmatrix} -2i & 2i \\ 1 & 1 \end{pmatrix}, \qquad \Phi^{(2)}_{12} = \left[\underline{u}^{(2)}_{2(+)},\, \underline{u}^{(2)}_{2(-)}\right], \tag{59c}$$

$$\Phi^{(1)}_{21} = \frac{\lambda^{-M}}{2}\,\frac{\lambda^2+1}{\lambda^2-1}\begin{pmatrix} 2i\widehat{\rho}^{M}_{1(+)} & -2i\widehat{\rho}^{M}_{1(-)} \\ \widehat{\rho}^{M}_{1(+)} & \widehat{\rho}^{M}_{1(-)} \end{pmatrix}, \qquad \Phi^{(2)}_{21} = \lambda^{-M}\left[\widehat{\rho}^{M}_{1(+)}\underline{v}^{(2)}_{1(+)},\, \widehat{\rho}^{M}_{1(-)}\underline{v}^{(2)}_{1(-)}\right], \tag{59d}$$

$$\Phi^{(0)}_{22} = \lambda^{M}\begin{pmatrix} \widehat{\rho}^{M}_{2(+)} & \widehat{\rho}^{M}_{2(-)} \\ -i\widehat{\rho}^{M}_{2(+)} & i\widehat{\rho}^{M}_{2(-)} \end{pmatrix}, \qquad \Phi^{(1)}_{22} = \lambda^{M}\left[\widehat{\rho}^{M}_{2(+)}\underline{u}^{(1)}_{2(+)},\, \widehat{\rho}^{M}_{2(-)}\underline{u}^{(1)}_{2(-)}\right]. \tag{59e}$$

Note that the main-order entries above are the same as their counterparts in (37), except that $\Phi_{12}^{(1)}$, $\Phi_{21}^{(1)}$ have the extra factor $(\lambda^2+1)/2$. It is this factor that will substantially change the behavior of modes with $\rho \approx e^{\pm i\pi/2}$ in the MoC-ME compared to those modes in the MoC-SE. Therefore, we will now discuss what difference in the analysis this factor makes.

When $\lambda^2+1 = O(1)$, i.e., when $\lambda \approx \rho \approx \pm i = \exp[\pm i\pi/2]$, the analysis of the MoC-ME proceeds exactly as in Section 4.3, with straightforward adjustment of some numeric coefficients. Most importantly, for modes with $\rho \approx \pm i$ (and for those with $\rho \approx \pm 1$; see Section 5.1), the main result of Section 4, Eq. (45): $|\lambda|^M = O(h)$, remains valid. With time, these modes decay as $O(h^{t/L})$ (see (47)), just as they do for the MoC-SE. However, the modes in the narrow interval where

$$\lambda^2 + 1 = O(h) \qquad \Rightarrow \qquad \rho = e^{\pm i\pi/2} + O(h) \tag{60}$$

will be shown to decay even faster, as $O(h^{2t/L})$. This, of course, will not change the fact that these modes are stable in the MoC-ME, as they are in the MoC-SE, but it *will* change the Fourier spectrum of the numerical error, as we will show below.

Let us explain why condition (60) affects $|\lambda|^M$. Under this condition, the two terms in each of the off-diagonal entries in (59a) both become $O(h^2)$: see (59c) and (59d). Therefore, in this case it is more appropriate to rewrite (59a) as

$$\mathbf{\Phi}(\lambda) \equiv \begin{pmatrix} \Phi_{11}^{(0)} + h\Phi_{11}^{(1)} & h^2\widetilde{\Phi}_{12}^{(2)} \\ h^2\,\lambda^{-M}\,\widetilde{\Phi}_{21}^{(2)} & \lambda^M\big(\widetilde{\Phi}_{22}^{(0)} + h\widetilde{\Phi}_{22}^{(1)}\big) \end{pmatrix}, \tag{61}$$

where by introducing the tilde-notations in the second row, we have also accounted for the λ^M-dependence. The calculation of $\widetilde{\Phi}_{12}^{(2)}$, $\widetilde{\Phi}_{21}^{(2)}$ would require finding the higher-order correction terms, denoted by $\boldsymbol{\xi}_{j(\pm)}^{(2)}$ in (58). This would be quite a tedious task, which would not provide any additional qualitative insight into the evolution of the modes (60). Therefore, we will not carry it out but will proceed with the general form (61). Furthermore, we will neglect the $O(h)$-terms in the diagonal entries of (61) because $\Phi_{11}^{(0)}$ and $\widetilde{\Phi}_{22}^{(0)}$ are nonsingular (see (59b) and (59e)). Then, repeating the steps that led to (39), we obtain:

$$\det\left[I - \frac{h^4}{\lambda^{2M}}\left(\widetilde{\Phi}_{22}^{(0)}\right)^{-1}\widetilde{\Phi}_{21}^{(2)}\left(\Phi_{11}^{(0)}\right)^{-1}\widetilde{\Phi}_{12}^{(2)}\right] = 0. \tag{62}$$

While we do not compute the explicit form of two out of the four matrices in (62), we still know that their entries are of order one. Therefore, (62) can hold only when the coefficient in front of the four-matrix product is also of order one, or, equivalently, when

$$|\lambda|^{2M} = O(h^4). \tag{63}$$

This is the result for modes (60) of the MoC-ME, announced earlier. It predicts a faster temporal decay of these modes than of the rest of the modes, satisfying (45). We will now verify this by direct numerical simulations. Note that unlike in the MoC-SE case (see (49)), here we will *not* attempt to resolve the numeric coefficient in the $O(h^4)$ notation.

5.4 Verification of (63) for modes (60)

We will perform this verification in two different ways. For both ways, we will make the numerical error periodic in space by multiplying it with a "window", as in Step 1 of Section 4.4. This will allow us to work with the Fourier spectrum of the so modified error.

5.4.1 Dependence of spectral components of error on time

The Fourier spectra of the error at equidistant times are shown in Fig. 7. As we announced after (63), modes near $kh = \pi/2$ decay faster than those farther away from $kh = \pi/2$. This faster decay of modes with $kh \approx \pi/2$ is manifested by the deepening "dip" in the spectra. More specifically, a comparison of (63) and (45) shows that the modes near $kh = \pi/2$ decay twice as fast as the modes farther away from $kh = \pi/2$. Indeed, (47) (which in the MoC-ME holds for modes with $kh \not\approx \pi/2$) and (63), for modes with $kh \approx \pi/2$, yield:

$$\ln \|F[\mathbf{s}](kh \not\approx \pi/2)\| \equiv y_{\text{away}} \propto \ln |\lambda_{\text{away}}|^{t/h} = t\,\frac{\ln h}{L} + \text{const}_{\text{away}}; \tag{64a}$$

$$\ln \|F[\mathbf{s}](kh \approx \pi/2)\| \equiv y_{\text{near}} \propto \ln |\lambda_{\text{near}}|^{t/h} = 2t\,\frac{\ln h}{L} + \text{const}_{\text{near}}. \tag{64b}$$

This implies that for any two moments of time, $t_{1,2}$, one is to have:

$$\frac{y_{\text{near}}(t_2) - y_{\text{near}}(t_1)}{y_{\text{away}}(t_2) - y_{\text{away}}(t_1)} = 2. \tag{65a}$$

This observation can be employed to verify the theoretical predictions (64) against the numerical results presented in Fig. 7. Using the numbers listed in that Figure, one has:

$$\frac{y_{\text{near}}(50) - y_{\text{near}}(25)}{y_{\text{away}}(50) - y_{\text{away}}(25)} \approx \frac{y_{\text{near}}(75) - y_{\text{near}}(50)}{y_{\text{away}}(75) - y_{\text{away}}(50)} \approx 1.8, \tag{65b}$$

which is indeed close to what (65a) predicts.

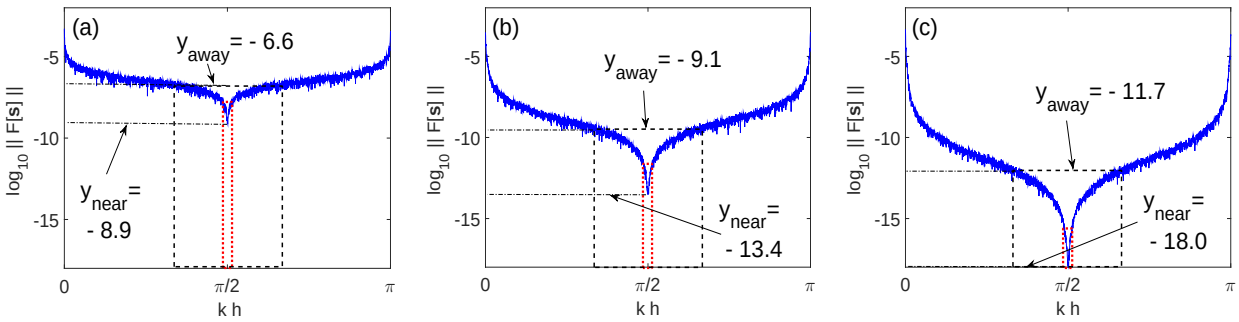

Figure 7: (Color online) Fourier spectra of the error of the MoC-ME for $t = 25$ (a), $t = 50$ (b), and $t = 75$ (c). Other simulation parameters: $L = 25$, $h = 0.05$. Narrower (red, dotted) and wider (black, dashed) boxes illustrate the averaging intervals with $m_{\text{ave}} = [0.025M]$ and $m_{\text{ave}} = [0.1M]$ (where $[\cdots]$ denotes the integer part), respectively, as explained after Eq. (66). Quantities y_{near} and y_{away} are defined in (64).

