

# A Unified Approach for Deriving Optimal Finite Differences

Komal Kumari<sup>a</sup>, Raktim Bhattacharya<sup>a,\*</sup>, Diego A. Donzis<sup>a,\*</sup>

<sup>a</sup>*Department of Aerospace Engineering, Texas A&M University, College Station, TX 77843, United States*

## Abstract

A unified approach to derive optimal finite differences is presented which combines three critical elements for numerical performance especially for multi-scale physical problems, namely, order of accuracy, spectral resolution and stability. The resulting mathematical framework reduces to a minimization problem subjected to equality and inequality constraints. We show that the framework can provide analytical results for optimal schemes and their numerical performance including, for example, the type of errors that appear for spectrally optimal schemes. By coupling the problem in this unified framework, one can effectively decouple the requirements for order of accuracy and spectral resolution, for example. Alternatively, we show how the framework exposes the tradeoffs between e.g. accuracy and stability and how this can be used to construct explicit schemes that remain stable with very large time steps. We also show how spectrally optimal schemes only bias odd-order derivatives to remain stable, at the expense of accuracy, while leaving even-order derivatives with symmetric coefficients. Schemes constructed within this framework are tested for diverse model problems with an emphasis on reproducing the physics accurately.

*Keywords:*

## 1. Introduction

Ordinary differential equations (ODEs) as well as partial differential equations (PDEs) are pervasive in science and engineering as they model accurately a large number of natural and man-made systems. Unfortunately, many of these equations are exceedingly complex at realistic conditions and analytical solutions are virtually impossible. Thus, significant advances in understanding these phenomena have relied on the use of computer simulations for which an appropriate numerical method needs to be used to assure certain degree of accuracy in the solution.

Perhaps the most widely used method to discretize these governing equations in order to solve them on a computer is the so-called finite differences. The derivation of explicit finite difference schemes is in general very well known and has been studied extensively [1]. The general idea is to linearly combine the values of the function to be differentiated at neighboring points. The weights in this linear combination are determined so as to minimize error in some sense. The specific choice of this objective function, we show here, has a critical effect on the resulting schemes.

More formally, the standard procedure starts with an approximation of the derivative of a function  $f$  at a point  $x_i$  of the form

$$f'_i = \left. \frac{\partial f}{\partial x} \right|_{x_i} \approx \frac{1}{\Delta x} \sum_{m=-M}^M a_m f_{i+m} + \mathcal{O}(\Delta x^{p+1}). \quad (1)$$

The last term indicates that the truncation error of the approximation is of order  $p + 1$ . In a traditional derivation one first selects the stencil size, that is the number of neighboring points to use in the approximation which is  $(2M + 1)$  in Eq. (1), and then finds the coefficients  $a_m$  such that  $p + 1$ , the order of the

\*Corresponding author.

Email addresses: [raktim@tamu.edu](mailto:raktim@tamu.edu) (Raktim Bhattacharya), [donzis@tamu.edu](mailto:donzis@tamu.edu) (Diego A. Donzis)

truncation error, is largest. This is done by eliminating all terms of order lower than  $p + 1$  in a Taylor expansion of the right-hand-side of Eq. (1).

However, in many problems, especially those which involve multi-scale phenomena, one is concerned not only with a global measure of the error but also with the error incurred at different scales or, in Fourier space, at different wavenumbers. The range of wavenumbers and, thus, the resolution requirements depend on non-dimensional parameters such as Reynolds number for turbulent flows, or the Prandtl and Schmidt numbers for mixing problems, as they define length and time scales of interest in the problem. It is therefore of consequence to use schemes that are optimal for the time and length scale posed by the physics of the problems. This concern is typically addressed by analyzing the spectral characteristics of scheme derived above by using, for example, von Neumann analysis [1] or a modified wavenumber approach [2]. The result of this analysis is used to determine if the scheme derived from an order-of-accuracy consideration is indeed appropriate to resolve all relevant scales in the problem at hand. Note that this is in general *a posteriori* evaluation of the scheme.

There has been efforts in the literature to devise schemes with general properties in terms of spectral accuracy. Early work on acoustic computations [3] showed that some desired spectral behavior can be obtained by solving an optimization problem where the objective function to be minimized is some measure of these errors. A number of different applications based on the same general approach have been presented in the literature [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. In all these studies specific requirements were put forth typically based on the physics of interest which often resulted in subjective criteria to account for those specific requirements. Some were limited in scope, for example, by limiting the results to approximating only the first derivative [3, 4, 8, 6, 10]. These optimizations were also limited in the sense that the nature of the unknown coefficients (e.g. whether they are symmetric or antisymmetric) is specified *a priori* which resulted in the objective function comprising either only the real part or only the imaginary part of the spectral error, depending upon the order of the derivative being computed. However, as we show below, that symmetry and antisymmetry of the coefficients for even and odd derivatives respectively, can be obtained as a *consequence* of the minimization problem without any externally imposed conditions. Some others [12] formulated an optimization based on the maximum norm to minimize spectral errors but the algorithm adheres to certain ad-hoc rules (such as the nature of the coefficients or the relative magnitude of the coefficients in a stencil) and presents challenges in finding the global optimum. Other specialized optimizations have been conducted to obtain schemes with lower errors that, for example, can resolve very strong gradients in fluid flow calculations [9, 5, 7], though some degree of subjectivity and trial-and-error was used in the formulations. Both [5, 7] utilize a two level optimization to achieve better resolving efficiency and [7] also incorporated a weight function to emphasize on the relevant scales. Another approach is presented in [11] where the unknown coefficient is selected according to the level of dissipation required without carrying out any formal optimization. A general conclusion one can arrive at from all these studies, is that more sensible choices than commonly made can lead to numerical schemes that can outperform standard finite differences [2, 3]. Here we support this idea and show that the precise meaning of this metric has a clear impact on the scheme obtained. Another observation is that formulations are typically tailored with specific applications and constraints in mind due to the different requirements dictated by the physics of interest. It is thus not surprising that there seems to be no general rigorous mathematical framework under which these particular cases can be derived. Here, in a first step, we provide such a framework along with some rigorous results on the nature of the error that result from optimization formulations. As we will show, the framework can incorporate the different requirements needed for different cases and conditions.

A third critical aspect when considering numerical schemes is their stability. Obviously, to be usable, a scheme must be numerically stable when utilized to solve an ODE or PDE. Again, this has traditionally been *a posteriori* undertaking: after selecting a scheme of a given order, with a desired spectral accuracy, one would check if the scheme is stable or not and under which conditions. This has been the case for standard or optimized schemes [3, 14]. This is also the case for the two-step optimization of [15] in which an optimal spatial scheme is obtained first, followed by an optimization to get a stable time marching scheme. This also highlights the importance of considering the relation between the space and time discretization to increase the computational efficiency of the spatial operators. In [13], the spatial and the temporal

schemes are decoupled and optimized separately in order to achieve maximum resolving efficiency for both the operators. The optimization of the temporal scheme is also subject to the stability constraint. Schemes that remain stable for a broader range in the appropriate parameter space (time step size, grid spacing, non-dimensional groups such as CFL number, etc.) are typically preferred as simulations with larger time steps are computationally less expensive. While the time step and grid spacing are conventionally subject to the stability and resolution requirements, [13] shows that optimal values for these that minimize computational cost for some error level can also be obtained.

An overriding question is, thus, whether it is possible to find optimal schemes of given order, that are stable and that minimize the spectral error in a suitably defined manner typically informed by the physical characteristics of the problem being solved. This is the main motivation of the present work. The mathematical framework in which this can be achieved reduces to an optimization problem with equality and inequality constraints which can be solved, under certain conditions, analytically. The importance of this work is that it allows us to express physically meaningful desired properties and constraints into a unified mathematical framework which results in highly-accurate schemes for a particular problem of interest. Another important aspect of the proposed framework is that it also exposes explicitly tradeoffs that can be profitably used in specific circumstances. For example, we show that it is possible to construct explicit schemes that can remain stable for very large time steps (even an order of magnitude larger than equivalent standard schemes) when constraints on accuracy at some scales can be relaxed. We can also show that unlike traditional finite differences, one can design “spectrally flat” schemes which present a more homogeneous accuracy distribution across wavenumbers. Furthermore, because of optimality in spectral properties, we also show that the resulting schemes present better performance in terms of important physical properties like its dispersion relation, and group and phase velocity [16, 10, 17].

The rest of the paper is organized as follows. In section 2 we present the fundamental ideas behind the framework which include order of accuracy, spectral accuracy and stability. In section 3 we present numerical results using optimal schemes constructed using the tools presented here. We present results on accuracy, stability and tradeoffs that can be explicitly incorporated into the design of finite differences. Final remarks and conclusions are discussed in section 4. Proof of two important theorems are included as appendices.

## 2. The framework for deriving finite differences

As described above, in this work we construct a framework to derive finite differences in which three important design characteristics, namely order of accuracy, spectral accuracy, and stability, are combined into a rigorous mathematical framework. We now present each in turn.

### 2.1. Order of accuracy

A generalization of the approximation in Eq. (1) to the  $d$ -th spatial derivative is given by

$$f_i^{(d)} = \frac{1}{(\Delta x)^d} \sum_{m=-M}^M a_m f_{i+m}, \quad (2)$$

where, as before, we have  $M$  points on either side of the  $i$ -th grid point where the derivative is sought. The stencil size is then  $S \equiv 2M + 1$ . A Taylor series for a term on the right-hand-side of Eq. (2) can be written as

$$f_{i+m} = f_i + (m\Delta x)f'_i + (m\Delta x)^2 f''_i / 2! + \dots$$

Upon constructing the entire sum in Eq. (2), a  $(p+1)$ -th order approximation of the  $d$ -th derivative requires that the term with the  $d$ -th derivative be equal to  $d!$  and that the rest of the terms up to order  $p$  be zero. After some algebra these constraints can be written as

$$\sum_{m=-M}^M m^q a_m = \begin{cases} 0 & q \neq d, \\ d! & q = d, \end{cases} \quad (3)$$

for  $q \leq d + p$ , or more compactly as

$$\mathbf{a}_d^T \mathbf{X}_d = \mathbf{y}_d, \quad (4)$$

where

$$\mathbf{a}_d^T := [a_{-M} \ a_{-M+1} \ \cdots \ a_{M-1} \ a_M], \quad (5)$$

$$\mathbf{m}^T := [-M \ -M+1 \ \cdots \ M-1 \ M], \quad (6)$$

$$\mathbf{X}_d := [\mathbf{1}_{S \times 1} \ \mathbf{m} \ \cdots \ \mathbf{m}^{d-1} \ \mathbf{m}^d \ \cdots \ \mathbf{m}^{d+p}], \quad (7)$$

$$\mathbf{y}_d := [\mathbf{0}_{1 \times d} \ d! \ \mathbf{0}_{1 \times p}]. \quad (8)$$

with  $\mathbf{1}_{S \times 1}$  is an  $S \times 1$  vector with ones as its elements and  $\mathbf{0}_{1 \times d}$  a  $1 \times d$  vector with zeros as its elements. The vectors  $\mathbf{m}^n$  in Eq. (7) are defined as vectors composed of each element of  $\mathbf{m}$  raised to the power  $n$ .

Equation Eq. (4) is the linear system that, for a given stencil size  $M$ , results in a finite difference scheme of order  $p + 1$ . This approach, though presented in different forms across the literature, forms the basis for standard derivation of finite differences when  $S = d + p$ . In this case, the number of unknown coefficients in  $\mathbf{a}_d$  equals the number of terms that need to be eliminated to maintain a certain order of accuracy. If on the other hand,  $S > d + p$  then the solution to Eq. (4) is not unique. The remaining degrees of freedom can then be used to, e.g., assure spectral accuracy. This is presented next.

## 2.2. Spectral accuracy

In order to understand the behavior of discrete differentiation operators at different scales or frequencies, it is common to evaluate them utilizing a Fourier representation. For simplicity consider a single mode in a spatial discrete Fourier series:

$$f(x) = \hat{f} e^{jkx} \quad (9)$$

where  $\hat{f}$  is the Fourier coefficient of the mode at wavenumber  $k$ , and  $j = \sqrt{-1}$ . Clearly, the exact  $d$ -th derivative is given by

$$f^{(d)}(x) = (jk)^d f(x). \quad (10)$$

Now consider a discrete approximation of the derivative of the form Eq. (2). Since,

$$f_i := f(x_i) = \hat{f} e^{jkx_i}, \quad (11)$$

$$f_{i+m} := f(x_i + m\Delta x) = \hat{f} e^{jkx_i} e^{jkm\Delta x} = f_i e^{jkm\Delta x} \quad (12)$$

equation Eq. (2) becomes

$$f_i^{(d)} = \left( \frac{1}{(\Delta x)^d} \sum_m a_m e^{jkm\Delta x} \right) f_i \quad (13)$$

in terms of this single Fourier mode.

Comparison between the numerical approximation Eq. (13) and exact differentiation Eq. (10) suggests a natural way to define the error at a given waveumber as

$$e(\eta) = \sum_m a_m e^{jm\eta} - (j\eta)^d, \quad (14)$$

where  $\eta := k\Delta x$  is a convenient normalized wavenumber in the interval  $[0, \pi]$ . More compactly, this can be expressed in matrix form as

$$e(\eta) = (\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)) \mathbf{a}_d - (j\eta)^d, \quad (15)$$

where

$$\mathbf{C}(\eta) := \begin{bmatrix} \cos(-M\eta) \\ \vdots \\ \cos(-\eta) \\ 1 \\ \cos\eta \\ \vdots \\ \cos(M\eta) \end{bmatrix}, \text{ and } \mathbf{S}(\eta) := \begin{bmatrix} \sin(-M\eta) \\ \vdots \\ \sin(-\eta) \\ 0 \\ \sin(\eta) \\ \vdots \\ \sin(M\eta) \end{bmatrix}. \quad (16)$$

Equation Eq. (14) can also be written in terms of the so-called modified wavenumber  $(j\tilde{\eta})^d = \sum_m a_m e^{jm\eta} = (\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)) \mathbf{a}_d$  as

$$e(\eta) = (j\tilde{\eta})^d - (j\eta)^d = j^d(\tilde{\eta}^d - \eta^d). \quad (17)$$

Clearly, the difference between  $\tilde{\eta}^d$  and  $\eta^d$  provides a measure of the spectral error at wavenumber  $\eta$ . The ratio of the modified wavenumber to the actual wavenumber is then

$$G_d := \frac{\tilde{\eta}^d}{\eta^d} = \frac{e(\eta)}{(j\eta)^d} + 1, \quad (18)$$

which is a complementary measure of error across wavenumbers and will be used later on when comparing different schemes.

A global figure of merit to assess how accurately the scheme captures spectral content can be defined as the weighted  $\mathcal{L}_2$  norm of the error  $e(\eta)$ :

$$\|e(\eta)\|_{\mathcal{L}_2}^2 := \int_0^\pi \gamma(\eta) e^*(\eta) e(\eta) d\eta =: \langle e^*(\eta) e(\eta) \rangle. \quad (19)$$

Here  $e^*(\eta)$  is the complex conjugate of  $e(\eta)$ , and  $\gamma(\eta)$  is a weighting function introduced to provide control over which wavenumbers are to be more accurately resolved. The selection of  $\gamma(\eta)$  would depend, in general, on the physical characteristics of the system of interest. For example, for PDEs with multi-scale broadband solutions a natural choice would be a constant  $\gamma(\eta)$  over the range of  $\eta$  of interest and zero elsewhere. For a system with two well defined wavenumber bands, on the other hand, one can define  $\gamma(\eta)$  presenting relatively large values around those bands but negligible values everywhere else. Examples on the impact of this choice will be provided in section 3.

We can now frame the problem of the derivation of finite difference schemes combining spectral resolution and order of accuracy. Formally, our goal is to determine  $\mathbf{a}_d$  such that  $\|e(\eta)\|_{\mathcal{L}_2}^2$  is minimized, subject to a given order of accuracy defined by Eq. (4), i.e.

$$\min_{\mathbf{a}_d \in \mathbb{R}^{2M+1}} \|e(\eta)\|_{\mathcal{L}_2}^2, \text{ subject to Eq. (4)}. \quad (20)$$

Equation Eq. (20) provides, then, the unifying formalism to find the coefficients in Eq. (2) that both provides a given order of accuracy and minimizes error in spectral space [3]. This formulation, written in different ways, has been used extensively as pointed out in the introduction, but here is presented in a very general form. Note that if the number of unknowns ( $S = 2M + 1$ ) is equal to the number of terms to be removed from the truncation error to achieve a given order, then Eq. (4) has a unique solution and no optimization is possible. This is the case of standard finite difference schemes. If, on the other hand, the stencil size makes the number of unknowns greater than those needed to achieve a given order, the system will utilize those degrees of freedom to minimize  $\|e(\eta)\|_{\mathcal{L}_2}^2$ .

We note here that in standard derivations of finite differences based only on order-of-accuracy considerations, the spectral behavior is found a posteriori. Spectral resolution is then coupled to (dependent on) order of accuracy. While typically increasing the formal order of accuracy of the scheme, leads to a better spectral resolution, standard techniques provide no mechanism to constrain the spectral behavior of the

resulting schemes. In the approach presented above, on the other hand, specifications on spectral accuracy are independent of order of accuracy. Thus, we see that by *coupling the two mathematical systems* into a unified formulation, we effectively *decouple the requirements* for order of accuracy and spectral resolution. The result is that when the stencil size is increased, the extra degrees of freedom can be used to either increase the order of accuracy further or improve specific spectral behavior. In section 3 we present specific examples of optimized low-order schemes that are shown to have better spectral resolution than standard high-order finite differences.

