A unified framework to generate optimized compact finite difference schemes

Vedang M. Deshpande[*] Raktim Bhattacharya[†] Diego A. Donzis[‡]

Aerospace Engineering, Texas A&M University
College Station, TX 77843-3141, USA.

Abstract

A unified framework to derive optimized compact schemes for a uniform grid is presented. The optimized scheme coefficients are determined analytically by solving an optimization problem to minimize the spectral error subject to equality constraints that ensure specified order of accuracy. A rigorous stability analysis for the optimized schemes is also presented. We also show that other types of schemes e.g., spatially explicit and biased finite differences, can be generated as special cases of the framework. Optimized schemes generated using this framework are tested on canonical partial differential equations, and numerical results are compared with the standard schemes.

Keywords: compact finite differences, optimized schemes, spectral error minimization, partial differential equations

1 Introduction

Finite differences have been extensively studied and used to solve ordinary differential equations (ODEs) and partial differential equations (PDEs) numerically. Finite differences, in a broad sense, are classified into two categories, namely, explicit and implicit (compact) finite differences. In the explicit formulation, approximation of a function derivative at a grid point depends only on the function values at that grid point and grid points adjacent to it. On the other hand, the implicit formulation approximates a function derivative at a grid point by using not only the function values but also the derivatives of the function at adjacent grid points. Each one of these two formulations has benefits associated with it. For example, explicit schemes are computationally more efficient than compact schemes. However, for a given stencil size, compact schemes can achieve better accuracy and stability than explicit schemes. Thus, depending upon the problem at hand, one can choose to use explicit or compact formulation.

Kumari, et al. (2019) proposed a unified approach to derive optimized explicit schemes in [1]. In this paper, we present a unified framework to derive optimized compact schemes as a natural extension and generalization of the approach presented in [1]. We also show that the optimized schemes derived using the explicit formulation can be recovered as special cases of the compact formulation presented in this paper.

One of the widely used formulations for deriving compact schemes is the 7 grid points stencil which leads to pentadiagonal finite differences [2–7]. Such formulations approximate the first derivative at i^{th} grid point as

$$\beta f'_{i-2} + \alpha f'_{i-1} + f'_i + \alpha f'_{i+1} + \beta f'_{i+2} = c\frac{f_{i+3} - f_{i-3}}{6\Delta x} + b\frac{f_{i+2} - f_{i-2}}{4\Delta x} + a\frac{f_{i+1} - f_{i-1}}{2\Delta x} \; + \; O(\Delta x^{p+1}), \quad (1)$$

[*]vedang.deshpande@tamu.edu
[†]raktim@tamu.edu
[‡]donzis@tamu.edu

where $f_i := f(x_i)$, $f'_i := \frac{\partial f}{\partial x}\Big|_{x=x_i}$, and $O(\Delta x^{p+1})$ indicates that the approximation error is of order $(p+1)$. Note that left hand side (LHS) and right hand side (RHS) stencil sizes for Eq.(1) are 5 and 7 respectively. After substituting Taylor series expansion of the terms in the above equation, relations between the unknown parameters a, b, c, α, and β are obtained by matching the coefficients. If the first unmatched coefficient is associated with Δx^{p+1}, then the formal order of accuracy for the discretization is said to be $p+1$. When f is periodic over the x-domain, derivative at each grid point can be calculated by solving a cyclic linear system of equations with formal truncation error as a global measure of error [2].

In real physical problems which involve a spectrum of scales, one is also keen to quantify and minimize errors incurred at different wavenumbers. The method of matching coefficients does not guarantee the desired spectral behavior of the compact scheme derived from Eq.(1). To ensure that all the relevant scales in a problem are properly resolved, it is crucial to have desired spectral accuracy at the wavenumbers of interest. Over the years, various frameworks were proposed to construct optimized compact schemes with desired spectral accuracy [3–7]. They all formulate compact schemes as generalizations of Padé approximation similar to Eq.(1), and calculate optimal coefficients by minimizing a cost function, most commonly, weighted $\mathcal{L}_2$ norm of suitably defined spectral error over the wavenumber space. Different techniques have been used to solve the minimization problem. For example, [3–6] used a weighting function constructed from unknown variables involved in the optimization problem to make the cost function integrable. Then optimal solution is calculated analytically by finding the local minimum of the cost function subject to order of accuracy constraints. On the other hand, [7] minimizes the cost function weighted equally across all wavenumbers and optimal solution is found using sequential quadratic programming (SQP). An extended optimized compact schemes framework for the implementation of boundary conditions was presented in [4]. To circumvent the computationally expensive inversion of a band matrix in case of compact schemes, [5, 6, 8] proposed prefactored optimized compact finite differences which use forward and backward biased schemes that can be solved explicitly. Frameworks devised specifically to derive optimized explicit schemes were presented in [9–12]. Unlike others, [12] defined the cost function as maximum (infinity) norm of the spectral error and used simulated annealing technique to find the optimal solution. Each afore discussed framework is restrictive in at least one of the following ways:

1. Frameworks for optimal compact schemes are limited to derive pentadiagonal (or tridiagonal) schemes. They can not be easily scaled up for larger stencil sizes.

2. Most of the frameworks rely on a carefully constructed weighting function to make the cost function integrable, which facilitates the analytical computation of optimal coefficients. Therefore, the weighting function has to have a special form, which restricts the choice of weighting function. Moreover, the weighting function is dependent upon the variables of the optimization problem itself. Therefore, nature of the weighting function is not completely known.

3. Frameworks are requirement specific and there is no unifying generalized framework. For example, [3–7] focus on pentadiagonal compact schemes, [1, 9–12] are exclusively for explicit finite difference, and [3, 9–11] approximate only the first derivative.

In this paper we present a unifying framework to derive optimized compact schemes, which overcomes limitations of the existing frameworks. The key features of the proposed framework are as follows:

1. Optimized coefficients of compact schemes to approximate a derivative of any order are determined analytically by solving a quadratic programming problem with equality constraints.

2. First, we develop a framework to derive optimized compact schemes of equal LHS and RHS stencil sizes. Numerical simulations show that the schemes with equal LHS and RHS stencil sizes perform better than those with unequal stencil sizes for a fixed degrees of freedom in the optimization.

3. This generalized framework also allows us to derive compact schemes with unequal LHS and RHS stencil sizes, explicit finite differences, and biased schemes for non-periodic domains by imposing additional equality constraints in the optimization problem, which are discussed as special cases of the framework.

4. The weighting function can be chosen independent of the cost function.

5. A rigorous stability analysis of the schemes is performed to maximize the time step Δt while guaranteeing the stability of discretization.

The organization of the paper is as follows. In Section 2 we present a generalized framework to derive central optimized compact schemes. We perform stability analysis of the schemes for semi-discrete and fully discrete case in Section 3. In Section 4 we discuss a few special cases of the framework. Numerical results obtained using the optimized schemes derived in this paper are presented in Section 5. Conclusions of this study are discussed in Section 6. Appendices at the end of the paper provide supplementary information.

2 Formulation of the unified framework

A generalization of Eq.(1) which is a Padé approximation of d^{th} spatial derivative, $\frac{\partial^d f(x)}{\partial x^d}$, with equal LHS and RHS stencil sizes is given as:

$$\sum_{m=-M}^{M} \mathrm{b}_m f_{i+m}^{(d)} = \frac{1}{(\Delta x)^d} \sum_{m=-M}^{M} \mathrm{a}_m f_{i+m} \; + \; O(\Delta x^{p+1}), \tag{2}$$

where $f_i := f(x_i)$, $f_i^{(d)} := \left.\frac{\partial^d f}{\partial x^d}\right|_{x=x_i}$, $M > 0$, and $N := 2M+1$ is a positive integer which denotes the stencil size. Note that index $m = 0$ corresponds to the i^{th} grid point at which derivative is to be approximated.

For convenience and clarity of discussion, we denote $(p+1)^{\text{th}}$ order accurate optimized compact finite difference schemes obtained using Eq.(2) as $\boldsymbol{\mathcal{O}}_M^M(p+1)$. The superscript M denotes the range of m ($-M$ to M) on the RHS of Eq.(2), and the subscript M is for the range of m on the LHS of Eq.(2). For example, we denote 4^{th} order accurate optimized scheme obtained using $M = 3$ by $\boldsymbol{\mathcal{O}}_3^3(4)$. Schemes derived by standard method of matching coefficients without any optimization, such as presented in [2], are denoted by $\boldsymbol{\mathcal{S}}_M^M(\cdot)$.

Using Taylor series expansion, we get

$$f_{i+m} = f_i + f_i' \, m\Delta x + \cdots + f_i^{(d)} \frac{(m\Delta x)^d}{d!} + \cdots + f_i^{(d+r)} \frac{(m\Delta x)^{(d+r)}}{(d+r)!} + \cdots,$$

and

$$f_{i+m}^{(d)} = f_i^{(d)} + f_i^{(d+1)} (m\Delta x) + \cdots + f_i^{(d+r)} \frac{(m\Delta x)^r}{r!} + \cdots.$$

Substituting them in Eq.(2), and matching coefficients we get the following linear constraints in a_m and b_m,

$$\sum_m \mathrm{a}_m m^j = 0, \text{ for } j = 0, \cdots, d-1;$$
$$\sum_m \left(\frac{m^{d+r}}{(d+r)!} \mathrm{a}_m - \frac{m^r}{r!} \mathrm{b}_m \right) = 0, \text{ for } r = 0, \cdots, p.$$

We use the definition $0^0 = 1$ for the equations above. For a more compact representation of the formulation, let $m \in \{-M, \cdots, M\}$ for periodic domains. Let us define vectors $\mathbf{a}_d \in \mathbb{R}^N$ and $\mathbf{b}_d \in \mathbb{R}^N$ as[1]

$$\mathbf{a}_d := \begin{bmatrix} \mathrm{a}_{-M} & \mathrm{a}_{-M+1} & \cdots & \mathrm{a}_{M-1} & \mathrm{a}_M \end{bmatrix}^T,$$
$$\mathbf{b}_d := \begin{bmatrix} \mathrm{b}_{-M} & \mathrm{b}_{-M+1} & \cdots & \mathrm{b}_{M-1} & \mathrm{b}_M \end{bmatrix}^T.$$

[1] $\mathbb{R}$ denotes the set of all real numbers and $\mathbb{C}$ denotes the set of all complex numbers.

Let us also define a vector

$$\mathbf{m} := \begin{bmatrix} -M & -M+1 & \cdots & M-1 & M \end{bmatrix}^T.$$

We can then write the above linear constraints in terms of $\mathbf{a}_d$ and $\mathbf{b}_d$ as

$$\mathbf{a}_d^T \mathbf{X}_d - \mathbf{b}_d^T \mathbf{Y}_d = \mathbf{0}_{1\times(d+p+1)} \tag{3}$$

where $\mathbf{0}_{m\times n} \in \mathbb{R}^{m\times n}$ denotes a zero-matrix. The matrices $\mathbf{X}_d \in \mathbb{R}^{N\times(d+p+1)}$, $\mathbf{Y}_d \in \mathbb{R}^{N\times(d+p+1)}$ are defined as

$$\mathbf{X}_d := \begin{bmatrix} \mathbf{1}_N & \mathbf{m} & \cdots & \mathbf{m}^{d-1} & \frac{\mathbf{m}^d}{d!} & \cdots & \frac{\mathbf{m}^{d+p}}{(d+p)!} \end{bmatrix}, \tag{4}$$

$$\mathbf{Y}_d := \begin{bmatrix} \mathbf{0}_{N\times d} & \mathbf{1}_N & \mathbf{m} & \cdots & \frac{\mathbf{m}^p}{p!} \end{bmatrix}. \tag{5}$$

where $\mathbf{1}_N \in \mathbb{R}^N$ denotes a vector whose all elements are unity, and the exponents of $\mathbf{m}$ are element wise. Eq.(3) represents $(d+p+1)$ constraints on $2(2M+1)$ variables. When the number of constraints is less than the number of variables, the system of linear equations can not be solved uniquely. In other words, the system has extra degrees of freedom which can be used to calculate a set of variables which minimizes a well defined cost function. We make use of this freedom to reduce the spectral error.

To study the spectral behavior of finite difference schemes at different wavenumbers, it is customary to work in the wavenumber space. Therefore, we consider a discrete Fourier mode as follows

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

where, $j := \sqrt{-1}$, and $\hat{f}$ is the Fourier coefficient of the mode corresponding to wavenumber k. Then the analytical d^{th} derivative is

$$f^{(d)}(x) = (jk)^d f(x) \tag{6}$$

Throughout the paper, we use the notation $\tilde{f}(\cdot)$ to denote the approximation of $f(\cdot)$. Therefore, the approximated d^{th} derivative is denoted by $\tilde{f}^{(d)}(x)$, i.e. $\tilde{f}^{(d)}(x) \approx f^{(d)}(x)$. The finite difference approximation of the d^{th} derivative follows from Eq.(2) as

$$\left(\sum_m \mathrm{b}_m e^{jkm\Delta x}\right) \tilde{f}_i^{(d)} = \frac{1}{(\Delta x)^d} \left(\sum_m \mathrm{a}_m e^{jkm\Delta x}\right) f_i.$$

Let us define the normalized wavenumber as $\eta := k\Delta x$, $\eta \in [0, \pi]$. Then the previous equation can be written in terms of η as

$$\Big(\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)\Big)\mathbf{b}_d \tilde{f}_i^{(d)} = \frac{1}{(\Delta x)^d} \Big(\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)\Big)\mathbf{a}_d f_i,$$

or,

$$\tilde{f}_i^{(d)} = \frac{1}{(\Delta x)^d} \frac{\Big(\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)\Big)\mathbf{a}_d}{\Big(\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)\Big)\mathbf{b}_d} f_i, \tag{7}$$

where

$$\begin{aligned} \mathbf{C}(\eta) &:= \begin{bmatrix} \cos(-M\eta) & \cdots & \cos(-\eta) & 1 & \cos\eta & \cdots & \cos(M\eta) \end{bmatrix}^T, \\ \mathbf{S}(\eta) &:= \begin{bmatrix} \sin(-M\eta) & \cdots & \sin(-\eta) & 0 & \sin(\eta) & \cdots & \sin(M\eta) \end{bmatrix}^T. \end{aligned} \tag{8}$$

Comparison of Eq.(7) with Eq.(6) leads us to define the spectral error $e(\eta)$ at wavenumber $\eta = k\Delta x$ as

$$e(\eta) := \frac{\Big(\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)\Big)\mathbf{a}_d}{\Big(\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)\Big)\mathbf{b}_d} - (j\eta)^d. \tag{9}$$

Let us denote the first term on RHS of Eq.(9) by $\frac{\left(\mathbf{C}^T(\eta)+j\mathbf{S}^T(\eta)\right)\mathbf{a}_d}{\left(\mathbf{C}^T(\eta)+j\mathbf{S}^T(\eta)\right)\mathbf{b}_d} =: (j\tilde{\eta})^d$, where $\tilde{\eta}$ is the *modified wavenumber*. Therefore, the spectral error can also be written as

$$e(\eta) = (j\tilde{\eta})^d - (j\eta)^d = j^d(\tilde{\eta}^d - \eta^d). \tag{10}$$

We compare $\tilde{\eta}^d$ with η^d to show the spectral accuracy of the optimized schemes in the sections to follow. The weighted $\mathcal{L}_2$ norm of a function $e(\eta)$, for $\eta \in [0, \pi]$, is defined as

$$\|e(\eta)\|^2_{\mathcal{L}_2} := \int_0^\pi \gamma(\eta)\bar{e}(\eta)e(\eta)d\eta =: \langle \bar{e}(\eta)e(\eta)\rangle\,. \tag{11}$$

Where, $\bar{e}(\eta)$ is the complex conjugate of $e(\eta)$, and $\gamma(\eta) \geq 0$ is a known weighting function. The objective here is to determine $\mathbf{a}_d$ and $\mathbf{b}_d$, which minimize $\|e(\eta)\|^2_{\mathcal{L}_2}$, subject to order accuracy constraint, i.e.

$$\min_{\mathbf{a}_d, \mathbf{b}_d \in \mathbb{R}^N} \|e(\eta)\|^2_{\mathcal{L}_2}, \text{ subject to Eq.(3).}$$

Before proceeding further, we present a definition and a lemma to be used in the subsequent discussion.

Definition 1. *A given vector $\mathbf{x} \in \mathbb{R}^n$, is said to be symmetric if $\mathbf{Jx} = \mathbf{x}$, or skew-symmetric if $\mathbf{Jx} = -\mathbf{x}$, where $\mathbf{J}$ is an anti-diagonal identity matrix of dimension n.*

Note that, $\mathbf{J}$ is an operator which reverses the order of elements in the vector. For example, let,

$$\mathbf{J} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}; \quad \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \text{then,} \quad \mathbf{Jx} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}; \quad \text{and} \quad \mathbf{x}^T\mathbf{J} = \begin{bmatrix} 3 & 2 & 1 \end{bmatrix}.$$

It can be easily verified that, $\mathbf{C}(\eta)$ and $\mathbf{S}(\eta)$ defined in Eq.(8) are respectively symmetric and skew-symmetric vectors, i.e., they satisfy

$$\mathbf{C}^T(\eta)\mathbf{J} = \mathbf{C}^T(\eta) \quad \text{and} \quad \mathbf{S}^T(\eta)\mathbf{J} = -\mathbf{S}^T(\eta). \tag{12}$$

Also note that, for any $\mathbf{x} \in \mathbb{R}^n$, we can define a symmetric vector $\mathbf{x}_s := (\mathbf{x} + \mathbf{Jx})/2$ and a skew-symmetric vector $\mathbf{x}_w := (\mathbf{x} - \mathbf{Jx})/2$ such that $\mathbf{x} = \mathbf{x}_s + \mathbf{x}_w$.

