ON SOME PROBLEMS IN SPARSE HYBRID IMAGING, NON-STANDARD FINITE DIFFERENCE METHODS, AND FOKKER-PLANCK FRAMEWORKS IN ESOPHAGEAL CANCER

by

MADHU GUPTA

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2022

ACKNOWLEDGEMENTS

First, I would like to express my deepest gratitude to my advisor Dr. Souvik Roy for his mentor ship, advises, and encouragement that have shaped and improved my ability to do research, and to excel in both my academics and research.

Words cannot express my gratitude to my co-advisor Dr. Hristo Kojouharov. He has been an amazing mentor throughout my entire graduate experience and I cannot thank him enough.

I would like to extend my sincere thanks to Dr. Gaik Ambartsoumian and Dr. Suvra Pal, for their interest in my research and for taking time to serve on my comprehensive committee and dissertation committee.

This endeavor would not have been possible without The University of Texas at Arlington (UTA) for financially supporting my PhD graduate program and The National Cancer Institute, National Institutes of Health, Grant Number: R21CA242933-01 for partly support my research on current density impedance imaging.

I would like to thank Dr. Jianzhong Su, Dr. Pedro Maia, Dr. Barbara Shipman, Dr. Michaela Vancliff, Dr. Chaoqun Liu, Dr. David Jorgensen, Mrs. Lona Donnelly, Mrs. Shanna Banda, Mr. Michael Schmidtkunz, and Mrs. Libby Carrol for all of their support and assistance.

It was a very nice experience of working with Fawaz Alalhareth. I am grateful to have amazing friends Rahul, Nitesh, Anuj, Sakshi, Shruti, Saumya, Fatma, Mugdha, Rodina, Shreshtha, Zach, Julio, Jose and many others.

Respect to my in laws Mr. R. C. Mishra and Mrs. Geeta Mishra and love to Tanu, Anurag, Prachi, Ayushi and Anshu.

I would like to express my deepest gratitude to my late father Mr. Vijay Kumar Gupta and my mother Mrs. Anita Gupta, who always believed in my ability to be successful in the academic arena. I would like to extend my gratitude to my late brother Pawan. Special thanks to my younger brother Yash for taking all the responsibilities in this young age, so that I could fulfill my dream.

Finally, I would like to thank my husband Dr. Rohit Kumar Mishra. Without his constant love and support I would have not been able to come this far.

August 1, 2022

ABSTRACT

ON SOME PROBLEMS IN SPARSE HYBRID IMAGING, NON-STANDARD FINITE DIFFERENCE METHODS, AND FOKKER-PLANCK FRAMEWORKS IN ESOPHAGEAL CANCER

Madhu Gupta, Ph.D.

The University of Texas at Arlington, 2022

Supervising Professor: Dr. Souvik Roy and Dr. Hristo V. Kojouharov

In this thesis, we first discuss nonlinear optimization frameworks for the sparsity-based nonlinear reconstruction of parameters in hybrid imaging modalities such as current density impedance imaging (CDII) and two-photon photoacoustic computed tomography (2P-PACT). The framework comprises minimizing an objective functional involving a least square fit and some regularization terms that promote sparsity patterns and enhance the edges to facilitate high contrast and resolution.

Next, we show the construction and analysis of the second-order nonstandard finite difference methods (NSFD) scheme for theta methods and explicit Runge-Kutta method. Finally, we present an application of the NSFD scheme for Fokker-Planck (FP) frameworks in esophageal cancer. We study a stochastic model of calcium signaling dynamics in the deterministic setup using the FP framework and solve this PDE using the NSFD scheme. We also present a detailed analysis of the numerical solution. To demonstrate the effectiveness of the theoretical studies, we show various numerical experiments.

TABLE OF CONTENTS

CHAPTER 1

Introduction

1.1 Hybrid imaging

Hybrid imaging refers to the amalgamation of two (or more) imaging modalities to form a new technique. It has attracted the research community a lot in the last few decades. The idea behind hybrid imaging methods is combining a high contrast modality and a high-resolution modality to get images with high contrast and resolution simultaneously. High contrast modalities like electrical impedance tomography (EIT) are used primarily for imaging electrical, optical, or elastic properties of biological tissues because these properties vary significantly between healthy and unhealthy tissues. On the other hand, modalities like magnetic resonance imaging (MRI) and ultrasound are used to provide better resolution. Therefore, the inversion process for hybrid imaging problems involves two steps coming from each modality discussed above. For a more detailed discussion on hybrid imaging techniques, please see the review articles [1, 2].

Examples of hybrid modalities include coupling optics or electromagnetism with ultrasound: Photo-Acoustic Tomography (PAT) and Thermo-Acoustic Tomography (TAT); coupling magnetic resonance with electrical currents: Magnetic Resonance EIT (MREIT) and Current Density Impedance Imaging (CDII) and many others.

1.2 Current density impedance imaging

Electrical impedance tomography (EIT) is an imaging modality, in which one attempts to recover the conductivity of a body from the boundary measurement of current and voltage [3]. The underlying inverse problem is highly ill-posed and non-linear yet very important due to its wide range of applications in the fields such as medical imaging [4] and engineering [5, 6]. The following conductivity equation gives the mathematical formulation of the EIT inverse problem

$$\begin{aligned} -\nabla \cdot (\sigma(x)\nabla u(x)) &= 0 \quad x \in \Omega, \\ \sigma(x)\frac{\partial u}{\partial \nu}(x) &= f(x),\ x \in \Gamma, \end{aligned} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^n$ is a convex and bounded domain with Lipschitz boundary and Γ is the boundary of Ω. In this model, σ is the electrical conductivity, u represents the electric potential and f is the current applied to the boundary. The reconstructions obtained through the EIT setup usually have high contrast but limited spatial resolution [7]. On the other hand, reconstructions obtained through ultrasound imaging have very high resolution but limited contrast [8, 9]. In recent years, attempts have been made to combine multiple imaging modalities to obtain image reconstructions with both high

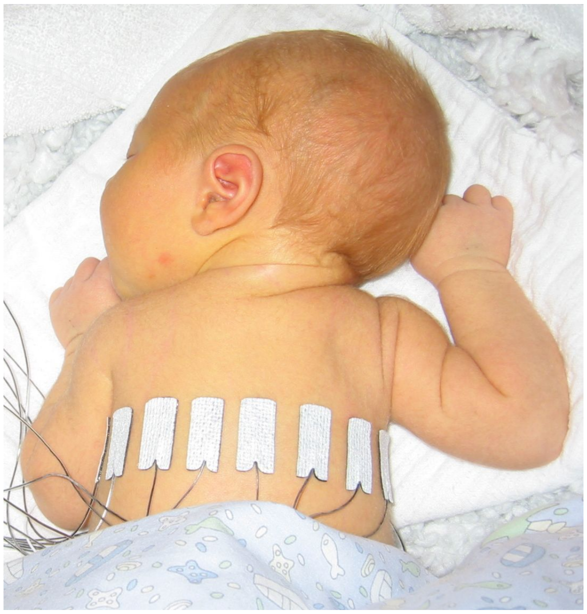


Figure 1.1: EIT (Source: Wikipedia)

contrast and high resolution. This led to the emergence of hybrid imaging methods that belong to class of coupled-physics imaging modalities to generate images of superior quality. One of such imaging methods, known as current density impedance imaging (CDII) combines the classical EIT setup with magnetic resonance (MR) scanning [10, 11]. It is alternatively known as magnetic resonance EIT (MREIT). Current or voltage is applied through the electrodes, which give rise to an interior electric field and the corresponding generated magnetic field, represented as $B = (B_x, B_y, B_z)$, is measured by the MR scanner. The corresponding inverse problem is to solve for the conductivity σ from B_z using the well-known iterative Harmonic B_z-algorithm [12, 13]. Convergence of the harmonic B_z algorithm has been well-studied [12, 13, 14]. In particular, it has been shown that for small contrast values of the target conductivity, the harmonic B_z-algorithm is stable and convergent, provided we have a good initial guess [13]. Thus, it is not clear that one can recover good quality images for high contrast objects through Harmonic B_z-algorithm.

An alternate approach to solve the CDII inverse problem is to use the knowledge of interior electric field, which is obtained from the magnetic field. Correspondingly, the magnitude of the interior electric field is also determined [15, 16], which is given by

$$H(\sigma(x)) = \sigma(x)|\nabla u(x)|, \quad x \in \Omega. \tag{1.2}$$

The formulation of reconstruction problem is as follows: Given the boundary data f for, possibly, several choices of boundary patterns and the corresponding interior measurement data H, find the conductivity distribution σ. In this framework, we use the internal function $H(\sigma)$ to replace σ in the EIT equation (1.1) to get the following nonlinear equation

$$\begin{aligned} \nabla \cdot \left(\frac{H}{|\nabla u|} \nabla u \right) &= 0 \text{ in } \Omega, \\ \frac{H}{|\nabla u|} \frac{\partial u}{\partial \nu} &= f \text{ on } \Gamma. \end{aligned} \tag{1.3}$$

For the CDII inverse problem, the solution to the boundary value problem (1.3) is crucial but it is difficult to use it in practice because of its highly nonlinear behaviour and also because the data represented by the measured values of H enter as a coefficient of the differential model [7]. Even with the additional measurements, analysis and application of the 1-Laplacian relies on an iterative localized algorithm, wherein one considers an approximation of the CDII problem. This subsequently led to several computational approaches in solving the CDII inverse problem. In [17], it was proved that the linearized problem is elliptic and hence solvable, if there are at least n set of measurements $\{H_i(\sigma)\}_{i=1}^n$ and corresponding to n boundary data $\{f_i\}_{i=1}^n$ such that ∇u_i and ∇u_j are nowhere collinear for $i \neq j$. It has been shown in [18] that the solution of the above 1-Laplacian equation with the Neumann boundary condition is non-existent unless additional measurements with different boundary current patterns are used. Recovery of isotropic conductivity in regions where the magnetic field is transversal using two internal current distributions was done using an explicit local formula [19]. Moreover, using the information of two internal current distributions, the authors in [18] uniquely determine the singular support of the conductivity function. In [20], the authors showed that the conductivity in the planer domain can be recovered from a single voltage-current on a part of boundary and the magnitude of one interior current density. In the same article, they also provide sufficient conditions on Dirichlet boundary data to guarantee unique recovery of conductivity. In [21], the recovery of Hölder continuous conductivities have been establised for domains with connected boundary from the interior measurement of the magnitude of one current density. Determination of isotropic conductivity variations from measurements of two current density vector fields was studied in [10]. In [22], authors showed the recovery of planar conductivities by solving the 1-Laplace equation with partial boundary data.

The well-known numerical reconstruction algorithm using the internal current distribution is an iterative J-substitution algorithm which was first introduced by [23] and subsequently considered in other works, see for e.g., in [24, 25, 21, 11]. It has been shown that the J-substitution algorithm is able to reconstruct the conductivity with high resolution [25, 18]. Another numerical reconstruction iterative method is the regularized D-bar method [26] that provides images with high resolution. In [27], the authors use an alternating split Bregman algorithm for solving a minimization problem related to the energy functional corresponding to the 1-Laplacian equation (1.3). Also, in [28], Picard and Newton type algorithms are implemented to solve the 1-Laplacian problem. But there is not enough evidence to suggest that these existing algorithms (linearized or localized iterative methods) can provide high contrast images, specially for objects with holes or inclusions, which are inherent to CDII reconstructions. In chapter 3, we will see a a new optimization framework for the sparse reconstruction of log-conductivity in CDII. The cost functional considered in this framework consist of a data-fitting term, $L^2 - L^1$ regularization term, and a Perona-Malik anisotropic diffusion filtering term.

1.3 Two-photon photoacoustic computed tomography

Photoacoustic tomography (PAT) is an another hybrid imaging modalities that couples electromagnetic waves together with ultrasound. PAT takes advantage of the photoacoustic effect to convert absorbed optical energy into acoustic waves. In

PAT, near infrared (NIR) light propagates into a medium of interest and a fraction of the incoming light energy is absorbed, which results in local heating and subsequent cooling of the medium. Due to this heating and cooling phenomenon, acoustic waves are generated that are recorded at the boundary of the medium. The inverse problem is to reconstruct the diffusion, absorption and Grüneisen coefficients from these acoustic measurements, for more details on the subject see [29, 30, 31, 32, 33, 34, 35, 36, 37, 38] and references therein. The PAT technology can be divided into two main categories based on the way images are formed, namely, photoacoustic microscopy (PAM) and photoacoustic computed tomography (PACT) [40, 41]. In PACT, unfocused light waves and a series of transducers are used to obtain cross-sectional images of an object from the photoacoustic data by solving an inverse problem. Moreover, the PACT mechanism provides a larger tomographic imaging penetration depth beyond one centimeter, but at the expense of inferior spatial resolution [41]. On the other hand, in PAM, focused laser waves are used along with the method of transverse raster-scanning to obtain cross-sectional images of an object from photoacoustic data in a direct way. However, PAM is known to provide high resolution within a tomographic imaging depth of several millimeters [40]. A modified PAM method, known as the acoustic resolution-PAM (AR-PAM), uses unfocused light on a larger

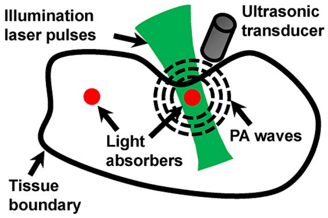


Figure 1.2: PAT [39]

region with similar intensity as that of the focused light to obtain greater tomographic imaging depth similar to PACT [40].

In recent years, a strong non-linear mechanism, called the two-photon absorption, has been observed and measured in the process of PAT reconstructions (see [42, 43, 44, 45, 46]. The two-photon absorption phenomenon is observed when an electron transfers to an excited state after simultaneously absorbing two photons. An imaging modality where one tries to recover optical properties of heterogeneous media (such as biological tissues) using the photoacoustic effect resulting from two photon absorption is known as two-photon photoacoustic tomography (2P-PACT) [42, 43, 47]. Even though the occurrence of two-photon absorption (in healthy biological tissues) is less frequent than single-photon absorption, two-photon absorption is extremely useful in practice. Similar to the PAT mechanism, 2P-PACT can be classified based on the different image formation ways: two-photon photoacoustic microscopy (2P-PAM) and two-photon photoacoustic computed tomography (2P-PACT). While, 2P-PAM is quite often used for deep tissue imaging, thick brain tissue imaging blood vessel imaging, liver biopsies (see [48, 49, 50, 51, 52, 53, 54] and references therein), there has not been any feasible demonstration of the application of 2P-PACT in medical imaging. However, since the one-photon and two-photon absorption effects are not easily separable from each other [44], a coupling of the two-photon effect due to unfocused strong light intensity mechanism, as in AR-PAM, and using an array of transducers to capture the photoacoustic wave data and solving an inverse problem to obtain the optical coefficients, as in PACT, is a promising and efficient way to use the 2P-PACT method in medical imaging (see [55] for a AR-PAM setup). Thus, it is of paramount importance to study mathematical numerical algorithms for obtaining reconstructions in 2P-PACT, which is the main aim of this chapter 4.

1.4 Nonstandard finite difference methods

Dynamical systems are important in many disciplines, including biology, economics, engineering, and chemistry. Because the majority of dynamical systems cannot be solved analytically, numerical methods are typically used to approximate their solutions. However, the stability properties of the corresponding numerical solutions are typically strongly dependent on the computational step size, particularly when standard numerical methods such as the explicit Euler and Runge-Kutta methods are used. R.E. Mickens [56] pioneered the use of nonstandard finite difference (NSFD) methods to overcome this dependency while numerically preserving important properties of exact solutions. Since then, NSFD methods have been developed and applied to a wide range of scientific and engineering problems [57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70]. Several classes of NSFD methods, in particular, have been developed based on standard theta methods [71, 72, 73, 67, 74] and standard two-stage explicit Runge-Kutta (ERK2) methods [71, 73]. However, these methods, while preserving the local dynamical properties of solutions near equilibrium points, are only first-order accurate. Recently, modified nonstandard theta and Runge-Kutta methods [75, 76, 77] that are not only elementary stable but also second-order accurate have been presented. But, those methods were only developed for one-dimensional autonomous dynamical systems. The previous theoretical results are extended in this paper, and new modified NSFD theta and modified NSFD ERK2 methods for solving n-dimensional autonomous dynamical systems are designed. The extensions are based on the use of novel denominator functions that account for both the elementary stability and the increased accuracy of the numerical methods.

1.5 Outline of thesis

In Chapter 2 we see the mathematical preliminaries, which is helpful to understand the later chapters. We discuss some concepts of functional analysis, optimization, and nonstandard finite difference methods used to develop the discussion in the thesis. Chapter 3 is dedicated to current density impedance imaging (CDII), where we see a new optimization framework that uses tools from PDE control models and anisotropic diffusion to obtain reconstructions with high resolution and good contrast. This framework is developed for reconstructing the log-conductivity in CDII. Chapter 4 focuses on a robust optimization framework for the sparsity-based nonlinear reconstruction of optical parameters in two-photon photoacoustic computed tomography. We discuss this framework to recover the optical properties of the medium Ω from the measured acoustic wave signals. In Chapter 5, we discuss the second order nonstandard finite difference methods for autonomous ordinary differential equations. In particular, we see the NSFD numerical schemes for theta methods and two-stage explicit Runge-Kutta methods. A natural generalization of NSFD to partial differential equations is discussed in Chapter 6. Additionally, we see a nonstandard finite difference Chang-Cooper scheme for calcium- signaling. The theoretical discussion is supported by numerical experiments at the end of every chapter, followed by discussions and conclusions on the current work and possible future developments.

CHAPTER 2

Mathematical Preliminaries

This chapter focuses on definitions and preliminaries, which will help us understand upcoming content. More specifically, we discuss some topics from functional analysis [78] (such as weak convergence, weak* convergence, reflexivity, Sobolev embedding, etc.), basic optimization and nonstandard finite difference methods concepts.

2.1 Weak and weak* convergence

In this section, we discuss several modes of convergence for sequences in a normed linear space X and in its dual X'.

Definition 2.1.1. (Weak Convergence) A sequence (x_n) in a normed space X is said to be weakly convergent if there exists $x \in X$ such that $x'(x_n) \to x'(x)$ in K for every $x' \in X'$. We then write it as, $x_n \overset{w}{\to} x$ in X.

Remark 1. In general, $x_n \overset{w}{\to} x$ in X does not imply that $x_n \to x$ in X. To see this, let $X = L^p([-\pi, \pi])$, $1 \leq p < \infty$, and $x_n(t) = e^{int}$, $t \in [-\pi, \pi]$, $n = 1, 2, 3, \ldots$. Then $\|x_n\|_p = (2\pi)^{1/p}$ for each n, and hence $x_n \nrightarrow 0$. Now if $x' \in X$, then by Riesz representation theorem for L^p spaces, there exists some $y \in L^q([-\pi, \pi])$ with $\frac{1}{p} + \frac{1}{q} = 1$ such that

$$x'(x) = \int_{-\pi}^{\pi} xy \; dm, \quad x \in X.$$

Thus,

$$x'(x_n) = \int_{-\pi}^{\pi} y(t)e^{int} dm(t) = 2\pi\hat{y}(-n) \to 0 \text{ as } n \to \infty$$

The last convergence implies (using Riemann-Lebesgue Lemma) $y \in L^q([-\pi, \pi])$, which is a subset of $L^1([-\pi, \pi])$. Thus $x_n \overset{w}{\to} 0$.

Definition 2.1.2. (Weak* Convergence) A sequence (x'_n) in X' is said to be weak* convergent if there is some $x' \in X'$ such that $x'_n(x) \to x'(x)$ in K for every $x \in X$. We then write it as, $x'_n \xrightarrow{w^*} x'$ in X'.

Remark 2. In general, $x^*_n \xrightarrow{w^*} x'$ in X' does not imply that $x'_n \to x'$ in X'. For instance, if $X = L^p([-\pi, \pi])$, $1 \leq p < \infty$,then X' is linearly isometric to $L^q([-\pi, \pi])$, where $\frac{1}{p} + \frac{1}{q} = 1$, and if we assume $y_n(t) = e^{int}$ for $t \in [-\pi, \pi]$, then $y_n \nrightarrow 0$ in $L^q([-\pi, \pi])$. But by Riemann-Lebesgue lemma (as described in previous example), $y_n \xrightarrow{w^*} 0$.

As X'' is the dual of the normed space X', we see that $x'_n \xrightarrow{w} x'$ in X' if and only if $x''(x'_n) \to x''(x')$ for every $x'' \in X''$. Let us now consider the canonical embedding $J : X \to X''$. If $x'_n \xrightarrow{w} x'$ in X', then for every $x \in X$, we get

$$x'_n(x) = J(x)(x'_n) \to J(x)(x') = x'(x)$$

and that means, $x'_n \xrightarrow{w^*} x'$ in X'. Now, we have three modes of convergence in the normed dual X':

- Norm convergence
- Weak convergence
- Weak* convergence

We also have,

$$\text{Norm convergence} \implies \text{Weak convergence} \implies \text{Weak}^*\text{convergence.}$$

Next, we will see an example of Banach space X such that three modes of convergence in X' are distinct, that is, reverse implication may not hold in general.

Let $X = l^1$, then X' can be identified with l^∞. Let us consider a sequence (e_n) in l^∞. Now,

$$\|e_n\|_\infty = 1, \text{ for every } n \implies \|e_n\| \nrightarrow 0.$$

Further, we will show that $e_n \xrightarrow{w} 0$ in l^∞. Suppose, this is not true. Then there exist some $f \in (l^\infty)'$ such that $f(e_n) \nrightarrow 0$.

That means, there are positive integers $n_1 < n_2 < \ldots$ and some $\delta > 0$ such that $|f(e_{n_j})| \geq 0$ for each $j = 1, 2, \ldots$. Now, for $m = 1, 2, \ldots$, define

$$x_m = \text{sgn} f(e_{n_1})e_{n_1} + \ldots + \text{sgn} f(e_{n_m})e_{n_m}.$$

Then for each m, $\|x_m\|_\infty \leq 1$, but

$$f(x_m) = |f(e_{n_1})| + \ldots |f(e_{n_m})| \geq m\delta,$$

which tends to ∞ as $m \to \infty$ and this goes to ∞ as m approaches to ∞. This gives a contradiction that f is continuous on l^∞. Hence $e_n \xrightarrow{w} 0$ in l^∞.

Now, we consider a sequence (a_n) in l^∞, where $a_n = (1, \ldots, 1, 0, 0, \ldots)$ with 1 occurring only in the first entries and let $a = (1, 1, \ldots)$. Then (a_n) is not weak convergent to a. To prove this, find $f \in (l^\infty)'$, such that $f(x) = 0$ for every $x \in c_0$ and $f(a) \neq 0$ using consequence of Hahn Banach extension theorem. Since each a_n is in c_0, so we see that $f(a_n) = 0$. Thus $f(a_n) \nrightarrow f(a)$.

However, we will see that $a_n \xrightarrow{w^*}$ in l^∞. Let F denote the linear isometry from l^∞ to l^∞ to $(l^1)'$. Then for each $x \in l^1$,

$$F(a_n)(x) = \sum_{j=1}^{\infty} x(j)a_n(j) = \sum_{j=1}^{n} x(j),$$

while

$$F(a)(x) = \sum_{j=1}^{\infty} x(j)a(j) = \sum_{j=1}^{\infty} x(j)$$

Thus, $F(a_n)(x) \to F(a)(x)$ for every $x \in l^1$, that is, $a_n \xrightarrow{w^*} a$. So, we conclude that in general, weak convergence is even weaker than norm convergence, and weak* convergence is even weaker than the weak convergence.

Theorem 2.1.1. *Let (x_n') be a sequence in a normed space X'. If*

(i) (x_n') is bounded and,

(ii) $(x_n'(x))$ is a Cauchy sequence in K for each x in a subset of X whose span is dense in X,

then (x_n') is weak convergent in X'. The converse holds if X is a Banach space. If $x_n' \to x'$ in X', then $\|x'\| \leq \liminf_{n\to\infty} \|x_n'\|$.*

2.1.0.1 Bolzano-Weirstrass Property

The statement of classical Bolzano-Weierstrass is that every bounded sequence in K has a convergent subsequence. If X is a normed space, then every bounded sequence in X has a convergent subsequence if and only if X is finite dimensional. Thus, the classical Bolzano-Weierstrass property does not hold for the norm convergence in X, when X is an infinite dimensional normed space. Hence, it is worth to investigate whether every bounded sequence in X has a weak convergent subsequence and whether every bounded sequence in X' has a weak* convergent subsequence.

Remark 3. We will now see an example of a bounded sequence which does not have a weak convergent subsequence.

Consider $X = l^1$ and a sequence (e_n), then $\|e_n\|_1 = 1$ for all n, but (e_n) does not have a weak convergent subsequence since $\|e_n - e_m\|_1 = 2$ for all $n \neq m$. We could also see in this way, if (e_{n_k}) is a subsequence of (e_n) and $e_{n_k} \xrightarrow{w} x_0$ in l^1, then $e_{n_k}(j) \to x_0(j)$ as $k \to \infty$, for each $j = 1, 2, \ldots$ this implies $x_0 = 0$. But if we let $f(x) = \sum_{j=1}^{\infty} x(j)$ for $x \in l^1$, then $f \in l^1$ and $f(e_{n_k}) \to 1$, whereas $f(x_0) = f(0) = 0$ and this contradiction proves that (e_n) has no weak convergent subsequence in l^1.

Remark 4. We will now see an example that not every bounded sequence in X' has a weak* convergent subsequence, let $X = l^\infty$ and for $n = 1, 2, \ldots,$ define

$f_n(x) = x(n), \quad x \in X$. Then $f_n \in X$ and $\|f_n\| = 1$ for all n. Let (f_{n_k}) be a subsequence of (f_n). Define $x \in X$ by

$$x(j) = \begin{cases} 1, & \text{if } j = n_k \text{ and k is odd} \\ 0, & \text{otherwise.} \end{cases}$$

Then $f_{n_k} = x(n_k) = 1$ for every odd k and $f_{n_k} = x(n_k) = 0$ for every even k. Since $(f_{n_k}(x))$ does not converge in K, (f_{n_k}) cannot be weak* convergent in X'.

No need to worry, we do have some positive result as well.

Theorem 2.1.2. *(Banach 1932) Let X be a separable normed space. Then every bounded sequence in X' has a weak* convergent subsequence.*

From, Theorem 2.1.1 and 2.1.2 we find that the closed unit ball of the dual of a separable normed space X is **weakly* sequentially compact**. We say that a normed X is weakly sequentially compact if $x'_n \in X'$ with $\|x'_n\| \leq 1$, then, there is a subsequence (x'_{n_k}) of (x'_n) such that $x'_{n_k} \xrightarrow{w^*} x'$, where $\|x'\| \leq \liminf_{k\to\infty} \|x'_{n_k}\| \leq 1$.

2.2 Reflexivity

This section is devoted to show how to embed a normed space in its second dual X''. For a fixed $x \in X$, define $J(x) : X' \to K$ by

$$J(x)(x') = x'(x), x' \in X'$$

Then $J(x) \in X''$, $\|J(x)\| = \|x\|$ and the map $J : X \to X''$ is linear. Since X'' is a Banach space, it follows that the subspace $J(X)$ is closed in X'' if and only if X is a Banach space.

Definition 2.2.1. (Reflexive space) A normed space is said to be reflexive if the canonical embedding J is surjective, that is if x'' is a continuous linear functional on X', then there exists some $x \in X$ such that $x'' = J(x)$.

Now, we will see some properties of reflexive normed space.

Theorem 2.2.1. *Let X be a reflexive normed space. Then*

(i) X is Banach and it remains reflexive in any equivalent form.

(ii) X' is reflexive.

(iii) Every closed subspace of X is reflexive.

(iv) X is separable if and only if X' is separable.

Example 2.2.1. If $1 < p < \infty$, $L^p([a, b])$ is reflexive.

Next result shows a relationship between reflexivity and weak convergence.

Theorem 2.2.2. *(Eberlein, 1947) Let X be a normed space. Then X is reflexive if and only if every bounded sequence in X has a weak convergent subsequence.*

In other words, a normed is reflexive if and only if its closed unit ball is "**weak sequentially compact**".

2.3 Sobolev spaces

This section is introduced to define some necessary function spaces which we would use later. More details about Sobolev spaces can be find in [79, 80, 81]. Let $\Omega \subset \mathbb{R}^n$ be open, bounded domain with Lipschitz boundary $\partial\Omega$. we take $p \in [1, +\infty]$

Definition 2.3.1. The Sobolev space $W^{k,p}(\Omega)$ is defined as

$$W^{k,p}(\Omega) = \{u \in W^k(\Omega) : D^\alpha u \in L^p, \forall |\alpha| \leq k\}.$$

It is clear that $W^{k,p}(\Omega)$ is a linear subspace of $(L^p()\Omega), \| \cdot \|_p$. We can define two equivalent norms in $W^{k,p}(\Omega)$:

$$\|u\|_{k,p} = \begin{cases} \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|_p^p\right)^{1/p}, & \text{if } p \in [1, +\infty), \\ \max_{|\alpha| \leq k}, \|D^\alpha u\|_\infty & \text{if } p = +\infty. \end{cases}$$

and

$$|||u|||_{k,p} \equiv \sum_{\alpha \leq k} \|D^\alpha u\|_p.$$

If $p = 2$, then the norm $\| \cdot \|_{k,2} \equiv W^{k,2}(\Omega)$ is the one associated to the inner product

$$(u, v)_{k,2} = \sum_{|\alpha| \leq k} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}.$$

Theorem 2.3.1. *The Sobolev space $W^{k,p}(\Omega)$ is a Banach space for $1 \leq p \leq +\infty$. Further,*

(i) $W^{k,p}(\Omega)$ is reflexive for $1 < p < +\infty$

(ii) $W^{k,p}(\Omega)$ is separable for $1 \leq p < +\infty$

In particular, $H^k(\Omega)$ is a separable Hilbert space.

We denote by $W_0^{k,p}(\Omega)$ closure of $C_0^\infty(\Omega)$ in the space $W^{k,p}(\Omega)$ with respect to $W^{k,p}(\Omega)-$norm. When $p = 2$, we also write $W_0^{k,p}(\Omega) = H_0^k(\Omega)$. Thus $u \in W_0^{k,p}(\Omega)$ if and only if there exists functions $u_m \in C_0^\infty(\Omega)$ such that $u_m \to u$ in $W^{k,p}(\Omega)$.
The space $H_0^1(\Omega)$ consists of functions $u \in H^1(\Omega)$ such that

$$u = 0 \quad \text{on } \partial\Omega$$

in the trace sense. We denote the dual of $H_0^1(\Omega)$ by $H^{-1}(\Omega)$.

2.4 Embedding theorems

In this section, we will see the well-known Sobolev and Rellich-Kondrachov embedding theorems. But, first, we understand the meaning of embedding and compact embedding.

Definition 2.4.1. Let $(X, \| \cdot \|_X)$ and $(Y, \| \cdot \|_Y)$ be normed spaces.

(i) The space X is said to be embedded in the space Y, we denote it by $X \hookrightarrow Y$, if there exists an injective linear and continuous operator from X into Y and the operator is called an embedding.

(ii) The space X is compactly embedded in the space Y, if there exists an embedding of X in Y which is compact and we denote it by $X \hookrightarrow Y$.

Embeddings of $W^{k,p}(\Omega)$ can be considered into the following three classes of spaces:

(i) $W^{j,q}(\Omega)$ with $0 \leq j \leq k$ $(W^{0,j}) \equiv L^q(\Omega)$ and q denotes the conjugate exponent of p, i.e., $\left(\frac{1}{q} + \frac{1}{p} = 1\right)$.

(ii) $C^j_B(\Omega)$, for $j \in \mathbb{N} \cup \{0\}$, i.e., the space of functions with continuous and bounded partial derivatives up to order j. This is a Banach space with the following norm:

$$\|u\|_{C^j_B(\Omega)} = \max_{0\leq|\alpha|\leq j} \sup_{x\in\Omega} |D^\alpha u(x)|$$

for every $u \in C^j_B(\Omega)$.

(iii) $C^{j,\nu}_{B,u}(\Omega)$, i.e., the space of functions with bounded and uniformly continuous partial up to order j in Ω and these partial derivatives of order j satisfy a Hölder condition with exponent $\nu \in (0, 1)$. It is also a Banach space with the norm

$$\|u\|_{C^{j,\nu}(\Omega)} = \|u\|_{C^j_B(\Omega)} + \sum_{|\alpha|=j} \sup_{\substack{x,y\in\Omega \\ x\neq y}} \frac{|D^\alpha u(x) - D^\alpha(y)|}{|x-y|^\nu}$$

for every $u \in C^{j,\nu}_{B,u}(\Omega)$. Clearly, $C^{j,\nu}_{B,u}(\Omega) \subset C^j_B(\Omega)$.

Now we will see the main embedding results.

Theorem 2.4.1. *Let $\Omega \subset \mathbb{R}^N$ be an open subset satisfying the cone condition. Consider also $k \in \mathbb{N} \cup 0$. We have the following embeddings.*

(i) If $k < \frac{N}{p}$, then $W^{j+k,p}(\Omega) \hookrightarrow W^{j,q}(\Omega)$ for every $q \in \left[p, \frac{Np}{N-kp}\right]$

(ii) If $k = \frac{N}{p}$, then $W^{j+k,p}(\Omega) \hookrightarrow W^{j,q}$ for every $p \leq q < \infty$. In addition, in the particular case $p = 1$ and $k = N$, we also have $W^{j+N,1} \hookrightarrow C^j_B(\Omega)$.

(iii) If $k > \frac{N}{p}$, then $W^{j+k,p} \hookrightarrow C^j_B(\Omega)$. Furthermore, if $\partial\Omega$ is of class C^1, then

(iv) If $k - 1 < \frac{N}{p}$, then $W^{j+k,p}(\Omega) \hookrightarrow C^{j,\nu}_{B,u}(\Omega)$ for every $\nu \in \left(0, k - \frac{N}{p}\right]$.

(v) If $k-1=\frac{N}{p}$, then $W^{j+k,p}(\Omega) \hookrightarrow C^{j,\nu}_{B,u}(\Omega)$ for every $\nu \in (0,1)$.

Theorem 2.4.2. *(Rellich-Kondrasov). Let $\Omega \subset \mathbb{R}^n$ be a bounded open set of class C^1. Then the following inclusions are compact.*

(i) if $p < n$, $W^{1,p}(\Omega) \to L^q(\Omega)$, $1 \leq q < p^*$,

(ii) if $p = n$, $W^{1,n}(\Omega) \to L^q(\Omega)$, $1 \leq q < \infty$,

(i) if $p > n$, $W^{1,p}(\Omega) \to C(\bar{\Omega})$, where $\bar{\Omega} = \Omega \cup \partial\Omega$.

2.5 Optimization

Let V be a Banach space and K be a non-empty subset of V. Let $J : V \to \mathbb{R}$, and we consider

$$\inf_{v \in K \subset V} J(v)$$

Definition 2.5.1. An element u is called a local minimizer of J on K if $u \in K$ and $\exists\ \delta > 0$ such that $\forall\ v \in K$

$$\|v-u\| < \delta \implies J(v) \geq J(u).$$

An element u is called a global minimizer of J on K if $u \in K$ and

$$J(v) \geq J(u) \quad \forall v \in K.$$

Definition 2.5.2. A minimizing sequence of a function J on the set K is a sequence $(u^n)_{n\in\mathbb{N}}$ such that

$$u^n \in K\ \forall\ n \text{ and } \lim_{n\to+\infty} J(u^n) = \inf_{v\in K} J(v).$$

By definition of the infimum value of J on K, there always exist minimizing sequences.

If V is a finite dimensional normed vector space (in particular $V = \mathbb{R}^n$). Then we have following theorem for the existence of minimizer:

Theorem 2.5.1. *Let K be a non-empty closed subset of $\mathbb{R}^N$ and J a continuous function from K to $\mathbb{R}$ satisfying the so-called "infinite at infinity" property, i.e.,*

$$\forall\ (u^n)_{n\geq 0} \text{ sequence in } K, \lim_{n\to+\infty} \|u^n\| = +\infty \implies \lim_{n\to+\infty} J(u^n) = +\infty$$

Then there exists at least one minimizer of J on K. Furthermore, from each minimizing sequence of J on K one can extract a subsequence which converges to a minimum of J on K.

The key idea which makes this theorem true is that the closed bounded sets are compact in finite dimensions. In general, the above result is not true for infinite dimensional spaces, for instance, see the following remark:

Remark 5. Let $V = H^1(0,1)$ with the norm given by $\|v\| = \left(\int_0^1 (v'^2(x) + v^2(x))dx\right)^{1/2}$, for $v \in H^1(0,1)$. Consider,

$$J(v) = \int_0^1 \left((|v'(x)| - 1)^2 + v^2(x)\right) dx.$$

One may check that this functional has no minimizer even though it satisfies the assumptions of 2.5.1.

Next, we add convexity assumption, and obtain the existence of minimizers.

Definition 2.5.3. A set $K \subset V$ is said to be convex if, for any $u, v \in K$ and for any $\theta \in [0,1]$,

$$\theta u + (1-\theta)v \in K.$$

Definition 2.5.4. Let K be convex subset of V, then a function $J : K \to \mathbb{R}$, is said to be convex on K if

$$J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(y),\ \forall\ u, v \in K,\ \forall\ \theta \in [0,1].$$

Further, J is said to be strictly convex if the above inequality is strict whenever $u \neq v$ and $\theta \in (0,1)$.

We have the following result for the existence of minimizer under the convexity assumption:

Theorem 2.5.2. *Let K be a non-empty closed set in a reflexive Banach space V, and J be a convex continuous function on K, which is "infinite at infinity" in K, i.e.,*

$$\forall\ (u^n)_{n\geq 0} \textit{ sequence in } K, \lim_{n\to+\infty} \implies \lim_{n\in\infty} J(u^n) = +\infty.$$

Then, there exists a minimizer of J in K.

Remark 6.

(i) V is reflexive Banach space if and only if $(V')' = V$ where V' is the dual of V

(ii) The theorem is still true if V is just the dual of a separable Banach space.

(iii) This assumption is satisfied for all functional spaces such as $L^p(\Omega)$ with $1 < p \leq +\infty$.

We will now see some uniqueness results under the same assumptions.

Proposition 2.5.3. *If J is strictly convex, then there exists at most one minimizer of J.*

Proposition 2.5.4. *If J is convex on the convex set K, then any local minimizer of J on K is a global minimizer.*

Remark 7.

(i) For convex functions there is no difference between local and global minimizers.

(ii) Convexity is not the only tool to prove existence of minimizers. Another method is, for example, compactness.

2.6 Differentiation in Banach spaces

We will now see the notions of derivatives in function spaces. Let $\mathcal{L}(A, B)$ denote the space of bounded the space of bounded linear operators from Banach space A to Banach space B. Let $(Z, \|\cdot\|_Z)$, $(V, \|\cdot\|_V)$ be real Banach spaces, U be an open subset of Z, $F : U \to V$ and $z \in U$.

Definition 2.6.1. (Directional derivative). F is said to be directionally differentiable at z if $\lim_{t\to 0} \frac{1}{t}(F(z+th) - F(z))$ exists in V for all $h \in Z$. If this limit exists, we denote

$$F'(z,h) := \lim_{t\to 0} \frac{1}{t}(F(z+th) - F(z))$$

and say that $F'(z,h)$ is the directional derivative of F at z in the direction h.

Definition 2.6.2. (Gâteaux derivative). F is said to be Gâteaux differentiable at z if its is directional derivative exists and $F'(z;h) = F'(z)h$ for $F'(z) \in \mathcal{L}(Z;V)$. We refer to $F'(z)$ as the Gâteaux derivative at z.

Next, we will see a stronger notation of derivation:

Definition 2.6.3. (Fréchet derivative). F is said to be Fréchet differentiable at z if and only if F is Gâteaux differentiable at z and the following holds:

$$F(z+h) = F(z) + F'(z)h + r(z,h) \text{ with } \frac{\|r(z,h)\|_V}{\|h\|_Z} \to 0 \text{ as } \|h\|_Z \to 0$$

Remark 8.

(i) If the Fréchet derivative exists then Gâteaux derivative also exists and they are same in this case. But, the converse is not true in general.

(ii) We say that F is continuously Gâteaux differentiable if $F'(\cdot)$ exists and $F'(\cdot)$ is continuous. In that case F is Fréchet differentiable.

(iii) Let $E = G(F(z))$ where F is Gâteaux differentiable at z and G is Fréchet differentiable at $F(z)$, then E is Gâteaux differentiable.

The case when $V = \mathbb{R}$, we obtain $\mathcal{L}(Z,V) = Z^*$. Further, If F is Gâteaux differentiable at z then we have

$$F'(z)h = \langle F'(z), h\rangle_{Z^*,Z},$$

where Z^* is the dual space of Z and $\langle \cdot, \cdot \rangle_{Z^*,Z}$ denote the duality pairing.

Example 2.6.1. Let $(H, (\cdot, \cdot)_H)$ be a real Hilbert space and $F : H \to \mathbb{R}$ defined as $F(z) := \|z\|_H^2 = (z, z)_H$, the $\forall z, h \in H$, we have

$$F(z + h) - F(z) = 2(z, h)_H + \|h\|_H^2$$

Thus,

$$F'(z)h = (2z, h)_H.$$

Using the Riesz Representation Theorem (identify H), we can write

$$(\nabla F(z), h)_H = \langle F'(z), h \rangle_{H^*,H},$$

where $\nabla F(z) \in H$ is the representative of $F'(z) \in H^*$. We denote $\nabla F(z) \in H$ as the gradient of F at z and then we have $\nabla F(z) = 2z$.

Remark 9. (i) The obtained expression to recognize $F'(z) \in H^*$ with an element of H is called the gradient of F.

(ii) We will use the notation $\nabla F(z)$ to denote the gradient.

(iii) We observe that the definition of the gradient depends on the underlying inner product.

A a brief overview of optimization problems with partial differential equation (PDE) constraints can be find in [82].