To develop the theory further, it is convenient to distinguish between odd-order and even-order derivatives which leads to different behavior in Fourier space. In both cases an analytical solution can be found and is presented next.

### 2.2.1. Even derivatives

For  $d = 2q$ ,  $q = \{1, 2, \dots\}$ ,  $e(\eta)$  becomes

$$e(\eta) = (\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d) + j\mathbf{S}(\eta)^T\mathbf{a}_d.$$

The  $\mathcal{L}_2$  norm of the error is therefore,

$$\|e(\eta)\|_{\mathcal{L}_2}^2 := \int_0^\pi \gamma(\eta) \left[ (\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d)^2 + (\mathbf{S}(\eta)^T\mathbf{a}_d)^2 \right] d\eta, \quad (21)$$

$$= \mathbf{a}_d^T \underbrace{(\langle \gamma(\eta)\mathbf{C}(\eta)\mathbf{C}^T(\eta) \rangle + \langle \gamma(\eta)\mathbf{S}(\eta)\mathbf{S}^T(\eta) \rangle)}_{\mathbf{Q}_d} \mathbf{a}_d - 2\mathbf{a}_d^T \underbrace{\langle \gamma(\eta)(-1)^q\eta^d\mathbf{C}(\eta) \rangle}_{\mathbf{r}_d} + \langle \gamma(\eta)\eta^{2d} \rangle, \quad (22)$$

and the optimization problem Eq. (20) can be written as

$$\min_{\mathbf{a}_d \in \mathbb{R}^{2M+1}} (\mathbf{a}_d^T \mathbf{Q}_d \mathbf{a}_d - 2\mathbf{a}_d^T \mathbf{r}_d), \text{ subject to Eq. (4)}, \quad (23)$$

where the constant term  $\langle \eta^{2d} \rangle$  is ignored in the minimization. The system Eq. (23) is a quadratic programming problem, with linear equality constraints. We can solve this problem analytically, which is determined from the Karush-Kuhn-Tucker (KKT) condition [18]. The optimal solution for the coefficients  $\mathbf{a}_d$ , which will be denoted with an asterisk ( $\mathbf{a}_d^*$ ), satisfies the following KKT condition

$$\underbrace{\begin{bmatrix} \mathbf{Q}_d & \mathbf{X}_d \\ \mathbf{X}_d^T & \mathbf{0}_{(d+p+1) \times (d+p+1)} \end{bmatrix}}_{:=\mathbf{K}} \begin{pmatrix} \mathbf{a}_d^* \\ \boldsymbol{\lambda}_d^* \end{pmatrix} = \begin{pmatrix} \mathbf{r}_d \\ \mathbf{y}_d^T \end{pmatrix}, \quad (24)$$

where  $\boldsymbol{\lambda}_d \in \mathbb{R}^{d+p+1}$  is the Lagrange multiplier associated with the constraint in Eq. (4). Since  $\mathbf{X}_d^T \in \mathbb{R}^{(d+p+1) \times S}$  has full row rank, then the KKT matrix  $\mathbf{K}$  is nonsingular and Eq. (24) has unique solution  $(\mathbf{a}_d^*, \boldsymbol{\lambda}_d^*)$  given by

$$\begin{pmatrix} \mathbf{a}_d^* \\ \boldsymbol{\lambda}_d^* \end{pmatrix} = \begin{bmatrix} \mathbf{Q}_d & \mathbf{X}_d \\ \mathbf{X}_d^T & \mathbf{0}_{(d+p+1) \times (d+p+1)} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{r}_d \\ \mathbf{y}_d^T \end{pmatrix}. \quad (25)$$

Since everything on the right-hand-side is known, Eq. (25) provides the coefficients  $\mathbf{a}_d^*$  for the finite difference approximation Eq. (2) of order  $p+1$  which minimize the spectral error Eq. (19).

As an example consider a second-order scheme for a second derivative. In this case, a 3-point stencil ( $S = 2M+1 = 3$ ) yields a unique solution to Eq. (4), namely the common approximation  $\mathbf{a}_2^* = [1 \ -2 \ 1]^T$  or  $f'_i \approx (f_{i-1} - 2f_i + f_{i+1})/\Delta x^2$ . If one now retains a second-order approximation but increases the stencil size  $S$ , the additional degrees of freedom are utilized to reduce, through the minimization process, the spectral error in the approximation. For illustration purposes assume we would like to resolve as accurately as possible all wavenumbers in the range  $\eta \in [0, 2.5]$ . In this case, one can naturally choose  $\gamma(\eta) = 1$  for  $\eta \in [0, 2.5]$  and  $\gamma(\eta) = 0$  otherwise. Upon solving Eq. (25), we obtain the coefficients of the resulting schemes which are shown in Fig. 1 and Table 1.

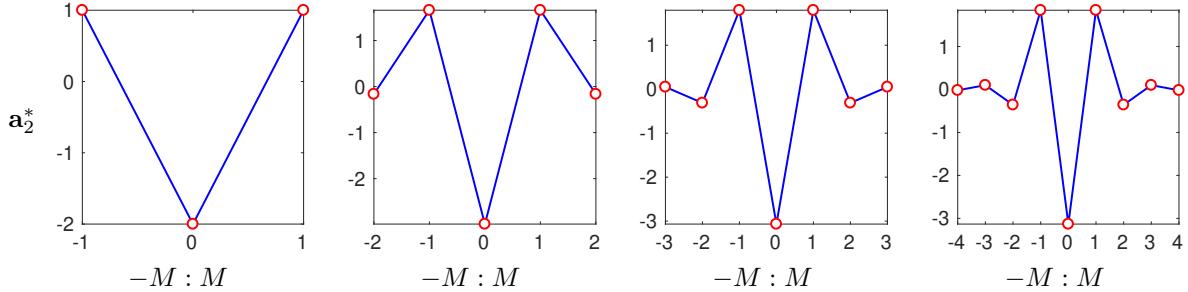


Figure 1: Optimal coefficients for various stencil sizes  $S = 2M + 1$ , for a second derivative with second order accuracy. The coefficients were obtained using  $\gamma(\eta) = 1$ , for  $\eta \in [0, 2.5]$  and  $\gamma(\eta) = 0$  otherwise.

	$M = 1$	$M = 2$	$M = 3$	$M = 4$
$a_0^*$	-2	-2.986945912146335	-3.067324780469417	-3.132525936497260
$a_1^*$	1	1.657963941430890	1.795865984254199	1.843958787844204
$a_2^*$		-0.164490985357722	-0.312793272384242	-0.357929955982910
$a_3^*$			0.050589678364752	0.099426449444277
$a_4^*$				-0.019192313056941

Table 1: Numerical coefficients for the schemes in Fig. 1. Note that since the resulting schemes are symmetric around the central grid point, only coefficients on one side are shown.

The spectral accuracy of these schemes is shown in Fig. 2 where we compare  $\tilde{\eta}^2$  (magenta) to  $\eta^2$  (dashed black) as in Eq. (17): the difference between these two curves correspond the error representing the derivative at a given frequency. The standard second order scheme,  $M = 1$ , is shown with the red line for comparison. In the top panels of the figure, we clearly see that as we increase  $M$  a better representation of the exact derivative is achieved in the range  $\eta \in [0, 2.5]$  as expected. There is also a stark difference between the standard and the optimized schemes which becomes more prominent as  $\eta$  or  $M$  increases.

Further information about the nature of the error can be obtained by analyzing the real and imaginary components of  $e(\eta)$ . This is shown in the middle and bottom panels of Fig. 2. While there are significant errors in the real component  $\Re[e(\eta)]$  for  $\eta \in [0, 2.5]$  when  $M = 1$ , they decrease substantially when  $M$  is increased, as expected. The errors outside of the support of  $\gamma(\eta)$  are marginally affected since those are not modes for which the minimization process seeks optimal solutions. The imaginary part  $\Im[e(\eta)]$  on the other hand presents zero error. This fact can indeed be proven and is presented here as,

**Lemma 1.** *Finite difference approximations Eq. (2) obtained from Eq. (20) with  $d$  even, yield a spectral error Eq. (14) which satisfies  $\Im[e(\eta)] = 0$ .*

*Proof.* See Appendix A. □

This result has implications in terms of the type of error expected (dispersive versus dissipative) when the scheme is used in a fully discretized PDE. In particular, from Eq. (18) we have  $G_2 = \tilde{\eta}^2/\eta^2 = (-e(\eta)/\eta^2 + 1)$  which implies that the modified wavenumber will not have a phase error since  $\Im[e(\eta)] = 0$ . Only dissipative errors are thus expected.

We note here that the error has also been defined using an additional arbitrary coefficient [9, 5] to emphasize the relative importance of the dissipative and dispersive errors. That is, one modifies Eq. (21) to  $\|e(\eta)\|_{\mathcal{L}_2}^2 = \int_0^\pi \gamma(\eta)[\sigma(\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q \eta^d)^2 + (1 - \sigma)(\mathbf{S}(\eta)^T \mathbf{a}_d)^2] d\eta$  where  $\sigma \in [0, 1]$ . However, from the proof of lemma 1, the minimum value of the error is obtained when one of the integrals is identically zero. Therefore the parameter  $\sigma$  will only appear as a constant and will have no effect on the optimal coefficients. The parameter  $\sigma$ , thus, is of relevance only when the grid is biased in one direction. In this scenario, neither

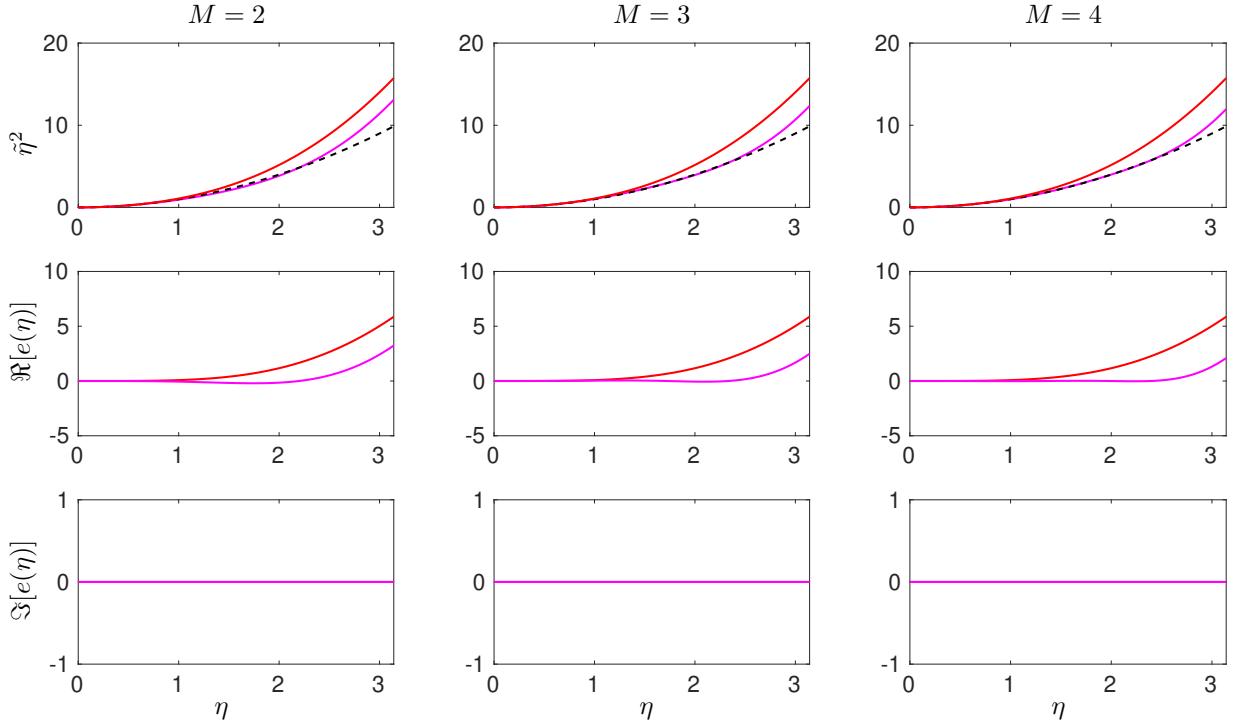


Figure 2: Spectral accuracy for the schemes presented in Fig. 1 and Table 1. Top, middle and bottom rows show the modified wavenumber, the real part, and the imaginary part of the spectral error, respectively. Left, middle and right columns correspond to  $M = 2, 3$ , and  $4$ , respectively. Dashed black line: exact differentiation ( $\tilde{\eta}^2 = \eta^2$ ). Magenta solid line:  $\tilde{\eta}$  for optimized schemes. In all plots, red solid line corresponds to standard second-order scheme ( $M = 1$ ) for comparison.

of the integrals in Eq. (21) will be 0, resulting in error with both real and imaginary part. The optimal solution will then be a minimum of the summation of dispersive and dissipative errors.

It is also of interest to understand how errors change with  $S$  (or  $M$ ). This is so because increasing  $N$  (number of grid points) or  $S$  (stencil size) leads to more computations and, thus, more computationally intensive simulations. For standard schemes, when one increases the number of points in the stencil ( $S$ ), one also increases the order of accuracy. In particular, the error for a given  $S$  (or  $M$ ) is proportional to  $\Delta x^{2M}$  where the proportionality constant also decreases with  $M$ . While there is no easy way to represent these constants in closed form, from their structure one may then still expect an approximately exponential decrease in the error as the stencil size increases. This is indeed what we see in Fig. 3 where we show the error  $\|e(\eta)\|_{L_2}^2$  as a function of  $M$  (black symbols). For optimized schemes of fixed order, on the other hand, it is not obvious a priori the rate of convergence with the number of grid points used in the stencil. However, we can readily evaluate the error numerically to assess this convergence rate. This is shown in Fig. 3 where we include the result of such numerical calculations with the second-order optimized schemes (blue symbols) in Fig. 1 as  $M$  is increased. It can be seen that the error also decreases approximately exponentially with  $M$ . While for  $M = 1$  the two curves coincide as expected since both approaches lead to the same scheme, optimized schemes (blue) present a better convergence rate as the stencil size increases.

### 2.2.2. Odd derivatives

For  $d = 2q + 1$ ,  $q = \{0, 1, \dots\}$ ,  $e(\eta)$  becomes

$$e(\eta) = \mathbf{C}^T(\eta)\mathbf{a}_d + j(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d).$$

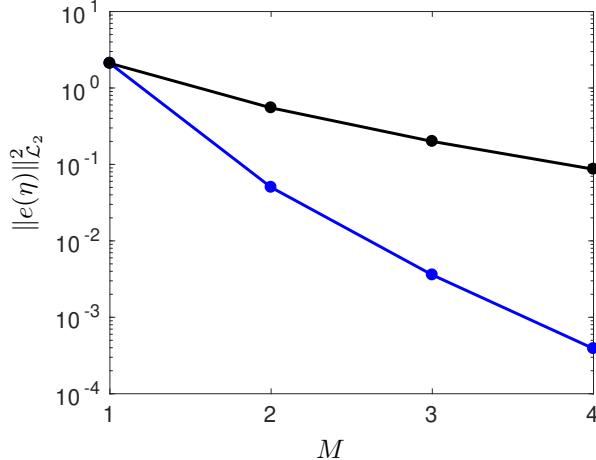


Figure 3: Optimal spectral error for various stencil size  $M$ , approximating second derivative. The blue line corresponds to the second order optimized schemes presented in Fig. 1 and Table 1. The black line corresponds to the standard schemes with stencil size  $M$  and  $2M$  order of accuracy.

The  $\mathcal{L}_2$ -norm of the error for this case is

$$\|e(\eta)\|_{\mathcal{L}_2}^2 := \int_0^\pi \gamma(\eta) \left[ (\mathbf{C}^T(\eta)\mathbf{a}_d)^2 + (\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q \eta^d)^2 \right] d\eta, \quad (26)$$

$$= \mathbf{a}_d^T \underbrace{\left( \langle \gamma(\eta) \mathbf{C}(\eta) \mathbf{C}^T(\eta) \rangle + \langle \gamma(\eta) \mathbf{S}(\eta) \mathbf{S}^T(\eta) \rangle \right)}_{\mathbf{Q}_d} \mathbf{a}_d - 2 \mathbf{a}_d^T \underbrace{\langle \gamma(\eta) (-1)^q \eta^d \mathbf{S}(\eta) \rangle}_{\mathbf{r}_d} + \langle \gamma(\eta) \eta^{2d} \rangle. \quad (27)$$

The optimization problem is the same as Eq. (23) with  $\mathbf{r}_d$  given by Eq. (27), and the optimal solution is given by Eq. (25).

As an example, we consider here again a second-order approximation but of the first derivative for increasing values of  $M$ . The resulting schemes are shown in Fig. 4 and their spectral behavior is shown in Fig. 5. Similar conclusions to the second derivative example shown above are observed. As  $M$  increases the wavenumbers where  $\gamma(\eta)$  is non-zero are increasingly well resolved which is seen as  $\tilde{\eta}$  becoming closer to  $\eta$  in the top panels of Fig. 5. We can also see that the real component of the spectral error (middle panels) is zero. This is due to the following lemma.