5.4.2 Dependence of spectral components of error on h

We will now confirm (63) in yet another way. Recall that when we extract $|\lambda|$ from numerically obtained spectra using formula (51), we perform averaging over m_{ave} nodes around the wavenumber of interest. If we focus on such a narrow vicinity of $kh = \pi/2$ that all modes in the averaging box satisfy condition (60) (which requires $m_{\text{ave}} \ll M/4$; see the narrower box in Fig. 7), then we expect that the numerically calculated $|\lambda|$ will satisfy (63). That relation is equivalent to

$$\ln |\lambda|^{2M} = 4 \ln h + \text{“const”}, \tag{66}$$

where the “const” may actually be a slow function of $\ln h$, as in (49), and tend to a “true” constant for $h \to 0$. Thus, the slope of the curve $\ln |\lambda|^{2M}$ vs. $\ln h$ should be (approximately) 4.

On the other hand, when m_{ave} in (51) becomes comparable to $M/4$, the averaging interval, represented by the wider box in Fig. 7, includes mostly the modes satisfying (45). Then the slope of $\ln |\lambda|^{2M}$ vs. $\ln h$ should be (approximately) 2. This is confirmed by Fig. 8, where we used $m_{\text{ave}} = [0.025M]$ and $m_{\text{ave}} = [0.1M]$. Thus, (66) and hence (63) are indeed confirmed.

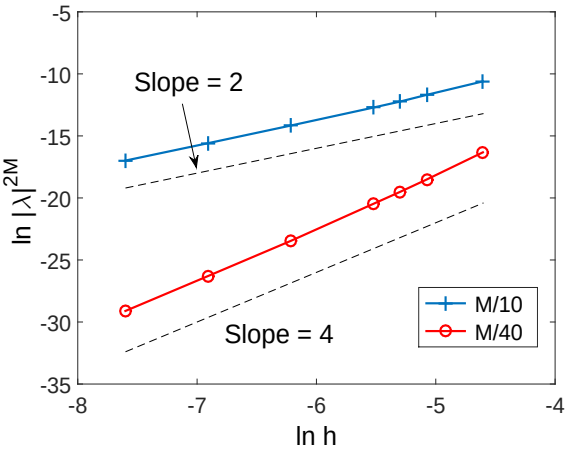

Figure 8: Verification of (66) for $L = 25$, as described in the text. The two sets of data pertain to $m_{\text{ave}} = [0.025M]$ and $m_{\text{ave}} = [0.1M]$. To smooth out the noise (which originates from the small noise added to the initial condition, as described in Section 4.4), each data point was obtained by averaging results from 10 simulations with different realizations of the initial noise. For reference, $\ln 0.01 \approx -4.6$ and $\ln 0.0005 \approx -7.6$.

6 Instability of modes with $\rho \approx \pm i$ for the MoC-LF

In Sections 4 and 5 we have seen that modes in the “middle” of the spectrum (i.e. having $\rho \approx \pm i$), which are unstable in the MoC-SE and MoC-ME schemes with periodic BC (see Fig. 1), are “made” stable by the nonreflecting BC. In contrast, as we announced in the Introduction, direct numerical simulations of the MoC-LF show that the scheme exhibits very similar instability (see Fig. 1(c)) for both periodic and nonreflecting BC. In this section we will explain why this is so. Since the

instability in the periodic case exists only for the modes with $kh \approx \pi/2$, or, equivalently, satisfying (60), we will proceed to find the amplification factor, λ, *only* for these modes.

The MoC-LF scheme is:

$$(S_j^{\pm})_m^{n+1} = (S_j^{\pm})_{m\mp 2}^{n-1} + 2h\, f_j^{\pm}\big((\underline{\mathbf{S}}^{+})_{m\mp 1}^{n},\, (\underline{\mathbf{S}}^{-})_{m\mp 1}^{n} \big), \qquad j = 1, 2, 3; \tag{67}$$

and its linearized form is:

$$(\mathbf{s})_m^{n+1} = \mathbf{\Gamma}_0 (\mathbf{s})_{m-2}^{n-1} + \mathbf{\Omega}_0 (\mathbf{s})_{m+2}^{n-1} + 2h \left(\mathbf{\Gamma}_1 (\mathbf{s})_{m-1}^{n} + \mathbf{\Omega}_1 (\mathbf{s})_{m+1}^{n} \right), \tag{68}$$

where $\mathbf{\Gamma}_{0,1}$ and $\mathbf{\Omega}_{0,1}$ are defined in (16b).

Since the numerical error satisfies the linearized equations, we will state the BC only for them. The MoC-LF scheme involves three time levels, and therefore the BC (14) for $(\underline{s}^{+})_0$ and $(\underline{s}^{-})_M$ need to be supplemented by conditions for $(\underline{s}^{+})_1$ and $(\underline{s}^{-})_{M-1}$. To preserve the second-order accuracy of the MoC-LF scheme, it is logical to compute these values by the MoC-ME. However, we have found in the simulations that using the MoC-SE, MoC-ME, or even a fourth-order approximation based on the classical Runge–Kutta method for finding $(\underline{s}^{+})_1$ and $(\underline{s}^{-})_{M-1}$ does not affect the observed instability of the MoC-LF in any perceptible way. Later on we will see why this is the case. In the meantime, for simplicity, we will consider the BC for $(\underline{s}^{+})_1$ and $(\underline{s}^{-})_{M-1}$ being computed by the MoC-SE:

$$(\underline{s}^{+})_1^{n+1} = (\underline{s}^{+})_0^{n} + h \left(P^{++} (\underline{s}^{+})_0^{n} + P^{+-} (\underline{s}^{-})_0^{n} \right) = h P^{+-} (\underline{s}^{-})_0^{n}\,; \tag{69a}$$

$$(\underline{s}^{-})_{M-1}^{n+1} = (\underline{s}^{-})_M^{n} + h \left(P^{-+} (\underline{s}^{+})_M^{n} + P^{--} (\underline{s}^{-})_M^{n} \right) = h P^{-+} (\underline{s}^{+})_M^{n}\,. \tag{69b}$$

In writing the last step in each of (69a) and (69b), we have used (14). Thus, (14) and (69) form the BC for the MoC-LF.

We will now follow the steps of Section 3 and arrive at a counterpart of the characteristic polynomial (24), whose roots will give the amplification factor λ of the MoC-LF. In Sections 6.1–6.3 we will follow the steps of finding λ as outlined at the end of Section 3.

A counterpart of (20), found from (68), (18), and (19), is:

$$\left(\lambda^{-1} \left(\rho^{-2} \mathbf{\Gamma}_0 + \rho^{2} \mathbf{\Omega}_0 \right) + 2h \left(\rho^{-1} \mathbf{\Gamma}_1 + \rho \mathbf{\Omega}_1 \right) - \lambda \mathbf{I} \right) \boldsymbol{\xi} = \mathbf{0}. \tag{70}$$