2.7 Nonstandard finite difference methods

An n-dimensional autonomous differential equation can be written as:

$$\vec{x}'(t) = \vec{f}(\vec{x}); \quad \vec{x}(t_0) = \vec{x}_0, \tag{2.1}$$

where $\vec{x}$ represents the vector-function $[x_1(t), \ldots, x_n(t)]^T$, $x_i : [t_0, T) \to \mathbb{R}$, $\vec{f} = [f_1, \ldots, f_n]^T \in C^2(\mathbb{R}^n; \mathbb{R}^n)$ is differentiable, $\vec{x}_0 \in \mathbb{R}^n$. It is assumed that System (2.1) has a finite number of only hyperbolic equilibria.

Definition 2.7.1. Let $\vec{x}^*$ be an equilibrium of System (2.1), $J(\vec{x}^*) = \left(\frac{\partial f_i}{\partial x_j}(\vec{x}^*)\right)_{1\leq i,j\leq n}$ be the Jacobian of system (2.1) at $\vec{x}^*$ and $\sigma(J(\vec{x}^*))$ denotes the spectrum of $J(\vec{x}^*)$. An equilibrium $\vec{x}^*$ of system (2.1) is called linearly stable if $Re(\lambda) < 0$, for $\lambda \in \sigma(J(\vec{x}^*))$ and linearly unstable if $Re(\lambda) > 0$ for some $\lambda \in \sigma(J(\vec{x}^*))$.

A general finite difference method which approximates the solution of System (2.1) on the interval $[t_0, T]$ can be written as:

$$D_{i,h}(\vec{x}^k) = F_{i,h}(f_i; \vec{x}^k), \ k = 0, \cdots, N_t, \tag{2.2}$$

where $D_{i,h}(\vec{x}^k) \approx x_i'\big|_{t=t_k}$, $F_{i,h}(f_i; \vec{x}^k) \approx f_i(\vec{x})$, $\vec{x}^k \approx \vec{x}(t_k)$, $t_k = t_0 + kh$, $k = 0, \cdots, N_t$, $i = 1, \cdots, n$, with mesh size $h > 0$.

The NSFD numerical methods discussed in this paper satisfy the following definitions introduced by Anguelov and Lubuma in [71] (see also [73, 83]):

Definition 2.7.2. The finite-difference method (2.2) for solving Equation (2.1) is a nonstandard finite difference (NSFD) method if at least one of the following conditions is satisfied for all $i = 1, 2, \ldots, n$:

- $D_{i,h}(\vec{x}^k) = \dfrac{x_i^{k+1} - x_i^k}{\varphi_i(h)}$, where $\varphi_i(h) = h + \mathcal{O}(h^2)$ is a non-negative function;
- $F_{i,h}(\vec{f}; \vec{x}^k) = g_i(\vec{x}^k, \vec{x}^{k+1}, h)$, where $g_i(\vec{x}^k, \vec{x}^{k+1}, h)$ is a non-local approximation of the i-th component of the right-hand side of System (2.1).

Definition 2.7.3 ([71, 73, 83])**.** A finite difference method is elementary stable if, for any value of the step-size h, its only fixed points $\vec{x}^*$ are the same as the equilibria of Equation (2.1) and the local stability properties of each $\vec{x}^*$ are the same for both the differential equation and the discrete method.

CHAPTER 3

Sparse Reconstruction of Log-Conductivity in Current Density Impedance Tomography

3.1 Introduction

[1] In the field of CDII imaging, we present a new optimization framework that uses tools from PDE control models and anisotropic diffusion theory to potentially obtain reconstructions with high resolution and contrast. Such a framework was first used in [85, 86] to reconstruct log-conductivity in acousto-electric tomography (AET). The results obtained demonstrated that such a framework was robust and accurate for imaging modalities arising through a partial differential equation (PDE). In this chapter, we will see a similar optimization framework developed in [86] for reconstructing the log-conductivity in CDII. formulate a minimization problem, where given interior electric field intensity data, we aim at determining the variation in conductivity from a known background conductivity. We, further, assume that this variation demonstrates a sparsity pattern. This is incorporated in our model through a $L^2 - L^1$ regularization term in our objective functional. To obtain sharp edges and, thus, improve spatial resolution of the reconstructed images, we use a Perona-Malik anisotropic diffusion filtering term in our functional. The resulting optimality system gives rise to an elliptic adjoint equation with a L^2 source term. Classical cell-nodal finite difference schemes are not applicable for solving such equations. We, thus, use a

[1] The content of this chapter has taken from [84], Gupta, M., Mishra, R. K., and Roy, S. Sparse reconstruction of log-conductivity in current density impedance tomography. Journal of mathematical imaging and vision, 2020, 62(2), 189-205.

averaged cell-nodal scheme to solve such equations. Finally, we solve the optimization problem using a variable inertial proximal scheme that efficiently handles the non-differentiable terms in the objective functional. We demonstrate through several examples that our method can be used to obtain superior quality reconstructions for objects with holes and inclusions.

This chapter is organized as follows: In the Section 3.2, we formulate the minimization problem for the CDII. In the Section 4.3, we present some theoretical results about our optimization problem. We also characterize the optimality system. The variable inertial proximal scheme and the averaged cell-nodal schemes to solve the optimization problem are discussed in Section 3.4. In the Section 3.5, we present simulation results of our CDII framework and compare them with the reconstructions obtained using the Picard scheme proposed in [28], which validate our framework for CDII and demonstrate the effectiveness of our method to reconstruct wide variety of objects with corners, holes and inclusions. A section on conclusions completes our work.

3.2 A minimization problem

We consider the conductivity equation in $\mathbb{R}^2$ arising in EIT

$$\begin{aligned} -\nabla \cdot (e^{\sigma(x,y)} \nabla u(x,y)) &= 0 \text{ in } \Omega, \\ u(x,y)|_{\Gamma} &= f_D(x,y), \end{aligned} \tag{3.1}$$

where $\Omega \subset \mathbb{R}^2$ is bounded, Γ is the boundary of Ω, e^{σ} is the conductivity coefficient and $u \in H^1_{f_D}(\Omega) = \{u \in H^1(\Omega) : u = f_D \text{ on } \Gamma\}$ is the electric potential.

We assume that σ is a sparse conductivity coefficient which we want to recover, given the fact that the conductivity of the background is 1. The conductivity equation (3.1) can also be written as

$$\mathcal{L}(u, \sigma, f_D) = 0,$$

where $\sigma(x,y) \in L_{ad} = \{\sigma \in H_0^1(\Omega) : \sigma_l \leq \sigma(x,y) \leq \sigma_u,\ \forall (x,y) \in \Omega\}$, $\sigma_u > 0$ and $\sigma_l = -\dfrac{1}{2}\sigma_t$.

We consider an optimization-based approach for reconstructing σ given $H_1(\sigma), H_2(\sigma)$, where

$$H(\sigma) = e^{\sigma}|\nabla u|$$

is the interior electric field corresponding to the voltage potential u. We consider the following cost functional

$$\begin{aligned} J(\sigma, u_1, u_2) = &\sum_{j=1}^{2} \frac{1}{2}\int_{\Omega} (H_j(x,y) - H_j^{\delta}(x,y))^2 \, dxdy + \frac{\beta}{2}\|\sigma\|^2_{L^2(\Omega)} \\ &+ \gamma \, \|\sigma\|_{L^1(\Omega)} + \frac{\delta}{2}\int_{\Omega} \log(1 + |\nabla\sigma(x,y)|^2) \, dxdy \end{aligned} \tag{3.2}$$

where u_1, u_2 satisfy (3.1) with boundary data f_D^1, f_D^2. We now consider the following minimization problem

$$\begin{aligned} &\min_{\sigma} \ J(\sigma, u_1, u_2), \\ &\text{s.t. } \ \mathcal{L}(u_1, \sigma, f_D^1) = 0, \\ &\qquad \mathcal{L}(u_2, \sigma, f_D^2) = 0. \end{aligned} \tag{P}$$

The term $\gamma \, \|\sigma\|_{L^1(\Omega)}$, $\gamma > 0$ in the functional, defined in (3.2), implements a L^1 regularization of the minimization problem that promotes sparsity patterns in the reconstruction of conductivity. Such a regularization method mirrors the well known compressed-sensing technique; see [87]. In recent past, optimal control with L^1 cost functionals has become a topic of major interest [88], because one obtains sparse controls through this procedure, which finds numerous applications. The motivation for sparse log-conductivity patterns is based on the assumption that the background conductivity is known to be 1 in a substantial part of the domain Ω after normalization and varies considerably from this value in correspondence to different kind of objects present within the domain.

The combined L^2-L^1 regularization allows for the reconstruction of conductivity, and thus the imaging of, possibly, irregular objects inside Ω. This does not serve the ultimate goal of reconstructing objects like tissues in medical imaging, which are more regular, save for the edges that eventually define them. We infuse this additional aprior knowledge into our model through the last term in our functional (3.2) that, commonly, appears in the field of anisotropic diffusion. Such a term plays an important role in dampening image noise while keeping significant parts of the image content such as edges and other anatomical details that are of utmost importance in the interpretation of the image. Anisotropic diffusion means non-uniform diffusion in different directions. The regions where $|\nabla\sigma|$ is very small corresponds to noise and thus, the process of smoothening occurs. At the edges or singularities of an object, where the value of $|\nabla\sigma|$ is large, there is a small amount of smoothening and this preserves the edges. A standard technique to implement anisotropic diffusion, in order to obtain a good contrast, is to use a total variation (TV) regularization [89, 90]. But, this regularization method gives rise to a non-differentiable term in the functional (3.2), thus requiring more sophisticated optimization algorithms. On the other hand, anisotropic diffusion is inherent to Perona-Malik (PM) filtering [91]. It is well-known that the diffusion process governed by the PM equation leads to a decrease in the total variation during its evolution [92]. We, thus, choose the energy functional of the Perona-Malik equation for anisotropic diffusion [91]. One can note that the PM regularization term is differentiable and, thus, easier to handle than the TV regularization term.

Mathematically, one can consider the PM filtering as the gradient flow generated by the non-convex and lower semi-continuous functional given by

$$J_{PM}(\sigma) = \int_{\Omega} \log(1 + |\nabla\sigma(x,y)|^2)\ dxdy.$$

We refer to [93, 92] for a general introduction to anisotropic diffusion and a detailed discussion on the PM functional. Further, in [86], the PM model was used in the reconstruction of log-conductivities in AET and it was observed that the reconstructions obtained demonstrated superior contrast and resolution. Thus, for the current setup in CDII, we use a similar PM anisotropic diffusion filter to facilitate high contrast and high resolution images.

3.3 Theory of the minimization problem

In this section, we discuss the existence of solutions of the minimization problem (P) and its characterization through a first-order optimality system. We refer to this minimization problem as the CDII sparse reconstruction problem (CDII-SR). Our analysis of this problem begins with the discussion concerning the existence and uniqueness of weak solutions of $\mathcal{L}(u, \sigma, f_D) = 0$, which can be proved by standard arguments of Riesz representation theorem [94, Chapter 8].

Proposition 3.3.1. *Let $\sigma \in L_{ad}$ and $f_D \in H^{1/2}(\partial\Omega)$. Then the problem* (3.1) *has a unique solution in $H^1_{f_D}(\Omega)$.*

The solvability of the CDII inversion problem depends on the type of Dirichlet boundary data f_D^j, $j = 1, 2$. In this, context, we have the following lemma from [95].

Lemma 3.3.2 (Boundary data)**.** *Let $\Omega \subset \mathbb{R}^2$ be a bounded simply connected open set, whose boundary Γ is a simple closed curve. Let $f = (f^1, f^2)$ be a mapping $\Gamma \to \mathbb{R}^2$ which is a homeomorphism of Γ onto a convex closed curve C, and let D denote the bounded convex domain bounded by C. Let $\sigma \in L^\infty(\Omega)$, and let $U = (u_1, u_2)$ be the e^σ-harmonic mapping whose components u_1 and u_2 are solutions to the Dirichlet problem* (3.1) *with $f_D = f_D^1$ and $f_D = f_D^2$, respectively, and $f_D^J \in H^1(\Omega) \cap C(\bar{\Omega})$, $J = 1, 2$. Then U is a homeomorphism of Ω onto D. In particular, for all $\omega \subset\subset \Omega$ we have either* $\det(\nabla u_1, \nabla u_1) > 0$ *or* $\det(\nabla u_1, \nabla u_1) < 0$ *almost everywhere in ω.*

In [17], the authors have shown that in 2D, the boundary condition pair $f_D^1 = x$ and $f_D^2 = y$ satisfies the conditions of Lemma 4.3.3 and, thus, the corresponding solutions to (3.1) u_1 and u_2 have no critical points and $\nabla u_1, \nabla u_2$ are not collinear in $\bar{\Omega}$. We will use these boundary conditions for our numerical experiments in Section 3.5.

Next, we consider the Fréchet differentiability of the mapping $u(\sigma)$.

Lemma 3.3.3. *The map $u(\sigma)$ defined by (3.1) is Fréchet differentiable as a mapping from L_{ad} to $H^1_{f_D}(\Omega)$.*

For the proof of this Lemma, we refer to [17]. Using Lemma 4.3.4, we introduce the reduced cost functional

$$\widehat{J}(\sigma) = J(\sigma, u_1(\sigma), u_2(\sigma)), \tag{3.3}$$

where $u_i(\sigma)$, $i = 1, 2$ denotes the unique solution of (3.1) given σ and $f_D^i, i = 1, 2$. The constrained optimization problem (P) can be formulated as an unconstrained one as follows

$$\min_{\sigma \in L_{ad}} \hat{J}(\sigma). \tag{3.4}$$

We next investigate the existence of a minimizer to the CDII-SR problem (P). We first consider the case when $\delta = 0$, i.e., the Perona-Malik term in the functional J is absent.

Proposition 3.3.4. *Let $f_D^1, f_D^2 \in H^{1/2}(\Omega)$ such that $|\nabla u_1| > 0, |\nabla u_2| > 0$ and let $\delta = 0$. Then there exists a triplet $(\sigma^*, u_1^*, u_2^*) \in L_{ad} \times H^1_{f_D^1}(\Omega) \times H^1_{f_D^2}(\Omega)$ such that $u_i^*, i = 1, 2$ are solutions to $\mathcal{L}(\sigma, u_i, f_D^i) = 0, i = 1, 2$ and σ^* minimizes $\hat{J}$ in L_{ad}.*

Proof. Boundedness from below of $\hat{J}$ guarantees the existence of a minimizing sequence (σ^m). Since L_{ad} is reflexive and $\hat{J}$ is sequentially weakly lower semi-continuous, this sequence is bounded. Therefore it contains a weakly convergent subsequence (σ^{m_l})

in L_{ad}, $\sigma^{m_l} \rightharpoonup \sigma^*$. Correspondingly, the sequence $(u_1^{m_l}, u_2^{m_l})$, where $u_i^{m_l} = u_i(\sigma^{m_l})$, is bounded in $H^1_{f_D^1}(\Omega) \times H^1_{f_D^2}(\Omega)$. Therefore the sequence converges weakly to (u_1^*, u_2^*). Now, using the Rellich Kondrachev compactness theorem in $\mathbb{R}^2$, we have that L_{ad} is compactly embedded in $L^2(\Omega)$. This results in a strong convergence of the subsequence σ^{m_l} in $L^2(\Omega)$ to σ^*. We, now, consider the weak formulation of the solutions of the elliptic problem (3.1) and, thus, focus on $\langle \nabla \cdot (\sigma^{m_l} \nabla u_i^{m_l}), \psi \rangle_{L^2(\Omega)}$ for any $\psi \in H_0^1(\Omega)$. Using integration by parts, we have $\langle \nabla \cdot (\sigma^{m_l} \nabla u_i^{m_l}), \psi \rangle_{L^2(\Omega)} = -\langle \sigma^{m_l} \nabla u_i^{m_l}, \nabla \psi \rangle_{L^2(\Omega)}$. From the above discussion, the sequence of products $\sigma^{m_l} \nabla u_i^{m_l}$ is weakly convergent in $L^2(\Omega)$, that is, $\langle \sigma^{m_l} \nabla u_i^{m_l}, \nabla \psi \rangle_{L^2(\Omega)} \to \langle \sigma^* \nabla u_i^*, \nabla \psi \rangle_{L^2(\Omega)}$. With this preparation and using the continuity of the maps $u_i(\sigma)$, it follows that $(u_1^*, u_2^*) = (u_1(\sigma^*), u_2(\sigma^*))$, and the triplet (σ^*, u_1^*, u_2^*) minimizes the objective $\hat{J}$. □

In the case $\delta \neq 0$, we first note that the function $\log(1 + z^2)$ is not convex. Therefore the PM functional, and, hence, the functional $\hat{J}$ in (3.2) is not weakly lower semi-continuous on $W^{1,p}(\Omega)$ for any $1 < p < \infty$. Nevertheless, $\hat{J}$ is a bounded below, lower semi-continuous Lipschitz functional, for which a minimizer exists, provided that L_{ad} is compact.

3.3.1 Characterization of local minima

To characterize the solution of our optimization problem through first-order optimality conditions, we write the reduced functional $\hat{J}$ as

$$\hat{J} = \hat{J}_1 + \hat{J}_2,\ J_i : L_{ad} \to \mathbb{R}^+,\ i = 1, 2,$$

where

$$\begin{aligned} \hat{J}_1(\sigma) &= \frac{\alpha_1}{2} \| e^\sigma |\nabla u_1| - g_1^\delta \|^2_{L^2(\Omega)} + \frac{\alpha_2}{2} \| e^\sigma |\nabla u_2| - g_2^\delta \|^2_{L^2(\Omega)} + \frac{\beta}{2} \|\sigma\|^2_{L^2(\Omega)}, \\ \hat{J}_2(\sigma) &= \gamma \|\sigma\|_{L^1(\Omega)}. \end{aligned} \tag{3.5}$$

Remark 10. The functional $\hat{J}_1$ is smooth and possibly non-convex, while $\hat{J}_2$ is non-smooth and convex.

We next state some properties of the reduced functional $\hat{J}_1(\sigma)$ which can be proved using the arguments in [17, Lemma 3.1].

Proposition 3.3.5. *The reduced functional $\hat{J}_1(\sigma)$ is weakly lower semi-continuous, bounded below and Fréchet differentiable.*

We now define the subdifferential of a non-smooth functional.

Definition 3.3.1 (Subdifferential)**.** If $\hat{J}$ is finite at a point σ, the Fréchet subdifferential of $\hat{J}$ at σ is defined as follows [96]

$$\partial \hat{J}(\bar{\sigma}) := \left\{ \phi \in L_{ad}^* : \liminf_{\sigma \to \bar{\sigma}} \frac{\hat{J}(\sigma) - \hat{J}(\bar{\sigma}) - \langle \phi, \sigma - \bar{\sigma} \rangle}{\|\bar{\sigma} - \sigma\|_2} \geq 0 \right\}, \tag{3.6}$$

where L_{ad}^* is the dual space of L_{ad}. An element $\phi \in \partial \hat{J}(\sigma)$ is called a subdifferential of $\hat{J}$ at σ.

In our setting, we have the following

$$\partial \hat{J}(\sigma) = \nabla \hat{J}_1(\sigma) + \partial \hat{J}_2(\sigma),$$

since $\hat{J}_1$ is Fréchet differentiable by Prop. 4.3.6. Moreover, for each $\alpha > 0$, it holds that

$$\partial(\alpha \hat{J}) = \alpha \partial \hat{J}.$$

The following proposition gives a necessary condition for a local minimum of $\hat{J}$ (see [86]).

Proposition 3.3.6 (Necessary condition)**.** *If $\hat{J} = \hat{J}_1 + \hat{J}_2$, with $\hat{J}_1, \hat{J}_2$ given by (4.14), attains a local minimum at $\sigma^* \in L_{ad}$, then*

$$0 \in \partial \hat{J}(\sigma^*),$$

or equivalently

$$-\nabla \hat{J}_1(\sigma^*) \in \partial \hat{J}_2(\sigma^*).$$

The following variational inequality holds for each $\lambda \in \partial \hat{J}_2(\sigma^*)$ (see [88]).

$$\langle \nabla \hat{J}_1(\sigma^*) + \lambda, \sigma - \sigma^* \rangle \geq 0, \qquad \forall \sigma \in L_{ad}. \tag{3.7}$$

Using the definition of $\hat{J}_2$ in (4.14) and the fact that L_{ad} is reflexive, the inclusion $\lambda \in \partial \hat{J}_2(\sigma^*)$ gives the following characterization of space of λ

$$\lambda \in \Lambda_{ad} := \{\lambda \in L^2(\Omega) : 0 \leq \lambda \leq \gamma, \text{ a.e. in } \Omega\}.$$

A pointwise analysis of the variational inequality (4.16) leads to the existence of a non-negative functions $\lambda^*_{\sigma_l}, \lambda^*_{\sigma_u} \in L^2(\Omega)$ that correspond to Lagrange multipliers for the inequality constraints in L_{ad}. We, thus, have the following first-order optimality system.

Proposition 3.3.7 (First-order necessary conditions)**.** *The optimal solution of the minimization problem (4.13) can be characterized by the existence of* $(\lambda^*, \lambda^*_{\sigma_l}, \lambda^*_{\sigma_u}) \in \Lambda_{ad} \times L^2(\Omega) \times L^2(\Omega)$ *such that*

$$\nabla_\sigma \hat{J}_1(\sigma^*) + \lambda^* + \lambda^*_{\sigma_u} - \lambda^*_{\sigma_l} = 0, \tag{3.8}$$

$$\lambda^*_{\sigma_u} \geq 0,\ \sigma_u - \sigma^* \geq 0,\ \langle \lambda^*_{\sigma_u}, \sigma_u - \sigma^* \rangle = 0, \tag{3.9}$$

$$\lambda^*_{\sigma_l} \geq 0,\ \sigma^* - \sigma_l \geq 0,\ \langle \lambda_{\sigma^*_l}, \sigma^* - \sigma_l \rangle = 0, \tag{3.10}$$

$$\lambda^* = \gamma \ \textit{a.e. on}\ \{x \in \Omega : \sigma^*(x) > 0\}, \tag{3.11}$$

$$0 \leq \lambda^* \leq \gamma \ \textit{a.e. on}\ \{x \in \Omega : \sigma^*(x) = 0\}. \tag{3.12}$$

The conditions (3.9)-(3.12) are known as the complementarity conditions for (σ^*, λ^*).

To determine the gradient $\nabla_\sigma \hat{J}_1$, we use the adjoint approach (see for e.g., [97, 98]). This gives the following reduced gradient of $\hat{J}_1$

$$\nabla_\sigma \hat{J}_1(\sigma^*) = (e^{\sigma^*}|\nabla u_1| - g_1^\delta)|\nabla u_1| + (e^{\sigma^*}|\nabla u_2| - g_2^\delta)|\nabla u_2| + \nabla u_1 \cdot \nabla v_1 + \nabla u_2 \cdot \nabla v_2 + \beta\sigma^*, \tag{3.13}$$

where u_1, u_2 satisfy the forward equations $\mathcal{L}(u_1, \sigma^*, f_D^1) = 0$, $\mathcal{L}(u_2, \sigma^*, f_D^2) = 0$, respectively, and v_1, v_2 satisfy the adjoint equations

$$\begin{aligned} -\nabla \cdot (e^{\sigma^*}\nabla v_1) &= \nabla \cdot \left[e^{\sigma^*}(e^{\sigma^*}|\nabla u_1| - g_1^\delta)(\nabla u_1) \right] \text{ in } \Omega, \\ v_1|_\Gamma &= 0, \end{aligned} \tag{3.14}$$

$$\begin{aligned} -\nabla \cdot (e^{\sigma^*}\nabla v_2) &= \nabla \cdot \left[e^{\sigma^*}(e^{\sigma^*}|\nabla u_2| - g_2^\delta)(\nabla u_2) \right] \text{ in } \Omega, \\ v_2|_\Gamma &= 0. \end{aligned} \tag{3.15}$$

The complementarity conditions (3.9)-(3.12) can be rewritten in a compact form as follows. Define

$$\mu^* = \lambda^* + \lambda^*_{\sigma_u} - \lambda^*_{\sigma_l}. \tag{3.16}$$

Then the triplet $(\lambda^*, \lambda^*_{\sigma_l}, \lambda^*_{\sigma_u})$ is obtained by solving the following equations

$$\begin{aligned} \lambda^* &= \min(\gamma, \max(0, \mu^*)), \\ \lambda^*_{\sigma_l} &= -\min(0, \mu^* + \gamma), \\ \lambda^*_{\sigma_u} &= \max(0, \mu^* - \gamma), \end{aligned} \tag{3.17}$$

(see [88]). For each $k \in \mathbb{R}^+$, define the following quantity

$$\begin{aligned} E(\sigma^*, \mu^*) = \sigma^* &- \max\{0, \sigma^* + k(\mu^* - \gamma)\} + \max\{0, \sigma^* - \sigma_u + k(\mu^* - \gamma)\} \\ &- \min\{0, \sigma^* + k(\mu^* + \gamma)\} + \min\{0, \sigma^* - \sigma_l + k(\mu^* + \gamma)\}. \end{aligned}$$

The following lemma determines the complementarity conditions (3.9)-(3.12) in terms of E (see [88, Lemma 2.2]).

Lemma 3.3.8. *The complementarity conditions (3.9)-(3.12) are equivalent to the following*

$$E(\sigma^*, \mu^*) = 0, \tag{3.18}$$

where μ is defined in (3.16).

Using the gradients in (4.27) and Lemma 4.3.9, the optimality conditions (3.8)-(3.12) for the CDII-SR problem can be rewritten as follows

Proposition 3.3.9. *A local minimizer (u_1, u_2, σ^*) of the problem (P) can be characterized by the existence of $(v_1, v_2, \mu^*) \in H_0^1(\Omega) \times H_0^1(\Omega) \times L_{ad}$, such that the following system is satisfied*

$$\begin{aligned}
-\nabla \cdot (e^{\sigma^*} \nabla u_1) &= 0 \ \textit{in}\ \Omega, \\
u_1|_\Gamma &= f_D^1, \\
-\nabla \cdot (e^{\sigma^*} \nabla v_1) &= \nabla \cdot \left[e^{\sigma^*} (e^{\sigma^*} |\nabla u_1| - g_1^\delta)(\nabla u_1) \right] \ \textit{in}\ \Omega, \\
v_1|_\Gamma &= 0, \\
-\nabla \cdot (e^{\sigma^*} \nabla u_2) &= 0 \ \textit{in}\ \Omega, \\
u_2|_\Gamma &= f_D^2, \\
-\nabla \cdot (e^{\sigma^*} \nabla v_2) &= \nabla \cdot \left[e^{\sigma^*} (e^{\sigma^*} |\nabla u_2| - g_2^\delta)(\nabla u_2) \right] \ \textit{in}\ \Omega, \\
v_2|_\Gamma &= 0,
\end{aligned}$$

$$(e^{\sigma^*} |\nabla u_1| - g_1^\delta) e^{\sigma^*} |\nabla u_1| + (e^{\sigma^*} |\nabla u_2| - g_2^\delta) e^{\sigma^*} |\nabla u_2| + \nabla u_1 \cdot \nabla v_1 + \nabla u_2 \cdot \nabla v_2 + \beta \sigma^* + \mu^* = 0, \ \textit{a.e. in}\ \Omega$$

$$E(\sigma^*, \mu^*) = 0, \qquad \textit{a.e. in}\ \Omega. \tag{3.19}$$

3.4 Numerical solution of the CDII-SR problem

In this section, we discuss numerical optimization and approximation schemes to solve the CDII-SR problem. In this context, we first discuss proximal methods that consists of identifying a smooth and a non-smooth part in the reduced objective $\hat{J}(\sigma)$. Thus, we consider the following optimization problem

$$\min_{\sigma \in L_{ad}} \hat{J}(\sigma) := \hat{J}_1(\sigma) + \hat{J}_2(\sigma). \tag{3.20}$$

We assume that $\nabla_\sigma \hat{J}_1(\sigma)$, given in (4.27) is Lipschitz continuous and the upper bound for the Lipschitz constant is obtained using a backtracking search scheme, which will be discussed later. Also, from (4.14), we have that $\hat{J}_2(\sigma)$ is a continuous, convex, and nondifferentiable functional. The formulation of proximal methods depends, essentially, on the following lemma [86]

Lemma 3.4.1. *Let $\hat{J}_1$ be differentiable with a Lipschitz continuous gradient with Lipschitz constant $L(\hat{J}_1)$. Then the following holds*

$$\hat{J}_1(\sigma) \leq \hat{J}_1(\widetilde{\sigma}) + \left\langle \nabla \hat{J}_1(\widetilde{\sigma}), \sigma - \widetilde{\sigma} \right\rangle + \frac{L}{2}\|\sigma - \widetilde{\sigma}\|^2, \quad \forall \sigma, \widetilde{\sigma} \in L_{ad}, \tag{3.21}$$

for all $L \geq L(\hat{J}_1) > 0$.

We note that $L := L(\hat{J}_1)$ represents the smallest value of L such that (3.21) holds true.

In a proximal scheme, one usually minimizes an upper bound of the objective functional at each iteration, instead of minimizing the functional directly. From Lemma 3.4.1, we obtain the following

$$\min_{\sigma \in L_{ad}} \left\{ \hat{J}_1(\sigma) + \hat{J}_2(\sigma) \right\} \leq \min_{\sigma \in L_{ad}} \left\{ \hat{J}_1(\widetilde{\sigma}) + \left\langle \nabla \hat{J}_1(y), \sigma - \widetilde{\sigma} \right\rangle + \frac{L}{2}\|\sigma - \widetilde{\sigma}\|^2 + \hat{J}_2(\sigma) \right\},$$

where equality holds if $\sigma = \widetilde{\sigma}$. Furthermore, we have the following equation

$$\arg\min_{\sigma \in L_{ad}} \left\{ \hat{J}_1(\widetilde{\sigma}) + \left\langle \nabla \hat{J}_1(\widetilde{\sigma}), \sigma - \widetilde{\sigma} \right\rangle + \frac{L}{2}\|\sigma - \widetilde{\sigma}\|^2 + \hat{J}_2(\sigma) \right\}$$

$$= \operatorname*{arg\,min}_{\sigma \in L_{ad}} \left\{ \frac{L}{2} \left\| \sigma - \left(\widetilde{\sigma} - \frac{1}{L} \nabla \hat{J}_1(\widetilde{\sigma}) \right) \right\|^2 + \hat{J}_2(\sigma) \right\}. \tag{3.22}$$

Using the definition of $\hat{J}_2(\sigma) = \gamma \|\sigma\|_{L^1(\Omega)}$, we have the following lemma from [99] that helps in characterizing the solution of (3.22).

Lemma 3.4.2. *The following equation holds*

$$\operatorname*{arg\,min}_{\sigma \in L_{ad}} \left\{ \tau \|\sigma\|_{L^1} + \frac{1}{2} \|\sigma - \widetilde{\sigma}\|^2 \right\} = \mathbb{S}_{\tau}^{L_{ad}}(\widetilde{\sigma}) \quad \textit{for any } \; \widetilde{\sigma} \in L^2(\Omega),$$

where the left-hand side represents the proximal function and the projected soft thresholding function on the right-hand side is defined as follows

$$\mathbb{S}_{\tau}^{L_{ad}}(\widetilde{\sigma}) := \begin{cases} \min\{\widetilde{\sigma} - \tau, \sigma_u\} & \textit{on } \{(x,y) \in \Omega : \, \widetilde{\sigma}(x,y) > \tau\} \\ 0 & \textit{on } \{(x,y) \in \Omega : \, |\widetilde{\sigma}(x,y)| \leq \tau\} \\ \max\{\widetilde{\sigma} + \tau, \sigma_l\} & \textit{on } \{(x,y) \in \Omega : \, \widetilde{\sigma}(x,y) < -\tau\} \end{cases}. \tag{3.23}$$

Using this lemma, the solution to (3.22) is given by

$$\operatorname*{arg\,min}_{\sigma \in L_{ad}} \left\{ \hat{J}_2(\sigma) + \frac{L}{2} \left\| \sigma - \left(\widetilde{\sigma} - \frac{1}{L} \nabla \hat{J}_1(\widetilde{\sigma}) \right) \right\|^2 \right\} = \mathbb{S}_{\frac{\gamma}{L}}^{L_{ad}} \left(\widetilde{\sigma} - \frac{1}{L} \nabla \hat{J}_1(\widetilde{\sigma}) \right).$$

This gives rise to the following iterative scheme

$$\sigma_{k+1} \leftarrow \mathbb{S}_{\frac{\gamma}{L}}^{L_{ad}} \left(\sigma_k - \frac{1}{L} \nabla \hat{J}_1(\sigma_k) \right),$$

starting from a given σ_0 and is known as the iterative shrinkage-thresholding algorithm (ISTA) scheme [99]. We note that the argument of $\mathbb{S}_{\frac{\gamma}{L}}^{L_{ad}}$ represents a gradient update in a steepest descent scheme with a fixed step size $s = 1/L$ in conjunction with a regularized PM filter [86]. Further, to accelerate the ISTA scheme described above, one can consider a sequence $\{t_k, v_k\}$ [99, 100] such that

$$t_0 = 1, \qquad t_k := 1 + \sqrt{1 + 4t_{k-1}^2}/2, \tag{3.24}$$

and

$$v_0 := \sigma_0, \qquad v_k := \sigma_k + \frac{(t_{k-1} - 1)}{t_k}(\sigma_k - \sigma_{k-1}). \tag{3.25}$$

This gives us the following update for the optimization variable σ_k

$$\sigma_{k+1} \leftarrow \mathbb{S}^{L_{ad}}_{\frac{\gamma}{L}} \left(v_k - \frac{1}{L} \nabla \hat{J}_1(v_k) \right). \tag{3.26}$$

Replacing v_k in (4.41) with (4.39), and assuming that $\nabla \hat{J}_1(\sigma_k) \approx \nabla \hat{J}_1(v_k)$, we obtain the following iterative scheme [100]

$$\sigma_{k+1} \leftarrow \mathbb{S}^{L_{ad}}_{\gamma\, s_k} \left(\sigma_k - s_k \nabla_\sigma \hat{J}_1(\sigma_k) + \theta_k \left(\sigma_k - \sigma_{k-1}\right) \right), \tag{3.27}$$

where $\sigma_{-1} = \sigma_0$.

The above discussion is valid for any $L \geq L(\hat{J}_1)$. However, since the quantity $s = 1/L$ represents the step size in a gradient update, we use a backtracking line search algorithm to determine the optimal step size in each iteration. This leads to the computation of an upper bound L_k that satisfies $L_k \geq L(\hat{J}_1)$ at each iteration step. Thus, we define our variable step size as $s_k = 1/L_k$ and substitute τ in (3.23) with γs_k. The variable step size causes the factor $\frac{(t_{k-1}-1)}{t_k}$ in (4.39) to be non-optimal and we replace it by the fixed inertial parameter θ. This leads to a variable inertial proximal (VIP) scheme, which is described in Algorithm **??**.

With our VIP scheme, we aim at determining an optimal $\sigma \in L_{ad} \subset H^1_0(\Omega)$. But in the update step of the algorithm, we have the argument of the thresholding function $\mathbb{S}^{L_{ad}}_{\frac{\gamma}{L}}$ as $\sigma_k - s_k \nabla_\sigma \hat{J}_1(\sigma_k)$. The term $\nabla_\sigma \hat{J}_1(\sigma_k)$ is only in $L^2(\Omega)$ and the resulting update gives us the argument of $\mathbb{S}^{L_{ad}}_{\frac{\gamma}{L}}$ in $L^2(\Omega)$, which is not desired. We, thus, use the H^1 gradient instead of the L^2 gradient, which are related by the equation $((\nabla_\sigma \hat{J}_1)_{H^1}, v)_{H^1(\Omega)} = (\nabla_\sigma \hat{J}_1, v)_{L^2(\Omega)}$ for all $v \in H^1(\Omega)$. But such a H^1 gradient results in a highly diffused σ with blurred edges. We, instead, consider a weighted H^1 product that represents a suitable denoising of the $\nabla_\sigma \hat{J}_1(\sigma)$. We apply the denoising operator $R(c) = (I - c\,\Delta)^{-1}$ with a small denoising parameter c (we take $c = 10^{-3}$) and define $(\nabla_\sigma \hat{J}_1)_{H^1} = R(c)\,\nabla_\sigma \hat{J}_1$. Note that a higher value of c results in a greater blurring of the edges along with noise removal. On the other hand, since the PM term in the functional J promotes better resolution with edge-enhancement, we choose the value of c in proportion to the weight η of the PM functional term (we choose $\eta = 10^{-2}$).

We summarize the variable inertial proximal (VIP) scheme for our CDII-SR setup in algorithm below

Variable inertial proximal (VIP) method

```
Input: $\beta$, $\hat{J}_1$, $\sigma_0 = \sigma_{-1}$, $L_{ad}$, $TOL$, $n > 1$, $L_0 > 0$
Initialize: $E_0 = 1$, $k = 0$, choose $\theta \in (0,1)$ and $c_1 < 2$ and $c_2 > 0$;
While$\|E_{k-1}\| > TOL$ do
Compute $\nabla_\sigma \hat{J}_1(\sigma_k)$
Backtracking: Find the smallest nonnegative integer $i$ such that with
    $\tilde{L} = n^i L_{k-1}$
    $$\hat{J}_1(\tilde{\sigma}) \leq \hat{J}_1(\sigma_k) + \left\langle \nabla_\sigma \hat{J}_1(\sigma_k), \tilde{\sigma} - \sigma_k \right\rangle + \frac{\tilde{L}}{2}\|\tilde{\sigma} - \sigma_k\|^2$$
    where $\tilde{\sigma} = \mathbb{S}^{L_{ad}}_{\gamma\,s}\left(\sigma_k - s\,(\nabla_\sigma \hat{J}_1)_{H^1}(\sigma_k) + \theta(\sigma_k - \sigma_{k-1})\right)$, $s = c_1(1-\theta)/(\tilde{L} + 2c_2)$,
Set $L_k = \tilde{L}$ and $s_k = c_1(1-\theta)/(L_k + 2c_2)$.
$\sigma_{k+1} = \mathbb{S}^{L_{ad}}_{\gamma\,s_k}\left(\sigma_k - s_k\,(\nabla_\sigma \hat{J}_1)_{H^1}(\sigma_k) + \theta(\sigma_k - \sigma_{k-1})\right)$
$\mu_k = -\alpha\sigma_k - (\nabla_\sigma \hat{J}_1)_{H^1}(\sigma_k)$
$E_k = E(\sigma_k, \mu_k)$
$k = k + 1$
end
```

In the VIP algorithm, we need to compute the reduced gradient $\nabla_\sigma \hat{J}_1$. This, in turn, requires an accurate numerical solution of the forward and the corresponding adjoint EIT problems as given in Proposition 4.3.10. For the forward EIT equation (3.1), we use the cell-nodal finite-difference approximation. We consider a sequence of uniform grids $\{\Omega_h\}_{h>0}$ given by

$$\Omega_h = \{(x_i, y_j) \in \mathbb{R}^2 : (x_i, y_j) = (a + ih, a + jh),\ (i, j) \in \{0, \dots, N\}^2\} \cap \Omega,$$

where N represents the number of cells in each direction and $h = \dfrac{(b-a)}{N}$ is the mesh size. The corresponding cell-nodal scheme for (3.1), at the grid point (x_i, y_j), is given as follows

$$\frac{1}{h^2}\left\{ (e^{\sigma_{i+1/2,j}} + e^{\sigma_{i-1/2,j}} + e^{\sigma_{i,j+1/2}} + e^{\sigma_{i,j-1/2}})u_{i,j} \right. \\ \left. - e^{\sigma_{i+1/2,j}}u_{i+1,j} - e^{\sigma_{i-1/2,j}}u_{i-1,j} - e^{\sigma_{i,j+1/2}}u_{i,j+1} - e^{\sigma_{i,j-1/2}}u_{i,j-1} \right\} = 0, \qquad 1 \leq i, j \leq N-1, \tag{3.28}$$

where $\sigma_{i\pm1,j} = \sigma(x_i \pm h, y_j)$, $\sigma_{i,j\pm1} = \sigma(x_i, y_j \pm h)$. The required intermediate values of σ are computed as follows

$$\sigma_{i\pm1/2,j} = \frac{1}{2}\Bigg(\sigma_{i\pm1,j} + \sigma_{i,j}\Bigg) \quad \text{and} \quad \sigma_{i,j\pm1/2} = \frac{1}{2}\Bigg(\sigma_{i,j\pm1} + \sigma_{i,j}\Bigg).$$

The Dirichlet boundary data f_D is included in the usual way in the right-hand side of the algebraic equation.