**Lemma 2.** *Finite difference approximations Eq. (2) obtained from Eq. (20) with  $d$  odd, yield a spectral error Eq. (14) which satisfies  $\Re[e(\eta)] = 0$ .*

*Proof.* See Appendix B. □

Since  $G_1 = e(\eta)/(j\eta) + 1$ , and  $e(\eta)$  has only an imaginary component, then  $G_1$  is real. As in the example before, errors are then expected to be dispersive in nature though the exact nature of the error would depend on the PDE in which such a scheme is used. As before, we found the global error to decrease exponentially with  $M$  as seen in Fig. 6 and faster than non-optimized schemes.

### 2.2.3. Effect of $\gamma(\eta)$

An important feature of the framework presented here for optimized schemes is that depending upon the physics of the problem, one can choose what scales need to be properly resolved. By setting  $\gamma(\eta)$  in Eq. (19) as unity, equal weight is given to all wavenumbers while solving the minimization problem. The formulation, however, is more general and allow us to change  $\gamma(\eta)$  to, for example, resolve a subset of wavenumbers more accurately than others. To illustrate this, consider an eight-order standard scheme ( $M = 4$ ) for the

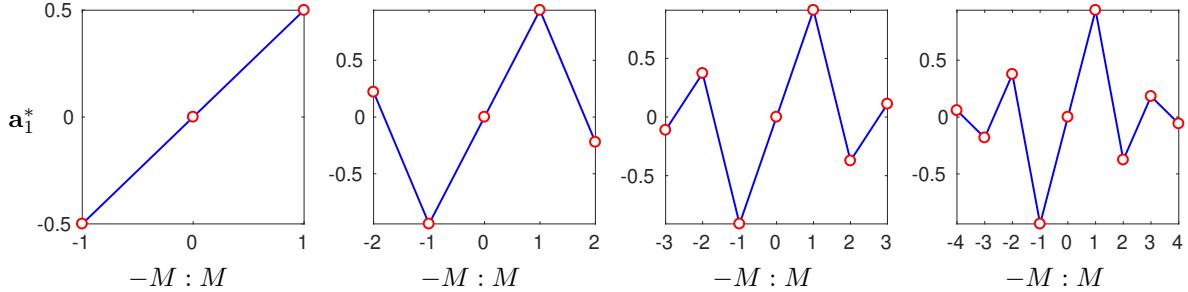


Figure 4: Optimal coefficients for various stencil size  $M$ , approximating first derivative with second order accuracy. The coefficients were obtained using  $\gamma(\eta) = 1$ , for  $\eta \in [0, 2.5]$ ; and  $\gamma(\eta) = 0$  otherwise.

	$M = 1$	$M = 2$	$M = 3$	$M = 4$
$a_1^*$	0.50	0.941502204636976	0.911624839168511	0.939273151104227
$a_2^*$		-0.220751102318488	-0.372951233396604	-0.376375957228243
$a_3^*$			0.111425875874899	0.182092697439389
$a_4^*$				-0.058199832241477

Table 2: Numerical coefficients for the schemes in Fig. 4. Note that since the resulting schemes are anti-symmetric ( $a_{-M}^* = -a_M^*$ ) around the central grid point, only coefficients on one side are shown.

second derivative whose modified wavenumber can be readily computed analytically. The relative error at each wavenumber, which with Eq. (18) can be written as  $|\tilde{\eta}^2 - \eta^2|/\eta^2 = G_2 - 1$ , is shown in Fig. 7 as a black line. We can see that standard schemes have very low errors at low  $\eta$  but become progressively worse at high  $\eta$ . In fact, there are about 15 orders of magnitude difference between the error incurred at low and high wavenumbers. This situation may present some challenges when these schemes are used to resolve multiscale problems where all wavenumbers contribute to the dynamics. In contrast, the optimized scheme with  $\gamma(\eta) = 1$  in  $\eta \in [0, 2.5]$  shown in magenta in the figure, results in a much flatter error in spectral space. The oscillatory nature of the error is due to the following. The optimized schemes minimize a global measure of the error based on the  $\mathcal{L}_2$  norm of difference between  $\tilde{\eta}^d$  and  $\eta^d$ . Pointwise however, the resulting scheme produces a  $\tilde{\eta}$  that can be above or below  $\eta$ ; only the appropriate integral is minimized. An easy-to-see example of this behavior is seen in Fig. 5 for  $M = 3$  (top-middle panel). The crossing points between  $\tilde{\eta}$  and  $\eta$  leads then to zero error which correspond to the down peaks in Fig. 7 for the corresponding scheme.

The other curves in Fig. 7 illustrate the effect of changing  $\gamma(\eta)$  on the spectral error. For this illustration we use  $\gamma(\eta) = e^{-\eta/\psi}$  where  $\psi$  is a constant that controls how quickly the weighting function drops to zero near the origin. A similar form has been used in some applications [9, 7] though for some  $\psi$ . As  $\gamma(\eta)$  becomes steeper (small  $\psi$ ), the error for lower and higher wavenumbers decreases and increases, respectively. The optimized schemes then get closer to the strongly non-uniform distribution of error in spectral space of standard schemes.

Clearly one can use, for example, a banded  $\gamma(\eta)$ , such that it is non-zero only for some intermediate wavenumbers. Or one may need to resolve to distinct bands in wavenumber space which can easily be accommodated by an appropriate choice of  $\gamma(\eta)$ . The choice of  $\gamma(\eta)$  provides a high flexibility to select how the available information (the value of the function at  $S$  grid points) to resolve only the scales relevant to the problem being solved.

### 2.3. Stability

As discussed in the introduction, it is common to first choose a scheme of given order and then verify that spectral accuracy is acceptable. That scheme is then used to discretize a PDE. However, before this numerical arrangement can be utilized, one needs to determine whether the fully discretized PDE is stable.

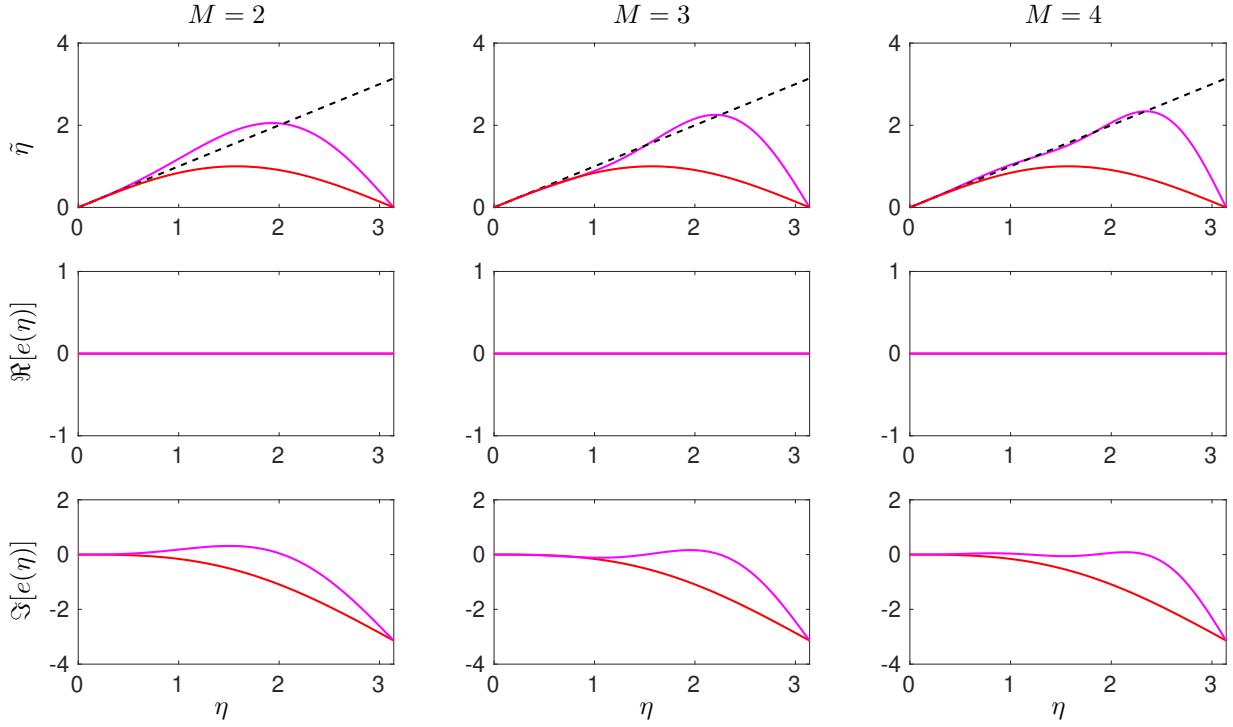


Figure 5: Spectral accuracy for the schemes presented in Fig. 4 and Table 2. Top, middle and bottom rows show the modified wavenumber, the real part, and the imaginary part of the spectral error, respectively. Left, middle and right columns correspond to  $M = 2, 3$ , and  $4$ , respectively. Dashed black line: exact differentiation ( $\tilde{\eta} = \eta$ ). Magenta solid line:  $\tilde{\eta}$  for optimized schemes. In all plots, red solid line corresponds to standard second-order scheme ( $M = 1$ ) for comparison.

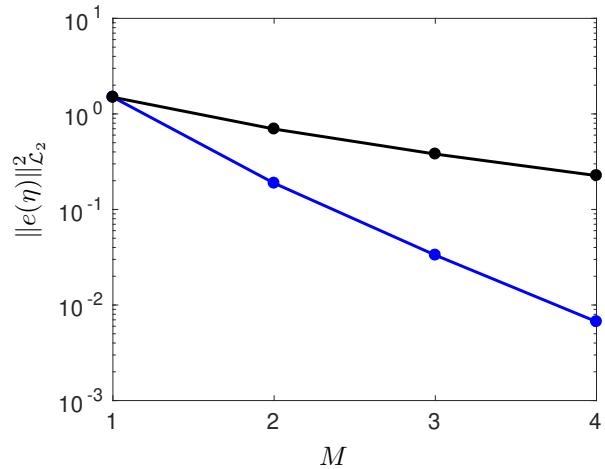


Figure 6: Optimal spectral error for various stencil size  $M$ , approximating first derivative. The blue line corresponds to the second order optimized schemes presented in Fig. 4 and Table 2. The black line corresponds to the standard schemes with stencil size  $M$  and  $2M$  order of accuracy.

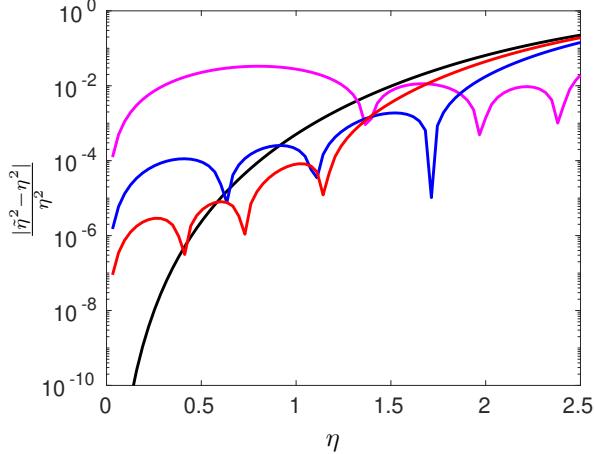


Figure 7: Spectral error for various  $\gamma(\eta)$  for second derivative. Standard eighth order scheme (black),  $\gamma(\eta) = 1$  (magenta),  $\gamma(\eta) = \exp(-\eta/0.1)$  (blue) and  $\gamma(\eta) = \exp(-\eta/0.06)$  (red) for  $\eta \in [0, 2.5]$  and 0 otherwise.

Different methods to assess stability of a given discretized equation have been discussed extensively [1]. Our objective here is not simply to determine whether the schemes developed above are stable or not. Rather is to incorporate the stability requirement into a unified formulation. This framework would thus provide, for a given  $M$ , a stable scheme of given order with the best spectral resolution possible for wavenumbers of interest.

For this, consider the general linear partial differential equation

$$\frac{\partial f}{\partial t} = \sum_{d=1}^D \beta_d \frac{\partial^d f}{\partial x^d}. \quad (28)$$

discretized over the entire domain with  $N := 2M_{\max} + 1$  grid points. As before, spatial derivatives are approximated using a stencil of size  $S = 2M + 1$ , where  $M$  can take values from 1 to  $M_{\max}$ . The  $d^{\text{th}}$  derivative at the  $i^{\text{th}}$  grid point is parameterized by  $\mathbf{a}_{i,d} \in \mathbb{R}^S$ . Define  $\mathbf{A}_d \in \mathbb{R}^{N \times S}$  to be the vertical stacking of  $\mathbf{a}_{i,d}^T$ , i.e.

$$\mathbf{A}_d := \begin{bmatrix} \mathbf{a}_{1,d}^T \\ \vdots \\ \mathbf{a}_{N,d}^T \end{bmatrix}. \quad (29)$$

The order accuracy constraint Eq. (4) for every  $\mathbf{a}_{i,d}$ , can then be compactly written as

$$\mathbf{A}_d \mathbf{X}_d = \mathbf{Y}_d, \quad (30)$$

for  $d = 1, \dots, D$ ; and  $\mathbf{Y}_d := \mathbf{1}_{N \times 1} \otimes \mathbf{y}_d$ .

The cost function to be minimized is the sum of the spectral error Eq. (19) at all locations  $i$ :

$$\begin{aligned} \sum_{i=1}^N \|e_i(\eta)\|_{\mathcal{L}_2}^2 &:= \sum_{i=1}^N \sum_{d=1}^D (\mathbf{a}_{i,d}^T \mathbf{Q}_d \mathbf{a}_{i,d} - 2\mathbf{a}_{i,d}^T \mathbf{r}_d) + N \langle \eta^{2d} \rangle, \\ &= \sum_{d=1}^D \mathbf{v}_d^T (\mathbf{I}_N \otimes \mathbf{Q}_d) \mathbf{v}_d - 2\mathbf{v}_d^T (\mathbf{1}_{N \times 1} \otimes \mathbf{r}_d), \end{aligned} \quad (31)$$

where  $\mathbf{v}_d := \text{vec}(\mathbf{A}_d^T)$ , and  $\text{vec}(\cdots)$  vectorizes a matrix by vertically stacking the columns. The minimization of Eq. (31) subjected to Eq. (30) leads to optimal schemes with a given order of accuracy. To include stability in the formulation, we first define the vectors

$$\mathbf{F} := \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix}, \text{ and } \mathbf{F}^{(d)} := \begin{pmatrix} f_1^{(d)} \\ \vdots \\ f_N^{(d)} \end{pmatrix}. \quad (32)$$

Using these definitions, the finite-difference approximation for Eq. (28) for all the grid points can be written compactly as

$$\mathbf{F}^{(d)} = \frac{1}{(\Delta x)^d} \mathbf{A}_d^\Phi \mathbf{F}, \quad (33)$$

where the matrix  $\mathbf{A}_d^\Phi$  contains the unknown coefficients  $\mathbf{a}_{i,d}$  arranged appropriately for the correct computation of the  $d^{\text{th}}$  derivative at  $i^{\text{th}}$  location. The matrix  $\mathbf{A}_d^\Phi$  can be constructed using shift operators the details of which are included in Appendix D. The important element here is that the spatial derivative is a linear operator acting on the value of the function  $f$  at all grid points.

### 2.3.1. Stability of Semi-discrete Scheme

With the spatial discretization from Eq. (33), Eq. (28) can be written as

$$\dot{\mathbf{F}} = \left( \sum_{d=1}^D \frac{1}{(\Delta x)^d} \beta_d \mathbf{A}_d^\Phi \right) \mathbf{F}, \quad (34)$$

whose analytical solution can be readily obtained as

$$\mathbf{F}(t) := \exp \left( \sum_{d=1}^D \frac{1}{(\Delta x)^d} \beta_d t \mathbf{A}_d^\Phi \right) \mathbf{F}_0, \quad (35)$$

where  $\mathbf{F}_0$  is the initial condition. Clearly the stability of Eq. (35) is governed by the coefficients  $\mathbf{A}_d^\Phi$ .