Lemma 1. *Let $f(x) : [q_1, q_2] \to \mathbb{C}$ and $g(x) : [q_1, q_2] \to \mathbb{C}$, $g(x) \neq 0$, where $q_1 \leq q_2 \in \mathbb{R}$. If $\bar{g}(x)g(x) \geq \epsilon^2$ for some $\epsilon > 0$, then the following is true*

$$\left\| \frac{f(x)}{g(x)} \right\|^2_{\mathcal{L}_2} \leq \frac{\|f(x)\|^2_{\mathcal{L}_2}}{\epsilon^2}.$$

Proof. By assumption, $\bar{g}(x)g(x) \geq \epsilon^2$. It follows immediately that,

$$\frac{1}{\bar{g}(x)g(x)} \leq \frac{1}{\epsilon^2} \implies \frac{\bar{f}(x)f(x)}{\bar{g}(x)g(x)} \leq \frac{\bar{f}(x)f(x)}{\epsilon^2}$$

$$\implies \int_{q_1}^{q_2} \gamma(x)\frac{\bar{f}(x)f(x)}{\bar{g}(x)g(x)}dx \leq \int_{q_1}^{q_2} \gamma(x)\frac{\bar{f}(x)f(x)}{\epsilon^2}dx,$$

for weight function $\gamma(x) \geq 0$. By using the definition of $\mathcal{L}_2$ norm from Eq.(11), we get the desired result. □

Now, we consider the minimization problem for even and odd derivatives separately, and solve it analytically.

2.1 Even derivatives

For even d, i.e. $d := 2q$, for $q = \{1, 2, 3, \cdots\}$, the spectral error from Eq.(9) is

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

Therefore,

$$\|e(\eta)\|_{\mathcal{L}_2} = \left\| \frac{\left(\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{C}^T(\eta)\mathbf{b}_d\right) + j\left(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{S}^T(\eta)\mathbf{b}_d\right)}{\left(\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)\right)\mathbf{b}_d} \right\|_{\mathcal{L}_2}.$$

Minimization of $\|e(\eta)\|^2_{\mathcal{L}_2}$ is not a convex problem. Therefore, we use Lemma 1 to relax the problem and find an upper bound on $\|e(\eta)\|^2_{\mathcal{L}_2}$. But, we have to ensure that $\bar{g}(\eta)g(\eta) \geq \epsilon^2$, where $g(\eta) := (\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta))\mathbf{b}_d$. Note that,

$$\begin{aligned}\bar{g}(\eta)g(\eta) &= (\mathbf{C}^T(\eta)\mathbf{b}_d)^2 + (\mathbf{S}^T(\eta)\mathbf{b}_d)^2 \\ &= (\mathbf{C}^T(\eta)\mathbf{b}_{d,s} + \mathbf{C}^T(\eta)\mathbf{b}_{d,w})^2 + (\mathbf{S}^T(\eta)\mathbf{b}_{d,s} + \mathbf{S}^T(\eta)\mathbf{b}_{d,w})^2,\end{aligned}$$

where $\mathbf{b}_d$ is decomposed into symmetric and skew-symmetric parts as per Definition 1. Using the fact that any symmetric and skew-symmetric vectors are orthogonal to each other, we get,

$$\bar{g}(\eta)g(\eta) = (\mathbf{C}^T(\eta)\mathbf{b}_{d,s})^2 + (\mathbf{S}^T(\eta)\mathbf{b}_{d,w})^2.$$

RHS of the previous equation is sum of squares and it can be zero only if each term is zero individually. If we ensure that at least one term is non-zero, then we can use Lemma 1. Consider the first term,

$$(\mathbf{C}^T(\eta)\mathbf{b}_{d,s})^2 = \mathbf{b}_{d,s}^T\mathbf{C}(\eta)\mathbf{C}^T(\eta)\mathbf{b}_{d,s}.$$

This term can be zero only if either $\mathbf{b}_{d,s} = \mathbf{0}$, or $\mathbf{b}_{d,s} \neq \mathbf{0}$ and $\mathbf{b}_{d,s} \in \text{null}(\mathbf{C}(\eta)\mathbf{C}^T(\eta))$. Because of the special structure of $\mathbf{C}(\eta)\mathbf{C}^T(\eta)$, it can be easily verified that no symmetric vector lies in its null space. This eliminates the second possibility. And we eliminate the first possibility by setting the central element of $\mathbf{b}_d$, i.e. b_0 to be a non-zero real constant κ_0, i.e.

$$\mathrm{b}_0 = \kappa_0 \neq 0.$$

This also ensures that, for such a $\kappa_0 \neq 0$, there exists an $\epsilon > 0$, such that $(\mathbf{C}^T(\eta)\mathbf{b}_{d,s})^2 \geq \epsilon^2$. Consequently,

$$\bar{g}(\eta)g(\eta) = (\mathbf{C}^T(\eta)\mathbf{b}_d)^2 + (\mathbf{S}^T(\eta)\mathbf{b}_d)^2 \geq \epsilon^2, \quad \text{if } \mathrm{b}_0 = \kappa_0 \neq 0.$$

Now we can use Lemma 1 to get

$$\begin{aligned}\|e(\eta)\|^2_{\mathcal{L}_2} &= \left\| \frac{\left(\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{C}^T(\eta)\mathbf{b}_d\right) + j\left(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{S}^T(\eta)\mathbf{b}_d\right)}{\left(\mathbf{C}^T(\eta) + j\mathbf{S}^T(\eta)\right)\mathbf{b}_d} \right\|^2_{\mathcal{L}_2} \\ &\leq \frac{\left\|\left(\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{C}^T(\eta)\mathbf{b}_d\right) + j\left(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{S}^T(\eta)\mathbf{b}_d\right)\right\|^2_{\mathcal{L}_2}}{\epsilon^2}. \qquad (13)\end{aligned}$$

We can minimize an upper bound on $\|e(\eta)\|^2_{\mathcal{L}_2}$ by minimizing the numerator of the previous inequality since ϵ is a constant. The actual values of ϵ and κ_0 are immaterial. The constant κ_0 can be chosen arbitrarily. In fact it can be verified empirically that numerical values of optimized $\mathbf{a}_d$ and $\mathbf{b}_d$ normalized by κ_0 are independent of its prescribed value. For simplicity, we select κ_0 and hence b_0 to be unity, as in [2]. This also

provides an intrinsic scaling for coefficients, where all coefficients are scaled with respect to b_0. Therefore, a constraint on b_0 is imposed as

$$\mathrm{b}_0 = 1; \quad \text{or,} \quad \boldsymbol{\delta}_N^T(M+1)\mathbf{b}_d = 1, \tag{14}$$

where $\boldsymbol{\delta}_N(i) \in \mathbb{R}^N$ is a unit vector whose i^{th} entry is 1.
We therefore minimize

$$\begin{aligned}
&\left\| \left(\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{C}^T(\eta)\mathbf{b}_d\right) + j\left(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{S}^T(\eta)\mathbf{b}_d\right)\right\|_{\mathcal{L}_2}^2 \\
&= \left\langle \left(\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{C}^T(\eta)\mathbf{b}_d\right)^2 + \left(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{S}^T(\eta)\mathbf{b}_d\right)^2 \right\rangle, \\
&= \begin{pmatrix}\mathbf{a}_d \\ \mathbf{b}_d\end{pmatrix}^T \underbrace{\left\langle \begin{bmatrix}\mathbf{C}(\eta) \\ -(-1)^q\eta^d\mathbf{C}(\eta)\end{bmatrix}\begin{bmatrix}\mathbf{C}(\eta) \\ -(-1)^q\eta^d\mathbf{C}(\eta)\end{bmatrix}^T + \begin{bmatrix}\mathbf{S}(\eta) \\ -(-1)^q\eta^d\mathbf{S}(\eta)\end{bmatrix}\begin{bmatrix}\mathbf{S}(\eta) \\ -(-1)^q\eta^d\mathbf{S}(\eta)\end{bmatrix}^T \right\rangle}_{=:\mathbf{Q}_d} \begin{pmatrix}\mathbf{a}_d \\ \mathbf{b}_d\end{pmatrix}, \\
&= \begin{pmatrix}\mathbf{a}_d \\ \mathbf{b}_d\end{pmatrix}^T \mathbf{Q}_d \begin{pmatrix}\mathbf{a}_d \\ \mathbf{b}_d\end{pmatrix}.
\end{aligned}$$

The optimization problem is therefore

$$\left.\begin{aligned} &\min_{\mathbf{a}_d, \mathbf{b}_d \in \mathbb{R}^N} \begin{pmatrix}\mathbf{a}_d \\ \mathbf{b}_d\end{pmatrix}^T \mathbf{Q}_d \begin{pmatrix}\mathbf{a}_d \\ \mathbf{b}_d\end{pmatrix}, \\ \text{subject to} \quad &\mathbf{a}_d^T\mathbf{X}_d - \mathbf{b}_d^T\mathbf{Y}_d = \mathbf{0}_{1\times(d+p+1)}, \\ &\boldsymbol{\delta}_N^T(M+1)\mathbf{b}_d = 1. \end{aligned}\right\} \tag{15}$$

The equality constraints can also be written as

$$\begin{bmatrix}\mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{1\times N} & \boldsymbol{\delta}_N^T(M+1)\end{bmatrix}\begin{pmatrix}\mathbf{a}_d \\ \mathbf{b}_d\end{pmatrix} = \begin{bmatrix}\mathbf{0}_{(d+p+1)\times 1} \\ 1\end{bmatrix}. \tag{16}$$

This is a quadratic programming problem with equality constraints. The solution to this problem can be found in many standard textbooks such as [13]. The analytical solution is given by the following Karush-Kuhn-Tucker (KKT) condition

$$\begin{bmatrix} 2\mathbf{Q}_d & \begin{pmatrix}\mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{1\times N} & \boldsymbol{\delta}_N^T(M+1)\end{pmatrix}^T \\ \begin{pmatrix}\mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{1\times N} & \boldsymbol{\delta}_N^T(M+1)\end{pmatrix} & \mathbf{0}_{(d+p+2)\times(d+p+2)} \end{bmatrix}\begin{pmatrix}\mathbf{a}_d^* \\ \mathbf{b}_d^* \\ \boldsymbol{\mu}_d^*\end{pmatrix} = \begin{bmatrix}\mathbf{0}_{2N\times 1} \\ \mathbf{0}_{(d+p+1)\times 1} \\ 1\end{bmatrix}, \tag{17}$$

where $\boldsymbol{\mu}_d$ is the Lagrange multiplier associated with Eq.(16), and the matrix on LHS is called the KKT matrix. Assuming the inverse exists, the optimal solution to Eq.(15) is given by

$$\begin{pmatrix}\mathbf{a}_d^* \\ \mathbf{b}_d^* \\ \boldsymbol{\mu}_d^*\end{pmatrix} = \begin{bmatrix} 2\mathbf{Q}_d & \begin{pmatrix}\mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{1\times N} & \boldsymbol{\delta}_N^T(M+1)\end{pmatrix}^T \\ \begin{pmatrix}\mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{1\times N} & \boldsymbol{\delta}_N^T(M+1)\end{pmatrix} & \mathbf{0}_{(d+p+2)\times(d+p+2)} \end{bmatrix}^{-1}\begin{bmatrix}\mathbf{0}_{2N\times 1} \\ \mathbf{0}_{(d+p+1)\times 1} \\ 1\end{bmatrix}. \tag{18}$$

Although $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$ are optimal for the optimization problem in Eq.(15), we note that they only minimize an upper bound on the $\|e(\eta)\|_{\mathcal{L}_2}^2$ given by Eq.(13) instead of the $\|e(\eta)\|_{\mathcal{L}_2}^2$ itself. Hereafter, we refer to $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$ as the optimized coefficients.

The optimized coefficients $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$ for even derivatives turn out to be symmetric, and the imaginary component of the spectral error vanishes. This is a well-known feature of the central difference schemes that is

provable via different approaches. For the sake of completeness, one such approach providing the analytical justification for symmetric coefficients from the optimization perspective considering the cost function in Eq.(15) is given in Appendix A.

Fig.1 shows the solution of the optimization problem Eq.(15) for approximating second derivative with fourth order accuracy, for various stencil sizes. As discussed above, both $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$ are symmetric. Numerical values of the coefficients are tabulated in Table 1 of Appendix B. Note that, for $M = 1$, there is no degree of freedom in the system for optimization. Consequently, for this case we recover the non-optimized standard scheme. In other words, $\boldsymbol{\mathcal{O}}_1^1(4)$ is identical to the non-optimized $\boldsymbol{\mathcal{S}}_1^1(4)$.

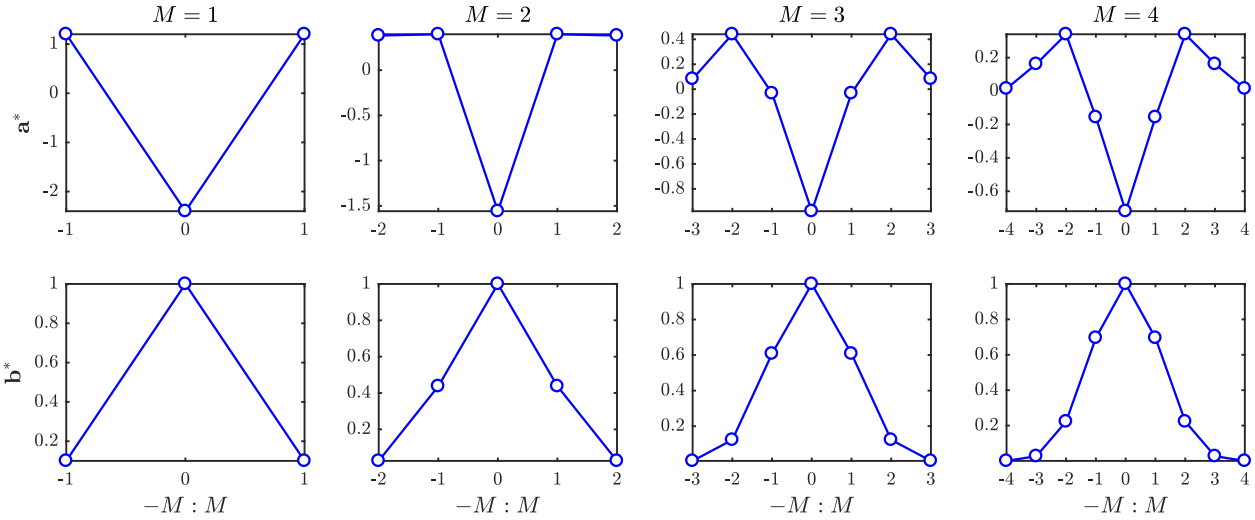

Figure 1: Optimized stencil coefficients for $\boldsymbol{\mathcal{O}}_M^M(4)$ approximating the second derivative. $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

For a specified value of M, we can derive standard schemes of order $4M$, i.e. $\boldsymbol{\mathcal{S}}_M^M(4M)$, to approximate the first and second derivatives. Therefore, as M is increased, order of accuracy of the schemes improves, and one may also expect improved spectral accuracy. However, these standard schemes are not optimized to reduce the spectral error. On the other hand, in case of the optimized schemes, we can achieve even better spectral accuracy by increasing M and holding the order of accuracy constant. A similar comparison of the convergence of spectral error (in log scale) for $\boldsymbol{\mathcal{S}}_M^M(4M)$ (magenta) and $\boldsymbol{\mathcal{O}}_M^M(4)$ (blue) is shown in Fig.2. Clearly, the optimized schemes present better convergence of the spectral error as M is increased. In other words, for a fixed computational cost (characterized by M), the optimized schemes provide better spectral accuracy than the standard schemes.

The $\mathcal{L}_2$ norm of $e(\eta)$ provides only a cumulative measure of the spectral error across all wavenumbers. Nature of the spectral error can be further analyzed as shown in Fig.3. The first row of Fig.3 shows the comparison of $\tilde{\eta}^2$ (defined in Eq.(10)) for different values of M, with η^2 which is calculated analytically (black dashed). In all subplots, blue and magenta lines correspond to $\boldsymbol{\mathcal{O}}_M^M(4)$ and $\boldsymbol{\mathcal{S}}_M^M(4M)$ respectively. It is evident that, $\tilde{\eta}^2$ for the optimized schemes is closer to analytical η^2 than the standard schemes, especially in the higher wavenumber region.