In Section 6.1 we will show that the corresponding characteristic equation has *eight* roots $\rho_j(\lambda)$. Since we are only interested in the modes satisfying (60), these roots have the form:

$$\rho_{(\pm,\, j)}(\lambda) = \pm i + h\beta_j(\lambda), \qquad j = 1, \ldots, 4, \tag{71}$$

with $\beta_j = O(1)$, which will be found in Section 6.1. The eigenvectors corresponding to $\rho_{(\pm,\, j)}$ will be denoted by $\boldsymbol{\xi}_{(\pm,\, j)}$. Seeking, similarly to (21),

$$(\mathbf{s})_m^{n} = \lambda^n \sum_{r=+,-} \sum_{j=1}^{4} C_{(r,j)} \boldsymbol{\xi}_{(r,j)}, \tag{72}$$

where $C_{(r,j)}$ are constants, one rewrites the BC (14) as:

$$\sum_{r=+,-}\sum_{j=1}^{4} C_{(r,j)}\underline{\xi}^{+}_{(r,j)} = \underline{0}\,, \tag{73a}$$

$$\sum_{r=+,-}\sum_{j=1}^{4} C_{(r,j)}\rho^{M}_{(r,j)}\underline{\xi}^{-}_{(r,j)} = \underline{0}\,; \tag{73b}$$

and the BC (69) as:

$$\sum_{r=+,-}\sum_{j=1}^{4} C_{(r,j)}(r\cdot i + h\beta_j)\,\underline{\xi}^{+}_{(r,j)} = hP^{+-}\sum_{r=+,-}\sum_{j=1}^{4} C_{(r,j)}\underline{\xi}^{-}_{(r,j)}\,, \tag{74a}$$

$$\sum_{r=+,-}\sum_{j=1}^{4} C_{(r,j)}(r\cdot i + h\beta_j)^{M-1}\,\underline{\xi}^{-}_{(r,j)} = hP^{-+}\sum_{r=+,-}\sum_{j=1}^{4} C_{(r,j)}(r\cdot i + h\beta_j)^{M}\,\underline{\xi}^{+}_{(r,j)}\,. \tag{74b}$$

In principle, arranging (73) and (74) into a matrix equation like (23b) will then allow one to determine the amplification factor $|\lambda|$ via a counterpart of (24). However, before we proceed along these lines, we will simplify the form of (74) in order to simplify subsequent calculations.

In Section 6.1 we will show that there is a set of modes for which $\beta_j \in \mathbb{R}$, as this will suffice for our purpose of explaining the instability of the MoC-LF. In what follows we will consider only such modes. Then, for $h \ll 1$,

$$\rho^{M}_{(r,j)} = (r\cdot i + h\beta_j)^{M} \approx r^{M} i^{M} e^{-r\cdot i\beta_j L}, \tag{75}$$

where we have used $M = L/h$. Next, later on we will show that all entries of $\boldsymbol{\xi}_{(r,j)}$ are $O(1)$. Therefore, coefficients multiplying $C_{(r,j)}$ on the lhs of (74a) have the form '$O(1)+O(h)$', while such coefficients on the rhs are $O(h)$. For this reason, *in the main order*, we will neglect all the $O(h)$ terms in (74a) and also the $O(h)$ terms on the rhs of (74b). Then, on the account of (75), the BC (74) become:

$$\sum_{r=+,-}\sum_{j=1}^{4} r\,C_{(r,j)}\underline{\xi}^{+}_{(r,j)} = \underline{0}\,, \tag{76a}$$

$$\sum_{r=+,-}\sum_{j=1}^{4} r^{M-1}e^{-r\cdot i\beta_j L}\,C_{(r,j)}\underline{\xi}^{-}_{(r,j)} = \underline{0}\,. \tag{76b}$$

Now, Eqs. (73) can be rewritten as

$$\boldsymbol{\Phi}_{(+)}\mathbf{C}_{(+)} + \boldsymbol{\Phi}_{(-)}\mathbf{C}_{(-)} = \mathbf{0}, \tag{77a}$$

where $\mathbf{C}_{(r)} = [C_{(r,1)},\ldots,C_{(r,4)}]^T$ and $\boldsymbol{\Phi}_{(r)}$ are defined as in (23b) using the components of the eigenvectors $\boldsymbol{\xi}_{(r,j)}$. Similarly, (76) can be written as

$$\boldsymbol{\Phi}_{(+)}\mathbf{C}_{(+)} - \boldsymbol{\Phi}_{(-)}\mathbf{C}_{(-)} = \mathbf{0}. \tag{77b}$$

Together, (77) imply that

$$\det \mathbf{\Phi}_{(+)}(\lambda) = 0 \qquad \text{and} \qquad \det \mathbf{\Phi}_{(-)}(\lambda) = 0, \tag{78}$$

which seems to impose two equations on λ. However, in Sections 6.2 and 6.3 we will show that $\boldsymbol{\xi}_{(-,j)}$ are related to $\boldsymbol{\xi}_{(+,j)}$ in such a way that the two equations in (78) are, in fact, the same. Therefore, solving, say, the first of them will yield the values of λ. *Our goal*, to be achieved in Section 6.3, will be to show that these values are very close to the eigenvalues of the unstable modes in the periodic problem.

Let us now note that the discarding of certain $O(h)$ terms, which has reduced (74) to (76), implies that *the order of approximation* with which one computes the boundary values $(\underline{s}^+)_1$ and $(\underline{s}^-)_{M-1}$ is inconsequential for the instability of the MoC-LF. This justifies our earlier corresponding statement, found before (69).

We will now follow the three steps, listed at the end of Section 3, of solving the characteristic equation (78). In our presentations we will focus on their differences from the corresponding calculations in Sections 4 and 5.

6.1 Step 1: Finding $\rho(\lambda)$ in (70)

The characteristic polynomial for (70) is:

$$\left(\rho - \frac{1}{\lambda}\right)^2 \left(\rho + \frac{1}{\lambda}\right)^2 (\rho - \lambda)^2 (\rho + \lambda)^2 + 4h^2\rho^2 \cdot \tag{79}$$
$$\left[\lambda^2 \left(\rho - \frac{1}{\lambda}\right)^2 \left(\rho + \frac{1}{\lambda}\right)^2 + \lambda^{-2} (\rho - \lambda)^2 (\rho + \lambda)^2 - 4\left(\rho - \frac{1}{\lambda}\right)\left(\rho + \frac{1}{\lambda}\right)(\rho - \lambda)(\rho + \lambda)\right] = 0.$$

In the order $O(1)$, it has four double roots:[8]

$$\rho_1^{(0)} = \lambda^{-1}, \quad \rho_2^{(0)} = \lambda, \quad \rho_3^{(0)} = -\lambda^{-1}, \quad \rho_4^{(0)} = -\lambda. \tag{80a}$$

Recall that we are specifically looking for $\rho \approx \pm i$ (see (71)), in which case (80a) implies $\lambda \approx \pm i$ and then

$$\rho_1^{(0)} \approx \rho_4^{(0)}, \qquad \rho_2^{(0)} \approx \rho_3^{(0)}. \tag{80b}$$

Due to this degeneracy, a perturbation expansion in h should proceed *not* similarly to (27), but instead similarly to (30b). Selecting the case $\lambda \approx +i$, we seek solutions of (79) in the form

$$\lambda = i(1 + h\alpha), \qquad \rho = \pm i + h\beta; \tag{81}$$

here α and β are $O(1)$ quantities to be determined below. A similar ansatz can be used for $\lambda \approx -i$.

Below we will present details for the '+' sign in (81) and will only state the final result for the '−' sign. Substituting (81) into (79), in the lowest nontrivial order, $O(h^4)$, one finds:

$$(\alpha^2 + \beta^2)^2 - 2\alpha^2 - 6\beta^2 = 0, \tag{82a}$$

[8] As common for a LF method, half of these roots correspond to the true dynamics of the solution (compare with (26)), while the other half are "parasitic". We will come back to this observation in Section 7.

whence

$$\beta^2 = (3-\alpha^2) \pm \sqrt{9-4\alpha^2}\,. \tag{82b}$$

For the '$-$' sign in (81), one obtains the same relation (82a). Note that (81) and (82) are to be interpreted as follows: For a given λ (or, equivalently, α), which will be found in Section 6.3, one has to find eight modes (i.e., eight values of ρ, with four corresponding to each of the two signs in (81)), which satisfy the nonreflecting BC (14), (69). This appears to be converse to the procedure that one follows in the case of periodic BC. Namely, there, one first finds the modes (Fourier harmonics) $\exp[ikhm]$ with k that are consistent with the periodic BC, and then for those modes finds λ; see, e.g., Section 5 in [1].

Now recall our goal: We want to explain why the instability of the MoC-LF is (almost) the same for the periodic and nonreflecting BC. This means that for the nonreflecting BC, we need to focus on the modes that are counterparts of the Fourier harmonics $\exp[ikhm],\ m = 0, 1, \ldots, M$ in the periodic case. For the ρ in (81) to mimic $\exp[ikh]$ (with $kh \approx \pi/2$), one needs to have $\beta \in \mathbb{R}$. For $\alpha \in \mathbb{R}$ (which will be justified by our results in Section 6.3), this occurs for

$$\alpha \in [\sqrt{2},\ 3/2]; \tag{83}$$

see (82b). Therefore, in what follows we will consider only the range (83) of values of α.