For the adjoint equations (4.28) and (4.29), we first note that the cell nodal finite difference scheme is not applicable to the right-hand side term in both the equations as they are of the form $G = \nabla \cdot F$, where F is in $L^2(\Omega)$. We modify the cell nodal scheme by replacing the nodal value of G at (x_i, y_j) with a cell average of G given as follows

$$G_a = \frac{1}{h^2}\int_{C_{ij}} G(x, y)\ dxdy,$$

where the cell C_{ij} is defined by

$$C_{ij} := \left(x_i - \frac{h}{2}, x_i + \frac{h}{2}\right) \times \left(y_j - \frac{h}{2}, y_j + \frac{h}{2}\right), \quad 1 \leq i, j \leq N - 1.$$

Since $G = \nabla \cdot F$, using the divergence theorem we have

$$G_a = \frac{1}{h^2} \int_{C_{ij}} \nabla \cdot F(x, y) \; dxdy = \frac{1}{h^2} \int_{\partial C_{ij}} F(x, y).n \; ds$$

The above integral can be approximated with a midpoint quadrature rule along each edge of C_{ij}. This results in the following approximation

$$G_a = \frac{F^1_{i+1/2,j} - F^1_{i-1/2,j}}{h} + \frac{F^2_{i,j+1/2} - F^2_{i,j-1/2}}{h},$$

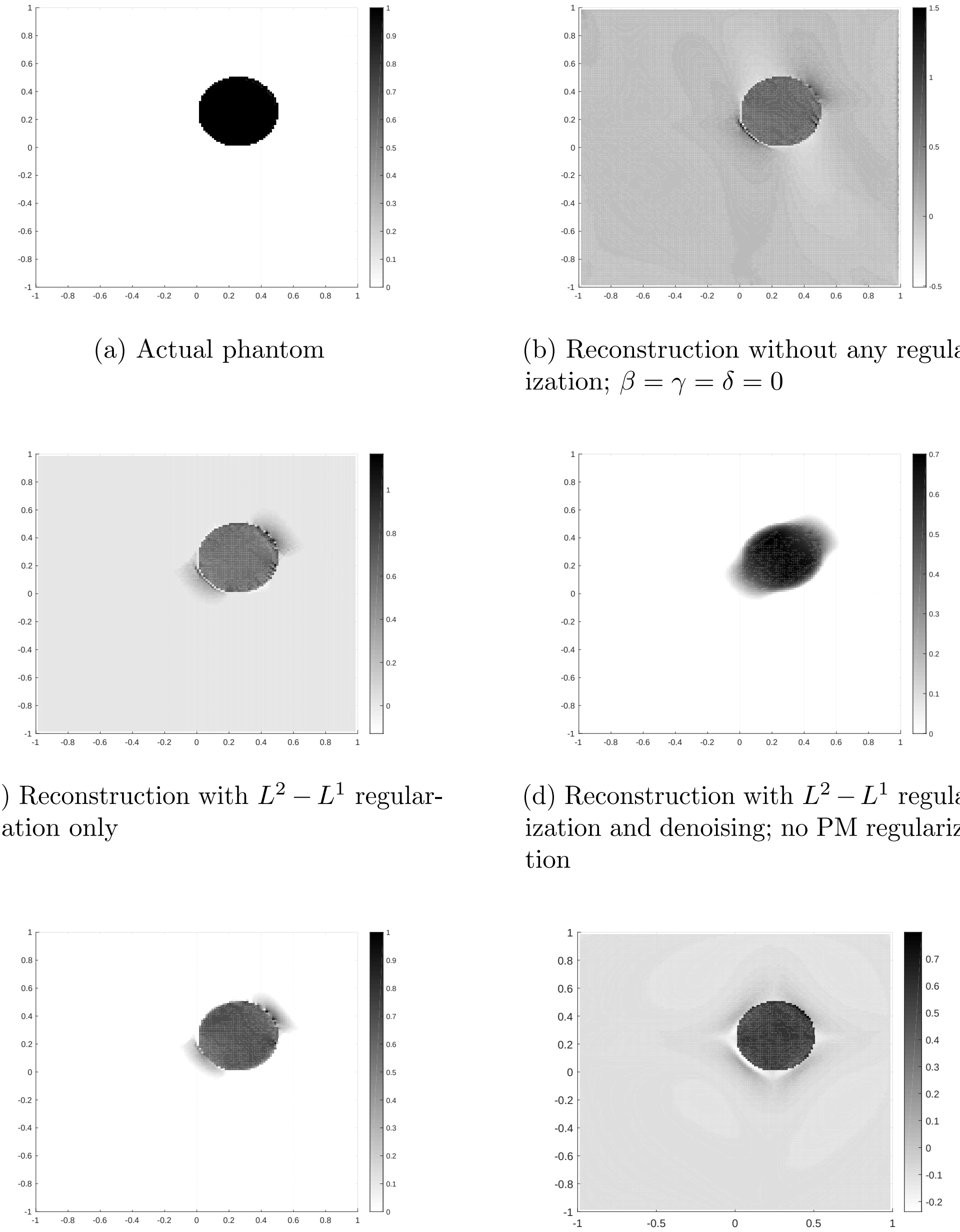
where $F = (F^1, F^2)$.

3.5 Numerical experiments

In this section, we validate our CDII-SR framework using different experiments that validate the choice of different features in our formulation and demonstrate its effectiveness in reconstructing a wide variety of objects. We choose the two boundary conditions as $f^1_D = x, \; f^2_D = y$ on Γ, which is the boundary of $\Omega = (-1, 1) \times (-1, 1)$. The weights in the functional (3.2) are chosen as follows: $\alpha_1 = \alpha_2 = 1.0$, $\beta = 0.3$, $\gamma = 0.01$, $\delta = 0.01$. The value of the denoising parameter is $c = 0.001$. The parameters of our VIP scheme are chose as $\theta = 0.5$, $c_1 = 1.9$, $c_2 = 0.001$, $TOL = 10^-4$ with the maximum number of iterations as 20. Even though there is a specified tolerance for the termination of the algorithm, due to the high non-linearity of the problem, our VIP scheme terminates due to the maximum number of iterations.

In all the experiments, the domain Ω is uniformly discretized into $N = 150$ subintervals in both the x and y directions with $h = 0.013$. The generation of the synthetic interior electric field data H^δ is done as follows: we first solve for u in (3.1) with given value of σ on a finer mesh with $N = 400$ using the finite difference method outlined in Section 3.4. Then, we compute ∇u with one-sided finite differences to obtain H^δ on the finer mesh. In the final step, we restrict the obtained H^δ onto the coarser mesh with $N = 150$ and choose this as our given data to which we also add noise in some of the experiments.

In Test Case 1, we consider a disk phantom for σ that is represented by a disk centered at $(0.25, 0.25)$ with radius 0.25. The value of σ inside the disk is 1 with the background value chosen as 0. The plots of the actual σ and the reconstructed σ are shown in Figure 4.1.



(a) Actual phantom

(b) Reconstruction without any regularization; $\beta = \gamma = \delta = 0$

(c) Reconstruction with $L^2 - L^1$ regularization only

(d) Reconstruction with $L^2 - L^1$ regularization and denoising; no PM regularization

(e) Reconstruction with $L^2 - L^1$ regularization, denoising and PM regularization

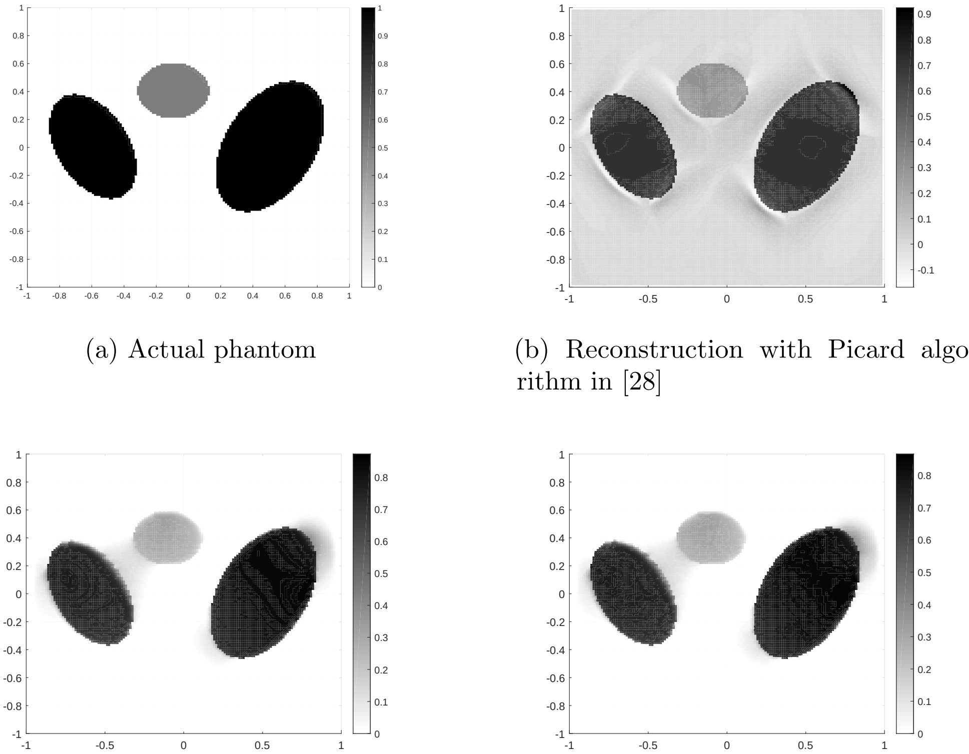
(f) Reconstruction with Picard algorithm in [28]

Figure 3.1: Test Case 1- The actual and reconstructed disk with different choices of the values of the regularization weights.

Figure 3.1b shows the reconstruction of $\sigma \in L_{ad}$ without any regularization terms, i.e. $\beta = \gamma = \delta = 0$ and no denoising, i.e., $c = 0$. Presence of strong artifacts can be observed in this case, which is inherent to the inverse problem and not the algorithm. A study of the pattern of such artifacts are very challenging and is out of the scope of the paper.

Figure 3.1c shows the result of CDII-SR reconstruction without the denoising and the Perona-Malik regularization term, $c = \delta = 0$, but with the $L^2 - L^1$ regularization. We observe that the artifacts are reduced to some extent, but are still present. Figure 3.1d shows the reconstruction with the $L^2 - L^1$ regularization and the H^1 denoising but without the PM filter. In this case, we observe that the artifacts diminish by a huge amount, but the edges are more blunt and the value of σ is lowered, leading to a loss of resolution and contrast. We correct this loss using the PM regularization term as can be observed in Figure 3.1e, where the edges are fairly well seen and the recovered parameter values are very close to the true ones. We also compare our results with the reconstruction in Figure 3.1f obtained with the Picard algorithm proposed in [28]. We observe a lot of artifacts and a significant loss of contrast in Figure 3.1f in comparison to the reconstruction shown in Figure 3.1e, which suggests that our CDII-SR scheme outperforms the Picard scheme.

In Test Case 2, we consider the heart and lung phantom for the true σ. It consists of two ellipses representing lungs with the value is 1 and a circular region representing the heart with the value is 0.5. The background value of $\sigma = 0$. The plots of the actual σ and the reconstructed σ are shown in Figure 3.2



(a) Actual phantom

(b) Reconstruction with Picard algorithm in [28]

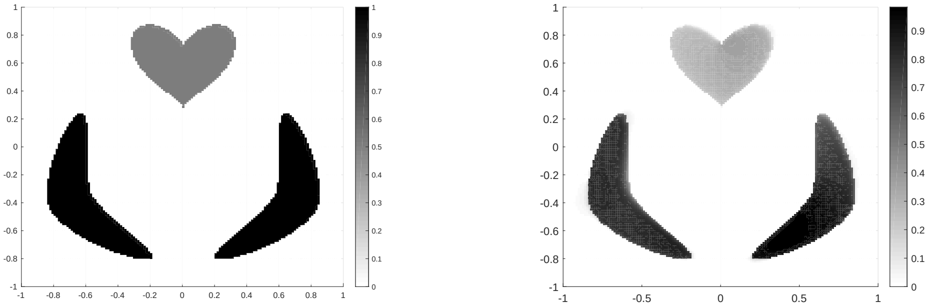
(c) Reconstruction using CDII-SR scheme

(d) Reconstruction using CDII-SR scheme with 10% Gaussian noise

Figure 3.2: Test Case 2- The actual and reconstructed heart and lung phantom with and without noisy data.

We see from Figure 3.2c that our CDII-SR algorithm results in reconstruction of σ with high resolution. Moreover, the values of the reconstructed σ are very close to the true values, with the background value exactly equals to 0, which implies a high contrast reconstruction. We further compare our results to the Picard algorithm proposed in [28]. The corresponding reconstruction is shown in Figure 3.2b. It can be seen that the CDII-SR provides a better contrast image, yet maintaining the same resolution as that of the Picard scheme. Also, there are far more artifacts through the Picard reconstruction method whereas the sparsity assumption and the H^1 denoising in CDII-SR scheme results in an image with very less artifacts. Further, to test the robustness of our method, we introduce 10% multiplicative Gaussian noise in the interior data H^δ, which is fed as input to our CDII-SR algorithm. The corresponding reconstruction is shown in Figure 3.2d, which again possesses high contrast and high resolution, demonstrating that the CDII-SR algorithm is robust even in the presence of noisy data.

In Test Case 3, we consider a more generalized form of the heart and lung phantom. The heart is represented by a cardioid with the value of $\sigma = 0.5$. The two lungs are represented by "boomerangs" with the value of $\sigma = 1.0$. The plots of the actual and the reconstructed σ are shown in Figure 3.3.



(a) Actual phantom

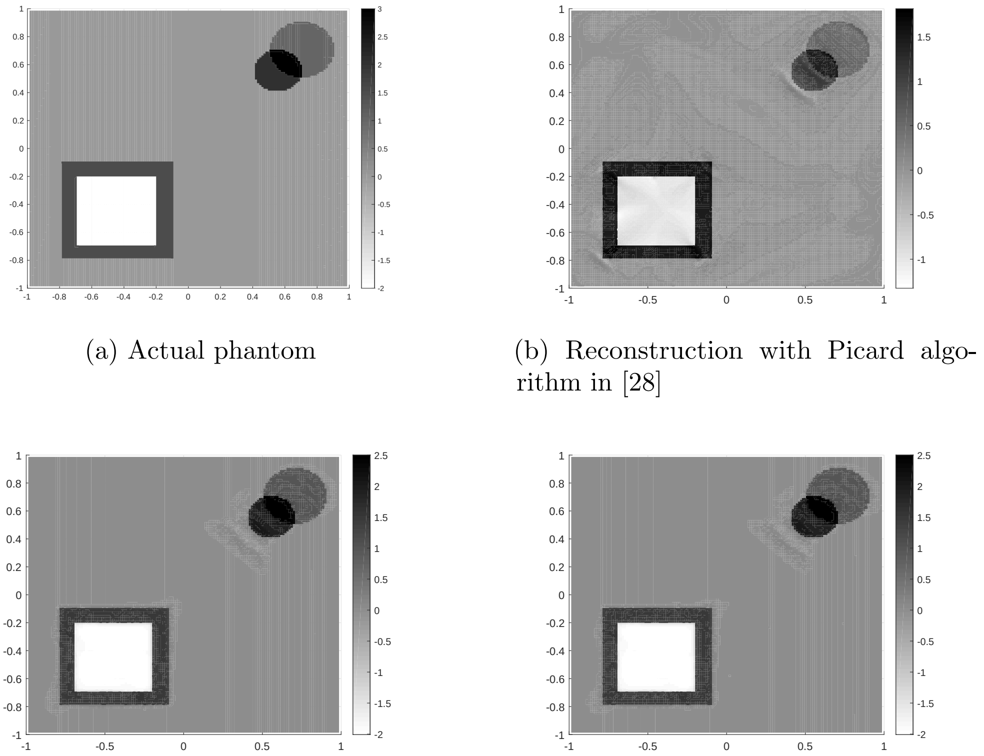
(b) Reconstruction using CDII-SR scheme

Figure 3.3: Test Case 3- The actual and reconstructed modified heart and lung phantom.

We again note good quality reconstructions, specially at the corners of the phantom, which demonstrates the effectiveness of the CDII-SR scheme for irregular shaped phantoms.

In Test Case 4, we consider a combination of phantoms, where one is supported on a square annulus $S_a = \{(x, y) \in \mathbb{R}^2 : -0.8 < x < -0.7, -0.2 < x < -0.1, -0.8 <$

$y < -0.7, -0.2 < y < -0.1\}$ with $\sigma = 3.0$; the other one consists of 2 disks centered at $(0.7, 0.7)$ with radius 0.2 and $\sigma = 1.0$ and at $(0.55, 0.55)$ with radius 0.15 and $\sigma = 2.0$. The value of σ inside the square annulus has a value -2.0. The plots of the actual and the reconstructed σ is shown in Figure 3.4



(a) Actual phantom

(b) Reconstruction with Picard algorithm in [28]

(c) Reconstruction using CDII-SR scheme

(d) Reconstruction using CDII-SR scheme with 10% Gaussian noise

Figure 3.4: Test Case 4- The actual and reconstructed mixed phantom with and without noisy data.

Figure 3.4c shows the reconstruction with the CDII-SR algorithm. We compare the results, as shown in Figure 3.4b, obtained with the Picard algorithm in [28] and note that the CDII-SR scheme results in a superior contrast image while preserving the same resolution. It should also be noted that the value inside the square annulus is recovered to be close to the actual negative value and the background is obtained to be exactly 0 due to the sparsity assumption, which suggests that the CDII-SR algorithm is robust and more effective in reconstructing objects with holes and inclusions. We observe similar features in the reconstruction even in the presence of 10% Gaussian noise in the interior data as shown in Figure 3.4d.

3.6 Conclusions

In this chapter, we propose a new framework to facilitate high contrast and high resolution reconstructions in CDII. Our framework is based on formulating the CDII inverse problem as a PDE-constrained optimization problem, where we minimize an objective functional comprising of least square interior data fitting terms corresponding to two boundary voltage measurements, a L^1 penalization term of the log-conductivity that helps promotes sparsity patterns and a PM filtering term to sharpen the edges.

CHAPTER 4

Sparsity-based nonlinear reconstruction of optical parameters in two-photon photoacoustic computed tomography

4.1 Introduction

[1] The mathematical formulation of 2P-PACT was first introduced in [44, 45], where the authors consider an optically absorbing and scattering medium $\Omega \subset \mathbb{R}^n$ ($n \geq 2$). Denoting the density of photons at a point $x \in \Omega$ as $u(x)$, it was shown that $u(x)$ solves the following semi-linear diffusion equation

$$\begin{aligned} -\nabla \cdot (D(x)\nabla u(x)) + \sigma(x)u(x) + \mu(x)|u(x)|u(x) = 0, \qquad & \text{in} \quad \Omega, \\ u(x) = g(x), \qquad & \text{on } \partial\Omega, \end{aligned} \tag{4.1}$$

where $D(x)$ denotes the diffusion coefficient, $\sigma(x)$ and $\mu(x)$ represent the single-photon and the two-photon absorption coefficients respectively, and the function $g(x)$ is the illumination pattern on the boundary $\partial\Omega$. The term $\mu(x)|u(x)|$ is the total two-photon absorption coefficient, where the absolute value of u is taken to ensure that the total two-photon absorption coefficient is non-negative [45].

The medium Ω heats up due to absorption of some portion of incoming photons that results in thermal expansion of the medium. The medium cools down after photons leave the medium and this results in contraction of the medium, which gives rise to acoustic waves. This effect is known as the photoacoustic effect. This

[1] The content of this chapter has taken from [101], Gupta, M., Mishra, R. K., and Roy, S. (2021). Sparsity-based nonlinear reconstruction of optical parameters in two-photon photoacoustic computed tomography. Inverse Problems, 37(4), 044001.

photoacoustic effect generates an acoustic wave pressure field $\mathcal{H}^{\sigma,\mu}$ is given by (see [31, 102])

$$\mathcal{H}^{\sigma,\mu}(x) = \Gamma(x)\left[\sigma(x)u(x) + \mu(x)|u(x)|u(x)\right], \quad \text{for } x \in \Omega, \tag{4.2}$$

where Γ is the Grüneisen coefficient that determines the efficiency of the photoacoustic effect. The aim is to recover the optical properties of the medium Ω from the measured acoustic wave signals on the surface of the medium. In this process, the first step involves the recovery of the initial acoustic wave pressure field $\mathcal{H}^{\sigma,\mu}$ from measured data, as usually done in a standard PAT. In the second step of 2P-PACT, the goal is to reconstruct the optical coefficients D, σ, μ and Γ from the information of internal data $\mathcal{H}^{\sigma,\mu}$. This step is usually known as the quantitative step. Recently, the experimental aspect of 2P-PACT have been studied by several authors and it has been shown that the effect of two-photon absorption can be measured accurately, we refer to [42, 43, 47, 103, 104, 105] for detailed discussions. Thus, we assume that the first step in the 2P-PACT process has been accomplished to obtain the initial acoustic wave pressure field $\mathcal{H}^{\sigma,\mu}$. For the second step of recovery of the optical coefficients, detailed mathematical and numerical analysis has been done in very few works [44, 45, 106]. It has been shown in [45] that simultaneous reconstruction of all the four coefficients D, σ, μ, Γ is not possible. In [44, 45], the authors show that given D, Γ, one of σ and μ can be reconstructed with internal data corresponding to one boundary illumination pattern and reconstruction of both coefficients require two sets of internal datum. The authors also present two reconstruction algorithms for reconstructing σ, μ.

There are three major drawbacks of the existing reconstruction algorithms for 2P-PACT: First, four sets of internal datum are used for reconstructing two coefficients. While this gives better reconstructions, it is not conforming with the theoretical requirement of only two sets of internal datum. Secondly, in the presence of 5% noise in the data, the reconstructions of μ exhibit severe artifacts. Thirdly, there is no evidence of the algorithms performing well to reconstruct complex objects with high contrast such as holes and inclusions. In this article, we aim at using a robust computational framework that has the ability to provide high contrast and high resolution reconstructions of objects with holes and inclusions. The framework is based on a non-linear PDE-constrained optimization technique, developed recently [107, 108, 86] to study the aforementioned hybrid inverse problem for 2P-PACT. We start by formulating a minimization problem where we aim to determine σ and μ given the interior acoustic wave pressure field $\mathcal{H}^{\sigma,\mu}$. Additionally, we also assume that the variations in the values of absorption coefficients from known background absorption coefficients demonstrate sparsity patterns. These patterns arise frequently in several tomographic imaging scenarios, for e.g. in blood vessel tomographic reconstructions [109]. One could use a Total Variation (TV) regularization for sparse reconstructions and denoising. However, though TV regularization is more appropriate for piecewise smooth patterns, for general objects like blood vessels, where optical coefficients are not piecewise constants, L^1 regularization has been observed to outperform TV regularization for obtaining sparsity (for e.g. see [110]). Thus, sparsity is incorporated in our model through an L^1 regularization term in our objective functional. An H^1 regularization term is also introduced in the functional that helps reducing artifacts. We provide a comprehensive theoretical analysis of our optimization framework. We provide a new proof for the existence of solutions of (4.1) with higher regularity, under the assumption that $g \geq 0$, using a fixed point approach. We also prove the existence of minimizers of our minimization problem. We solve the optimization problem using a variable inertial proximal scheme that efficiently handles the non-differentiable L^1 regularization term in the objective functional. Finally, we demonstrate the applicability of our reconstruction approach by implementing scheme to several examples.

The chapter is organized as follows: In Section 4.2, we formulate the minimization problem for the 2P-PACT reconstruction problem. In Section 4.3, we present some theoretical results about our optimization problem and we also characterize the optimality system. The numerical schemes to solve the forward problem and the optimization problem are discussed in Section 4.4. In Section 4.5, we present simulation results of our 2P-PACT framework. A section on conclusions completes our work.

4.2 A minimization problem

In this section, we describe the minimization problem corresponding to the 2P-PACT reconstruction problem. We assume Ω to be bounded domain in $\mathbb{R}^2$. The authors in [45] show that, under the assumptions of the boundary function $g \geq 0$, there exists a non-negative solution u of (4.1) in $H^1_g(\Omega)$. Since g represents the density of photons, g is non-negative. Therefore, instead of the photon propagation equation (4.1), we consider the following boundary value problem

$$\begin{aligned} -\nabla \cdot (D(x)\nabla u(x)) + \sigma(x)u(x) + \mu(x)u^2(x) &= 0, \qquad \text{in } \Omega, \\ u(x) &= g(x) \quad \text{on } \partial\Omega \end{aligned} \tag{4.3}$$

as the model for photon propagation in Ω. We assume that the diffusion coefficient $D \in W^{1,\infty}(\Omega)$ is known. Throughout the chapter, we assume that the absorption coefficients σ and μ belong to the function spaces L^σ_{ad} and L^μ_{ad} respectively, where

$$L^\sigma_{ad} = \{q(x) \in H^1(\Omega) : a_\sigma \leq q(x) \leq b_\sigma,\ \forall x \in \Omega,\ a_\sigma, b_\sigma > 0\},$$

$$L^\mu_{ad} = \{q(x) \in H^1(\Omega) : a_\mu \leq q(x) \leq b_\mu,\ \forall x \in \Omega,\ a_\mu, b_\mu > 0\}.$$

Then the aim is to recover both absorption coefficients σ and μ from the knowledge of two sets boundary illumination functions g_1, g_2 and the corresponding initial acoustic wave pressure field $\mathcal{H}_1^{\sigma,\mu}, \mathcal{H}_2^{\sigma,\mu}$, where

$$\mathcal{H}^{\sigma,\mu}(x) = \Gamma(x)\left[\sigma(x)u(x) + \mu(x)u^2(x)\right], \quad \text{for } x \in \Omega. \tag{4.4}$$

For a known diffusion coefficient D, the equation (4.1) can be represented as follows

$$\mathcal{L}(u, \sigma, \mu, g) = 0. \tag{4.5}$$

We will use an optimization based approach to reconstruct the coefficients $\sigma(x)$ and $\mu(x)$. We start by defining the following cost functional

$$\begin{aligned} J(\sigma, \mu, u_1, u_2) = \sum_{j=1}^{2} \frac{\alpha_j}{2} \|\mathcal{H}_j^{\sigma,\mu} - G_j^\delta\|^2 + \frac{\xi_1}{2}\|\sigma - \sigma_b\|^2_{H^1(\Omega)} + \frac{\xi_2}{2}\|\mu - \mu_b\|^2_{H^1(\Omega)} \\ + \gamma_1\|\sigma - \sigma_b\|_{L^1} + \gamma_2\|\mu - \mu_b\|_{L^1}, \end{aligned} \tag{4.6}$$

where u_1, u_2 satisfy (4.1) with boundary source functions g_1, g_2 respectively, σ_b, μ_b are known background absorption coefficients and $G_j^\delta,\ j = 1, 2$ are the (possibly noisy) measured initial acoustic wave pressure fields.

We now consider the following constrained minimization problem associated to the above cost functional

$$\min_{\sigma,\mu} J(\sigma, \mu, u_1, u_2), \tag{4.7}$$

$$\text{s.t. } \mathcal{L}(u_1, \sigma, \mu, g_1) = 0, \tag{4.8}$$

$$\mathcal{L}(u_2, \sigma, \mu, g_2) = 0. \tag{P}$$

The first term in the functional (4.6) represents a least-square data fitting term for obtaining σ, μ such that $\mathcal{H}_j^{\sigma,\mu} \approx G_j^\delta,\ j = 1, 2$. The regularization terms $\|\sigma - \sigma_b\|_{L^1}$ and $\|\mu - \mu_b\|_{L^1}$ in the above functional (4.6) implement L^1 regularization of the minimization problem that helps promote sparsity patterns in the reconstruction of absorption coefficients. The use of such L^1 regularization terms has been shown to obtain high contrast in the reconstructions [108, 86]. The H^1 regularization terms $\|\sigma - \sigma_b\|_{H^1}^2$ and $\|\mu - \mu_b\|_{H^1}^2$ help in denoising and removal of artifacts, thus, promoting high resolution.

4.3 Theory of the minimization problem

In this section, we analyze the existence of a solution to the minimization problem (4.7) and, further, characterize this solution through a first-order optimality system. We refer to this minimization problem as the 2P-PACT sparse reconstruction problem (2PPACT-SR). We begin our discussion with the analysis of the solution of (4.3). The existence of solution $u \in H_g^1(\Omega)$ for the boundary value problem (4.3) has been established in [45] under the assumptions that the coefficients D, σ, μ are bounded above and below by some positive constants and the boundary function g is the restriction of a continuous function $\varphi \in C^0(\bar{\Omega})$. The authors also showed the existence of a regular solution $u \in H_g^3(\Omega)$ under extra assumptions D, σ, μ are in $H^1(\Omega)$ and g comes from $\varphi \in C^3(\bar{\Omega})$. Further, the authors show that u is non-negative corresponding to a non-negative boundary function g is non-negative.

To prove the existence of minimizer of (4.6), we need $u \in H^2(\Omega)$. For this purpose, we impose weaker assumptions on the coefficients of (4.3) and boundary function g compared to the assumptions used in [45]. We present a new proof to the existence and uniqueness of solution $u \in H^2(\Omega)$ for the boundary value problem (4.3). We first recall the following well known fixed point theorem, for reference see [111, Theorem 4, Section 9.2].

Theorem 4.3.1 (Schaefer's Fixed Point Theorem)**.** *Suppose $A : X \longrightarrow X$ is a continuous and compact mapping. Assume further that the set*

$$\{u \in X : u = \lambda A[u] \textit{ for some } 0 \leq \lambda \leq 1\}$$

is bounded. Then A has a fixed point.

The following theorem gives the existence and uniqueness of solution $u \in H^2(\Omega)$ of (4.3).

Theorem 4.3.2. *Let Ω be a bounded domain in $\mathbb{R}^2$. Assume $D(x) \in W^{1,\infty}(\Omega)$, $(\sigma(x), \mu(x)) \in L^{\sigma}_{ad} \times L^{\mu}_{ad}$ and $g \in H^{3/2}(\partial\Omega)$ are given. Then the boundary value problem* (4.3) *has a unique solution u in $H^1_0(\Omega) \cap L^4(\Omega)$. Further, any weak solution u of* (4.3) *is also a strong solution, that is, $u \in H^2(\Omega)$.*

Proof. In order to solve above equation (4.3), we start by reducing it to a homogeneous boundary value problem by putting $u = v+\varphi$, where $\varphi \in H^2(\Omega)$ is a possible extension of g from boundary $\partial\Omega$ to whole Ω. Then, we can verify that the function v satisfies the equation:

$$-\nabla \cdot (D(x)\nabla v(x)) + \vartheta(x)v + \mu(x)v^2 = f(x), \qquad \text{in } \Omega, \tag{4.9}$$

$$v(x) = 0, \qquad \text{on } \partial\Omega. \tag{4.10}$$

where $\vartheta = \sigma + 2\mu\varphi$ and $f = \nabla \cdot (D(x)\nabla\varphi) - \sigma\varphi - \mu\varphi^2$.

For a given $v \in H^1_0(\Omega) \cap L^4(\Omega)$, define

$$F(x) := -\mu(x)v^2(x) + f(x).$$

Using conditions on φ, D, σ and μ together with $v \in L^4(\Omega)$, we see $F \in L^2(\Omega)$. Hence there exists a unique $w \in H^1_0(\Omega)$ (dependent on v) satisfying the following linear boundary value problem, see [112, Chapter 9] and [113, Chapter 3, Section 7]

$$-\nabla \cdot (D(x)\nabla w(x)) + \vartheta(x)w(x) = F(x), \qquad \text{in } \Omega,$$

$$w(x) = 0, \qquad \text{on } \partial\Omega$$

with the estimate

$$\|w\|_{H^2(\Omega)} \leq C\|F\|_{L^2(\Omega)}$$

for some constant C (dependent only on coefficient functions and the domain Ω).

This motivates us to define the the operator $A : H_0^1(\Omega) \cap L^4(\Omega) \to H_0^1(\Omega) \cap L^4(\Omega)$ given by $A[v] = w$, where w and v are related in the same manner as above. Further, we have

$$\|A[v]\|_{H^2(\Omega)} \leq C\|F\|_{L^2(\Omega)} \leq C\left(\|v\|_{L^4(\Omega)} + \|f\|_{L^2(\Omega)}\right). \tag{4.11}$$

Note that any fixed point of A will solve (4.3) which means to obtain a solution of (4.3) it is enough to verify the conditions of Theorem 4.3.1 for A, i.e., we need to show that the operator A is continuous, compact and the set $\{v \in H_0^1(\Omega) \cap L^4(\Omega) : v = \lambda A[v] \text{ for some } 0 \leq \lambda \leq 1\}$ is bounded.

To show continuity of A, let us start with a sequence

$$v_k \to v, \qquad \text{in} \qquad H_0^1(\Omega) \cap L^4(\Omega)$$

then by the inequality (4.11), we have

$$\sup_k \|w_k\|_{H^2(\Omega)} < \infty, \quad \text{where} \quad w_k = A[v_k], \text{ for } k = 1, \ldots$$

Thus there is a subsequence $\{w_{k_j}\}_{j=1}^{\infty}$ and a function $w \in H_0^1(\Omega) \cap L^4(\Omega)$ with

$$w_{k_j} \to w, \quad \text{in} \quad H_0^1(\Omega) \cap L^4(\Omega).$$

Now,

$$\int_\Omega \left(D(\nabla w_{k_j} \cdot \nabla\chi) + \vartheta w_{k_j}\chi\right) dx = -\int_\Omega \left(\mu v_{k_j}^2 \chi - f\chi\right) dx, \quad \forall \chi \in H_0^1(\Omega).$$

Taking the limit $k_j \to \infty$ we get

$$\int_\Omega \left(D(\nabla w \cdot \nabla \chi) + \vartheta w \chi\right) dx = -\int_\Omega \left(\mu v^2 \chi - f\chi\right) dx, \quad \forall \chi \in H_0^1(\Omega).$$

Hence $w = A[v]$. This shows the continuity of A. The compactness of A also follows by a similar argument, indeed if $\{v_k\}$ is a bounded sequence in $H_0^1(\Omega) \cap L^4(\Omega)$, the estimate (4.11) shows $\{A[v_k]\}_{k=1}^{\infty}$ is bounded in $H^2(\Omega)$ and hence possess a strongly convergent subsequence. The only thing remains to prove is the boundedness of the set:

$$Y = \left\{v \in H_0^1(\Omega) \cap L^4(\Omega) : v = \lambda A[v] \text{ for some } 0 \le \lambda \le 1\right\}.$$

Let $v \in H_0^1(\Omega) \cap L^4(\Omega)$ such that

$$v = \lambda A[v], \quad \text{for some } 0 \le \lambda \le 1.$$

Then $v/\lambda = A[v] \in H^2(\Omega) \cap H_0^1(\Omega) \cap L^4(\Omega)$ and

$$-\nabla \cdot (D(x)\nabla v(x)) + \vartheta(x)v(x) = -\lambda\mu v^2 + \lambda f, \quad \text{a.e. in } \Omega.$$

Multiplying the above relation with v and integrating over Ω to get

$$\begin{aligned}\int_\Omega D|\nabla v|^2 + \vartheta|v|^2 &= -\int_\Omega \lambda\mu v^3 dx + \int_\Omega \lambda f v dx \\ &\le \int_\Omega f v dx = \int_\Omega \left(\frac{1}{\epsilon} f\right)(\epsilon v)\, dx, \quad \text{for any } \epsilon > 0 \\ &\le \frac{\epsilon^2}{2}\int_\Omega v^2 dx + \frac{1}{2\epsilon^2}\int_\Omega f^2 dx.\end{aligned}$$

This gives

$$\int_\Omega D|\nabla v|^2 + \left(\vartheta - \frac{\epsilon^2}{2}\right)|v|^2 \le \frac{1}{2\epsilon^2}\int_\Omega f^2 dx.$$

Choose an $\epsilon > 0$ such that $\left(\vartheta - \frac{\epsilon^2}{2}\right)$ is bounded below by positive constant. Using this information together with the fact D is bounded below by a positive constant, we verified that the set Y is bounded. Hence by Schaefer's Theorem 4.3.1, we conclude that the operator A has a fixed point $v \in H^2(\Omega) \cap H^1_g(\Omega) \cap L^4(\Omega)$.

To show the uniqueness of the solution u, let u_1 and u_2 be two non-negative solutions of the boundary value problem (4.3). Then $w = u_1 - u_2$ satisfies the following boundary value problem

$$
\begin{aligned}
-\nabla \cdot (D(x)\nabla w(x)) + \sigma(x)w(x) + \mu(x)w(x)(u_1(x) + u_2(x)) = 0, &\qquad \text{in } \Omega, \\
w(x) = 0, &\qquad \text{on } \partial\Omega.
\end{aligned}
$$

Multiplying above equation by w and integrating by part , we get

$$\int_\Omega D(x)(\nabla w(x))^2 + \sigma(x)w^2(x) + \mu(x)w^2(x)(u_1(x) + u_2(x))dx = 0.$$

Since all coefficients are positive and solutions u_1, u_2 are non-negative therefore the above relation entails $w \equiv 0$. This proves the uniqueness of solution for boundary value problem (4.3). □

Remark 11. The result in Theorem 4.3.2 ensures that the initial acoustic wave pressure field $\mathcal{H}^{\sigma,\mu}$ given by (4.4) belongs to $L^2(\Omega) \cap L^4(\Omega)$. Thus, the functional J given by (4.6) is well-defined.

The solvability of the 2PPACT-SR inversion problem depends on the type of Dirichlet boundary data $g_j,\ j = 1, 2$. In this context, we have the following lemma from [45]

Lemma 4.3.3 (Boundary data)**.** *Let $g_i,\ i = 1, 2$ be two sets of boundary conditions with $g_i > 0$ and $g_1 - g_2 > 0$. Then $u_1 \neq u_2$ almost everywhere in Ω and one can uniquely reconstruct (σ, μ) from the two sets of initial acoustic wave pressure fields $\mathcal{H}^{\sigma,\mu}_i,\ i = 1, 2$.*

Next, we state the following lemma about the Fréchet differentiability of the mapping $u(\sigma, \mu)$ which will be needed later. For proof of this lemma, we refer to [45, Proposition 2.5].

Lemma 4.3.4. *The map $u(\sigma, \mu)$ defined by (4.1) is Fréchet differentiable with respect to σ and μ as a mapping from $L^{\sigma}_{ad} \times L^{\mu}_{ad}$ to $H^1_g(\Omega)$.*

Using Lemma 4.3.4, we introduce the reduced cost functional

$$\widehat{J}(\sigma, \mu) = J(\sigma, \mu, u_1(\sigma, \mu), u_2(\sigma, \mu)), \tag{4.12}$$

where $u_i(\sigma, \mu)$, $i = 1, 2$ denotes the unique solution of (4.5) given σ, μ and $g_i, i = 1, 2$. The constrained optimization problem (4.7) can be formulated as an unconstrained one as follows

$$\min_{(\sigma,\mu) \in L^{\sigma}_{ad} \times L^{\mu}_{ad}} \hat{J}(\sigma, \mu). \tag{4.13}$$

We next investigate the existence of a minimizer to the 2PPACT-SR problem (4.7).

Proposition 4.3.5. *Let $g_1, g_2 \in H^{1/2}(\Omega)$. Then there exists a quadruplet $(\sigma^*, \mu^*, u_1^*, u_2^*) \in L^{\sigma}_{ad} \times L^{\mu}_{ad} \times H^1_{g_1}(\Omega) \times H^1_{g_2}(\Omega)$ such that $u_i^*, i = 1, 2$ are solutions to $\mathcal{L}(\sigma, \mu, u_i, g_i) = 0, i = 1, 2$ and (σ^*, μ^*) minimizes $\hat{J}$ in $L^{\sigma}_{ad} \times L^{\mu}_{ad}$.*

Proof. We observe that $\hat{J}$ is bounded below. This implies there exists a minimizing sequence $(\sigma_m, \mu_m) \in L^{\sigma}_{ad} \times L^{\mu}_{ad}$. Since $\hat{J}$ is coercive in $L^{\sigma}_{ad} \times L^{\mu}_{ad}$, we have that the sequence (σ_m, μ_m) is bounded. Since $L^{\sigma}_{ad} \times L^{\mu}_{ad}$ is a closed subspace of a Hilbert space, it is reflexive. Thus, the sequence (σ_m, μ_m) has a weakly convergent subsequence $(\sigma_{m_l}, \mu_{m_l}) \rightharpoonup (\sigma^*, \mu^*)$. Consequently, the sequences $u_i(\sigma_{m_l}, \mu_{m_l}) \rightharpoonup u^*$ in $H^2(\Omega) \subset H^1_{g_i}(\Omega)$, $i = 1, 2$. Due to the fact that $H^2(\Omega)$ is compactly embedded in $H^1_{g_i}(\Omega)$, we have $u_i(\sigma_{m_l}, \mu_{m_l}) \to u^* \in H^1_{g_i}(\Omega)$. Again, since $H^2(\Omega)$ is compactly embedded in $L^4(\Omega)$, we additionally have $u_i(\sigma_{m_l}, \mu_{m_l}) \to u^* \in L^4(\Omega)$. We next aim at showing that $u^* = u(\sigma^*, \mu^*) \in H^1_{g_i}(\Omega)$. For this purpose, we consider the weak formulation of the solution of (4.1). The first term in the weak formulation we need to consider is $\langle \sigma_{m_l} u_i(\sigma_{m_l}, \mu_{m_l}), \psi \rangle_{L^2(\Omega)}$. By the preceding discussion, we have $\langle \sigma_{m_l} u_i(\sigma_{m_l}, \mu_{m_l}), \psi \rangle_{L^2(\Omega)} \to \langle \sigma^* u_i^*, \psi \rangle_{L^2(\Omega)}$. The second term we need to analyze is $\langle \mu_{m_l} u_i^2(\sigma_{m_l}, \mu_{m_l}), \psi \rangle_{L^2(\Omega)}$. Since, $\mu_{m_l} \rightharpoonup \mu^*$ in $L^2(\Omega)$ and $u_i(\sigma_{m_l}, \mu_{m_l}) \to u^* \in L^4(\Omega)$, we have $\langle \mu_{m_l} u_i^2(\sigma_{m_l}, \mu_{m_l}), \psi \rangle_{L^2(\Omega)} \to \langle \mu^* (u_i^*)^2, \psi \rangle_{L^2(\Omega)}$.

Thus, (σ^*, μ^*, u_i^*) solves (4.1) with boundary condition g_i and by continuity of the map $u(\sigma, \mu)$, we have $u^* = u(\sigma^*, \mu^*)$. Since $\hat{J}$ is sequentially weakly lower semi-continuous, we have that $(\sigma^*, \mu^*, u_1^*, u_2^*)$ minimizes $\hat{J}$ in $L^{\sigma}_{ad} \times L^{\mu}_{ad} \times H^1_{g_1}(\Omega) \times H^1_{g_2}(\Omega)$. □

4.3.1 Characterization of local minima

To characterize the solution of our optimization problem through first-order optimality conditions, we write the reduced functional $\hat{J}$ as follows

$$\hat{J} = \hat{J}_1 + \hat{J}_2, \;\; \hat{J}_i : L^{\sigma}_{ad} \times L^{\mu}_{ad} \to \mathbb{R}^+, \;\; i = 1, 2,$$

where

$$\begin{aligned} \hat{J}_1(\sigma, \mu) &= \sum_{j=1}^{2} \frac{\alpha_j}{2} \|\mathcal{H}_j^{\sigma,\mu} - G_j^{\delta}\|^2 + \frac{\xi_1}{2} \|\sigma - \sigma_b\|^2_{H^1(\Omega)} + \frac{\xi_2}{2} \|\mu - \mu_b\|^2_{H^1(\Omega)}, \\ \hat{J}_2(\sigma, \mu) &= \gamma_1 \|\sigma - \sigma_b\|_{L^1} + \gamma_2 \|\mu - \mu_b\|_{L^1}. \end{aligned} \tag{4.14}$$

Remark 12. The functional $\hat{J}_1$ is smooth and possibly non-convex, while $\hat{J}_2$ is non-smooth and convex.