In order to understand the conditions under which the scheme defined by  $\mathbf{A}_d^\Phi$  is stable consider a single Fourier mode as in Eq. (11). The approximate  $d^{\text{th}}$  derivative can be conveniently written in terms of the modified wavenumber as  $(j\tilde{\eta})^d \hat{f}$  and the original PDE becomes

$$\frac{d\hat{f}}{dt} = \sum_{d=1}^D \beta_d (j\tilde{\eta})^d \hat{f}. \quad (36)$$

The solution, with  $\hat{f}_0$  being the value of  $\hat{f}$  at  $t = 0$ , is given by

$$\hat{f} = \exp \left[ \sum_{d=1}^D \beta_d (j\tilde{\eta})^d t \right]. \quad (37)$$

If the solution to the original PDE is non-increasing in time, then the discretization is considered stable if no Fourier mode grows in time. This is clearly satisfied if

$$\Re \left\{ \sum_{d=1}^D \beta_d (j\tilde{\eta})^d \right\} \leq 0 \quad (38)$$

From lemma 1 and 2, we can conclude that

$$(j\tilde{\eta})^d = (\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)) \mathbf{a}_d = \begin{cases} j\mathbf{S}^T(\eta) \mathbf{a}_d, & \text{for odd derivative,} \\ \mathbf{C}(\eta)^T \mathbf{a}_d, & \text{for even derivative,} \end{cases}$$

because  $\mathbf{C}^T(\eta)\mathbf{a}_d = 0$  for odd derivatives and  $\mathbf{S}^T(\eta)\mathbf{a}_d = 0$  for even derivatives. Then,

$$\begin{aligned} \sum_{d=1}^D \beta_d(j\tilde{\eta})^d &= j\beta_1\mathbf{S}^T(\eta)\mathbf{a}_1 + \beta_2\mathbf{C}^T(\eta)\mathbf{a}_2 + j\beta_3\mathbf{S}^T(\eta)\mathbf{a}_3 + \beta_4\mathbf{C}^T(\eta)\mathbf{a}_4 + \cdots, \\ &= (\beta_2\mathbf{C}^T(\eta)\mathbf{a}_2 + \beta_4\mathbf{C}^T(\eta)\mathbf{a}_4 + \cdots) + j(\beta_1\mathbf{S}^T(\eta)\mathbf{a}_1 + \beta_3\mathbf{S}^T(\eta)\mathbf{a}_3 + \cdots). \end{aligned} \quad (39)$$

Therefore, Eq. (38) implies

$$\beta_2\mathbf{C}^T(\eta)\mathbf{a}_2 + \beta_4\mathbf{C}^T(\eta)\mathbf{a}_4 + \cdots \leq 0 \quad (40)$$

which provides a general constraint to assure stability of the semidiscrete system.

Clearly stability for the semi-discrete system depends only on the even derivatives. In that case  $d$  can be written as  $2q$  for  $q = 1, 2, \dots$ , and  $(j\tilde{\eta})^d = (-1)^q\tilde{\eta}^{2q}$ . Since the optimization of spectral error guarantees  $(j\tilde{\eta})^d = \mathbf{C}^T(\eta)\mathbf{a}_{2q}$ , we have  $\mathbf{C}^T(\eta)\mathbf{a}_{2q} = (-1)^q\tilde{\eta}^{2q}$ . Consequently, the sign of  $\mathbf{C}^T(\eta)\mathbf{a}_{2q}$  alternates with  $q$ , i.e.

$$\begin{aligned} \text{for } q = 1, \mathbf{C}^T(\eta)\mathbf{a}_2 &= -\tilde{\eta}^2 \leq 0, \\ \text{for } q = 2, \mathbf{C}^T(\eta)\mathbf{a}_4 &= \tilde{\eta}^4 \geq 0, \\ \text{for } q = 3, \mathbf{C}^T(\eta)\mathbf{a}_6 &= -\tilde{\eta}^6 \leq 0, \end{aligned}$$

and so on. Therefore, if the coefficients of the original PDE can be written as  $\beta_{2q} := (-1)^{q+1}\gamma_{2q}^2$  for some  $\gamma_{2q}$ , then Eq. (40) is implicitly satisfied as it reduces to a sum of negative numbers. In general, if  $\beta_{2q}$  does not have sign as required by  $(-1)^{q+1}\gamma_{2q}^2$ , then, while that particular derivative is unstable, the system is still stable if

$$-\beta_2\tilde{\eta}^2 + \beta_4\tilde{\eta}^4 - \beta_6\tilde{\eta}^6 + \cdots \leq 0. \quad (41)$$

### 2.3.2. Stability of Fully-Discrete Scheme

While semi-discrete analyses provide some information about the temporal behavior of the discretized spatial derivatives in the PDEs, one is ultimately interested in the stability of the fully discretized system. For concreteness assume that the time discretization of Eq. (28) is done using a *forward difference*. With Eq. (33) for the spatial discretization we write Eq. (28) as:

$$\mathbf{F}^{k+1} = \left( \mathbf{I}_N + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{A}_d^\Phi \right) \mathbf{F}^k. \quad (42)$$

This is a linear discrete-time system in  $\mathbf{F}$ , where the system matrix is linearly dependent on the stencil coefficients  $\mathbf{A}_d$ , discretization parameters  $\Delta t$  and  $\Delta x$ , and the coefficients  $\beta_d$ . In what follows we assume  $\Delta x$  and  $\beta_d$  are given which is a typical situation in simulations of physical systems. The objective then is to determine  $\Delta t$  and  $\mathbf{A}_d$  so that stability is achieved subjected to specifications in order of accuracy and spectral resolution.

In general, stability is guaranteed if the spectral radius of the evolution matrix is bounded by unity [1]:

$$\lambda_{\max} \left( \mathbf{I}_N + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{A}_d^\Phi \right) \leq 1. \quad (43)$$

Because the spectral radius is bounded by matrix norms, we guarantee stability by bounding the 2-norm, i.e.

$$\left\| \mathbf{I}_N + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{A}_d^\Phi \right\|_2 \leq 1. \quad (44)$$

Using Schur complements, this inequality is equivalent to

$$\begin{bmatrix} \mathbf{I}_N & \left( \mathbf{I}_N + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{A}_d^\Phi \right)^T \\ \left( \mathbf{I} + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{A}_d^\Phi \right) & \mathbf{I}_N \end{bmatrix} \geq 0. \quad (45)$$

This (inequality) constraint coupled with the order of accuracy (equality) constraint Eq. (30) completes the set of constraints for the optimization problem aimed at minimizing the spectral error given by Eq. (31).

Unfortunately, the matrix inequality in Eq. (45) has products of  $\Delta t$  and  $\mathbf{A}_d$ , which makes the problem non convex. Thus, more general techniques than convex optimization have to be employed to solve the problem. Here we present two approaches. In the first approach, we determine  $\mathbf{A}_d$  analytically that minimizes the spectral error and then maximize  $\Delta t$  for which stability is achieved. In the second approach, we assume a value for  $\Delta t$  and determine  $\mathbf{A}_d$  which minimizes spectral error and guarantees stability. Both these approaches are discussed in detail below.

### 2.3.3. Given spectrally optimal $\mathbf{A}_d$ , maximize $\Delta t$ and guarantee stability simultaneously

In this approach, we solve for the optimal  $\mathbf{a}_{i,d}$  analytically that, for a given order, minimizes the spectral error and then maximize  $\Delta t$  for which stability is guaranteed. Optimization of the spectral error with given order of accuracy is given by previous formulation as

$$\min_{\{\mathbf{A}_d\}_{d=1}^D} \sum_{d=1}^D \mathbf{v}_d^T (\mathbf{I}_N \otimes \mathbf{Q}_d) \mathbf{v}_d - 2\mathbf{v}_d^T (\mathbf{1}_{N \times 1} \otimes \mathbf{r}_d),$$

subject to

$$\mathbf{A}_d \mathbf{X}_d = \mathbf{Y}_d, \text{ for } d = 1, \dots, D,$$

where  $\mathbf{v}_d := \text{vec}(\mathbf{A}_d^T)$ .

We observe that the cost and constraint functions are separable with respect to  $d$ , and thus can be independently optimized using,

$$\min_{\mathbf{A}_d} \sum_{d=1}^D \mathbf{v}_d^T (\mathbf{I}_N \otimes \mathbf{Q}_d) \mathbf{v}_d - 2\mathbf{v}_d^T (\mathbf{1}_{N \times 1} \otimes \mathbf{r}_d), \text{ subject to } \mathbf{A}_d \mathbf{X}_d = \mathbf{Y}_d, \quad (46)$$

for  $d = 1, \dots, D$ . The linear constraint  $\mathbf{A}_d \mathbf{X}_d = \mathbf{Y}_d$  can be written as

$$(\mathbf{I}_N \otimes \mathbf{X}_d^T) \mathbf{v}_d = \mathbf{Y}_d^T.$$

Therefore, the optimization problem in Eq. (46) can be written as

$$\min_{\mathbf{v}_d} \mathbf{v}_d^T (\mathbf{I}_N \otimes \mathbf{Q}_d) \mathbf{v}_d - 2\mathbf{v}_d^T (\mathbf{1}_{N \times 1} \otimes \mathbf{r}_d), \quad (47)$$

$$\text{subject to } (\mathbf{I}_N \otimes \mathbf{X}_d^T) \mathbf{v}_d = \mathbf{Y}_d^T, \quad (48)$$

which has the analytical solution

$$\begin{pmatrix} \mathbf{v}_d \\ \mathbf{\Lambda}_d \end{pmatrix}^* = \begin{bmatrix} \mathbf{I}_N \otimes \mathbf{Q}_d & \mathbf{I}_N \otimes \mathbf{X}_d \\ \mathbf{I}_N \otimes \mathbf{X}_d^T & \mathbf{I}_N \otimes \mathbf{0} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{1}_{N \times 1} \otimes \mathbf{r}_d \\ \mathbf{Y}_d^T \end{pmatrix}, \quad (49)$$

where  $\mathbf{\Lambda}_d$  is the vector of Lagrange multipliers associated with the constraints. Finally, we can recover  $\mathbf{A}_d^*$ , the coefficients of the scheme that minimizes the spectral error and is stable, from  $\mathbf{v}_d^*$ . While intuitively it is clear that this solution should be the same at all grid points in the domain, here we present, for completion, a proof as:

**Lemma 3.** *Optimal explicit symmetrical finite-difference approximation is invariant of the grid point location in the domain.*

*Proof.* See Appendix C. □

From Lemma 3, we write the complete system as a stacking of local solutions:

$$\mathbf{A}_d^* = \mathbf{1}_{N \times 1} \otimes \mathbf{a}_d^{*T}. \quad (50)$$

Therefore, the maximum  $\Delta t$  for which stability is guaranteed is obtained by solving the optimization problem Eq. (45), that is

$$\max_{\Delta t} \text{ subject to } \begin{bmatrix} \mathbf{I}_N & \left( \mathbf{I} + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{A}_d^\Phi \right)^T \\ \left( \mathbf{I} + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{A}_d^\Phi \right) & \mathbf{I}_N \end{bmatrix} \geq 0, \quad (51)$$

where  $\mathbf{A}_d^\Phi$  is determined using Eq. (D.5) and  $\mathbf{A}_d^*$  from Eq. (50). The optimization in Eq. (51) is a convex optimization problem in  $\Delta t$  and can be efficiently solved numerically using software such as `cvx`[19]. This formulation can then be used to obtain stability limits of the fully-discrete system.

As an example, consider an advection-diffusion equation, that is Eq. (28) with  $D = 2$ . For given values of  $\beta_1, \beta_2 > 0$ , and  $\Delta x$ , one can solve the optimization problem to obtain the largest  $\Delta t$  for which a previously obtained optimal  $\mathbf{A}_d^*$  remain stable. Note that the stability of the scheme will depend only on the non-dimensional parameters  $\Delta t \beta_d / \Delta x^d$ . For  $D = 2$ , these are commonly called convective and diffusive CFL, that is,  $r_c := \Delta t \beta_1 / \Delta x$  and  $r_d := \Delta t \beta_2 / \Delta x^2$ , respectively. Stability regions can then be obtained by e.g. sweeping values of  $\Delta x$  and plot results in terms of  $r_c$  and  $r_d$ .

This is what we show in Fig. 8 for the advection-diffusion equation for different values of  $M$ . For  $M = 1$  the numerically evaluated stability region is the same as the analytical form found in textbooks for the standard second-order 3-point stencil for first and second derivatives, namely  $r_c^2 \leq 2r_d \leq 1$  [1]. As  $M$  is increased the stability region becomes smaller. This is also generally consistent with standard schemes for which as the order (and stencil size) increases the stability region shrinks. In both cases, this is related to the decreasing dissipation at high wavenumbers as spectral resolution improves which could trigger instabilities.

In summary, the formulation here allows us to obtain, for optimal schemes of given order and maximum spectral resolution, the largest step size which guarantees stability or, more generally, regions of stability.

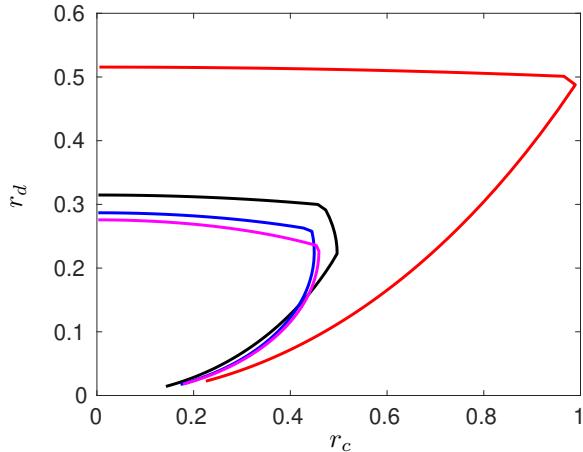


Figure 8: Stability region for the advection-diffusion equation with optimal coefficients  $\mathbf{a}_1^*$  and  $\mathbf{a}_2^*$  for different stencil sizes:  $M = 1$  (red),  $M = 2$  (black),  $M = 3$  (blue),  $M = 4$  (magenta).

#### 2.3.4. Given $\Delta t$ , optimize spectral error and guarantee stability simultaneously

In this approach, for a given  $\Delta t$  we solve for  $\mathbf{A}_d$  that simultaneously minimizes spectral error, achieves given order accuracy, and guarantees stability. This approach is of practical relevance as the time-step  $\Delta t$

would depend on the fastest physical process in the problem. For example, when chemical reactions are present in a flow, the Damkohler number (the ratio of flow time scales to chemistry time scales) can be very high which means the time scales of reactions are considerably smaller than flow time scales which forces  $\Delta t$  to be small for an accurate solution.

The formulation here is done by combining the results in previous sections to write the following optimization problem,

$$\left. \begin{array}{l} \min_{\{\mathbf{A}_d\}_{d=1}^D} \sum_{d=1}^D \sum_{i=1}^N (\mathbf{a}_{i,d}^T \mathbf{Q}_d \mathbf{a}_{i,d} - 2\mathbf{a}_{i,d}^T \mathbf{r}_d), \\ \text{subject to} \\ \mathbf{A}_d \mathbf{X}_d = \mathbf{Y}_d, \text{ for } d = 1, \dots, D; \\ \left[ \begin{array}{cc} \mathbf{I}_N & \left( \mathbf{I} + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{A}_d^\Phi \right)^T \\ \left( \mathbf{I} + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{A}_d^\Phi \right) & \mathbf{I}_N \end{array} \right] \geq 0. \end{array} \right\} \quad (52)$$

Equation Eq. (52) is a convex optimization problem [20] in variables  $\{\mathbf{A}_d\}_{d=1}^D$ , with  $\mathbf{A}_d \in \mathbb{R}^{N \times S}$ , and can be efficiently solved numerically using `cvx` [19]. The time step  $\Delta t$  can be maximized by iteratively solving Eq. (52) with increasing  $\Delta t$  until the problem becomes unfeasible. Since the spectra of a matrix is continuous in terms of the elements, the maximum  $\Delta t$  can be determined using a bisection algorithm [21].

In Fig. 9 we show the result of such a computation again for an advection-diffusion equation discretized with second-order approximations. In particular, we show contours of the sum of the spectral error  $\|e(\eta)\|_{L_2}^2$  in the first and second derivatives for different values of  $M$  normalized by the value for  $M = 1$ . Colored areas represent regions where schemes are stable. For  $M = 1$ , the stable region is well known as discussed above and presents values of 1.0 due to the normalization chosen. Interestingly, as  $M$  increases the stability area increases instead of decreasing as in Fig. 8. This illustrates an interesting aspect of the formulation. By providing additional degrees of freedom (bigger stencil size) but fixing the formal order of accuracy, Eq. (52) minimizes spectral error with the condition that the scheme be stable when some fixed  $\Delta t$  is used. As can be seen, for  $M = 4$ , schemes can remain stable for very large CFL numbers (an order of magnitude larger than for standard schemes) though with larger spectral error. In other words, the framework allows one to trade off error with stability. If one is interested in stationary states, for example, one can use the optimal scheme with  $M = 4$  with very large time steps in a fully explicit arrangement to solve transients. After the desired steady state is attained one can reduce  $\Delta t$  with corresponding optimized schemes to reduce errors. In fact, by comparing part (a) and (d) in the figure, we can see that for regions where  $M = 1$  is stable,  $M = 4$  provides errors which could be two orders of magnitude smaller. Conversely, for similar accuracy, schemes with  $M = 4$  can be used with much larger  $\Delta t$  than a scheme with  $M = 1$ . Note that this approach could prove beneficial at very large levels of parallelism where implicit schemes in time, while providing good stability characteristics for large time steps, become challenging due to the necessity to invert large matrices.

It is interesting to observe how optimal coefficients change to maintain stability for very large time steps. Since biased (upwind) schemes tend to be more dissipative and stable, one would expect that gains in stability are mediated by losses in accuracy due to biasing. This is, in fact, what is observed. In order to quantify deviations from symmetry ( $a_i = a_{-i}$ , where  $a_i$ s are the coefficients in Eq. (2) that are obtained from  $\mathbf{a}_d$ ) for even derivatives or anti-symmetry ( $a_i = -a_{-i}$ ) for odd derivatives we define the metric:

$$A = \sum_{i=1}^M \frac{|a_{-i} + (-1)^{d+1} a_i|}{|a_{-i}| + |a_i|}, \quad (53)$$

for the  $d$ -th derivative. Clearly  $A = 0$  when the coefficients are symmetric (for even derivatives) or anti-symmetric (for odd derivatives). Larger values of  $A$  are associated with increasingly biased approximations. The denominator in Eq. (53) is included such that the contribution from all coefficients are of the same order. This is needed because coefficients tend to decrease in magnitude with distance from the point where derivatives are computed.