Absolute values of the real and imaginary components of $e(\eta)$ are shown in the bottom two rows of Fig.3. For a better visualization, $|\Re[e(\eta)]|$ is plotted in log scale. The standard schemes show lower errors in lower wavenumber region, but these errors increase quickly with the wavenumber. In contrast, the optimized schemes show almost similar accuracy across the wavenumber space due to the weighting function $\gamma(\eta) = 1$ that weights all wavenumbers equally, and provide a better resolving efficiency. Note that the spectral error for the optimized schemes is similar to the standard schemes for $\eta > 3$ since it lies outside the optimization bandwidth of the optimized schemes. As expected, the imaginary component $\Im[e(\eta)]$ is zero for the second

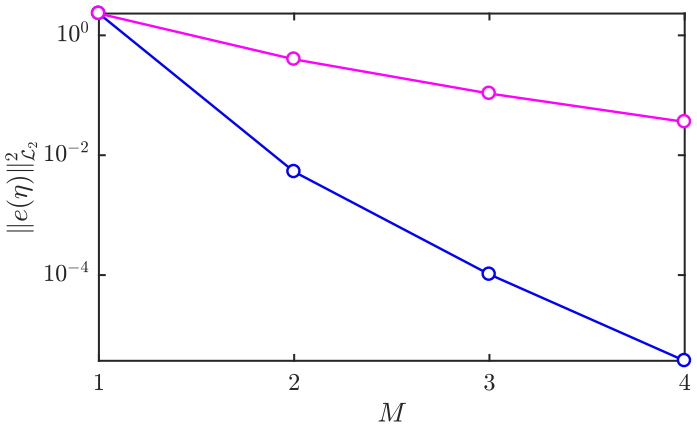

Figure 2: Convergence of $\mathcal{L}_2$ norm of the spectral error for $\boldsymbol{\mathcal{O}}_M^M(4)$ (blue) and $\boldsymbol{\mathcal{S}}_M^M(4M)$ (magenta) approximating the second derivative. $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

derivative.

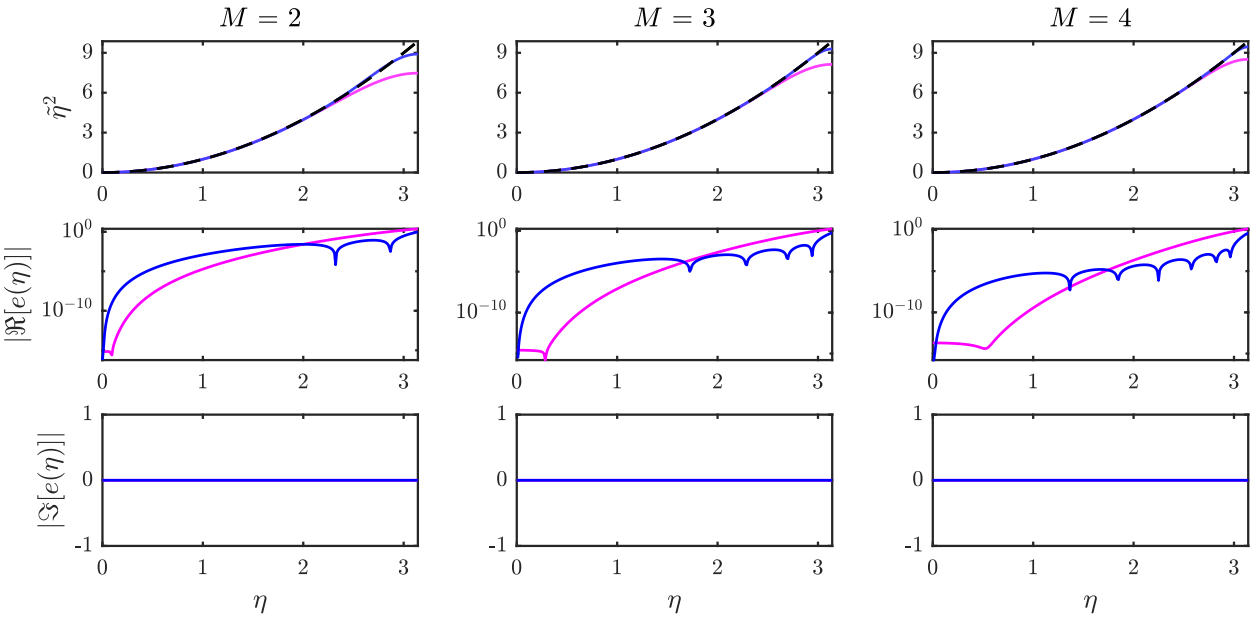

Figure 3: Comparison of $\tilde{\eta}^2$, and real and imaginary components of spectral error for $\boldsymbol{\mathcal{O}}_M^M(4)$ (blue) and $\boldsymbol{\mathcal{S}}_M^M(4M)$ (magenta) approximating the second derivative. $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise. Analytical η^2 is shown by the black-dashed line.

2.2 Odd derivatives

For odd d, i.e. $d := 2q + 1$, for $q = \{0, 1, 2, \cdots\}$, the spectral error from Eq.(9) is

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

Using arguments similar to the case for even d, here we minimize

$$\left\| \left(\mathbf{C}^T(\eta)\mathbf{a}_d + (-1)^q\eta^d\mathbf{S}^T(\eta)\mathbf{b}_d\right) + j\left(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{C}^T(\eta)\mathbf{b}_d\right)\right\|_{\mathcal{L}_2}^2,$$
$$= \begin{pmatrix}\mathbf{a}_d\\ \mathbf{b}_d\end{pmatrix}^T \mathbf{Q}_d \begin{pmatrix}\mathbf{a}_d\\ \mathbf{b}_d\end{pmatrix},$$

$$\text{where,}\quad \mathbf{Q}_d = \left\langle \begin{bmatrix}\mathbf{C}(\eta)\\ (-1)^q\eta^d\mathbf{S}(\eta)\end{bmatrix}\begin{bmatrix}\mathbf{C}(\eta)\\ (-1)^q\eta^d\mathbf{S}(\eta)\end{bmatrix}^T + \begin{bmatrix}\mathbf{S}(\eta)\\ -(-1)^q\eta^d\mathbf{C}(\eta)\end{bmatrix}\begin{bmatrix}\mathbf{S}(\eta)\\ -(-1)^q\eta^d\mathbf{C}(\eta)\end{bmatrix}^T \right\rangle. \tag{19}$$

The optimization problem is the same as Eq.(15), with $\mathbf{Q}_d$ defined by Eq.(19), and the analytical solution is given by Eq.(18).

As discussed in Appendix A, a well-known result for odd derivatives shows that the optimized coefficients $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$ are respectively skew-symmetric and symmetric, and the real component of the spectral error becomes zero.

Fig.4 shows the solution of the optimization problem Eq.(15) for approximating the first derivative with fourth order accuracy, for various stencil sizes. Numerical values of the coefficients are tabulated in Table 2. As in the case of second derivative, here also we recover the standard non-optimized scheme, $\boldsymbol{\mathcal{S}}_1^1(4)$, for $M = 1$.

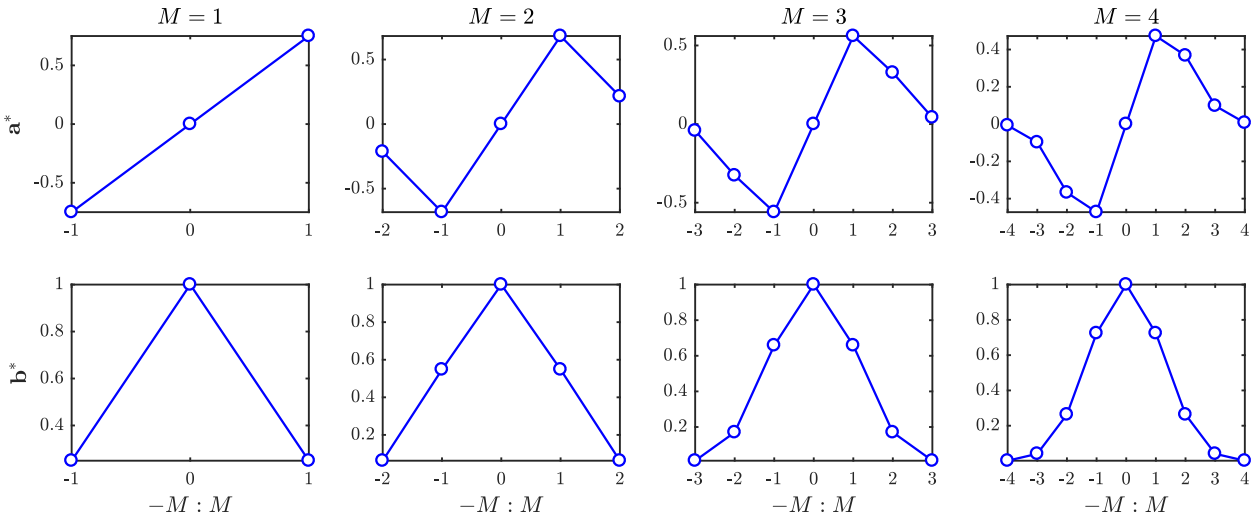

Figure 4: Optimized stencil coefficients for $\boldsymbol{\mathcal{O}}_M^M(4)$ approximating the first derivative. $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

Again, as in the case of second derivative, similar observations can be made for the spectral accuracy of optimized schemes approximating the first derivative. Fig.5 and Fig.6 show that the optimized schemes provide better resolving efficiency than the standard schemes for a fixed computational cost.

Quadratic programs such as Eq.(15) are known to suffer from the rank deficient KKT matrix, especially when dimension of the problem is large. Therefore, solving Eq.(17) may not be trivial for large M when the KKT matrix is not invertible. Numerous works exist in the literature which discuss special algorithms to solve large quadratic programs. For instance, see [14] and the references therein. However, a detailed discussion on such algorithms is out of the scope of this paper. We observed empirically that the KKT matrix in Eq.(18) becomes rank deficient for $M \geq 6$ and $\eta \in [0, 3]$. We also observed that the minimum M at which the KKT matrix becomes rank deficient depends on the range of η. However, the schemes of practical relevance with smaller M, such as tabulated in Appendix B, can be computed easily with any

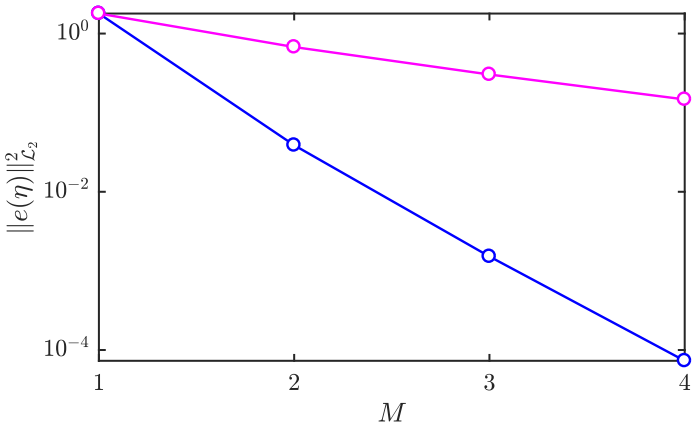

Figure 5: Convergence of $\mathcal{L}_2$ norm of the spectral error for $\boldsymbol{\mathcal{O}}_M^M(4)$ (blue) and $\boldsymbol{\mathcal{S}}_M^M(4M)$ (magenta) approximating the first derivative. $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

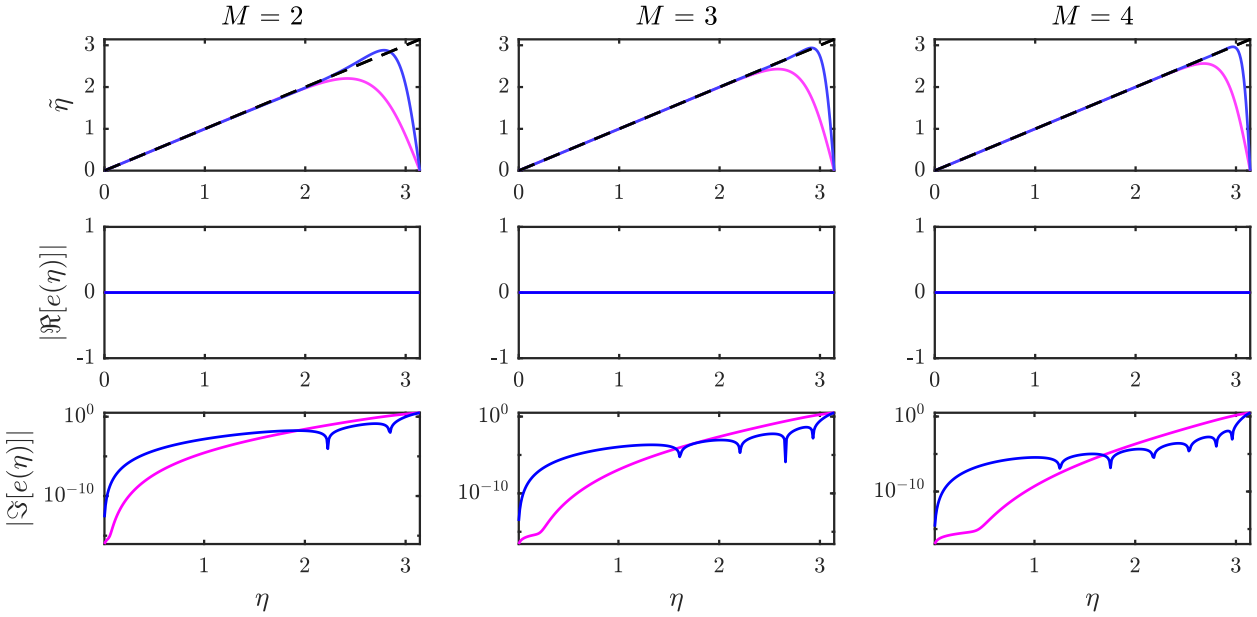

Figure 6: Comparison of $\tilde{\eta}$, and real and imaginary components of spectral error for $\boldsymbol{\mathcal{O}}_M^M(4)$ (blue) and $\boldsymbol{\mathcal{S}}_M^M(4M)$ (magenta) approximating the first derivative. $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise. Analytical η is shown by the black-dashed line.

standard linear algebra package on a personal computer. The explicit schemes discussed in Section 4.2 do not suffer from this problem.

Next, we consider the effect of $\gamma(\eta)$ on spectral accuracy of the optimized schemes.

2.3 Effect of $\gamma(\eta)$

Recall that we used weighting function $\gamma(\eta) \geq 0$ when we defined $\mathcal{L}_2$-norm of error in Eq.(11). In the optimal solutions presented so far, we have used $\gamma(\eta) = 1$ for $\eta \in [0, 3]$ and $\gamma(\eta) = 0$ otherwise, i.e. equal weight was

given to all wavenumbers within $[0, 3]$. We can use $\gamma(\eta)$ to weight some preferred wavenumbers more than others, and the choice is typically motivated by the physics involved in the PDE. This essentially means that, we can keep the error lower at preferred wavenumbers at the expense of higher errors at other wavenumbers that are not vital to the physics of problem.

To study the effect of weighting function on spectral behavior of the optimized schemes, let us consider a candidate function $\gamma(\eta) = \exp(\alpha\eta) \geq 0$ where $\alpha \in \mathbb{R}$. By choosing different values of α, one can assign different weights to different wavenumbers. For example, $\alpha < 0$ weights lower wavenumbers more than the higher ones, $\alpha = 0$ i.e., $\gamma(\eta) = 1$ weights all wavenumbers equally, and $\alpha > 0$ weights higher wavenumbers more that the lower ones. This spectral behavior of optimized schemes is illustrated in Fig.7.

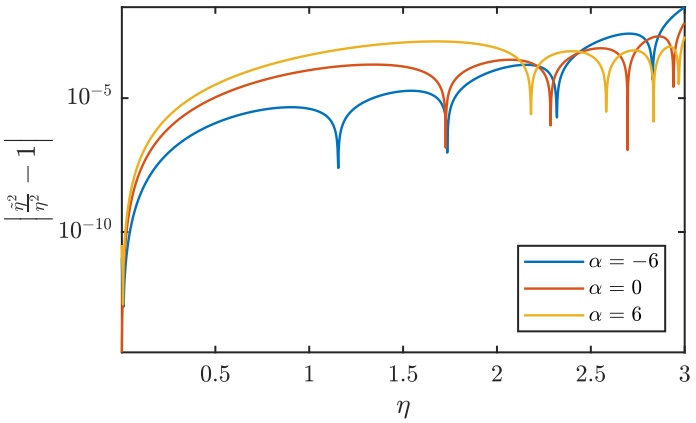

Figure 7: Spectral error for $\boldsymbol{\mathcal{O}}_3^3(4)$ approximating the second derivative for different $\gamma(\eta) = \exp(\alpha\eta)$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

The choice of exponential for weighting function is purely for the purpose of illustration in Fig.7. One can also weight wavenumbers only over a finite number of intervals, e.g., $\gamma(\eta) = \sin(\eta)$ for $\eta \in [0, 1]$, $\gamma(\eta) = 1$ for $\eta \in [2, 3]$, and $\gamma(\eta) = 0$ otherwise. From this discussion it is evident that, the flexibility of choosing $\gamma(\eta)$ can be leveraged to improve the accuracy of solution at wavenumbers that are relevant to the physics of problem.

The unified framework presented in this section allows us to derive central compact schemes with specified order of accuracy. The schemes are also optimized using $\gamma(\eta)$ to have desired spectral accuracy over the wavenumber space. The third important aspect of discretization, in addition to order of accuracy and spectral error, is the temporal stability, which is discussed next.