The following property can be proved using arguments in [114].

Proposition 4.3.6. *The reduced functional $\hat{J}_1(\sigma, \mu)$ is weakly lower semi-continuous, bounded below and Fréchet differentiable with respect to σ, μ.*

Next, we are going to define the subdifferential of a non-smooth functional.

Definition 4.3.1 (Subdifferential)**.** If $\hat{J}$ is finite at a point (σ, μ), the Fréchet subdifferential of $\hat{J}$ at (σ, μ) is defined as follows [96]

$$\partial\hat{J}(\bar{\sigma}, \bar{\mu}) := \left\{ \phi \in (L^{\sigma}_{ad} \times L^{\mu}_{ad})^* : \liminf_{(\sigma,\mu)\to(\bar{\sigma},\bar{\mu})} \frac{\hat{J}(\sigma, \mu) - \hat{J}(\bar{\sigma}, \bar{\mu}) - \langle \phi, (\sigma, \mu) - (\bar{\sigma}, \bar{\mu})\rangle}{\|(\bar{\sigma}, \bar{\mu}) - (\sigma, \mu)\|_2} \geq 0 \right\}, \tag{4.15}$$

where $(L^{\sigma}_{ad} \times L^{\mu}_{ad})^*$ is the dual space of $L^{\sigma}_{ad} \times L^{\mu}_{ad}$. An element $\phi \in \partial\hat{J}(\sigma, \mu)$ is called a subdifferential of $\hat{J}$ at (σ, μ).

In our setting, we have the following

$$\partial\hat{J}(\sigma, \mu) = \nabla_{(\sigma,\mu)}\hat{J}_1(\sigma, \mu) + \partial\hat{J}_2(\sigma, \mu),$$

since $\hat{J}_1$ is Fréchet differentiable by Prop. 4.3.6. Moreover, for each $\alpha > 0$, it holds that

$$\partial(\alpha\hat{J}) = \alpha\partial\hat{J}.$$

The following proposition gives a necessary condition for a local minimum of $\hat{J}$ (see [86]).

Proposition 4.3.7 (Necessary condition)**.** *If $\hat{J} = \hat{J}_1 + \hat{J}_2$, with $\hat{J}_1, \hat{J}_2$ given by (4.14), attains a local minimum at $(\sigma^*, \mu^*) \in L^{\sigma}_{ad} \times L^{\mu}_{ad}$, then*

$$\boldsymbol{0} \in \partial\hat{J}(\sigma^*, \mu^*),$$

or equivalently

$$-\nabla_{(\sigma,\mu)}\hat{J}_1(\sigma^*, \mu^*) \in \partial\hat{J}_2(\sigma^*, \mu^*).$$

The following variational inequality holds for each $\lambda \in \partial\hat{J}_2(\sigma^*, \mu^*)$ (see [88]).

$$\langle \nabla\hat{J}_1(\sigma^*, \mu^*) + \lambda, (\sigma, \mu) - (\sigma^*, \mu^*)\rangle \geq 0, \qquad \forall(\sigma, \mu) \in L^{\sigma}_{ad} \times L^{\mu}_{ad}. \tag{4.16}$$

Using the definition of $\hat{J}_2$ in (4.14) and the fact that $L^{\sigma}_{ad} \times L^{\mu}_{ad}$ is reflexive, the inclusion $\lambda \in \partial\hat{J}_2(\sigma^*, \mu^*)$ gives the following characterization of space of λ

$$\lambda = (\lambda_1, \lambda_2),\ \lambda_i \in \Lambda^i_{ad} := \{\kappa \in L^2(\Omega) : 0 \leq \kappa \leq \gamma_i,\ \text{a.e. in } \Omega\},\ i = 1, 2.$$

A pointwise analysis of the variational inequality (4.16) leads to the existence of a non-negative functions $\lambda^*_{i,a}, \lambda^*_{i,b} \in L^2(\Omega),\ i = 1, 2$ that correspond to Lagrange multipliers for the inequality constraints in $L^\sigma_{ad} \times L^\mu_{ad}$. We, thus, have the following first-order optimality system.

Proposition 4.3.8 (First-order necessary conditions)**.** *The optimal solution of the minimization problem (4.13) can be characterized by the existence of* $(\lambda^*_1, \lambda^*_2, \lambda^*_{1,a}, \lambda^*_{2,a}, \lambda^*_{1,b}, \lambda^*_{2,b}) \in (\Lambda_{ad})^2 \times (L^2(\Omega))^4$ *such that*

$$\nabla_\sigma \hat{J}_1(\sigma^*, \mu^*) + \lambda^*_1 + \lambda^*_{1,b} - \lambda^*_{1,a} = 0, \tag{4.17}$$

$$\nabla_\mu \hat{J}_1(\sigma^*, \mu^*) + \lambda^*_2 + \lambda^*_{2,b} - \lambda^*_{2,a} = 0, \tag{4.18}$$

$$\lambda^*_{1,b} \geq 0,\ b - \sigma^* \geq 0,\ \langle \lambda^*_{1,b}, b - \sigma^* \rangle = 0, \tag{4.19}$$

$$\lambda^*_{1,a} \geq 0,\ \sigma^* - a \geq 0,\ \langle \lambda_{1,a}, \sigma^* - a \rangle = 0, \tag{4.20}$$

$$\lambda^*_{2,b} \geq 0,\ b - \mu^* \geq 0,\ \langle \lambda^*_{2,b}, b - \mu^* \rangle = 0, \tag{4.21}$$

$$\lambda^*_{2,a} \geq 0,\ \mu^* - a \geq 0,\ \langle \lambda_{2,a}, \mu^* - a \rangle = 0, \tag{4.22}$$

$$\lambda^*_1 = \gamma_1 \ \textit{a.e. on}\ \{x \in \Omega : \sigma^*(x) > 0\}, \tag{4.23}$$

$$\lambda^*_2 = \gamma_2 \ \textit{a.e. on}\ \{x \in \Omega : \mu^*(x) > 0\}, \tag{4.24}$$

$$0 \leq \lambda^*_1 \leq \gamma_1 \ \textit{a.e. on}\ \{x \in \Omega : \sigma^*(x) = 0\}, \tag{4.25}$$

$$0 \leq \lambda^*_2 \leq \gamma_2 \ \textit{a.e. on}\ \{x \in \Omega : \mu^*(x) = 0\}. \tag{4.26}$$

The conditions (4.19)-(4.26) are known as the complementarity conditions for $(\sigma^*, \mu^*, \lambda^*_1, \lambda^*_2)$.

To determine the gradient $\nabla_\sigma \hat{J}_1, \nabla_\mu \hat{J}_1$, we use the adjoint approach (see for e.g., [97, 98]). This gives the following reduced gradients of $\hat{J}_1$

$$\begin{aligned}
\nabla_\sigma \hat{J}_1(\sigma^*, \mu^*) =& \alpha_1(\mathcal{H}_1^{\sigma^*,\mu^*} - G_1^\delta)\Gamma u_1 + \alpha_2(\mathcal{H}_2^{\sigma^*,\mu^*} - G_2^\delta)\Gamma u_2 + u_1 v_1 + u_2 v_2 + \xi_1 \sigma^* \\
\nabla_\mu \hat{J}_1(\sigma^*, \mu^*) =& \alpha_1(\mathcal{H}_1^{\sigma^*,\mu^*} - G_1^\delta)\Gamma u_1^2 + \alpha_2(\mathcal{H}_2^{\sigma^*,\mu^*} - G_2^\delta)\Gamma u_2^2 + u_1^2 v_1 + u_2^2 v_2 + \xi_2 \mu^*
\end{aligned} \tag{4.27}$$

where u_1, u_2 satisfy the forward equations $\mathcal{L}(u_1, \sigma^*, \mu^*, g_1) = 0,\ \mathcal{L}(u_2, \sigma^*, \mu^*, g_2) = 0$, respectively, and v_1, v_2 satisfy the adjoint equations

$$\begin{gathered} -\nabla \cdot (D\nabla v_1) + \sigma^* v_1 + 2\mu^* u_1 v_1 = -\alpha_1 \Gamma(\sigma^* u_1 + \mu^* u_1^2 - G_1^\delta) \cdot (\sigma^* + 2u_1) \text{ in } \Omega, \\ v_1 = 0, \quad \text{ on } \partial\Omega \end{gathered} \tag{4.28}$$

$$\begin{gathered} -\nabla \cdot (D\nabla v_2) + \sigma^* v_2 + 2\mu^* u_2 v_2 = -\alpha_2 \Gamma(\sigma^* u_2 + \mu^* u_2^2 - G_2^\delta) \cdot (\sigma^* + 2|u_2|) \text{ in } \Omega, \\ v_2 = 0, \quad \text{ on } \partial\Omega. \end{gathered} \tag{4.29}$$

The complementarity conditions (4.19)-(4.26) can be rewritten in a compact form as follows. Define

$$\begin{aligned} c_1^* &= \lambda_1^* + \lambda_{1,b}^* - \lambda_{1,a}^*, \\ c_2^* &= \lambda_2^* + \lambda_{2,b}^* - \lambda_{2,a}^*. \end{aligned} \tag{4.30}$$

Then the triplets $(\lambda_1^*, \lambda_{1,a}^*, \lambda_{1,b}^*), (\lambda_2^*, \lambda_{2,a}^*, \lambda_{2,b}^*)$ are obtained by solving the following equations

$$\begin{aligned} \lambda_i^* &= \min(\gamma_i, \max(0, c_i^*)), \\ \lambda_{i,a}^* &= -\min(0, c_i^* + \gamma_i), \\ \lambda_{i,b}^* &= \max(0, c_i^* - \gamma_i), \end{aligned} \tag{4.31}$$

for $i = 1, 2$ (see [88]). For each $k \in \mathbb{R}^+$, define the following quantity

$$\begin{aligned} E_1(\sigma^*, c_1^*) = \sigma^* &- \max\{0, \sigma^* + k(c_1^* - \gamma_1)\} + \max\{0, \sigma^* - b + k(c_1^* - \gamma_1)\} \\ &- \min\{0, \sigma^* + k(c_1^* + \gamma_1)\} + \min\{0, \sigma^* - a + k(c_1^* + \gamma_1)\}. \\ E_2(\mu^*, c_2^*) = \mu^* &- \max\{0, \mu^* + k(c_2^* - \gamma_2)\} + \max\{0, \mu^* - b + k(c_2^* - \gamma_2)\} \\ &- \min\{0, \mu^* + k(c_2^* + \gamma_2)\} + \min\{0, \mu^* - a + k(c_2^* + \gamma_2)\}. \end{aligned}$$

The following lemma determines the complementarity conditions (4.19)-(4.26) in terms of E_1, E_2 (see [88, Lemma 2.2]).

Lemma 4.3.9. *The complementarity conditions (4.19)-(4.26) are equivalent to the following*

$$E_1(\sigma^*, c_1^*) = 0 = E_2(\mu^*, c_2^*), \tag{4.32}$$

where $c_i,\ i = 1, 2$ *are defined in (4.30).*

Using the gradients in (4.27) and Lemma 4.3.9, the optimality conditions (4.28)-(4.26) for the 2PPACT-SR problem can be rewritten as follows

Proposition 4.3.10. *A local minimizer* $(u_1, u_2, \sigma^*, \mu^*)$ *of the problem (4.7) can be characterized by the existence of* $(v_1, v_2, c_1^*, c_2^*) \in H_0^1(\Omega) \times H_0^1(\Omega) \times L_{ad}^{\sigma} \times L_{ad}^{\mu}$*, such that the following system is satisfied*

$$\begin{aligned}
-\nabla \cdot (D\nabla u_1) + \sigma^* u_1 + \mu^* u_1^2 = 0, &\qquad \text{in } \Omega,\\
u_1 = g_1, &\quad\ \ \text{on } \partial\Omega,\\
-\nabla \cdot (D\nabla v_1) + \sigma^* v_1 + 2\mu^* u_1 v_1 = -\alpha_1 \Gamma(\sigma^* u_1 + \mu^* u_1^2 - G_1^\delta) \cdot (\sigma^* + 2u_1) &\ \text{in } \Omega,\\
v_1 = 0, &\quad\ \ \text{on } \partial\Omega,\\
-\nabla \cdot (D\nabla u_2) + \sigma^* u_2 + \mu^* u_2^2 = 0, &\qquad \text{in } \Omega,\\
u_2 = g_2, &\quad\ \ \text{on } \partial\Omega,\\
-\nabla \cdot (D\nabla v_2) + \sigma^* v_2 + 2\mu^* u_2 v_2 = -\alpha_2 \Gamma(\sigma^* u_2 + \mu^* u_2^2 - G_2^\delta) \cdot (\sigma^* + 2u_2) &\ \text{in } \Omega,\\
v_2 = 0, &\quad\ \ \text{on } \partial\Omega,\\
\alpha_1(\mathcal{H}_1^{\sigma^*,\mu^*} - G_1^\delta)\Gamma u_1 + \alpha_2(\mathcal{H}_2^{\sigma^*,\mu^*} - G_2^\delta)\Gamma u_2 + u_1 v_1 + u_2 v_2 + \xi_1 \sigma^* &= 0,\\
\alpha_1(\mathcal{H}_1^{\sigma^*,\mu^*} - G_1^\delta)\Gamma u_1^2 + \alpha_2(\mathcal{H}_2^{\sigma^*,\mu^*} - G_2^\delta)\Gamma u_2^2 + u_1^2 v_1 + u_2^2 v_2 + \xi_2 \mu^* &= 0,\\
E_1(\sigma^*, c_1^*) &= 0,\\
E_2(\mu^*, c_2^*) &= 0.
\end{aligned} \tag{4.33}$$

4.4 Numerical schemes for solving the 2PPACT-SR inverse problem

4.4.1 Picard type method to solve the forward problem

In this section we propose a Picard type iterative scheme to solve the semi-linear boundary value problem (4.3). The algorithm is given as follows.

We now show the convergence of the Picard algorithm **??** to the solution of (4.1).

```
Picard-type algorithm
  1. Input: Initial guess $u_0$, $D$, $\sigma$, $\mu$, $g$, $N$ and $TOL$
     Initialize: $err_0 = 1$, $k = 0$
  2. While $err_k > TOL$ and $k < N$ do
  3. Solve the following linear elliptic boundary value problem
        $-\nabla \cdot (D(x)\nabla u_{k+1}(x)) + \sigma(x)u_{k+1}(x) + \mu(x)u_k(x)u_{k+1}(x) = 0, \quad \text{in } \Omega,$
        $u_{k+1}(x) = g(x), \quad \text{on } \partial\Omega$
     to get $u_{k+1}$ for $k \geq 0$
  4. $err_{k+1} = \|u_{k+1} - u_k\|_2$
  5. $k = k + 1$
  6. end
```

Theorem 4.4.1. *Let D, σ, μ be non-negative functions in $L^\infty(\Omega)$ and g be non-negative function in $C^0(\partial\Omega)$. Then the iterative sequence $\{u_k\}$, we obtained from the above Picard's method, converges in $H^1(\Omega)$ and the limit u is a solution of the following semi-linear elliptic boundary value problem*

$$-\nabla \cdot (D(x)\nabla u(x)) + \sigma(x)u(x) + \mu(x)u^2(x) = 0, \qquad \textit{in } \Omega,$$

$$u(x) = g(x), \qquad \textit{on } \partial\Omega.$$

Proof. By completeness of $H^1(\Omega)$ to show the convergence of sequence $\{u_k\}$ in $H^1(\Omega)$, we only need to show that the sequence $\{u_k\}$ is a Cauchy sequence in $H^1(\Omega)$. To achieve this goal, we will show the following contraction type relation for any $k \geq 1$

$$\|u_{k+1} - u_k\|_{H^1(\Omega)} \leq \gamma \|u_k - u_{k-1}\|_{H^1(\Omega)} \leq \cdots \leq \gamma^k \|u_1 - u_0\|_{H^1(\Omega)}, \qquad \text{for some } \gamma < 1.$$

We start with u_2 and u_1, recall from above Picard's type algorithm **??** that the iterates u_1 and u_2 satisfy the following two BVP's respectively

$$\begin{aligned} -\nabla \cdot (D(x)\nabla u_1(x)) + \sigma(x)u_1(x) + \mu(x)u_0(x)u_1(x) &= 0, && \text{in } \Omega, \\ u_1(x) &= g(x), && \text{on } \partial\Omega. \end{aligned} \tag{4.34}$$

$$\begin{aligned} -\nabla \cdot (D(x)\nabla u_1(x)) + \sigma(x)u_2(x) + \mu(x)u_1(x)u_2(x) &= 0, && \text{in } \Omega, \\ u_2(x) &= g(x), && \text{on } \partial\Omega. \end{aligned} \tag{4.35}$$

Then by direct substitution, we see that the difference $\bar{u} = u_2 - u_1$ solves

$$\begin{aligned} -\nabla \cdot (D(x)\nabla \bar{u}(x)) + \sigma(x)\bar{u}(x) + \mu(x)\bar{u} = \mu u_2(u_0 - u_1), && \text{in } \Omega, \\ \bar{u}(x) = 0, && \text{on } \partial\Omega. \end{aligned}$$

With the help of regularity estimates for elliptic boundary value problem, we get

$$\|\bar{u}\|_{H^1(\Omega)} \leq \|\bar{u}\|_{H^2(\Omega)} \leq C\|\mu u_2(u_0 - u_1)\|_{L^2(\Omega)}.$$

Consider the right hand side of the above inequality

$$\begin{aligned} \|\mu u_2(u_0 - u_1)\|_{L^2(\Omega)} &= \left(\int_\Omega |\mu|^2 |u_2|^2 |u_0 - u_1|^2 dx\right)^{\frac{1}{2}} \\ &\leq \underbrace{\|\mu\|_{L^\infty}\|g\|_{L^\infty}}_{\widetilde{C}} \|(u_0 - u_1)\|_{L^2(\Omega)}. \end{aligned}$$

Using this inequality, we have

$$\|u_2 - u_1\|_{H^1(\Omega)} \leq \|u_2 - u_1\|_{H^2(\Omega)} \leq \underbrace{C\widetilde{C}}_{\gamma} \|(u_0 - u_1)\|_{L^2(\Omega)} \leq \gamma \|(u_0 - u_1)\|_{H^1(\Omega)}.$$

By exactly same argument, we get

$$\|u_{k+1} - u_k\|_{H^1(\Omega)} \leq \gamma \|u_k - u_{k-1}\|_{H^1(\Omega)}.$$

Thus, we have the required relation

$$\|u_{k+1} - u_k\|_{H^1(\Omega)} \leq \gamma^k \|(u_0 - u_1)\|_{H^1(\Omega)}.$$

We can make $\gamma < 1$ by choosing appropriate g and μ. Hence, the sequence $\{u_k\}$ is a Cauchy sequence in $H^1(\Omega)$ and hence converges to a limit u in $H^1(\Omega)$. To complete the proof of our theorem, the only thing remain to show is that u solve

$$\begin{aligned} -\nabla \cdot (D(x)\nabla u(x)) + \sigma(x)u(x) + \mu(x)u^2(x) = 0, \qquad & \text{in } \Omega, \\ u(x) = g(x), \quad & \text{on } \partial\Omega. \end{aligned}$$

We know each u_k satisfies

$$\int_\Omega D\nabla u_k \cdot \nabla \varphi dx + \int_\Omega \sigma u_k \varphi dx + \int_\Omega \mu u_{k-1} u_k \varphi dx = 0, \quad \text{for all} \quad \varphi \in C_0^\infty(\Omega).$$

The convergence of $u_k \longrightarrow u$ in $H^1(\Omega)$ implies the convergence $\nabla u_k \longrightarrow \nabla u$ in $L^2(\Omega)$ and the convergence $\sigma u_k \longrightarrow \sigma u$ in $H^1(\Omega)$ as $k \to \infty$. Additionally, the strong convergence of $\{u_k\}$ in $H^1(\Omega)$ will guarantee the weak convergence of $u_{k-1}u_k \rightharpoonup u^2$ in $L^2(\Omega)$. Thus we have

$$\int_\Omega D\nabla u_k \cdot \nabla \varphi dx + \int_\Omega \sigma u_k \varphi dx + \int_\Omega \mu u_{k-1} u_k \varphi dx \longrightarrow \int_\Omega D\nabla u \cdot \nabla \varphi dx + \int_\Omega \sigma u \varphi dx + \int_\Omega \mu u^2 \varphi dx$$

for all $\varphi \in C_0^\infty(\Omega)$. Therefore

$$\int_\Omega D\nabla u \cdot \nabla \varphi dx + \int_\Omega \sigma u \varphi dx + \int_\Omega \mu u^2 \varphi dx = 0, \quad \text{for all} \quad \varphi \in C_0^\infty(\Omega).$$

This completes the proof of the theorem. □

4.4.2 Variable inertial proximal method for solving the optimality system

For solving the optimality system (4.33), we use a class of iterative schemes known as the proximal method. The fundamental idea behind a proximal scheme is to minimize an upper bound of the objective function $\hat{J}_1$, and, thus, $\hat{J}$, instead of directly minimizing the functional. Assuming that the gradients of $\hat{J}$, given in (4.27) are Lipschitz continuous, the formulation of the proximal method depends on the proximal operators $\mathbb{S}_{\gamma_1}^{L^\sigma_{ad}}(\widetilde{\sigma})$ and $\mathbb{S}_{\gamma_2}^{L^\mu_{ad}}(\widetilde{\mu})$, which are defined and characterized as follows

$$\underset{\sigma\in L^\sigma_{ad}}{\arg\min}\left\{\gamma_1\|\sigma\|_{L^1}+\frac{1}{2}\|\sigma-\widetilde{\sigma}\|^2\right\}=\mathbb{S}_{\gamma_1}^{L^\sigma_{ad}}(\widetilde{\sigma})\quad\text{for any }\ \widetilde{\sigma}\in L^2(\Omega),$$
$$\underset{\mu\in L^\mu_{ad}}{\arg\min}\left\{\gamma_2\|\mu\|_{L^1}+\frac{1}{2}\|\mu-\widetilde{\mu}\|^2\right\}=\mathbb{S}_{\gamma_2}^{L^\mu_{ad}}(\widetilde{\mu})\quad\text{for any }\ \widetilde{\mu}\in L^2(\Omega),$$

where the left-hand sides represents the proximal functions corresponding to σ and μ, respectively and the associated projected soft thresholding functions on the right-hand side are defined as follows

$$\mathbb{S}_{\gamma_1}^{L^\sigma_{ad}}(\widetilde{\sigma}):=\begin{cases}\min\{\widetilde{\sigma}-\gamma_1,b_\sigma\} & \text{on }\{(x,y)\in\Omega:\ \widetilde{\sigma}(x,y)>\gamma_1\}\\ 0 & \text{on }\{(x,y)\in\Omega:\ |\widetilde{\sigma}(x,y)|\leq\gamma_1\}\\ \max\{\widetilde{\sigma}+\gamma_1,a_\sigma\} & \text{on }\{(x,y)\in\Omega:\ \widetilde{\sigma}(x,y)<-\gamma_1\}\end{cases}\cdot \tag{4.36}$$

$$\mathbb{S}_{\gamma_2}^{L^\mu_{ad}}(\widetilde{\mu}):=\begin{cases}\min\{\widetilde{\mu}-\gamma_2,b_\mu\} & \text{on }\{(x,y)\in\Omega:\ \widetilde{\mu}(x,y)>\gamma_2\}\\ 0 & \text{on }\{(x,y)\in\Omega:\ |\widetilde{\mu}(x,y)|\leq\gamma_2\}\\ \max\{\widetilde{\mu}+\gamma_2,a_\mu\} & \text{on }\{(x,y)\in\Omega:\ \widetilde{\mu}(x,y)<-\gamma_2\}\end{cases}\cdot \tag{4.37}$$

Using (4.36)-(4.37), we get the following iterative schemes

$$\sigma_{k+1}\leftarrow\mathbb{S}_{\frac{\gamma_1}{L_\sigma}}^{L^\sigma_{ad}}\left(\sigma_k-\frac{1}{L}\nabla_\sigma\hat{J}_1(\sigma_k)\right),$$

$$\mu_{k+1} \leftarrow \mathbb{S}^{L^{\mu}_{ad}}_{\frac{\gamma_2}{L_\mu}} \left(\mu_k - \frac{1}{L} \nabla_\mu \hat{J}_1(\mu_k) \right),$$

starting from a given σ_0, where L is the Lischitz constants for the gradient function $\nabla \hat{J}_1$. The above schemes form the iterative shrinkage-thresholding algorithm (ISTA) method. to accelerate the ISTA scheme described above, one can consider a sequence $\{t_k, v_k\}$ such that

$$t_0 = 1, \qquad t_k := 1 + \sqrt{1 + 4t_{k-1}^2}/2, \tag{4.38}$$

and

$$v_0^\sigma := \sigma_0, \qquad v_k^\sigma := \sigma_k + \frac{(t_{k-1} - 1)}{t_k}(\sigma_k - \sigma_{k-1}), \tag{4.39}$$

$$v_0^\mu := \mu_0, \qquad v_k^\mu := \mu_k + \frac{(t_{k-1} - 1)}{t_k}(\mu_k - \mu_{k-1}). \tag{4.40}$$

This gives us the following update for the optimization variable σ_k

$$\begin{aligned} \sigma_{k+1} &\leftarrow \mathbb{S}^{L^{\sigma}_{ad}}_{\frac{\gamma_1}{L_\sigma}} \left(\sigma_k - \frac{1}{L} \nabla_\sigma \hat{J}_1(v_k^\sigma) \right), \\ \mu_{k+1} &\leftarrow \mathbb{S}^{L^{\mu}_{ad}}_{\frac{\gamma_2}{L_\mu}} \left(\mu_k - \frac{1}{L} \nabla_\mu \hat{J}_1(v_k^\mu) \right). \end{aligned} \tag{4.41}$$

We note that the arguments of the proximal operators in (4.41) represents a gradient step with fixed step size $s = 1/L$. Replacing v_k^σ, v_k^μ in (4.41) with (4.39), and assuming that $\nabla_\sigma \hat{J}_1(\sigma_k) \approx \nabla_\sigma \hat{J}_1(v_k^\sigma)$ and $\nabla_\mu \hat{J}_1(\mu_k) \approx \nabla_\mu \hat{J}_1(v_k^\mu)$, we obtain the following iterative scheme

$$\begin{aligned} \sigma_{k+1} &\leftarrow \mathbb{S}^{L^{\sigma}_{ad}}_{\gamma_1 s} \Big(\sigma_k - s\nabla_\sigma \hat{J}_1(\sigma_k) + \theta_k \left(\sigma_k - \sigma_{k-1}\right) \Big), \\ \mu_{k+1} &\leftarrow \mathbb{S}^{L^{\mu}_{ad}}_{\gamma_2 s} \Big(\mu_k - s\nabla_\mu \hat{J}_1(\mu_k) + \theta_k \left(\mu_k - \mu_{k-1}\right) \Big), \end{aligned} \tag{4.42}$$

where $\mu_{-1} = \mu_0$. The schemes given in (4.42) are known as the fast iterative shrinkage-thresholding algorithm (FISTA) proximal gradient method.

The computations done above are valid for any $L \geq L(\hat{J}_1)$. However, because the quantity $s = 1/L$ represents the step size in a gradient update, we use a backtracking line search algorithm to determine the optimal step size in each iteration. We then compute an upper bound L_k that satisfies $L_k \geq L(\hat{J}_1)$ at each iteration step. This leads to the definition of the variable step size as $s_k = 1/L_k$. The variable step size causes the factor $\frac{(t_{k-1}-1)}{t_k}$ in (4.39) to be non-optimal and we replace it by the fixed inertial parameter θ and, thus, the resulting scheme is known as Variable Inertial Proximal Gradient Method (VIP) (see[108] for more detail). We summarize the VIP scheme in the algorithm below

4.5 Numerical results

We first demonstrate the convergence of the Picard scheme given in Algorithm **??** for solving (4.1). We use the method of manufactured solutions to construct an exact solution for (4.1) with a non-zero source term $f(x_1, x_2)$ on the right hand side. We set $D(x_1, x_2) = 1.0,\ \sigma(x_1, x_2) = \sin(x_1)\sin(x_2), \mu = 1$. Further, we choose $\Omega = (0, 1) \times (0, 1)$. The boundary condition is given as $g(x_1, x_2) = \sin(x_1)\sin(x_2)$ and the right-hand side $f(x_1, x_2) = 2\sin(x_1)\sin(x_2) + 2(\sin(x_1)\sin(x_2))^2$. With the preceding choices of the parameters, the exact solution is given as $u_{ex} = \sin(x_1)\sin(x_2)$. The solution error is evaluated based on the following discrete L^1 norm

$$\|u\|_1 = h^2 \sum_{i,j=0}^{N_x} |u_{i,j}|,$$

which we identify with L^1_h. The discrete L^1 error is defined as follows

$$Err = \|u - u_{ex}\|_1.$$

Table 4.1 shows the results of experiments that demonstrate the convergence of the Picard algorithm. We see that the resulting order of convergence is $\mathcal{O}(h)$.

N_x	Err	Order
25	1.70e-3	–
50	8.77e-4	0.96
100	4.39e-4	0.99
200	2.20e-4	1.00

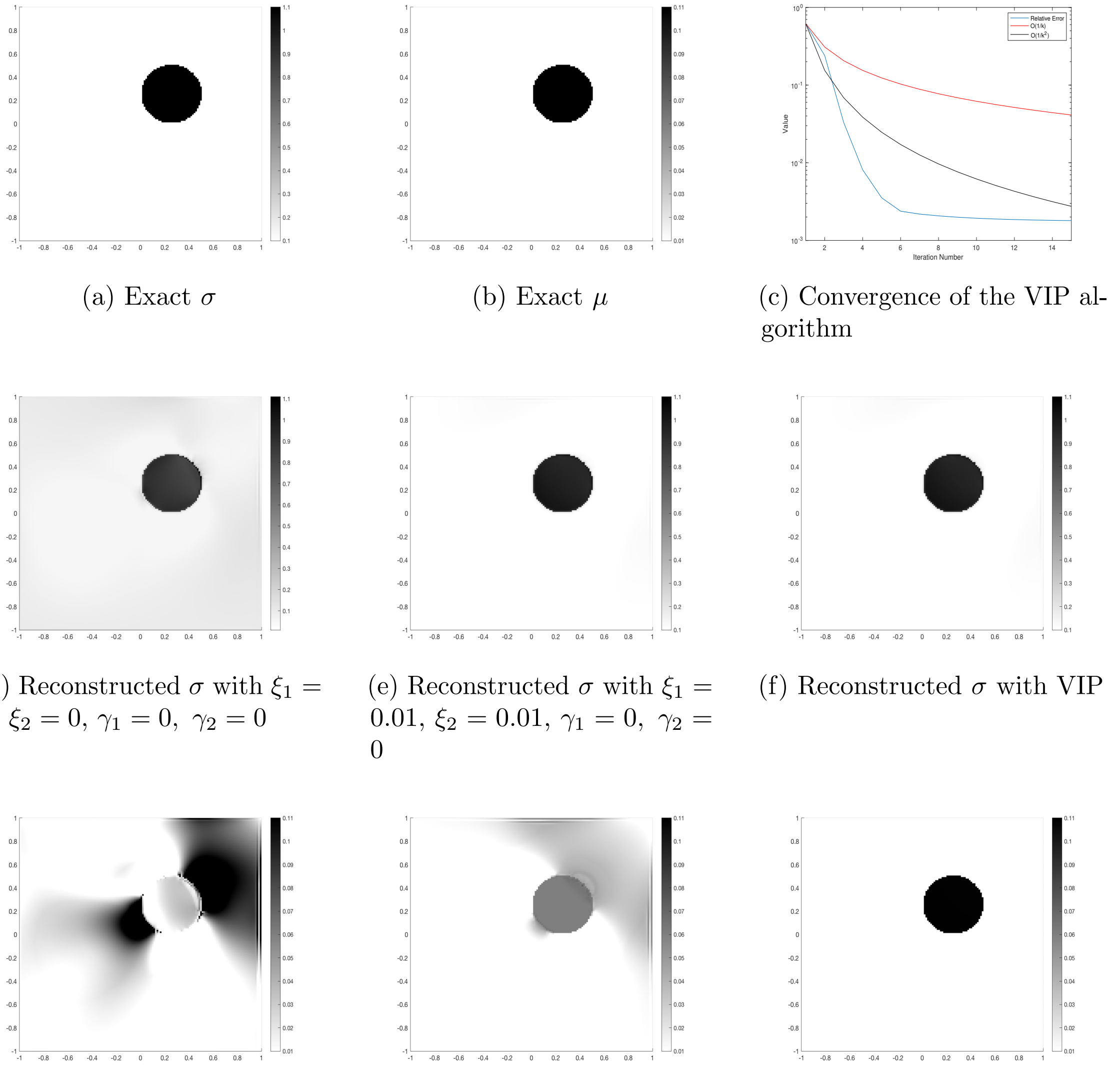
Table 4.1: Convergence of the Picard algorithm given in Algorithm **??**

We now present the results of numerical experiments obtained using the VIP scheme to solve the 2PPACT-SR reconstruction problem. We choose our domain in the experiments below as $\Omega = (-1, 1) \times (-1, 1)$. We discretize Ω into 150 equally spaced points in both x and y directions. The boundary illuminations for solving (4.1) to generate two sets of initial acoustic wave pressure field data are chosen as $g_1(x, y) = 1.0, \ g_2(x, y) = 2.0$. Such a choice of boundary conditions are consistent with Lemma 4.3.3 that ensure unique solvability of the 2PPACT-SR reconstruction problem. The range of the true values of σ and μ are chosen to lie in $(0.1, 1.1)$ and $(0.01, 0.11)$, respectively, except for the vascular phantom in test case 4. While these ranges are chosen for experimentation purposes, the ratio between the optical coefficients are based on the true experiments done in [49, 115]. Thus, upto a scaling factor, the chosen true values of the optical coefficients are similar to the experimentally observed values. The background values σ_b and μ_b are chosen to be 0.1 and 0.01 respectively, unless otherwise mentioned and D is chosen to be 0.1σ while generating the data with a known σ. For test case 4 with a brain vascular phantom, we choose the range of the values of $\sigma \in (0.1, 0.2)$, which is consistent with the values for σ in brain blood vessels [116], and use the aforementioned ratios to set the values of μ and D. The weights of the functional J given in (4.6) are chosen as $\alpha_1 = \alpha_2 = 1, \xi_1 = 0.01, \xi_2 = 0.01, \gamma_1 = 0.1, \gamma_2 = 0.1$. We remark that large values of

ξ_1, ξ_2 lead to smoothening of edges due to the presence of the H^1 regularization term whereas very small values lead to preservation of artifacts. Furthermore, large values of γ_1, γ_2 also lead to greater sparsity patterns in the reconstructions whereas very small values doesn't help in removing the artifacts. Thus, the choices of the weights are purely experimental and we have observed that with values of $\xi_1, \xi_2 \in (0.005, 0.1)$ and $\gamma_1, \gamma_2 \in (0.01, 0.5)$, we obtain similar reconstructions. The value of the Grüneisen coefficient is chosen to be 1.0. To generate the data $G_i^\delta,\ i = 1, 2$, we first solve for u_i in (4.1) with given test values of σ, μ and boundary illumination data g_i on a finer mesh with $N = 400$ using the Picard iterative scheme given in Algorithm **??**. We then compute G_i^δ on the finer mesh using the values of σ, μ, u_i from (4.4). Finally, we restrict G_i^δ onto the coarser mesh with $N = 150$ and use this as our given data.

In test case 1, we consider a phantom represented by a disk centered at $(0.25, 0.25)$ and having radius 0.25. The value of σ inside the disk is .11 and outside is 0.1. The corresponding value of μ inside the disk is 0.11 and outside is 0.01. With this test case, we demonstrate the need for each of the regularization term in the functional J. Further, we also validate the convergence of the VIP algorithm. The results are shown in Figure 4.1.

Figures 4.1a and 4.1b represent the exact phantoms for σ and μ, respectively. Figures 4.1d and 4.1g show the reconstructions without any regularization term in the functional J in (4.6). We observe the presence of strong artifacts in the reconstructions. With the H^1 regularization, the artifacts are severely reduced as seen in Figures 4.1e and 4.1h. One could choose a higher weight of the H^1 regularization term that would result in further removal of the artifacts. However, we observed that in addition to artifact suppression, it also resulted in smoothing of the edges and loss of contrast. The addition of the L^1 regularization term circumvents this issue as seen in Figures 4.1f and 4.1i. We also see the $\mathcal{O}(1/k^2)$ convergence of the relative error using the VIP algorithm in Figure 4.1c.



(a) Exact σ

(b) Exact μ

(c) Convergence of the VIP algorithm

(d) Reconstructed σ with $\xi_1 = 0,\ \xi_2 = 0,\ \gamma_1 = 0,\ \gamma_2 = 0$

(e) Reconstructed σ with $\xi_1 = 0.01,\ \xi_2 = 0.01,\ \gamma_1 = 0,\ \gamma_2 = 0$

(f) Reconstructed σ with VIP

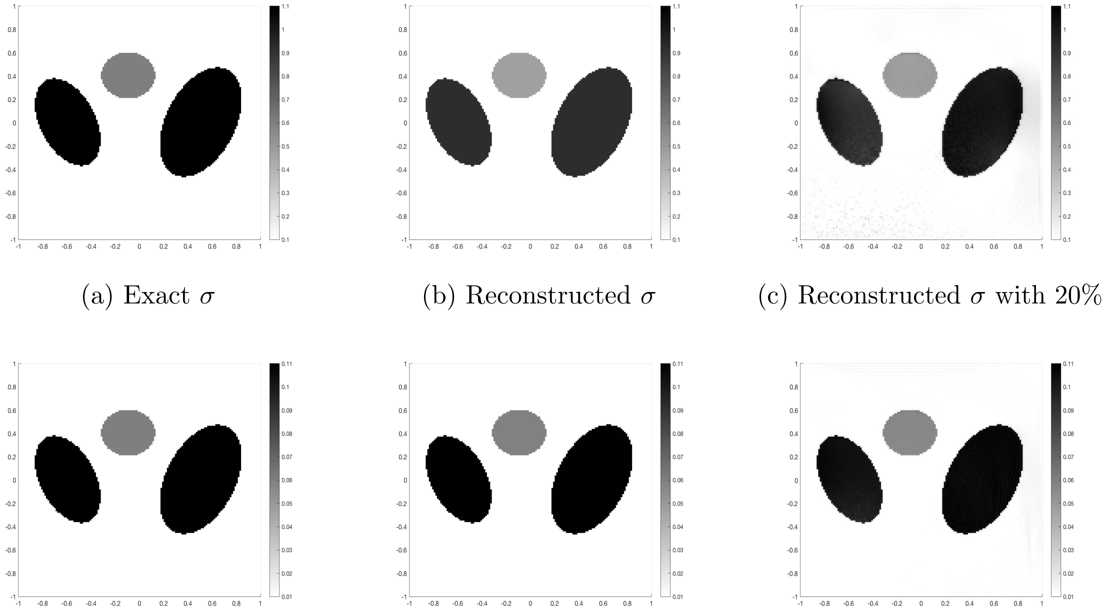
(g) Reconstructed μ with $\xi_1 = 0,\ \xi_2 = 0,\ \gamma_1 = 0,\ \gamma_2 = 0$

(h) Reconstructed μ with $\xi_1 = 0.01,\ \xi_2 = 0.01,\ \gamma_1 = 0,\ \gamma_2 = 0$

(i) Reconstructed μ with VIP

Figure 4.1: Test Case 1-Reconstructions of the disk phantom with the 2PPACT-SR framework

In test case 2, we consider a heart lung phantom for both σ and μ. For σ, the background value of the phantom is 0.1 that is perturbed into two ellipses that represent the lungs with value 1.1 and into a disk representing heart with value 0.5. The corresponding value of μ inside the ellipses is 0.11, in the disks is 0.05 and 0 elsewhere.The plots of the exact and the reconstructed phantoms are shown in Figure 4.2.



(a) Exact σ

(b) Reconstructed σ

(c) Reconstructed σ with 20%

(d) Exact μ

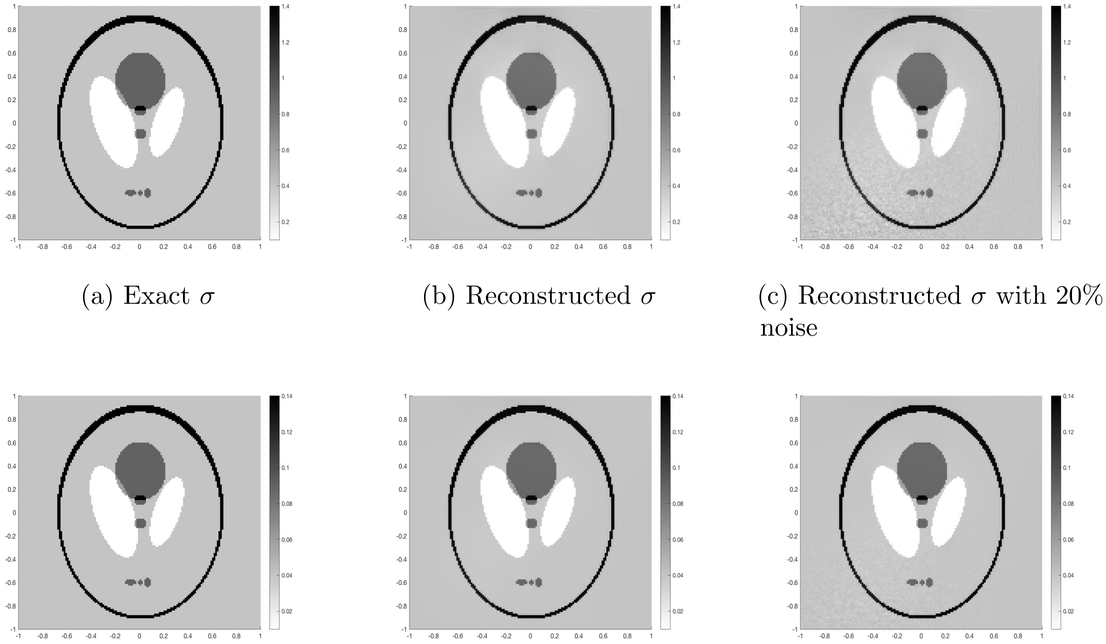
(e) Reconstructed μ

(f) Reconstructed μ with 20% noise

Figure 4.2: Test Case 2-Reconstructions of the heart and lung phantom with the 2PPACT-SR framework

We again see from Figures 4.2b and 4.2e that the reconstructions of σ, μ are of high contrast and high resolution. To test the robustness of our method, we added 20% multiplicative Gaussian noise to the interior data $\mathcal{H}^{\sigma,\mu}$ and use it for our 2PPACT-SR inversion algorithm. We also modify the value of the regularization parameters $\xi_1 = 0.1, \xi_2 = 0.1, \gamma_1 = 0.3, \gamma_2 = 0.3$, in order to counter the noisy data. The results can be seen in Figure 4.2c and 4.2f. We see that the reconstruction of σ contains a few artifacts but still is of good quality. The reconstruction of μ demonstrates very little artifacts. This shows that our 2PPACT-SR reconstruction framework is robust and accurate even in the presence of noisy data.