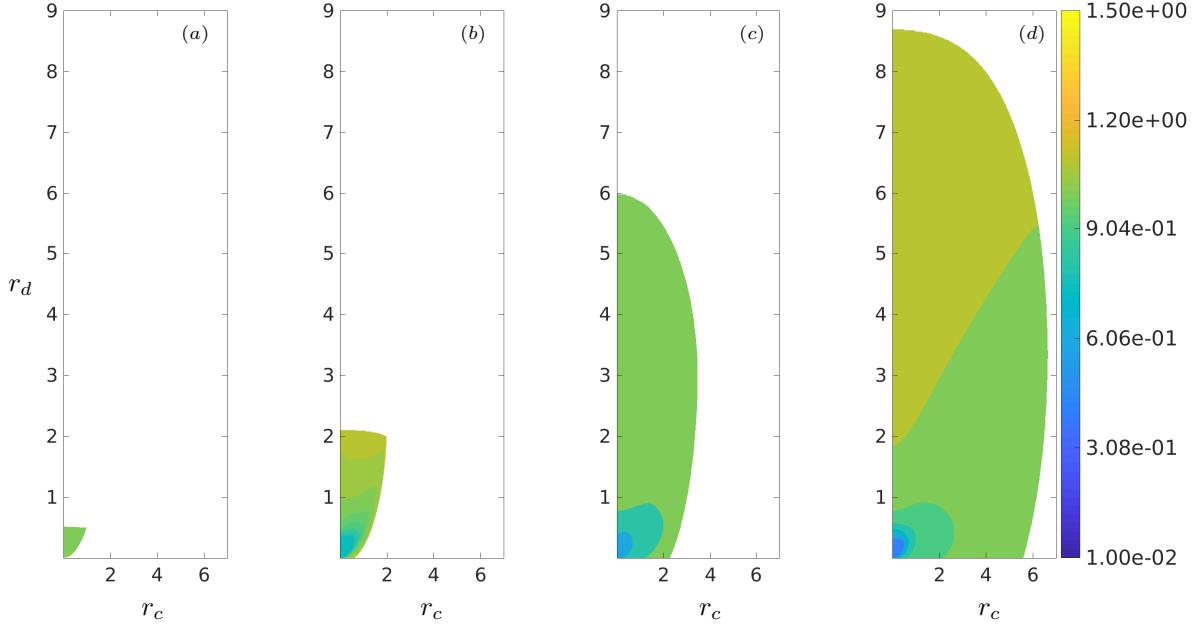


Figure 9: Tradeoff between stability and spectral accuracy with respect to  $(r_d, r_c)$ , for  $\beta_1 = 10$  and  $\beta_2 = 1$ . Contours represent sum of spectral errors  $\|e(\eta)\|_{L_2}^2$  for first and second derivatives using optimal coefficients obtained as a solution of Eq. (52) with fixed  $\Delta t$  for the advection-diffusion equation. Errors are normalized by the spectral error obtained for  $M = 1$ . Different stencil sizes shown in (a)  $M = 1$ , (b)  $M = 2$ , (c)  $M = 3$ , and (d)  $M = 4$ .

In Fig. 10(a) we plot  $A$  for the first (a) and second (b) derivatives, respectively for  $M = 2$ . As before, colored areas correspond to stable conditions. The black line in Fig. 10(a) corresponds to the stability region given in Fig. 8 for  $M = 2$ . When we are within this boundary, the value of  $A$  for the first derivative is very small which suggests that all extra degrees of freedom are used to minimize the spectral error for which symmetric stencils are generally better. However, as we move outside of this area, these extra degrees of freedom are needed to satisfy the stability constraint. By losing symmetry or anti-symmetry error increases (as seen in Fig. 9) but stability improves. In fact, asymmetry for the first derivative implies that the error has both real and imaginary parts as seen in Fig. 11. The biased nature of coefficients leads not only to dispersion errors (as assured by lemma 2 with symmetric coefficients) but also to dissipation errors as  $\Re[e(\eta)] \neq 0$ . Both dissipation and dispersion error increase as we increase the value of  $r_c$  and  $r_d$ .

A different behavior is observed for the second derivative whose contours of  $A$  are shown in Fig. 10(b). Here we see that  $A$  is very small for any  $r_c$  and  $r_d$  within the limits of stability. The implication of this is that stability is essentially governed by the first derivative, which may be intuitive. It is indeed common to stabilize fluid flow simulation codes by special treatment of the convective terms while leaving diffusive terms approximated by standard central differences. What our results show, beyond heuristic stabilization considerations is that, indeed, biasing convective terms and keeping central schemes for diffusion is the spectrally optimal way of achieving stability. They also show that, unlike [16], so-called anti-diffusion does not lead to instability due to the compensating effect of the other term in the equation. For completion, in Fig. 10(c) we show the change in optimal coefficients for the first derivative for the conditions represented by different symbols in part (a). We clearly see how the scheme changes from a completely anti-symmetric configuration (red circles) to a more biased set of coefficients. We emphasize this stable configurations are optimal in spectral space.

We close this section with two remarks about the stability limits computed here. First, in terms of

computational cost we note that as a consequence of Lemma 3 for periodic domains, the optimization problem in Eq. (52) has a total of  $n = Sd$  degrees of freedom, which is independent of the total number of grid points  $N$ . Since the computational cost in terms of time and memory for state-of-the-art optimization algorithms is  $\mathcal{O}(n)$  [22], obtaining the numerical scheme represents a small fraction of the total computational cost. Furthermore, since the time step size is typically controlled either by  $r_c$  or  $r_d$ , the computational cost of solving the equation scales as  $N^2$  or  $N^3$ , respectively, making the optimization cost increasingly smaller as the problem size increases. Second, we note that the stability limits shown in Fig. 9 were obtained numerically with machine precision. Thus, caution needs to be exercised when selecting conditions in the  $r_c$ - $r_d$  plane very close to the stability boundaries.

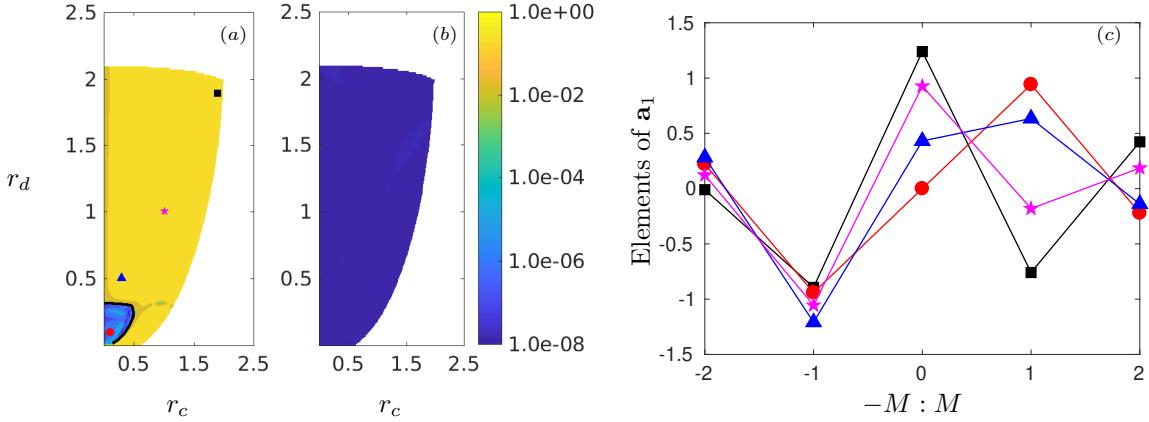


Figure 10: Contours of measure  $A$  for  $M = 2$  for (a) first derivative (b) second derivative. (c) Optimal coefficients for different values of  $r_c$  and  $r_d$  as marked, with the same symbols, in (a).

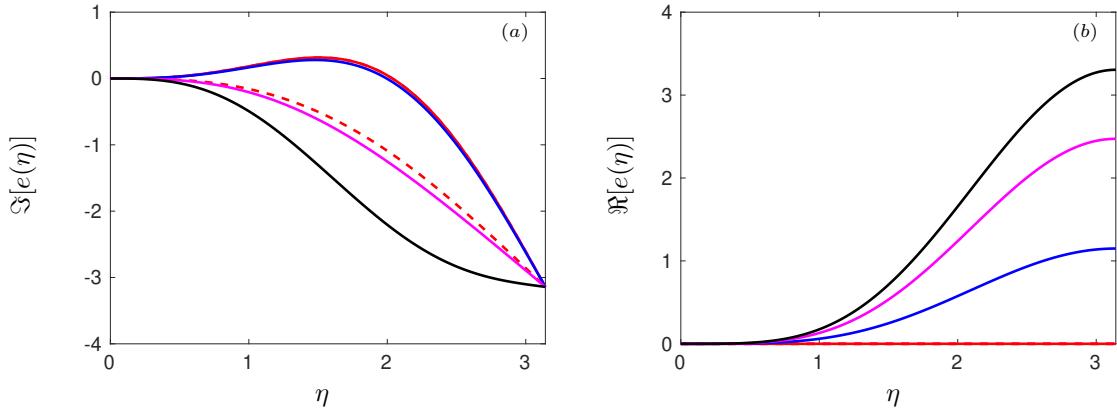


Figure 11: Dispersion error (a) and dissipation error (b) for the first derivative with  $M = 2$  at different values of  $r_c$  and  $r_d$  as marked with the same color in Fig. 10(a). The dashed red line corresponds to standard second order scheme with  $M = 1$  and the dashed magenta line (hidden) corresponds to optimized scheme obtained for  $M = 2$  without stability constraint.

### 2.3.5. Higher-order temporal schemes

In the previous section we illustrated how stability can be incorporated into the framework using a first-order forward temporal discretization. However, in practice higher order schemes are typically used. Thus,

we generalize Eq. (42) to include higher order multi-step temporal schemes, such as Adams-Bashforth. For this we define a vector  $\underline{\mathbf{F}}$  as a stack of the solution at  $L$  consecutive time levels,

$$\underline{\mathbf{F}}^{k+1} = [\mathbf{F}^{k+1} \quad \mathbf{F}^k \quad \dots \quad \mathbf{F}^{k-L+1}]^T. \quad (54)$$

The evolution equation can then be written as

$$\underline{\mathbf{F}}^{k+1} = \left( \mathbf{A}_t + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{B}_t \right) \underline{\mathbf{F}}^k, \quad (55)$$

where

$$\mathbf{A}_t := \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{B}_t := \begin{bmatrix} \mathbf{b}^T \otimes \mathbf{A}_d^\Phi \\ \mathbf{0}_{NL \times N(L+1)} \end{bmatrix}, \quad (56)$$

$$\mathbf{A}_1 := \mathbf{e}_{1 \times (L+1)} \otimes \mathbf{I}_{N \times N} \quad \text{and} \quad \mathbf{A}_2 := [\mathbf{I}_{NL \times NL} \quad \mathbf{0}_{NL \times N}], \quad (57)$$

with  $\mathbf{e} = [1 \ 0 \ \dots \ 0]$ . The vector  $\mathbf{b}^T$  is the vector of the coefficients of the temporal scheme that uses  $L+1$  time levels. By examining Eq. (55) it is clear that multistep methods have the same structure as the example in section 2.3.4 where, for a given  $\Delta t$ , the unknowns are contained in  $\mathbf{B}_t$ . Thus, the stability condition can then be written as

$$\lambda_{max} \left( \mathbf{A}_t + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{B}_t \right) \leq 1 \quad (58)$$

Following the same procedure as before, this condition, and thus stability, is guaranteed by bounding the 2-norm, i.e.

$$\left\| \mathbf{A}_t + \sum_d \frac{\Delta t}{(\Delta x)^d} \beta_d \mathbf{B}_t \right\|_2 \leq 1. \quad (59)$$

which is the generalization for the inequality constraint Eq. (44).

Thus, the unified generalized formulation for arbitrary multi-step temporal discretizations consists of an optimization problem that minimizes the spectral error given by Eq. (31) subjected to the order of accuracy (equality) constraint Eq. (30), and the stability (inequality) constraint Eq. (59).

The two approaches discussed in sections 2.3.3 and 2.3.4, can then be applied to the generalized multi-step schemes as well. That is, we can compute maximum  $\Delta t$  for given  $\mathbf{A}_d^\Phi$  or we can compute  $\mathbf{A}_d^\Phi$  by fixing  $\Delta t$ . We do note, however, that bounding the spectral radius by the 2-norm can be unnecessarily restrictive for some multistep schemes especially for long temporal stencils which leads to long vectors resulting from the stacking of increasingly large number of time levels.

### 3. Numerical results

In order to test the theoretical results from previous sections we conducted several test on model PDEs of increasing complexity. The focus would be in a comparison between the schemes developed here and standard schemes. In particular, we will compare our optimized schemes against

- (a) standard scheme of the same order
- (b) standard scheme with the same stencil size

While (a) allows us to assess how the additional degrees of freedom are used to increase spectral accuracy and/or maintain stability, (b) provides comparison between two schemes with the same computational cost in computing spatial derivatives since both schemes use the same stencil size. As pointed out later on, however, optimal schemes of lower order may indeed provide a computational advantage when they are coupled with a temporal scheme of matching order to solve a PDE.

We begin our analysis with the diffusion equation and the linear advection-diffusion equation for which exact analytical solutions are known and the error in numerical solutions can be evaluated accurately. We then turn to the non-linear advection-diffusion (Burgers) equation which is a widely used proxy to study important features of fluid flow motion governed by the Navier-Stokes equations. We will conclude the numerical section by a brief analysis of the wave equation as well as a discussion of the effect on dispersion relations.

For short, we will refer to the standard and optimized finite difference schemes as SFD and OFD, respectively in what follows.

### 3.1. Diffusion equation

Consider the equation:

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2}, \quad (60)$$

where  $\alpha$  is the diffusivity. Since this equation is linear, different Fourier modes do not interact and wavenumbers present in the solution are only due to them being present in the initial conditions. The dissipative action of the second derivative causes the decay of amplitude of all modes with time. This decay becomes more prominent as the wavenumber increases. Equation Eq. (60) is solved in a periodic domain of length  $L = 2\pi$ . The initial condition is a superimposition of sinusoidal waves,

$$u(x, 0) = \sum_k A(k) \sin(kx + \phi_k), \quad (61)$$

where  $k$  denotes the wavenumber,  $\phi_k$  is a random phase angle corresponding to each wavenumber, and  $A(k)$  is the amplitude of each mode taken here to be represented as a power law of the form  $A(k) = A(1)k^\sigma$ . The value of the exponent was chosen to be  $\sigma = -1/6$  which, by being small, corresponds to a shallow spectrum representative of a solution with a wide range of energetic modes. The reason for this is to critically assess the ability of schemes to represent accurately a wide range of scales.

The analytical solution of Eq. (60) is known:

$$u_a(x, t) = \sum_k e^{-\alpha k^2 t} A(k) \sin(kx + \phi_k) \quad (62)$$

For the semi-discrete analysis, we discretize Eq. (60) using an optimized second order scheme (OFD2) with  $M = 4$  in space whose coefficients can be found in Tables 1 and 2. The numerical results so obtained are compared with standard second order (SFD2) and standard eighth order (SFD8) in space. For the fully discrete system, we match the order of accuracy of time and space discretization. Time and space step sizes ( $\Delta t, \Delta x$ ) are related through a diffusive CFL condition ( $r_d = \alpha \Delta t / \Delta x^2$ ). Thus, we use a forward first order discretization in time for OFD2 and SFD2, and a fourth-order five-stage Runge-Kutta scheme for SFD8.

To assess the error across scales, we compute the relative difference between the energy at individual Fourier modes (at a given wavenumber) of the numerical and analytical solutions, that is  $|\hat{u}(\eta)|^2 - |\hat{u}_a(\eta)|^2|/|\hat{u}_a(\eta)|^2$  where as before  $\eta = k\Delta x$ . In Fig. 12 we show this error for the semi-discrete system (solid lines) at a normalized time of  $t\alpha k_0^2 \approx 0.002$  where  $k_0 = 1$  is the lowest wavenumber in the simulation. This small normalized time was selected to ensure that the solution has evolved enough to present measurable errors while, at the same time, energy in high wavenumbers is not completely dissipated. In terms of the highest wavenumber in the simulation he normalized time is  $t\alpha k_{max}^2 \approx 20$ . The error for SFD2 is larger than OFD2 and grows drastically with increasing wavenumber, illustrating the inability of low-order standard schemes to capture rapid spatial fluctuations. However, this is clearly not the case of optimized second-order schemes. For SFD8, the error is very small at low wavenumbers but also increases significantly at high wavenumbers. In fact, we see that the error spans more than fifteen orders of magnitude showing a dramatic disparity in resolution capabilities for multiscale problems.