6.2 Step 2: Finding $\boldsymbol{\xi}$ in (70)

Substituting (81) in (70), one finds that all terms in the order $O(1)$ cancel out, and then in the order $O(h)$ one finds:

$$(\,(i\beta-\alpha)\boldsymbol{\Gamma}_0 - (i\beta+\alpha)\boldsymbol{\Omega}_0 + \boldsymbol{\Omega}_1 - \boldsymbol{\Gamma}_0\,)\,\boldsymbol{\xi} = \mathbf{0}. \tag{84}$$

Using the explicit form of $\boldsymbol{\Gamma}_{0,1}$ and $\boldsymbol{\Omega}_{0,1}$ from (16b) and (7) and relation (82a) between α and β, one finds the eigenvector $\boldsymbol{\xi}$. Below we state the result for both signs in (81):

$$\boldsymbol{\xi}_{(\pm)} = \begin{pmatrix} \pm\frac{\alpha\pm i\beta}{\alpha\mp i\beta}\cdot\frac{\alpha\mp 3i\beta}{\alpha^2+\beta^2} \\ -\frac{\alpha\pm i\beta}{\alpha\mp i\beta} \\ \mp\frac{\alpha\pm 3i\beta}{\alpha^2+\beta^2} \\ 1 \end{pmatrix}. \tag{85}$$

As explained before (83), we consider the case $\alpha, \beta \in \mathbb{R}$; then (85) yields:

$$\boldsymbol{\xi}_{(-)} = -\begin{pmatrix} \sigma_3 & \mathcal{O} \\ \mathcal{O} & \sigma_3 \end{pmatrix} \boldsymbol{\xi}^*_{(+)}, \tag{86a}$$

where $\sigma_3 = \mathrm{diag}(1,-1)$ is a Pauli matrix. Also, for future use, note that from (75) it follows that

$$\rho^M_{(-,j)} = \left(\rho^M_{(+,j)}\right)^*, \qquad j = 1, \ldots, 4, \tag{86b}$$

where we have used the fact that all β_j that we consider are real.

6.3 Step 3: Finding amplification factor of MoC-LF from (78)

Here we will obtain the main result of Section 6. Unlike in the previous two sections, here λ will not have an analytical expression, but can be easily found by graphing an analytically defined function.

We will begin by using (85) to show that the two equations in (78) are equivalent. Given the definition of $\mathbf{\Phi}_{(\pm)}$, found after (77a), and using (86), we have:

$$\mathbf{\Phi}_{(-)}(\lambda) = -\begin{pmatrix} \sigma_3 & \mathcal{O} \\ \mathcal{O} & \sigma_3 \end{pmatrix} \mathbf{\Phi}^*_{(+)}(\lambda); \tag{87}$$

hence we have $\det \mathbf{\Phi}_{(-)} = -\big(\det \mathbf{\Phi}_{(+)}\big)^*$. This justifies the above statement, and therefore we will consider only the first equation in (78), $\det \mathbf{\Phi}_{(+)}(\lambda) = 0$.

Recall that our goal is to find those λ, or, equivalently, α in (81), which make (78) hold true. They are the eigenvalues of the original problem (70), (73), (76), which determines stability of the modes with $\rho \approx \pm i$ of the MoC-LF scheme. Since this eigenvalue problem has a finite dimension, the set of its eigenvalues is discrete.[9] These eigenvalues α are found as follows. First, one related α to β via (82b) and then to ρ_j via (75). Second, one forms the 4×4 matrix $\mathbf{\Phi}_{(+)}$ as explained after (77a), using the found ρ_j and the $\boldsymbol{\xi}^{\pm}_{(+)}$ from (85), where the superscript '$\pm$' notation was defined in (22). Third, one numerically solves the first equation in (78) to find α (see the next paragraph). Finally, one finds the eigenvalues λ via (81); they yield the amplification factor of the modes in question, as defined in (19).

A result of this calculation for a representative pair of values L and h is shown in Fig. 9. The indicated values of α from the range (83) are those points where the curve touches the x-axis; they are marked with circles. Marked with stars are the values corresponding (via (81)) to the eigenvalues of the periodic problem. They are found from Eq. (30) of [1] with the same L and h (see the footnote after Eq. (87) above). One can see that the eigenvalues with largest magnitude, which primarily determine the observed dynamics of unstable modes, are only 1% apart for the problems with periodic and nonreflecting BC. This explains the numerical observation, stated in the first paragraph of this section, that the instability growth rates of the MoC-LF with periodic and nonreflecting BC are very similar.

7 Summary and discussion

7.1 Summary and an example of an unusual numerical instability

The main contributions of this study are as follows. First, we have presented an approach to carry out a stability analysis of the MoC for a PDE system with constant coefficients while placing emphasis on its having non-periodic BC. The two ingredients of this approach have been known before. One ingredient was Ziółko's paper [6] presenting the problem in terms of a block-tridiagonal

[9] This is similar to why it is discrete in the periodic problem. Namely, there, the discrete modes with $\rho = \exp[ikh]$, $k \in \mathbb{Z}$, determine λ via setting the characteristic polynomial of (70) to zero.

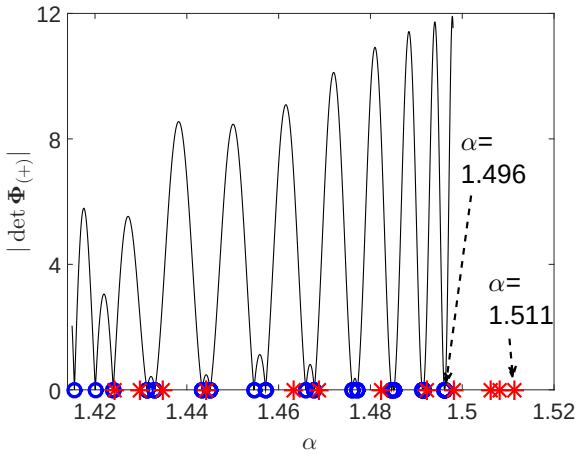

Figure 9: (Color online) A diagram which allows one to visualize, via the first equation in (78), the eigenvalues of MoC-LF with nonreflecting BC; they are marked as blue circles. Red stars mark the locations of the eigenvalues of the problem with periodic BC; they are found as explained in Section 6.3. The arrows point at the largest α's in the problems with nonreflecting and periodic BC. Parameters: $L = 50$, $h = 0.01$.

Toeplitz matrix, similar to our Eq. (17a). However, he did not use its block-tridiagonal Toeplitz structure to obtain its eigenvalues. For that reason, it would be far from straightforward to apply his analysis, done for the case of two variables (corresponding to $\tilde{\mathbf{u}}$ in our Eq. (1) being a 2-component vector) to a multi-variable case along each characteristic, which we have considered here. Moreover, Ziółko focused on the case of a dissipative hyperbolic system (i.e., matrix $\mathbf{P}$ in our (1) having all of its eigenvalues negative), for which the stability of an Euler-type ODE solver, that he considered, is substantially different than for a non-dissipative system considered here. The other ingredient of our analysis was the well-known method of finding eigenvalues of a band Toeplitz matrix; see, e.g., [11, 13, 14]. In Section 3 we outlined how these two ingredients can be combined to enable a stability analysis of the MoC applied to a constant-coefficient PDE with non-periodic BC. While we have considered a special kind of such BC — nonreflecting, — the approach can be straightforwardly generalized to any other kind of BC.

Second, we have applied our analysis to three "flavors" of the MoC, where the ODEs along the characteristics were solved by the SE, ME, and LF methods. For all three of them, a previous study [1] has found, for *periodic* BC, a conspicuous instability for modes in the "middle" of the Fourier spectrum, i.e., for $kh \approx \pi/2$. Here, we have shown that this instability is *suppressed* for the MoC-SE and MoC-ME schemes when nonreflecting BC are applied. As we announced in the Introduction, this finding contradicts a statement made in textbooks [3], [5] that an instability of a scheme with periodic BC always implies its instability for other types of BC. We will provide a conceptual explanation of this contradiction in Section 7.3. Thus, we have shown that for some "flavors" of the MoC schemes, a numerical instability found by the von Neumann *does not always imply* that the given scheme will be unstable for non-periodic BC, contrary to what is stated in [3],

[5].

Third, we have found that for $Lh \lesssim 1$, the modes stabilized by the nonreflecting BC have a decay rate

$$\gamma \propto (\ln h)/L; \tag{88}$$

see (64). In other words, the evolution[10] of the amplitudes of the numerical error's modes satisfies (again, for $Lh \lesssim 1$):

$$\text{amplitude}(t) \propto h^{t/L}; \tag{89a}$$

see (47). This holds for all modes with $|kh| \not\approx 0,\ \pi$ in the MoC-SE and for modes with $|kh| \not\approx 0,\ \pi,\ \pi/2$ in the MoC-ME. The modes in the MoC-ME with $|kh| \approx \pi/2$ decay in time even faster:

$$\text{amplitude}(t) \propto h^{2t/L}, \tag{89b}$$

as follows from comparison of (45) with (63). Thus, the modes' decay rate depends both on the discretization step h and on the domain length L in an unusual way. To our knowledge, such features have not been previously reported for other numerical schemes applied to PDEs with spatially-constant coefficients.