In test case 3, we consider σ as the Shepp-Logan phantom given in [117]. The background σ_b is chosen to be 0.4 in this case. We compute $\mu = 0.1\sigma$ and the background value of μ_b is chosen as 0.04. The plots of the exact and reconstructed phantoms are shown in Figure 4.3.



(a) Exact σ

(b) Reconstructed σ

(c) Reconstructed σ with 20% noise

(d) Exact μ

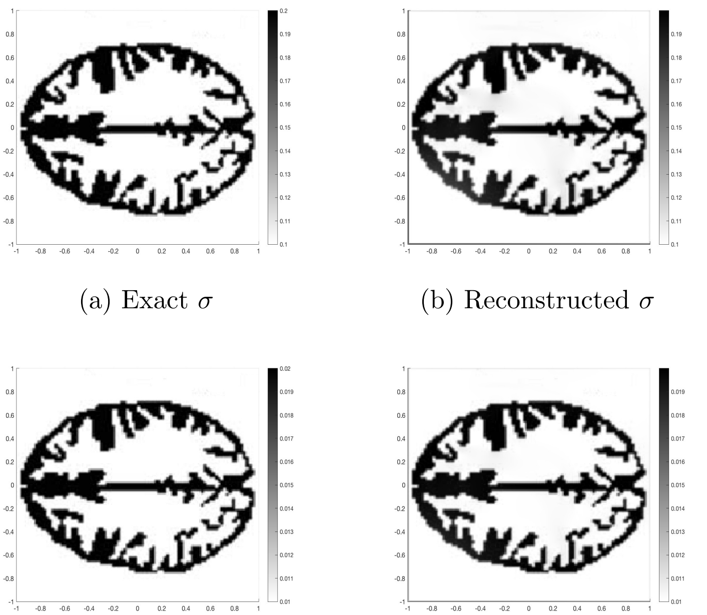
(e) Reconstructed μ

(f) Reconstructed μ with 20% noise

Figure 4.3: Test Case 3-Reconstructions of the Shepp-Logan phantom with the 2PPACT-SR framework

We again see from Figures 4.3b and 4.3e that the 2PPACT-SR reconstruction framework gives superior quality reconstructions even for objects with high contrast values and with holes and inclusions. The reconstructions with 20% noise in the interior data are shown in Figures 4.3c and 4.3f with the modified regularization parameter values as in the previous test case. We see that the reconstructions are still of high quality with very less artifacts.

In the final test case, we consider a brain vascular phantom taken from [118]. The value of σ lies between 0.1 and 0.2 and the value of μ lies between 0.01 and 0.02. The exact and the reconstructed phantoms are shown in Figure 4.4.



(a) Exact σ

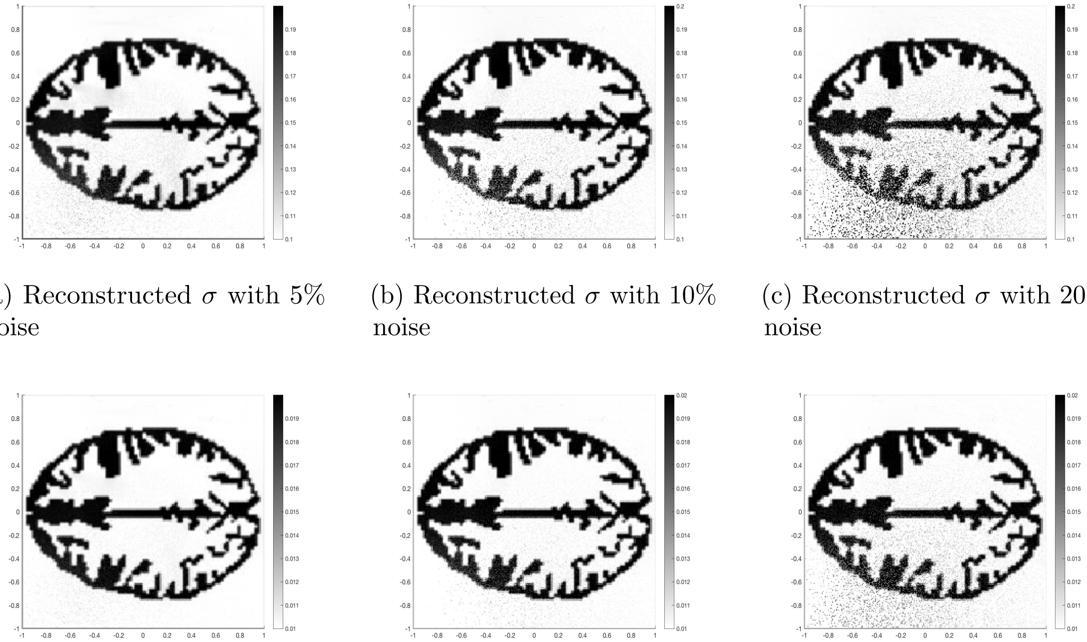
(b) Reconstructed σ

(c) Exact μ

(d) Reconstructed μ

Figure 4.4: Test Case 4-Reconstructions of brain vascular phantom with the 2PPACT-SR framework

Figures 4.4b and 4.4d demonstrates the capability of our 2PPACT-SR framework to provided superior reconstructions in vascular imaging. Our method is able to capture the fine detail of the structures without loss of resolution or contrast. We also tested the performance of our algorithm to provide reconstructions in presence of different levels of noise in the interior data. The results are shown in Figure 4.5.



(a) Reconstructed σ with 5% noise

(b) Reconstructed σ with 10% noise

(c) Reconstructed σ with 20% noise

(d) Reconstructed μ with 5% noise

(e) Reconstructed μ with 10% noise

(f) Reconstructed μ with 20% noise

Figure 4.5: Test Case 4-Reconstructions of brain vascular phantom with the 2PPACT-SR framework

Figures 4.5a-4.5c and 4.5d-4.5f show the reconstructions of the vascular phantom in presence of 5%, 10% and 20% white multiplicative Gaussian noise in the interior data. We observe that as the noise level increases, the artifacts, especially in the reconstruction of σ, also increases. However, our algorithm is able to provide high contrast and high resolution images for upto 10% noise. With 20% noisy data, our algorithm still provided high contrast and high resolution images for μ but the reconstruction of σ shows the presence of moderate speckled artifacts at the corner that reduces the resolution, but still has good contrast. This shows the robustness of our algorithm, even in the presence of large noise in the data.

Remark 13. For the proposed nonlinear reconstruction framework, the underlying assumption that has been used in the reconstruction procedure is the availability of interior acoustic wave pressure field measurements in the whole of the domain Ω. However, in practice, such measurements are limited and sparse due to the restriction of the number of detectors that can be used. Thus, it is of paramount importance to devise a computational algorithm for recovering the optical coefficients from sparse interior data. For this purpose, we remark that in the part of the domain where the sparse interior data can be obtained, multiple sets of such measurements are available. Using this fact, one possible approach to obtain the optical coefficients from sparse data would be to first solve a data completion problem using a Nash games framework (see for e.g., [119]), where missing data is filled in from the given multiple sets of sparse data and then, one can apply our proposed method to obtain superior reconstructions.

4.6 Conclusions

In this work, we have presented a new reconstruction framework in 2P-PACT for determining the optical coefficients from two-photon photoacoustic data. The framework comprises of a PDE-constrained optimization problem that promotes sparsity patterns in the reconstructions of the single and two photon absorption coefficients. We present a new theoretical analysis of the existence and uniqueness of a solution to a semi-linear elliptic PDE arising in 2P-PACT. Further, we present a proximal scheme using a Picard solver for the semi-linear PDE and its adjoint to solve the optimization problem. Several numerical results demonstrate that the proposed framework is able to achieve reconstructions with high contrast and high resolution for objects including holes and inclusions.

```
1. Input: $\beta$, $\hat{J}_1$, $\sigma_0 = \sigma_{-1}$, $\mu_0 = \mu_{-1}$, $TOL$, $n > 1$, $L_0 > 0$
   **Initialize:** $E_1^0 = E_2^0 = 1$, $k = 0$, choose $\theta \in (0,1)$ and $c_1 < 2$ and $c_2 > 0$;
2. While $\|E_1^{k-1}\| + \|E_2^{k-1}\| > TOL$ do
3. Compute $\nabla_\sigma \hat{J}_1(\sigma_k, \mu_k)$, $\nabla_\mu \hat{J}_1(\sigma_k, \mu_k)$
4. Backtracking: Find the smallest non-negative integer $i$ such that with
     $\tilde{L} = n^i L_{k-1}$
     $\hat{J}_1(\tilde{\sigma}, \tilde{\mu}) \leq \hat{J}_1(\sigma_k, \mu_k) + \Big\langle \nabla_\sigma \hat{J}_1(\sigma_k, \mu_k), \tilde{\sigma} - \sigma_k \Big\rangle + \Big\langle \nabla_\mu \hat{J}_1(\sigma_k, \mu_k), \tilde{\mu} - \mu_k \Big\rangle$
         $+ \frac{\tilde{L}}{2} \left( \|\tilde{\sigma} - \sigma_k\|^2 + \|\tilde{\mu} - \mu_k\|^2 \right)$
      where $\tilde{\sigma} = \mathbb{S}_{\gamma_1 s}^{L_{ad}^\sigma} \Big( \sigma_k - s\,(\nabla_\sigma \hat{J}_1)(\sigma_k, \mu_k) + \theta(\sigma_k - \sigma_{k-1}) \Big)$
            $\tilde{\mu} = \mathbb{S}_{\gamma_2 s}^{L_{ad}^\mu} \Big( \sigma_k - s\,(\nabla_\mu \hat{J}_1)(\sigma_k, \mu_k) + \theta(\mu_k - \mu_{k-1}) \Big)$,
            $s = c_1(1-\theta)/(\tilde{L} + 2c_2)$,
5. Set $L_k = \tilde{L}$ and $s_k = c_1(1-\theta)/(L_k + 2c_2)$
6. $\sigma_{k+1} = \mathbb{S}_{\gamma_1 s_k}^{L_{ad}\sigma} \Big( \sigma_k - s_k\,(\nabla_\sigma \hat{J}_1)(\sigma_k, \mu_k) + \theta(\sigma_k - \sigma_{k-1}) \Big)$
   $\mu_{k+1} = \mathbb{S}_{\gamma_2 s_k}^{L_{ad}^\mu} \Big( \mu_k - s_k\,(\nabla_\mu \hat{J}_1)(\sigma_k, \mu_k) + \theta(\mu_k - \mu_{k-1}) \Big)$
7. $c_1^k = -(\nabla_\sigma \hat{J}_1)(\sigma_k, \mu_k)$, $c_2^k = -(\nabla_\mu \hat{J}_1)(\sigma_k, \mu_k)$
8. $E_1^k = E(\sigma_k, c_1^k)$, $E_2^k = E(\mu_k, c_2^k)$
9. $k = k + 1$
10. end
```

CHAPTER 5

Second-order nonstandard methods for ordinary differential equations

5.1 Introduction

Nonstandard finite difference (NSFD) methods provide an efficient way to solve many problems numerically appearing in engineering and science and also known to provide several advantages over classical techniques. In recent years, some NSFD methods were constructed that are elementary stable. However, they are only first order accurate, in general. In this chapter, we propose a new class of non-standard finite difference methods, which are both elementary stable and of higher order accuracy.

This chapter is structured as follows: We first discuss a second order nonstandard finite difference Explicit-Euler scheme to approximate one-dimensional autonomous dynamical system. We further discuss the one-stage and two-stage modified nonstandard theta methods for $n-$ dimensional autonomous dynamical system. We see some applications to classical mathematical biology models in Section 5.3 to numerically validate the theoretical results. In the last section, some concluding remarks are made and future research directions are outlined and the content of this chapter has been taken from [120, 121, 122], [1] .

[1] Second-order modified nonstandard theta and Runge-Kutta methods for n-dimensional autonomous differential equations, Hristo V. Kojouharov, Souvik Roy, Madhu Gupta and Fawaz K. Alalhareth (Submitted).

5.2 Main Results

5.2.1 One-dimensional

A one-dimensional autonomous differential equation can be written as

$$\frac{dx}{dt} = f(x); \quad x(t_0) = x_0, \tag{5.1}$$

where $x : [t_0, T) \to \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ is differentiable and $x_0 \in \mathbb{R}$

In this section, we present a NSFD scheme for Equation (5.1) that is second order accurate and elementary stable.

We have assumed that Equation (5.1) has a finite number of equilibria and each of them is hyperbolic. For this purpose, we need the following result.

Lemma 5.2.1. *Let $f_x \in C^1(\mathbb{R})$ and $f_x(x^*) \neq 0$. Then there exists an $\epsilon(x^*) > 0$ such that*

$$\|f_x(x) - f_x(x^*)\|_{\mathbb{R}} < \frac{|f_x(x^*)|}{2}, \ \forall \ \|x - x^*\|_{\mathbb{R}} < \epsilon(x^*).$$

In the following theorem, we now describe the NSFD scheme and the conditions under which it is second order accurate and elementary stable.

Theorem 5.2.2. *Let $f \in C^1(\mathbb{R})$ and let $\varphi : \mathbb{R}^+ \times \mathbb{R} \to \mathbb{R}^+$ satisfying the following conditions:*

(i) $\varphi(h, x) = h + f_x(x)\dfrac{h^2}{2} + \mathcal{O}(h^3)$

(ii) $0 < \varphi(h, x) < \dfrac{2}{|f_x(x^*)|}$ *for all hyperbolic equilibria x^* of (5.1) with $h > 0$, and for all $x \in \mathbb{R}$ with $\|x - x^*\|_{\mathbb{R}} < \epsilon(x^*)$ and $\epsilon(x^*)$ as obtained from Lemma 5.2.1.*

Then the following nonstandard Explicit Euler's method

$$\frac{x_{n+1} - x_n}{\varphi(h, x_n)} = f(x_n), \tag{5.2}$$

for approximating the solution of problem (5.1), is accurate of second order and elementary stable, where x_n is the numerical approximation of the exact solution $x(t_n)$.

Proof. We first prove the second order accuracy of (5.2). For this purpose, we use a Taylor series expansion about t_n to obtain

$$\begin{aligned} & x(t_{n+1}) - [x(t_n) + \varphi(h, x(t_n))f(x(t_n))] \\ &= \left[x(t_n) + hx'(t_n) + \frac{h^2}{2}x''(t_n) + \mathcal{O}(h^3)\right] - \left[x(t_n) + \varphi(h, x(t_n))f(x(t_n))\right] \qquad (5.3) \\ &= hf(x(t_n)) + \frac{h^2}{2}f_x(x(t_n))f(x(t_n)) + \mathcal{O}(h^3) - \varphi(h, x(t_n))f(x(t_n)) \end{aligned}$$

For second order accuracy of (5.2), we need the right hand side of (5.9) to be of $\mathcal{O}(h^3)$. Substituting the expression of φ from condition (i) into equation (5.9), we obtain

$$x(t_{n+1}) - [x(t_n) + \varphi(h, x(t_n))f(x(t_n))] = \mathcal{O}(h^3),$$

which implies that the numerical method (5.2) is of second order.

To show that the scheme in (5.2) is elementary stable, we need to show that $|1 + f_x(x^*)\varphi(h, x_n)| < 1$, whenever x^* is a stable equilibrium point of (5.1) and $|1 + f_x(x^*)\varphi(h, x_n)| > 1$, whenever x^* is an unstable equilibrium point of (5.1) (see [**?**]).

First, let x^* be a stable equilibrium point. Then $f_x(x^*) < 0$. From condition (ii), we obtain,

$$0 < \varphi(h, x_n) < \frac{2}{|f_x(x^*)|}.$$

This implies

$$-2 < f_x(x^*)\varphi(h, x_n) < 0, \text{ since } f_x(x^*) < 0.$$

Thus, we obtain

$$-1 < 1 + f_x(x^*)\varphi(h, x_n) < 1$$

that gives us $|1 + f_x(x^*)\varphi(h, x_n)| < 1$.

If x^* is an unstable equilibrium point, then $f_x(x^*) > 0$. Thus, $|1+f_x(x^*)\varphi(h, x_n)| > 1$, since $\varphi(h, x_n) > 0$. □

We next explore a class of denominator functions $\varphi(h, x_n)$ that satisfy conditions (i) and (ii) of Theorem 5.2.3. In this context we have the following lemma. Let $\phi : \mathbb{R} \to \mathbb{R}$ satisfy the following conditions:

(i) $0 < \phi(h) < 1$ for all $h > 0$.

(ii) $\phi(h) = h - \dfrac{h^2}{2} + \mathcal{O}(h^3)$

Then, the denominator function $\varphi(h, x_n)$ defined as follows:

$$\varphi(h, x_n) = \begin{cases} \dfrac{\phi(hq)}{q}, & q = -f_x(x_n) \neq 0, \\ h, & f_x(x_n) = 0 \end{cases}$$

guarantees that the NSFD Euler's scheme (5.2) is second order accurate and elementary stable.

Proof. It is enough to show that the conditions (i) and (ii) of Theorem 5.2.3 are satisfied by φ. Since $\phi(h) = h - \dfrac{h^2}{2} + \mathcal{O}(h^3)$, we have

$$\varphi(h, x_n) = \frac{\phi(hq)}{q} = \frac{hq - \dfrac{h^2q^2}{2} + q\ \mathcal{O}(h^3)}{q} = h + f_x(x_n)\frac{h^2}{2} + \mathcal{O}(h^3)$$

This implies $\varphi(h, x_n)$ satisfies condition (i) of Theorem 5.2.3.

For verifying condition (ii) of Theorem 5.2.3, we first note that since x^* is a hyperbolic equilibrium, $f_x(x^*) \neq 0$, by Lemma 5.2.1, $\|x_n - x^*\|_{\mathbb{R}} < \epsilon(x^*)$ implies $f_x(x_n) \neq 0$. Thus, we assume $f_x(x_n) \neq 0$ for the forthcoming discussion. We first consider the case when x^* is a stable equilibrium. Then for $\|x_n - x^*\|_{\mathbb{R}} < \epsilon(x^*)$ with $\epsilon(x^*)$ as obtained from Lemma 5.2.1, we have

$$\|f_x(x_n) - f_x(x^*)\|_{\mathbb{R}} < -\frac{f_x(x^*)}{2}.$$

This implies

$$-\frac{f_x(x^*)}{2} < -f_x(x_n) = q. \tag{5.4}$$

Since $0 < \phi(hq) < 1$, we have

$$0 < \varphi(h, x_n) < \frac{1}{q}.$$

Using (5.4), we have

$$0 < \varphi(h, x_n) < -\frac{2}{f_x(x^*)} = \frac{2}{|f_x(x^*)|}, \tag{5.5}$$

since $f_x(x^*) < 0$. Next let x^* be an unstable equilibrium. Then

$$\|f_x(x_n) - f_x(x^*)\|_{\mathbb{R}} < \frac{f_x(x^*)}{2},$$

implies

$$\frac{f_x(x^*)}{2} < f_x(x_n) = -q.$$

This gives us

$$0 < \varphi(h, x_n) < -\frac{1}{q} < \frac{2}{f_x(x^*)} = \frac{2}{|f_x(x^*)|}. \tag{5.6}$$

Equations (5.4) and (5.6) imply that $\varphi(h, x_n)$ satisfy condition (ii) of Theorem 5.2.3, which proves that linear stabiity of each equilibrium x^* of Equation (5.1) is same as the linear stability of x^* as a fixed point of Method (5.2).

Also, definition of numerical method (5.2) assures that all of its fixed points are equilibria of Equation (5.1) and vice versa.

Thus, the NSFD scheme (5.2) with the denominator function $\varphi(h, x_n) = \dfrac{\phi(hq)}{q}$ for $q = -f_x(x_n) \neq 0$ is second order accurate and elementary stable. □

Remark 14. The function $\phi(h) = 1 - e^{-h}$ satisfies the conditions of Lemma 5.2.1 and ensures a second order accurate and elementary stable scheme (5.2).

5.2.2 General second-order modified nonstandard theta method

The new second-order modified nonstandard theta method is given in the following theorem:

Theorem 5.2.3. *Let $\vec{f} \in C^2(\mathbb{R}^n; \mathbb{R}^n)$ and let $\varphi_i : \mathbb{R}^+ \times \mathbb{R}^n \to \mathbb{R}^+$, for $i = 1, \dots, n$, satisfies the following conditions:*

(I) $\varphi_i(h, \vec{x}) = h + (1 - 2\theta)\dfrac{\langle \nabla_x f_i(\vec{x}), \vec{f}(\vec{x})\rangle}{f_i(\vec{x})}\dfrac{h^2}{2} + \mathcal{O}(h^3)$, *for all* $1 \leqslant i \leqslant n$,

(II) $0 < \varphi_i(h, \vec{x}) < \dfrac{2|\mathrm{Re}(\lambda_i)|}{|2\theta - 1||\lambda_i|^2}$, $0 \leqslant \theta \leqslant 1$, $\theta \neq \frac{1}{2}$, *for all hyperbolic equilibria* $\vec{x}^*$ *of Equation* (2.1) *with* $h > 0$ *and for all* $\vec{x} \in \mathbb{R}^n$.

Then the following modified nonstandard theta method:

$$\frac{x_i^{k+1} - x_i^k}{\varphi_i(h, \vec{x}^k)} = f_i\left(\theta \vec{x}^{k+1} + (1 - \theta)\vec{x}^k\right) \tag{5.7}$$

and the modified nonstandard two-stage theta method [76]:

$$\frac{x_i^{k+1}-x_i^k}{\varphi_i(h,\vec{x}^k)} = \theta f_i(\vec{x}^{k+1}) + (1-\theta) f_i(\vec{x}^k), \tag{5.8}$$

$i=1,\dots,n$, *for approximating the solution of Equation* (2.1), *are both accurate of second order and elementary stable.*

Proof. The second-order accuracy of the modified nonstandard one-stage theta method (5.7) is proven using the Taylor series expansion about t_k which yields:

$$\begin{aligned}
&x_i(t_{k+1}) - \left[x_i(t_k) + \varphi_i(h,\vec{x}(t_k)) f_i\left(\theta \vec{x}(t_{k+1}) + (1-\theta)\vec{x}(t_k)\right)\right] \\
&= \left[x_i(t_k) + h x_i'(t_k) + \frac{h^2}{2} x''(t_k) + \mathcal{O}(h^3)\right] \\
&\quad - \left[x_i(t_k) + \varphi_i(h,\vec{x}(t_k)) f\left(\theta\vec{x}(t_k) + \theta h \vec{x}'(t_k) + \theta\frac{h^2}{2}\vec{x}''(t_k) + \mathcal{O}(h^3) + \vec{x}(t_k) - \theta\vec{x}(t_k)\right)\right] \\
&= h x_i'(t_k) + \frac{h^2}{2} x_i''(t_k) - \varphi_i(h,\vec{x}(t_k)) f\left(\vec{x}(t_k) + \theta h\vec{x}'(t_k) + \theta\frac{h^2}{2}\vec{x}''(t_k) + \mathcal{O}(h^3)\right) + \mathcal{O}(h^3).
\end{aligned} \tag{5.9}$$

Introducing the notation $\vec{H} = \theta h \vec{x}'(t_k) + \theta \frac{h^2}{2}\vec{x}''(t_k) + \mathcal{O}(h^3)$ gives

$$\begin{aligned}
&x_i(t_{k+1}) - \left[x_i(t_k) + \varphi_i(h,\vec{x}(t_k)) f\left(\theta \vec{x}(t_{k+1}) + (1-\theta)\vec{x}(t_k)\right)\right] \\
&= h x_i'(t_k) + \frac{h^2}{2} x_i''(t_k) - \varphi_i(h,\vec{x}(t_k)) f_i\left(\vec{x}(t_k) + \vec{H}\right) + \mathcal{O}(h^3) \\
&= h x_i'(t_k) + \frac{h^2}{2} x_i''(t_k) - \varphi_i(h,\vec{x}(t_k)) \left(f_i(\vec{x}(t_k)) + \langle \vec{H}, \nabla_x f_i(\vec{x}(t_k))\rangle + \mathcal{O}(h^2)\right) + \mathcal{O}(h^3) \\
&= h x_i'(t_k) + \frac{h^2}{2} x_i''(t_k) \\
&\quad - \left[h + (1-2\theta)\frac{\langle \nabla_x f_i(\vec{x}(t_k)), \vec{f}(\vec{x}(t_k))\rangle}{f_i(\vec{x})}\frac{h^2}{2}\right]\left(f_i(\vec{x}(t_k)) + \langle \vec{H}, \nabla_x f_i(\vec{x}(t_k))\rangle + \mathcal{O}(h^2)\right) + \mathcal{O}(h^3)
\end{aligned}$$

$$= \mathcal{O}(h^3).$$

Therefore, the numerical method (5.7) is of second-order accuracy.

To prove the elementary stability of the NSFD method (5.7), let $\vec{x}^*$ be a hyperbolic equilibrium of System (2.1) and $J = J(x^*)$ be the Jacobian evaluated at $\vec{x}^*$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. The corresponding linear system can then be given as follows:

$$\vec{x}' = J\vec{x} \tag{5.10}$$

If Λ is a Jordan form of J, then $J = S^{-1}\Lambda S$, where S is a non-singular complex $n \times n$ matrix. In general, Λ has the following bi-diagonal form:

$$\begin{pmatrix} \lambda_1 & \alpha_1 & & & \\ & \lambda_2 & \alpha_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & \alpha_{n-1} \\ & & & & \lambda_n \end{pmatrix}$$

where $\lambda_i \in \sigma(J)$. $i = 1, 2, \ldots n$, and $\alpha_i = \{0, 1\}$. Therefore, the linear system can be written as: $\vec{x}' = S^{-1}\Lambda S\ \vec{x}$ and the change of variables $\vec{y} = S\vec{x}$ yields the following new system

$$\vec{y}' = \Lambda\vec{y}.$$

Applying the numerical method (5.7) on the above system results in the following:

$$\begin{bmatrix} \dfrac{y_1^{k+1} - y_1^k}{\varphi_1(h, \vec{y^k})} \\ \dfrac{y_2^{k+1} - y_2^k}{\varphi_2(h, \vec{y^k})} \\ \vdots \\ \dfrac{y_n^{k+1} - y_n^k}{\varphi_n(h, \vec{y^k})} \end{bmatrix} = \Lambda \begin{bmatrix} \theta y_1^{k+1} + (1-\theta) y_1^k \\ \theta y_2^{k+1} + (1-\theta) y_2^k \\ \vdots \\ \theta y_n^{k+1} + (1-\theta) y_n^k \end{bmatrix} \tag{5.11}$$

and, equivalently, the following vector formulation:

$$\vec{y^{k+1}} = (I - V\Lambda\theta)^{-1}(I + V(1-\theta)\Lambda)\vec{y^k}, \tag{5.12}$$

where V is the diagonal matrix:

$$V = \begin{pmatrix} \varphi_1(h, \vec{y^k}) & & & & \\ & \varphi_2(h, \vec{y^k}) & & & \\ & & \ddots & & \\ & & & \varphi_{n-1}(h, \vec{y^k}) & \\ & & & & \varphi_n(h, \vec{y^k}) \end{pmatrix}.$$

Note that the matrix $(I - V\Lambda\theta)^{-1}(I + V(1-\theta)\Lambda)$ is upper triangular and its eigenvalues are given by $\mu_i(h, \vec{y^k}) = \dfrac{1 + \lambda_i(1-\theta)\varphi_i(h, \vec{y^k})}{1 - \lambda_i\theta\varphi_i(h, \vec{y^k})}$, where $\lambda_i \in \sigma(J)$, $i = 1, 2, \dots, n$. Observe that $\vec{x}^*$ being a stable fixed point of System (5.12) is equivalent to

$$\left|\frac{1 + \lambda_i(1-\theta)\varphi_i(h, \vec{y^k})}{1 - \lambda_i\theta\varphi_i(h, \vec{y^k})}\right| < 1. \tag{5.13}$$

Inequality (5.13) can be rewritten as

$$|1 + \lambda_i(1-\theta)\varphi_i(h, \vec{y^k})| < |1 - \lambda_i\theta\varphi_i(h, \vec{y^k})|,$$

and therefore

$$|1+\varphi_i(h,\vec{y}^{k})(1-\theta)\operatorname{Re}(\lambda_i)+\varphi_i(h,\vec{y}^{k})(1-\theta)\operatorname{Im}(\lambda_i)|^2 < |1-\varphi_i(h,\vec{y}^{k})\operatorname{Re}(\lambda_i)\theta-\varphi_i(h,\vec{y}^{k})\theta\operatorname{Im}(\lambda_i)|^2.$$

From here, a straightforward algebraic manipulation shows that Inequality (5.13) is equivalent to

$$\varphi_i(h,\vec{y}^{k})(1-2\theta) < \frac{-2\operatorname{Re}(\lambda_i)}{|\lambda_i|^2}. \tag{5.14}$$

Now consider the following two cases:

1. If $0 \leqslant \theta < 1/2$, then $1-2\theta > 0$. Condition (II) implies

   $$0 < \varphi_i(h,\vec{y}^{k}) < \frac{2|\operatorname{Re}(\lambda_i)|}{(1-2\theta)|\lambda_i|^2},$$

   and multiplying both sides by $(1-2\theta)$ yields

   $$0 < \varphi_i(h,\vec{y}^{k})(1-2\theta) < \frac{2|\operatorname{Re}(\lambda_i)|}{|\lambda_i|^2}.$$

   If $\vec{x}^*$ is a locally stable equilibrium, then $|\operatorname{Re}(\lambda_i)| = -\operatorname{Re}(\lambda_i)$. Thus from the above inequality, it can be seen that Inequality (5.14) holds. Hence, $\vec{x}^*$ is a stable fixed point.

   On the other hand, if $\vec{x}^*$ is an unstable equilibrium, then there is $j_0 \in \{1,\dots,n\}$ such that $\operatorname{Re}(\lambda_{j_0}) > 0$. This implies

   $$\varphi_{j_0}(h,\vec{y}^{k})(1-2\theta) > 0 > \frac{-2\operatorname{Re}(\lambda_{j_0})}{|\lambda_{j_0}|^2},$$

   since $\varphi_{j_0}(h,\vec{y}^{k}) > 0$. Therefore, Inequality (5.14) is strictly not satisfied, when $\vec{x}^*$ is an unstable equilibrium.

2. If $1/2 < \theta \leqslant 1$, then $1 - 2\theta < 0$. Multiplying the denominator function by $(1 - 2\theta)$ yields

$$\varphi_i(h, \vec{y}^*)(1 - 2\theta) < 0,$$

since $\varphi_i(h, \vec{y}^*) > 0$. If $\vec{x}^*$ is a locally stable equilibrium, then $\mathrm{Re}(\lambda_i) < 0$. Hence,

$$\varphi_i(h, \vec{y}^*)(1 - 2\theta) < 0 < \frac{-2\,\mathrm{Re}(\lambda_i)}{|\lambda_i|^2}.$$

Therefore, Inequality (5.14) is satisfied, and the fixed point $\vec{x}^*$ is stable. If $\vec{x}^*$ is an unstable equilibrium, then there is $j_0 \in \{1, \ldots, n\}$, such that $\mathrm{Re}(\lambda_{j_0}) > 0$. Condition (II) implies

$$0 < \varphi_{j_0}(h, \vec{x}) < \frac{2\,\mathrm{Re}(\lambda_{j_0})}{(2\theta - 1)|\lambda_{j_0}|^2},$$

and multiplying both sides by $-(2\theta - 1)$ yields

$$\varphi_{j_0}(h, \vec{x})(1 - 2\theta) > \frac{-2\,\mathrm{Re}(\lambda_{j_0})}{|\lambda_{j_0}|^2}.$$

Hence, Inequality (5.14) is strictly not satisfied. As a result, $\vec{x}^*$ is an unstable fixed point. Therefore, the numerical scheme (5.7) is elementary stable.

Next, the second-order accuracy of the modified nonstandard two-stage theta method (5.8) is similarly proven using the Taylor series expansion about t_k which yields:

$$
\begin{aligned}
&x_i(t_{k+1}) - \Big[x_i(t_k) + \varphi_i(h, \vec{x}(t_k))\Big\{\theta f_i\left(\vec{x}(t_{k+1})\right) + (1-\theta) f_i\left(\vec{x}(t_k)\right)\Big]\\
&= \Big[x_i(t_k) + hx_i'(t_k) + \frac{h^2}{2}x_i''(t_k) + \mathcal{O}(h^3)\Big] - \Big[x_i(t_k) + \varphi_i(h, \vec{x}(t_k))\Big\{\theta\Big[x_i'(t_k) + hx_i''(t_k)\\
&+ \frac{h^2}{2}x_i'''(t_k) + \mathcal{O}(h^3)\Big] + (1-\theta)f_i(\vec{x}(t_k))\Big\}\Big]\\
&= hf_i(\vec{x}(t_k)) + \frac{h^2}{2}\langle\nabla_x f_i(\vec{x}(t_k)), \vec{f}(\vec{x}(t_k))\rangle\\
&- \varphi_i(h, \vec{x}(t_k))\left[f_i(\vec{x}(t_k)) + \theta h\langle\nabla_x f_i(\vec{x}(t_k))\vec{f}(\vec{x}(t_k))\rangle\right] + \mathcal{O}(h^3) = \mathcal{O}(h^3),
\end{aligned}
$$

which implies the second order accuracy of the numerical method (5.8).

The proof of the elementary stability for the modified nonstandard two-stage theta method (5.8) uses the same arguments as the above proof for Method (5.7) and it is omitted here. □

Lemma 5.2.4. *Let $\phi_{i_1} : \mathbb{R}_+ \to \mathbb{R}_+$ and $\phi_{i_2} : \mathbb{R} \to \mathbb{R}_+$ satisfy the following conditions:*

(a) $0 < \phi_{i_1}(h) < 1$, for all $h > 0$, and $\phi_{i_1}(h) = h - \dfrac{h^2}{2} + \mathcal{O}(h^3)$.

(b) $0 < \phi_{i_2}(h) < M$, for all $h \in \mathbb{R}$ and some $M > 0$, and $\phi_{i_2}(h) = 1 + h + \mathcal{O}(h^3)$.

Then, the functions

$$\varphi_i(h, \vec{x}) = \frac{\phi_{i_1}(\alpha h)}{\alpha}\phi_{i_2}\left(\frac{\alpha - q_i(\vec{x})}{2}h\right),\ i = 1, \ldots, n,$$

with $\alpha > \dfrac{M}{2} \max_{\Omega} \dfrac{|2\theta - 1||\lambda|^2}{|Re(\lambda)|}$*, where* $\Omega = \bigcup_{\vec{x}^* \in \Gamma} \sigma(J(\vec{x}^*))$ *and* Γ *denotes the set of all equilibria of System* (2.1)*, and* $q_i(\vec{x}) = -(1-2\theta)\dfrac{\langle \nabla_x f_i(\vec{x}), \vec{f}(\vec{x})\rangle}{f_i(\vec{x})}$*, satisfy the conditions* (*I*) *and* (*II*) *of Theorem* (5.2.3)*.*

Proof. Notice that

$$\frac{\phi_{i_1}(\alpha h)}{\alpha} = h - \frac{\alpha h^2}{2} + \mathcal{O}(h^3), \quad \text{and}$$
$$\phi_{i_2}\left(\frac{\alpha - q_i(\vec{x})}{2} h\right) = 1 + \frac{\alpha - q_i(\vec{x})}{2} h + \mathcal{O}(h^3).$$

Therefore,

$$\varphi_i(h, \vec{x}) = h + (1-2\theta)\frac{\langle \nabla_x f_i(\vec{x}), \vec{f}(\vec{x})\rangle}{f_i(\vec{x})}\frac{h^2}{2} + \mathcal{O}(h^3),$$

which proves Condition (I). Next, since $0 < \phi_{i_1}(h) < 1$ and $0 < \phi_{i_2}(h) < M$, then one can easily see that

$$0 < \frac{\phi_{i_1}(\alpha h)}{\alpha}\phi_{i_2}\left(\frac{\alpha - q_i(\vec{x})}{2} h\right) < \frac{M}{\alpha} < \frac{2|Re(\lambda_i)|}{|2\theta - 1||\lambda_i|^2},$$

therefore, Condition (II) is also satisfied. □

Remark 15. There exists a variety of functions ϕ_{i_1} and ϕ_{i_2} that satisfy the conditions of Lemma 5.2.4. One such set of functions is $\phi_{i_1}(h) = 1 - e^{-h}$ and $\phi_{i_2}(h) = 1 + \tanh(h)$, which can be used to construct the denominator functions $\varphi_i(h, \vec{x})$ in Theorem 5.2.3.

5.2.3 General second-order modified nonstandard ERK2 method

The following result holds for the new two-stage modified nonstandard ERK2 method:

Theorem 5.2.5. *Let* $\vec{f} = [f_1, \ldots, f_n] \in C^2(\mathbb{R}^n; \mathbb{R}^n)$ *and let* $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ *satisfy the following conditions*

(I) $\varphi(h) = h + \mathcal{O}(h^3)$,

(II) $0 < \varphi(h) < \dfrac{1}{q}$, *for all* $\lambda \in \Omega$, *where* $q > \max_\Omega \frac{|\lambda|^2}{2|Re(\lambda)|}$, $\Omega = \bigcup_{\vec{x}^* \in \Gamma} \sigma(J(\vec{x}^*))$ *and* Γ *denotes the set of all hyperbolic equilibria* $\vec{x}^*$ *of System* (2.1).

Then the following two-stage modified nonstandard ERK2 method for approximating the solution of Equation (2.1)*:*

$$x_i^{k+1} = x_i^k + \varphi(h)\left\{(1-\omega)f_i(\vec{x}^k) + \omega f_i\left(\vec{x}^k + \frac{1}{2\omega}\vec{f}(\vec{x}^k)\varphi(h)\right)\right\},\ 0 < \omega \leqslant 1, \quad (5.15)$$

$i = 1, \ldots, n$, *is conservative, elementary stable and of second-order accuracy, provided the method does not introduce additional fixed points other than those of Equation* (2.1).

Proof. The second-order accuracy of the two-stage modified nonstandard ERK2 method (5.15) is similarly proven using the Taylor series expansion about t_k which yields:

$$\begin{aligned}
&x_i(t_{k+1}) - \left[x_i(t_k) + \varphi(h)\left\{(1-\omega)f_i(\vec{x}(t_k)) + \omega f_i\Big(\vec{x}(t_k) + \frac{1}{2\omega}\vec{f}(\vec{x}(t_k))\varphi(h)\Big)\right\}\right] \\
&= \left[x_i(t_k) + hx_i'(t_k) + \frac{h^2}{2}x_i''(t_k) + \mathcal{O}(h^3)\right] - \Bigg[x_i(t_k) + \varphi(h)\Bigg\{(1-\omega)f_i(\vec{x}(t_k)) \\
&\quad + \omega\left(f_i(\vec{x}(t_k)) + \frac{1}{2\omega}\sum_{j=1}^{n}\varphi(h)f_i(\vec{x}(t_k))\frac{\partial f_i}{\partial x_j} + \mathcal{O}(h^2)\right)\Bigg\}\Bigg] \\
&= x_i(t_k) + hf_i(\vec{x}_{t_k}) + \frac{h^2}{2}\sum_{j=1}^{n}\frac{\partial f_i}{\partial x_j}f_i(\vec{x}(t_k)) - \Bigg[x_i(t_k) + \varphi(h)\Bigg\{(1-\omega)f_i(\vec{x}(t_k)) \\
&\quad + \omega\left(f_i(\vec{x}(t_k)) + \frac{1}{2\omega}\varphi(h)\sum_{j=1}^{n}f_i(\vec{x}(t_k))\frac{\partial f_i}{\partial x_j} + \mathcal{O}(h^3)\right)\Bigg\}\Bigg] \\
&= (h-\varphi(h))f_i(\vec{x}(t_k)) + \left(\frac{h^2}{2} - \frac{\varphi^2(h)}{2}\right)\sum_{j=1}^{n}\frac{\partial f_i}{\partial x_j}f_i(\vec{x}(t_k)) + \mathcal{O}(h^3) = \mathcal{O}(h^3),
\end{aligned}$$

which implies the second order accuracy of the numerical method (5.15).