The optimized scheme OFD2, on the other hand, shows a more uniform error distribution across the entire wavenumber space shown. In fact, OFD2 presents much smaller errors than even the eighth order SFD8 at high wavenumbers. By fixing the formal order the scheme to two, the additional degrees of freedom

provided are used to increase the resolution capabilities of increasingly large regions of wavenumber space. The fact that OFD2 is more *spectrally flat* than the other schemes stems for the choice we have made for the function  $\gamma(\eta)$  in Eq. (19) as unity in the region of interest  $\eta \in [0, 2.5]$  and zero otherwise. This assigns equal weights to all wavenumbers in that interval and, thus, leads to spectrally flat schemes. Such performance is highly desired when one wants to study multiscale physical processes like turbulence. As we have shown in section 2.2.3, however, the formulation allows for non-uniform weights which can be used to obtain better resolution in one or more arbitrary regions of the wavenumber space.

In the figure we also include the fully-discrete system integrated with a very small time step ( $r_d = 0.0005$ ) as a dashed line. This is seen to be very similar to the semi-discrete case. However, if we increase the time-step, that is, increase  $r_d$ , then the error increases and the time discretization errors may dominate the solution. Note that this is true both for optimized and standard schemes.

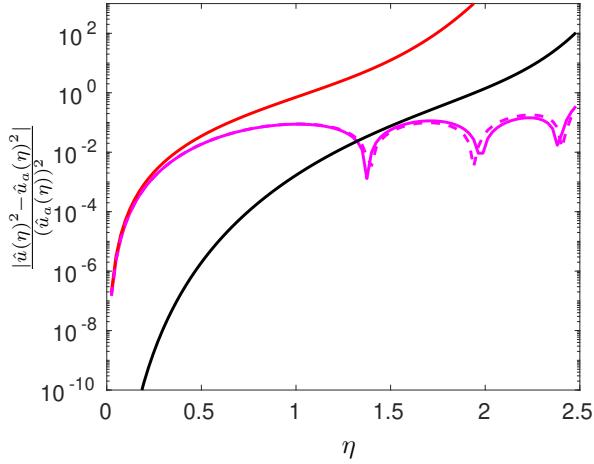


Figure 12: Normalized error in spectral energy for diffusion equation. Solid lines correspond to semi-discrete integration with different numerical schemes: SFD2 (red), SFD8 (black), OFD2 (solid magenta). Dashed magenta line is OFD2 for the fully-discrete system using forward first order in time with  $r_d = 0.0005$ .

### 3.2. Linear advection-diffusion equation

We now consider, next in complexity, the linear advection-diffusion equation which has both the first and second derivatives in space which, as shown in previous sections, typically present different types of errors:

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = \alpha \frac{\partial^2 u}{\partial x^2}, \quad (63)$$

In this case, Fourier modes are convected at the velocity  $c$  and dissipated at a rate determined by the diffusivity  $\alpha$ . Here we use the same initial condition Eq. (61) as in the previous section. The analytical solution for Eq. (63) is given by

$$u_a(x, t) = \sum e^{-\alpha k^2 t} A(k) \sin(k(x - ct) + \phi_k) \quad (64)$$

Because of the presence of both first and second derivatives we expect the numerical solution to be affected by both dispersion and dissipation errors. Dispersion error typically due to the first derivative, distorts phase relations between different waves and tend to create distorted shapes. Dissipation errors, as explained before, affect the amplitude of different waves. In Fig. 13 we show again the relative error of spectral energy which represents dissipation error across the wavenumber space. The trend is similar to what we observed for the diffusion equation.

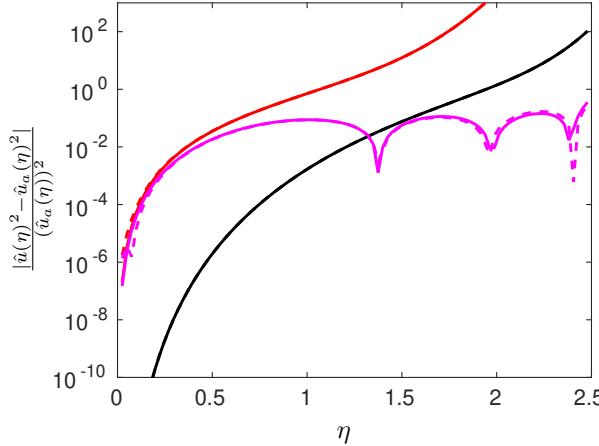


Figure 13: Normalized error in spectral energy for advection-diffusion equation. Different lines correspond to different numerical schemes: standard second order in space (red), standard eighth order in space (black), optimized ( $M=4$ ) second order in space (solid magenta) and optimized ( $M=4$ ) second order in space and forward first order in time (dashed magenta)

It is also possible to quantify dispersion errors by computing the effective propagation speed  $c^*$  of Fourier modes at wavenumber  $k$  which can readily shown to be given by  $c^*(k) = \arg[\hat{u}_t(k)/\hat{u}_0(k)]/kt$ , where  $\hat{u}_0(k)$  and  $\hat{u}_t(k)$  are the Fourier coefficients at the beginning of the simulations and at a time  $t$ , respectively. The ratio of the numerical speed  $c^*$  to the actual propagation speed  $c$  is a measure of phase (dispersive) errors. This ratio is unity if there is no phase error which implies that the numerical solution travels with the same speed as the actual solution. In fig. 14(a) we show results for a normalized time of  $ck_0t = 0.1$ , or in terms of the highest wavenumber in the simulation  $ck_{max}t \approx 10$ . We see that, for SDF2, the numerical wave speed becomes much smaller compared to the actual speed  $c$  as the wavenumber increases, resulting in large phase errors. For OFD2 the numerical speed remains much closer to the actual speed, though with some oscillations. OFD2 is also seen to be better than even the eighth order scheme SFD8 for which, as the wavenumber increases, the numerical speed decreases monotonically leading to very large phase error. The normalization of fig. 14(b) highlights this trend in the phase error. In particular, it shows that, because of the particular choice of  $\gamma(\eta)$ , the numerical scheme presents a more spectrally flat response.

A complementary assessment of the numerical performance of these schemes can be obtained by comparing the time evolution of different Fourier modes. In the fully discrete system, order of accuracy in time is chosen to be consistent with the spatial discretization. With a fixed diffusive CFL, this implies that the second and eight order schemes in space need a first and fourth order temporal discretization, respectively. In the examples below we use, thus, forward Euler (first order) and RK4 (fourth order) for time integration.

For illustration purposes we consider waves at the extreme ends ( $k = 10, 100$ ) of the wavenumber interval over which the schemes have been optimized, and an intermediate wave-number ( $k = 50$ ). These correspond to  $\eta = \{0.245, 1.227, 2.454\}$ . In fig. 15 we show the evolution of energy for these three wavenumbers. The decay in energy occurs at a faster rate as the wavenumber increases. At lower  $\eta$ , the energy decay is well captured by all three schemes. As we increase  $\eta$ , SFD2 over-predicts the energy content and fails to capture the actual dissipation. The disparity between SFD2 and OFD2 increases with increasing wavenumber with the latter remaining close to the analytical solution. For very high wavenumbers  $\eta = 2.454$ , the second-order optimized scheme is in fact visibly closer to the analytical solution than a standard eighth-order scheme. Clearly by using the additional information provided by neighboring points to increase spectral accuracy instead of formal order of accuracy leads to better resolved physics, especially at high wavenumbers where dissipation is strongest.

We have also compared schemes derived with our framework with those in [8, 3]. For example, we have computed fourth-order optimized schemes for  $M = 4$  with the same conditions as those in those references and found essentially the same coefficients. Thus, our general framework can reproduce other particular

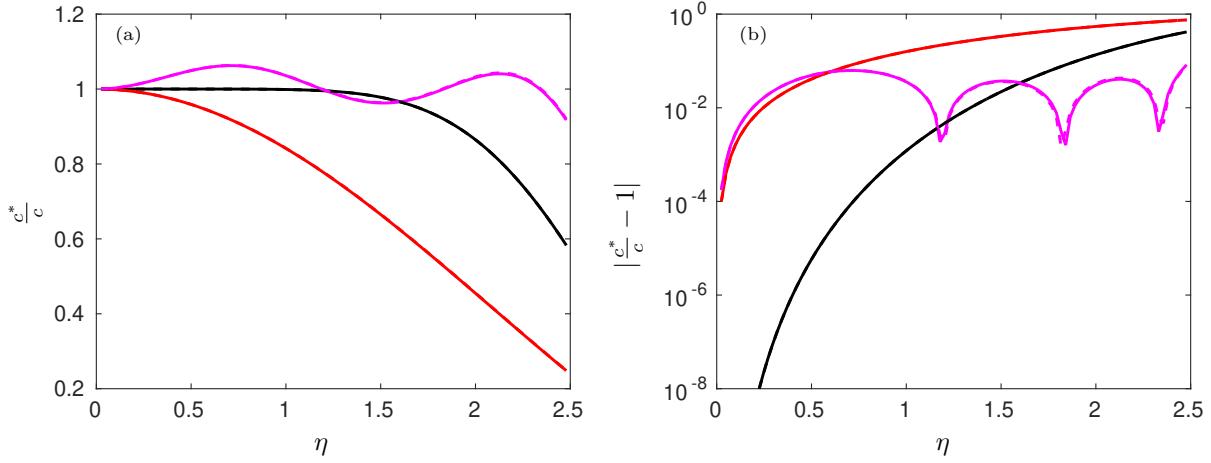


Figure 14: (a) Ratio of numerical to actual speed. (b) Normalized phase error for advection-diffusion equation. Different lines correspond to different numerical schemes: SFD2 (red), SFD8 (black), semi-discrete with OFD2 in space (solid magenta) and OFD2 in space and forward first order in time (dashed magenta).

results in the literature.

### 3.3. Non-linear advection-diffusion equation

Of fundamental and practical interest is the non-linear advection-diffusion equation as it resembles the one-dimensional version of the Navier-Stokes equation that governs the motion of fluid flows,

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \alpha \frac{\partial^2 u}{\partial x^2}. \quad (65)$$

Here  $u(x, t)$  is the velocity and  $\alpha$  is viscosity. The non-linear term causes interaction between Fourier modes which redistributes energy among the different scales (wavenumbers) in the solution and produces new scales of motion. Because of the absence of a constant input of energy, the amplitude of different modes decay with time due to the dissipative action which becomes more effective at smaller scales.

Eq. (65) is solved in a periodic domain of length  $L = 2\pi$ , subject to initial conditions given by Eq. (61). The value of the exponent in Eq. (61) was chosen to be  $\sigma = -5/6$ . This exponent, which corresponds to an energy spectrum decaying as  $k^{-5/3}$  consistent with fully developed turbulence, ensures that the spectrum is shallow enough to ensure high energy content at high wavenumbers while remaining stable.

Although Eq. (65) is non-linear, we can apply the Cole-Hopf transformation and find an analytical solution to the problem [23, 24]. Define a transformation variable  $\phi$ , such that

$$u = -2\alpha \frac{1}{\phi} \frac{\partial \phi}{\partial x}, \quad (66)$$

Then Eq. (65) reduces to a simple diffusion equation in  $\phi$ , which can be readily solved analytically. The result in terms of the primitive variable  $u(x, t)$  is

$$u(x, t) = \frac{\int_{-\infty}^{\infty} \frac{(x-y)\mathcal{I}(y, x)}{t} dy}{\int_{-\infty}^{\infty} \mathcal{I}(y, x) dy}. \quad (67)$$

where  $\mathcal{I}(y, x) = e^{-\frac{(x-y)^2}{4\alpha t}} \phi(y, 0)$ . and the initial condition for  $\phi$ , is computed from the initial condition for  $u$  as

$$\phi(x, 0) = e^{(-\int_0^x \frac{u(y, 0)}{2\alpha} dy)}. \quad (68)$$

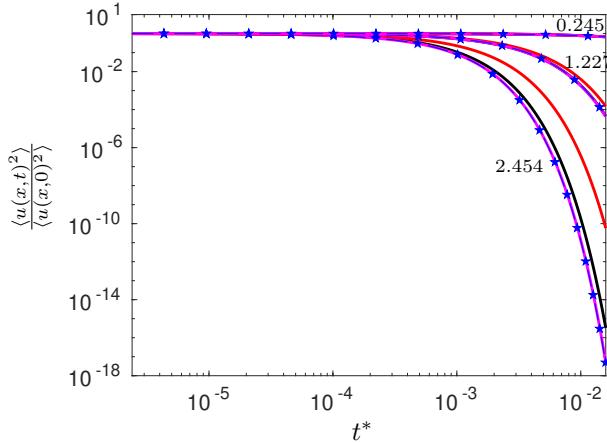


Figure 15: Evolution of space-averaged energy normalized by the initial space-averaged energy with normalized time  $t^* = tc/L$  for three different wave-numbers ( $\eta = 0.245, 1.227, 2.454$ ). Different line styles correspond to different numerical schemes: SFD8 with RK4 in time (black), OFD2 with forward Euler in time (solid magenta), SFD2 with forward Euler in time (red). The blue dashed-star line is the analytical solution.

While Eq. (67) is the exact solution, the integrals involved are computed numerically with standard integral techniques which were tested for grid convergence. This is compared with the numerical solution of Eq. (65) using, as before, SFD2, SFD8, and OFD2. The fully discrete system is formed with compatible temporal scheme as described in previous section. We performed grid convergence studies and found that e.g. the space-averaged kinetic energy evolution in time becomes independent of resolution at  $N = 256$ . This is the resolution used for the comparisons that follow.

In Fig. 16 we show the energy spectrum obtained for these three schemes along with the analytical solution Eq. (67), at  $t/t_0 \approx 0.464$ , where  $t_0 = K_0/\epsilon_0$  is a characteristic time scale defined by the initial energy ( $K_0 \equiv \langle u_0^2 \rangle / 2$ , where angular brackets denote space averages and a subscript 0 denotes initial conditions) and the energy dissipation rate ( $\epsilon_0 \equiv \alpha \langle (\partial u_0 / \partial x)^2 \rangle$ ). We observe that OFD2 and SFD8 agree closely with each other and with the analytical solution throughout the range of  $\eta$ . Results for SFD2, however, exhibit clear departures especially at high wavenumbers. The ability of OFD2 to capture high wavenumbers accurately is expected since, as shown above, this scheme presents a more spectrally flat response. In fact, this optimized scheme presents better resolution than SFD8 at very high wavenumbers though the converse is true at low wavenumbers.

Because of the better small-scale resolution of OFD2, one would expect it to capture quantities that depend sensitively on the small scales more accurately than SFD8 as well. An example of such a quantity is the dissipation rate introduced above which is proportional to the second order moment of the velocity gradient and thus is dominated by high wavenumber activity. This is indeed observed in Fig. 17(a) where we show the evolution of  $\langle \epsilon \rangle$  as the flow decays along with the analytical value (dashed-star line) which can be computed easily by taking derivative of Eq. (67),  $\langle \epsilon \rangle = \alpha \langle (du/dx)^2 \rangle = \alpha \langle \left( \int_{-\infty}^{\infty} \mathcal{I}(y, x) dy \int_{-\infty}^{\infty} \frac{(2\alpha t - (x-y)^2) \mathcal{I}(y, x)}{2\alpha t^2} dy \right. \right. - \left. \left. \int_{-\infty}^{\infty} \frac{(x-y) \mathcal{I}(y, x)}{t} dy \int_{-\infty}^{\infty} \frac{(y-x) \mathcal{I}(y, x)}{2\alpha t} dy \right) \left( \int_{-\infty}^{\infty} \mathcal{I}(y, x) dy \right)^{-2} \rangle$ . We can clearly see that initially OFD2 is very close to the analytical value followed by SFD8 and SFD2. As time evolves, diffusive effects damp high wavenumbers faster than small scales and the main contribution to dissipation moves to lower wavenumbers where the three schemes present similar resolution capabilities. The same conclusion holds for the more challenging higher order moments. This is seen in Fig. 17(b), where we show the fourth-order moment of velocity gradients. Again OFD2 is more accurate than second and eighth order standard schemes.

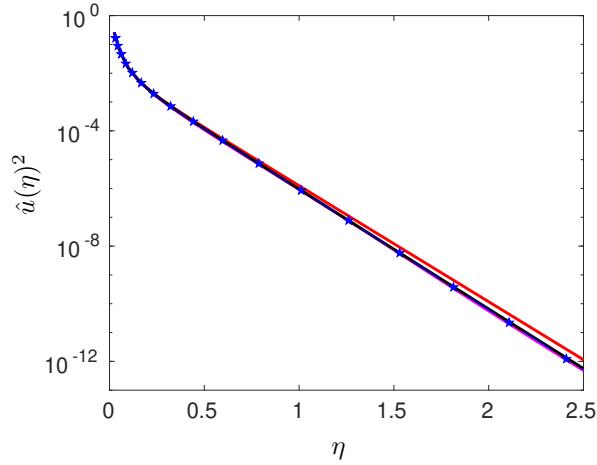


Figure 16: Energy spectrum for the non-linear advection-diffusion equation at  $t^* = t/t_0 \approx 0.464$ . Different lines correspond to different numerical schemes: SFD8 with RK4 in time (black), OFD2 with forward first-order in time (magenta), and SFD2 with forward first-order in time (red). The blue dashed-star line is the analytical solution.