3 Stability

Let us consider the general linear partial differential equation

$$\frac{\partial f}{\partial t} = \sum_{d=1}^{D} \beta_d \frac{\partial^d f}{\partial^d x}. \tag{20}$$

Suppose the spatial periodic domain is discretized in N_p grid points. The optimized coefficients for the i^{th} grid point are derived using Eq.(2). Since i is arbitrary, it is obvious that the optimized coefficients must be identical for all i in the domain. Approximation of the d^{th} derivative at all grid points using Eq.(2) results in a cyclic linear system of equations [2]. These linear equations can be written compactly in the following

matrix form

$$\mathbf{B}_d^{\mathbf{\Phi}}\tilde{\mathbf{F}}^{(d)} = \frac{1}{(\Delta x)^d}\mathbf{A}_d^{\mathbf{\Phi}}\mathbf{F}, \tag{21}$$

where, $\mathbf{F}$ is the vector of known function values at the grid points, and $\tilde{\mathbf{F}}^{(d)}$ is the vector of approximated d^{th} derivatives at the respective grid points, i.e.

$$\mathbf{F} := \begin{bmatrix} f_1 & f_2 & \cdots & f_{N_p} \end{bmatrix}^T, \quad \text{and} \quad \tilde{\mathbf{F}}^{(d)} := \begin{bmatrix} \tilde{f}_1^{(d)} & \tilde{f}_2^{(d)} & \cdots & \tilde{f}_{N_p}^{(d)} \end{bmatrix}^T.$$

The matrices $\mathbf{A}_d^{\mathbf{\Phi}}$, $\mathbf{B}_d^{\mathbf{\Phi}} \in \mathbb{R}^{N_p \times N_p}$ are sparse circulant matrices whose rows contain cyclically shifted optimized coefficients $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$. The superscript $\mathbf{\Phi}$ indicates that the matrices are circulant, and we use this notation for the consistency with [1].

Therefore, finite-difference approximation for the d^{th} derivative over the entire domain follows from Eq.(21) as

$$\tilde{\mathbf{F}}^{(d)} = \frac{1}{(\Delta x)^d}(\mathbf{B}_d^{\mathbf{\Phi}})^{-1}\mathbf{A}_d^{\mathbf{\Phi}}\mathbf{F}. \tag{22}$$

subject to the existence of the inverse. For all the schemes presented in this paper, $\mathbf{B}_d^{\mathbf{\Phi}}$ turns out to be invertible.

3.1 Stability of semi-discrete scheme

Similar to the stability analysis presented in [1] for explicit spatial schemes, herein we generalize it for implicit schemes. When the Eq.(20) is integrated temporally, the solution suffers from approximation errors even if the initial condition is exactly known because we approximate the spatial derivatives using Eq.(22). Therefore, we denote the approximate solution at time t, at the grid points as a vector $\tilde{\mathbf{F}}(t)$. Using spatial discretization Eq.(22), we can convert PDE Eq.(20) into an ODE which is continuous in time as follows

$$\dot{\tilde{\mathbf{F}}}(t) = \underbrace{\left(\sum_{d=1}^{D} \frac{1}{(\Delta x)^d}\beta_d(\mathbf{B}_d^{\mathbf{\Phi}})^{-1}\mathbf{A}_d^{\mathbf{\Phi}}\right)}_{\mathbf{\Lambda}:=} \tilde{\mathbf{F}}(t) = \mathbf{\Lambda}\tilde{\mathbf{F}}(t), \tag{23}$$

with solution

$$\tilde{\mathbf{F}}(t) = \exp(t\mathbf{\Lambda})\mathbf{F_0}, \tag{24}$$

where $\mathbf{F_0}$ is the given initial condition. For a single Fourier mode, the approximate d^{th} derivative can be expressed in terms of the modified wavenumber as $(j\tilde{\eta})^d\hat{\tilde{f}}$, where $\hat{\tilde{f}}$ is the Fourier coefficient calculated from the approximate solution. Then the original PDE becomes,

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

The solution is given by

$$\frac{\hat{\tilde{f}}}{\hat{f}_0} = \exp\left[\sum_{d=1}^{D}\beta_d(j\tilde{\eta})^d t\right],$$

where $\hat{f}_0$ is the initial condition at $t = 0$. 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 satisfied if

$$\Re\left\{\sum_{d=1}^{D}\beta_d(j\tilde{\eta})^d\right\} \leq 0. \tag{25}$$

From Eq.(9), Eq.(57), and Eq.(59)

$$(j\tilde{\eta})^d = \frac{\mathbf{C}^T(\eta)\mathbf{a}_d}{\mathbf{C}^T(\eta)\mathbf{b}_d}, \quad \text{for even } d, \text{ and}$$
$$(j\tilde{\eta})^d = j\left(\frac{\mathbf{S}^T(\eta)\mathbf{a}_d}{\mathbf{C}^T(\eta)\mathbf{b}_d}\right), \quad \text{for odd } d.$$

Therefore,

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

Then Eq.(25) implies,

$$\Re\left\{\sum_{d=1}^{D}\beta_d(j\tilde{\eta})^d\right\} = \beta_2\frac{\mathbf{C}^T(\eta)\mathbf{a}_2}{\mathbf{C}^T(\eta)\mathbf{b}_2} + \beta_4\frac{\mathbf{C}^T(\eta)\mathbf{a}_4}{\mathbf{C}^T(\eta)\mathbf{b}_4} + ... \leq 0. \tag{26}$$

For even d, i.e. $d = 2q$, the optimization of spectral error guarantees,

$$(j\tilde{\eta})^d = (-1)^q\tilde{\eta}^{2q} = \frac{\mathbf{C}^T(\eta)\mathbf{a}_{2q}}{\mathbf{C}^T(\eta)\mathbf{b}_{2q}}.$$

Therefore, the sign of $\frac{\mathbf{C}^T(\eta)\mathbf{a}_{2q}}{\mathbf{C}^T(\eta)\mathbf{b}_{2q}}$ changes alternatively with q as follows,

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

and so on. If coefficients β_d for $d = 2q$ satisfy $\beta_{2q} = (-1)^{q+1}\zeta_{2q}^2$ for $\zeta_{2q} \in \mathbb{R}$, then Eq.(26) is implicitly satisfied. In general, if

$$-\beta_2\tilde{\eta}^2 + \beta_4\tilde{\eta}^4 - \beta_6\tilde{\eta}^6 + ... \leq 0, \tag{27}$$

then the overall discretization is stable.

3.2 Stability of fully discretized scheme

After temporal discretization of Eq.(23) using a given temporal scheme, we get a fully discretized system. We analyze the stability of such system in this section. Our goal is to find maximum Δt for which the overall discretization is stable. We begin with the forward Euler method.

3.2.1 Forward Euler method

Using forward Euler method for $\partial f/\partial t$ and the optimized finite difference approximation for $\partial^d f/\partial x^d$, from Eq.(23) we get the following

$$\tilde{\mathbf{F}}^{n+1} = \left(\mathbf{I}_{N_p} + \sum_{d=1}^{D} \frac{\Delta t}{(\Delta x)^d} \beta_d (\mathbf{B}_d^{\mathbf{\Phi}})^{-1} \mathbf{A}_d^{\mathbf{\Phi}} \right) \tilde{\mathbf{F}}^n. \tag{28}$$

where $\tilde{\mathbf{F}}^n$ is the approximated function at n^{th} time step, and $\mathbf{I}_{N_p}$ is an identity matrix of dimension N_p. This is a linear discrete-time system in $\tilde{\mathbf{F}}$, where the system matrix is dependent on the stencil coefficients $\mathbf{A}_d^{\mathbf{\Phi}}$ and $\mathbf{B}_d^{\mathbf{\Phi}}$. For stability, magnitudes of the eigenvalues of $\left(\mathbf{I}_{N_p} + \Delta t \sum_d \frac{1}{(\Delta x)^d} \beta_d (\mathbf{B}_d^{\mathbf{\Phi}})^{-1} \mathbf{A}_d^{\mathbf{\Phi}} \right)$ should be less than one, which can be achieved by constraining the 2-norm as

$$\left\| \mathbf{I}_{N_p} + \Delta t \sum_{d=1}^{D} \frac{1}{(\Delta x)^d} \beta_d (\mathbf{B}_d^{\mathbf{\Phi}})^{-1} \mathbf{A}_d^{\mathbf{\Phi}} \right\|_2 \leq 1. \tag{29}$$

We use the analytical solution for $\mathbf{A}_d^{\mathbf{\Phi}}$ and $\mathbf{B}_d^{\mathbf{\Phi}}$ from spectral error optimization, and maximize Δt subject to the stability constraint Eq.(29). The reader is referred to [1] for the solution of this maximization problem using software packages such as `cvx` [15]. Solving optimization problem by this approach may become difficult if temporal schemes other than forward Euler method are used. Therefore, we present an approach to determine maximum Δt for generalized temporal schemes in the following section.

3.2.2 Generalized temporal scheme

Runge-Kutta (RK) methods are widely used for temporal discretization of ODEs to attain higher order of accuracy. Many temporal schemes can be expressed as special cases of implicit RK method, e.g., see Appendix C for different temporal schemes written in Butcher tableau [16] form of RK methods. Thus, herein, we consider implicit RK method as a generalized temporal scheme to determine the maximum Δt which guarantees the stability. An implicit s-stage RK scheme with the following Butcher tableau

$$\begin{array}{c|c} \boldsymbol{c} & \boldsymbol{A} \\ \hline & \boldsymbol{b}^T \end{array} := \begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \dots & a_{1s} \\ c_2 & a_{21} & a_{22} & \dots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\ \hline & b_1 & b_2 & \dots & b_s \end{array} \tag{30}$$

is given by

$$\tilde{f}^{n+1} \quad = \quad \tilde{f}^n + \Delta t \sum_{i=1}^{s} b_i k_i, \tag{31}$$

$$k_i = \dot{\tilde{f}} \left(t_n + c_i \Delta t, \quad \tilde{f}^n + \Delta t \sum_{j=1}^{s} a_{ij} k_j \right),$$

where, $\tilde{f}^n$ is the approximate value of function $f(t)$ at n^{th} time step and $\dot{\tilde{f}}(\cdot)$ is the approximate time derivative of $f(t)$. For a scalar equation such as $\dot{f} = \lambda f$, $\tilde{f}^{n+1}$ can be written in terms of the *stability function* of the RK scheme $r(z)$, $z \in \mathbb{C}$, as [16]

$$\tilde{f}^{n+1} = r(\lambda \Delta t) \tilde{f}^n, \tag{32}$$

$$\text{where,} \quad r(z) = 1 + z \boldsymbol{b}^T (\mathbf{I}_s - z\boldsymbol{A})^{-1} \mathbf{1}_s = \frac{\det(\mathbf{I}_s + z(\mathbf{1}_s \boldsymbol{b}^T - \boldsymbol{A}))}{\det(\mathbf{I}_s - z\boldsymbol{A})},$$

$\det(\cdot)$ denotes the determinant of the matrix. Thus, $r(z)$ is a polynomial-over-polynomial function of z. The sequence formed by Eq.(32) is bounded iff $|r(z)| \leq 1$, which is the stability condition for the RK scheme. The equality $|r(z)| = 1$ defines a closed *stability curve* in the complex plane. Generally, for the explicit RK schemes, area enclosed by the stability curve defines the stability region. While in the case of implicit RK schemes, stability region may lie either inside or outside the closed curve. If stability curve is the imaginary axis itself, then the scheme is stable for all z which lie in the left half of the complex plane. Note that, $\Delta t > 0$ is a real number while λ can be complex, so that $z = \lambda \Delta t$ is complex. The timestep Δt just *scales* the radius of λ while keeping the argument same in the complex plane. We are interested in finding the maximum scaling Δt for which the product $\lambda \Delta t$ lies in the stability region.

As in the forward Euler case, here also we assume that $\mathbf{A}_d^{\mathbf{\Phi}}$ and $\mathbf{B}_d^{\mathbf{\Phi}}$ are determined analytically and hence $\mathbf{\Lambda}$ is known. Now, let us consider multi-dimensional semi-discrete system defined by Eq.(23). We reduce this N_p-dimensional system into N_p scalar decoupled systems using the similarity transformation. Let $\mathbf{D}$ be a diagonal matrix with eigenvalues of the $\mathbf{\Lambda}$ as its diagonal elements and $\mathbf{M}$ be the modal matrix with eigenvectors of $\mathbf{\Lambda}$ as its columns, such that $\mathbf{D} = \mathbf{M}^{-1}\mathbf{\Lambda}\mathbf{M}$. Let us define the transformation $\mathbf{L} := \mathbf{M}\tilde{\mathbf{F}}$. Then Eq.(23) can be written as

$$\mathbf{M}\dot{\mathbf{L}} = \mathbf{\Lambda}\mathbf{M}\mathbf{L} \implies \dot{\mathbf{L}} = \mathbf{M}^{-1}\mathbf{\Lambda}\mathbf{M}\mathbf{L} = \mathbf{D}\mathbf{L}. \tag{33}$$

Since $\mathbf{D}$ is diagonal, Eq.(33) represents N_p scalar decoupled systems as follows

$$\dot{l}_i = \lambda_i l_i, \quad i = 1 \dots N_p,$$

where λ_i, are N_p eigenvalues of $\mathbf{\Lambda}$. The i^{th} scalar system can be fully discretized using the RK scheme, and from Eq.(32) it follows that

$$l_i^{n+1} = r(\lambda_i \Delta t) l_i^n, \quad i = 1 \dots N_p.$$

For overall stability of Eq.(33), it is required that the products $\lambda_i \Delta t$ lie within stability region of the RK scheme. Thus, for known λ_i and $r(z)$, the maximum Δt can be calculated numerically to guarantee the stability of the system.

The stability curves for different RK schemes tabulated in Appendix C are shown in Fig.8. The asterisks show the eigenvalues of $\mathbf{\Lambda}$ obtained for $\boldsymbol{\mathcal{O}}_4^4(4)$, $\beta_1 = -0.1$, $\beta_2 = 0.2$, and $N_p = 31$. In subplots (a), (b), and (c), the area enclosed by the curves defines the stability region. Therefore, maximum allowable Δt is determined by ensuring that the eigenvalues λ_i after scaling, i.e. the products $\lambda_i \Delta t$ lie within the closed curves. On the other hand, in subplot (d), the region outside the curve defines the stability region and all eigenvalues are already in that region. Therefore, this scheme is stable for all $\Delta t > 0$. If we increase the number of grid points, or equivalently decrease Δx, the magnitude of eigenvalues of $\mathbf{\Lambda}$ will increase as Δx^d appears in the denominator of Eq.(23). Consequently, the stability limit or the maximum allowable Δt will decrease. The dependence of Δt on Δx is also affected by the PDE coefficients β_d. Therefore, for linear PDEs, the stability is often discussed in terms of a normalized parameter, namely, Courant-Friedrichs-Lewy (CFL) number, which is discussed next.

The CFL number with respect to d^{th} derivative is defined as $r_d := |\beta_d| \Delta t / \Delta x^d$. By varying Δx and calculating the corresponding maximum Δt, we can calculate and plot CFL numbers r_1 and r_2 for the advection-diffusion equation (see Eq.(48)). A similar plot obtained for the optimized schemes of different stencil sizes for forward Euler and ERK4 method is shown in Fig.9. $M = 1$ (blue) corresponds to the standard scheme $\boldsymbol{\mathcal{S}}_1^1(4)$, and $M = 2$ onwards correspond to the optimized schemes, $\boldsymbol{\mathcal{O}}_M^M(4)$. There is a drastic reduction in stability region from $M = 1$ to $M = 2$. As M is increased further, the stability region decreases.

We next consider the relationship between λ_i and the derivatives in a PDE. For even derivatives, both $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$ are symmetric. Consequently, both $\mathbf{A}_d^{\mathbf{\Phi}}$ and $\mathbf{B}_d^{\mathbf{\Phi}}$ are real symmetric circulant matrices. Therefore, if a PDE consists of only even derivatives, then $\mathbf{\Lambda}$ defined in Eq.(23) is also symmetric and hence all λ_i are real. Real parts of λ_i cause only amplitude decay (diffusion) or amplification but phase of the solution remains constant. On the other hand, for odd derivatives, $\mathbf{a}_d^*$ is skew-symmetric and $\mathbf{b}_d^*$ is symmetric. Therefore,

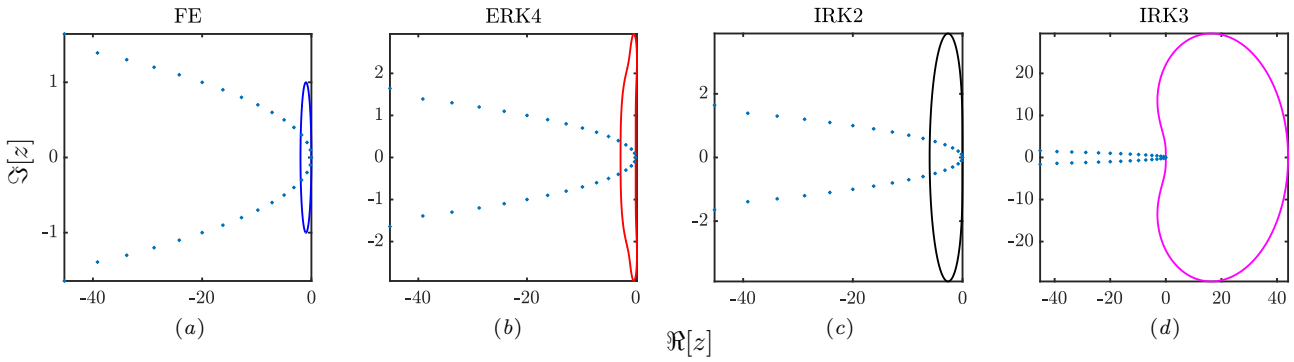

Figure 8: Closed curve shows stability regions for different RK schemes. Asterisks are eigenvalues of $\mathbf{\Lambda}$ for $M = 4$, $\beta = [-0.1, 0.2]$, $N_p = 31$. (a), (b) and (c): Inside the curve is stable. (d): Outside the curve is stable.