To prove the elementary stability of the NSFD method (5.15), let $\vec{x}^*$ be an hyperbolic equilibrium of System (2.1) and J be the Jacobian evaluated at $\vec{x}^*$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Since J is diagonalizable, using a similar argument as in the prrof of Theorem 5.2.3, the numerical method (5.15) can be applied to

$$\vec{y}' = \Lambda\vec{y}, \tag{5.16}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $J = S^{-1}\Lambda S$ and $\vec{y} = S\vec{x}$. This yields

$$y_i^{k+1} = y_i^k + \varphi(h)\left\{(1-\omega)\lambda_i y_i^k + \omega\lambda_i\left(y_i^k + \frac{1}{2\omega}\lambda_i y_i^k\varphi(h)\right)\right\},$$

which results in

$$\begin{aligned}
y_i^{k+1} &= \left[1+\varphi(h)(1-\omega)\lambda_i + \omega\lambda_i\varphi(h)\left(1+\frac{1}{2\omega}\lambda_i\varphi(h)\right)\right] y_i^k \\
&= \left[1+\varphi(h)(1-\omega)\lambda_i + \omega\lambda_i\varphi(h) + \frac{1}{2}\lambda_i^2\varphi^2(h)\right] y_i^k \\
&= \left[1+\varphi(h)\lambda_i + \frac{1}{2}\lambda_i^2\varphi^2(h)\right] y_i^k.
\end{aligned}$$

Therefore, to show that $\vec{x}^*$ is a stable equilibrium is equivalent to showing that

$$|1+\varphi(h)\lambda_i + \frac{1}{2}\lambda_i^2\varphi^2(h)| < 1. \tag{5.17}$$

Inequality (5.17) corresponds to

$$\left(1+\frac{1}{2}\lambda_i^2\varphi^2(h) + \mathrm{Re}(\lambda_i)\varphi(h)\right)^2 + (\mathrm{Im}(\lambda_i)\varphi(h))^2 < 1,$$

which is equivalent to

$$\frac{1}{4}|\lambda_i|^4\varphi^3(h) + \mathrm{Re}(\lambda_i)|\lambda_i|^2\varphi^2(h) + 2|\lambda_i|^2\varphi(h) + 2\,\mathrm{Re}(\lambda_i) < 0.$$

Denote $r_i(t) = \frac{1}{4}|\lambda_i|^4 t^3 + \mathrm{Re}(\lambda_i)|\lambda_i|^2 t^2 + 2\,\mathrm{Re}(\lambda_i)$. Thus, Inequality (5.17) is equivalent to $r_i(\varphi(h)) < 0$ for each λ_i, where $i = 1, 2, \ldots, n$. Similarly, to show that $\vec{x}^*$ is an unstable equilibrium point is equivalent to show that there exists an i such that $r_i(\varphi(h)) > 0$, and the rest of the proof follows from [67].

Finally, the conservative property of the two-stage modified nonstandard ERK2 method (5.15) is proven. First, observe that the denominator function $\varphi(h)$ is independent of x and hence, is the same for each component $i = 1, \ldots, n$, of the numerical method. Next, summing over all $i = 1, \ldots, n$, yields the following

$$\sum_{i=1}^{n} x_i^{k+1} = \sum_{i=1}^{n} x_i^k + \varphi(h)\left\{(1-\omega)\sum_{i=1}^{n} f_i(\vec{x}^k) + \omega\sum_{i=1}^{n} f_i\left(\vec{x}^k + \frac{1}{2\omega}\vec{f}(\vec{x}^k)\varphi(h)\right)\right\}$$

and, therefore,

$$\sum_{i=1}^{n} x_i^{k+1} = \sum_{i=1}^{n} x_i^k,$$

since $\sum_{i=1}^{n} f_i(\vec{x}^k) = \sum_{i=1}^{n} f_i\left(\vec{x}^k + \frac{1}{2\omega}\vec{f}(\vec{x}^k)\vec{\varphi}(h)\right) = 0.$ □

Remark 16. There exists a variety of functions $\varphi(h)$ which satisfy the conditions of above Theorem 5.2.5, and hence ensures a second-order accurate and elementary stable method (5.15). One such denominator function is $\varphi(h) = \dfrac{\tanh(qh)}{q}$, where $q > \max_{\Omega} \frac{|\lambda|^2}{2|Re(\lambda)|}$.

5.3 Numerical Simulations

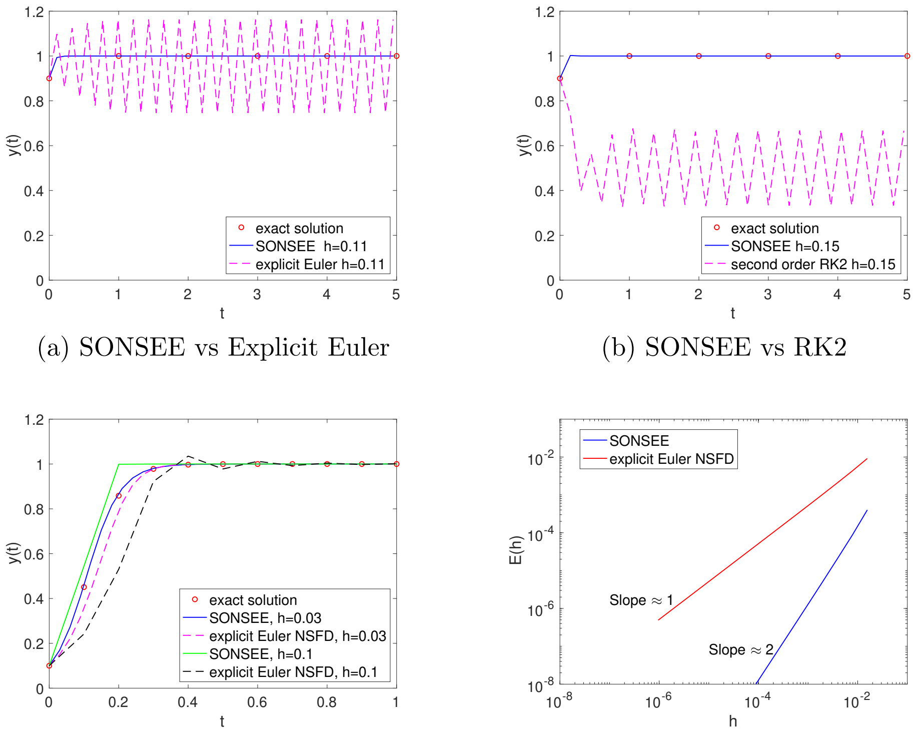
To illustrate our NSFD scheme, we consider the following logistic growth model:

$$\frac{dx}{dt} = ax\left(1 - \frac{x}{K}\right) \tag{5.18}$$

where K is carrying capacity and a represents growth rate. Hyperbolic Equilibrium points of this ODE are 0 (unstable equilibrium point) and K (stable equilibrium point). Numerical solution of equation (5.18) with $\varphi(h) = \frac{1-e^{-q(x_n)h}}{q(x_n)}$ where $q(x_n) = -f_x(x_n)$, $a = 20$ and $x(0) = 0.9$ and $\theta = 0$ support Theorem 5.2.3 for all h. We present a set of numerical simulations for $a = 2$ and $K = 1$.

In figure (a) we compare second order nonstandard explicit Euler (SONSEE) method with the explicit Euler method. We note that when $h = 0.11$, the explicit Euler method diverges away from the equilibrium point. Figure (b), for $h = 0.15$ represents the comparison of SONSEE with second order Runge Kutta (RK2) which is of order two but not elementary stable. In figure (c), for $h = 0.1$ and $h = 0.03$ we compared

SONSEE method with explicit Euler NSFD with fix $q = 0.5$ which is elementary stable but of first order. Figure (d) actually shows that SONSEE is of second order while explicit Euler NSFD is of first order. Above comparisons verifies all claims numerically which has been stated in the theorems about SONSEE.



(a) SONSEE vs Explicit Euler

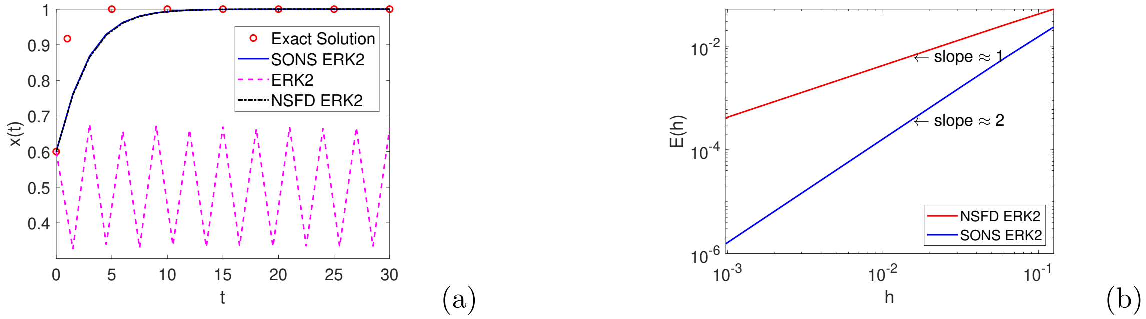
(b) SONSEE vs RK2

(c) SONSEE vs NSFD

(d) SONSEE vs NSFD error plot

Figure 5.1: Comparison of SONSEE with different methods

The new second-order modified nonstandard explicit Runge-Kutta (SONS ERK2) method (5.15) with $\omega = \dfrac{1}{2}$, is numerically compared to the standard ERK2



Comparison of SONS ERK2, NSFD ERK2, and ERK2

Error plot: SONS ERK2 vs NSFD ERK2

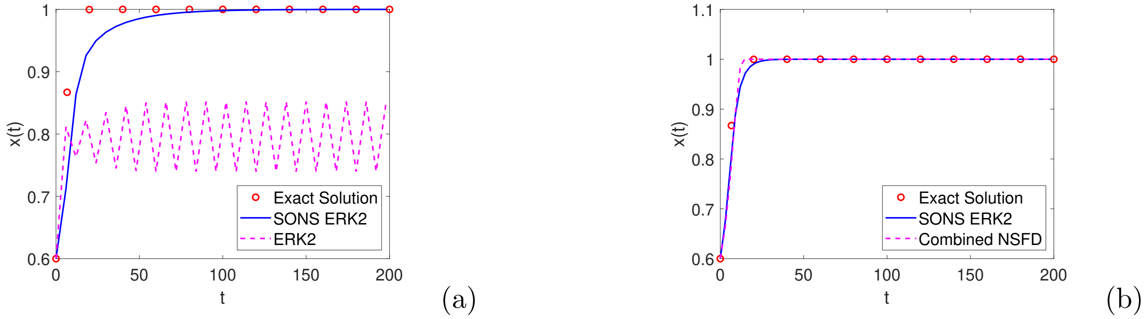
Figure 5.2: Numerical solutions of Equation (5.18) with $q = 2.5,\ x_0 = 0.6$, and using $h = 1.5$ in (a)

method and the NSFD ERK2 method [71, 73]. We use the nonstandard denominator function

$$\varphi(h, x) = \frac{\tanh(qh)}{q}, \text{ with } q = 2.5 > \frac{\max\{|f_x(0)|, |f_x(1)|\}}{2} = 1.$$

The SONS ERK2 method is of second-order accuracy and elementary stable, while the ERK2 method is also second-order accurate but unstable for $h > \dfrac{2}{a} = 1$, and the NSFD ERK2 method is elementary stable but only first-order accurate. Accordingly, for $h = 1.5$, we see in Figure 5.2(a) that the ERK2 method does not converge to the exact solution whereas both the SONS ERK2 and NSFD ERK2 methods correctly mimic the behavior of the exact solution. To better visualize the second-order accuracy of the new SONS ERK2 method (5.15), we denote the numerical solution for a given mesh size h as x^h. Let us define the l^∞ error as

$$E(h) = \|x^h - x\|_\infty,$$



Comparison of SONS ERK2 and ERK2

Comparison of SONS ERK2 and Combined NSFD

Figure 5.3: Numerical solutions of Equation (5.19) for $b = \dfrac{1}{2}$, with $q = 0.26, x_0 = 0.6$, and using $h = 6$ in (a) and $h = 3$ in (b)

where

$$\|y\|_\infty = \max_{k=0,\cdots,N_t} |y_k|$$

represents the discrete l_∞ norm of the vector y, and x represents the exact solution of Equation (2.1). Figure 5.2(b) shows the error plot for NSFD ERK2 and SONS ERK2 methods, where the slopes of the error lines are 1 and 2, respectively. This numerically verifies that the SONS ERK2 method is second-order accurate while the NSFD ERK2 method is only first-order accurate.

As a second example, we consider the following differential equation which is a modification of the predator pit model in population ecology ([123], p. 115):

$$\frac{dx}{dt} = -\left(x - b + \frac{1}{2}\right)(x - b)\left(x - b - \frac{1}{2}\right). \tag{5.19}$$

Equation (5.19) has $x^* = b$ as an unstable equilibrium while $x^* = b \pm \frac{1}{2}$ are both stable equilibria, with $\max\{|f_x(x^*)|\} = \frac{1}{2}$. To support the results of Theorem 5.2.5, we perform numerical simulations using the nonstandard denominator function

$$\varphi(h, x) = \frac{\tanh(qh)}{q}, \text{ with } q = 0.26 > \frac{\max\{|f_x(x^*)|\}}{2} = 0.25.$$

First, we consider $b = \dfrac{1}{2}$, which results in the right-hand side function

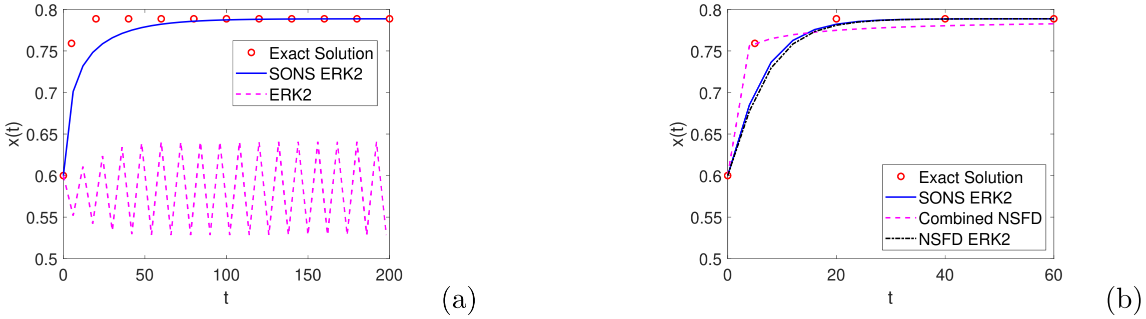
$$f(x) = -x^3 + \frac{3}{2}x^2 - \frac{1}{2}x.$$

Figure 5.3(a) shows a comparison of the SONS ERK2 method with the standard ERK2 method, for $h = 6$ and initial condition $x_0 = 0.6$. Simulations show that the ERK2 method does not converge to the exact solution for large values of h, while the SONS ERK2 method preserves the local stability properties of the equilibrium $x^* = 1$ for any value of the step-size h. Figure 5.3(b) shows a comparison of the SONS ERK2 method with the combined NSFD method [124] for $h = 3$. The two numerical methods reproduce the correct behavior of the exact solution as they are both of second order accuracy and elementary stable, however the combined NSFD method is implicit in nature and, therefore, not as computationally easy to implement.

Next, we consider $b = \dfrac{1}{2\sqrt{3}}$, that results in the right-hand side function

$$f(x) = -x^3 + \frac{\sqrt{3}}{2}x^2 - \frac{1}{12\sqrt{3}}.$$

A similar set of numerical comparisons was performed as in the case with $b = \frac{1}{2}$ and the same results were obtained, as shown in Figure 5.4. In this case, the combined NSFD method, which is of second-order accuracy and elementary stable, does not



Comparison of SONS ERK2 and ERK2

SONS ERK2, Combined NSFD, and NSFD ERK2

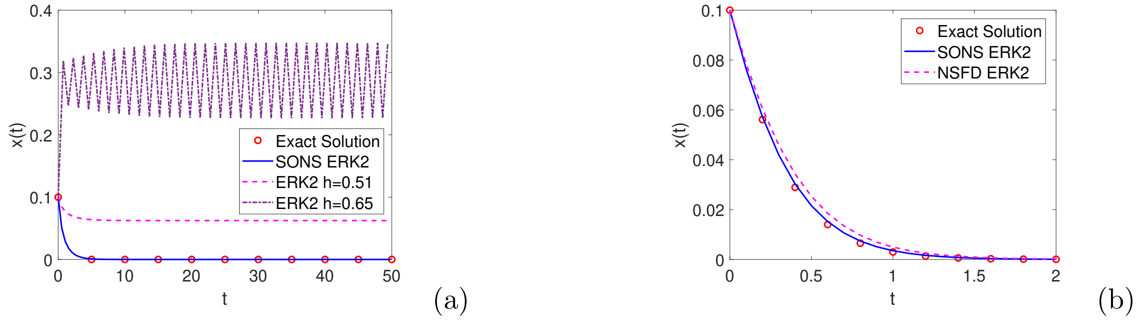
Figure 5.4: Numerical solutions of Equation (5.19) for $b = \dfrac{1}{2\sqrt{3}}$, with $x_0 = 0.6$, and using $h = 6, q = 0.26$ in (a) and $h = 4, q = 0.45$ in (b)

require a nonstandard denominator function, since the right-hand side function $f(x)$ does not contain a first term [124]. However, it is again an implicit method, and therefore still not as computationally easy to implement as the explicit SONS ERK2 method. In the third example, we consider the Michaelis-Menten model (Allen), where the rate of change of the nutrient concentration $x(t)$ used by a cell for growth and development is modeled by the following differential equation:

$$\frac{dx}{dt} = -\frac{k_{max}x}{k_n + x}. \tag{5.20}$$

Here, the parameter $k_{\max} > 0$ is the maximum rate of uptake by the cell of the nutrient and $k_n > 0$ is the half-saturation constant. Given that $f(x) = -\dfrac{k_{max}x}{k_n + x}$, yields $f'(0) = -k_{\max}/k_n < 0$ and, therefore, $x^* = 0$ is a stable equilibrium of Equation (5.20). In the numerical simulations, we take $k_n = 0.2, k_{\max} = 0.8$, with an initial condition $x(0) = 0.1$, and $q = 0.25$ for the comparison of our method with the NSFD ERK2 method.

Figure 5.5(a) shows a comparison of the SONS ERK2 method with the ERK2 method,



Comparison of SONS ERK2 and ERK2

SONS ERK2 and NSFD ERK2 for $q = 0.25, h = 0.1$

Figure 5.5: Numerical solutions of Equation (5.20) with $x_0 = 0.1$

where we see that the ERK2 method introduces artificial equilibria for $h = 0.51$ and becomes unstable when $h = 0.65$, while the SONS ERK2 method behaves very well for arbitrary large values of h. Figure 5.5(b) shows a comparison of our method with the nonstandard ERK2 method which is elementary stable but only of first-order accuracy, and therefore, the numerical solution of the SONS ERK2 method converges faster to the stable equilibrium $x^* = 0$. Next, we will discuss the performance of the proposed n-dimensional new modified NSFD methods, the new second-order modified nonstandard theta method (5.8) with $\theta = 0$ and two-stage modified nonstandard ERK2 method (5.15) with $\omega = 1/2$ are chosen. Note that for this value of θ, the modified NSFD method (5.8) is the same as Method (5.7), which results in the modified NSFD explicit Euler (modified NSFD EE) method [75]. Furthermore, the two-stage modified nonstandard ERK2 method (5.15) with $\omega = 1/2$ is henceforth referred to as the modified NSFD ERK2 method. The modified NSFD methods are compared to other standard and nonstandard finite difference methods for solving two specific biological systems.

For the numerical test cases, the MSEIR epidemiological model in [125, 126] is first considered, with the notation $\vec{x} = (x_1, x_2, x_3, x_4, x_5) = (m, s, e, i, r)$:

$$\begin{aligned} \frac{dx_1}{dt} &= d(x_3 + x_4 + x_5) - \delta x_1, \\ \frac{dx_2}{dt} &= -\beta x_2 x_4 + \delta x_1, \\ \frac{dx_3}{dt} &= \beta x_2 x_4 - (\epsilon + d)x_3, \\ \frac{dx_4}{dt} &= \epsilon x_3 - (\gamma + d)x_4, \\ \frac{dx_5}{dt} &= \gamma x_4 - dx_5. \end{aligned} \tag{5.21}$$

The following initial conditions $x_1(0) = 0.1, x_2(0) = 0.05, x_3(0) = 0.05, x_4(0) = 0.1, x_5(0) = 0.7$ and parameter values $d = 1/(40 \times 365), \beta = 0.14, \gamma = 1/7, \delta = 1/180, \epsilon = 1/14$ are used in numerical simulations.

The novel nonstandard denominator functions φ_i for the modified NSFD EE method are selected using Remark 15 as follows:

$$\varphi_i(h, \vec{x}) = \left(\frac{1 - \exp(-\alpha h)}{\alpha}\right)\left(1 + \tanh\left(\frac{\alpha - q_i(\vec{x})}{2}h\right)\right),$$

for $i = 1, \ldots, 5$. Here, the parameters q_i are given according to Lemma 5.2.4:

$$q_i = -\frac{(f_1\frac{\partial f_i}{x_1} + f_2\frac{\partial f_i}{\partial x_2} + f_3\frac{\partial f_i}{\partial x_3} + f_4\frac{\partial f_i}{\partial x_4} + f_5\frac{\partial f_i}{\partial x_5})}{f_i},$$

$i = 1, \ldots, 5$, where

$$f_1(\vec{x}) = d(x_3 + x_4 + x_5) - \delta x_1$$

$$f_2(\vec{x}) = -\beta x_2 x_4 + \delta x_1$$

$$f_3(\vec{x}) = \beta x_2 x_4 - (\epsilon + d)x_3$$

$$f_4(\vec{x}) = \epsilon x_3 - (\gamma + d)x_4$$

$$f_5(\vec{x}) = \gamma x_4 - dx_5.$$

The Jacobian matrix has the form:

$$\begin{pmatrix} -\delta & 0 & d & d & d \\ \delta & -\beta x_4 & 0 & -\beta x_2 & 0 \\ 0 & \beta x_4 & -(\epsilon + d) & \beta x_2 & 0 \\ 0 & 0 & \epsilon & -(\gamma + d) & 0 \\ 0 & 0 & 0 & \gamma & -d \end{pmatrix}.$$

and the eigenvalues evaluated at the epidemic equilibrium are: $\lambda_1 = -0.214422$, $\lambda_2 = -0.00555541$, $\lambda_3 = -0.000300252$, $\lambda_4 = 0.000232743$ and $\lambda_5 = 1.95907 \times 10^{-18}$. Accordingly, the value $\alpha = 0.3 > 0.214422 = \max\left\{\frac{|\lambda_i|^2}{|Re(\lambda_i)|} : i = 1, \ldots, 5\right\}$ has been used in the denominator function of the modified NSFD EE method. For the modified NSFD ERK2 method, the denominator function is chosen as $\varphi(h) = \dfrac{\tanh(qh)}{q}$, with $q = 0.25 > 0.107211 = \max\left\{\frac{|\lambda_i|^2}{2|Re(\lambda_i)|} : i = 1, \ldots, 5\right\}$.

Figure (5.6a) compares the modified NSFD EE method with the explicit Euler (EE) method [127] for $h = 16$. For illustration purposes, only the infected population x_4 plots are shown. It is known that the EE method does not preserve the local stability of the equilibria [127]. As it can be seen from the figure, the numerical solution from the EE method oscillates and diverges from the exact solution, while the numerical solution from the modified NSFD EE method behaves well and converges to the exact solution. Figure (5.6b) compares the modified NSFD ERK2 method with the standard explicit second-order Runge Kutta (ERK2) method [127] for $h = 16$. While the ERK2 method is second-order accurate, it is not elementary stable. It can be seen that the numerical solution from the modified NSFD ERK2 method converges to the exact solution while the numerical solution from the ERK2 method diverges and eventually blows up to infinity for large values of the step-size h. Figure (5.6c) compares the NSFD EE method [67] with the modified NSFD EE method. For step-size $h = 0.95$, the numerical solution from the second-order modified NSFD EE method converges to the exact solution faster than from the first-order NSFD EE method. In Figure (5.6d), for $h = 0.95$, it can also be seen that the numerical solution from the second-order modified NSFD ERK2 method is more accurate than from the NSFD ERK2 method [67]. Next, the two new modified nonstandard methods are compared with two of the nonstandard numerical methods presented in [125]. Note that the method in [125] for $\theta = 0, \hat{\theta} = 1$, is an explicit NSFD method, but with a different approximation of the nonlinear right-hand side terms and a denominator function $\varphi(h) = (1 - \exp(-Qh))/Q$, where $Q = \max\{\delta, \epsilon + d, \gamma + d, \beta\}$. Also, the method in [125] for $\theta = \hat{\theta} = \frac{1}{2}$ is a second-order NSFD method with a denominator function $\varphi(h) = \dfrac{\tanh(Qh)}{Q}$, where $Q = \frac{1}{2}\max\{\delta, \epsilon + d, \gamma + d, \beta\}$. In Figure (5.6e) these two numerical methods are denoted by NSFD $\theta = 0, \hat{\theta} = 1$ and NSFD $\theta = \hat{\theta} = \frac{1}{2}$, respectively. It can be seen in the figure, the behavior of the numerical solutions from the modified NSFD EE and the modified NSFD ERK2 methods is superior to those produced by the NSFD methods in [125]. In addition, the absolute error plots of the modified NSFD EE and the modified NSFD ERK2 methods are presented for $h = 1$. The absolute errors are calculated by using as a benchmark the numerical solution obtained by the MATLAB® ode45 solver. As can be seen in Figure (5.6f), the error in the numerical solution from the modified NSFD ERK2 method is less than the error from the modified NSFD EE method.

For the next numerical test case, the predator-prey system with Beddington-DeAngelis functional response in [67, 73] is considered, with the notation $\vec{x} = (x_1, x_2) = (x, y)$:

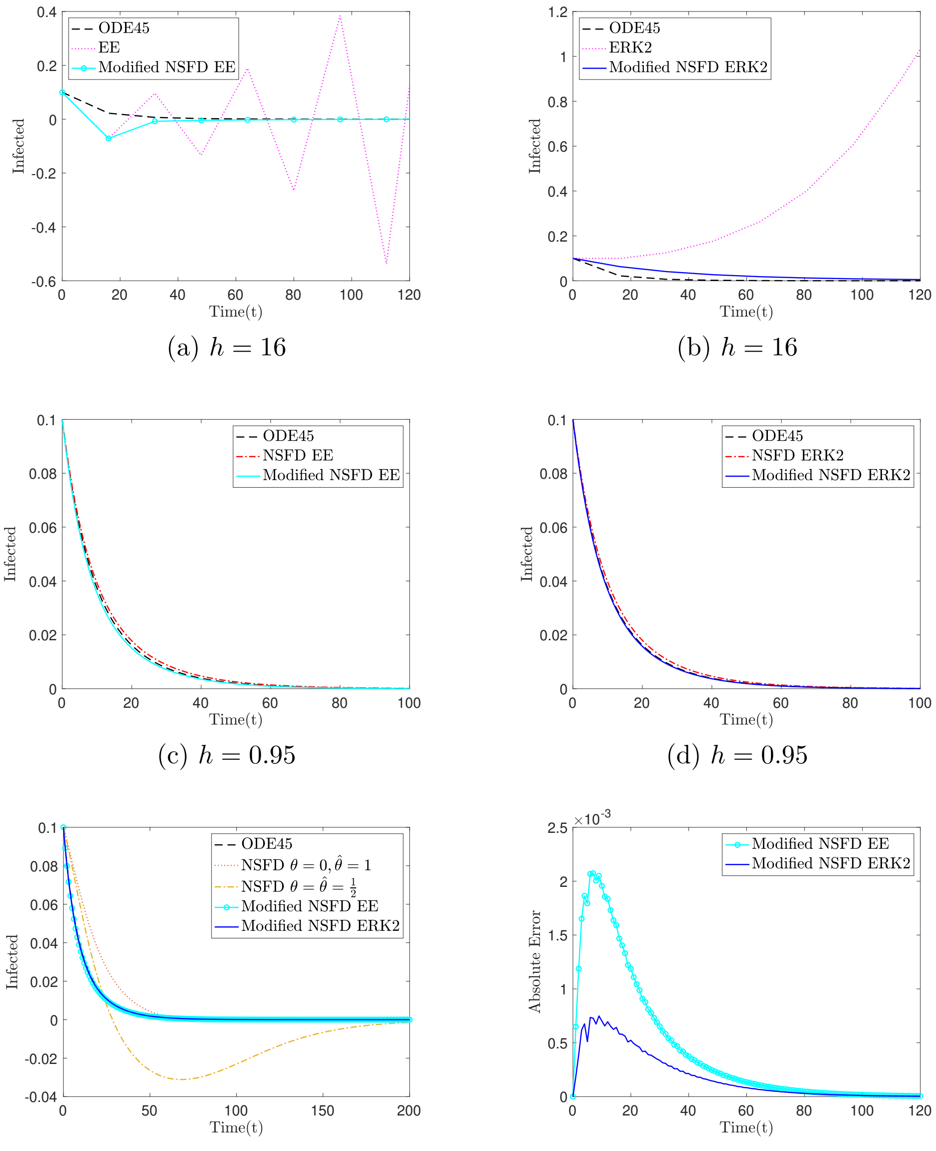
$$\begin{aligned} \frac{dx_1}{dt} &= x_1 - \frac{Ax_1x_2}{1 + x_1 + x_2}, \\ \frac{dx_2}{dt} &= \frac{Ex_1x_2}{1 + x_1 + x_2} - Dx_2, \end{aligned} \tag{5.22}$$

where x_1 and x_2 represent the prey and predator population sizes, respectively. The following parameter values $A = 6.0, D = 5.0$, and $E = 7.5$ are used in the numerical simulations.

Stability analysis of System (5.22) reveals that there exist two equilibria [73]: $(0, 0)$ and $\left(\frac{AD}{AE-E-AD}, \frac{E}{AE-E-AD}\right) = (4, 1)$. The eigenvalues of the Jacobian matrix evaluated at $(0, 0)$ are $\lambda_1 = 1$ and $\lambda_2 = -5$, while the eigenvalues evaluated at $(4, 1)$ are $\lambda_{3,4} = -\frac{1}{12} \pm i\frac{\sqrt{119}}{12}$, with $|\lambda_{3,4}| = 0.9129$. Therefore, the coexistence equilibrium $(4, 1)$ is globally asymptotically stable in the interior of the first quadrant, while $(0, 0)$ is unstable.

The novel nonstandard denominator functions φ_i for the modified NSFD EE method are selected using Remark 15 as follows:

$$\varphi_i(h, \vec{x}) = \left(\frac{1 - \exp(-\alpha h)}{\alpha}\right)\left(1 + \tanh\left(\frac{\alpha - q_i(\vec{x})}{2}h\right)\right),$$



(a) $h = 16$

(b) $h = 16$

(c) $h = 0.95$

(d) $h = 0.95$

(e) $h = 1$

(f) Absolute error plots for $h = 1$

Figure 5.6: Comparison of the modified NSFD EE and the modified NSFD ERK2 methods to other numerical methods, applied to Model (5.21), using different values of the step-size h. For illustration purposes, only the infected population plots are shown.

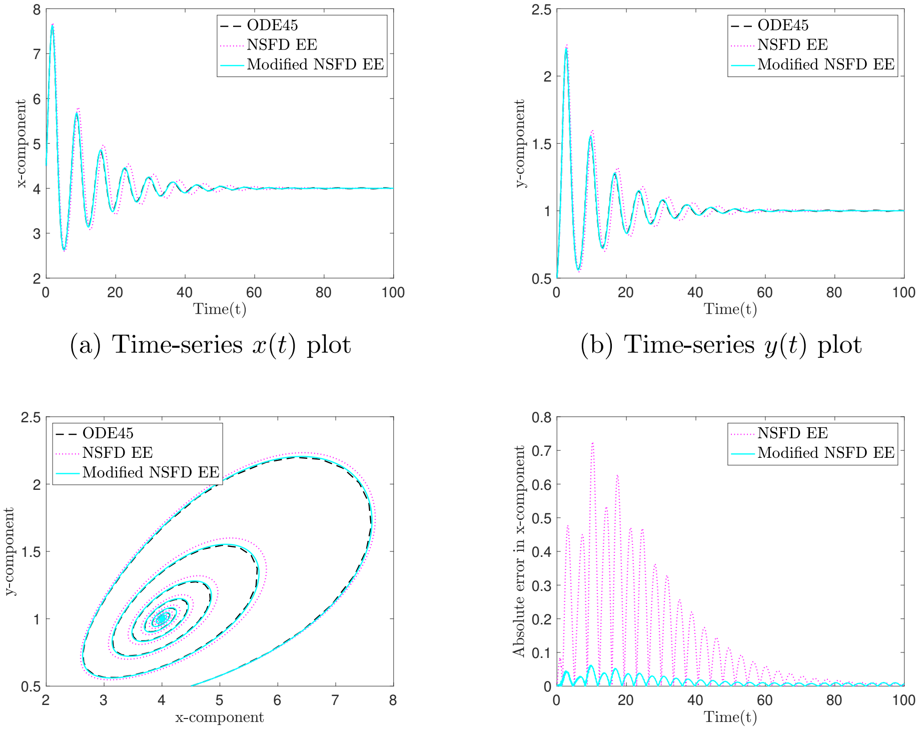
for $i = 1, 2$, using

$$q_i(\vec{x}) = -\frac{\left(f_1(x_1, x_2)\frac{\partial f_i}{\partial x_1} + f_2(x_1, x_2)\frac{\partial f_i}{\partial x_2}\right)}{f_i(x_1, x_2)},$$

where $f_1(x_1, x_2) = x_1 - \dfrac{Ax_1x_2}{1 + x_1 + x_2}$ and $f_2(x_1, x_2) = \dfrac{Ex_1x_2}{1 + x_1 + x_2} - Dx_2$, and $\alpha = 10.1 > \max_\Omega \frac{|\lambda|^2}{|Re(\lambda)|}$. For the modified NSFD ERK2 method, the denominator function is chosen as $\varphi(h) = \dfrac{\tanh(qh)}{q}$, with $q = 5.1 > \max_\Omega \frac{|\lambda|^2}{2|Re(\lambda)|}$, where $\Omega = \bigcup_{\vec{x}^* \in \Gamma} \sigma(J(\vec{x}^*))$, and Γ denotes the set of all hyperbolic equilibria $\vec{x}^*$ of System (2.1). Figure 5.7 compares the modified NSFD EE method with the NSFD EE method for $h = 0.02$. As can be seen in Figures 5.7(a)-(c), there is a slight horizontal shift in both components of the numerical solution from the first-order NSFD EE method, while the numerical solution from the modified NSFD EE method converges much more accurately to the exact solution. In addition, absolute error plots are presented in Figure 5.7(d), where the new modified NSFD EE method clearly outperforms the NSFD EE method. Again, the numerical solution obtained by the MATLAB® ode45 solver has been used as a benchmark. Similar comparison of numerical methods is presented in Figure 5.8, where the second-order modified NSFD ERK2 method outperforms the first-order NSFD ERK2 method for $h = 0.05$.

5.4 Conclusion

In this paper, two new classes of second-order modified nonstandard theta and Runge-Kutta methods for multi-dimensional autonomous dynamical systems have been presented and analyzed. The fundamental idea underlying the numerical methods' development is the use of a novel modified nonstandard denominator function



(a) Time-series $x(t)$ plot

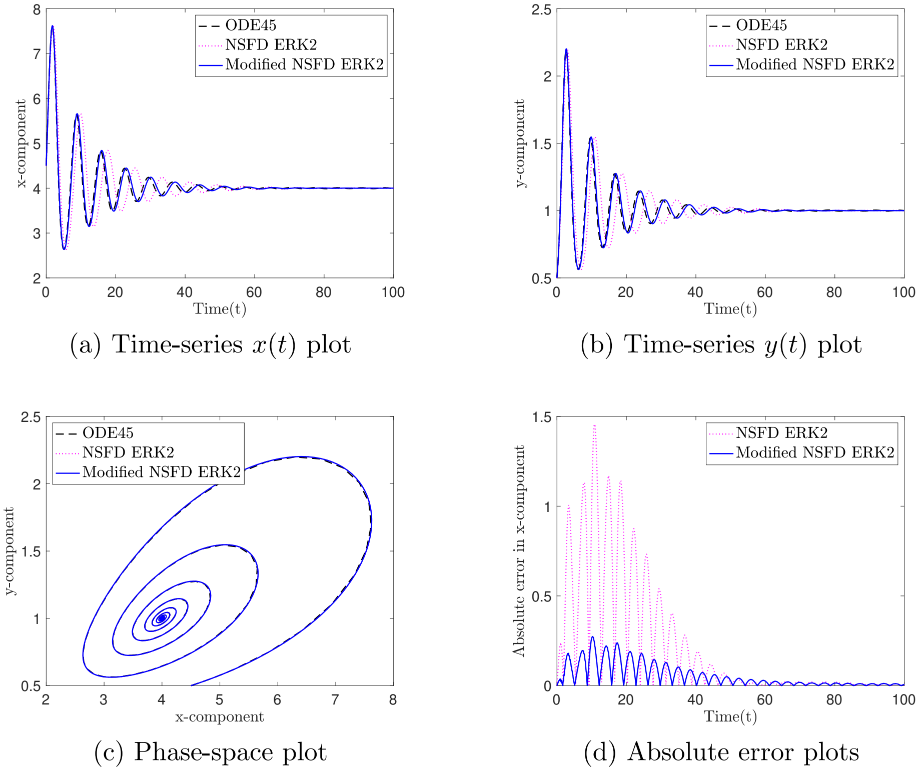
(b) Time-series $y(t)$ plot

(c) Phase-space plot

(d) Absolute error plots

Figure 5.7: Comparison of the modified NSFD EE method to the NSFD EE method, applied to Model (5.22), using $h = 0.02$. The numerical solution obtained by the MATLAB® ode45 solver has been used as a benchmark.

in the discretization of the derivative. In the case of the modified NSFD theta methods, the denominator function is a product of two special functions. One of the functions satisfies the methods' second-order accuracy property, while the other function satisfies the stability criteria of Theorem 5.2.3. For the modified nonstandard Runge-Kutta method, a single denominator function suffices to satisfy the criteria of accuracy and elementary stability. Examples of denominator functions have been presented that also serve as recipes for the choice of the generic ones. Next, the proposed



(a) Time-series $x(t)$ plot

(b) Time-series $y(t)$ plot

(c) Phase-space plot

(d) Absolute error plots

Figure 5.8: Comparison of the modified NSFD ERK2 method to the NSFD ERK2 method, applied to Model (5.22), using $h = 0.05$. The numerical solution obtained by the MATLAB® ode45 solver has been used as a benchmark.

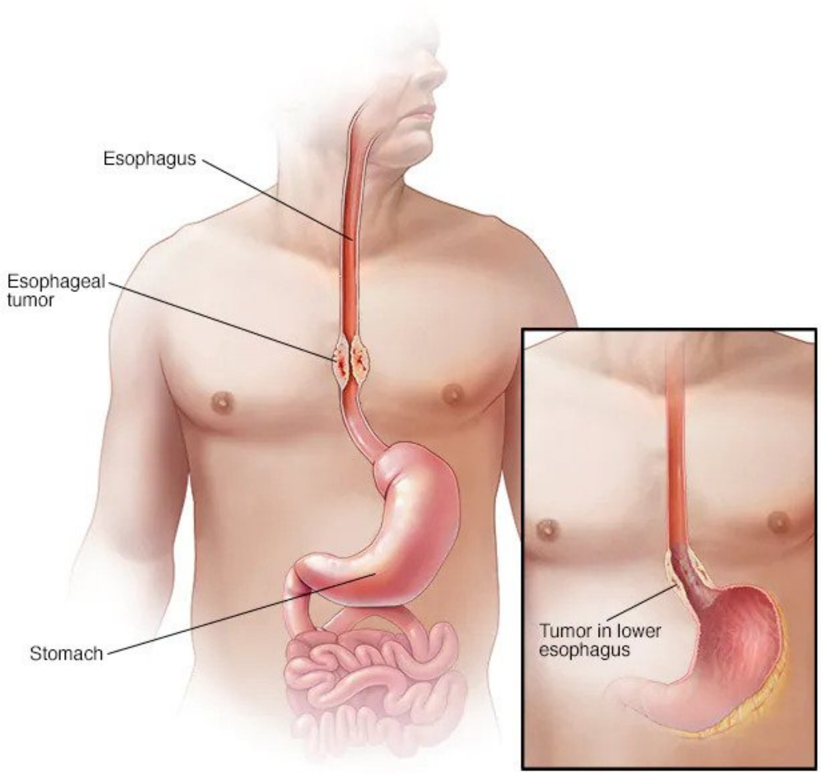
modified nonstandard methods have been applied to solve an MSEIR system and a predator-prey system with Beddington-DeAngelis functional response. The results obtained using the new numerical methods have been compared to existing standard and nonstandard finite difference methods, and it has been observed that the new methods demonstrate high accuracy and better stability properties. Future research directions include the development of modified nonstandard numerical methods that are not only higher-order accurate and elementary stable but also preserves other important properties of the exact solution, such as positivity.

CHAPTER 6

NSFD Scheme for Fokker-Planck framework in esophageal cancer

In previous chapter, we have discussed the NSFD schemes for a class of ordinary differential equations. In this cahpter, we will discuss an application of NSFD for Fokker-Planck framework in esophageal cancer. Calcium signaling plays important role in esophageal cancer.

Esophageal Cancer: The esophagus is a muscular tube connecting the throat (pharynx) with the stomach.The esophagus is about 8 inches long and is lined by moist pink tissue called mucosa and it runs behind the windpipe (trachea) and heart, and in front of the spine.



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Esophageal cancer usually begins in the cells that line the inside of the esophagus. Esophageal cancer can occur anywhere along the esophagus. More men than women get esophageal cancer. This cancer is the 6th most common cause of cancer mortality globally. Incidence rates vary within different geographic locations. In some regions, higher rates of esophageal cancer may be attributed to tobacco and alcohol use or particular nutritional habits and obesity. In a cancer patient, one of the primary dysregulated intracellular Ca2+ signaling pathways is the store operated Ca2+ entry (SOCE). Fluctuations in SOCE is linked to tumor cell proliferation, metastasis and impedance to apoptosis. In accordance to experimental evidence, the messenger IP_3 allows calcium to be released to the cytosol from an IP_3 sensitive store which is endoplasmic reticulam (ER). Furthermore, calcium is pumped out of the cytosol to the ER and, also calcium leaks into the cytosol from outside the cell.