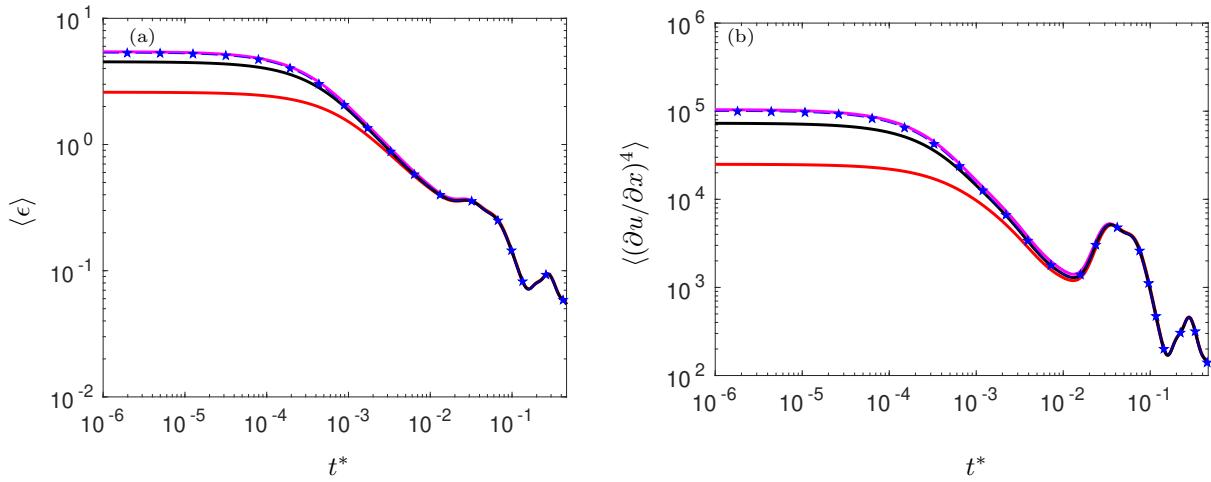


Figure 17: (a) Evolution of space-averaged dissipation (b) Evolution of space-averaged fourth order moment of velocity gradient with normalized time ( $t^* = t/t_0$ ). Different lines correspond to different numerical schemes: SFD8 with RK4 in time (black), OFD2 with forward first-order in time (magenta), and SFD2 with forward first-order in time (red). The blue dashed-star line is the analytical solution.

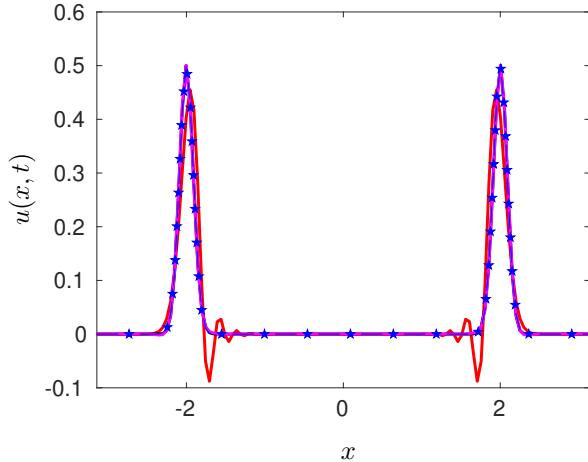


Figure 18: Solution of the wave-equation for  $t = 2$  with a Gaussian initial condition. Different lines correspond to different numerical schemes: SFD8 with RK4 in time (black), OFD2 with RK2 in time (magenta), and SFD2 with RK2 in time (red). The blue dashed-star line is the analytical solution.

### 3.4. Wave Equation

Another equation that has been used to assess the performance of numerical schemes is the second order linear PDE for the description of waves [15], commonly used in acoustics and given by,

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \quad (69)$$

where  $c$  is the propagation speed. The general solution to this equation is a standing wave formed by the superposition of two travelling waves moving with velocity  $c$  to the right and  $-c$  to the left respectively. Following [15], this equation can be expressed as the following system of equations,

$$\begin{pmatrix} \partial u / \partial t \\ \partial v / \partial t \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ c^2 \partial^2 / \partial x^2 & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}. \quad (70)$$

We solved this system for a Gaussian initial condition,  $u_0(x) = 0.2e^{-64x^2}$ , that has a narrow width in the physical domain and therefore the spectrum spans a wide range of wavenumbers in Fourier space. We use the optimized second-order OFD2 for the second derivative in space computed with a weight function,  $\gamma(w) = e^{-w^2/256}$  for  $w = [0, 2]$  and  $\gamma(w) = 0$  elsewhere in Eq. (21). This was done to emphasize wavenumbers relevant to the problem. Because of the use of a *CFL* condition given by  $r_c = c\Delta t/\Delta x = 0.1$ , we also employ second-order temporal discretization (RK2). The solution was advanced until a physical time of  $t = 2$  for periodic boundary conditions and the results have been plotted in Fig. 18. We can see that the SFD2 scheme has already developed significant oscillations that trigger instabilities, whereas the OFD2 is comparable to both SFD8 and the exact solution. The space averaged 2-norm of the error for OFD2 is an order of magnitude smaller than the error for SFD2. The 2-norm of the error for OFD2 scheme is comparable to the error obtained for SFD8. Besides this, the OFD2 scheme is three times computationally less expensive as compared to the SFD8.

### 3.5. Space-time errors

In addition to spectral accuracy, other performance metrics have been utilized in the literature to assess the appropriateness of a numerical scheme to reproduce physics of interest. For strongly convective problems, such as in acoustics, the dispersion relation provides important information about propagation speeds and characteristics. Thus, the interest in so-called dispersion-relation preserving schemes [3].

Consider again the convection-diffusion equation:

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = \alpha \frac{\partial^2 u}{\partial x^2}, \quad c, \alpha > 0, \quad (71)$$

This linear equation propagates the initial condition to the right at the speed  $c$  and because of the diffusive term the amplitude decreases with time at a rate determined by the diffusivity coefficient  $\alpha$ . Using  $u = \hat{u}e^{-j\omega t}e^{j k x}$  it is easy to obtain the exact dispersion relation for this equation,  $\omega = ck - j\alpha k^2$ , where  $k$  is the wavenumber. Two physically meaningful quantities related to the dispersion relation are the phase and group velocities ( $c_p$  and  $c_g$  respectively) which are given by the real part of  $c_p = \omega/k$  and  $c_g = d\omega/dk$ . Note that both are equal to  $c$  for the analytical solution. The equivalent numerical dispersion relation is, on the other hand, given by  $\omega^* = c^*k - j\alpha^*k$ , where  $c^*$  and  $\alpha^*$  are the numerical velocity and diffusivity respectively. One is thus interested in how the phase and group velocities from the computed solution ( $c_p^*$  and  $c_g^*$ ) compares to their exact values.

A convenient way to obtain dispersion relations and derived quantities is through the so-called amplification factor  $G^*$  defined as the ratio of Fourier modes of the numerical solution at two consecutive time steps. In general, one can write  $G^* = |G^*|e^{-j\beta}$ , where  $|G^*|$  is the magnitude of the amplification factor and the phase  $\beta$  is related to the numerical phase speed ( $c_p^*$ ). Note that the exact amplification factor is given by  $G = e^{-i\omega\Delta t}$ . It is then readily shown that the numerical group velocity can be written as [10],

$$\frac{c_g^*}{c_g} = \frac{1}{r_c \Delta x} \frac{d\beta}{dk}. \quad (72)$$

where, as noted above,  $c_g = c$ . In a dispersion-relation preserving scheme, this ratio as well as  $c_p^*/c_p$  should be close to unity.

As an illustration, consider Eq. (71) discretized with a forward difference in time and the spatial schemes in Table 1 and Table 2:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{-c}{\Delta x} \sum_{m=-M}^M a_{m,1} u_{i+m} + \frac{\alpha}{\Delta x^2} \sum_{m=-M}^M a_{m,2} u_{i+m}. \quad (73)$$

Since  $c_g^*/c_g$  and  $c_p^*/c_p$  depend on both  $k\Delta x$  and  $\omega\Delta t$ , one can assess performance by measuring the area in the  $k\Delta x - \omega\Delta t$  plane in which these ratios are within some percentage of unity, say 5% [16]. The larger this area, the larger the range of wavenumbers that preserve the dispersion relation. And since the phase errors are more prominent in convection dominated problems, we will consider relatively small values of  $r_d$ . From table 3, we can see that when the optimized schemes derived above are coupled with forward difference in time, the region in the  $k\Delta x - \omega\Delta t$  for which  $c_g^*/c_g$  is close to unity, increases considerably for different values of  $r_d$ . For  $M = 4$ , this area is five times larger than that for the standard second order scheme. We can also see that the ratio  $c_p^*/c_p$  is close to unity for a much larger area for optimized schemes. This is in fact not unexpected as the objective function in the minimization problem is designed to bring the modified wavenumber close to the actual wavenumber, resulting in more accurate derivatives leading, in general, to lower errors in phase and group velocities. Similar results were also observed for the diffusivity ratio  $\alpha^*/\alpha$ .

We close this section by noting that although optimized schemes seem to naturally preserve dispersion relations better, constraints on these dispersion relations can be directly included in the general framework proposed here. For example, one can construct objective functions as in [3] where one can emphasize either propagation characteristics or damping characteristics. In the present framework, this can be done in such a way to ensure, at the same time, stability of the resulting scheme. Some of the effects of such an implementation were discussed in section 2.2.1.

#### 4. Conclusions and final remarks

Standard finite difference schemes are commonly derived to maximize its formal order of accuracy for a given stencil. The spectral accuracy and stability of the schemes so obtained are typically checked *a*

$M$	$r_d = 0$		$r_d = 0.02$	
	$c_g^*/c_g$	$c_p^*/c_p$	$c_g^*/c_g$	$c_p^*/c_p$
1(S)	0.58	1.73	0.62	1.84
2(O)	2.29	6.87	2.27	6.89
3(O)	1.82	5.13	1.87	5.34
4(O)	2.91	10.09	3.04	10.59

Table 3: Percentage area in the  $k\Delta x - \omega\Delta t$  plane for which  $c_g^*/c_g$  and  $c_p^*/c_p$  is within 5% of unity for  $r_d = 0$  and  $r_d = 0.02$ . In parenthesis, O stands for optimized scheme, and S for standard scheme. The schemes used here are those in Table 1 and Table 2.

*posteriori*. While efforts have been devoted to merge order-of-accuracy constraints and spectral accuracy, no general mathematical framework has been put forth which, perhaps more importantly, introduced stability as an additional constraint. Here we develop such a framework to derive finite differences that accounts for order of accuracy, spectral resolution and stability. The most general formulation is given by Eq. (52).

While order of accuracy is defined as the power of the lowest order term in the truncation error, spectral accuracy is defined through an objective function that minimizes the spectral error in some specific way. This definition of this error is rather general and includes a weighting function  $\gamma(\eta)$  which can be used to emphasize different scales relevant to the physical problem being solved. The fusion of order-of-accuracy constraints and spectral accuracy leads to a minimization problem which is convex and thus leads to a global minimum. The optimal coefficients, which are obtained analytically from the minimization problem, were shown to be symmetric for the even derivatives and anti-symmetric for the odd derivatives. This was shown to correspond to vanishing imaginary and real part of the error  $e(\eta)$ . This is the case, regardless of the functional form of the weighting function  $\gamma(\eta)$ . In other words, we have shown that the minimization of spectral errors leads to symmetric or anti-symmetric stencils for even and odd derivatives respectively, regardless of which range of wavenumber is optimized.

We have also incorporated stability into the unified framework. The semi-discrete system is shown to depend upon the sign of the parameters  $\beta_d$  for even  $d$  in the PDE. For the fully-discrete system, stability is assured by requiring the spectral radius of the evolution matrix to be smaller than unity which can in turn be bounded by its 2-norm. The latter can be written as a linear matrix inequality. This provides an additional constraint imposed on the minimization problem to ensure that the resulting scheme is stable. The final unified mathematical framework consists on the minimization of an objective function representative of spectral error constrained by given order of accuracy and stability guarantees. Due to the non-linearity inherent in this constraint, this problem is no longer convex and therefore cannot be solved directly using standard optimization tools. Two approaches were proposed to tackle this non-linear optimization.

In the first approach, given an optimal scheme (subjected to constraints in order and spectral accuracy) one finds the largest  $\Delta t$  for which the scheme remains stable. This approach is similar to standard practices in which given a scheme, one aims at finding the largest time step that assures stability. Here, however, the largest time step results from a convex optimization which gives global extrema and can be solved efficiently. Through an example, we showed that the stability region decreased with increasing stencil size  $M$  consistent with a reduction of (artificial) dissipation at higher wavenumbers. This approach illustrates how accuracy and stability are separate requirements in the formulation: the additional degrees of freedom available for longer stencils are used to maximize spectral resolution regardless of stability. The result is a reduced stability region in the  $r_d$ - $r_c$  space. In the second approach, the three elements are combined: optimal coefficients are obtained with both order of accuracy and stability constraints for a given time-step  $\Delta t$ . This is a common situation when the time step is set by physical considerations (e.g. shortest time scale in the problem). In this approach, the additional information provided by neighboring grid points is used to extend stability which make the use of much larger time-step feasible with spectral error comparable to the standard schemes. We showed that explicit schemes both in time and space, can remain stable for very large time steps. This can provide significant advantages for massively parallel simulations for which implicit schemes become increasingly challenging at large processor counts.

Several numerical results were presented to illustrate the numerical performance against standard finite differences of different orders. In particular, we compared optimized schemes against standard schemes of the same order and the same stencil size. While the latter presents the same computational cost in terms of spatial derivatives, the fully discrete system may be less expensive if the order of the temporal discretization is to be compatible with spatial order. Introducing the effect of temporal discretization in the unified framework presented here is part of our own ongoing research. Another application of the framework presented here is the inclusion of more general boundary conditions. Although we have limited our results to periodic domains, the framework can straightforwardly be extended to solve problems with non-periodic boundary conditions. The overall minimization problem and the constraints remain the same but the structure of some matrices has to be changed to restrict the stencil used close to boundaries. In particular, the first few and last elements of  $\mathbf{A}_d$  in Eq. (29) will be biased so that only grid points to the right and left of the boundary are used respectively. The size of each stencil as well as the order or accuracy can be set individually through the corresponding entries in  $\mathbf{X}_d$  in Eq. (30). As commonly done in simulations of complex flows, a progressive reduction of order of accuracy close to the boundary can thus be easily accommodated in this formulation. The stability constraint is identical to that presented for periodic domains. Thus, here too Eq. (52) will yield spectrally optimal schemes for a given  $\Delta t$  and Eq. (49) will yield the maximum  $\Delta t$  for which a spectrally optimal scheme remains stable.

We finally make some remarks about further potential generalizations of the framework. The stability constraint used here ensure that solutions are non-growing and are applied to the entire PDE which may involve terms of different characteristics (convection, diffusion, etc). While this type of constraint is standard practice, it seems possible to extend the framework to other approaches that can capture other details of the error dynamics [25]. This is part of future work. Furthermore, our focus was on spatial discretization with a given temporal scheme which could be of arbitrary order. Because of the potential additional effects when coupling space-time operators, it is a natural next step to optimize spatial and temporal schemes simultaneously. Unfortunately, this coupled optimization problem yields non-linear objective functions [15] which makes the mathematical problem much more challenging. This is also part of our ongoing research and will be discussed in detail elsewhere.

In summary, we integrated order of accuracy, spectral resolution and stability in the derivation of finite differences in a unified framework. We have shown specific properties of the resulting schemes in terms of the kind of error expected. The coupling of these three critical elements in a unified formulation allows one to decouple requirements in terms of e.g. order of accuracy and spectral accuracy. This coupling also manifests itself in the trading of accuracy with stability. We showed, for example, how spectrally optimal finite differences bias odd order derivatives to maintain stability at the expense of accuracy. Other tradeoffs can be analyzed similarly within the framework presented here.

## 5. Acknowledgments

The authors gratefully acknowledge NSF (Grant OCI-1054966 and CCF-1439145) for financial support. The authors also thank XSEDE for computer time on their systems.

## Appendix A. Imaginary component of the optimal spectral error $e(\eta)$ is zero for even derivatives

*Proof.* The spectral error for even derivative is

$$e(\eta) = (\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d) + j\mathbf{S}(\eta)^T\mathbf{a}_d.$$

Therefore,

$$\|e(\eta)\|_{\mathcal{L}_2}^2 = \int_0^\pi \gamma(\eta) \left[ (\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d)^2 + (\mathbf{S}(\eta)^T\mathbf{a}_d)^2 \right] d\eta,$$

or

$$\min_{\mathbf{a}_d} \|e(\eta)\|_{\mathcal{L}_2}^2 = \min_{\mathbf{a}_d} \int_0^\pi \gamma(\eta) \left[ (\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q \eta^d)^2 \right] d\eta + \min_{\mathbf{a}_d} \int_0^\pi \gamma(\eta) (\mathbf{S}(\eta)^T \mathbf{a}_d)^2 d\eta.$$

Since the cost function is sum of squares, it is minimized if and only if individual terms are minimized.