Moreover, our analysis has predicted yet another *unusual phenomenon*: Merely by increasing the product Lh, one can turn stable modes of the MoC-SE with nonreflecting BC into unstable ones; see (43) or (49). We illustrate this now via slightly altering the case shown in Fig. 2, which has stable modes in the "middle" of the spectrum. The only difference between that case and that shown in Fig. 10(a) is that in the latter, L is four times greater. According to (49), the magnitude of the modes in the "middle" of the spectrum is to increase by a factor of approximately 8 over time $t = L$, and this agrees well with the results seen in Fig. 10(a). For the MoC-ME we do not have an estimate of $|\lambda|^M$ that would have been accurate up to a factor $\exp[\,\text{const} \cdot Lh]$, but it is reasonable to expect that such a factor is indeed present in (63) (for modes with $\rho \approx \pm i$) and in (45) (for modes with $\rho \not\approx \pm i$). This expectation is borne out by the result shown in Fig. 10(b), which should be compared with the stable evolution of the MoC-ME error in Fig. 7. By trial-and-error we have determined that for $L = 300$ the scheme is still stable (as it is also in the case shown in Fig. 7), while for $L = 400$ the growth rate, while positive, is quite small. Therefore, we showed the case with $L = 600$ and hence a higher growth rate. Let us note that we are not aware of any previous reports where merely changing the length of the spatial domain would alter stability of some of the modes and, as in the MoC-ME example above, even of the entire scheme.[11]

7.2 Applicability of the above analysis to other models

Let us emphasize that the second and third contributions stated above are specific to the energy-preserving PDEs whose linearized form is (1). In fact, such hyperbolic PDEs are fairly common.

[10] after "straightening out" the staircase shape; see Step 3 in Sec. 4.4

[11] Let us mention in passing that in [15, 16], the growth rate of unstable modes for a PDE with *spatially-varying* coefficients was also found to depend on L, but it did so *not* monotonically and via a completely different mechanism.

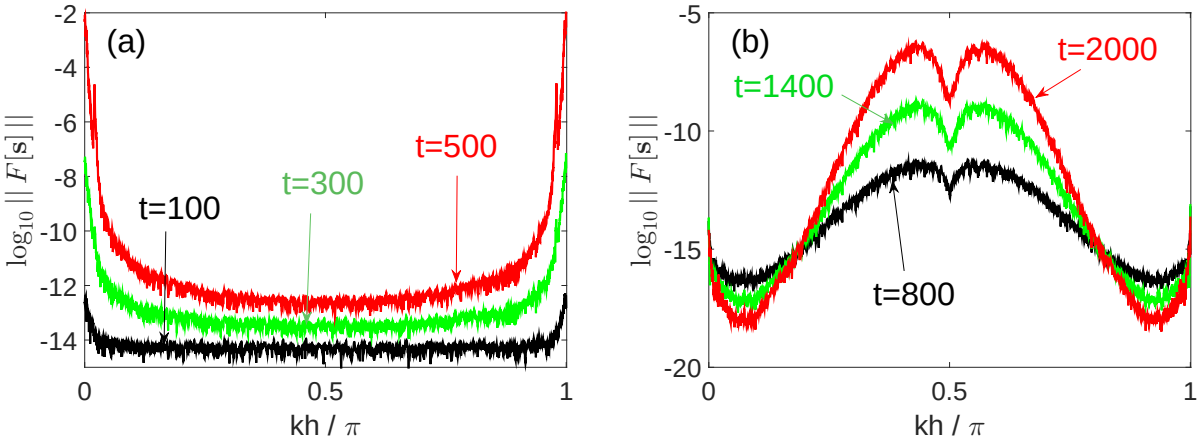

Figure 10: (Color online) Spectra of the numerical error for MoC-SE with $h = 0.02$, $L = 200$ (a) and for MoC-ME with $h = 0.02$, $L = 600$ (b). Time values are indicated in the plots. As explained in Section 4.4, these times are to be separated by an integer multiple of L in order to show monotonic evolution of the error. See end of Section 7.1 for a discussion.

For example, as stated in Section 2, they arise in the theory of birefringent optical fibers [10], with one particular example given by Eqs. (3). More generally, they arise in any model that involves two coupled waves (or two distinct, e.g., counter-propagating, groups of waves); see Section 2 of Ref. [1] for a more detailed discussion and references.

As one example of such a model arising outside of optics, we present, in Appendix C, the Gross–Neveu model from the relativistic field theory, whose linearized equations have the form (1). The soliton (i.e., localized pulse) solution of that model has received considerable attention in the past decade from both the analytical and numerical perspectives; see Refs. [17], [18], [37], [38], and [67]–[71] in [1] and earlier works cited there. Since the soliton solution is not constant in space, our analysis can only be applied to it in a non-rigorous sense of the "principle of frozen coefficients". Nonetheless, in the remainder of this section we will demonstrate that *all* conclusions which we obtained in Section 5 about the behavior of the numerical error for the constant-coefficient model also hold for the soliton of the Gross–Neveu model (99).

Below we will use this example of the Gross–Neveu model to illustrate the most practically important application of this work. As follows from the second contribution listed in Section 7.1, the MoC-ME scheme, which is unstable for periodic BC, gets stabilized by the imposition of nonreflecting BC. Thus, it becomes stable for simulation times up to $O(h^{-3})$ (see the second footnote in the Introduction), which suffices for most practical purposes.

To confirm this, in Fig. 11(a)[12] we show the Fourier spectrum of the numerical error obtained when a soliton of the Gross–Neveu model is simulated with the MoC-ME with periodic BC. The unstable modes' amplitudes are seen to have the same profile as for the constant solution (4) of Eqs. (3): see Fig. 1(b). (The peak near $kh = 0$ in Fig. 11(a) is not related to numerical instability; its origin is explained in [1].) If we now simulate the soliton of the Gross–Neveu model with the

[12] which is equivalent to Fig. 5(c) from [1]

same parameters as for Fig. 11(a) but with nonreflecting BC (101), the Fourier spectrum of the numerical error acquires the shape shown in Fig. 11(b). The error is seen to be several orders of magnitude higher than the initial error (which has the order 10^{-10} in the spectral domain). However, this occurs *not* because of its exponential growth, but because a non-periodicity at the boundaries, created by the nonreflecting BC (see Fig. 11(c)), led to a spectrum that decays with k very slowly. To confirm that the error does not grow in time (but, in fact, slightly decays), we presented the spectra for $t = 1000$ and $t = 5000$ in Fig. 11(b). Thus, as predicted by our theory for the simpler model based on Eqs. (3), (4), the MoC-ME scheme for the soliton of the Gross–Neveu model is also stabilized by nonreflecting BC.

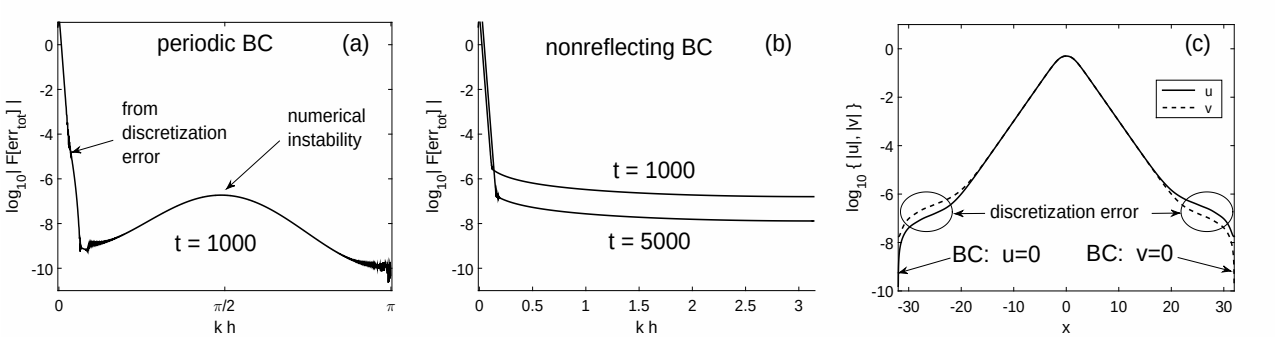

Figure 11: Fourier spectra of the numerical error of the MoC-ME for the soliton (100) with $\Omega = 0.7$ of the Gross–Neveu model (99) for periodic (a) and nonreflecting (b) BC. Parameters: $L = 64$, $h = L/2^{12} \approx 0.016$. The initial condition is the soliton plus a white noise in space of magnitude 10^{-12}. Note that the numerical error in (b) was *not* made spatially periodic before taking the Fourier transform (see main text). Panel (c): the soliton obtained for nonreflecting BC at $t = 5000$. The non-periodicity of the discretization error of magnitude $O(10^{-8})$ is seen, in agreement with the lower curve in panel (b).