6.1 Mathematical Model

To illustrate the aim, we consider the dynamics based on very famous Atri models, that represents the dynamics of Ca2+ ions, inositol trisphosphate (IP3) pathway .

$$\frac{dc}{dt} = \underbrace{k_f \frac{p}{k_\mu + p} h \frac{bk_1 + c}{k_1 + c}}_{J_{channel}} - \underbrace{\frac{\gamma}{k_\gamma + c}}_{J_{pump}} + \underbrace{\beta}_{J_{leak}} \tag{6.1}$$

$$\tau_h \frac{dh}{dt} = \underbrace{\frac{k_2^2}{k_2^2 + c^2}}_{h_\infty} - h \tag{6.2}$$

- c is the cytosolic calcium concentration
- $J_{channel}$ models the flux of calcium from the ER into the cytosol through the IP_3 receptors, assuming that calcium activates the IP_3 receptors quickly but inactivates them on a slower timescale;
- J_{pump} models the calcium pumped out of the cytoplasm back to the ER or out through the plasma membrane, and
- $J_{leak} = \beta$ models the calcium leaking into the cytosol from outside the cell.
- h is the fraction of IP_3 receptors on the ER that have not been closed (inactivated) by calcium. In other words h is the rate at which Ca^{2+} can activate IP3R.
- k_f is the maximum total Ca^{2+} flux through all IP_3Rs when all IP_3Rs are open and activated
- β is the constant rate of Ca^{2+} influx into the cytosol from the outside
- γ is the maximum rate of Ca^{2+} pumping from the cytosol
- k_γ is the concentration of Ca^{2+} at which the rate of Ca^{2+} pumping from the cytosol is at half maximum.

- h represents the fraction of IP_3 receptors on the ER that have not been inactivated by calcium.
- The function h_∞ represents the steady state of h as a function of c
- τ_h is the time constant for the dynamics of h, the proportion of IP_3Rs is not closed by Ca^{2+}.

with the following values of parameters:

Parameter	Value
c	$0.01\mu M$
h	$0-1$
b	0.111
k_1	$0.7\mu M$
k_2	$0.7\mu M$
k_f	$16.2\mu M/s$
k_γ	$0.1\mu M$
k_μ	$0.7\mu M$
γ	$2\mu M/s$
τ_h	$2s$
β	$0.01\mu M/s$ (Range : $0-0.02$)

where μM represents micromolar,

6.2 Nondimensionalisation of the dynamical model

To non-dimensionalise the equation, we do the following substitution: $c = \bar{s}_c\bar{c}, h = \bar{s}_h\bar{h}, t = \bar{s}_t\bar{t}$, where dimension of $\bar{s}_c$ is μM , $\bar{s}_h$ is dimensionless scaling factor and dimension of $\bar{s}_t$ is second.

$$\frac{d\bar{c}}{d\bar{t}} = \frac{\bar{s}_t\bar{s}_h}{\bar{s}_c}k_f\frac{\frac{p}{k_\mu}}{1+\frac{p}{k_\mu}}\bar{h}\frac{b\frac{k_1}{s_c}+\bar{c}}{\frac{k_1}{\bar{s}_c}+\bar{c}} - \frac{s_t\gamma}{s_c}\frac{\bar{c}}{\frac{k_\gamma}{\bar{s}_c}+\bar{c}} + \frac{\bar{s}_t}{\bar{s}_c}\beta \tag{6.3}$$

$$\frac{d\bar{h}}{d\bar{t}} = \frac{\bar{s}_t}{\bar{s}_h\tau_h}\frac{(\frac{k_2}{\bar{s}_c})^2}{(\frac{k_2}{\bar{s}_c})^2+\bar{c}^2} - \frac{\bar{s}_t}{\tau_h}\bar{h} \tag{6.4}$$

Let us now put $\bar{k}_f = \frac{k_f\bar{s}_t\bar{s}_h}{\bar{s}_c}$,$\bar{P} = \frac{p}{k_\mu}$, $\bar{k}_1 = \frac{k_1}{\bar{s}_c}$, $\bar{k}_\gamma = \frac{k_\gamma}{\bar{s}_c}$, $\bar{\gamma} = \frac{\gamma\bar{s}_t}{\bar{s}_c}$, $\bar{\beta} = \frac{\bar{s}_t\beta}{\bar{s}_c}$, $\bar{k}_2 = \frac{k_2}{\bar{s}_c}$, $\bar{\tau}_h = \frac{\tau_h\bar{s}_h}{\bar{s}_t}$, and we obtain,

$$\frac{d\bar{c}}{d\bar{t}} = \bar{k}_f\mu(\bar{p})\bar{h}\frac{b\bar{k}_1+\bar{c}}{\bar{k}_1+\bar{c}} - \frac{\bar{\gamma}\bar{c}}{\bar{k}_\gamma+\bar{c}} + \bar{\beta} \tag{6.5}$$

$$\bar{\tau}_h\frac{d\bar{h}}{d\bar{t}} = \frac{\bar{k}_2^2}{\bar{k}_2^2+\bar{c}^2} - s_h\bar{h} \tag{6.6}$$

For the sake of convenience, dropping the bar notation and the equation as following:

$$\frac{dc}{dt} = k_f\mu(p)h\frac{bk_1+c}{k_1+c} - \frac{\gamma c}{k_\gamma+c} + \beta \tag{6.7}$$

$$\tau_h\frac{dh}{dt} = \frac{k_2^2}{k_2^2+c^2} - s_h h \tag{6.8}$$

6.3 Fokker-Planck framework

Let us write our original equations with initial conditions:

$$\frac{dc}{dt} = k_f\mu(p)h\frac{bk_1+c}{k_1+c} - \frac{\gamma c}{k_\gamma+c} + \beta, \quad c(0) = c_0 \tag{6.9}$$

$$\tau_h\frac{dh}{dt} = \frac{k_2^2}{k_2^2+c^2} - s_h h, \quad h(0) = h_0 \tag{6.10}$$

We will now write above equations in compact form as follows:

$$\frac{dX(t)}{dt} = u(X(t),t), \quad X(t_0) = X_0$$

where $X(t) = [x_1(t), x_2(t)]^t$ and $x_1(t) = c(t)$ and $x_2(t) = h(t)$, $X(0) = [x_1(0), x_2(0)]^t = [c(0), h(0)]^t$ corresponds the initial condition, and

$$u_1(X(t), t) = k_f \mu(p) h \frac{bk_1 + x_1}{k_1 + x_1} - \frac{\gamma x_1}{k_\gamma + x_1} + \beta, \tag{6.11}$$

$$u_2(X(t), t) = \frac{k_2^2}{k_2^2 + x_1^2} - s_h x_2 \tag{6.12}$$

Since in the experimental cases, trajectory does not behave as predicted or the captured dynamics is not accurate. Therefore, we would like to include some randomness in the system to explain the disturbance.

And, the dynamics for the continuous-time stochastics process can be modelled as following:

$$dX(t) = u(X(t), t)dt + \sigma dW(t), \quad X(t_0) = X_0, \tag{6.13}$$

where $dW(t) = [dW_1(t), dW_2(t)]^t$ shows random infinitesimal increments of two stochastically independent normalized Wiener process. We assume that the process is occurring in a bounded convex domain and boundaries of the domain are Lipschitz and hence $X(t) \in \Omega \subset \mathbb{R}^2$. We also assume that $\partial\Omega$ works as a reflecting barrier. We, now write the Fokker-Planck equation corresponding to equation (6.13) that evolves the probability density function (PDF) of the process, we have

$$\begin{aligned} &\partial_t f(x,t) - \frac{\sigma^2}{2} \sum_{i=1}^{2} \partial^2_{x_i x_i} f(x,t) + \sum_{i=1}^{2} \partial_{x_i} (u_i(x,t) f(x,t)) = 0 \\ &f(x,0) = f_0(x) \end{aligned} \tag{6.14}$$

where $f = f(x,t)$ is the PDF of the individual to be in x at time t. The distribution of the initial position X_0 of the process is represented by $f_0(x)$ and it is represented by the initial PDF distribution. It also satisfies the following conditions:

$$f_0 \geqslant 0, \quad \int_\Omega f_0(x) dx = 1. \tag{6.15}$$

Domain of the FP problem is $Q = \Omega \times (0, T)$ and we assume zero flux boundary conditions for the above FP equation. Also, notice that (6.14) can be written in flux form as follows:

$$\partial_t f(x,t) = \nabla \cdot F, \quad f(x,0) = f_0(x) \tag{6.16}$$

where the flux F is given component-wise by

$$F_j(x,t;f) = \frac{\sigma^2}{2}\partial_{x_j} f - u_j(x,t)f. \tag{6.17}$$

and '$\nabla\cdot$' denotes the divergence operator. Flux zero condition can be given as:

$$F.\hat{n} = 0 \text{ on } \partial\Omega \times (0,T), \tag{6.18}$$

where $\hat{n}$ is the unit outward normal on $\partial\Omega$.

6.4 Discretization of the Fokker Planck equation

In this section, we discuss the we solve the forward FP equation using the second-order accurate Chang-Cooper (CC) scheme. For the temporal discretization, we use the NSFD method. We focus on the 2D case and consider a square domain $\Omega \equiv (-a,a) \times (-a,a)$ and a sequence of uniform grids $\{\Omega_h\}_{h>0}$ which is given by

$$\Omega_h = \{(x_{1i}, x_{2j}) \in \mathbb{R}^2 : (x_{1i}, x_{2j}) = (x_{10} + ih, x_{20} + jh), (i,j) \in \{0, \ldots, N_x\}^2\} \cap \Omega$$

where N_x represents the number of grid points in each direction and h is the mesh size. Further, h is chosen in such a way that the boundaries of Ω coincide with the grid points. Let $\delta t = T/N_t$ be the time step size and N_t denotes the number of time steps. Define

$$Q_{h,\delta t} = \{(x_i, y_j, t_m) : (x_i, y_j) \in \Omega_h, t_m = m\delta t, 0 \leqslant m \leqslant N_t\}$$

On the grid $Q_{h,\delta t}$, $f^m_{i,j}$ represents the value of the grid function in Ω_h at (x_i, y_j) and time t_m.

Using the CC scheme, the term $\nabla.F$ in Equation () at time t_m can be discretized as follows

$$\nabla.F = \frac{1}{h}\left\{\left(F^m_{i+\frac{1}{2},j} - F^m_{i-\frac{1}{2},j}\right) + \left(F^m_{i,j+\frac{1}{2}} - F^m_{i,j-\frac{1}{2}}\right)\right\}$$

where $F^m_{i+\frac{1}{2},j}$ and $F^m_{i,j+\frac{1}{2}}$ represents the flux in the ith and jth direction, respectively, at the point (x_{1i}, x_{2j}). These fluxes are as follows:

$$F^m_{i+\frac{1}{2},j} = \left[-(1-\delta^m_{i+\frac{1}{2},j})u^{m,1}_{i+\frac{1}{2},j} + \frac{\sigma^2}{2h}\right] f^m_{i+1,j} - \left[\frac{\sigma^2}{2h} + \delta^m_{i+\frac{1}{2},j}u^{m,1}_{i+\frac{1}{2},j}\right] f^m_{i,j}$$

$$F^m_{i-1\frac{1}{2},j} = \left[-(1-\delta^m_{i-\frac{1}{2},j})u^{m,1}_{i-\frac{1}{2},j} + \frac{\sigma^2}{2h}\right] f^m_{i,j} - \left[\frac{\sigma^2}{2h} + \delta^m_{i-\frac{1}{2},j}u^{m,1}_{i-\frac{1}{2},j}\right] f^m_{i-1,j}$$

and

$$F^m_{i,j+\frac{1}{2}} = \left[-(1-\delta^m_{i,j+\frac{1}{2}})u^{m,2}_{i,j+\frac{1}{2}} + \frac{\sigma^2}{2h}\right] f^m_{i,j+1} - \left[\frac{\sigma^2}{2h} + \delta^m_{i,j+\frac{1}{2}}u^{m,2}_{i,j+\frac{1}{2}}\right] f^m_{i,j}$$

$$F^m_{i,j-\frac{1}{2}} = \left[-(1-\delta^m_{i,j-\frac{1}{2}})u^{m,2}_{i,j-\frac{1}{2}} + \frac{\sigma^2}{2h}\right] f^m_{i,j} - \left[\frac{\sigma^2}{2h} + \delta^m_{i,j+\frac{1}{2}}u^{m,2}_{i,j-\frac{1}{2}}\right] f^m_{i,j-1}$$

where

$$u^{m,1}_{i+\frac{1}{2},j} = -u_1\left(x_{i+\frac{1}{2}}, y_j, t_m\right)$$

and

$$u^{m,2}_{i,j+\frac{1}{2}} = -u_2\left(x_i, y_{j+\frac{1}{2}}, t_m\right)$$

and

$$\delta^m_{i+\frac{1}{2},j} = \frac{1}{w^m_{i+\frac{1}{2},j}} - \frac{1}{\exp\left(w^m_{i+\frac{1}{2},j}\right) - 1}, \quad w^m_{i+\frac{1}{2},j} = -2hu^{m,1}_{i+\frac{1}{2},j}/\sigma^2, \tag{6.19}$$

$$\delta^m_{i,j+\frac{1}{2}} = \frac{1}{w^m_{i,j+\frac{1}{2}}} - \frac{1}{\exp\left(w^m_{i,j+\frac{1}{2}}\right) - 1}, \quad w^m_{i,j+\frac{1}{2}} = -2hu^{m,2}_{i,j+\frac{1}{2}}/\sigma^2. \tag{6.20}$$

6.5 The Chang-Cooper scheme with first-order NSFD time differencing

The NSFD-CC scheme can be written as follows:

$$\frac{f_{i,j}^{m+1} - f_{i,j}^{m}}{\phi(\delta t)} = \frac{1}{h}\left(F_{i+\frac{1}{2},j}^{m} - F_{i-\frac{1}{2},j}^{m}\right) + \frac{1}{h}\left(F_{i,j+\frac{1}{2}}^{m} - F_{i,j-\frac{1}{2}}^{m}\right) \tag{6.21}$$

for all $(i, j) \in \{1, \ldots, N_x - 1\}$.

The flux zero boundary condition at the discrete level is given by

$$F(i, N_x - 1/2, t_m) = 0, F(i, 1/2, t_m) = 0 \quad \forall i = 0, \ldots, N_x, \tag{6.22}$$

$$F(N_x - 1/2, j, t_m) = 0, F(1/2, j, t_m) = 0 \quad \forall j = 0, \ldots, N_x, \tag{6.23}$$

for all $m = 0, 1, 2, \ldots, N_t$.

Lemma 6.5.1. *The NSFD-CC scheme* (6.21)-(6.22) *is conservative .*

Proof. To see this, writing NSFD-CC scheme as follows:

$$\frac{f_{i,j}^{m+1} - f_{i,j}^{m}}{\phi(\delta t)} = \frac{1}{h}\left(F_{i+\frac{1}{2},j}^{m} - F_{i-\frac{1}{2},j}^{m}\right) + \frac{1}{h}\left(F_{i,j+\frac{1}{2}}^{m} - F_{i,j-\frac{1}{2}}^{m}\right) \tag{6.24}$$

Now, summing over all i, j, we get

$$\sum_{i,j} \frac{f_{i,j}^{m+1} - f_{i,j}^{m}}{\phi(\delta t)} = \sum_{i,j} \left[\frac{1}{h}\left(F_{i+\frac{1}{2},j}^{m} - F_{i-\frac{1}{2},j}^{m}\right) + \frac{1}{h}\left(F_{i,j+\frac{1}{2}}^{m} - F_{i,j-\frac{1}{2}}^{m}\right)\right] \tag{6.25}$$

We observe, the right hand side of (6.25) is a telescoping series and this summation yields following:

$$\sum_{i,j} \frac{f_{i,j}^{m+1} - f_{i,j}^{m}}{\phi(\delta t)} = \sum_{i,j} \left[\frac{1}{h}\left(F_{i+\frac{1}{2},j}^{m} - F_{i-\frac{1}{2},j}^{m}\right) + \frac{1}{h}\left(F_{i,j+\frac{1}{2}}^{m} - F_{i,j-\frac{1}{2}}^{m}\right)\right] \tag{6.26}$$

$$= 0 \quad (\text{using}(6.22)) \tag{6.27}$$

This provides

$$\sum f_{i,j}^{m+1} = \sum_{i,j} f_{i,j}^{m}, \quad \forall m = 0, \ldots, N_t - 1, \tag{6.28}$$

which shows that NSFD-CC scheme is conservative. □

Further, to investigate the positivity and error estimate properties for (6.21), we define the following

$$\alpha_{i,j}^{m} = \frac{\sigma^2}{2h} + \delta_{i+\frac{1}{2},j}^{m} u_{i+\frac{1}{2},j}^{1,m} = -\frac{u_{i+\frac{1}{2},j}^{1,m}}{\bar{w}_{i+\frac{1}{2},j}^{m} - 1}, \quad 1 \leqslant i, j \leqslant N_x - 1, \tag{6.29}$$

$$\beta_{i,j}^{m} = \frac{\sigma^2}{2h} + \delta_{i,j+\frac{1}{2}}^{m} u_{i,j+\frac{1}{2}}^{2,m} = -\frac{u_{i,j+\frac{1}{2}}^{2,m}}{\bar{w}_{i,j+\frac{1}{2}}^{m} - 1}, \quad 1 \leqslant i, j \leqslant N_x - 1, \tag{6.30}$$

$$\alpha_{0,j}^{m} = 0, \quad 1 \leqslant j \leqslant N_x - 1 \tag{6.31}$$

$$\beta_{i,0}^{m} = 0, \quad 1 \leqslant i \leqslant N_x - 1, \tag{6.32}$$

where $\delta_{i+\frac{1}{2},j}^{m}$, $\delta_{i,j+\frac{1}{2}}^{m}$ are defined in (6.19) and (6.20) and $\bar{w}_{i+\frac{1}{2},j}^{m} = \exp(w_{i+\frac{1}{2},j}^{m})$, $\bar{w}_{i,j+\frac{1}{2}}^{m} = \exp(w_{i,j+\frac{1}{2}}^{m})$. We remark that $\alpha_{i,j}^{m}$, $\beta_{i,j}^{m}$ are positive.

Now to investigate stability and consistency of the NSFD-CC scheme with first-order time difference, we define the following:

$$\bar{D}_x f_{i,j}^{m} = D_x^{+} C_{i-\frac{1}{2},j}^{m} D_x^{-} f_{i,j}^{m} + D_x^{+} B_{i-\frac{1}{2},j}^{m} M_{\delta,x} f_{i,j}^{m} \tag{6.33}$$

$$\bar{D}_y f_{i,j}^{m} = D_y^{+} C_{i,j-\frac{1}{2}}^{m} D_y^{-} f_{i,j}^{m} + D_y^{+} B_{i,j-\frac{1}{2}}^{m} M_{\delta,y} f_{i,j}^{m} \tag{6.34}$$

where,

$$D_x^{+} f_{i,j} = \frac{f_{i+1,j} - f_{i,j}}{h}, \tag{6.35}$$

$$D_x^{-} f_{i,j} = \frac{f_{i,j} - f_{i-1,j}}{h}, \tag{6.36}$$

$$D_y^{+} f_{i,j} = \frac{f_{i,j+1} - f_{i,j}}{h}, \tag{6.37}$$

$$D_y^{-} f_{i,j} = \frac{f_{i,j} - f_{i,j-1}}{h}, \tag{6.38}$$

$$M_{\delta,x} = (1 - \delta_{i-\frac{1}{2}}, j) f_{i,j} + \delta_{i-\frac{1}{2},j} f_{i,j} \tag{6.39}$$

$$M_{\delta,y} = (1 - \delta_{i,j-\frac{1}{2}}) f_{i,j} + \delta_{i,j-\frac{1}{2}} f_{i,j} \tag{6.40}$$

$$\tag{6.41}$$

Writing the NSFD-CC scheme using the above setup,

$$\frac{f_{i,j}^{m+1}-f_{i,j}^{m}}{\varphi(\delta t)} = D_x^+ C_{i-\frac{1}{2},j}^m D_x^- f_{i,j}^m + D_x^+ B_{i-\frac{1}{2},j}^m M_i^\delta f_{i,j}^m + g_{i,j}^m \tag{6.42}$$

where

$$\begin{aligned}
D_x^+ C_{i-\frac{1}{2},j}^m D_x^- f_{i,j}^m &= D_x^+ C_{i-\frac{1}{2},j}^m \left(\frac{f_{i,j}^m - f_{i,j-1}^m}{h}\right)\\
&= \frac{1}{h}\left\{C_{i+\frac{1}{2},j}^m \left(\frac{f_{i+1,j}^m - f_{i,j}^m}{h}\right) - C_{i-\frac{1}{2},j}^m \left(\frac{f_{i,j}^m - f_{i-1,j}^m}{h}\right)\right\}\\
&= \frac{1}{h}\left\{\frac{1}{h}C_{i+\frac{1}{2},j}^m f_{i+1,j}^m - \frac{1}{h}\left(C_{i+\frac{1}{2},j}^m + C_{i-\frac{1}{2},j}^m\right) f_{i,j}^m + \frac{1}{h}C_{i-\frac{1}{2},j}^m f_{i-1,j}^m\right\},\\
D_x^+ B_{i-\frac{1}{2},j}^m M_{\delta,x} f_{i+1,j}^m &= D_x^+ \left((1-\delta_{i-\frac{1}{2},j}^m) B_{i-\frac{1}{2},j}^m f_{i,j}^m + \delta_{i-\frac{1}{2},j}^m B_{i-\frac{1}{2},j}^m f_{i-1,j}^m\right)\\
&= \frac{1}{h}\left\{(1-\delta_{i,j}^m) B_{i+\frac{1}{2},j}^m f_{i+1,j}^m - (1-\delta_{i-1,j}^m) B_{i-\frac{1}{2},j}^m f_{i,j}^m\right\}\\
&+ \frac{1}{h}\left\{\delta_{i,j}^m B_{i+\frac{1}{2},j}^m f_{i,j}^m - \delta_{i-1,j}^m B_{i-\frac{1}{2},j}^m f_{i-1,j}^m\right\}
\end{aligned}$$

Now, we introduce the following discrete L^1 norm, $\|f\|_1 = \sum_{i,j} h^2 |f_{i,j}|$ and the discrete L^2 norm is given by $\|f\| = \sqrt{\sum_{i,j} h^2 |f_{i,j}|^2}$.

Theorem 6.5.2. *Let $\varphi : \mathbb{R}^+ \to \mathbb{R}^+$ satisfy the following conditions:*

(I) $\varphi(\delta t) = dt + \mathcal{O}(\delta t^2)$,

(II) $0 < \varphi(\delta t) < \frac{2}{2\gamma - 1}$, where γ is the Lipschitz constant.

then the NSFD-CC (6.21) *is satisfies the following inequality:*

$$\|f^m\| \leqslant 3^{m/2}\|f^0\| + C\delta t \sum_{k=0}^{m} \max(\|g^k\|, \|g^{k+1}\|), \tag{6.43}$$

where C is a constant.

Proof. We take discrete L^2 inner product of (6.21) with f^m, we have

$$\left(\frac{f^{m+1}-f^m}{\varphi(\delta t)}, f^m\right) = \left(D_x^+ C_{\frac{1}{2}}^m D_x^- f^m, f^m\right) + \left(D_x^+ B_{\frac{1}{2}}^m M_{\delta,x} f^m, f^m\right)$$
$$+ \left(D_y^+ C_{\frac{1}{2}}^m D_y^- f^m, f^m\right) + \left(D_y^+ B_{\frac{1}{2}}^m M_{\delta,y} f^m, f^m\right) + (g^m, f^m)$$

Similar to the computation in [128], we get

$$\begin{aligned}\left(D_x^+ C_{\frac{1}{2}}^m D_x^- f^m, f^m\right) &= \sum_{i=0}^{N} (D_x^+ C_{i-\frac{1}{2},j}^m D_x^- f^m) f_{i,j}^m h \\ &\leqslant C_{N+\frac{1}{2},j}^m \left(\frac{f_{N+1,j}^m - f_{N,j}^m}{h}\right) f_{N,j}^m - C_{-\frac{1}{2},j}^m \left(\frac{f_{0,j}^m - f_{-1,j}^m}{h}\right) f_{i-1,j}^m\end{aligned}$$

Next, we have

$$\begin{aligned}\left(D_x^+ B_{\frac{1}{2}}^m M_{\delta,x} f^m, f^m\right) &= \sum_{i=0}^{N} ((1-\delta_{i,j}^m) B_{i+\frac{1}{2},j}^m f_{i+1,j}^m f_{i,j}^m - (1-\delta_{i-1,j}^m) B_{i-1,j}^m (f_{i,j}^m)^2 \\ &\quad + \delta_{i,j}^m (f_{i,j}^m)^2 - \delta_{i-1,j}^m B_{i-\frac{1}{2},j}^m f_{i-1,j}^m f_{i,j}^m) \\ &= \sum_{i=0}^{N} ((1-\delta_{i,j}^m) B_{i+\frac{1}{2},j}^m f_{i+1,j}^m f_{i,j}^m - \sum_{i=-1}^{N-1} \delta_{i,j}^m B_{i+\frac{1}{2},j}^m f_{i,j}^m f_{i+1,j}^m \\ &\quad + \sum_{i=0}^{N} \delta_{i,j}^m B_{i+\frac{1}{2}}^m (f_{i,j}^m)^2 + \sum_{i=-1}^{N-1} (\delta_j^m - 1) B_{i+\frac{1}{2},j}^m (f_{i+1,j}^m)^2 \\ &= \sum_{i=0}^{N} (1-\delta_{i,j}^m) B_{i+\frac{1}{2},j}^m f_{i+1,j}^m f_{i,j}^m - \sum_{i=-1}^{N-1} \delta_{i,j}^m B_{i+\frac{1}{2},j}^m f_{i,j}^m f_{i+1,j}^m \\ &\quad + (1-\delta_{N,j}) B_{N+\frac{1}{2},j}^m f_{N+1,j}^m - \delta_{-1}^m B_{-\frac{1}{2},j}^m f_{-1,j}^m f_{0,j}^m \\ &\quad + \sum_{i=0}^{N-1} \delta_{i,j}^m B_{i+\frac{1}{2},j}^m (f_{i,j}^m)^2 + \sum_{i=0}^{N-1} (\delta_{i,j}^m - 1) B_{i+\frac{1}{2},j}^m (f_{i+1,j}^m)^2 \\ &\quad + \delta_{N,j}^m B_{N=\frac{1}{2}}^m (f_N^m, j)^2 + (\delta_{-1}^m - 1) B_{-\frac{1}{2}}^m (f_{0,j}^m)^2\end{aligned}$$

Applying the zero-flux boundary conditions $F_{-\frac{1}{2},0}^m = 0$ and $F_{N+\frac{1}{2},j}^m = 0$ for all $j = 0, \ldots, N$ which is given by

$$B_{-\frac{1}{2},j}^m ((1-\delta_{-1}^m) f_{0,j}^m + \delta_{-1,j}^m f_{-1,j}^m) + C_{-\frac{1}{2},j}^m \left(\frac{f_{0,j}^m - f_{-1,j}^m}{h}\right) = 0$$

and

$$B_{N+\frac{1}{2},j}^m ((1-\delta_N^m) f_{N+1,j}^m + \delta_{N,j}^m f_{N,j}^m) + C_{N+\frac{1}{2},j}^m \left(\frac{f_{N+1,j}^m - f_{N,j}^m}{h}\right) = 0$$

which yields

$$D_x^+ C_{i-\frac{1}{2},j}^m D_x^- f_{i,j}^m + \left(D_x^+ B_{\frac{1}{2}}^m M_{\delta,x} f^m, f^m\right)$$

$$
\begin{aligned}
&\leqslant \sum_{i=0}^{N-1} \left[(f_{i,j}^m)^2 + (f_{i+1,j}^m)^2\right] B_{i+1,j}^m \left(\frac{1-2\delta_{i,j}^m}{2}\right) \\
&\quad + \sum_{i=0}^{N-1} \delta_{i,j}^m B_{i+\frac{1}{2},j}^m (f_{i,j}^m)^2 + \sum_{i=0}^{N-1} (\delta_{i,j}^m - 1) B_{i+\frac{1}{2}}^m (f_{i+1,j}^m)^2 \\
&\leqslant \sum_{i=0}^{N-1} \frac{1}{2} B_{i+\frac{1}{2},j}^m (f_{i,j}^m)^2 - \sum_{i=0}^{N-1} \frac{1}{2} B_{i+\frac{1}{2},j}^m (f_{i+1,j}^m)^2 \\
&\leqslant \sum_{i=0}^{N-1} \frac{1}{2} B_{i+\frac{1}{2},j}^m (f_{i,j}^m)^2 - \sum_{i=0}^{N} \frac{1}{2} B_{i-\frac{1}{2},j}^m (f_{i,j}^m)^2 \\
&\quad + \frac{1}{2} B_{N+\frac{1}{2},j}^m (f_{N,j}^m)^2 - \frac{1}{2} B_{-\frac{1}{2},j}^m (f_{0,j}^m)^2 && (6.44) \\
&= \sum_{i=0}^{N} \frac{1}{2} B_{i+\frac{1}{2},j}^m (f_{i,j}^m)^2 - \sum_{i=0}^{N} \frac{1}{2} B_{i-\frac{1}{2},j}^m (f_{i,j}^m)^2 && (6.45) \\
&\leqslant \sum_{i=0}^{N} \frac{1}{2} |B_{i+\frac{1}{2},j}^m - B_{i-\frac{1}{2},j}^m| |f_{i,j}^m|^2 && (6.46) \\
&\leqslant \frac{1}{2} \gamma \sum_{i=0}^{N} |f_{i,j}^m|^2 h && (6.47) \\
&= \frac{1}{2} \gamma \|f^m\|^2 && (6.48)
\end{aligned}
$$

Therefore, we obtain the following estimate

$$\left(\frac{f^{m+1} - f^m}{\varphi(\delta t)}, f^m\right) \leqslant \frac{1}{2}\gamma \|f^m\|^2 + \|g^m\| \|f^m\| \tag{6.49}$$

And from the calculation in [128], we have

$$\left(\frac{f^{m+1} - f^m}{\varphi(\delta t)}, f^{m+1}\right) \leqslant \frac{1}{2}\gamma \|f^{m+1}\|^2 + \|g^{m+1}\| \|f^{m+1}\| \tag{6.50}$$

Adding above two equations, we get

$$\left(\frac{f^{m+1} - f^m}{\varphi(\delta t)}, f^{m+1} + f^m\right) \leqslant \frac{1}{2}\gamma \|f^m\|^2 + \frac{1}{2}\gamma \|f^{m+1}\|^2 + \|g^m\| \|f^m\| + \|g^{m+1}\| \|f^{m+1}\| \tag{6.51}$$

On the other hand, we have

$$\left(\frac{\|f^{m+1}\|^2 - \|f^m\|^2}{\varphi(\delta t)}\right) = \left(\frac{f^{m+1} - f^m}{\varphi(\delta t)}, f^{m+1} + f^m\right) \tag{6.52}$$

Thus, we get

$$\left(\frac{\|f^{m+1}\|^2 - \|f^m\|^2}{\varphi(\delta t)}\right) \leqslant \frac{1}{2}\gamma\|f^m\|^2 + \frac{1}{2}\gamma\|f^{m+1}\|^2 + \|g^m\|\|f^m\| + \|g^{m+1}\|\|f^{m+1}\| \tag{6.53}$$

which gives

$$\begin{aligned}
\|f^{m+1}\|^2 &\leqslant \|f^m\|^2 + \frac{1}{2}\gamma\varphi(\delta t)\|f^m\|^2 + \frac{1}{2}\gamma\varphi(\delta t)\|f^{m+1}\|^2 + \varphi(\delta t)\|g^m\|\|f^m\| + \varphi(\delta t)\|g^{m+1}\|\|f^{m+1}\| \\
\|f^{m+1}\|^2 &\leqslant \|f^m\|^2 + \frac{1}{2}\gamma\varphi(\delta t)^2\|f^m\|^2 + \frac{1}{2}\gamma\varphi(\delta t)\|f^{m+1}\|^2 + \frac{1}{2}(\varphi(\delta t)^2\|g^m\|^2 + \|f^m\|^2) \\
&\quad + \frac{1}{2}(\varphi(\delta t)^2\|g^{m+1}\|^2 + \|f^{m+1}\|^2) \\
\Big(1 - \frac{1}{2}&\varphi(\delta t)(\gamma - 1)\Big)\|f^{m+1}\|^2 \\
&\leqslant \left(1 + \frac{1}{2}\varphi(\delta t)(\gamma + 1)\right)\|f^m\|^2 + \frac{1}{2}\varphi(\delta t)^2(\|g^m\|^2 + \|g^{m+1}\|^2) \\
&\leqslant \frac{\left(1 + \frac{1}{2}\varphi(\delta t)(\gamma + 1)\right)}{\left(1 - \frac{1}{2}\varphi(\delta t)(\gamma - 1)\right)}\|f^m\|^2 + \frac{1}{2}\varphi(\delta t)^2(\|g^m\|^2 + \|g^{m+1}\|^2) \\
\|f^{m+1}\|^2 &\leqslant \frac{\left(1 + \frac{1}{2}\varphi(dt)(\gamma + 1)\right)}{\left(1 - \frac{1}{2}\varphi(\delta t)(\gamma - 1)\right)}\|f^m\|^2 + \frac{1}{2}\varphi(\delta t)^2(\|g^m\|^2 + \|g^{m+1}\|^2 + 2\|g^m\|\|g^{m+1}\|) \\
&= \frac{\left(1 + \frac{1}{2}\varphi(\delta t)(\gamma + 1)\right)}{\left(1 - \frac{1}{2}\varphi(\delta t)(\gamma - 1)\right)}\|f^m\|^2 + \frac{1}{2}\varphi(\delta t)^2(\|g^m\| + \|g^{m+1}\|)^2 \\
&= \frac{\left(1 + \frac{1}{2}\varphi(\delta t)(\gamma + 1)\right)}{\left(1 - \frac{1}{2}\varphi(\delta t)(\gamma - 1)\right)}\|f^m\|^2 + \frac{1}{2}\varphi(\delta t)^2(\|g^m\| + \|g^{m+1}\|)^2 \\
&\quad + 2\varphi(\delta t)\left\{\frac{\left(1 + \frac{1}{2}\varphi(\delta t)(\gamma + 1)\right)}{2\left(1 - \frac{1}{2}\varphi(\delta t)(\gamma - 1)\right)}\right\}^{\frac{1}{2}}\|f^m\|(\|g^m\| + \|g^{m+1}\|) \\
&\leqslant \left\{\frac{\left(1 + \frac{1}{2}\varphi(\delta t)(\gamma + 1)\right)}{\left(1 - \frac{1}{2}\varphi(\delta t)(\gamma - 1)\right)}\right\}^{\frac{1}{2}}\|f^m\| + \sqrt{2}\varphi(\delta t)\max(\|g^m\|, \|g^{m+1}\|) \\
&\leqslant \left\{\frac{\left(1 + \frac{1}{2}\varphi(\delta t)(\gamma + 1)\right)}{\left(1 - \frac{1}{2}\varphi(\delta t)(\gamma - 1)\right)}\right\}^{\frac{1}{2}}\|f^m\| + \sqrt{2}\delta t(1 + \mathcal{O}(\delta t))\max(\|g^m\|, \|g^{m+1}\|)
\end{aligned}$$

We want to emphasize here that

$$\frac{\big(1+\frac{1}{2}\varphi(\delta t)(\gamma+1)\big)}{\Big(1-\frac{1}{2}\varphi(\delta t)(\gamma-1)\Big)} > 0.$$

and in fact it is greater that 1, hence the term inside the bracket makes sense. It can be seen as follows

$$\Big(1+\frac{1}{2}\varphi(\delta t)(\gamma+1)\Big) > \Big(1-\frac{1}{2}\varphi(\delta t)(\gamma-1)\Big),$$

which gives $\varphi(\delta t)\gamma > 0$, and this is true since φ and $\gamma > 0$. Further , under the assumption $\varphi(\delta t) < \frac{2}{2\gamma-1}$, we get

$$\frac{\big(1+\frac{1}{2}\varphi(\delta t)(\gamma+1)\big)}{\Big(1-\frac{1}{2}\varphi(\delta t)(\gamma-1)\Big)} < 3$$

This gives,

$$\|f^{m+1}\| \leqslant \sqrt{3}\|f^m\| + C\max(\|g^m\|, \|g^{m+1}\|) \tag{6.54}$$

This recursion relation gives us,

$$\|f^m\| \leqslant 3^{m/2}\|f^0\| + C\delta t\sum_{k=0}^{m}\max(\|g^k\|, \|g^{k+1}\|). \tag{6.55}$$

where $C = \sqrt{2}\sum_{k=0}^{m}(\sqrt{3})^{\frac{m+k}{2}}$. □

To explore the order of accuracy of the NSFD-CC scheme, we assume $f \in C^2([0,T], C^3(\Omega))$ and compute the truncation error.

Lemma 6.5.3. *The truncation error of the discretization scheme* (6.21) *is of order* $\mathcal{O}(\delta t + h^2)$.

We write the Taylor series expansion of f in t

$$\begin{aligned} f(x_i, y_j, t_{m+1}) &= f(x_i, y_j, t_m) + \delta t\frac{\partial f}{\partial t}(x_i, y_j, t_m) + \frac{\delta t^2}{2}\frac{\partial^2 f}{\partial t^2}(x_i, y_j, \mu_k) \\ &= f(x_i, y_j, t_m) + \delta t\nabla\cdot F(x_i, y_j, t_m) + \frac{\delta t^2}{2}\frac{\partial^2 f}{\partial t^2}(x_i, y_j, \mu_k) \end{aligned} \tag{6.56}$$

since $\frac{\partial f}{\partial t}(x_i, y_j, t_m) = \nabla \cdot F(x_i, y_j, t_m)$.

Now consider the following,

$$\begin{aligned}
&f(x_i, y_j, t_{m+1}) - \{f(x_i, y_j, t_m) + \varphi(\delta t)\nabla \cdot F(x_i, y_j, t_m)\} \\
&= f(x_i, y_j, t_m) + \delta t \frac{\partial f}{\partial t}(x_i, y_j, t_m) + \frac{\delta t^2}{2}\frac{\partial^2 f}{\partial t^2}(x_i, y_j, \mu_k) \\
&- \{f(x_i, y_j, t_m) + \varphi(\delta t)\nabla \cdot F(x_i, y_j, t_m)\} \\
&= (\delta t - \varphi(\delta t))\nabla \cdot F(x_i, y_j, t_m) + \frac{\delta t^2}{2}\frac{\partial^2 f}{\partial t^2}(x_i, y_j, \mu_k) \\
&= \mathcal{O}(\delta t^2) \qquad (6.57)
\end{aligned}$$

as $\varphi(\delta t) = \delta t + \mathcal{O}(\delta t^2)$

Further, doing the similar calculations as in [128], we get

$$\begin{aligned}
D_x^+ C^m_{i-\frac{1}{2},j} D_x^- f^m_{i,j} - \partial_x (C^m \partial_x f)^{t_m}_{x_i,y_j} &= \mathcal{O}(h^4) \\
D_y^+ C^m_{i-\frac{1}{2},j} D_y^- f^m_{i,j} - \partial_x (C^m \partial_y f)^{t_m}_{x_i,y_j} &= \mathcal{O}(h^4) \\
D_x^+ B^m_{i-\frac{1}{2},j} M_{\delta,x} f^m_{i,j} - \partial_x (B f)^{t_m}_{x_i,y_j} &= \mathcal{O}(h^2) \\
D_y^+ B^m_{i,j-\frac{1}{2}} M_{\delta,y} f^m_{i,j} - \partial_y (B f)^{t_m}_{x_i,y_j} &= \mathcal{O}(h^2) \qquad (6.58)
\end{aligned}$$

and hence,

$$\begin{aligned}
\bar{D}_x f^m_{i,j} &= D_x^+ C^m_{i-\frac{1}{2},j} D_x^- f^m_{i,j} + D_x^+ B^m_{i-\frac{1}{2},j} M_{\delta,x} f^m_{i,j} \\
&= \mathcal{O}(h^2) \qquad (6.59) \\
\bar{D}_y f^m_{i,j} &= D_y^+ C^m_{i,j-\frac{1}{2}} D_y^- f^m_{i,j} + D_y^+ B^m_{i,j-\frac{1}{2}} M_{\delta,y} f^m_{i,j} \\
&= \mathcal{O}(h^2) \qquad (6.60)
\end{aligned}$$

Using above calculations, the truncation error can be given as follows:

$$\begin{aligned}
\tau^{m+1}_{i,,j} &= \frac{f(x_i, y_j, t_{m+1}) - f(x_i, y_j, t_m)}{\varphi(\delta t)} + \bar{D}_x f(x_i, y_j, t_m) + \bar{D}_x f(x_i, y_j, t_m) \\
&= \mathcal{O}(\delta t + h^2) \qquad (6.61)
\end{aligned}$$

Next, we define the global error as follows:

$$e_{i,j}^m = f(x_i, y_j, t_m) - f_{i,j}^m, \quad i,j = 0, \dots, N_x,\ m = 1, \dots, N_t. \tag{6.62}$$

Theorem 6.5.4. *The NSFD-CC scheme* (6.21) *converges with an error of order* $\mathcal{O}(dt + h^2)$ *under the CFL condition in the discrete* L^1 *norm.*

Proof. By the definition of truncation error, we have

$$\begin{aligned}
&\frac{e_{i,j}^{m+1} - e_{i,j}^m}{\varphi(\delta t)} - (\bar{D}_x + \bar{D}_y)e_{i,j}^m \\
&= \frac{f(x_i, y_j, t_{m+1}) - f_{i,j}^{m+1} - f(x_i, y_j, t_m) + f_{i,j}^m}{\varphi(\delta t)} \\
&- (\bar{D}_x + \bar{D}_y)(f(x_i, y_j, t_{m+1}) - f_{i,j}^{m+1} - f(x_i, y_j, t_m) + f_{i,j}^m) \\
&= \frac{f(x_i, y_j, t_{m+1}) - f(x_i, y_j, t_m)}{\varphi(\delta t)} - \frac{(f_{i,j}^{m+1} - f_{i,j}^m)}{\varphi(\delta t)} \\
&- (\bar{D}_x + \bar{D}_y)(f(x_i, y_j, t_{m+1}) - f(x_i, y_j, t_m)) + (\bar{D}_x + \bar{D}_y)(f_{i,j}^{m+1} - f_{i,j}^m) \\
&= \tau_{i,j}^{m+1}
\end{aligned}$$

since $\frac{f_{i,j}^{m+1} - f_{i,j}^m}{\varphi(\delta t)} - (\bar{D}_x + \bar{D}_y)(f_{i,j}^{m+1} - f_{i,j}^m) = 0$.