Without loss of generality, we can write  $\mathbf{a}_d := \mathbf{a}_d^s + \mathbf{a}_d^{as}$ , where  $\mathbf{a}_d^s$  is symmetrical about the central element, and  $\mathbf{a}_d^{as}$  is anti-symmetric about central element. Therefore,

$$\begin{aligned}\mathbf{S}^T(\eta)\mathbf{a}_d &= \mathbf{S}^T(\eta)(\mathbf{a}_d^s + \mathbf{a}_d^{as}) = \mathbf{S}^T(\eta)\mathbf{a}_d^{as}, \\ \mathbf{C}^T(\eta)\mathbf{a}_d &= \mathbf{C}^T(\eta)(\mathbf{a}_d^s + \mathbf{a}_d^{as}) = \mathbf{C}^T(\eta)\mathbf{a}_d^s,\end{aligned}$$

since  $\mathbf{S}^T(\eta)\mathbf{a}_d^s = 0$  for symmetric coefficients and  $\mathbf{C}^T(\eta)\mathbf{a}_d^{as} = 0$  for anti-symmetric coefficients.

Consequently,

$$\begin{aligned}&\min_{\mathbf{a}_d} \int_0^\pi \gamma(\eta) \left[ (\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q \eta^d)^2 \right] d\eta + \min_{\mathbf{a}_d} \int_0^\pi \gamma(\eta) (\mathbf{S}(\eta)^T \mathbf{a}_d)^2 d\eta, \\ &= \min_{\mathbf{a}_d^s} \int_0^\pi \gamma(\eta) \left[ (\mathbf{C}^T(\eta)\mathbf{a}_d^s - (-1)^q \eta^d)^2 \right] d\eta + \min_{\mathbf{a}_d^{as}} \int_0^\pi \gamma(\eta) (\mathbf{S}(\eta)^T \mathbf{a}_d^{as})^2 d\eta.\end{aligned}$$

Therefore, the two optimizations are independent of each other. For a positive real valued function  $\gamma(\eta)$ , the second term is zero if and only if  $\mathbf{a}_d^{as} = 0$ . Consequently, for  $\mathcal{L}_2$  optimal spectral errors, the imaginary part of the spectral error is zero for even derivatives, and the optimal  $\mathbf{a}_d$  is symmetrical about central element.

We next analyze the feasibility of the order constraint with  $\mathbf{a}_d^{as} = 0$ . Let  $\mathbf{T}_s := (\mathbf{I} + \mathbf{J})/2$ , and  $\mathbf{T}_{as} := (\mathbf{I} - \mathbf{J})/2$ , where  $\mathbf{J}$  represents anti-diagonal matrix. With these transformation matrices, we can write  $\mathbf{a}_d^s := \mathbf{T}_s \mathbf{a}_d$  and  $\mathbf{a}_d^{as} := \mathbf{T}_{as} \mathbf{a}_d$ . Therefore,

$$\mathbf{a}_d^T \mathbf{X}_d = (\mathbf{a}_d^s + \mathbf{a}_d^{as})^T \mathbf{X}_d = (\mathbf{a}_d^s)^T \mathbf{X}_d = \mathbf{a}_d^T \mathbf{T}_s^T \mathbf{X},$$

and the order accuracy constraint can be written as

$$\mathbf{a}_d^T \mathbf{T}_s^T \mathbf{X} = \mathbf{y}_d.$$

We observe that the structure of  $\mathbf{X}_d$  is such that the odd columns are symmetric and the even columns are anti-symmetric, about the central element. Therefore, for the even columns,  $\mathbf{T}_s^T(\mathbf{X}_d)_i = 0$ . Noting that the even columns of  $\mathbf{y}_d$  are zero, we can conclude that the constraints corresponding to the even columns are trivially satisfied for symmetric coefficients. For the odd columns of  $\mathbf{X}_d$ , we observe that  $\mathbf{T}_s^T(\mathbf{X}_d)_i = (\mathbf{X}_d)_i$ . That is, the constraints corresponding to the odd columns are unaffected. Consequently, if  $\mathbf{a}_d^T \mathbf{X}_d = \mathbf{y}_d$  is feasible, then  $\mathbf{a}_d^T \mathbf{T}_s^T \mathbf{X} = \mathbf{y}_d$  is also feasible.  $\square$

## Appendix B. Real component of the optimal spectral error $e(\eta)$ is zero for odd derivatives

*Proof.* The proof follows similarly to above. The spectral error in this case is

$$e(\eta) := \mathbf{C}^T(\eta)\mathbf{a}_d + j(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q \eta^d).$$

Using the same decomposition for  $\mathbf{a}_d$  as above, we get

$$\begin{aligned}\min_{\mathbf{a}_d} \|e(\eta)\|_{\mathcal{L}_2}^2 &= \min_{\mathbf{a}_d} \int_0^\pi \gamma(\eta) (\mathbf{C}^T(\eta)\mathbf{a}_d)^2 d\eta + \min_{\mathbf{a}_d} \int_0^\pi \gamma(\eta) [\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q \eta^d]^2 d\eta, \\ &\min_{\mathbf{a}_d^s} \int_0^\pi \gamma(\eta) (\mathbf{C}^T(\eta)\mathbf{a}_d^s)^2 d\eta + \min_{\mathbf{a}_d^{as}} \int_0^\pi \gamma(\eta) [\mathbf{S}^T(\eta)\mathbf{a}_d^{as} - (-1)^q \eta^d]^2 d\eta.\end{aligned}$$

Using similar arguments as above, the optimal solution will guarantee  $\mathbf{a}_d^s = 0$  and consequently, the real part of the spectral error is zero for odd derivatives. With  $\mathbf{a}_d^s = 0$ , the optimal  $\mathbf{a}_d$  will be anti-symmetric about its central element.

The proof for feasibility of the accuracy order constraint, with  $\mathbf{a}_d^s = 0$ , is similar to the feasibility proof for the even derivative.  $\square$

### Appendix C. $\mathbf{A}_d$ Invariance of grid point location

*Proof.* We know that the inverse of a partitioned matrix can be written as

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix}. \quad (\text{C.1})$$

Using this result, let

$$\begin{bmatrix} \mathbf{Q}_d & \mathbf{X}_d \\ \mathbf{X}_d^T & \mathbf{0} \end{bmatrix}^{-1} := \begin{bmatrix} \mathbf{M}_1 & \mathbf{M}_2 \\ \mathbf{M}_3 & \mathbf{M}_4 \end{bmatrix}, \quad (\text{C.2})$$

for suitably defined  $\mathbf{M}_i$  using Eq. (C.1). Using the kronecker product result

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}, \quad (\text{C.3})$$

and Eq. (C.1), we can write

$$\begin{bmatrix} \mathbf{I}_N \otimes \mathbf{Q}_d & \mathbf{I}_N \otimes \mathbf{X}_d \\ \mathbf{I}_N \otimes \mathbf{X}_d^T & \mathbf{I}_N \otimes \mathbf{0} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I}_N \otimes \mathbf{M}_1 & \mathbf{I}_N \otimes \mathbf{M}_2 \\ \mathbf{I}_N \otimes \mathbf{M}_3 & \mathbf{I}_N \otimes \mathbf{M}_4 \end{bmatrix}. \quad (\text{C.4})$$

Therefore, the optimal solution

$$\begin{aligned} \left( \begin{bmatrix} \mathbf{v}_d \\ \boldsymbol{\lambda}_d \end{bmatrix} \right)^* &= \left( \begin{array}{l} (\mathbf{I}_N \otimes \mathbf{M}_1)(\mathbf{1}_{N \times 1} \otimes \mathbf{r}_d) + (\mathbf{I}_N \otimes \mathbf{M}_2)(\mathbf{1}_{N \times 1} \otimes \mathbf{y}_d^T) \\ (\mathbf{I}_N \otimes \mathbf{M}_3)(\mathbf{1}_{N \times 1} \otimes \mathbf{r}_d) + (\mathbf{I}_N \otimes \mathbf{M}_4)(\mathbf{1}_{N \times 1} \otimes \mathbf{y}_d^T) \end{array} \right), \\ &= \begin{pmatrix} \mathbf{1}_{N \times 1} \otimes (\mathbf{M}_1 \mathbf{r}_d + \mathbf{M}_2 \mathbf{y}_d^T) \\ \mathbf{1}_{N \times 1} \otimes (\mathbf{M}_3 \mathbf{r}_d + \mathbf{M}_4 \mathbf{y}_d^T) \end{pmatrix}, \\ &= \begin{pmatrix} \mathbf{1}_{N \times 1} \otimes \mathbf{a}_d^* \\ \mathbf{1}_{N \times 1} \otimes \boldsymbol{\lambda}_d^* \end{pmatrix}, \end{aligned}$$

where  $(\mathbf{a}_d^*, \boldsymbol{\lambda}_d^*)$  is the solution of Eq. (25). Therefore, the optimal solution is identical for all grid points.  $\square$

### Appendix D. Construction of coefficient matrix $\mathbf{A}_d^\Phi$

*Proof.* Define a shift operator  $\Phi_k$ , which is an  $N \times N$  matrix, with elements

$$\Phi_{k_{ij}} := \delta((i - j - k) \bmod N), \quad (\text{D.1})$$

where  $\delta(\cdot)$  is the Kronecker delta function defined as

$$\delta(i) = \begin{cases} 0 & \text{if } i \neq 0, \\ 1 & \text{if } i = 0. \end{cases}$$

For a column vectors, the operator  $\Phi_k$  cyclically shifts the elements down,  $k$  times. For example, for

$$\mathbf{v} := \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \quad \Phi_1 \mathbf{v} = \begin{pmatrix} 4 \\ 1 \\ 2 \\ 3 \end{pmatrix},$$

where

$$\Phi_1 := \delta((i - j - 1) \bmod 4) = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

For a row vector, the operator cyclically shifts the elements left,  $k$  times. That is,

$$\mathbf{v}^T := (1 \ 2 \ 3 \ 4), \mathbf{v}^T \Phi_1 = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 1 \end{pmatrix},$$

From the definition of vector  $\mathbf{F}$  and  $\mathbf{F}^{(d)}$

$$\mathbf{F} := \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix}, \text{ and } \mathbf{F}^{(d)} := \begin{pmatrix} f_1^{(d)} \\ \vdots \\ f_N^{(d)} \end{pmatrix}. \quad (\text{D.2})$$

we can write the finite difference approximation at the  $i$ -th grid point as

$$f_i^{(d)} = \frac{1}{(\Delta x)^d} \mathbf{a}_{i,d}^T \mathbf{T} \Phi_{M_{\max}+i-1} \mathbf{F},$$

for  $i = \{1, \dots, N\}$ , and  $\mathbf{T} \in \mathbb{R}^{S \times N}$  is a transformation matrix defined by

$$\mathbf{T} := [\mathbf{0}_{S \times (M_{\max}-M)} \ \ \mathbf{I}_S \ \ \mathbf{0}_{S \times (M_{\max}-M)}]. \quad (\text{D.3})$$

The matrix  $\mathbf{T} \Phi_{M_{\max}+i-1}$  is a linear operator, or simply a mask, that picks the correct elements from  $\mathbf{F}$  in determining the derivative at the  $i^{\text{th}}$  location.

Now, let  $\mathbf{A}_d^\Phi$  be the vertical stacking of  $\mathbf{a}_{i,d}^T \mathbf{T} \Phi_{M_{\max}+i}$ , for  $i = \{1, \dots, N\}$ , i.e.

$$\mathbf{A}_d^\Phi := \begin{bmatrix} \mathbf{a}_{1,d}^T \mathbf{T} \Phi_{M_{\max}} \\ \vdots \\ \mathbf{a}_{N,d}^T \mathbf{T} \Phi_{M_{\max}+N-1} \end{bmatrix} := \sum_i^N \boldsymbol{\delta}_i \boldsymbol{\delta}_i^T \mathbf{A}_d \mathbf{T} \Phi_{M_{\max}+i-1}, \quad (\text{D.4})$$

where  $\boldsymbol{\delta}_i \in \mathbb{R}^n$  is a vector whose  $k^{\text{th}}$  element is defined by  $\delta(i-k)$ , i.e. the  $i^{\text{th}}$  element of  $\boldsymbol{\delta}_i \in \mathbb{R}^n$  is equal to one and the rest are zero. The vector  $\boldsymbol{\delta}_i$  in Eq. (D.5) is defined for  $n = N$ .

The definition of  $\mathbf{A}_d^\Phi$  can be compactly written as

$$\mathbf{A}_d^\Phi = \mathbf{M}_1 (\mathbf{I}_N \otimes \mathbf{A}_d) \mathbf{M}_2, \quad (\text{D.5})$$

where

$$\mathbf{M}_1 := [\boldsymbol{\delta}_1 \boldsymbol{\delta}_1^T \ \ \cdots \ \ \boldsymbol{\delta}_N \boldsymbol{\delta}_N^T], \quad \mathbf{M}_2 := \begin{bmatrix} \mathbf{T} \Phi_{M_{\max}} \\ \vdots \\ \mathbf{T} \Phi_{M_{\max}+N-1} \end{bmatrix}. \quad (\text{D.6})$$

Equation Eq. (D.5) shows that  $\mathbf{A}_d^\Phi$  is linear in  $\mathbf{A}_d$ .

Thus, the finite-difference approximation for the  $d^{\text{th}}$  derivative for all the grid points is

$$\mathbf{F}^{(d)} = \frac{1}{(\Delta x)^d} \mathbf{A}_d^\Phi \mathbf{F}. \quad (\text{D.7})$$

□

## References

- [1] C. Hirsch, Numerical computation of internal and external flows, volume 1, Wiley, New York, 1994.
- [2] S. K. Lele, Compact finite-difference schemes with spectral-like resolution, *J. Comp. Phys.* 103 (1992) 16–42.
- [3] C. K. Tam, J. C. Webb, Dispersion-relation-preserving finite difference schemes for computational acoustics, *J. Comp. Phys.* 107 (1993) 262 – 281.
- [4] J. W. Kim, D. J. Lee, Optimized compact finite difference schemes with maximum resolution, *AIAA Journal* 34 (1996) 887–893.
- [5] Z. Wang, R. Chen, Optimized weighted essentially nonoscillatory schemes for linear waves with discontinuity, *J. Comp. Phys.* 174 (2001) 381 – 404.
- [6] M. Zhuang, R. F. Chen, Applications of high-order optimized upwind schemes for computational aeroacoustics, *AIAA Journal* 40 (2002) 443–449.
- [7] D. Ponziani, S. Pirozzoli, F. Grasso, Development of optimized weighted-ENO schemes for multiscale compressible flows, *Int. J. Numer. Meth. Fluids* 42 (2003) 953–977.
- [8] C. Bogey, C. Bailly, A family of low dispersive and low dissipative explicit schemes for flow and noise computations, *J. Comp. Phys.* 194 (2004) 194 – 214.
- [9] M. P. Martin, E. M. Taylor, M. Wu, V. G. Weirs, A bandwidth-optimized WENO scheme for the effective direct numerical simulation of compressible turbulence, *J. Comp. Phys.* 220 (2006) 270–289.
- [10] T. K. Sengupta, S. K. Sircar, A. Dipankar, High accuracy schemes for DNS and acoustics, *Journal of Scientific Computing* 26 (2006) 151–193.
- [11] J. Fang, Z. Li, L. Lu, An optimized low-dissipation monotonicity-preserving scheme for numerical simulations of high-speed turbulent flows, *J. Sci. Comput.* 56 (2013) 67–95.
- [12] J.-H. Zhang, Z.-X. Yao, Optimized explicit finite-difference schemes for spatial derivatives using maximum norm, *J. Comp. Phys.* 250 (2013) 511 – 526.
- [13] S. Pirozzoli, Performance analysis and optimization of finite-difference schemes for wave propagation problems, *J. Comp. Phys.* 222 (2007) 809–831.
- [14] G. Ashcroft, X. Zhang, Optimized prefactored compact schemes, *J. Comp. Phys.* 190 (2003) 459 – 477.
- [15] Z. Haras, S. Ta’asan, Finite difference schemes for long-time integration, *J. Comp. Phys.* 114 (1994) 265 – 279.
- [16] T. Sengupta, G. Ganeriwal, S. De, Analysis of central and upwind compact schemes, *J. Comp. Phys.* 192 (2003) 677 – 694.
- [17] T. K. Sengupta, A. Bhole, Error dynamics of diffusion equation: Effects of numerical diffusion and dispersive diffusion, *Journal of Computational Physics* 266 (2014) 240 – 251.
- [18] D. P. Bertsekas, Nonlinear programming, Athena scientific Belmont, 1999.
- [19] M. Grant, S. Boyd, Y. Ye, CVX: Matlab software for disciplined convex programming, 2008.
- [20] A. Ben-Tal, A. Nemirovski, Lectures on modern convex optimization: analysis, algorithms, and engineering applications, SIAM, 2001.

- [21] J. Stoer, R. Bulirsch, Introduction to numerical analysis, volume 12, Springer Science & Business Media, 2013.
- [22] R. Y. Zhang, J. Lavaei, Efficient algorithm for large-and-sparse LMI feasibility problems, IEEE CDC (2018) 6868–6875.
- [23] H. Eberhard, The partial differential equation  $u_t + uu_x = \mu u_{xx}$ , Communications on Pure and Applied Mathematics 3 (1950) 201–230.
- [24] J. D. Cole, On a quasi-linear parabolic equation occurring in aerodynamics, Quarterly of Applied Mathematics 9 (1951) 225–236.
- [25] Y. Shokin, K. Roesner, The Method of Differential Approximation, Scientific Computation, Springer Berlin Heidelberg, 2011.