In this and the next paragraphs we will comment on a few subtle features of the evolution of the MoC-ME error for the Gross–Neveu model, which agree with our analysis of the simpler model in Section 5. The reader who is not interested in such nuances may skip these two paragraphs. The reader may notice that the spectral shape of the error in Fig. 11(b) is missing the "dip" around $kh = \pi/2$, which is seen in Fig. 7 and which was predicted by our analysis of the constant-coefficient model in Section 5. This occurs due to the same reason, related to the non-periodicity of the numerical solution at the boundaries, as illustrated in Fig. 11(c). In order to recover the shape of the spectrum as seen in Fig. 7, we have made the numerical error periodic by multiplying it (but *not* the numerical solution) by a spatial super-Gaussian "window", as described in Step 1 of Section 4.4.[13] In Fig. 12 we show the evolution of the so modified, periodic error. Figure 12(a) shows the case for $L = 128$ and $h = L/2^{12} \approx 0.031$. Note that we had to increase both L and h compared to the case shown in Fig. 11 since otherwise, in agreement with estimates (89), the

[13] In fact, the error used in plotting Fig. 7 was made periodic in the same way, as explained at the beginning of Section 5.4.

error would reach the size of the computer round-off error very quickly. For the same reason, in the initial condition we added to the soliton a white noise of the size 10^{-6} as opposed to 10^{-12}. The profile with a "dip", seen in Fig. 7, is also observed in Fig. 12(a). Also in agreement with our analysis for the simpler model, the "dip" decreases approximately twice as fast as the "wings" of the spectrum: see Fig. 12(b).

Moreover, as we pointed out at the end of Section 7.1, it should be possible to make the MoC-ME unstable again by simply increasing the length L of the computational domain. This, indeed, also holds true for the soliton of the Gross–Neveu model. We chose the same $L = 600$ as in Fig. 10(b), kept the *same* h as in Fig. 12(a), and added to the initial soliton a white noise of magnitude 10^{-12} (since the error will now grow, not decay). The corresponding Fourier spectra at different times are shown in Fig. 12(c). The shapes of the spectra for our simpler model (3), (4) (Fig. 10(b)) and for the soliton of the Gross–Neveu model (Fig. 12(c)) are seen to be very similar.

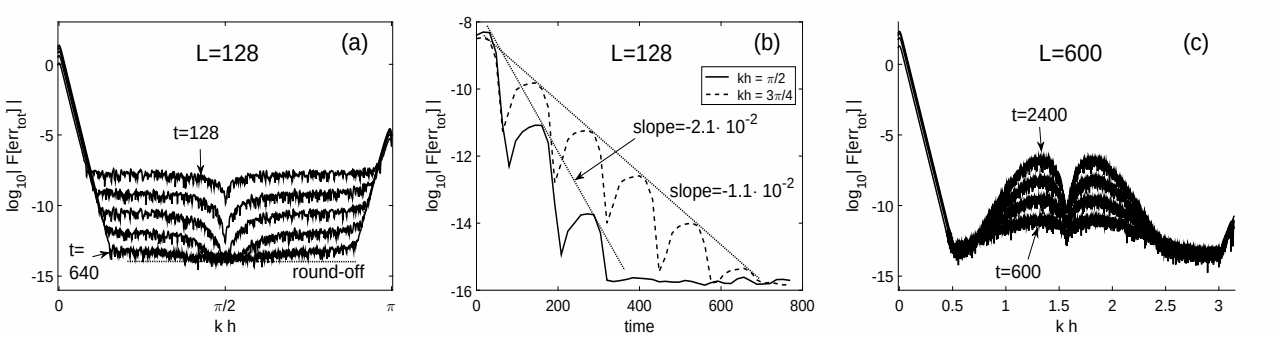

Figure 12: (a) Fourier spectra of the *periodic* numerical error of the MoC-ME with nonreflecting BC for the same soliton solution as in Fig. 11, but for $L = 128$; see text for details. The times increase from top to bottom with the increment $t_{\rm incr} = L = 128$, as discussed in Section 4.4. (b) Time evolution of amplitudes of the Fourier modes near the "dip" and at the "wing" of the spectrum from panel (a). The amplitudes are averaged over a narrow "box" to produce smooth curves. (c) Same as in (a), but for $L = 600$. The times increase from bottom to top with the increment $t_{\rm incr} = L = 600$.

The stable behavior of the MoC-ME scheme with nonreflecting BC has been used in a recent work [17] by one of the authors to resolve a controversy regarding the Gross–Neveu soliton with *small* Ω (i.e., $\Omega \lesssim 0.3$). As reviewed in [17], prior analytical studies had proved that Gross–Neveu solitons with *any* Ω are linearly stable, but long-time *simulations using several different numerical methods (excluding the MoC-ME)* had detected an instability of solitons with small Ω. The author of [17] used the MoC-ME scheme with nonreflecting BC and was able to simulate the soliton with $\Omega = 0.1$ for several thousand time units,[14] which was significantly longer than any previous studies had been able to achieve for this Ω. This resolution of the aforementioned controversy in favor of the analytical prediction was made possible by using the appropriate combination of the numerical scheme (MoC-ME) and BC (nonreflecting). Using a different combination, e.g., MoC-ME with

[14]Even longer stable propagation of solitons with even smaller Ω was possible with additional absorption at the boundaries.

periodic BC, or MoC-LF with nonreflecting BC (see Section 6), would not have allowed one to simulate the Gross–Neveu soliton stably.

Let us point out that these observed similarities of (in)stability properties of the MoC-ME scheme for problems with constant and spatially-varying coefficients, and *especially* the dependence of the stability of the MoC-ME scheme on L, provide an indirect evidence that conclusions of our analysis are *valid for an even broader class of models*. More specifically, a sufficient condition for the same (in)stability behavior of a numerical scheme appears to be the presence of the constant, linear coupling between the forward- and backward-propagating components of the solution, given by the last terms on the rhs of Eqs. (99). These terms are generic for coupled-mode equations and hence appear in a wide range of physical applications; see, e.g., Refs. [4], [5], [22], [24]–[26], [28], [29], [56]–[61], [63], [64] in [1]. Let us now explain the statement made two sentences above. From the analysis in Section 4.3 (and a similar analysis in Section 5.3) it follows that the dependence of λ on h occurs due to the linearization matrix $\mathbf{P}$ in (7) (and, more generally, in (1)) being nonzero. For Eqs. (99), entries of the corresponding $\mathbf{P}$ have two contributions: from the nonlinear terms and from the linear terms on the rhs. Now note that the nonlinear terms are essentially nonzero only in an interval of width $O(1)$ where the soliton is essentially nonzero. Thus, their contributions to entries of $\mathbf{P}$ are nonzero only in that $O(1)$-long interval, and hence they cannot affect λ by a factor that is related to the length $L \gg 1$ of the entire spatial domain. On the other hand, the linear terms on the rhs of (99) make spatially constant contributions to entries of $\mathbf{P}$ and thus must be solely responsible for the changes of the (in)stability behavior of the numerical scheme with L.

Thus, we have shown that our analysis for a constant-coefficient, non-dissipative hyperbolic system whose linearization is given by (1) has predictive value for another non-dissipative hyperbolic model with *non*-constant coefficients. An extension of our stability result to dissipative hyperbolic PDEs, or non-dissipative ones whose linearization in some way substantially differs from (1)[15] or to more general BC (e.g., those considered in [6]), remains an open problem. However, the approach will remain the same as that outlined in Section 3.

7.3 Why von Neumann analysis may incorrectly "predict" instability for non-periodic BC

We will now explain why the instability of the modes with $kh \not\approx 0, \pi$ in the MoC-SE and MoC-ME with periodic BC did *not* imply an instability of their counterparts when the BC are changed to nonreflecting. Let us recall that the opposite statement was made in textbooks [3]–[5], and thus, our contradicting that common knowledge may be the most unexpected conclusion of this study. We will begin by stating the reason behind this contradiction in general terms. Our system (1) (or (6a)) has the same form as that considered in [3]–[5]. However, one of the *underlying assumptions* made in [3]–[5] does *not* actually hold for the spatial modes of the MoC-SE and MoC-ME.