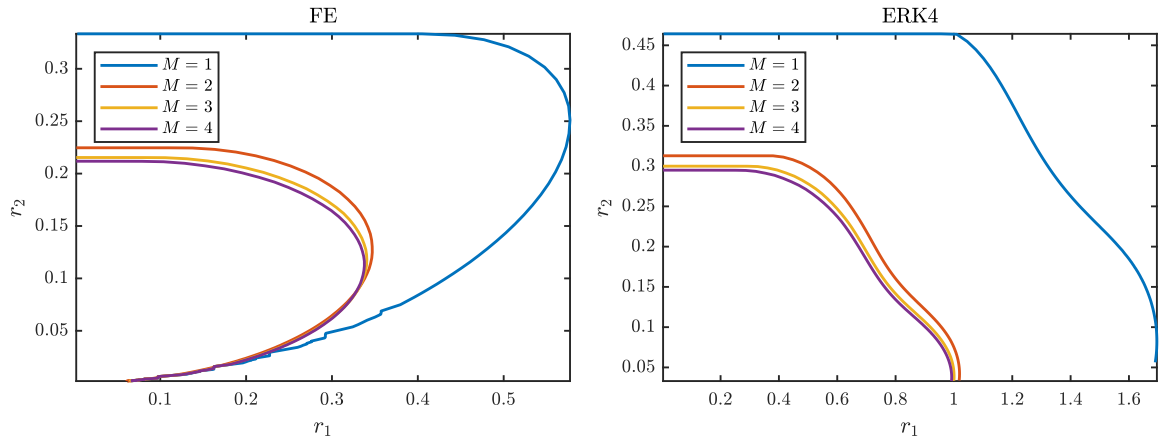

Figure 9: r_2 vs r_1 for forward Euler and ERK4 method. First and second derivatives approximated using $\boldsymbol{\mathcal{O}}_M^M(4)$, $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

circulant matrices $\mathbf{A}_d^{\mathbf{\Phi}}$ and $\mathbf{B}_d^{\mathbf{\Phi}}$ become skew-symmetric and symmetric respectively, which in turn make $\mathbf{\Lambda}$ a skew-symmetric matrix. Consequently, if a PDE consists of only odd derivatives, then all λ_i are pure imaginary which affect only phase of the solution preserving the magnitude of amplitudes, leading to pure dispersion. It is clear that even derivatives contribute to the real part of λ_i and hence diffusion of the numerical solution, and odd derivatives cause dispersion by means of the imaginary part of λ_i. However, this is true only for central schemes. It is shown in the subsequent section that in the case of biased schemes, both even and odd derivatives individually contribute to real and imaginary components of λ_i.

In what follows next, we extend the stability analysis discussed in this section, and framework presented in Section 2 to derive and analyze special types of optimized schemes.

4 Special cases of the framework

We generalize Eq.(2) further as

$$\sum_{m=-M_L^B}^{M_R^B} \mathrm{b}_m f_{i+m}^{(d)} = \frac{1}{(\Delta x)^d} \sum_{m=-M_L^A}^{M_R^A} \mathrm{a}_m f_{i+m} \; + \; O(\Delta x^{p+1}), \tag{34}$$

where, each one of M_L^B, M_R^B, M_L^A and M_R^A can be chosen independently. In Section 2, we have presented a general framework to derive optimized implicit finite difference approximations. These approximations are obtained for equal LHS and RHS stencil sizes, i.e. $M_L^B = M_R^B = M_L^A = M_R^A = M$. In this section, we show that other types of schemes, namely, compact schemes with unequal LHS and RHS stencil sizes, spatially explicit, and biased finite difference approximations can be derived as special cases of the framework presented in Section 2. These special cases are derived by imposing additional constraints on the coefficients $\mathbf{a_d}$ and $\mathbf{b_d}$ while solving the minimization problem defined in Eq.(15).

In the most generalized form, we denote $(p+1)^{\text{th}}$ order accurate optimized schemes derived using Eq.(34) as $\boldsymbol{\mathcal{O}}^{M_L^A, M_R^A}_{M_L^B, M_R^B}(p+1)$. For example, 5^{th} order accurate optimized scheme obtained using $M_L^A = 1, M_R^A = 2, M_L^B = 3$ and $M_R^B = 4$ is denoted by $\boldsymbol{\mathcal{O}}^{1,2}_{3,4}(5)$. See Fig.10 as an illustration of this. The i^{th} grid point at which derivative is to be approximated is shown by the green asterisk, and the neighboring grid points are shown by black circles. Blue and red braces respectively show stencil points used in RHS and LHS of Eq.(34).

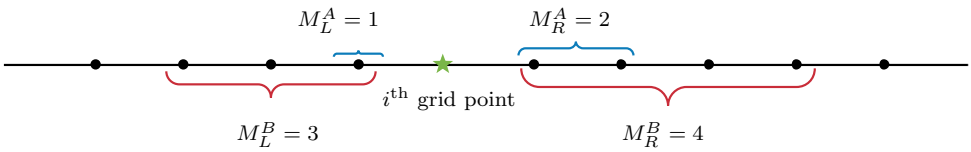

Figure 10: Illustration for the notation of optimized schemes in the most generalized form.

4.1 Central compact schemes with unequal LHS and RHS stencil sizes

In Eq.(2), indices of both a_m and b_m vary from $-M$ to M. We can also have different ranges of index m for a_m and b_m. Let $M_L^B = M_R^B = M^B$ and $M_L^A = M_R^A = M^A$. i.e., Eq.(34) becomes,

$$\sum_{m=-M^B}^{M^B} \mathrm{b}_m f_{i+m}^{(d)} = \frac{1}{(\Delta x)^d} \sum_{m=-M^A}^{M^A} \mathrm{a}_m f_{i+m} \; + \; O(\Delta x^{p+1}), \tag{35}$$

where $M^B \neq M^A$.

Continuing the use of the notation defined in Section 2, we refer to optimized schemes derived using Eq.(35) as $\boldsymbol{\mathcal{O}}^{M^A}_{M^B}(\cdot)$. On the other hand, if a scheme is derived by the *standard* method of coefficient matching without optimization, then it is referred to as $\boldsymbol{\mathcal{S}}^{M^A}_{M^B}(\cdot)$. When Eq.(1) is compared with Eq.(35), we observe that $M^A = 3$ and $M^B = 2$ for Eq.(1). Therefore, pentadiagonal optimized schemes are denoted by $\boldsymbol{\mathcal{O}}^3_2(\cdot)$, and pentadiagonal standard schemes are denoted by $\boldsymbol{\mathcal{S}}^3_2(\cdot)$.

The stencil size for the right side of Eq.(35) is defined as $N^A := 2M^A + 1$, and similarly the stencil size for the left side is $N^B := 2M^B + 1$. Let us define *augmented* parameters as $\hat{M} := \max\{M^A, M^B\}$ and $\hat{N} := 2\hat{M} + 1$.

First, we consider the case $M^A > M^B$, and hence $\hat{M} = M^A$. This $\boldsymbol{\mathcal{O}}^{M^A}_{M^B}(\cdot)$ scheme can be treated as an $\boldsymbol{\mathcal{O}}^{M^A}_{M^A}(\cdot)$ scheme with augmented stencil size $\hat{N} = N^A = 2M^A + 1$, and an extra imposed constraint that $\mathrm{b}_m = 0$ for $m = \{-M^A, -M^A+1, \cdots, -(M^B+1), M^B+1, M^B+2, \cdots, M^A\}$. In other words, we are imposing a constraint that $(N^A - N^B)$ elements of $\mathbf{b}_d$ which are located symmetrically about the central element, to be zero. Therefore, we can use the central difference framework presented in Section 2 with $M = M^A$ and the following additional constraint

$$\underbrace{\begin{bmatrix} \boldsymbol{\delta}_n^T(1) \\ \boldsymbol{\delta}_n^T(2) \\ \vdots \\ \boldsymbol{\delta}_n^T(M^A - M^B) \\ \boldsymbol{\delta}_n^T(M^A + M^B + 2) \\ \boldsymbol{\delta}_n^T(M^A + M^B + 3) \\ \vdots \\ \boldsymbol{\delta}_n^T(\hat{N}) \end{bmatrix}}_{\boldsymbol{\Delta}^T :=} \mathbf{b}_d = \mathbf{0}_{(\hat{N}-N^B)\times 1}, \tag{36}$$

for $n = \hat{N}$, and as defined earlier, $\boldsymbol{\delta}_n(i) \in \mathbb{R}^n$ is a unit vector whose i^{th} entry is 1. Therefore, the constraint can be written as

$$\begin{bmatrix} \mathbf{0}_{(\hat{N}-N^B)\times\hat{N}} & \boldsymbol{\Delta}^T \end{bmatrix} \begin{bmatrix} \mathbf{a}_d \\ \mathbf{b}_d \end{bmatrix} = \mathbf{0}_{(\hat{N}-N^B)\times 1}. \tag{37}$$

The minimization problem defined in Eq.(15) is solved analytically using the KKT condition constructed by incorporating the additional constraint Eq.(37). Then the optimal solution similar to Eq.(18) is given by

$$\begin{pmatrix} \mathbf{a}_d^* \\ \mathbf{b}_d^* \\ \boldsymbol{\mu}_d^* \end{pmatrix} = \mathbf{P} \begin{bmatrix} \mathbf{0}_{2\hat{N}\times 1} \\ \mathbf{0}_{(d+p+1)\times 1} \\ 1 \\ \mathbf{0}_{(\hat{N}-N^B)\times 1} \end{bmatrix}, \tag{38}$$

where,

$$\mathbf{P} = \begin{bmatrix} 2\mathbf{Q}_d & \begin{pmatrix} \mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{1\times\hat{N}} & \boldsymbol{\delta}_{\hat{N}}^T(\hat{M}+1) \\ \mathbf{0}_{(\hat{N}-N^B)\times\hat{N}} & \boldsymbol{\Delta}^T \end{pmatrix}^T \\ \begin{pmatrix} \mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{1\times\hat{N}} & \boldsymbol{\delta}_{\hat{N}}^T(\hat{M}+1) \\ \mathbf{0}_{(\hat{N}-N^B)\times\hat{N}} & \boldsymbol{\Delta}^T \end{pmatrix} & \mathbf{0}_{(\hat{N}-N^B+d+p+2)\times(\hat{N}-N^B+d+p+2)} \end{bmatrix}^{-1}.$$

In a similar manner, we can calculate optimized coefficients for the complementary case $M^A < M^B$ by setting $(N^B - N^A)$ elements of $\mathbf{a}_d$ to zero.

Note that by setting $M^A = 3$ and $M^B = 2$ in Eq.(35), we get the pentadiagonal compact schemes such as presented in [2] and [3]. These frameworks restrict the stencil size and require that $M^A > M^B$. On the other hand, the proposed framework is more general because M^A and M^B can be chosen independently.

In Fig.11, we show the comparison of spectral error $\|e(\eta)\|^2_{\mathcal{L}_2}$ for different optimized schemes $\boldsymbol{\mathcal{O}}^{M^A}_{M^B}(4)$ approximating the first and second derivatives. Let us consider the schemes $\boldsymbol{\mathcal{O}}^1_2(4)$ and $\boldsymbol{\mathcal{O}}^2_1(4)$ which are located symmetrically about the diagonal dotted line, for $d = 1$. Although both the schemes utilize equal number of grid points and hence the same number of degrees of freedom in the optimization, the scheme

$\boldsymbol{\mathcal{O}}_1^2(4)$ in the lower half of the plot presents smaller error than the $\boldsymbol{\mathcal{O}}_2^1(4)$ ($0.258 < 0.308$). In fact, all the schemes in lower half of the plot with $M^A > M^B$ present smaller errors than their complementary schemes with $M^A < M^B$ located symmetrically in the upper half of the plot. A similar observation can be made for $d = 2$ as well. Therefore, based on this data, we can say that adding more grid points on RHS of Eq.(35) is more effective for improving the spectral accuracy than on the LHS.

Now lets consider the schemes $\boldsymbol{\mathcal{O}}_1^3(4)$, $\boldsymbol{\mathcal{O}}_2^2(4)$, and $\boldsymbol{\mathcal{O}}_3^1(4)$ all of which have the same number of degrees of freedom in the optimization. We already discussed that, $\boldsymbol{\mathcal{O}}_1^3(4)$ performs better than $\boldsymbol{\mathcal{O}}_3^1(4)$. However, note that $\boldsymbol{\mathcal{O}}_2^2(4)$ presents the smallest spectral error among these three optimized schemes, and similar observation can be made about other schemes as well. Therefore, we can say that among the family of optimized schemes with equal number of degrees of freedom, the scheme with $M^A = M^B$ (if exists) which lies on the dotted line produces the smallest spectral error.

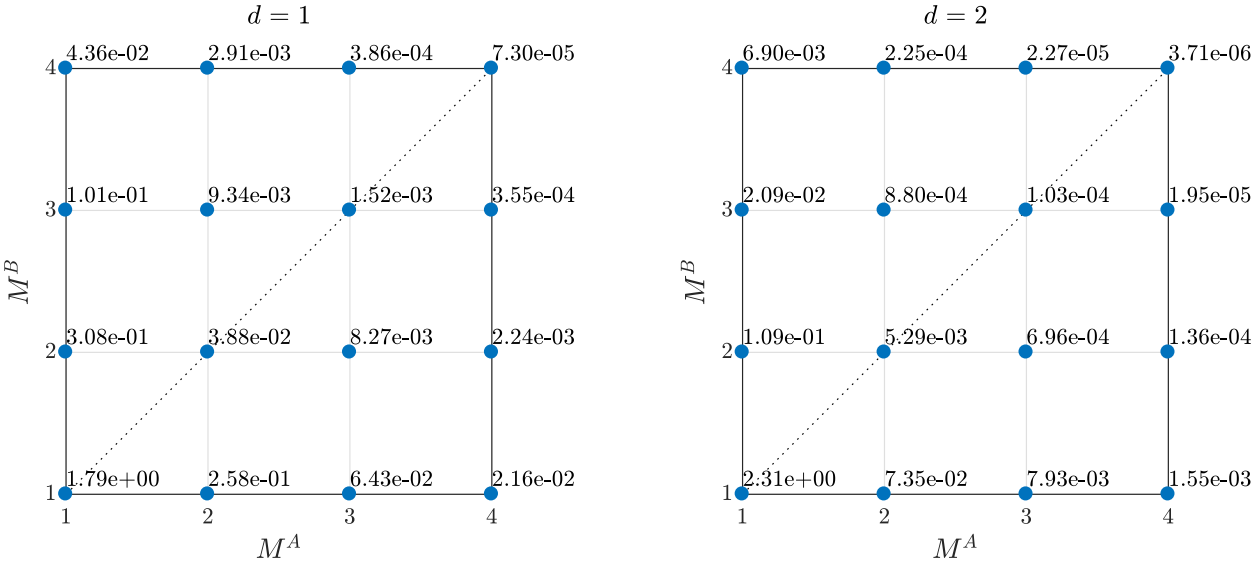

Figure 11: The spectral error $\|e(\eta)\|_{\mathcal{L}_2}^2$ (shown by numerical values) for different $\boldsymbol{\mathcal{O}}_{M^B}^{M^A}(4)$ schemes approximating the first and second derivatives. $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

Also, note that the constraint Eq.(36) sets $\mathrm{b}_m = 0$ symmetrically about the central element of $\mathbf{b}_d$. Therefore, the discussion on (skew-)symmetry of the optimized coefficients in Appendix A, and the stability analysis presented in Section 3 are valid for this special case.