Thus

$$\frac{e_{i,j}^{m+1} - e_{i,j}^m}{\varphi(\delta t)} = (\bar{D}_x + \bar{D}_y)e_{i,j}^m + \tau_{i,j}^{m+1}$$

□

Thus, the solution error $e_{i,j}^m$ satisfies the discretized FP equation discussed above with the right hand side given by the truncation error function. Hence, using the Theorem 6.5.2 , we have

$$\|e^m\| \leqslant 3^{m/2}\|e^0\| + C\delta t \sum_{k=0}^{m} \max(\|\tau^k\|, \|\tau^{k+1}\|),$$

$$
\begin{aligned}
&\leqslant 3^{m/2}\|e^0\| + C\max(\|\tau^k\|, \|\tau^{k+1}\|)M\delta t\\
&= CT\max(\|\tau^k\|, \|\tau^{k+1}\|).
\end{aligned}
$$

Therefore, from Lemma 6.5.3 , the NSFD-CC scheme converges with order $\mathcal{O}(dt + h^2)$ in the discrete L^1 norm.

Next, we will investigate the positivity property of NSFD-CC scheme. Let $f^m = (f^m_{1,1}, \ldots, f^m_{1,N_x-1}, \ldots, f^m_{2,1}, \ldots, f^m_{2,N_x-1}, \ldots, f^m_{N_x-1,1}, \ldots, f^m_{N_x-1,N_x-1})$, then we can rewrite the scheme (6.21) as follows:

$$
f^{m+1} = \mathcal{A}f^m + g^m_{i,j}, \quad m = 0, \ldots, N_t - 1, \tag{6.63}
$$

where $\mathcal{A}$ is a block diagonal matrix and is given by

$$
\begin{pmatrix}
P^{m,1} & R^{m,1} & & & & \\
Q^{m,1} & P^{m,2} & R^{m,2} & & & \\
 & Q^{m,2} & P^{m,3} & R^{m,3} & & \\
 & & \ddots & \ddots & \ddots & \\
 & & & & & P^{m,N_x-1}
\end{pmatrix}
\tag{6.64}
$$

where, each block $P^{m,k}$, for $k = 1, 2, \ldots, N_x - 1$ is a tridiagonal matrix and block $Q^{m,k}$ and $R^{m,k}$ are diagonal matrices,

$$
\begin{aligned}
\text{super diagonal entries of } P^{m,k} &= \frac{\delta t}{h}\alpha^m_{i,j}\bar{w}^m_{i+\frac{1}{2},j},\\
\text{sub diagonal entries of } P^{m,k} &= \frac{\delta t}{h}\alpha^m_{i-1,j}\\
\text{diagonal entries of } P^{m,k} &= 1 - \frac{\delta t}{h}(\alpha^m_{i-1,j}\bar{w}^m_{i,j} + \beta^m_{i-1,j}\bar{w}^m_{i,j} + \alpha^m_{i,j} + \beta^m_{i,j})\\
\text{diagonal entries of } Q^{m,k} &= \frac{\delta t}{h}\beta^m_{i-1,j}\\
\text{diagonal entries of } R^{m,k} &= \frac{\delta t}{h}\beta^m_{i,j}\bar{w}^m_{i+\frac{1}{2},j},
\end{aligned}
$$

From (6.29) and (6.30), we see that super diagonal and sub diagonal entries of $P^{m,k}$, diagonal entries of $Q^{m,k}$ and $R^{m,k}$ are positive. Diagonal entries of $P^{m,k}$ are also non-negative under the assumption that $\frac{\delta t}{h}(\alpha^m_{i-1,j}\bar{w}^m_{i,j} + \beta^m_{i-1,j}\bar{w}^m_{i,j} + \alpha^m_{i,j} + \beta^m_{i,j}) < 1$. Since, u is bounded, so let $M = \max\{\alpha^m_{i-1,j}\bar{w}^m_{i,j} + \beta^m_{i-1,j}\bar{w}^m_{i,j} + \alpha^m_{i,j} + \beta^m_{i,j}, 0 \leqslant i, j \leqslant N\}$, then $\delta t < \frac{h}{M}$.

Thus the CFL-like condition is

$$\delta t < \frac{h}{M} \tag{6.65}$$

shows the non-negativity of diagonal entry of $P^{m,k}$. Thus, we see that the NSFD-CC (6.21) is positive.

Theorem 6.5.5. *The NSFD-CC scheme* (6.21)*, is* L^1 *stable under the CFL-like condition,*

$$\|f^m\|_1 = \|f^0\|_1, \quad m = 0, \dots N_t - 1.$$

Proof. From Lemma (6.5.1), we get

$$\sum_{i,j=0}^{n} f^m_{i,j} = \sum_{i,j=0}^{n} f^0_{i,j}, \quad \forall\ m = 1, \dots, N_t, \tag{6.66}$$

Using positivity result, above equation can be written as

$$\sum_{i,j=0}^{n} |f^m_{i,j}| = \sum_{i,j=0}^{n} |f^0_{i,j}|, \quad \forall m = 1, \dots, N_t, \tag{6.67}$$

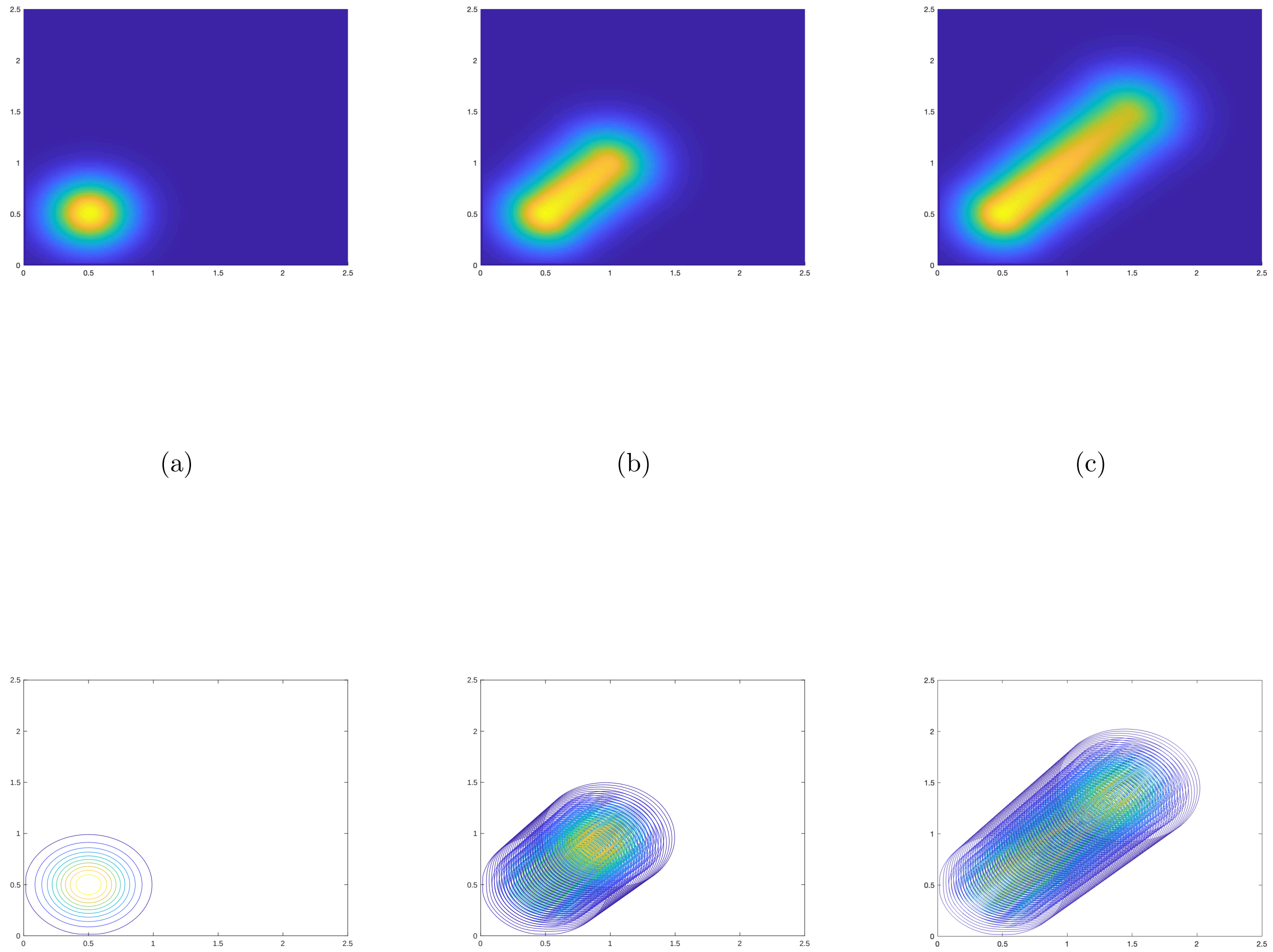
□

6.6 Numerical results

In this section, we present results of numerical experiments to validate our proposed NSFD-CC scheme. The initial PDF $f_0(x)$ is given as follows

$$f_0(x) = \hat{C} \exp(-(x_1 - A_1)^2 - (x_2 - A_2)^2) \tag{6.68}$$

where $(A_1, A_2) = \bar{x}(0)$ is the starting point of the trajectory $\bar{x}$ and $\hat{C}$ is the normalisation constant such that $\int_\Omega f(x)dx = 1$. For test case I, we consider the drift coefficient



(a)

(b)

(c)

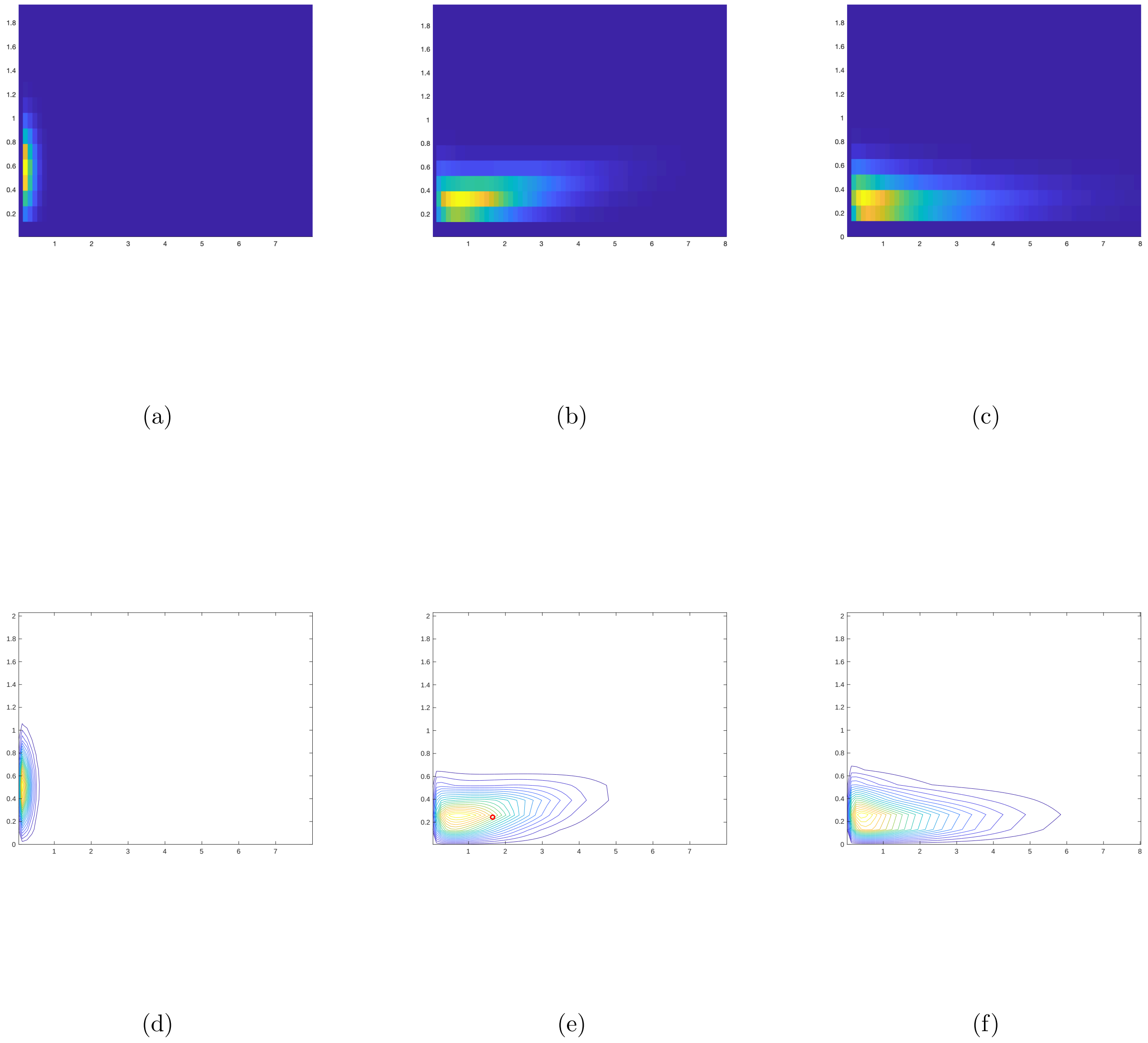
(d)

(e)

(f)

Figure 6.1

$(1, 1)$, hence the expected motion is in the direction of a straight line $x_2 = x_1$, which is evident from figure. The total number of spatial points is $N_x = 150$ and the number of temporal grid points is $N_t = 1000$ and $q = 2$.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 6.2

For the next simulation, we consider the drift term as the right hand side of ode and is given in Equation (6.11).

6.7 Conclusion

A Fokker-Planck framework for calcium signaling is discussed. To compute the approximate solution of FP equation, an NSFD-CC scheme is used to discretize the system. The scheme was shown to be conservative, positivity preserving, stable, and first accurate in time while second order accurate in space. To demonstrate the effectiveness of the numerical scheme, various numerical simulations are also discussed.

CHAPTER 7

Conclusion

The main aim of the thesis was to study nonlinear optimization framework to reconstruct the electrical and optical properties in hybrid imaging modalities, nonstandard finite difference methods and Fokker-Planck framework in esophageal cancer. In the field of hybrid imaging, we developed a PDE-constraint optimization problem to reconstruct the electrical conductivity in current density impedance imaging and optical coefficients in two-photon photoacoustic computed tomography.

We characterized the solution of the optimization problem through an optimality system that was solved with using a proximal scheme, coupled with various regularizer for denoising to remove artifacts in the reconstructions and obtain images which are of high contrast and high resolution. We then demonstrated the effectiveness of our proposed scheme through several numerical experiments and compared our results with an existing scheme. Our scheme facilitated reconstructions of a wide variety of conductivity patters with good contrast and resolution. The developed framework is useful for other kind of hybrid imaging modalities.

In nonstandard finite difference methods, we first discussed two-classes of modified nonstandard theta methods and Runge-Kutta methods. The new numerical methods were developed based on modifications of the nonstandard denominator functions used in NSFD methods. The methods were shown to be of second order accuracy, which is an improvement in the accuracy of their NSFD counterparts, while preserving their stability properties. Using a set of numerical simulations, the two-stage modified nonstandard explicit Runge-Kutta methods were compared to the

NSFD ERK2 method, the standard ERK2 method, and the combined NSFD method, which verified the theoretical results and demonstrated the strengths of the proposed new numerical methods. We would like to further develop NSFD schemes which are higher order accurate. In Chapter 6, we discussed a Fokker-Planck framework for esophageal cancer. To solve FP equation, we presented an NSFD scheme based on explicit Euler method for temporal discretization and Chang-Cooper scheme for spatial resolution. Next, we showed that the NSFD-CC scheme is conservative, positive, and L^1 stable. We further showed that NSFD-CC is of first order accurate in time and second order accurate in spatial discretization. To show the effectiveness of our method, we also discussed various numerical simulations In this direction, future work involves to develop NSFD-CC schemes which are based on ERK2 method.

REFERENCES

[1] G. Bal, Hybrid inverse problems and internal information, Inside Out, Cambridge University Press, Cambridge, UK, G. Uhlmann, Editor (2012).

[2] P. Kuchment, Mathematics of hybrid imaging: A brief review, The mathematical legacy of Leon Ehrenpreis (2012) 183–208.

[3] M. Cheney, D. Isaacson, J. C. Newell, Electrical impedance tomography, SIAM Rev. 41 (1) (1999) 85–101.

[4] N. Zain, K. Chelliah, Breast imaging using electrical impedance tomography: correlation of quantitative assessment with visual interpretation, Asian Pacific Journal of Cancer Prevention 15 (3) (2014) 1327–1331.

[5] M. V. P. Kruger, Tomography as a metrology technique for semiconductor manufacturing, University of California, Berkeley, 2003.

[6] R. C. Waterfall, R. He, C. M. Beck, Visualizing combustion using electrical impedance tomography, Chemical engineering science 52 (13) (1997) 2129–2138.

[7] B. D. Seagar AD, B. BH, Theoretical limits to sensitivity and resolution in impedance imaging, Clinical Physics and Physiological Measurement 8 (Suppl A) (1987) 13–31.

[8] G. Ambartsoumian, R. Gouia-Zarrad, V. P. Krishnan, S. Roy, Image reconstruction from radially incomplete spherical radon data, European Journal of Applied Mathematics 29 (3) (2018) 470–493.

[9] G. Ambartsoumian, R. Gouia-Zarrad, V. P. Krishnan, S. ROY, An efficient numerical algorithm for the inversion of an integral transform arising in ultrasound imaging, Journal of Mathematical Imaging and Vision, 53 (2015) 78—-91.

[10] K. F. Hasanov, A. W. Ma, A. I. Nachman, M. L. Joy, Current density impedance imaging, IEEE Transactions on Medical Imaging 27 (9) (2008) 1301–1309.

[11] A. Nachman, A. Tamasan, A. Timonov, Current density impedance imaging, Tomography and inverse transport theory 559 (2011) 135–149.

[12] D. Garmatter, B. Harrach, Magnetic resonance electrical impedance tomography (mreit): Convergence and reduced basis approach, SIAM Journal on Imaging Sciences 11 (1) (2018) 863–887.

[13] J. Liu, J. K. Seo, M. Sini, E. J. Woo, On the convergence of the harmonic b_z algorithm in magnetic resonance electrical impedance tomography, SIAM Journal on Applied Mathematics 67 (5) (2007) 1259–1282.

[14] J. Liu, J. Seo, E. Woo, A posteriori error estimate and convergence analysis for conductivity image reconstruction in mreit, SIAM Journal on Applied Mathematics 70 (8) (2010) 2883–2903.

[15] G. Scott, M. Joy, R. Armstrong, R. Henkelman, Measurement of nonuniform current density by magnetic resonance, IEEE transactions on medical imaging 10 (3) (1991) 362–374.

[16] J. K. Seo, E. J. Woo, Magnetic resonance electrical impedance tomography (mreit), SIAM review 53 (1) (2011) 40–68.

[17] P. Kuchment, D. Steinhauer, Stabilizing inverse problems by internal data, Inverse Problems 28 (8) (2012) 084007.

[18] S. Kim, O. Kwon, J. K. Seo, J.-R. Yoon, On a nonlinear partial differential equation arising in magnetic resonance electrical impedance tomography, SIAM journal on mathematical analysis 34 (3) (2002) 511–526.

[19] J.-Y. Lee, A reconstruction formula and uniqueness of conductivity in mreit using two internal current distributions, Inverse Problems 20 (3) (2004) 847.

[20] A. Nachman, A. Tamasan, A. Timonov, Conductivity imaging with a single measurement of boundary and interior data, Inverse Problems 23 (6) (2007) 2551.

[21] A. Nachman, A. Tamasan, A. Timonov, Recovering the conductivity from a single measurement of interior data, Inverse Problems 25 (3) (2009) 035014.

[22] A. Tamasan, J. Veras, Conductivity imaging by the method of characteristics in the 1-laplacian, Inverse Problems 28 (8) (2012) 084006.

[23] O. Kwon, E. J. Woo, J.-R. Yoon, J. K. Seo, Magnetic resonance electrical impedance tomography (mreit): simulation study of j-substitution algorithm, IEEE Transactions on Biomedical Engineering 49 (2) (2002) 160–167.

[24] H. S. Khang, B. I. Lee, S. H. Oh, E. J. Woo, S. Y. Lee, M. H. Cho, O. Kwon, J. R. Yoon, J. K. Seo, J-substitution algorithm in magnetic resonance electrical impedance tomography (mreit): phantom experiments for static resistivity images, IEEE transactions on medical imaging 21 (6) (2002) 695–702.

[25] Y. J. Kim, O. Kwon, J. K. Seo, E. J. Woo, Uniqueness and convergence of conductivity image reconstruction in magnetic resonance electrical impedance tomography, Inverse Problems 19 (5) (2003) 1213.

[26] K. Knudsen, M. Lassas, J. L. Mueller, S. Siltanen, Regularized d-bar method for the inverse conductivity problem, Inverse Problems & Imaging 3 (4) (2009) 599.

[27] A. Moradifam, A. Nachman, A. Timonov, A convergent algorithm for the hybrid problem of reconstructing conductivity from minimal interior data, Inverse Problems 28 (8) (2012) 084003.

[28] K. Hoffmann, K. Knudsen, Iterative reconstruction methods for hybrid inverse problems in impedance tomography, Sensing and imaging 15 (1) (2014) 1–27.

[29] H. Ammari, An introduction to mathematics of emerging biomedical imaging, Vol. 62, Springer, 2008.

[30] G. Bal, K. Ren, Multi-source quantitative photoacoustic tomography in a diffusive regime, Inverse Problems 27 (7) (2011) 075003.

[31] G. Bal, G. Uhlmann, Inverse diffusion theory of photoacoustics, Inverse Problems 26 (8) (2010) 085010.

[32] P. Kuchment, L. Kunyansky, Mathematics of thermoacoustic and photoacoustic tomography, preprint (2007).

[33] G. Langer, K.-D. Bouchal, H. Grün, P. Burgholzer, T. Berer, Two-photon absorption-induced photoacoustic imaging of rhodamine b dyed polyethylene spheres using a femtosecond laser, Optics Express 21 (19) (2013) 22410–22422.

[34] L. V. Wang, Ultrasound-mediated biophotonic imaging: a review of acousto-optical tomography and photo-acoustic tomography, Disease markers 19 (2-3) (2004) 123–138.

[35] L. V. Wang, Tutorial on photoacoustic microscopy and computed tomography, IEEE Journal of Selected Topics in Quantum Electronics 14 (1) (2008) 171–179.

[36] Y. Xu, L. V. Wang, G. Ambartsoumian, P. Kuchment, Limited view thermoacoustic tomography, in: Photoacoustic imaging and spectroscopy, CRC Press, 2017, pp. 61–74.

[37] Y. Xu, L. V. Wang, G. Ambartsoumian, P. Kuchment, Limited view thermoacoustic tomography, in: Photoacoustic imaging and spectroscopy, CRC Press, 2017, pp. 61–74.

[38] M. Xu, L. V. Wang, Photoacoustic imaging in biomedicine, Review of scientific instruments 77 (4) (2006) 041101.

[39] J. Yao, L. V. Wang, Photoacoustic tomography: fundamentals, advances and prospects, Contrast media & molecular imaging 6 (5) (2011) 332–345.

[40] L. V. Wang, L. Gao, Photoacoustic microscopy and computed tomography: from bench to bedside, Annual review of biomedical engineering 16 (2014) 155.

[41] J. Xia, J. Yao, L. V. Wang, Photoacoustic tomography: principles and advances, Electromagnetic waves (Cambridge, Mass.) 147 (2014) 1.

[42] Y.-H. Lai, S.-Y. Lee, C.-F. Chang, Y.-H. Cheng, C.-K. Sun, Nonlinear photoacoustic microscopy via a loss modulation technique: from detection to imaging, Optics Express 22 (1) (2014) 525–536.

[43] G. Langer, K.-D. Bouchal, H. Grün, P. Burgholzer, T. Berer, Two-photon absorption-induced photoacoustic imaging of rhodamine b dyed polyethylene spheres using a femtosecond laser, Optics Express 21 (19) (2013) 22410–22422.

[44] P. Bardsley, K. Ren, R. Zhang, Quantitative photoacoustic imaging of two-photon absorption, Journal of biomedical optics 23 (1) (2018) 016002.

[45] K. Ren, R. Zhang, Nonlinear quantitative photoacoustic tomography with two-photon absorption, SIAM Journal on Applied Mathematics 78 (1) (2018) 479–503.

[46] Y. Yamaoka, Y. Kimura, Y. Harada, T. Takamatsu, E. Takahashi, Fast focus-scanning head in two-photon photoacoustic microscopy with electrically controlled liquid lens, in: Photons Plus Ultrasound: Imaging and Sensing 2018, Vol. 10494, SPIE, 2018, pp. 170–174.

[47] B. E. Urban, J. Yi, V. Yakovlev, H. F. Zhang, Investigating femtosecond-laser-induced two-photon photoacoustic generation, Journal of biomedical optics 19 (8) (2014) 085001.

[48] W. Denk, J. H. Strickler, W. W. Webb, Two-photon laser scanning fluorescence microscopy, Science 248 (4951) (1990) 73–76.

[49] C. J. de Grauw, M. M. van Zandvoort, M. oude Egbrink, D. W. Slaaf, H. C. Gerritsen, Two-photon lifetime imaging of blood and blood vessels, in: Mul-

tiphoton Microscopy in the Biomedical Sciences, Vol. 4262, SPIE, 2001, pp. 171–176.

[50] P. T. So, Two-photon fluorescence light microscopy, eLS (2001).

[51] L. V. Wang, Photoacoustic tomography: Deep tissue imaging by ultrasonically beating optical diffusion, in: Make Life Visible, Springer, 2020, pp. 3–7.

[52] J. Ying, F. Liu, R. Alfano, Spatial distribution of two-photon-excited fluorescence in scattering media, Applied optics 38 (1) (1999) 224–229.

[53] R. M. W. W. R. Zipfel, W. W. Webb, Nonlinear magic: multiphoton microscopy in the biosciences, Nature Biotechnology 21 (2003) 1369–1377.

[54] S. G. Stanciu, S. Xu, Q. Peng, J. Yan, G. A. Stanciu, R. E. Welsch, P. T. So, G. Csucs, H. Yu, Experimenting liver fibrosis diagnostic by two photon excitation microscopy and bag-of-features image classification, Scientific reports 4 (1) (2014) 1–12.

[55] S. Park, J.-C. Vial, K. Kyhm, Optical sectioning in optical resolution photo acoustic microscopy, Optics Express 25 (16) (2017) 18917–18928.

[56] R. Mickens, Nonstandard Finite Difference Models of Differential Equations, World Scientific, Singapore, 1994.

[57] B. M. Chen, H. V. Kojouharov, Non-standard numerical methods applied to subsurface biobarrier formation models in porous media, Bulletin of Mathematical Biology 61 (4) (1999) 779 – 798.

[58] J. M.-S. Lubuma, K. C. Patidar, Contributions to the theory of non-standard finite-difference methods and applications to singular perturbation problems, World Scientific, 2005, Ch. 12, pp. 513–560.

[59] D. Dimitrov, H. Kojouharov, Nonstandard finite-difference methods for predator-prey models with general functional response, Mathematics and Computers in Simulation 78 (3) (2008) 1–11.

[60] J. Benz, A. Meister, P. Zardo, A conservative, positivity preserving scheme for advection-diffusion-reaction equations in biochemical applications, in: Proceedings of Symposia in Applied Mathematics, American Mathematical Society, 2009, p. 399.

[61] A. Suryanto, W. Kusumawinahyu, I. Darti, I. Yanti, Dynamically consistent discrete epidemic model with modified saturated incidence rate, Computational and Applied Mathematics 32 (2) (2013) 373–383.

[62] H. Obaid, R. Ouifki, K. Patidar, An unconditionally stable nonstandard finite difference method applied to a mathematical model of HIV infection, International Journal of Applied Mathematics and Computer Science 23 (2) (2013) 357–372.

[63] I. Martines, H. Kojouharov, J. Grover, A chemostat model of resource competition and allelopathy, Applied Mathematics and Computation 215 (2) (2009) 573–582.

[64] D. Dimitrov, H. Kojouharov, Analysis and numerical simulation of phytoplankton-nutrient systems with nutrient loss, Mathematics and Computers in Simulation 70 (1) (2005) 33–43.

[65] D. T. Dimitrov, H. V. Kojouharov, Nonstandard numerical methods for a class of predator-prey models with predator interference, Electronic Journal of Differential Equations 15 (2007) 67–75.

[66] D. Dimitrov, H. Kojouharov, Positive and elementary stable nonstandard numerical methods with applications to predator-prey models, Journal of Computational and Applied Mathematics 189 (1–2) (2006) 98–108.

[67] D. Dimitrov, H. Kojouharov, Stability-preserving finite-difference methods for general multi-dimensional autonomous dynamical systems, International Journal of Numerical Analysis and Modeling 4 (2) (2007) 280–290.

[68] D. Wood, D. Dimitrov, H. Kojouharov, A nonstandard finite difference method for n-dimensional productive-destructive systems, Journal of Difference Equations and Applications 21 (3) (2015) 240–254.

[69] D. Dimitrov, H. Kojouharov, Dynamically consistent numerical methods for general productive-destructive systems, Journal of Difference Equations and Applications 17 (12) (2011) 1721–1736.

[70] R. Anguelov, Y. Dumont, J. Lubuma, M. Shillor, Dynamically consistent nonstandard finite difference schemes for epidemiological models, Journal of Computational and Applied Mathematics 255 (2014) 161 – 182.

[71] R. Anguelov, J. Lubuma, Contributions to the mathematics of the nonstandard finite difference method and applications, Numerical Methods for Partial Differential Equations 17 (5) (2001) 518–543.

[72] J. M.-S. Lubuma, A. Roux, An improved theta-method for systems of ordinary differential equations, Journal of Difference Equations and Applications 9 (11) (2003) 1023–1035.

[73] D. Dimitrov, H. Kojouharov, Nonstandard finite-difference schemes for general two-dimensional autonomous dynamical systems, Applied Mathematics Letters 18 (2005) 769–774.

[74] R. Anguelov, T. Berge, M. Chapwanya, J. Djoko, P. Kama, J. M.-S. Lubuma, Y. Terefe, Nonstandard finite difference method revisited and application to the ebola virus disease transmission dynamics, Journal of Difference Equations and Applications 26 (6) (2020) 818–854.

[75] M. Gupta, J. Slezak, F. Alalhareth, S. Roy, H. Kojouharov, Second-order nonstandard explicit euler method, AIP Conference Proceedings 2302 (1) (2020) 110003.

[76] H. V. Kojouharov, S. Roy, M. Gupta, F. Alalhareth, J. M. Slezak, A second-order modified nonstandard theta method for one-dimensional autonomous differential equations, Applied Mathematics Letters 112 (2021) 106775.

[77] M. Gupta, J. Slezak, F. Alalhareth, S. Roy, H. V. Kojouharov, Second-order modified nonstandard explicit runge-kutta and theta methods for one-dimensional autonomous differential equations, Applications and Applied Mathematics: An International Journal (AAM) 16 (2021) 788–803.

[78] B. V. Limaye, Functional analysis, 2nd Edition, New Age International Publishers Limited, New Delhi, 1996.

[79] A. Ambrosetti, D. Arcoya, An introduction to nonlinear functional analysis and elliptic problems, Vol. 82 of Progress in Nonlinear Differential Equations and their Applications, Birkhäuser Boston, Ltd., Boston, MA, 2011.

[80] S. Kesavan, Topics in functional analysis and applications, John Wiley & Sons, Inc., New York, 1989.

[81] H. Brezis, H. Brézis, Functional analysis, Sobolev spaces and partial differential equations, Vol. 2, Springer, 2011.

[82] H. Antil, D. Leykekhman, A brief introduction to pde-constrained optimization, in: Frontiers in PDE-constrained optimization, Springer, 2018, pp. 3–40.

[83] D. T. Wood, H. V. Kojouharov, D. T. Dimitrov, Universal approaches to approximate biological systems with nonstandard finite difference methods, Mathematics and Computers in Simulation 133 (2017) 337 – 350.

[84] M. Gupta, R. K. Mishra, S. Roy, Sparse reconstruction of log-conductivity in current density impedance tomography, Journal of mathematical imaging and vision 62 (2) (2020) 189–205.

[85] B. J. Adesokan, K. Knudsen, V. P. Krishnan, S. Roy, A fully non-linear optimization approach to acousto-electric tomography, Inverse problems 34 (10) (2018) 104004.

[86] S. Roy, M. Annunziato, A. Borzì, A fokker–planck feedback control-constrained approach for modelling crowd motion, Journal of Computational and Theoretical Transport 45 (6) (2016) 442–458.

[87] E. J. Candes, J. K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences 59 (8) (2006) 1207–1223.

[88] G. Stadler, Elliptic optimal control problems with l 1-control cost and applications for the placement of control devices, Computational Optimization and Applications 44 (2) (2009) 159–181.

[89] T. F. Chan, S. Esedoglu, M. Nikolova, Algorithms for finding global minimizers of image segmentation and denoising models, SIAM journal on applied mathematics 66 (5) (2006) 1632–1648.

[90] L. I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms, Physica D: nonlinear phenomena 60 (1-4) (1992) 259–268.

[91] P. Perona, J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Transactions on pattern analysis and machine intelligence 12 (7) (1990) 629–639.

[92] O. Scherzer, J. Weickert, Relations between regularization and diffusion filtering, Journal of Mathematical Imaging and Vision 12 (1) (2000) 43–63.

[93] F. Catté, P.-L. Lions, J.-M. Morel, T. Coll, Image selective smoothing and edge detection by nonlinear diffusion, SIAM Journal on Numerical analysis 29 (1) (1992) 182–193.

[94] D. Gilbarg, N. S. Trudinger, D. Gilbarg, N. Trudinger, Elliptic partial differential equations of second order, Vol. 224, Springer, 1977.

[95] G. Alessandrini, V. Nesi, Univalent σ-harmonic mappings, Arch. Ration. Mech. Anal. 158 (2) (2001) 155–171.

[96] I. Ekeland, R. Témam, Convex analysis and variational problems, classics in appl, Math. SIAM, Philadelphia, PA, english ed (1999).

[97] S. Roy, M. Annunziato, A. Borzì, A fokker–planck feedback control-constrained approach for modelling crowd motion, Journal of Computational and Theoretical Transport 45 (6) (2016) 442–458.

[98] S. Roy, M. Annunziato, A. Borzì, C. Klingenberg, A fokker–planck approach to control collective motion, Computational Optimization and Applications 69 (2) (2018) 423–459.

[99] A. Schindele, A. Borzì, Proximal methods for elliptic optimal control problems with sparsity cost functional, Applied Mathematics 7 (9) (2016).

[100] A. Schindele, A. Borzì, Proximal schemes for parabolic optimal control problems with sparsity promoting cost functionals, International Journal of Control 90 (11) (2017) 2349–2367.

[101] M. Gupta, R. K. Mishra, S. Roy, Sparsity-based nonlinear reconstruction of optical parameters in two-photon photoacoustic computed tomography, Inverse Problems 37 (4) (2021) 044001.

[102] A. R. Fisher, A. J. Schissler, J. C. Schotland, Photoacoustic effect for multiply scattered light, Physical Review E 76 (3) (2007) 036604.

[103] Y. Yamaoka, M. Nambu, T. Takamatsu, Frequency-selective multiphoton-excitation-induced photoacoustic microscopy (mepam) to visualize the cross sections of dense objects, in: Photons Plus Ultrasound: Imaging and Sensing 2010, Vol. 7564, SPIE, 2010, pp. 592–600.

[104] Y. Yamaoka, M. Nambu, T. Takamatsu, Fine depth resolution of two-photon absorption-induced photoacoustic microscopy using low-frequency bandpass filtering, Optics express 19 (14) (2011) 13365–13377.

[105] Y. Yamaoka, T. Takamatsu, Enhancement of multiphoton excitation-induced photoacoustic signals by using gold nanoparticles surrounded by fluorescent dyes, in: Photons Plus Ultrasound: Imaging and Sensing 2009, Vol. 7177, SPIE, 2009, pp. 572–580.

[106] T. Vu, D. Razansky, J. Yao, Listening to tissues with new light: recent technological advances in photoacoustic imaging, JournaYul of Optics 21 (10) (2019) 103001.

[107] B. J. Adesokan, K. Knudsen, V. P. Krishnan, S. Roy, A fully non-linear optimization approach to acousto-electric tomography, Inverse problems 34 (10) (2018) 104004.

[108] M. Gupta, R. K. Mishra, S. Roy, Sparse reconstruction of log-conductivity in current density impedance tomography, Journal of mathematical imaging and vision 62 (2) (2020) 189–205.

[109] M. Li, H. Yang, H. Kudo, An accurate iterative reconstruction algorithm for sparse objects: application to 3d blood vessel reconstruction from a limited number of projections, Physics in Medicine & Biology 47 (15) (2002) 2599.

[110] E. Farnell, H. Kvinge, J. R. Dupuis, M. Kirby, C. Peterson, E. C. Schundler, Total variation vs l1 regularization: a comparison of compressive sensing optimization methods for chemical detection, in: Computational Imaging V, Vol. 11396, SPIE, 2020, pp. 64–76.

[111] L. C. Evans, Partial differential equations (graduate studies in mathematics, vol. 19), Instructor 67 (2009).

[112] H. Brezis, H. Brézis, Functional analysis, Sobolev spaces and partial differential equations, Vol. 2, Springer, 2011.

[113] H. Beckert, Oa ladyzhenskaya and nn ural'tseva, linear and quasilinear ellipltic equations.(mathematics in science and engineering, volume 46). xviii+ 495 s. new york/london 1968. academic press. preis geb., Zeitschrift Angewandte Mathematik und Mechanik 51 (2) (1971) 155–155.

[114] P. Kuchment, D. Steinhauer, Stabilizing inverse problems by internal data, Inverse Problems 28 (8) (2012) 084007.

[115] X. Zhang, Y. Xia, R. H. Friend, Single-photon pumping and two-photon probing spectroscopy for the determination of absorption cross-sections in an organic semiconductor, Optics Express 13 (26) (2005) 10873–10881.

[116] J. L. Sandell, T. C. Zhu, A review of in-vivo optical properties of human tissues and its impact on pdt, Journal of biophotonics 4 (11-12) (2011) 773–787.

[117] L. A. Shepp, B. F. Logan, The fourier reconstruction of a head section, IEEE Transactions on nuclear science 21 (3) (1974) 21–43.

[118] R. L. Harrison, B. F. Elston, D. W. Byrd, A. M. Alessio, J. Jacobs, R. C. Rockne, A. Hawkins-Daarud, M. Muzi, S. K. Johnston, P. R. Jackson, et al., A digital reference object for the 3d hoffman brain phantom for characterization of pet neuroimaging quality, in: 2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (2013 NSS/MIC), IEEE, 2013, pp. 1–4.

[119] R. Chamekh, A. Habbal, M. Kallel, N. Zemzemi, A nash game algorithm for the solution of coupled conductivity identification and data completion in cardiac electrophysiology, Mathematical Modelling of Natural Phenomena 14 (2) (2019) 201.

[120] H. V. Kojouharov, S. Roy, M. Gupta, F. Alalhareth, J. M. Slezak, A second-order modified nonstandard theta method for one-dimensional autonomous differential equations, Applied Mathematics Letters 112 (2021) 106775.

[121] M. Gupta, J. M. Slezak, F. K. Alalhareth, S. Roy, H. V. Kojouharov, (r1504) second-order modified nonstandard runge-kutta and theta methods for one-dimensional autonomous differential equations, Applications and Applied Mathematics: An International Journal (AAM) 16 (2) (2021) 1.

[122] M. Gupta, J. Slezak, F. Alalhareth, S. Roy, H. Kojouharov, Second-order nonstandard explicit euler method, in: AIP Conference Proceedings, Vol. 2302, AIP Publishing LLC, 2020, p. 110003.

[123] E. Yeargers, R. Shonkwiler, J. Herod, An Introduction to the Mathematics of Biology: Computer Algebra Models, Birkhauser, Boston, 1996.

[124] B. M. Chen-Charpentier, D. T. Dimitrov, H. V. Kojouharov, Combined nonstandard numerical methods for ODEs with polynomial right-hand sides, Mathematics and Computers in Simulation 73 (2006) 105–113.

[125] R. Anguelov, Y. Dumont, J.-S. Lubuma, M. Shillor, Dynamically consistent nonstandard finite difference schemes for epidemiological models, Journal of Computational and Applied Mathematics 255 (2014) 161–182.

[126] R. Anguelov, Y. Dumont, J.-S. Lubuma, M. Shillor, Comparison of some standard and nonstandard numerical methods for the mseir epidemiological model, AIP Conference Proceedings 1168 (2009) 1209–1212.

[127] A. Quarteroni, R. Sacco, F. Saleri, Numerical Mathematics, Springer-Verlag Berlin Heidelberg, Springer-Verlag Berlin Heidelberg, 2007.

[128] M. Mohammadi, A. Borzì, Analysis of the chang–cooper discretization scheme for a class of fokker–planck equations, Journal of Numerical Mathematics 23 (3) (2015) 271–288.

BIOGRAPHICAL STATEMENT

Madhu Gupta was born in Bahraich, Uttar Pradesh, India in 1991. She did her B.Sc. degree from Kisan Post Graduate College and earned degree from Dr. Ram Manohar Lohia Avadh University, Faizabad, India and M.Sc. degree in mathematics from Indian Institute of Technology, Bombay, India. Her research interests focus on nonlinear optimization frameworks for hybrid imaging problems, nonstandard finite difference methods and Fokker-Planck frameworks.