We will now explain this in detail. In [3], [5] it was assumed that all modes in a problem with

[15] E.g., there are more than two groups of waves propagating with distinctly different group velocities.

non-periodic BC fell into two groups.[16] Modes in one group are localized near the boundary (or either of the boundaries). These modes become exponentially small in the "bulk" of the domain. In particular, the mode localized near the left boundary does not "feel" the right boundary, and vice versa; see Fig. 13(a). Modes in the other group do not exhibit a monotonic decay away from the boundary(ies) of the spatial domain. In sufficiently large domains, such modes resemble Fourier harmonics (i.e., modes of the periodic problem) in the "bulk" of the domain; see Fig. 13(a). (Near the boundary(ies), their profile would, of course, be modified to satisfy the BC.) Given this resemblance, one can reasonably expect the (in)stability of modes in this group to be similar to that of the corresponding Fourier harmonics. Therefore, if any of the modes in the periodic problem is unstable, so should be the corresponding mode from this "second group" of the non-periodic problem. It does not matter in this case whether modes from the "first group" are stable or not, for the numerical scheme is already unstable due to modes from the "second group". This is the reason why it was stated in [3]–[5] that an instability of a scheme predicted by the von Neumann analysis implies an instability of that scheme with *arbitrary* BC.

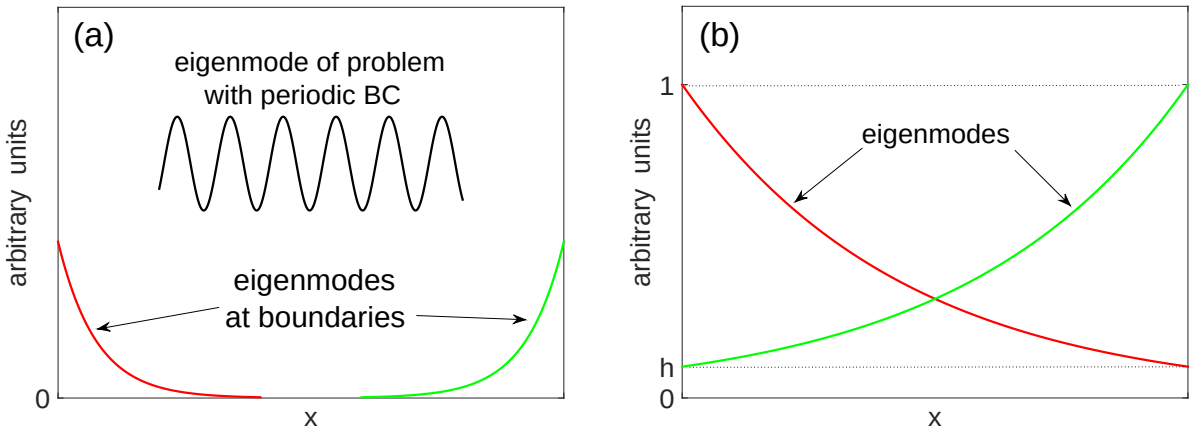

Figure 13: Schematics for the explanation of when the von Neumann analysis predicts instability of a numerical scheme (a) or does not predict it (b). See Section 7 for details. The dotted horizontal lines in (b) are guides for the eye. Clarification about panel (b): for a given λ, there are two spatially growing and two decaying modes, according to the number of roots ρ in (28). We stress that, according to (23a), all four of those modes are required to satisfy the nonreflecting BC at each boundary.

The reason why the above statement does not hold for the MoC-SE and MoC-ME is schematically shown in Fig. 13(b). Namely, the modes in the problem with nonreflecting BC do not split into the two groups described in the previous paragraph. Thus, the underlying assumption that led to the conclusion in textbooks [3]–[5] does not hold for the MoC-SE and MoC-ME.

[16] This follows from the so-called Babenko–Gelfand procedure ([5], p. 379); the more widely known Godunov–Ryabenkii stability criterion uses the same assumption ([3]; [5], pp. 397, 402). Textbook [4] does not explicitly mention these two groups of modes, but states, on p. 276, that in the stability analysis of a scheme, one can replace the finite computational interval (i.e., that with *two* boundaries) by two separate semi-infinite intervals (each having one boundary either on the left or on the right). This step is essentially equivalent to the Babenko–Gelfand procedure.

Indeed, the spatial dependence of the eigenmodes in these methods is given by (18), which via (28) yields:

$$\|(\mathbf{s})_m\| \approx \lambda^{-m}\|(\mathbf{s})_0\| \qquad \text{or} \qquad \|(\mathbf{s})_m\| \approx \lambda^{m}\|(\mathbf{s})_0\| \,. \tag{90}$$

(Here we have ignored a contribution of the order $\exp[\text{const} \cdot Lh]$, which could come from factors $\widehat{\rho}_{1,2(\pm)}$, as such a contribution does not change the general conclusion of this discussion.) Then, according to (45):

$$\|(\mathbf{s})_0\| \,/\, \|(\mathbf{s})_M\| = O(h) \qquad \text{or} \qquad \|(\mathbf{s})_M\| \,/\, \|(\mathbf{s})_0\| = O(h) \,, \tag{91}$$

as depicted in Fig. 13(b). In particular, the mode decaying away from the left boundary is *not* negligibly small at the right boundary *no matter how large the spatial domain is*; and vice versa. (Obviously, such modes could not have been found if the analysis had assumed (as stated in [4], p. 276) a semi-infinite spatial domain instead of a finite one.) Even more importantly, there are *no* modes in the problem with nonreflecting BC that resemble Fourier harmonics in the "bulk" of the domain. Indeed, all modes either decay or grow monotonically in magnitude. This exponential decay or growth[17] appears to make these modes sufficiently different from Fourier harmonics, so that a prediction of the von Neumann analysis about instability of Fourier harmonics has no bearing for the modes of the problem with nonreflecting BC.

To conclude this discussion, let us note that our results do *not* contradict those of the celebrated paper [18] on the stability of numerical schemes for boundary value problems. That paper does not require the von Neumann stability as a necessary condition for stability of a scheme with arbitrary BC; rather, it *assumes* that the scheme is von Neumann-stable. This is a subtle, but important difference: If a scheme is not von Neumann-stable (as is the case of the MoC-ME), then conclusions of [18] simply do not apply rather than lead one to infer that the scheme should be unstable for other BC. Incidentally, Ref. [14] provides an insight into why a von Neumann-unstable scheme can be stable with, say, Dirichlet BC. The reason is that the spectrum of the scheme in the latter case lies strictly inside the spectrum of the periodic problem, at least in the limit $M \equiv L/h \to \infty$. Hence, the former spectrum can be stable even though the latter is unstable; see the Figures and Ref. [9] in [14].

7.4 When von Neumann analysis can correctly predict instability for a MoC scheme with non-periodic BC

Finally, let us give a qualitative reason why the above explanation did *not* apply to the MoC-LF scheme, for which, as we showed in Section 6, the nonreflecting BC did not suppress the instability that existed for periodic BC. The reason is precisely that (unstable) modes of the "second group", which resemble Fourier harmonics in the "bulk" of the domain, *do* exist in the MoC-LF scheme. This, in turn, occurs due to the following difference between the LF and Euler ODE solvers. Unlike the latter, the LF solver involves more than two time levels. This leads, in the limit $h \to 0$, to the

[17] with the same exponent (88) as that governing the time evolution

existence of "parasitic" amplification factors, $\rho_{\text{parasitic}} = -\lambda,\ -\lambda^{-1}$, in addition to the true ones, $\rho_{\text{true}} = \lambda,\ \lambda^{-1}$. For $\rho = \pm i$, which is where the modes of the periodic problem are found to be unstable, each of the "parasitic" factors coincides with a true one: $-\lambda = \lambda^{-1}$, $-\lambda^{-1} = \lambda$. In this situation and for $0 < h \ll 1$, it is possible to have

$$|\lambda| = 1 + O(h) > 1 \qquad \text{but} \qquad |\rho| = 1; \tag{92}$$

see (81) and the paragraph before (83), where we specified that $\alpha,\ \beta \in \mathbb{R}$. Condition (92), which is possible for the MoC-LF but impossible for the MoC-SE and MoC-ME (see (28)), means that for the MoC-LF, there *do* exist modes which resemble Fourier harmonics in the "bulk" of the spatial domain ($|\rho| = 1$) and are strongly unstable ($|\lambda| = 1 + O(h) > 1$). This is why the von Neumann analysis for the MoC-LF predicts nearly the same instability that exists for that scheme with non-periodic BC. Thus, a necessary ingredient for this to occur for an arbitrary MoC scheme is that its ODE solver must have some $\rho_{\text{parasitic}}$ that coincides with a ρ_{true} for $h \to 0$.

Acknowledgment

This work was supported in part by the NSF grant DMS-1217006.