4.2 Central spatially explicit schemes

A finite difference approximation is said to be spatially explicit if the coefficients b_m in Eq.(2) are

$$b_m = \begin{cases} 1 \text{ if } m = 0, \\ 0 \text{ if } m \neq 0. \end{cases}$$

i.e.,

$$f_i^{(d)} = \frac{1}{(\Delta x)^d} \sum_{m=-M^A}^{M^A} \mathrm{a}_m f_{i+m} \; + \; O(\Delta x^{p+1}). \tag{39}$$

It is obvious that central approximations which are spatially explicit can be obtained by substituting $M^B = 0$ in Eq.(35). Therefore, an optimized explicit central schemes is represented by $\boldsymbol{\mathcal{O}}_0^{M^A}(\cdot)$. Constraints Eq.(36)

and Eq.(14) can be combined together as

$$\underbrace{\begin{bmatrix} \boldsymbol{\delta}_n^T(1) \\ \boldsymbol{\delta}_n^T(2) \\ \vdots \\ \boldsymbol{\delta}_n^T(M^A+1) \\ \vdots \\ \boldsymbol{\delta}_n^T(\hat{N}-1) \\ \boldsymbol{\delta}_n^T(\hat{N}) \end{bmatrix}}_{\boldsymbol{\Delta}^T :=} \mathbf{b}_d = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix} = \boldsymbol{\delta}_n(M^A+1), \tag{40}$$

for $n = \hat{N}$, where $\hat{N} = 2M^A + 1$. Since $\boldsymbol{\Delta}^T$ in the above equation is an identity matrix, the constraint reduces to

$$\mathbf{b}_d = \boldsymbol{\delta}_n(M^A+1). \tag{41}$$

Therefore, optimal solution can be found analytically as

$$\begin{pmatrix} \mathbf{a}_d^* \\ \mathbf{b}_d^* \\ \boldsymbol{\mu}_d^* \end{pmatrix} = \begin{bmatrix} 2\mathbf{Q}_d & \begin{pmatrix} \mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{\hat{N}\times\hat{N}} & \mathbf{I}_{\hat{N}} \end{pmatrix}^T \\ \begin{pmatrix} \mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{\hat{N}\times\hat{N}} & \mathbf{I}_{\hat{N}} \end{pmatrix} & \mathbf{0}_{(\hat{N}+d+p+1)\times(\hat{N}+d+p+1)} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0}_{2\hat{N}\times 1} \\ \mathbf{0}_{(d+p+1)\times 1} \\ \boldsymbol{\delta}_{\hat{N}}(M^A+1) \end{bmatrix}. \tag{42}$$

The only difference between spatially implicit and explicit schemes is the structure of matrix $\mathbf{B}_d^{\boldsymbol{\Phi}}$. Since in the case of explicit schemes $\mathbf{b}_d$ is given by Eq.(40), $\mathbf{B}_d^{\boldsymbol{\Phi}}$ becomes an identity matrix. Therefore stability analysis discussed in Section 3 holds true for explicit schemes as well with the substitution $\mathbf{B}_d^{\boldsymbol{\Phi}} = \mathbf{I}$ in Eq.(22).

A similar framework exclusively for spatially explicit schemes has been presented in [1] and the optimized coefficients calculated using this framework match exactly with the ones obtained using Eq.(42). In [1], authors extensively discuss error characteristics and stability analysis for the optimized explicit schemes.

4.3 Biased schemes

In the discussion so far, we have assumed the domain to be periodic, which enables us to use central schemes at all grid points in the domain. Central schemes can not be used if the domain is not periodic or there are specified boundary conditions. This is because of the fact that there are unequal number of usable grid points on either side of a point which lies near boundary of the domain. Thus, we have to use *one-sided* or *biased* schemes near boundaries. A finite difference approximation is said to be biased if the number of grid points used on the left and right side of a point at which derivative is calculated is not equal. For the sake of discussion, in this special case we assume that $M_L^B = M_L^A = M_L$ and $M_R^B = M_R^A = M_R$. Therefore, Eq.(34) becomes

$$\sum_{m=-M_L}^{M_R} \mathrm{b}_m f_{i+m}^{(d)} = \frac{1}{(\Delta x)^d} \sum_{m=-M_L}^{M_R} \mathrm{a}_m f_{i+m} \; + \; O(\Delta x^{p+1}), \tag{43}$$

Thus, actual or true stencil size is $\bar{N} := M_L + M_R + 1$. The augmented parameters are $\hat{M} := \max\{M_L, M_R\}$ and $\hat{N} := 2\hat{M}+1$. Note that, a biased scheme is said to be *backward* finite difference if $M_L \neq 0$ and $M_R = 0$, and *forward* finite difference if $M_L = 0$ and $M_R \neq 0$.

Let $M_L > M_R$, and hence $\hat{M} = M_L$. We refer to this type of approximation as *left-biased*. This biased scheme can be treated as a central scheme with augmented stencil size $\hat{N} = 2M_L + 1$, and an extra imposed

constraint that the rightmost $(\hat{N}-\bar{N})$ coefficients to be zero. Therefore, we can use the central difference framework presented in Section 2 with $M = M_L$ and the following additional constraint

$$\underbrace{\begin{bmatrix} \boldsymbol{\delta}_n^T(\bar{N}+1) \\ \boldsymbol{\delta}_n^T(\bar{N}+2) \\ \vdots \\ \boldsymbol{\delta}_n^T(\hat{N}) \end{bmatrix}}_{\boldsymbol{\Delta}^T:=} \mathbf{a}_d = \underbrace{\begin{bmatrix} \boldsymbol{\delta}_n^T(\bar{N}+1) \\ \boldsymbol{\delta}_n^T(\bar{N}+2) \\ \vdots \\ \boldsymbol{\delta}_n^T(\hat{N}) \end{bmatrix}}_{\boldsymbol{\Delta}^T:=} \mathbf{b}_d = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{44}$$

for $n = \hat{N}$. Therefore, the constraint can be written as

$$\begin{bmatrix} \boldsymbol{\Delta}^T & \mathbf{0}_{(\hat{N}-\bar{N})\times\hat{N}} \\ \mathbf{0}_{(\hat{N}-\bar{N})\times\hat{N}} & \boldsymbol{\Delta}^T \end{bmatrix} \begin{bmatrix} \mathbf{a}_d \\ \mathbf{b}_d \end{bmatrix} = \mathbf{0}_{2(\hat{N}-\bar{N})\times 1}. \tag{45}$$

Now, we can write the analytical optimal solution for Eq.(15) subject to Eq.(45) as follows

$$\begin{pmatrix} \mathbf{a}_d^* \\ \mathbf{b}_d^* \\ \boldsymbol{\mu}_d^* \end{pmatrix} = \mathbf{P} \begin{bmatrix} \mathbf{0}_{2\hat{N}\times 1} \\ \mathbf{0}_{(d+p+1)\times 1} \\ 1 \\ \mathbf{0}_{2(\hat{N}-\bar{N})\times 1} \end{bmatrix}, \tag{46}$$

where,

$$\mathbf{P} = \begin{bmatrix} 2\mathbf{Q}_d & \begin{pmatrix} \mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{1\times\hat{N}} & \boldsymbol{\delta}_{\hat{N}}^T(\hat{M}+1) \\ \boldsymbol{\Delta}^T & \mathbf{0}_{(\hat{N}-\bar{N})\times\hat{N}} \\ \mathbf{0}_{(\hat{N}-\bar{N})\times\hat{N}} & \boldsymbol{\Delta}^T \end{pmatrix}^T \\ \begin{pmatrix} \mathbf{X}_d^T & -\mathbf{Y}_d^T \\ \mathbf{0}_{1\times\hat{N}} & \boldsymbol{\delta}_{\hat{N}}^T(\hat{M}+1) \\ \boldsymbol{\Delta}^T & \mathbf{0}_{(\hat{N}-\bar{N})\times\hat{N}} \\ \mathbf{0}_{(\hat{N}-\bar{N})\times\hat{N}} & \boldsymbol{\Delta}^T \end{pmatrix} & \mathbf{0}_{(2(\hat{N}-\bar{N})+d+p+2)\times(2(\hat{N}-\bar{N})+d+p+2)} \end{bmatrix}^{-1}.$$

Similarly, we can determine the optimal *right-biased* approximation for the case $M_L < M_R$ by setting the leftmost $(\hat{N}-\bar{N})$ coefficients to zero. It can be proved that the optimized coefficients of left-biased scheme $(\mathbf{a}_{d_L}^*, \mathbf{b}_{d_L}^*)$ with $M_L = M_1$, $M_R = M_2$ (where, $M_1 > M_2$) and its complementary right-biased scheme $(\mathbf{a}_{d_R}^*, \mathbf{b}_{d_R}^*)$ with $M_L = M_2$, $M_R = M_1$ satisfy

$$\begin{aligned} \mathbf{a}_{d_R}^* &= \mathbf{J}\mathbf{a}_{d_L}^*; \quad \mathbf{b}_{d_R}^* = \mathbf{J}\mathbf{b}_{d_L}^*, \quad \text{for even derivatives,} \\ \mathbf{a}_{d_R}^* &= -\mathbf{J}\mathbf{a}_{d_L}^*; \quad \mathbf{b}_{d_R}^* = \mathbf{J}\mathbf{b}_{d_L}^*, \quad \text{for odd derivatives,} \end{aligned}$$

where $\mathbf{J}$ is an anti-diagonal identity matrix of the appropriate dimension.

Fig.12 and Fig.13 compare modified wavenumbers of different implicit left-biased schemes ($M_L = 4, 5, 6$, red) with the implicit central scheme ($M = 3$, blue) of equal true stencil size ($\bar{N} = 2M+1 = 7$) for second and first derivatives respectively. Numerical values of the optimized coefficients are tabulated in Table 4 and 5.

From the middle row of Fig.12, we observe that the variation of the real component $\Re[e(\eta)]$ (shown in log scale) for biased schemes (red) is very close to that of the central scheme of same stencil size (blue).

In case of central scheme (blue), the imaginary component $\Im[e(\eta)]$ for even derivatives is zero, which is seen in the last row of Fig.12 (see Fig.3). And as expected, the imaginary component $\Im[e(\eta)]$ for biased schemes (red) which approximate even derivatives is not zero. This is due to the fact that the discussion in Appendix A does not hold for biased schemes as constraint Eq.(45) imposes asymmetry on the optimized

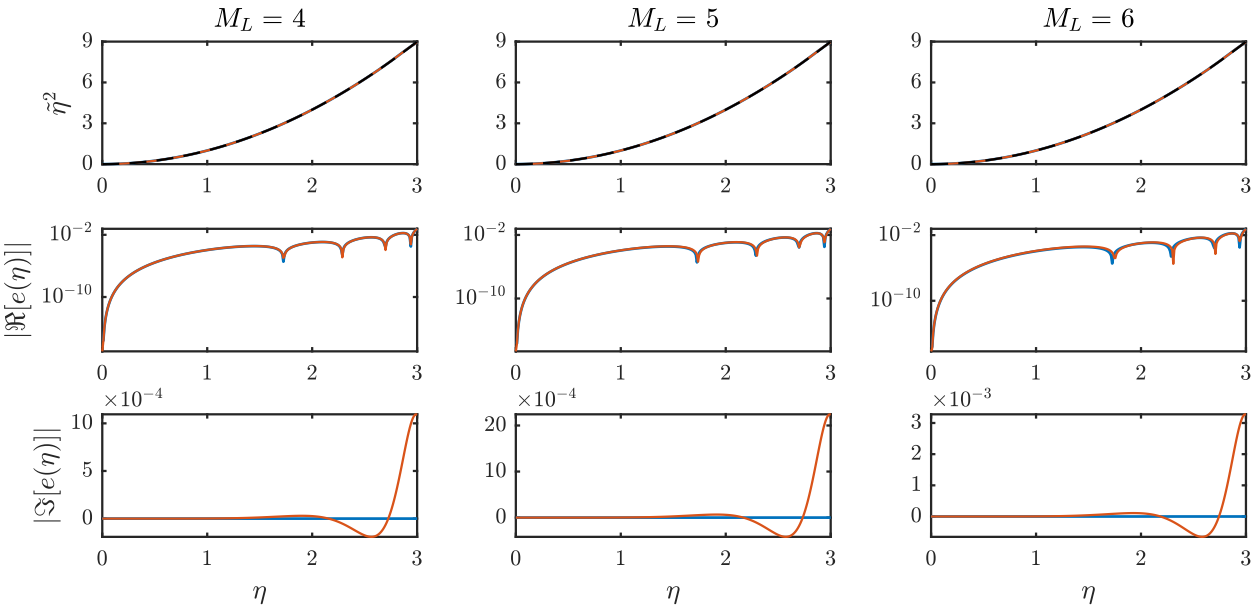

Figure 12: Modified wavenumber, and real and imaginary components of the spectral error for left-biased $\mathcal{O}^{M_L,M_R}_{M_L,M_R}(4)$ (red) approximating the second derivative. $\bar{N} = 7$, $\gamma(\eta) = 1$ for $\eta \in [0,3]$, and $\gamma(\eta) = 0$ otherwise.

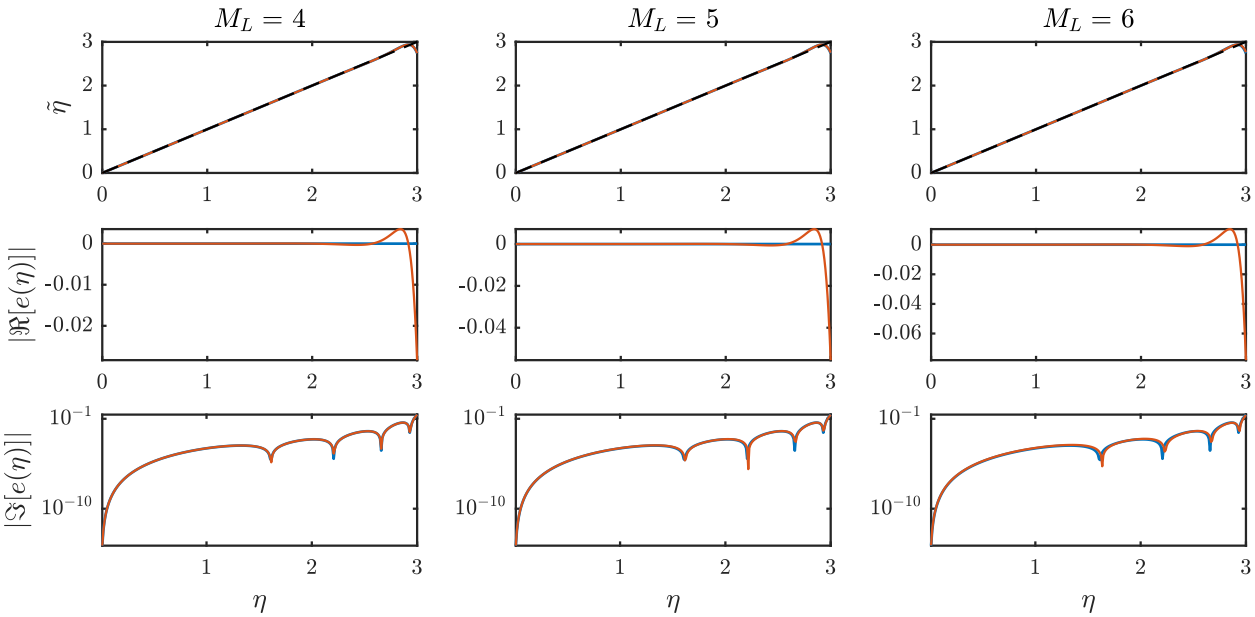

Figure 13: Modified wavenumber, and real and imaginary components of the spectral error for left-biased $\mathcal{O}^{M_L,M_R}_{M_L,M_R}(4)$ (red) approximating the first derivative. $\bar{N} = 7$, $\gamma(\eta) = 1$ for $\eta \in [0,3]$, and $\gamma(\eta) = 0$ otherwise.

coefficients. For fixed true stencil size, as M_L increases, $M_R = \bar{N} - (M_L + 1)$ decreases and schemes become more biased. It can be observed from the last row of Fig.12 that the imaginary component $\Im[e(\eta)]$ increases

with increase in biasedness of the scheme.

Similarly, the middle row of Fig.13 shows that unlike the central scheme (blue), real component $\Re[e(\eta)]$ for biased schemes (red) which approximate odd derivatives is not zero, and its magnitude increases with the increase in biasedness of the scheme. However, the variation of imaginary component $\Im[e(\eta)]$ (shown in log scale) for biased schemes (red) is very close to that of the central scheme (blue).

While solving PDEs numerically, we use a central scheme of stencil size $(2M+1)$ at $(N_p - 2M)$ points within the domain and biased schemes at M points adjacent to each one of left and right boundaries of the domain. Therefore, $(N_p - 2M)$ middle rows of $\mathbf{A}_d^{\mathbf{\Phi}}$ and $\mathbf{B}_d^{\mathbf{\Phi}}$ defined in Eq.(21) contain the same central scheme coefficients, last M rows contain left-biased scheme coefficients and first M rows contain complementary right-biased scheme coefficients. Semi-discrete stability discussed in Section 3.1 is not valid for biased approximations since coefficients do not satisfy the properties described in Appendix A. However, Eq.(22) and Eq.(23) hold true for non-periodic discretization. Therefore, fully discretized stability analysis presented in Section 3.2.2 is valid for this case.

Special cases discussed in this section are a few examples which demonstrate how to derive optimized finite difference schemes with a desired structure by imposing suitable constraints on the coefficients. Since all constraints are linear, one can impose any number and combination of the constraints given by Eq.(36), Eq.(41) and Eq.(44) to obtain the desired schemes. For example, we can impose constraints in Eq.(41) and Eq.(44) simultaneously and solve the resulting optimization problem to derive biased schemes which are spatially explicit. However, to ensure the invertibility of the matrix $\mathbf{P}$, such as in Eq.(46), duplicate constraints should be removed while constructing the KKT condition.

5 Numerical Results

In this section we use the optimized schemes derived in preceding sections to solve some PDEs numerically and compare the results with the standard schemes of the same stencil sizes. In particular, we compare the numerical solutions obtained using the optimized scheme $\boldsymbol{\mathcal{O}}_2^3(4)$ (see Table 3) with the standard scheme $\boldsymbol{\mathcal{S}}_2^3(10)$.