Appendix A: Why ρ in (18) is to be a scalar

Let us argue by contradiction: suppose that $\boldsymbol{\rho}$ is a 4×4 matrix and substitute (18) into (15a). The result,

$$\left(\boldsymbol{\rho}^{-1}\boldsymbol{\Gamma} - \lambda \mathbf{I} + \boldsymbol{\rho}\boldsymbol{\Omega} \right) \boldsymbol{\rho}^m \, (\mathbf{s})_0 = \mathbf{0}, \tag{93}$$

implies two corollaries. First, the eigenvalue problem (93) for any given $\boldsymbol{\rho}$ can have no more than 4 eigenvectors. Then, considering the factor $\boldsymbol{\rho}^m$, one concludes that

$$\boldsymbol{\rho}^4 = c\,\mathbf{I}, \tag{94}$$

where c is a scalar. Second, if some $(\mathbf{s})_0$ is an eigenvector of (93), then so are $\boldsymbol{\rho}^l \, (\mathbf{s})_0$, $l = 1, 2, 3$. Thus, we have arrived at a situation where four eigenvectors of a rather general matrix (the one in parentheses in (93)) are to be related to one another by a fourth root of the 4×4 identity matrix. In addition, these eigenvectors are to satisfy a certain condition, which follows from the BC (see (23), (24)). While this may be possible for some special values of h and the coefficients in the PDE system in question, in general this situation does not hold. Thus, $\boldsymbol{\rho}$ cannot be a matrix and hence must be a scalar.

Appendix B: Derivation of (40)

Expressions $\widehat{\rho}_{1,2(\pm)}$ from (28) can be written in the form

$$\widehat{\rho} = 1 + i\epsilon h + (c_R + i c_I)h^2, \tag{95}$$

where $\epsilon = +1$ or -1 and the coefficients $c_{R,\,I} \in \mathbb{R}$ depend on λ. Equivalently,

$$\begin{aligned} \widehat{\rho} &= (1 + c_R h^2)\sqrt{1 + \frac{h^2(\epsilon + c_I h^2)^2}{(1 + c_R h^2)^2}} \cdot \exp\left[\, i\left(\epsilon h + c_I h^2 + O(h^3)\right)\right] \\ &= \left(1 + \left(c_R + \frac{1}{2}\right) h^2 + O(h^3)\right) \cdot \exp\left[\, i\left(\epsilon h + c_I h^2 + O(h^3)\right)\right] . \end{aligned} \tag{96}$$

Consequently, for $M = L/h$,

$$\widehat{\rho}^{\,M} = \exp\left[\left(c_R + \frac{1}{2}\right) Lh + O(Lh^2)\right] \cdot \exp\left[\, i\left(\epsilon L + c_I Lh + O(Lh^2)\right)\right] . \tag{97}$$

Relations (40) are derived for the modes satisfying (50), because it is only for those modes that we chose to verify our analytical results in Section 4.4. For these modes, one has, with the subscript notations of (28):

$$\epsilon_{1(\pm)} = \mp 1, \;\; (c_R)_{1(\pm)} \approx 1, \;\; (c_I)_{1(\pm)} \approx 0; \qquad \epsilon_{2(\pm)} = \pm 1, \;\; (c_R)_{2(\pm)} \approx -2, \;\; (c_I)_{2(\pm)} \approx 0. \tag{98}$$

Substituting (97) with (98) into the fractions in (40) and neglecting terms $O(Lh^2)$, we obtain their values as stated there.

Appendix C: Soliton solution of the Gross–Neveu model

The Gross–Neveu model in the notations convenient for this work has the form (see [1] and references therein):

$$\begin{aligned} u_t + u_x &= i(\,|v|^2 u + v^2 u^*\,) - iv, \\ v_t - v_x &= i(\,|u|^2 v + u^2 v^*\,) - iu. \end{aligned} \tag{99}$$

Its standing soliton solution has the form:

$$\{u, v\} \,=\, \{U(x), V(x)\}\, \exp[-i\Omega t], \qquad \Omega \in (0, 1); \tag{100a}$$

$$\{U(x), V(x)\} \,=\, \sqrt{1 - \Omega}\, \frac{\cosh(\beta x) \pm i\mu\, \sinh(\beta x)}{\cosh^2(\beta x) - \mu^2\, \sinh^2(\beta x)}; \tag{100b}$$

with $\beta = \sqrt{1 - \Omega^2}$ and $\mu = \sqrt{(1 - \Omega)/(1 + \Omega)}$. This solution for $\Omega = 0.7$ is shown in Fig. 5(a) of [1]. For reasons explained there, in this work we will focus on the soliton with this value of Ω, but will also mention results pertaining to solitons with smaller Ω. The initial condition is taken as that soliton plus a very small (of order 10^{-6} or 10^{-12}; see Section 7.2) white noise, which is added to make the numerical error grow (or decay) from a level other than the computer's round-off.

The nonreflecting BC for this model are taken as:

$$u(x = -L/2,\; t) = 0; \qquad v(x = L/2,\; t) = 0, \tag{101}$$

where the computational domain is $x \in [-L/2,\; L/2]$. While it may be formally more correct to use the respective values $U(-L/2)$ and $V(L/2)$ instead of zeros in (101), in practice that makes no difference, because those values are at the level of the computer's round-off error.

The numerical error, referred to in Section 7.2, was defined as:

$$\mathrm{err}_{\mathrm{tot}} = \left(\sum_{m=0}^{M} \left| u_m^n - U(x_m) e^{-i\Omega t_n} \right|^2 + \left| v_m^n - V(x_m) e^{-i\Omega t_n} \right|^2 \right)^{1/2} . \tag{102}$$

This error does not satisfy periodic BC because u and v do not satisfy them.

References

[1] T.I. Lakoba, Z. Deng, Stability analysis of the numerical Method of characteristics applied to energy-preserving systems. Part I: Periodic boundary conditions, J. Comput. Appl. Math. https://doi.org/10.1016/j.cam.2019.01.027.

[2] D.F. Griffiths, D.J. Higham, Numerical methods for ordinary differential equations, Springer-Verlag, London, 2010; Chaps. 13 and 15.

[3] R.D. Richtmyer, K.W. Morton, Difference methods for initial-value problems, Interscience, New York, 1967; Sec. 6.7, pp. 163–164.

[4] J.C. Strikwerda, Finite difference schemes and partial differential equations, 2nd Ed., SIAM, Philadelphia, 2004; Sec. 11.3, p. 290 (proof of Theorem 11.3.1).

[5] V.S. Ryaben'kii, S.V. Tsynkov, A theoretical introduction to numerical analysis, Chapman & Hall/CRC, Boca Raton, 2007; Sec. 10.5.1, pp. 379, 382 and Sec. 10.5.2, pp. 397, 402.

[6] M. Ziółko, Stability of method of characteristics, Appl. Num. Math. 31 (1999) 463–486.

[7] V.E. Zakharov, A.V. Mikhailov, Polarization domains in nonlinear media, JETP Lett. 45 (1987) 349–352.

[8] S. Pitois, G. Millot, S. Wabnitz, Nonlinear polarization dynamics of counterpropagating waves in an isotropic optical fiber: theory and experiments, J. Opt. Soc. B 18 (2001) 432–443.

[9] S. Wabnitz, Chiral polarization solitons in elliptically birefringent spun optical fibers, Opt. Lett. 34 (2009) 908–910.

[10] V.V. Kozlov, J. Nuño, S. Wabnitz, Theory of lossless polarization attraction in telecommunication fibers, J. Opt. Soc. B 28 (2011) 100–108.

[11] G.D. Smith, Numerical solution of partial differential equations: Finite difference methods, Clarendon Press, Oxford, 1978; Chap. 3.

[12] R. Haberman, Applied partial differential equations with Fourier series and boundary value problems, 4th Ed., Pearson/Prentice Hall, Upper Saddle River, 2003.

[13] W.F. Trench, On the eigenvalue problem for Toeplitz band matrices, Lin. Alg. Appl. 64 (1985) 199–214.

[14] R.M. Beam, R.F. Warming, The asymptotic spectra of banded Toeplitz and quasi-Toeplitz matrices, SIAM J. Sci. Comput. 14 (1993) 971–1006.

[15] T.I. Lakoba, Instability analysis of the split-step Fourier method on the background of a soliton of the nonlinear Schrödinger equation, Num. Meth. Part. Diff. Eqs. 28 (2012) 641–669.

[16] T.I. Lakoba, Instability of the finite-difference split-step method applied to the nonlinear Schrödinger equation. II. moving soliton, Num. Meth. Part. Diff. Eqs. 32 (2016) 1024–1040.

[17] T.I. Lakoba, Numerical study of solitary wave stability in cubic nonlinear Dirac equations in 1D, Phys. Lett. A 382 (2018) 300–308.

[18] B. Gustafsson, H.-O. Kreiss, A. Sundström, Stability theory of difference approximations for mixed initial boundary value problems. II, Math. Comp. 26 (1972) 649–686.