Initial condition for solving PDEs is taken to be the superposition of multiple sinusoidal waves with wavenumbers k, given by Eq.(47). $A(k)$ is amplitude of the wave corresponding to wavenumber k, and $\phi_k \in [0, 2\pi]$ is the corresponding phase initialized randomly. We use the same initial condition for solving different PDEs in sections to follow.

$$f_0(x) := f(x, t=0) = \sum_k A(k) \sin(kx + \phi_k). \tag{47}$$

5.1 Advection-diffusion equation

Consider the following simplified version of Eq.(20), where coefficients of only first two spatial derivatives are non-zero.

$$\frac{\partial f}{\partial t} = \beta_1 \frac{\partial f}{\partial x} + \beta_2 \frac{\partial^2 f}{\partial x^2}. \tag{48}$$

To satisfy Eq.(27), we have the diffusivity $\beta_2 > 0$. Analytical solution for Eq.(48) is known and given by

$$f_t(x) := f(x, t>0) = \sum_k \exp(-\beta_2 k^2 t) A(k) \sin(k(x + \beta_1 t) + \phi_k), \tag{49}$$

In this case, the amplitude decays with time due to non-zero β_2 and the phase changes due to non-zero β_1.

We used CFL number to express Δt and Δx as a normalized quantity. Similarly, we define normalized time with respect to d^{th} derivative as $t_d^* := |\beta_d| t k_{\max}^d$, where $k_{\max}$ is the largest wavenumber present in the initial condition Eq.(47). Eq.(48) is solved numerically using the optimized and standard spatial finite

difference schemes. For a fixed diffusive CFL (r_2), we have used matching order explicit Runge-Kutta (ERK) schemes for the temporal discretization, i.e for spatial schemes which are 4^{th} and 10^{th} order accurate, we have used 2^{nd} and 5^{th} order ERK for the temporal discretization respectively.

The spectral energy at wavenumber k at time t is defined to be $|\hat{f}_t(k)|^2$, where $\hat{f}_t(k)$ is the discrete Fourier coefficient corresponding to wavenumber k at time t. To assess the dissipative numerical errors at different wavenumbers, we plot normalized error in spectral energy of the numerical solution with respect to the analytical solution in Fig.14. Note, $\hat{\tilde{f}}_t(k)$ is the Fourier coefficient calculated from the numerical solution $\tilde{f}_t(k)$. These results are obtained for the diffusive CFL number $r_2 = 0.01$ at $t_2^* \approx 25$.

The error is small for the standard pentadiagonal scheme, $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta), in low wavenumber region. However, it demonstrates steep variation of error as the wavenumber increases. On the other hand, the optimized scheme $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) shows relatively slower variation of error with increasing wavenumber. Again, this behavior is expected due to equally weighted wavenumbers ($\gamma(\eta) = 1$) in the optimization. $\boldsymbol{\mathcal{O}}_2^3(4)$ has larger error than $\boldsymbol{\mathcal{S}}_2^3(10)$ at lower wavenumbers. But at higher wavenumbers, the 4^{th} order accurate $\boldsymbol{\mathcal{O}}_2^3(4)$ demonstrates much better accuracy than the 10^{th} order accurate $\boldsymbol{\mathcal{S}}_2^3(10)$.

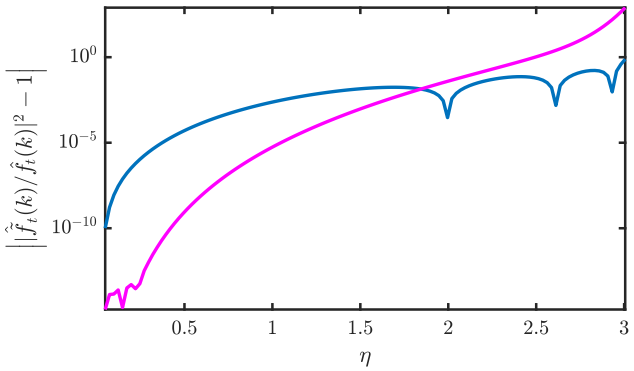

Figure 14: Diffusion error for advection-diffusion equation: $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) and $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta). $r_1 \approx r_2 = 0.01$, $t_2^* \approx 25$, $A(k) = 1$, $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

While Fig.14 shows numerical dissipation error, the dispersion error is shown in Fig.15. Dispersion error is quantified by the calculation of numerical speed, $\tilde{c}(k) := \arg[\hat{\tilde{f}}_t(k)/\hat{f}_0(k)]/kt$. The analytical speed is $c := \beta_1$ for all wavenumbers. Normalized numerical speed, $c^*(k) := \tilde{c}(k)/c$, is unity when there is no dispersion error.

The observations made for Fig.14 are consistent with Fig.15 as well. From Fig.15, it is evident that the normalized numerical speed for the standard schemes $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta), deviates from unity at lower wavenumbers leading to large dispersion error at higher wavenumbers. The normalized numerical speed for the optimized scheme $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) is close to unity even at higher wavenumbers. As in the case of dissipative error, here also we observe that the standard scheme shows steep variation of the dispersion error, while the optimized scheme is relatively slower in variation.

5.2 Non-linear advection-diffusion equation

Now let us consider the non-linear advection-diffusion equation which is also known as Burgers' equation and widely studied in various areas of applied mathematics.

$$\frac{\partial f}{\partial t} = -f\frac{\partial f}{\partial x} + \beta_2 \frac{\partial^2 f}{\partial x^2}. \tag{50}$$

Because of the non-linear term, different Fourier modes interact with each other and produce new wavenumbers in the solution. Therefore, to ensure that all wavenumbers are well resolved, we performed grid convergence study and observed that space-averaged energy ($K := \sum_{i=1}^{N_p} f_i^2(x,t)/N_p$) becomes independent of the number of grid points in the domain at $N_p = 256$.

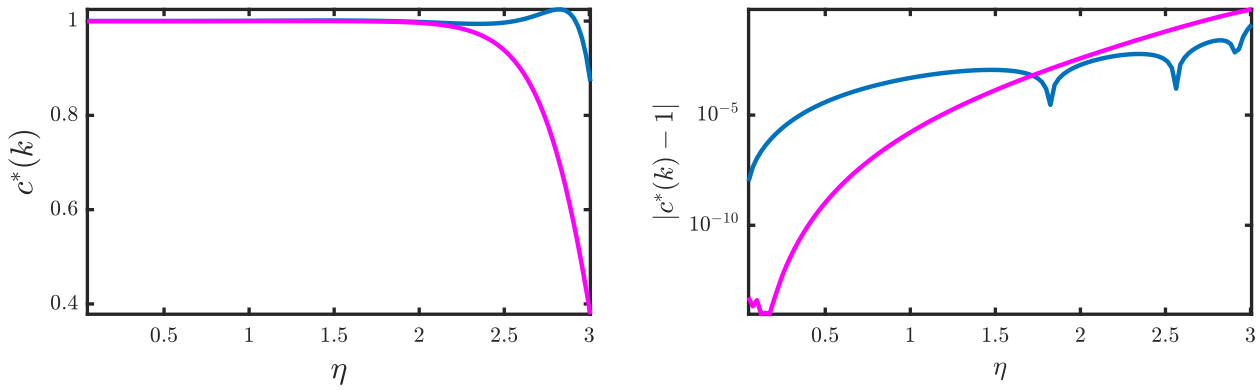

Figure 15: Dispersion error for advection-diffusion equation: $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) and $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta). $r_1 \approx r_2 = 0.01$, $t_2^* \approx 25$, $A(k) = 1$, $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

Numerical solution of Eq.(50) is obtained in a periodic domain $x \in [0, 2\pi]$ with initial condition given by Eq.(47), where $A(k) = k^{-1/2}$. The negative exponent ensures that higher wavenumbers have sufficient energy content without being unstable. For this particular PDE, we define timescale as $t = K/\epsilon$, where ϵ is space averaged energy dissipation rate ($\epsilon := \sum_{i=1}^{N_p} (\partial f_i(x,t)/\partial x)^2/N_p$). Then time is expressed as a normalized quantity $t^* := t/t_0$ where $t_0 = K_0/\epsilon_0$ is the initial timescale.

The analytical solution of Eq.(50) is obtained by applying Cole-Hopf transformation as follows [17, 18]. First, let us define the transformation

$$f = -2\beta_2 \frac{1}{\phi} \frac{\partial \phi}{\partial x}.$$

With this substitution, Eq.(50) becomes diffusion equation and the analytical solution for $f(x,t)$ is given by

$$f_t(x) := f(x,t) = \frac{\int_{-\infty}^{\infty} \frac{x-y}{t} \phi(y,0) \exp\left(\frac{(x-y)^2}{4\beta_2 t}\right) dy}{\int_{-\infty}^{\infty} \phi(y,0) \exp\left(\frac{(x-y)^2}{4\beta_2 t}\right) dy}, \tag{51}$$

where initial $\phi(y,0)$ is obtained using

$$\phi(y,0) = \exp\Big(- \int_0^y \frac{f(z,0)}{2\beta_2} dz \Big).$$

Eq.(51) is the exact solution for non-linear advection-diffusion equation. However, integrals involved are computed numerically.

Fig.16 shows the spectral energy content normalized with initial energy for different wavenumbers. All numerical results for this PDE are obtained for $\beta_2 = 0.04$, $r_2 = 0.01$ at $t^* \approx 148$. It can be clearly seen that the spectral energy content for the standard schemes $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta) deviates from analytical solution (black-dashed) at high wavenumbers. In the same wavenumber region, the optimized scheme $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) is relatively closer to the analytical solution.

Similar to Fig.14, we show the normalized error in spectral energy for the non-linear advection-diffusion equation in Fig.17. Due to the interaction of Fourier modes with each other, in general, we do not expect the spectral behavior of different schemes for non-linear PDE to be in exact agreement with the behavior observed for the linear case. Unlike the linear case (Fig.14), Fig.17 shows that the optimized schemes $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) has better accuracy than $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta) in both low and high wavenumber region.

Because of the non-linearity, there is no well-defined analytical speed in this case and therefore, the notion of numerical speed can not be used to quantify the dispersion error. However, we can calculate and compare

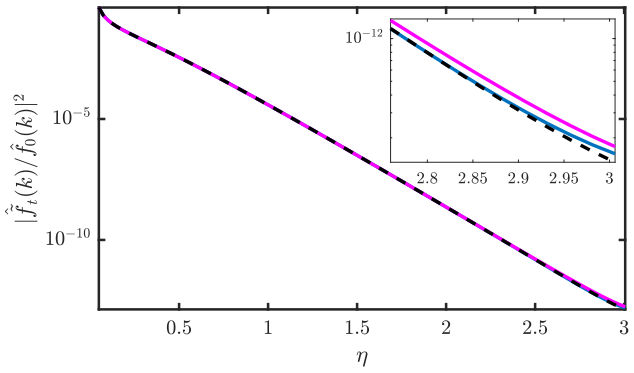

Figure 16: Spectral energy content for non-linear advection-diffusion equation normalized with initial energy: Analytical solution (black-dashed), $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) and $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta). $\beta_2 = 0.04$, $r_2 = 0.01$, $A(k) = k^{-1/2}$, $t^* \approx 148$, $k_{\max} = 121$, $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

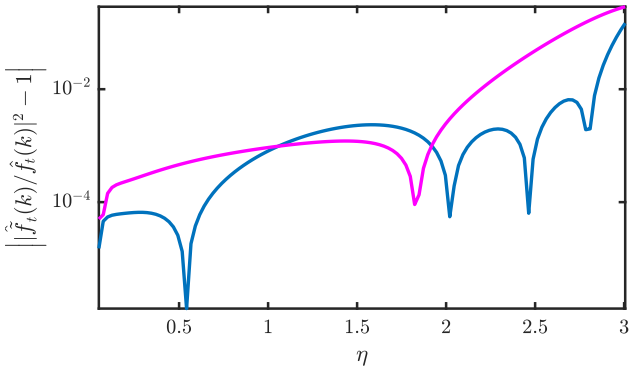

Figure 17: Diffusion error for non-linear advection-diffusion equation: $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) and $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta). $\beta_2 = 0.04$, $r_2 = 0.01$, $A(k) = k^{-1/2}$, $t^* \approx 148$, $k_{\max} = 121$, $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

the argument of Fourier coefficients, $\hat{\theta}_k := \arg[\hat{f}_t(k)]$, for different wavenumbers. The normalized error in $\hat{\theta}_k$ is shown in Fig.18 where $\hat{\tilde{\theta}}_k$ is the argument of the numerical solution. Clearly, the optimized scheme $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) has smaller phase error than the standard scheme $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta) across all wavenumbers. Whereas, in the linear case (Fig.15) $\boldsymbol{\mathcal{S}}_2^3(10)$ demonstrates better accuracy than the optimized scheme in the low wavenumber region.

Fig.17 and Fig.18 independently quantify the errors in $|\hat{f}_t(k)|$ and $\hat{\theta}_k$. Fig.19 accounts for $|\hat{\tilde{f}}_t(k) - \hat{f}_t(k)|$, which quantifies the combined effect of error in $|\hat{f}_t(k)|$ and $\hat{\theta}_k$ in a single plot. When we look at the combined error in the non-linear case, it is observed from Fig.19 that the optimized scheme $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) has better accuracy than the standard schemes $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta) at all wavenumbers. Thus, for solving practical physical problems which are governed by non-linear PDEs, the optimized scheme seem to provide much better spectral resolution than the non-optimized standard scheme.

6 Conclusions

In this paper, we presented a generalized unified framework to derive compact optimized schemes. The optimized coefficients are determined analytically by solving a quadratic program subject to linear equality

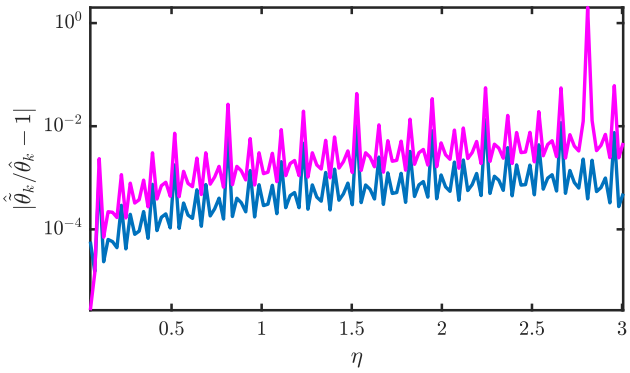

Figure 18: Phase error for non-linear advection-diffusion equation: $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) and $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta). $\beta_2 = 0.04$, $r_2 = 0.01$, $A(k) = k^{-1/2}$, $t^* \approx 148$, $k_{\max} = 121$, $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

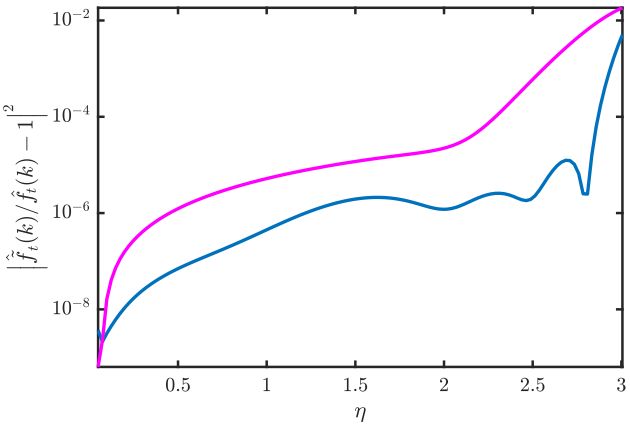

Figure 19: Amplitude error for non-linear advection-diffusion equation: $\boldsymbol{\mathcal{O}}_2^3(4)$ (blue) and $\boldsymbol{\mathcal{S}}_2^3(10)$ (magenta). $\beta_2 = 0.04$, $r_2 = 0.01$, $A(k) = k^{-1/2}$, $t^* \approx 148$, $k_{\max} = 121$, $\gamma(\eta) = 1$ for $\eta \in [0, 3]$, and $\gamma(\eta) = 0$ otherwise.

constraints to minimize an upper bound on $\mathcal{L}_2$ norm of the spectral error. We have shown that for a given order of accuracy, we can increase the stencil size and hence the number of degrees of freedom in the optimization to achieve better spectral resolution. The freedom of selecting weighting function ($\gamma(\eta)$) allows us to achieve better resolution of the wavenumbers which are important to the physics of problem by giving them more weight in the optimization problem.

We have shown that the special cases, namely, central schemes with unequal LHS and RHS stencil sizes (e.g. pentadiagonal), explicit, and biased schemes can be derived by imposing additional linear constraints in the optimization problem. Using these three special cases as building blocks, an optimized scheme with any desired structure can be derived as all constraints are linear and therefore, they can be imposed in any number and combination.

We have also presented a rigorous stability analysis for a generalized implicit RK scheme. We showed that by increasing the stencil size the spectral resolution can be improved, but simultaneously it results in reduced stability region in $r_1 - r_2$ space.

The numerical results show that the optimized schemes show better accuracy than the standard scheme of the same stencil size with larger formal order of accuracy in the high wavenumber region. The better resolution in high wavenumber region is achieved at the expense of relatively larger error at the lower

wavenumbers, because $\gamma(\eta) = 1$ was used in the simulations which weights all wavenumbers equally. However, in the case of non-linear PDE, we observed that the optimized scheme shows overall better accuracy than the standard scheme across the wavenumber spectrum.

Empirical observations show that KKT matrix used in determining the optimized coefficients suffer from rank deficiency at certain stencil size (M). However, explicit schemes are unaffected by this problem. This issue, perhaps, hints at the fundamental limitation of implicit schemes to scale up beyond certain stencil size, and it will be investigated in our future studies.

Although our framework allows us to derive biased schemes, for the brevity of discussion we have restricted numerical results presented in this work on periodic domains. Also, implementation of boundary conditions without compromising the accuracy of numerical solution is an important area of research and will be addressed in our future work.

A (Skew-)Symmetry of optimized coefficients for central differences

Even derivatives:

The optimized coefficients both $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$ are symmetric, and imaginary component of the spectral error $e(\eta)$ is zero.

In Eq.(15), the cost function is given by

$$\left\langle \left(\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{C}^T(\eta)\mathbf{b}_d\right)^2 + \left(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{S}^T(\eta)\mathbf{b}_d\right)^2 \right\rangle. \tag{52}$$

Now, let us consider the integrand function in Eq.(52)

$$f(\mathbf{a}_d, \mathbf{b}_d) = \gamma(\eta)\left[\left(\mathbf{C}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{C}^T(\eta)\mathbf{b}_d\right)^2 + \left(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{S}^T(\eta)\mathbf{b}_d\right)^2\right]. \tag{53}$$

After substituting $(\mathbf{J}\mathbf{a}_d, \mathbf{J}\mathbf{b}_d)$ in Eq.(53) we get,

$$f(\mathbf{J}\mathbf{a}_d, \mathbf{J}\mathbf{b}_d) = \gamma(\eta)\left[\left(\mathbf{C}^T(\eta)\mathbf{J}\mathbf{a}_d - (-1)^q\eta^d\mathbf{C}^T(\eta)\mathbf{J}\mathbf{b}_d\right)^2 + \left(\mathbf{S}^T(\eta)\mathbf{J}\mathbf{a}_d - (-1)^q\eta^d\mathbf{S}^T(\eta)\mathbf{J}\mathbf{b}_d\right)^2\right].$$

Clearly from Eq.(12),

$$f(\mathbf{J}\mathbf{a}_d, \mathbf{J}\mathbf{b}_d) = f(\mathbf{a}_d, \mathbf{b}_d). \tag{54}$$

This is a property of the function $f(\mathbf{a}_d, \mathbf{b}_d)$ and holds true for any admissible pair $(\mathbf{a}_d, \mathbf{b}_d)$. Let $(\mathbf{a}_d^*, \mathbf{b}_d^*)$ be an optimal pair which minimizes the cost function. It immediately follows from Eq.(54),

$$f(\mathbf{J}\mathbf{a}_d^*, \mathbf{J}\mathbf{b}_d^*) = f(\mathbf{a}_d^*, \mathbf{b}_d^*).$$

For $(\mathbf{a}_d^*, \mathbf{b}_d^*)$ being a unique minimizer, the following must be true,

$$\mathbf{J}\mathbf{a}_d^* = \mathbf{a}_d^* \quad \text{and} \quad \mathbf{J}\mathbf{b}_d^* = \mathbf{b}_d^*. \tag{55}$$

By Definition 1, both $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$ are symmetric. To show that $(\mathbf{a}_d^*, \mathbf{b}_d^*)$ that satisfies Eq.(55) is also a feasible solution for the order of accuracy constraints, consider Eq.(3),

$$\mathbf{a}_d^T\mathbf{X}_d - \mathbf{b}_d^T\mathbf{Y}_d = \mathbf{0}_{1\times(d+p+1)}. \tag{56}$$

Substitute $(\mathbf{a}_d, \mathbf{b}_d)$ with $(\mathbf{J}\mathbf{a}_d, \mathbf{J}\mathbf{b}_d)$ to get,

$$\mathbf{a}_d^T\mathbf{J}^T\mathbf{X}_d - \mathbf{b}_d^T\mathbf{J}^T\mathbf{Y}_d = \mathbf{0}_{1\times(d+p+1)}.$$

Note that, $\mathbf{J}^T = \mathbf{J}$, and it operates on the columns of $\mathbf{X}_d$ and $\mathbf{Y}_d$. Let $\mathbf{X}_i$ and $\mathbf{Y}_i$ be i^{th} columns of $\mathbf{X}_d$ and $\mathbf{Y}_d$. From equations Eq.(4) and Eq.(5), it is evident that $\mathbf{X}_i$ and $\mathbf{Y}_i$ are symmetric vectors for an odd i. It implies that $\mathbf{a}_d^T\mathbf{J}\mathbf{X}_i = \mathbf{a}_d^T\mathbf{X}_i$ and $\mathbf{b}_d^T\mathbf{J}\mathbf{Y}_i = \mathbf{b}_d^T\mathbf{Y}_i$. Clearly, constraint Eq.(56) is satisfied for odd columns of $\mathbf{X}_d$ and $\mathbf{Y}_d$. Similarly, for an even i, it can be shown that $\mathbf{a}_d^T\mathbf{J}\mathbf{X}_i = -\mathbf{a}_d^T\mathbf{X}_i$ and $\mathbf{b}_d^T\mathbf{J}\mathbf{Y}_i = -\mathbf{b}_d^T\mathbf{Y}_i$, which again satisfies Eq.(56) since right hand side is zero. Hence, it can be concluded that, if $(\mathbf{a}_d, \mathbf{b}_d)$ satisfies the order constraint, then $(\mathbf{J}\mathbf{a}_d, \mathbf{J}\mathbf{b}_d)$ also satisfies the constraint Eq.(56). Therefore, $(\mathbf{a}_d^*, \mathbf{b}_d^*)$ which satisfy Eq.(55) is a feasible solution for Eq.(56).

Recalling that symmetric and skew-symmetric vectors are orthogonal, we get,

$$\mathbf{S}^T(\eta)\mathbf{a}_d^* = 0 \quad \text{and} \quad \mathbf{S}^T(\eta)\mathbf{b}_d^* = 0. \tag{57}$$

Thus, spectral error from Eq.(9) becomes, $e(\eta) = \frac{\mathbf{C}^T(\eta)\mathbf{a}_d^*}{\mathbf{C}^T(\eta)\mathbf{b}_d^*} - (-1)^q\eta^d$. Therefore, for even derivatives, the optimized coefficients make imaginary component of the spectral error $e(\eta)$ zero.

Odd derivatives:

The optimized coefficients $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$ are respectively skew-symmetric and symmetric, and real component of the spectral error is zero.

For odd derivatives, the integrand function is

$$g(\mathbf{a}_d, \mathbf{b}_d) = \gamma(\eta)\left[\left(\mathbf{C}^T(\eta)\mathbf{a}_d + (-1)^q\eta^d\mathbf{S}^T(\eta)\mathbf{b}_d\right)^2 + \left(\mathbf{S}^T(\eta)\mathbf{a}_d - (-1)^q\eta^d\mathbf{C}^T(\eta)\mathbf{b}_d\right)^2\right].$$

Using similar arguments as for the case of even derivatives, it can be easily shown that,

$$g(-\mathbf{J}\mathbf{a}_d, \mathbf{J}\mathbf{b}_d) = g(\mathbf{a}_d, \mathbf{b}_d).$$

Uniqueness of the optimal solution requires that

$$-\mathbf{J}\mathbf{a}_d^* = \mathbf{a}_d^* \quad \text{and} \quad \mathbf{J}\mathbf{b}_d^* = \mathbf{b}_d^*. \tag{58}$$

By definition, $\mathbf{a}_d^*$ and $\mathbf{b}_d^*$ are respectively skew-symmetric and symmetric. To establish the feasibility of such a solution for the order of accuracy constraints, consider Eq.(3) and substitute $(\mathbf{a}_d, \mathbf{b}_d)$ with $(-\mathbf{J}\mathbf{a}_d, \mathbf{J}\mathbf{b}_d)$. Then, using similar arguments as for the case of even derivatives, it can be shown that if $(\mathbf{a}_d, \mathbf{b}_d)$ satisfies Eq.(3), then $(-\mathbf{J}\mathbf{a}_d, \mathbf{J}\mathbf{b}_d)$ also satisfies the same constraint for odd d. Therefore, $(\mathbf{a}_d^*, \mathbf{b}_d^*)$ which satisfy Eq.(58) is a feasible solution for Eq.(3).

Using the orthogonality property,

$$\mathbf{C}^T(\eta)\mathbf{a}_d^* = 0 \quad \text{and} \quad \mathbf{S}^T(\eta)\mathbf{b}_d^* = 0. \tag{59}$$

The spectral error follows from Eq.(9) as $e(\eta) = j\left(\frac{\mathbf{S}^T(\eta)\mathbf{a}_d^*}{\mathbf{C}^T(\eta)\mathbf{b}_d^*} - (-1)^q\eta^d\right)$, indicating that the real component is zero.

B Optimized coefficients

	$M=1$	$M=2$	$M=3$	$M=4$
a_0^*	-2.4	-1.55920152194026	-0.979288292571078	-0.719422653838933
a_1^*	1.2	0.396897309677732	-0.033306701818875	-0.156640799785708
a_2^*		0.382703451292396	0.440495791275238	0.340037669820826
a_3^*			0.0824550568291757	0.161702448995973
a_4^*				0.0146120078883761
b_0^*	1	1	1	1
b_1^*	0.1	0.437358728499431	0.607804000534683	0.69537501810989
b_2^*		0.0264968289242269	0.122983617052232	0.223008838348136
b_3^*			0.00459836978541528	0.0272452165890212
b_4^*				0.000682950290635715

Table 1: Coefficients for the second derivative corresponding to Fig.1. Note that $a_{-m}^* = a_m^*$ and $b_{-m}^* = b_m^*$.

	$M=1$	$M=2$	$M=3$	$M=4$
a_0^*	0	0	0	0
a_1^*	0.75	0.682194069313335	0.560054939856331	0.472419664132013
a_2^*		0.214144479273011	0.326746645436286	0.367572867069987
a_3^*			0.0418602478971568	0.0980340659498803
a_4^*				0.00699750157631073
b_0^*	1	1	1	1
b_1^*	0.25	0.547827381201651	0.658367308183134	0.72407136413065
b_2^*		0.0626556466577058	0.170094141092335	0.26326428439027
b_3^*			0.0106675251449049	0.0407179433494389
b_4^*				0.00160401055651088

Table 2: Coefficients for the first derivative corresponding to Fig.4. Note that $a_{-m}^* = -a_m^*$ and $b_{-m}^* = b_m^*$.

	$d=1$	$d=2$
a_0^*	0	-1.22993292260472
a_1^*	0.630815603923759	0.130644921343958
a_2^*	0.273862907724336	0.459462898620059
a_3^*	0.00849109849909385	0.0248586413383436
b_0^*	1	1
b_1^*	0.599688672582508	0.531049490588671
b_2^*	0.104326042287205	0.0650626533459724
b_3^*	0	0

Table 3: Coefficients for $\boldsymbol{\mathcal{O}}_2^3(4)$ approximating the first and second derivatives. Note that $b_{-m}^* = b_m^*$, and $a_{-m}^* = a_m^*$ for $d=2$, and $a_{-m}^* = -a_m^*$ for $d=1$.

Note that, in Table 4 and 5, $^{\#}$ denotes coefficient of the grid point at which derivative is to be approximated.

	$M_L=4$	$M_L=5$	$M_L=6$
	0.135141927552199	0.662304984252634	17.3670624080996
	0.722707534591416	3.53558092392018	92.3021288782114
	-0.0524729395599328	-0.250337371007728	-6.49624407257855
$\mathbf{a}^*$	-1.60731259496048	-7.85976328865994	-204.957850510534
	$^{\#}$-0.0583231083517459	-0.310357190338834	-8.85074823578218
	0.72408652155438	$^{\#}$3.54961920557169	92.8470054538251
	0.13617265917416	0.672952736261999	$^{\#}$17.7886460787584
	0.0075365800956553	0.0369784406883765	0.971865865002116
	0.201615926044887	0.987464737949076	25.8562746553017
	0.99713483322273	4.88096076338054	127.576933509124
$\mathbf{b}^*$	1.64238167997833	8.04522660766854	210.335351055833
	$^{\#}$1	4.91039583089326	128.733779558004
	0.202831788738001	$^{\#}$1	26.3512560574889
	0.00760492052475932	0.0376863400465079	$^{\#}$1

Table 4: Coefficients for the left-biased schemes which approximate second derivative corresponding to Fig.12. Corresponding central scheme is $M=3$ from Table 1.

	$M_L = 4$	$M_L = 5$	$M_L = 6$
	-0.0621972998530267	-0.232619531619737	-3.52690296300194
	-0.488868232296276	-1.83780260017807	-27.9116587382461
	-0.846178148476397	-3.2098376453028	-49.0922987950105
$\mathbf{a}^*$	-0.00918720281492579	-0.0700718215089091	-1.57497223752058
	$^{\#}$0.844294225659752	3.19534746352695	48.7613719684532
	0.497811414459153	$^{\#}$1.90604922449128	29.4470383302118
	0.0643252433217213	0.248934910591297	$^{\#}$3.89742243511407
	0.0158345798757476	0.0591989978487381	0.898171816719291
	0.253744537333521	0.951710004840582	14.4387550915639
	0.987857487221426	3.72365571735861	56.675421516294
$\mathbf{b}^*$	1.50948607590456	5.7229905302476	87.5412499132427
	$^{\#}$1	3.81632520991078	58.7608049690781
	0.260059400465721	$^{\#}$1	15.5296377878293
	0.016417216370275	0.0636716245952377	$^{\#}$1

Table 5: Coefficients for the left-biased schemes which approximate first derivative corresponding to Fig.13. Corresponding central scheme is $M = 3$ from Table 2.

C Butcher tableaux of RK schemes

1. Forward Euler method (FE)

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$$

   Following RK schemes are taken from [16].

2. Explicit four stage RK method (ERK4)

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array}$$

3. Implicit two stage RK method (IRK2)

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 2/3 & 1/3 & 1/3 \\ \hline & 1/4 & 3/4 \end{array}$$

4. Implicit three stage RK method (IRK3)

$$\begin{array}{c|ccc} 0.158984 & 0.158984 & 0 & 0 \\ 0.579492 & 0.420508 & 0.158984 & 0 \\ 1 & 0.348023 & 0.492993 & 0.158984 \\ \hline & 0.348022 & 0.492994 & 0.158984 \end{array}$$

Acknowledgements

Funding: This work was supported by the National Science Foundation [grant numbers 1762825, 1439145].

References

[1] Komal Kumari, Raktim Bhattacharya, and Diego A. Donzis. A unified approach for deriving optimal finite differences. *Journal of Computational Physics*, 399, 2019.

[2] Sanjiva K Lele. Compact finite difference schemes with spectral-like resolution. *Journal of Computational Physics*, 103(1):16–42, 1992.

[3] Jae Wook Kim and Duck Too Lee. Optimized compact finite difference schemes with maximum resolution. *AIAA Journal*, 34(5):887–893, 1996.

[4] Jae Wook Kim. Optimised boundary compact finite difference schemes for computational aeroacoustics. *Journal of Computational Physics*, 225(1):995–1019, 2007.

[5] Graham Ashcroft and Xin Zhang. Optimized prefactored compact schemes. *Journal of Computational Physics*, 190(2):459–477, 2003.

[6] Hongbo Zhou and Guanquan Zhang. Prefactored optimized compact finite-difference schemes for second spatial derivatives. *GEOPHYSICS*, 76(5):WB87–WB95, 2011.

[7] Zhanxin Liu, Qibai Huang, Zhigao Zhao, and Jixuan Yuan. Optimized compact finite difference schemes with high accuracy and maximum resolution. *International Journal of Aeroacoustics*, 7:123–146, 2008.

[8] A. Rona, I. Spisso, E. Hall, M. Bernardini, and S. Pirozzoli. Optimised prefactored compact schemes for linear wave propagation phenomena. *Journal of Computational Physics*, 328:66–85, 2017.

[9] C. K. Tam and J. C. Webb. Dispersion-relation-preserving finite difference schemes for computational acoustics. *Journal of Computational Physics*, 107:262–281, 1993.

[10] M. Zhuang and R. F. Chen. Applications of high-order optimized upwind schemes for computational aeroacoustics. *AIAA Journal*, 40(3):443–449, 2002.

[11] C. Bogey and C. Bailly. A family of low dispersive and low dissipative explicit schemes for flow and noise computations. *Journal of Computational Physics*, 194:194–214, 2004.

[12] Jin-Hai Zhang and Zhen-Xing Yao. Optimized explicit finite-difference schemes for spatial derivatives using maximum norm. *Journal of Computational Physics*, 250:511–526, 2013.

[13] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[14] Nicholas I. M. Gould, Mary E. Hribar, and Jorge Nocedal. On the solution of equality constrained quadratic programming problems arising in optimization. *SIAM Journal on Scientific Computing*, 23(4):1376–1395, 2001.

[15] Michael Grant, Stephen Boyd, and Yinyu Ye. Cvx: Matlab software for disciplined convex programming, 2008.

[16] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons Ltd., 2008.

[17] Hopf Eberhard. The partial differential equation $u_t + uu_x = \mu u_{xx}$. *Communications on Pure and Applied Mathematics*, 3(3):201–230, 1950.

[18] Julian D. Cole. On a quasi-linear parabolic equation occurring in aerodynamics. *Quarterly of Applied Mathematics*, 9(3):225–236, 1951.